FOUNDATIONS OF REALISTIC RENDERING

A MATHEMATICAL APPROACH

MATHIAS M. LANG
COMPUTER GRAPHICS GROUP
SAARLAND UNIVERSITY
SAARBRÜCKEN, GERMANY

DISSERTATION ZUR ERLANGUNG DES GRADES
DOKTOR DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER NATURWISSENSCHAFTLICH-TECHNISCHEN FAKULTÄT I
DER UNIVERSITÄT DES SAARLANDES

DATUM DES KOLLOQUIUMS: —DATE OF COLLOQUIUM
28. AUGUST 2013—28. AUGUST 2013

DEKAN DER FAKULTÄT—DEAN OF THE FACULTY:
PROF. DR. MARK GROVES
UNIVERSITÄT DES SAARLANDES, SAARBRÜCKEN, GERMANY

PRÜFUNGSAUSSCHUSS—COMMITTEE:
VORSITZENDER—CHAIR OF THE COMMITTEE:
PROF. DR. JOACHIM WEICKERT,
UNIVERSITÄT DES SAARLANDES, SAARBRÜCKEN, GERMANY

BERICHTERSTATTER—REVIEWERS:

PROF. DR.-ING. PHILIPP SLUSALLEK,
UNIVERSITÄT DES SAARLANDES, SAARBRÜCKEN, GERMANY

DR.-ING. HABIL. KAROL MYSZKOWSKI,
MAX-PLANCK-INSTITUT FÜR INFORMATIK, SAARBRÜCKEN, GERMANY

PROF. DR. LÁSZLÓ SZIRMAY-KALOS,
TECHNICAL UNIVERSITY OF BUDAPEST, BUDAPEST, HUNGARY

AKADEMISCHER BEISITZER—ACADEMIC ASSITANT:

VINCENT PEGORARO, PH.D.,
UNIVERSITÄT DES SAARLANDES, SAARBRÜCKEN, GERMANY

BETREUENDER HOCHSCHULLEHRER—SUPERVISOR:
Prof. Dr.-Ing. Philipp Slusallek,
Universität des Saarlandes, Saarbrücken, Germany

GUTACHTER—REVIEWERS:
Prof. Dr.-Ing. Philipp Slusallek,
Universität des Saarlandes, Saarbrücken, Germany

Dr.-Ing. habil. Karol Myszkowski,
Max-Planck-Institut für Informatik, Saarbrücken, Germany

Prof. Dr. László Szirmay-Kalos,
Technical University of Budapest, Budapest, Hungary

DEKAN DER FAKULTÄT—DEAN OF THE FACULTY:
Prof. Dr. Mark Groves
Universität des Saarlandes, Saarbrücken, Germany

EINGEREICHT AM—THESIS SUBMITTED:
12. Dezember 2012—12. December 2012

Mathias M. Lang
Universität des Saarlandes
Lehrstuhl für Computergraphik, Geb. E 1 1
66123 Saarbrücken, Germany
mathias_lang@icloud.com

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Bliesransbach, 01.10.2012

x

IN MEMORY TO MY PARENTS, HERMINE AND MATHIAS LANG,
AND TO MY FRIEND JOHANNES, WHO DIED MUCH TOO EARLY

SUMMARY

The available doctoral thesis is not a usual paper but it is conceived as a text book for realistic rendering, made for students in upper courses, as well as for researchers but also interested people.

The textbook as such fulfills the principles of modern didactic theories, as it—in contrast to most of mathematical text books—demonstrates the necessary mathematical principles that are followed immediately by examples just behind their definition. It illustrates them and shows their practical use for realistic rendering. As an interdisciplinary hinge textbook which joins mathematics to realistic rendering as a branch of computer sciences, it attaches great importance to the exact mathematical formulation of the problem to be solve, i.e. of the formulation of the light transport equation in a vacuum as in participating media. Of course, it values highly the imperative mathematical tools and strategies needed therefore.

Mathematics is the science of essential importance for computer science in general, for the fast progress of its development in particular. That is a matter of fact that our book especially reflects, courting the favor and the motivation of mathematically interested readers and future researchers.

By its structure, it satisfies the needs of its readers: It is logically structured by chapters. So, first it mentions a problem and corresponding approaches for solving it. Furthermore it distinguishes the fundamental course contents from more special contents to consolidate. The contents are analyzed referring to their importance for applications. They are presented in a clear and scientific way.

From mathematical view, *realistic rendering* means solving the stationary light transport equation, a complicated *Fredholm Integral equation of* 2^{nd} *kind*. Its exact solution exists—if possible at all—only in an infinite dimensional functional space. Whereas implementation of approaches for solving problems are in the center of attention in the existing textbooks that treat global illumination theory, we are more interested in familiarizing our reader with the mathematical tools which permit them to formulate the global illumination problem in accordance with strong mathematical principles and last but not least to solve it.

New, more efficient and more elegant algorithms to calculate approximate solutions for the light transport equation and its existing variants must be developed in the context

of deep and complete understanding of the light transport equation. As the problems of realistic rendering are deeply rooted in different mathematical disciplines, the complete comprehension of all those areas must precede. There are evolving principles of functional analysis, theory of integral equations, measure and integration theory, as well as probability theory.

Let us consider for an example *Monte Carlo light tracing*. With knowledge on functional analysis Monte Carlo light tracing can directly be interpreted as the dual version of the standard algorithm of realistic rendering: *Monte Carlo path tracing*. The mathematical foundation for this is based on the concept of the adjoint of an operator equation from functional analysis.

Another example is Eric Veach's *path integral formulation*, an extremely elegant method which can be considered as the basis of a series of efficient rendering algorithms. Without the concept of the measure—commonly introduced in special courses on measure theory, which are rarely visited by students of computer science—the understanding, respectively the derivation of this elegant formulation is not conceivable. Even so fundamental concepts as the probability space or the random variable, defined as a special measure space, respectively, measurable function, require knowledge on σ-algebras, and countable or uncountable sets. A still deeper understanding of mathematical principles on measure and integration theory requires the study of continuous probability spaces. Since we search just in those probability spaces for solutions of the light transport equation, the handling with the concept of the *Lebesgue integral* is inevitable. Without knowledge on the Lebesgue integral and the underlying Lebesgue function spaces it is also not possible to make variance analysis, which gives statements on the quality of *Monte Carlo estimators*. Moreover, the Lebesgue integral serves—due to its properties, which let it become a much more powerful tool than the ordinary Riemann integral—as the modern integral notion on which the theory of integral equation is based on.

Let us also mention as a further example Eric Veach's *Metropolis light transport algorithm*, a rendering procedure, which can be used to simulate difficult lighting conditions in a scene to be rendered. It is based on the $\mathrm{M(RT)}^2$ algorithm—developed in the fifties of the last century—and adapted to the path integral formulation of light transport in a vacuum. This adaptation requires, apart from solid knowledge in probability theory, also knowledge from the theory of Markov processes, which are based on special transition kernels, constructs from measure and integration theory. Understanding the theory of Markov processes also requires a new machinery of convergence statements, which are not covered by the commonly presented limit theorems from probability theory. Finally, let us build a bridge to the *finite element method* based rendering algorithms. All radiosity methods are based on square Lebesgue-integrable function spaces, so-called Hilbert spaces, which possess very nice properties, and therefore justify the right usage of this fine rendering technique. We could extend this short list with other examples, such as the idea to replace random samples by deterministic chosen points in *quasi-Monte Carlo methods*, or the Neumann series approach for solving Fredholm type integral equations of the 2^{nd}

kind, which, if it is mentioned at all, in most lectures on realistic image synthesis simply falls from sky.

As we can already conclude from the above list of examples, realistic rendering is intertwined with many branches of mathematics. We have set ourselves the target, to remerge this bundle of fluff of mathematical concepts and principles, to represent them to the students in an understandable manner, and to give them, if required, exhaustive information. That is why our book is a new and unique approach.

ZUSAMMENFASSUNG

Die vorliegende Dissertation ist keine gewöhnliche Abhandlung, sondern sie ist als Lehrbuch zum *Realistisic Rendering* für Studenten im zweiten Studienabschnitt, für Forscher aber auch alle am Thema Interessierten konzipiert.

Das Buch als solches entspricht den Prinzipien moderner Lerntheorien, indem es—im Unterschied zu den meisten mathematische Lehrbüchern—die benötigten mathematischen Prinzipien in Anschluss an die Definition direkt an Beispielen aufzeigt, diese illustriert und auf ihren praktischen Nutzen für das realistische Rendering hinweist. Als fachübergreifendes Scharnierwerk zwischen Mathematik und realistischem Rendering legt es besonderen Wert auf die exakte mathematische Formulierung des zu lösenden Problems—nämlich der stationären Lichttransportgleichung in einem Vakuum aber auch in partizipierenden Medien—sowie der dazu verwendeten mathematischen Tools und Strategien.

Die Mathematik als Wissenschaft hat eine grundlegende Bedeutung für die Informatik im Allgemeinen, für deren schnell fortschreitende Entwicklung im Besonderen. Dieser Tatsache, die auch motivationalen Charakter für mathematisch interessierte Leser und zukünftige Forscher hat, trägt das Buch in besonderem Maße Rechnung. Mit seinem Aufbau antwortet es auf die Art und Weise der Vermittlung: Es stellt das oben genannte Problem und entsprechende Lösungsansätze in logisch aufeinander folgenden Kapiteln dar, unterscheidet das grundlegend zu Lernende vom Vertiefenden. Die Inhalte sind auf ihre Bedeutung für Anwendungen analysiert worden und werden sachgerecht präsentiert.

Aus mathematischer Sicht versteht man unter realistischem Rendering das Lösen der stationären Lichttransportgleichung, einer komplizierten Fredholm Integralgleichung der 2^{ten} Art, deren exakte Lösung, wenn überhaupt berechenbar, nur in einem unendlichdimensionalen Funktionenraum existiert. Während in den existierenden Bücher, die sich mit globaler Beleuchtungstheorie beschäftigen, meist vorwiegend die praktische Implementierung von Lösungsansätzen im Vordergrund steht, sind wir eher daran interessiert den Leser mit den mathematischen Hilfsmitteln vertraut zu machen mit welchen das globale Beleuchtungsproblem streng mathematisch formuliert und letzendlich auch gelöst werden kann.

Neue, effizientere und elegantere Algorithmen zur Berechnung zumindest approximativer Lösungen der Lichttransportgleichung und ihrer unterschiedlichen Varianten können nur im Kontext mit einem vertieften Verständnis der Lichttransportgleichung entwickelt

werden. Da die Probleme des realistischen Renderings tief in verschiedenen mathematische Disziplinen verwurzelt sind, setzt das vollständige Verständnis des globalen Beleuchtungsproblems Kenntnisse aus verschiedenen Bereichen der Mathematik voraus. Als zentrale Konzepte kristallisieren sich dabei Prinzipien der Funktionalanalysis, der Theorie der Integralgleichungen, der Maß- und Integrationstheorie sowie der Wahrscheinlichkeitstheorie heraus.

Betrachten wir als Beispiel Monte Carlo light tracing. Mit Kenntnissen aus der Funktionalanalysis lässt sich Monte Carlo light tracing direkt als duale Version des Standardalgorithmus im realistischen Rendering, nämlich Monte Carlo path tracing, interpretieren. Die mathematische Grundlage hierfür liegt nämlich im dem Konzept der Adjungierten einer Operatorgleichung aus der Funktionalanalysis.

Oder betrachten wir Eric Veach's Pfadintegralformulierung, eine äußerst elegante Methode auf der eine Reihe effizienter Rendering Algorithmen basieren. Ohne das Konzept des Maßes—gewöhnlich in speziellen Vorlesungen zur Maßtheorie eingeführt, die von Informatikstudenten fast nie besucht werden—ist das Verständnis bzw. die Herleitung dieser eleganten Formulierung nicht vorstellbar. Selbst solch fundamentale Konzepte wie die des Wahrscheinlichkeitsraums oder der Zufallsvariablen, definiert als spezieller Maßraum bzw. als messbare Funktion, erfordern bereits Kenntnisse über σ-Algebren, abzählbar unendliche und überabzählbar-unendliche Mengen.

Ein noch tieferes Verständnis mathematischer Prinzipien aus Maß- und Integrationstheorie erfordert das Studium stetiger Wahrscheinlichkeitsräume. Da wir gerade in diesen Wahrscheinlichkeitsräumen stochastisch nach Lösungen der Lichttransportgleichung suchen, ist der Umgang mit dem Konzept des Lebesgue Integrals unumgänglich. Ohne Kenntnisse über das Lebesgue Integral und die ihm unterliegenden Lebesgueschen Funktionenräume sind selbst Varianzanalysen, die Aussagen über die Qualität von Monte Carlo Schätzern liefern, nicht möglich. Zudem dient das Lebesgue Integral aufgrund seiner Eigenschaften, die es zu einem wesentlich mächtigeren Instrument werden lassen als das gewöhnliche Riemann Integral, als der moderne Integralbegriff, auf dem die Theorie der Integralgleichungen aufgebaut ist.

Erwähnen wir als weiteres Beispiel Eric Veach's Metropolis light transport Algorithmus, eine Rendering Methode, die sich zur Simulation schwieriger Lichtverhältnisse in einer Szene einsetzen lässt. Sie basiert auf dem $\mathrm{M(RT)}^2$ Algorithmus, entwickelt in den 50er Jahren des letzten Jahrhunderts, angepasst auf die Pfadintegralformulierung des Lichttransports im Vakuum. Diese Anpassung erfordert neben fundierten Kenntnissen in Wahrscheinlichkeitstheorie insbesondere auch Kenntnisse aus der Theorie der Markov Prozesse, denen spezielle Übergangskerne zugrunde liegen. Das Erfassen der Theorie der Markov Prozesse erfordert auch eine neue Maschinerie an Konvergenzaussagen, die über die gewöhnlichen Grenzwertsätze der Wahrscheinlichkeitstheorie nicht abgedeckt sind.

Schlagen wir letztendlich noch eine Brücke zu den auf der finite Elemente Methode basierenden Rendering Algorithmen. Hier sind es die quadratisch Lebesgue-integrierbaren Funktionenräume, die dem Prinzip aller Radiosity Algorithmen unterliegen.

Wie man bereits aus dieser Reihe von Beispielen schließen kann, so ist das realistische Rendering engverwoben mit vielen Bereichen der Mathematik. Wir haben uns zum Ziel gesetzt, dieses Knäuel an mathematischen Konzepten zu entflechten, sie für Studenten gegenüber verständlich darzustellen und ihnen bei Bedarf und je nach speziellem Interesse erschöpfend Auskunft zu geben. Unser Buch wählt erstmalig diesen grundlegenden interdisziplinären Ansatz.

ABSTRACT

The available doctoral thesis is not a usual paper but it is conceived as a text book for realistic rendering, made for students in upper courses, as well as for researchers and interested people.

From mathematical point of view, realistic rendering means solving the stationary light transport equation, a complicated Fredholm Integral equation of 2^{nd} kind. Its exact solution exists—if possible at all—in an infinite dimensional functional space. Whereas practical implementation of approaches for solving problems are in the center of attention in the existing textbooks that treat global illumination theory, we are more interested in familiarizing our reader with the mathematical tools which permit them to formulate the global illumination problem in accordance with strong mathematical principles and last but not least to solve it.

New, more efficient and more elegant algorithms to calculate approximate solutions for the light transport equation and their different variants must be developed in the context of deep and complete understanding of the light transport equation. As the problems of realistic rendering are deeply rooted in different mathematical disciplines, there must precede the complete comprehension of all those areas. There are evolving principles of functional analysis, theory of integral equations, measure and integration theory as well as probability theory.

We have set ourselves the target to remerge this bundle of fluff of mathematical concepts and principles, to represent them to the students in an understandable manner, and to give them, if required, exhaustive information.

KURZFASSUNG

Die vorliegende Dissertation ist keine gewöhnliche Abhandlung, sondern sie ist als Lehrbuch zum realistischen Rendering für Studenten im zweiten Studienabschnitt, sowie Forscher und am Thema Interessierte konzipiert.

Aus mathematischer Sicht versteht man unter realistischem Rendering das Lösen der stationären Lichttransportgleichung, einer komplizierten Fredholm Integralgleichung der 2^{ten} Art, deren exakte Lösung, wenn überhaupt berechenbar, nur in einem unendlichdimensionalen Funktionenraum existiert. Während in den existierenden Büchern, die sich mit globaler Beleuchtungstheorie beschäftigen, vorwiegend die praktische Implementierung von Lösungsansätzen im Vordergrund steht, sind wir eher daran interessiert, den Leser mit den mathematischen Hilfsmitteln vertraut zu machen, mit welchen das globale Beleuchtungsproblem streng mathematisch formuliert und letzendlich auch gelöst werden kann.

Neue, effizientere und elegantere Algorithmen zur Berechnung zumindest approximativer Lösungen der Lichttransportgleichung und ihrer unterschiedlichen Varianten können nur im Kontext mit einem vertieften Verständnis der Lichttransportgleichung entwickelt werden. Da die Probleme des realistischen Renderings tief in verschiedenen mathematischen Disziplinen verwurzelt sind, setzt das vollständige Verständnis des globalen Beleuchtungsproblems Kenntnisse aus verschiedenen Bereichen der Mathematik voraus. Als zentrale Konzepte kristallisieren sich dabei Prinzipien der Funktionalanalysis, der Theorie der Integralgleichungen, der Maß- und Integrationstheorie sowie der Wahrscheinlichkeitstheorie heraus.

Wir haben uns zum Ziel gesetzt, dieses Knäuel an mathematischen Konzepten zu entflechten, sie für Studenten verständlich darzustellen und ihnen bei Bedarf und je nach speziellem Interesse erschöpfend Auskunft zu geben.

MOTIVATION

Why this book has been written? At the whole beginning there was a USA journey, a vacation at the end of the 1990s in New York City. I rummaged in the bookstore *Barnes & Noble* for mathematical and computer science literature, as I have always been interested in problems of classical mathematics and computer science. I wanted to deal with parallelizing sequential algorithms, once a rising branch of computer science. Looking for good corresponding literature I found a reference book for computer graphics. Its title is *Radiosity and Global Illumination*, it is written by François Sillion and Claude Puech. Probably it was one of the first works that appeared in this area.

At that time I was not up to date on the development of computer graphics because twenty years ago I had begun to be active for the German software enterprise SAP, thus far away and cut off of the research. As many of my colleagues, as well as many laymen, I suited with great interest and fascination the rapid development which took computer science in the nineties. So, further development of the classic ray tracing procedure and radiosity methods permitted an incredibly fast progress when generating photo-realistic images: *Toy Story* and *A Bug's Life*, two animation films from the software manufacturer *Pixar* had been created. They exited millions of people. Their technology and their know-how to create images got me fascinated. I wanted to know how they were generated.

The famous *Toy Story* and *A Bug's Life* gave me the opportunity to estimate the standard of research for the industry: Without any doubt, the computer graphics community had made much progress. However films as the two mentioned above could easily be recognized as artificial—there were too hard shadows, there were obvious deficits, as for example in the representation of participating media such as smoke or fog.

However, in Sillion and Puech's *Radiosity and Global Illumination* I had already found very realistic illustrations, for example, the cover of the book, rendered with a radiosity method, which can not hardly be distinguished from a real photo. Obviously, with radiosity methods it is possible to represent any kind of light effects.

Radiosity and Global Illumination profited, in a much more extensive way than other reference books in computer science, of much more concepts from mathematical analysis, linear algebra, numerics, and probability theory. Also the connection of computer science to physics, particularly optics, and transport theory, was completely new.

If all these disciplines played a role for the further development of realistic rendering, then it should be possible—by means of deeper understanding of mathematical concepts—to produce more efficient, perhaps new, and rather exact procedures to produce realistic images. My interest in calculating and optimizing images of evidently little convincing *Toy Story quality* arose.

I would have liked to understand instantly at that time how to produce such pictures by computers. As a matter of fact, I lacked the special knowledge, imperative for the comprehension of rendering. So, I dropped my project.

Until, one year later, in my next vacation, I discovered, just in the same city and in the same bookstore, Andrew Glassner's *Principles of Digital Image Synthesis*. I enjoyed very much discovering there many background information referring to the development of rendering algorithms. I enjoyed less the great number of omissions or misprints in it. With the aim of understanding the principles of rendering, I read Glassner's treatise simultaneously to Sillion and Puech's *Radiosity and Global Illumination*.

The problem in computer graphics consists in resolving the light transport equation, which is a complicated integro-differential equation that describes the illumination of a point in a scene, in an exact or in an approximative way.

I mainly dealt with differential equations during my university studies, the concept of the integro-differential equation, which is based in functional analysis, is not used in computer science but rather in physics, where it helps to describe physical processes in mathematical terms.

The general approach to solve the light transport equation consists in transforming it to an integral equation in order to solve it numerically or stochastically. However, I wondered why not to try to transform the light transport equation into a differential equation: there are considerably more solution methods for solving a differential than an integral equation.

Once again I regarded the approach of the radiosity method, which consists in partitioning a scene to be rendered into a finite net of patches. This approach didn't go out of my head, thus I recognized a connection to my master thesis, in which I examined *algebraic multigrid methods* suitable to solve partial differential equations. To do this, we discretized the continuous problem and tried to solve the resulting system of equations on a much coarser grid than where it was defined with the help of numerical procedures. Then, the coarse solution was transformed via a special operator to the original grid, resulting in a good approximation to the exact solution.

Comparing the two procedures, the common approach to solve the light transport equation as an integral equation seemed to have some disadvantages. My own classical mathematical approach could help to solve better the light transport equation.

From my comparison, I concluded that we could perhaps consult *algebraic or geometric multigrid methods* for solving this equation. I thought we could apply a geometric or algebraic multigrid approach to the rendering equation. Since hierarchical radiosity

methods refine the original net, this procedure seemed to me very similar to the multigrid approach and would offer a solution strategy for the rendering equation.

Around the turn of the millennium I decided to contact Philipp Slusallek, who was building up the chair of computer graphics in Saarbrücken at that time and who considered my intellectual approach as interesting and worthy to research. As a specialist for global illumination he used to solve the rendering equation by the *Monte Carlo method*. So I began studying this method, too, building up a thesis entitled already: *Realistic Rendering—A Mathematical Approach*.

Many technical discussions followed between him as a PhD tutor and me as a doctoral candidate. Departing from the discussions originally focused on my thesis, there arose, by the time, the exchange between him as a college lecturer and me as a teacher at the high school. Gradually our subjects were converging to our common professional interest, namely our pupils respectively students. Remark that my pupils would be his students! We wanted to do common work to make benefit the next generation. We decided to exchange our experiences and to bundle our knowledge.

At the beginning we faced questions: Were my pupils pre-formed and trained adequately well when they came from high school to university? Were his students, who arrived at university after eight years of high school lessons, able to follow his lectures and to understand his exercises? Could we help them providing them with a demanding textbook but which fits the needs of very young persons? How should be looking such a textbook? In order to know how modern textbooks are made, I examined the available textbooks like [36, Cohen & Wallace 1993], [67, Glassner 1995], [185, Shirley 2000], [50, Dutré & al. 2003] and [158, Pharr & Humphreys 2004]. I noticed that they wouldn't need any innovation or improvement. They are excellent textbooks of computer science with the focus on practical implementation.

Our textbook should have another aim: It should be mathematically orientated, and it should follow a strongly mathematical structure. It should be exact and sufficiently thorough, it should be accessible and well understandable for young students. It should give them information, and it should be exhaustive if one's special interest requires more details. Of course, it should systematize and synthesize. In particular, our book should characterize functional analysis, measure, and integration theory, as well as probability theory as the three pillars on which the field of realistic rendering is mathematically based. The transport of light, in vacuum but also in participating media, should be built on a deeply anchored, mathematically based framework. So, each student should have the chance to develop his understanding for the mathematical perspective and the fundamental aspects of mathematics, the science on which are based all calculations in the field of realistic rendering.

If the students understand the areas of functional analysis, of measure, and integration theory as well as probability theory, they will find out one day other algorithms or completely new approaches suitable for arriving at even more exact approximate solutions to the light transport equation and their different variants, whether in stationary or

non-stationary form, expressed in terms of scalar or vector valued functions, or valid in a vacuum respectively in participating media.

The textbook is available in its first setting now. So, we hope that many students will benefit from our enquiries, our research, and compositions and that they will have a deep insight into the substantial mathematical bases for the solution of the light transport equation. As well we hope that they will enjoy representation, illustrations, and explanations.

STRUCTURE OF THE BOOK

Mainly due to the continuous increase in processing power of computers, in the meantime computer graphics has become one of the most dynamic areas of computer science. Keywords such as virtual reality, computer animated movies or pictures, and scientific visualization apparently show this trend. Thus, theoretical results, which—due to their huge computational and storage requirements—a few years ago are unthinkable for practical applications, determine the state of the art of the leading manufacturers of graphics chips and graphics-based software components.

In particular, this holds for the field of realistic rendering, the field of computer graphics dedicated to generate photorealistic images by using computers. Therefore, the knowledge of the methods and procedures of the underlying theory and systems, related technology and its applications is an important prerequisite for the successful career of any computer scientist, natural scientist, and engineer.

The present book is intended to familiarize the reader with the basic concepts, principles, but also with the most advanced techniques from the field of realistic rendering. Here, we are particularly interested in the mathematics behind it. Specifically we will present Monte Carlo rendering and radiosity methods as well as their recent evolutions for solving the so-called global illumination problem, central in the field of realistic rendering. It will always be our primarily interest to investigate the mathematics that carry these methods to obtain a solid and compact fundament for further research.

The very formal, theoretical, and mathematical structure of our book seems to require knowledge of mathematical foundations. So, for reading the book, prerequisites about functional analysis, as well as measure, integration, and probability theory are useful, but not required. All needed maths is included in Chapter 2 and the Appendix A. We only assume that the reader has any familiarity with calculus and linear algebra. Despite the fact that many branches of higher mathematics are covered by our book in a short and compact manner—in particular with respect to their significance for the field of global illumination—we are sure that an in depth study of Chapter 2 satisfies to provide the reader with the basic building blocks from so many branches of mathematics useful for a full understanding of photorealistic rendering. This should be achieved in particular due to the methodical and didactical framework of the book. Additionally, we expect from

reader a basic knowledge from some lectures about computer graphics and large interest to look behind the facades of rendering algorithms, in expectation to be supplied with the appropriate tools for developing new, more efficient rendering algorithms. So, our book is mainly designed for graduate students, but it should also be accessible to a wide spectrum of students as an accompanying textbook for a course at the graduate level. We hope that apart from graduate students, the researcher, and the computer enthusiast may also profit from the presentation.

The chapters of our book are organized as follows: At the begin of every chapter, within an introductory paragraph we familiarize the reader with the content of the chapter. Here, you find out how the chapter is organized, and often, the prerequisites required for understanding the content of the chapter are already summarized. In the whole book, important definitions, theorems, concepts, constructs, and principles are accessible over and over again via links in the marginal area of the text. This ensures at one hand that the knowledge, necessary to understand a chapter, at the appropriate time is present and on the other hand that the joy of the learner is maintained in spite of some abstract content.

Chapter 1 serves as an introduction. It presents the global illumination problem and should make aware the reader of mathematics that we will use to formulate and solve the integral equations underlying the global illumination problem.

In Chapter 2, we introduce the most important mathematical concepts that are needed to understand the principles of rendering and radiosity methods from field of computer graphics. This material is fundamental for those who are not familiar with functional analysis, measure, integration, and probability theory. The chapter is divided in four parts: In the first part, we introduce fundamental results from functional analysis that are needed to understand the theory and the derivation of solution methods of integral equation, such as: function spaces, linear operators, as well as linear and adjoint operator equations. The second part of Chapter 2 is devoted to the Lebesgue integral. It is the mathematical basis of many physical processes. Without this integral notation a strong mathematical treatment of many physical and stochastic problems is not possible. Afterwards, we talk about integral equations and the role they play in the field of global illumination theory. So, we present analytically based iteration methods for solving integral equations, and we will show how the concept of the adjoint can lead to new solution approaches for Fredholm type equations. Last but not least, the fourth part of the chapter contains an in-depth description of probability theory. Here, we introduce the concept of the probability space based on the notion of the measure space from measure theory. We present random variables as measurable functions from abstract probability spaces into the well-known Borel sets over $\mathbb{R}^n$, and we will show how the expected value and the variance of a random variable can be used to make statements about the convergence behavior of sequences of random variables. The chapter is finished with a small overview on Markov chains and Markov processes.

To familiarize the reader with the physical processes involved in the simulation of global illumination, Chapter 3 deals with the measurement of light. So, we will talk about

the transport of abstract particles and photons, and we give an overview of the most relevant concepts of radiometry used in the field of transport theory that are needed to formulate our light transport equations. With the radiometric quantity radiance, then we introduce the basic concept from radiometry on which all other radiometric quantities can be defined. Since photometry is closely coupled with radiometry, we also shortly talk about photometry.

Based on principles of transport theory, mathematical derivations of the various particle transport equations follows in Chapter 4. This chapter can also be seen as being divided into four parts. First, we derive the stationary particle transport equations in scalar form. The second part of Chapter 4 is devoted to the scattering behavior of light at object surfaces. Here, we will formulate these processes via a series of various mathematical functions, so-called bidirectional distribution functions. After that, we derive different formulations of the equations of light and importance transport. These are the integral equations to be solved using rendering algorithms based on Monte Carlo methods. Last but not least, we will present the measurement equation, the mathematical formulation of the global illumination problem which has to be solved in realistic rendering.

In Chapter 5 we will develop various mathematical models of light and importance transport. Here, we discuss in detail two models in detail. The first is based on the functional analytical concept of the operator, as a mapping between linear spaces. The other is the so-called path integral model of light transport, which is based on a measure theoretical approach. The entire chapter is of a very formal and theoretical character, since we try to attempt, for the first time in the field of global illumination, to define all needed concepts and constructs in a strict mathematical manner. So, we will exactly describe the underlying function spaces and operators, and we will also show the construction of the path measures in more detail, starting with the σ-algebras on which they are defined. Although this approach is of a very theoretical nature, we are convinced that it is the right way to a complete understanding for the mathematical models of light transport, that have delivered so many fruitful ideas for solving the light transport equations until today.

Chapter 6 and Chapter 7 are devoted to Monte Carlo and quasi-Monte integration. These probabilistic as well as deterministic methods can be used to solve the light transport equations—which are all Fredholm integral equations of the second kind—involved in various rendering and radiosity methods. First we show why common numerical integration is not a suitable method for solving high dimensional integrals, but why Monte Carlo method should be used. Then we introduce the concept of the Monte Carlo estimator as an approximate solver of high dimensional integrals. We talk about the convergence of Monte Carlo methods and present some techniques for sampling random numbers, which are crucial in the construction of efficient Monte Carlo algorithms for the evaluation of integrals. So, we will present the transformation methods, talk about acceptance-rejection sampling, and give an insight into Markov chain Monte Carlo methods. We also discuss a series of variance reduction techniques to be able to construct fast evaluable estimators

with a variance as small as possible, and last but not least, we show how Monte Carlo methods can be used to solve Fredholm integral equations of the 2^{nd} kind. With respect to quasi-Monte Carlo methods we declare the concept of discrepancy as a measure for the derivation of a point set from its ideal distribution. Afterwards, we present the construction of so-called low-discrepancy sequences, such as Halton, Hammersley, and Zaremba sequences, or (t, s)-sequences, and (t, m, s)-nets. With the help of the Fourier analysis, we then compare the quality of sampling patterns resulting from low-discrepancy sequences or sampling strategies form Monte Carlo methods.

Chapter 9 focuses on the classic ray based rendering algorithms. Here, we will start with ray casting and classic Whitted-style ray tracing, and we will discuss in more detail distribution ray tracing, the first on stochastic principles based rendering algorithm.

In Chapter 9, we talk about Monte Carlo rendering algorithms for solving Fredholm type integral equations—here in particular the light transport equation in free space. We present Monte Carlo path tracing, and the associated dual algorithm Monte Carlo light tracing and show that both procedures can be simulated via discrete-time, continuous-space Markov processes. After that, we introduce bidirectional path tracing, in some sense a combination of path tracing and light tracing. With the Metropolis light transport, we present a first Markov chain Monte Carlo approach for solving the light transport problem. Last but it not least, we will discuss photon mapping, a very efficient two-pass algorithm, and we will give a short overview about instant global illumination, which allows to simulate the most important illumination effects at realtime rates.

In Chapter 10, we then deal with radiosity methods. Here, starting with the rendering equation, we derive the radiosity equation and present the classical radiosity algorithm. Then, we will discuss the concept of form factors, present properties of form factors, and show how form factors can easily and efficiently be computed. Furthermore, we present techniques from numerical analysis that can be used to compute an approximative solution to the radiosity integral equation. The chapter will be finished with a short review on the general finite element radiosity approach.

For this students that are not familiar with basic calculus and basic linear algebra, the book also contains, in the Appendix A a complete refresher to basic concepts from linear algebra and calculus which are absolutely needed to understand the mathematical foundations of realistic rendering. Students familiar with basic calculus and linear algebra can skip the appendix since it must be considered as an introductory preparation for the main sections of the book.

We have not included separate sections with exercises to practice and deepen the discussed materials, as many of these things are already included in the numerous examples. Nevertheless, we give some useful homework assignments to the interested reader at different places within the book.

In order to not destroy the flow of reading, we have summarized the sources of our workouts in reference literature and further reading sections. Here, you can also find many

hints and further sources of literature that deal with the content of the chapter and serve to enhance the discussion in the chapter.

Let us still mention that a version of the book will be soon available as an eBook.

How to Use this Book

Both, a textbook or a lecture, are constructed as good literature trying to achieve one goal: They are designed to spark interest in the reader or listener and to keep it. The last-mentioned are requested to deepen the subject by their intrinsic motivation. Nowadays, lectures are mainly presented with the help of slides created on a computer. The presentation of a lecture with the help of slides—the voice of the speaker at the same time for the visual representation as a help, as a second channel of reception—is an excellent technique for stage-managing, in particular, if it is supported by a textbook, where the material is deepened and explained.

Our book is designed to be used either as a textbook for an advanced course on realistic rendering or as a very useful mathematical supplement to all current standard texts in global illumination, such as [51, Dutré & al. 2006] and in particular [159, Pharr & Humphreys 2010]. It is suitable for a one-semester course meeting four hours per week.

In the following, we give a coarse schedule of nearly 20 lectures. There is also sufficient time to talk about more recent methods which can be applied if the book is used as basis of an advanced course in realistic image synthesis:

- An advanced course in realistic image synthesis should begin with a short overview of the topic, followed by a short review on the most useful mathematical concepts from linear algebra and calculus.
- Then, we recommend talking about linear function spaces, here in particular on the function spaces of great relevance for the concerns of global illumination theory, that is, the ray spaces.
- With the notion of the function space at hand, the functional analytical concepts of the linear operator, the linear operator equation, and the adjoint equations should be discussed.
- A deeper discussion should be assigned to the concepts of the integral and the integral equation. As the Lebesgue integral can be derived intuitively, as we did in Section 2.2.1, the Fredholm integral equation, its representation as linear integral operator equation and the associated solution techniques should be discussed in more detail.

- Probability theory must be our next topic to be dealt with: probability spaces, random variables, and stochastic processes.
- Afterwards, the field of radiometry could build the basis for a further lecture, i.e. counting photons, and introducing the radiometric quantities: flux, radiance, irradiance, radiosity, and radiant intensity.
- In the next larger block a mathematical formulation of the stationary light transport equation can be derived. Here, we recommend starting with the derivation of the stationary particle transport equation in integro-differential form, followed by the process of transforming this integro-differential equation into an integral equation.
- As the concept of the BRDF is central for the field of global illumination, we suggest to talk about the BRDF in more detail. So, the BRDF could be derived from the concept of the BSSRDF. Afterwards it is just the right time to talk about physical properties of the BRDF and the BTDF, and to introduce the concepts of reflectance and transmittance. Also the representation and the measurement of BRDFs could be mentioned.
- Then, the most frequently used BRDF models in computer graphics, that is, the idealized BRDF models, as well as the most known examples of phenomenological and physical based BRDF models, should be presented. Considering the light transport in participating media, we cannot ignore the concept of the phase function.
- The block about a mathematical formulation of the light transport equation can be finished by introducing the different formulations of the light transport equation, that is, the light transport equation valid in participating media respectively valid in a vacuum, in its spherical and 3-point forms. The importance and the measurement equation must also be introduced.
- For its mathematical demands, the block on deriving a mathematical model of light and importance transport is quite hard to understand. So, we recommend beginning with a discussion of the light transport in a vacuum. Then, the importance transport can simply be derived via the adjoint equation. If the students interest lies also on the light transport in participating media, a mathematical model of light transport in participating media can be discussed.
- The sections 6.2 and 6.3 are fundamental for the understanding of Monte Carlo methods. So, we recommend transforming the content of these two sections in one lecture. Additionally, the convergence of the Monte Carlo methods should be discussed.
- As a Monte Carlo algorithm succeeds or fails according to the chosen samples, the most efficient sampling techniques must be presented, that is, we have to talk on the transformation method, acceptance-rejection sampling, and—if the MLT algorithm will later be discussed—also the Metropolis sampling algorithm.

- As they are fundamental for Monte Carlo methods applied to the light transport equation, the most promising variance reduction techniques—use of expected values, importance sampling, control variates, stratified sampling, as well as LHS and Jittered sampling—should be presented. Here we recommend in particular the application of these strategies for solving the rendering equation.
- Multiple importance sampling can then be presented in a separate lecture.
- With a detailed discussion on the application of Monte Carlo methods for solving Fredholm integral equations of the 2^{nd} kind, the block on Monte Carlo integration can be finished.
- If the student is interested in quasi-Monte Carlo methods, the concept of discrepancy is necessary, the classical constructs of Halton sequence and Hammersley points should be shown, and more advanced techniques such as the (t, m, s)-nets and (t, s)-sequences as well as their randomized variants can also be given. Last but not least, Fourier analysis as a technique for comparing sampling patterns can be discussed.
- The classic rendering algorithm based on the principle of ray tracing, i.e. ray casting, Whitted-ray tracing, and distribution ray tracing should be the topic of a more practically oriented lecture.
- Afterwards, the Markov process based rendering algorithms can be discussed in more detail. So, we recommend talking intensively on pure-Monte Carlo ray tracing and Monte Carlo ray tracing with next event estimation. The dual algorithm can be omitted, or it can be shortly mentioned. Bidirectional path tracing in connection with the path integral formulation—if it was not introduced when deriving a model of light transport—should also be discussed in more detail now.
- Afterwards, we recommend discussing the Metropolis light transport algorithm, followed by the photon-mapping algorithm, and instant global illumination.
- The block on finite element based rendering algorithms can be opened via the derivation of the classical radiosity formulation and its associated discretization. Then, the concept of the form factor should be introduced. The lecture should also cover talking about properties of form factors and characterizing solution strategies.
- Finally the classical relaxation methods must be addressed as solution techniques for the discrete radiosity equation. This last block should be finished with a short review of the finite element radiosity approach and a short comparison of image and object-based rendering procedures.

ACKNOWLEDGEMENTS

This thesis would not have been possible unless Prof. Dr.-Ing. Philipp Slusallek, who has made available his support in a number of ways. I came to him as external and asked him to supervise a thesis. He agreed although not knowing me, and he accepted me in the inner circle of his staff members. He involved his colleagues in the mentoring of this thesis and gave me the opportunity to map his lecture and even to lecture parts of it at his place.

He cared about interesting his students in my work as well as he cared about interesting me in the work of his students. So that is how it came about my first thesis version became a textbook: Future generations of students should benefit from it.

My original idea was to develop a geometric and even algebraic multigrid approach for solving the global illumination problem, but Philipp countered and convinced me to deal with the mathematical principles beyond Monte Carlo methods rather from a pedagogical view, as a more appropriate approach that ideally fits the conditions of young students.

Philipp met me very often and very regularly, in order to discuss and to reconsider my thesis, to agree about further proceeding and not least to give me advices. He is responsible, too, for the inductive method I chose structuring the resulting book. Without him, I would have preferred a deductive procedure, i.e. a structure as it is usually known from books on mathematics.

Finally he helped me with his comprehensive knowledge of physics, and brought me up-to-date with regard to the recent trends and evolutions in computer graphics.

I am indebted to him for all this support, but not last for reading tirelessly the single chapters and parts of chapters, for structuring and restructuring this work, aiming at a didactically optimal result. Thanks for deskewing and rectifying some statements that had been to complicatedly described, and, of sure, for the corrections of my language faults.

His friendly and constructive attitude bore me even in charged periods of finishing.

I am grateful, too, to his colleague and agent, Dr. Vincent Pegoraro, who read regularly parts of my work, he revised and adapted a series of formulations.

I would like to show my gratitude to Philipp's secretary, Léa Schäfer, who reached our agreements or if necessary coordinated and arranged new dates.

Furthermore I would like to thank my teacher colleagues at the Marienschule, Holger Christmann, Uli Jager, and Gabriele Piro-Johanns, who asked for the progress in my

thesis, who attended me friendly and answered gently to my questions concerning English grammar and syntax.

Finally I would thank my brother Johannes, who gave me support in perspective drawing. He also had the idea for the cover of the book version and designed it for me.

Last but not least I owe my deepest gratitude to my wife, Maria, who supported me in times of doubt and cared all these years for me, so that I could work independently.

Contents

List of Figures

List of Tables

CHAPTER ONE

INTRODUCTION

From a mathematical point of view, *realistic rendering* is equivalently to solving the so-called *global illumination equation*, also denoted as the *light transport equation*, and in computer graphics better known as the *rendering equation*. This equation describes the light distribution at all points and in all directions within a scene to be rendered. As solution of a Fredholm type integral equation, the unknown light distribution must be interpreted as a continuous function living in an infinite-dimensional function space and occurring in an equation inside and outside of an integral.

Until today a lot of work was put into the study of solutions to discrete versions of the light transport equation—in particular into Monte Carlo and finite element methods for solving the rendering equation—but only little is known about the continuous equation beyond the existence and uniqueness of its solution as well as the interaction of the different mathematical disciplines that deal with the topic. From a practical point of view, lying the focus of the research to finding solution methods for discrete versions of the rendering equation should be justified, since computers, as finite-space and finite-state machines, cannot represent an infinite-dimensional solution in finite time. Nevertheless, new, more efficient, and elegant algorithms for finding finite-dimensional approximate solutions to the light transport equation can only be fully understood in the context of the continuous global illumination problem and its interaction with the corresponding fields of mathematics.

Here many branches of mathematics, such as functional analysis, measure, and integration theory as well as probability theory come into play. In functional analysis—as the study of algebraic and topological properties of abstract spaces, particularly infinite-dimensional function spaces—a problem such as the light transport equation is first reformulated as an operator equation in an abstract infinite-dimensional function space. In case of the light transport equation, this transform leads to a linear integral operator equation. Functional analysis then provides a series of theoretical solution approaches that can be implemented on a computer to deliver a practically usable solution. Such algorithms are mainly based on *quadrature* or *projection methods*. Now, in case of the light transport equation, where the resulting operator equation is of such a high dimension, ordinary analytic solution methods are often less suitable due to efficiency reasons. Here, stochastic

approaches lead to better results. This means, that the analysis and the derivation of already existing or new algorithms requires apart from a deeper insight into the *theory of solving integral equations* from functional analysis also requires a deeper insight into the *probability theory*. As the theory for solving integral equations, it is—similar as probability theory—based on the concept of the *Lebesgue integral*. So, we cannot circumvent also to talk about *measure* and *integration theory*.

As the main goal of any realistic rendering algorithm is the creation of physically accurate synthetic images from complete scene descriptions, apart from the large field of mathematics, also many concepts are required from physics, and here in particular, from optics and radiometry. It is the goal of this book to cover all these requirements, which are used for a deeper understanding of the various light transport equations and corresponding solution techniques, in a clear and strictly mathematically based manner. So, as a starting point into the study of realistic rendering, we will give the reader in this introductory chapter a short overview about the content of our book, that is,

- we will formulate the global illumination problem,
- present first algorithms for solving the underlying light transport equation, and
- address shortly the mathematics as their fundamental scientific base.

Section 1.1 OVERVIEW OF THIS CHAPTER. We begin the chapter with some opening remarks to *local* and *global reflection models*. Then, we introduce the *global illumination problem*, and present with *classic Whitted-style ray tracing* a first ray based algorithm for finding approximate solutions to simple variants of the global illumination problem. Based on
Section 1.2 functional analytical approaches, we present two different classes of numerical methods for solving or at least approximate solving the global illumination problem: the *Neumann series approach* and the *finite element method*. Afterwards, we introduce from each of these two classes the most promising solution algorithm for the global illumination
Section 1.3 problem. This will be *Monte Carlo path tracing*, a stochastic solution procedure, based on the principle of the *random walk*, and the *radiosity method*, an algorithm that has its origin into the finite element method.

1.1 REALISTIC RENDERING

Quite generally considered, realistic rendering is a field of computer graphics where techniques are developed for generating photorealistic images of real objects or scenes by means of a computer—see Figure 1.1. As a basis for realistic rendering, the modeling of interaction between light and object has emerged. This modeling requires apart from the exact geometric description of the scene, and an associated visibility tool also a so-called *illumination model*, that is, a procedure for computing the color of a point on the image plane.

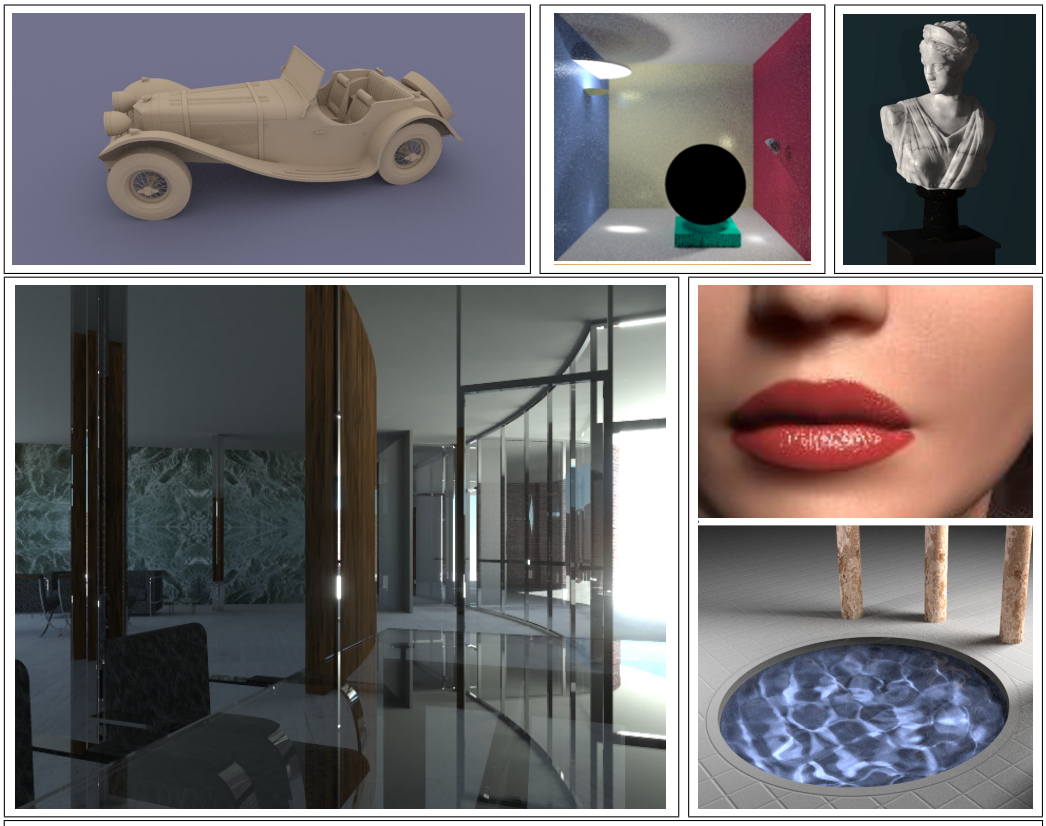


FIGURE 1.1: REALISTIC RENDERING. Algorithms from the field of realistic rendering, such as the Markov process based procedures of Monte Carlo path tracing, Monte Carlo light trancing, bidirectional path tracing, or photon mapping, can simulate all phenomena of light within a scene.

The model of the Jaguar in the upper left image consists of 140,000 triangles and was rendered with Monte Carlo path tracing. The image in the middle is rendered with Monte Carlo light tracing, the dual algorithm to Monte Carlo path tracing. Image courtesy of Frank Suykens-De Laet, Department of Computer Science, KU Leuven. The upper right image, rendered with the photon mapping algorithm shows a translucent marble bust, *Diana the Huntress*. It illustrates how the soft appearance of marble can simulated via subsurface scattering. The lower left image is from the animation "The Light of Mies van der Rohe" and demonstrates how photon mapping can be used to compute global illumination in a complex model. The face model was rendered with photon mapping using a BSSRDF that does account for subsurface light transport, e.g. light enters and leaves at different locations on the surface and gives the skin a more natural translucent appearance. The last image shows a pool of water with small waves and caustics at the ground. The image was rendered with MLT, a probabilistic rendering algorithm based on the Metropolis algorithm, image courtesy of Eric Veach, Pixar. All other images are courtesies of Henrik Wann Jensen, UCSD.

1.1.1 A BIT ABOUT LOCAL AND GLOBAL ILLUMINATION

In computer graphics, an illumination model can be interpreted as the description of the geometry and the reflection behavior of the surfaces within a scene together with the influence made by existing light sources due to their position, size, and other properties. That is, apart from the requirement, that it is easily be computable, an illumination model has also to describe the process of light propagation as well as the interaction of light at object surfaces. Here, we distinguish between two types of illumination: *local* and *global* illumination.

LOCAL ILLUMINATION. Introduced in the early seventies of the last century, the idea behind a local illumination model is to generate images with a computer as realistic as possible but with minimal effort. Thus, local illumination models are not based on strict physical regularities. As the name already suggests, in a local illumination model the focus lies on *local illumination*. That is, for shading a surface point, we consider only the light that comes directly from existing light sources within a scene.

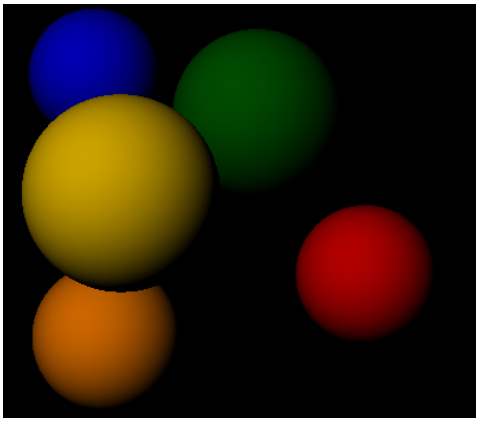
DEFINITION 1.1 (Local Illumination) *The contribution of light that arrives directly at a surface point* s *from a light source and is reflected at* s *is called* local illumination. *That is, local illumination corresponds to the single-light, single-surface interaction of light, where the shading of point* s *on any surface is independent from the shading of all other surfaces.*

Due to the above definition, local illumination only takes into account the relationship between light sources and a single object, it does not consider the effects that result from the presence of multiple objects. Thus for example, if a light source is blocked by another object, in a local illumination the surface under observation does not contain light from this source, although light can be contributed to it due to reflection from some other objects.

Similar to the approach of many modern sciences, computer graphics achieved to consolidate the initially made good progress in a practical and usable local illumination model, the *Phong illumination model*, see Figure 1.2. Although the Phong model does not describe a physically plausible reflection model—it is a phenomenological illumination model, which often reflects more light than it receives—the Phong model has become the most commonly used illumination model in computer graphics until today.

Phong Illumination Model (351)
Section 4.2.2.5

To make the results of a local illumination model appear more realistic, the indirect reflected, transmitted, or scattered fraction of light within a scene is represented by a so-called *ambient term*, which is assumed to be constant for all points of the scene, see Figure 1.2. Now, as this term is constant, it cannot account for the positions of objects to the observer nor other objects that block the light coming from light sources or neighboring objects. As direct illumination in some situations contributes only a small fraction to the measured amount of light at a point in a real scene, the embedding of an ambient



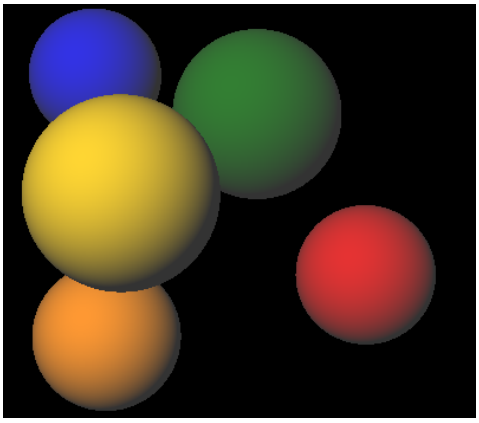


FIGURE 1.2: PHONG SHADING. A set of spheres rendered with the local illumination model by Phong, for details see Section 4.2.2.5.2. To make the results of Phong reflection model appear more realistic, the indirect reflected, transmitted, or scattered fraction of light within a scene is represented by a constant term: the ambient light.

term in a local illumination model for simulating global light interaction delivers only unsatisfactorily result.

GLOBAL ILLUMINATION. Another problem in local illumination models is the generation of shadows. As local illumination takes into account only the interaction of direct light with the objects in a scene, the generation of shadows, a phenomenon of global interaction, is not involved in any local illumination model. The light intensity in a shadow area can be determined only by global interaction, as such areas are generated by indirect illumination, thus indirectly reflected, transmitted, or scattered light from objects in the scene.

DEFINITION 1.2 (Global Illumination) *Apart from the local illumination at a surface point* $\mathbf{s}$, global illumination *accounts for shading a point* $\mathbf{s}$ *also the light reflected, transmitted, or scattered between all objects within a scene. Thus, global illumination models the interchange of light between all surfaces within a scene model.*

Due to the above definition, a global illumination model combines the light resulting from local reflection with the light that is reflected or refracted from other surfaces to the current surface. Thus, a global illumination model consists of a local illumination model and a process that gathers the light incoming at an observation point due to multiple reflections onto objects in a scene. This means, that an illumination model, based on global illumination, is more comprehensive, more physically correct, and produces more realistic images, but it is also more computationally expensive. Chapter 8

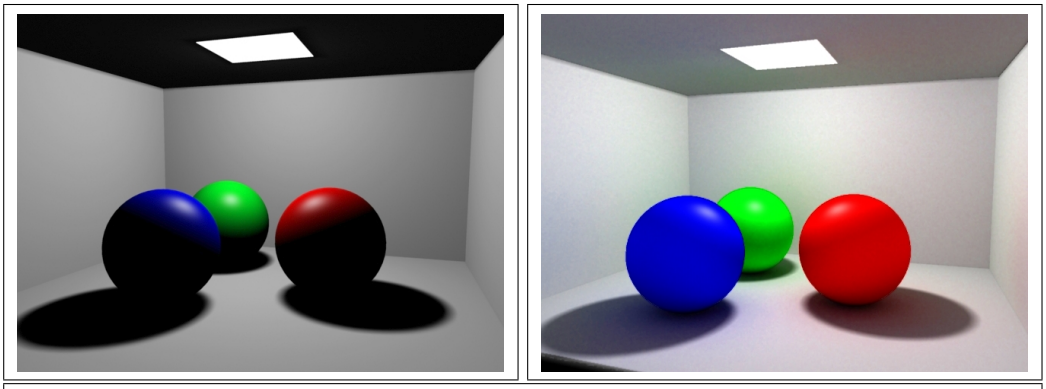


FIGURE 1.3: LOCAL ILLUMINATION VS GLOBAL ILLUMINATION. A set of spheres rendered with a local and a global illumination model. While local illumination models only consider the interaction of light with objects in a scene coming directly from light sources, global illumination models take into account also the light reflected, transmitted, or scattered between all objects in a scene.

1.1.2 THE GLOBAL ILLUMINATION PROBLEM

In computer graphics, one understands under the term of *global illumination* the computation of light distribution at all points within a 3-dimensional virtual scene model. Such a scene model is usually constructed via a very large set of surface primitives like triangles, polygons, spheres, etc. Since our goal is photorealistic rendering, it is required that a virtual scene model must also be described in a realistic way as much as possible. This means, that apart from the geometry of the model, we must also specify the types of existing light sources, the optical and physical properties of the scene objects, thus the color, nature, and reflectance behavior of the object surfaces, as well as the type of the detector used to produce an image of the scene. We call such a specification the *global illumination problem*. Based on [104, Keller 1998], [208, Szirmay-Kalos 1999], we define the global illumination problem mathematically as follows:

$\mathcal{V}$ (41) Radiance (250) Section 4.3 Section 4.2 Section 4.5

DEFINITION 1.3 (The Global Illumination Problem) *Let $\mathcal{V}$ be the union of a finite number of 3-dimensional volumes within $\mathbb{R}^3$, L_e be the radiance, loosely speaking the light emitted from a finite set of light sources existing in the scene—thus points, surfaces, or volumetric light sources—furthermore let $\bigcup_{j=1}^{m} f_{s_j}$ be a set of functions characterizing the color, roughness, and reflectance behavior at the surfaces of volumes from $\mathcal{V}$, or at volumetric points, and let W_e represent the specification of a detector. Then, the* global illumination problem *is given by the quadruple*

$$\left(\mathcal{V}, L_e, \bigcup_{j=1}^{m} f_{s_j}, W_e \right). \tag{1.1}$$

Now, the main task in computer graphics is to find a solution to the global illumination problem, in other words, the construction of an image that visualizes a scene, specified by the global illumination problem.

Solving the global illumination problem can be considered as a two step process: First, we have to determine which region of the scene is relevant for the final image. If it is clear, which part of the scene should be mapped to the film plane, then, we can compute the illumination at all points within this region.

As we will see in the following chapters, the solution of the global illumination problem can mathematically be reduced to the evaluation of the so-called *measurement equation*,

Measurement Equation (416)

$$\mathcal{M} \stackrel{\text{def}}{=} \langle W_e(\mathbf{s},\omega), L_i(\mathbf{s},\omega)\rangle. \tag{1.2}$$

Defined as an inner product, it uses on the one hand the specification of the detector, W_e, and on the other hand the illumination, $L_i(\mathbf{s},\omega)$, at all points relevant for the final image. Combining these two quantities then results in the final image.

Inner Product (859)
W_e (416)

Now, the detector specification, occurring in the measurement equation, is already given via the global illumination problem $\left(\mathcal{V}, L_e, \bigcup_{j=1}^{m} f_{s_j}, W_e\right)$, but the illumination L_i at points $\mathbf{s}$ in direction ω is still unknown. That is, solving the measurement equation requires information about the illumination at points which are relevant for the image. This information can be achieved by solving—depending on the specification of the global illumination problem—one of the so-called *stationary light transport equations*. These equations, that record the light distribution within a scene, can be expressed in form of *linear Fredholm integral equations of the* 2^{nd} *kind*, such as,

L_i (250)
Light Transport Equation (295)
Section 2.3

$$L_i(\mathbf{x},\omega_i) = \beta(\mathbf{s}\to\mathbf{x})\,\epsilon_b(\mathbf{s},\omega_o) + \int_{[0,\|\mathbf{s}-\mathbf{x}\|]} \beta(\mathbf{x}'\to\mathbf{x})\,Q_o(\mathbf{x}',\omega_o)\,d\mu(\alpha), \tag{1.3}$$

and are also often denoted as *global illumination equations*. They describe the light incident at a point $\mathbf{x}$ coming from direction ω_i by means of two terms. The first term on the right hand side of Equation (1.3) describes the light incident at point $\mathbf{x}$ coming from the nearest surface point $\mathbf{s}$ in direction ω_i, while the second term describes the light, that comes from volumetric points of a participating medium lying on the straight line to the surface point $\mathbf{s}$, see Figure 1.4

REMARK 1.1 *Usually the radiometric quantity radiance, denoted by* L, *corresponds to a vector from the RGB-color system, consisting of components* $(R,G,B)^T$, *which correspond to the intensities for the selected wavelengths of red, green and blue light, used for firing the electron guns of a CRT. From our discussions about vector spaces in the next chapter, it will be clear that this system could easily be replaced by any other color system without changing the algorithms that are described in the following. Therefore, and sake of simplicity we use radiance in all of our equations and algorithms as a scalar quantity, which naturally implies, that, whenever we write an equation using* L, *this equation must be interpreted in the sense that it only holds for a single component of the color base.*

Radiance (250)

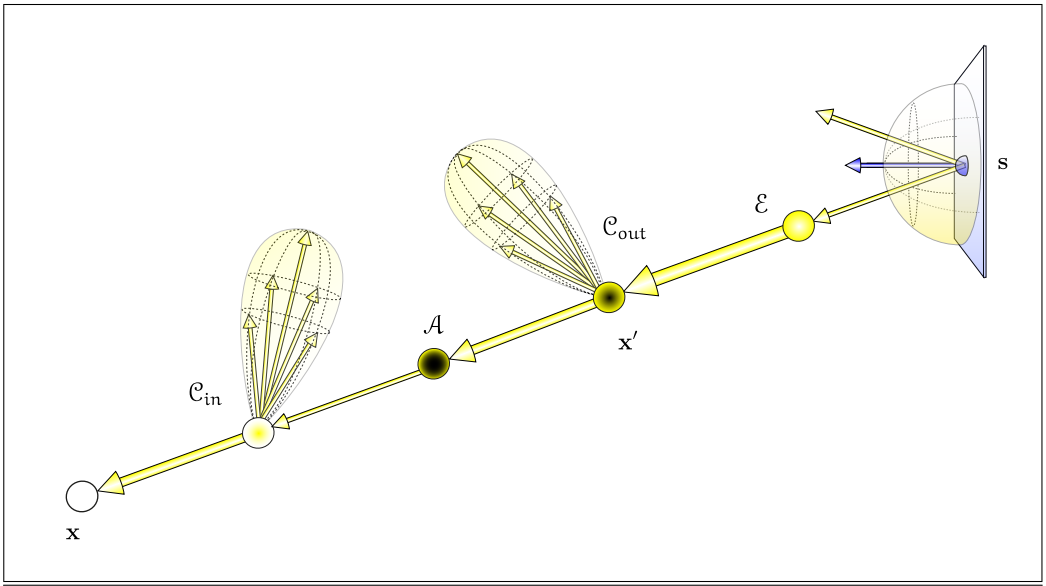


FIGURE 1.4: VISUALIZATION OF THE STATIONARY LIGHT TRANSPORT WITHIN A SCENE. During its travel through a scene light is subjected to different phenomena. This can occur at small particles within a participating medium or at object surfaces. Thus for example, light can be out-scattered at a particle in many different directions, or a fraction of its original energy can be absorbed on its way through the scene. Light from other directions can also be in-scattered or new light particles can be emitted at a point in the current direction. As shown by the thickness of the arrows, out-scattering and absorption lead to a decrease in the light flow, while in-scattering and emission processes increase the amount of light energy within a system.

REMARK 1.2 (Solutions to the Global Illumination Problem) *Later we will see, that the global illumination problem is given by two equations:* The stationary light transport equation, SLTE, *describing the distribution of light in a scene, and the* measurement equation, *which can be interpreted as the exposure of an image. Solving the global illumination problem then means the calculation of the illumination at relevant points which are specified via the measurement equation, and combining both results in a proper way.*

SLTE (296)
Measurement Equation (416)
Chapter 5

Now, the goal of any realistic rendering algorithm results in computing solutions to the equations (1.2) and (1.3) or in other words, the best possible and realistic visualization of a 3-dimensional scene to be rendered. In this case, one speaks also of generating photorealistic images, whereas this objective can obviously only be achieved if all physical phenomena of light—such as reflection, transmission, polarization, interference, and the diffraction of light as well as phosphorescence and fluorescence effects—can be modeled as accurately as possible. For the models of global illumination, this would mean that

Chapter 5
Chapter 4

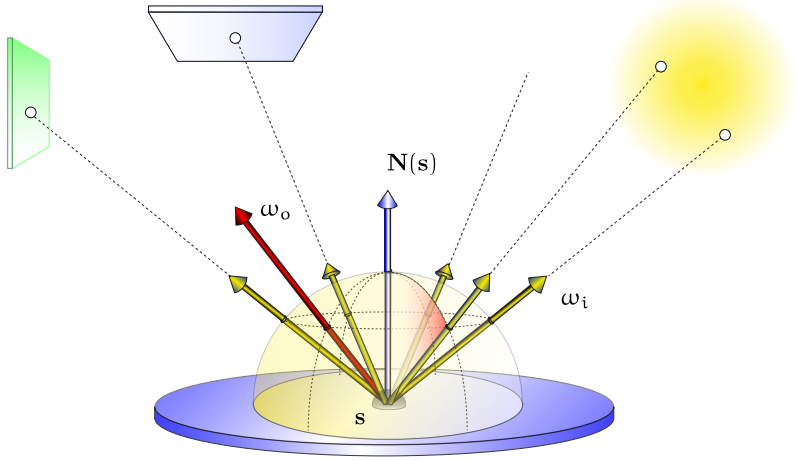


FIGURE 1.5: THE GEOMETRY FOR DERIVING THE STATIONARY LIGHT TRANSPORT IN A VACUUM. A point s on a surface is illuminated by an area light source. The fraction of light, exitant from s in direction ω_o depends on the light incident at s coming from all directions over the hemisphere at s and the reflection properties of the surface.

they must satisfy in addition to the laws of geometrical and physical optics, also quantum-optical phenomena, and the energy conservation laws of physics.

It has been found that this requirements, in relation to the development of appropriate algorithms, are too strict and for the currently available hardware computationally too costly. Thus, we are only interested in approximate solutions to the global illumination problem, where complicated light phenomena such as polarization, interference, and the diffraction of light as well as phosphorescence and fluorescence effects are neglected. Under the further restriction, that we consider only the stationary light transport in a vacuum, an approximated global illumination problem can be described by the *stationary light transport equation in free space, the SLTEV,* given by: SLTEV (398)

$$L_o(s, \omega_o) = L_e(s, \omega_o) + \int_{S^2(s)} f_s(s, \omega_i \to \omega_o)\, L_i(s, \omega_i)\, d\sigma_s^{\perp}(\omega_i), \tag{1.4}$$

in computer graphics also known as the *rendering equation.* In this equation, the functions L_o, L_i and L_e represent the outgoing, the incoming, as well as the emitted light at point s in the corresponding directions ω_o and ω_i, and the function f_s returns the fraction of light incident at s from direction ω_i that is reflected or refracted in direction ω_o, see Figure 1.5. Rendering Equation (400)

As the name already suggests, the light transport equation in a vacuum describes

how light propagates under vacuum conditions in a real scene. Even this equation is still a complicated Fredholm integral equation of the 2^{nd} kind, but much simpler than the light transport equation from (1.3). With respect to this equation, global illumination then means the simulation of light interplay with the objects in a 3-dimensional scene under vacuum conditions.

Section 2.3

Due to its high complexity and its requirements to the power of digital computers, the simulation of light interaction with objects in a 3-dimensional scene, even in a vacuum, was only difficult to achieve for a long time. Thus, classical numerical methods were essentially inapplicable, due to the high complexity and the discontinuities of the integrands, induced by different optical and geometric properties of the objects within the scene.

Section ??

With the development of so-called *Monte Carlo rendering algorithms*, based on *discrete-time Markov processes*, and *radiosity algorithms*, derived from the *finite element method*, then the problem of global illumination was getting under control. Both methods achieve, in their most commonly used forms, only partially global interaction of light with the objects involved in a scene. That is why the generated images only correspond to approximate solutions of the complete *global illumination problem*.

Chapter 9

Chapter 10

Global Illumination Problem (6)

While Monte Carlo algorithms are based on stochastic principles from probability theory—they try to solve an integral by interpreting the integral as the expected value of a continuous random variable—radiosity methods are based on the finite element approach, which transforms infinite-dimensional integral equations, such as the stationary light transport equations, into a systems of finite-dimensional equations.

Chapter 10

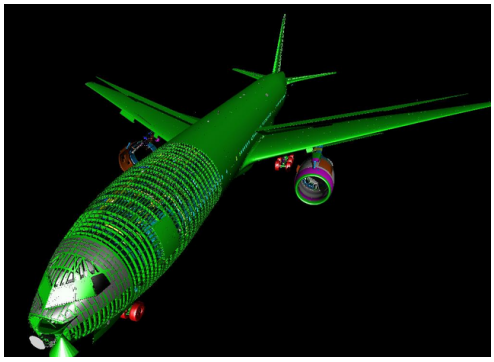
Section 2.4

Section 2.3.3.2.3

In combination with the continuous and rapid development, particularly in terms of processor performance in computer systems and the increasing interest in photorealism, then the development of global illumination models has also found its way into computer graphics. Thus, in particular, there are many application areas in industry, technique, and science, which all require realistic computations of light distributions in a predefined scene.

Thus, the automotive industry uses these procedures in order to develop vehicles faster and avoid mistakes in planning from the outset. According to well-known car manufacturers the cost and time involved in the development, can be decreased by a substantial amount. Also the aircraft industry has great interest in this new technology to represent complete planes in 3D. Thus, the original CAD model of a *777* requires the interactive processing of a data volume of more than 30 GB. This new technology makes it possible, that designers can move interactively through the virtual plane and check every detail to the point of the smallest bolt and rivet. Thus, new aircraft models can be tested as a whole and potential problems can be detected already before the construction [226, Wald & al. 2003], see Figure 1.6.

Realistic rendering methods can also be used to simulate the light propagation in biological tissues and other scattering materials, which are used for the development of biomedical devices and in medicine to produce meaningful data from images. But the



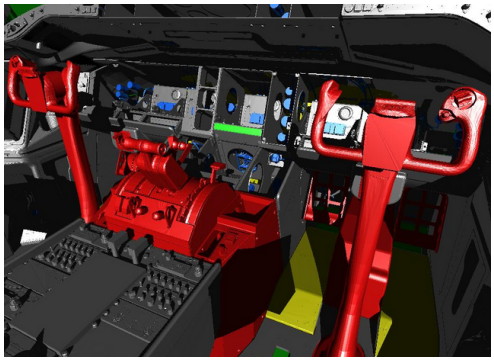


FIGURE 1.6: CG IN AUTOMOTIVE AND AIRCRAFT INDUSTRY. The complete *Boeing 777* model, which consists of 350 million triangles and more than 30GB scene data on disk, rendered with shadows on a single PC and a zoom into the cockpit.

largest scope of realistic rendering techniques is probably in the film and television industry, particularly in fantasy, science fiction and commercials, see Figure 1.7. If things are to be shown that don't exist in reality, but which should contribute most genuinely, then more often Monte Carlo rendering and radiosity methods are used.

1.1.3 RAY TRACING - A FIRST DETERMINISTIC APPROACH FOR SOLVING THE GLOBAL ILLUMINATION PROBLEM

Before we shortly speak about Monte Carlo algorithms and radiosity methods, we present a simple technique, which has been used for a long time to solve the global illumination problem in an approximative manner: *Ray tracing*.

Section 1.3

Section 8.3

The basics of any ray tracing algorithm are the principles of geometric optics, i.e. the field of optics, that allows to describe light by the mathematical concept of the *light ray*.

Section 4.2.1

DEFINITION 1.4 (The Concept of a Mathematical Ray) *Let* $\mathbf{x}$ *be a point within the Euclidean space* $\mathbb{R}^3$, α *a positive real number, and* ω *a direction over the unit sphere around point* $\mathbf{x}$. *Then, a* ray $\mathbf{r}$ *is defined by*

Euclidean Space (830)

Direction (834)

$$\mathbf{r} \stackrel{\text{def}}{=} \mathbf{x} + \alpha\,\omega, \tag{1.5}$$

thus, the set of all points lying on the line starting at point $\mathbf{x}$ *in direction* ω, *see Figure 1.8.*

REMARK 1.3 *In classic Whitted-style ray tracing, rays are emitted from points typically selected on object surfaces in any arbitrary direction* ω *of the upper or lower*

Section 8.3

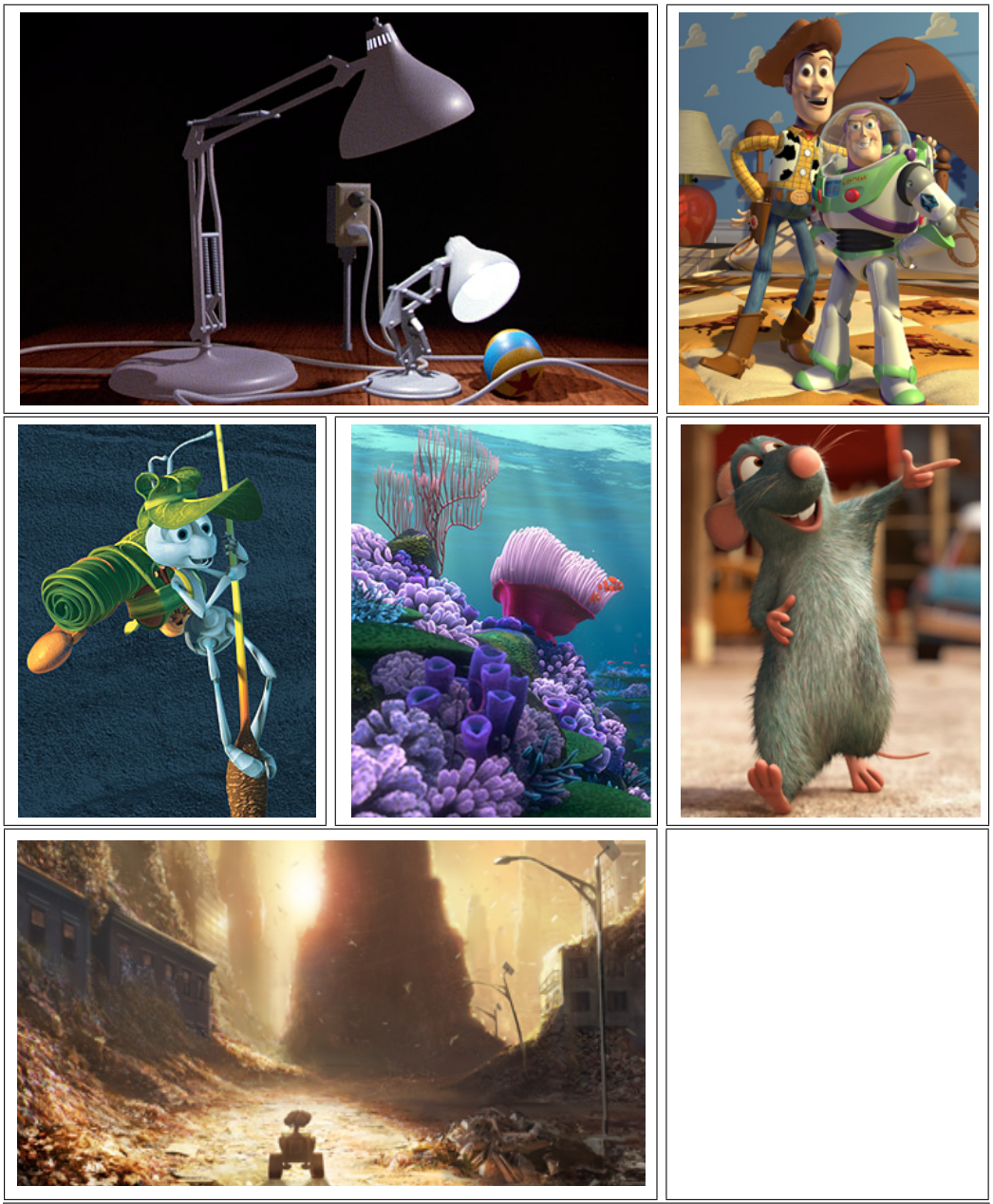


FIGURE 1.7: FILM AND VISUAL EFFECTS. Pixar's history of animation movies: Ray traced pictures with hard shadows within the upper row. The other images are rendered with global illumination algorithms simulating soft shadows, light traveling through participating media, skin, fur, and dust. Image courtesy of Pixar Animation Studios.

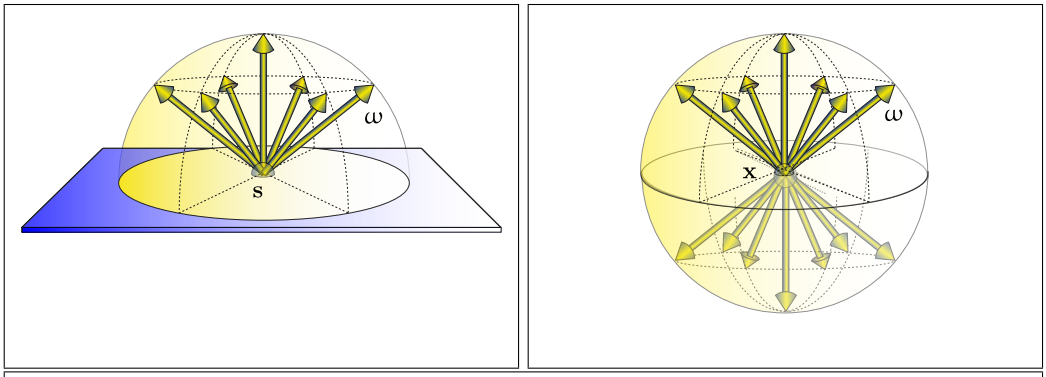


FIGURE 1.8: THE GEOMETRY OF A MATHEMATICAL RAY. A direction ω from the upper hemisphere and a point $\mathbf{s}$ at a surface patch are required to define a mathematical ray starting from an opaque surface object. Rays starting from points in space $\mathbb{R}^3$ are given via a point $\mathbf{x} \in \mathbb{R}^3$ and a direction ω over the unit sphere.

hemisphere about $\mathbf{s}$. *Only in the case of involved participating media the starting point of a ray is a point from* $\mathbb{R}^3$.

Simple variants of ray tracing, such as the classic Whitted-style ray tracing algorithm, [236, Whitted 1980], computes only approximations to the global illumination problem, that is, they only account for *light coming directly from light sources* within the scene, or light that is reflected or refracted at specular surfaces. *Indirect light*, diffusely *reflected, transmitted,* or *scattered* at objects within the scene, is simulated by means of a constant ambient term.

Section 8.3
Global Illumination Problem (6)

In principle, a ray tracing algorithm works as if one makes a photo, see Figure 1.9. There is a scene, a viewer with, in the simplest case, a pinhole camera, and a light source, which illuminates the scene. In a ray tracing algorithm, the viewer or the camera are replaced by a virtual camera, and the scene and the light sources are described by a model of 3-dimensional objects in a strictly defined manner. The final image on the film plane then corresponds to the projection of the scene onto a 2-dimensional pixel array, or perhaps on a computer screen.

A ray tracing algorithm, see Figure 1.10, sends rays from the eye through every pixel of the image plane into the scene to be rendered and computes the first hit of such a ray with an object. If there isn't a hit with an object, then the pixel gets the color of the background of the scene, that is, it will be black. If the ray hits a specular object, then—due to the *law of reflection* —the algorithm computes a reflection ray, which will be traced recursively through the scene. Is the object transparent, then a ray tracing algorithm generates, in addition to the *reflection ray*, also a *refracted ray* and traces both rays *recursively* through the scene. Eventually occurring shadow areas are computed by

Law of Reflection (300)
Law of Refraction (305)

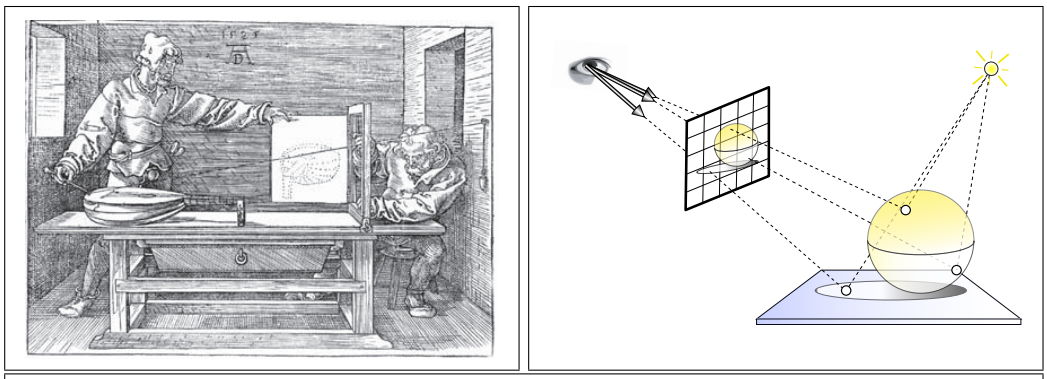


FIGURE 1.9: THE PRINCIPLE OF RAY TRACING. The principle of ray tracing as it can be interpreted in fine arts and in computer graphics.

so-called *shadow rays*. Such rays are fired from every hit point of the original rays with scene objects towards existing light sources. If a shadow ray hits a blocking object on its way to a light source, then the starting point of the shadow ray lies in the shadow area of this light source, that is: The point does not get lit from this light source.

Now, as ray tracing algorithms are based on the law of reflection and refraction, they are capable of simulating a wide variety of optical effects such as reflection, refraction or scattering, see Figure 1.11, but for the treatment of diffuse phenomena of light, they require additional effort.

REMARK 1.4 *One of the main reasons for the success of ray tracing lies in its natural extensibility. The above primitive method is inadequate for present requirements of image synthesis. With increasing computing power and increasing inspiration from physics—in particular the fields of optics and radiometry—more and more extensions and variants of ray tracing come into play, and we will also introduce some of them in this book.*

Chapter 8

Chapter 9 *Basically it holds: every extension of the algorithm leads to better quality of the rendered images but also to an increase in the run time of the algorithm.*

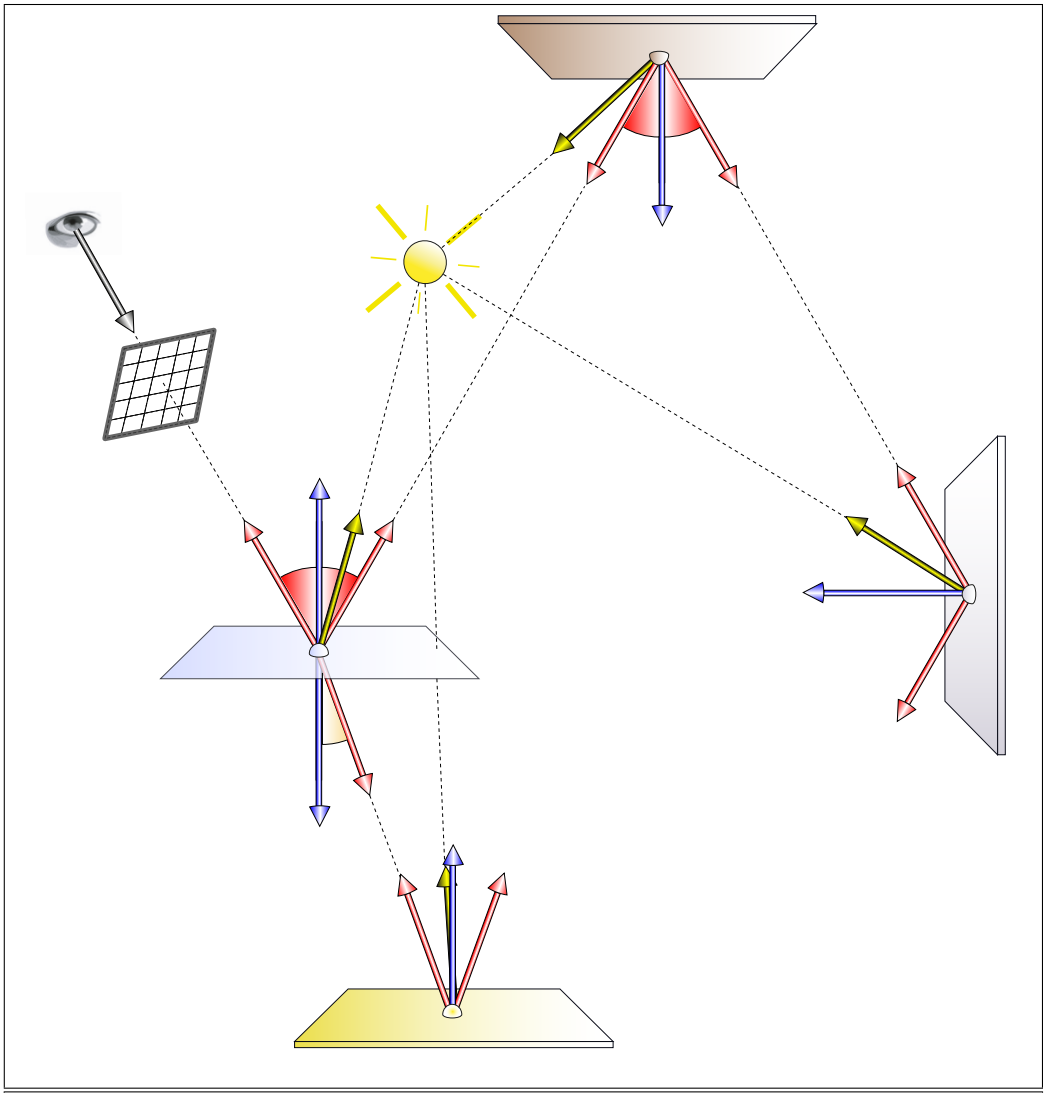


FIGURE 1.10: CLASSIC WHITTED-STYLE RAY TRACING. The algorithm starts with generating a primary ray from a sensor, typically the eye of an observer or a virtual camera, through a pixel of the image plane into the scene to be rendered. At the first hit point of this ray with the closest scene object, the algorithm can generate, depending on the properties of the material of the concerned surface, up to three new types of rays: a reflection ray, a refraction ray, and a shadow ray. The algorithm estimates the incoming light at the intersection point of the primary ray with an object and combines this information to a contribution to the final color of the pixel. The computation of the light contributions of the reflected as well as the refracted ray are taken recursively until a diffuse surface is hit, the ray doesn't intersect an object within the scene, or the intensity of the ray is below a threshold value respectively the recursion depth of ray generation exceeds a predefined value.

FIGURE 1.11: RAY TRACING. Typical ray-traced images. The ray tracing algorithm can render hard shadows and specular reflections, but it does not simulate indirect illumination of diffuse surfaces. Image Courtesy of Gilles Tran and Jaime Vives Piqueres.

1.2 FUNCTIONAL ANALYTICAL APPROACHES FOR SOLVING THE GLOBAL ILLUMINATION PROBLEM

Let us consider once more the stationary light transport equation valid in a vacuum from Relation (1.4), thus,

$$L_o(\mathbf{s}, \omega_o) = L_e(\mathbf{s}, \omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i), \tag{1.6}$$

but now a little bit closer.

Linear Integral Equations (126)
Linear Function Spaces (27)

As already mentioned, it is a linear Fredholm integral equations of the 2^{nd} kind, where the functions L_o, L_e, as well as L_i, and f_s are elements of abstract function spaces. As we see further below, the above integral is constructed over spaces, which are more complicated than the usual space of real numbers, i.e. it is not a Riemann integral, but it must be interpreted as a so-called *Lebesgue integral*.

Lebesgue Integral (105)

Now, the Lebesgue integral is the integration concept of modern mathematics and for this type of integral some statements hold that do not hold for the classic Riemann integral. These properties are also responsible for ensuring that the Lebesgue integral is the basis for the functional analytical theory of integral equations. That is, for an exact mathematical formulation of light transport and the derivation of solutions to light transport equations, fundamental knowledge about *measure*- and *integration theory* are absolutely necessary.

Light Transport Equation (295)
Section 2.2

Unfortunately, only few integral equations arising from practical applications prove to be analytically solvable. This holds in particular to our case of the light transport equation. For equations of such kind there exists, if at all, only approximative solutions. For deriving procedures to approximate solutions of linear Fredholm integral equations of the 2^{nd} kind, now we have to construct the function spaces underlying the light transport equation. In order to derive consistent solution methods, which can be used on computers, we then need on the one hand abstract existence and uniqueness proofs for integral equations, but also *convergence proofs* including *error estimates*, that show theoretically that approximate procedures exist and converge to the real solution to the problem. This is the main task of *functional analysis*.

Linear Function Spaces (27)
Section 2.3.3
Section 2.1

For that purpose, in a first step the attempt is made to represent the problem to be solved in form of a *linear operator equation* and then apply the results about existence and uniqueness of solutions to linear operator equations. If the general conditions are satisfied in the given problem, then functional analysis provides us with methods for solving the problem. In this context, *abstract, infinite dimensional function spaces*, such as the *Lebesgue spaces* defined over a given base set, play a central role. Without the concept of the Lebesgue integral it would not be possible to discuss the global illumination problem in infinitely dimensional spaces in a strict mathematical way.

Linear Operator Equations (61)
Section 2.1
Section 2.2.4

As the Lebesgue spaces of global illumination theory allow to represent incident functions by exitant functions, and vice versa—we will discuss this in a later chapter - —Equation (1.4) can also be written only in terms of the functions L_o, L_e, and f_s, that is,

Chapter 5

$$L_o(s,\omega_o) = L_e(s,\omega_o) + \int_{S^2(s)} f_s(s,\omega_i \to \omega_o)\, L_o(\gamma(s,\omega_i),-\omega_i)\, d\sigma_s^{\perp}(\omega_i). \tag{1.7}$$

The unknown function L_o then appears on the left hand side and on the right hand side of the equal sign under the integrand. This recursive structure now implies, as we will see this in more detail in the next section, a first idea for finding approximate solution methods for integral equations of the above type.

1.2.1 THE NEUMANN SERIES APPROACH

Let us now apply the above mentioned approach from functional analysis to the global illumination problem, namely: To reformulate the vacuum light transport equation into a linear operator equation on an abstract infinite dimensional function space. As demonstrated in one of the next sections in more detail, the integral occurring in the light transport equation can then be replaced by a so-called *integral operator* and the original integral equation can be written in form of an operator equation as follows:

Global Illumination Problem (6)
Linear Operator Equation (61)
Linear Function Space (28)
Linear Integral Equation (127)

$$\underbrace{L_o(s,\omega_o)}_{L_o} = \underbrace{L_e(s,\omega_o)}_{L_e} + \underbrace{\int_{S^2(s)} f_s(s,\omega_i \to \omega_o)\, L_o(\gamma(s,\omega_i),-\omega_i)\, d\sigma_s^{\perp}(\omega_i)}_{\mathbf{K}L_o} \tag{1.8}$$

$$L_o = L_e + \mathbf{K}\, L_o, \tag{1.9}$$

Linear Integral Operator (130) where $\mathbf{K}$ denotes a so-called *integral operator*, exactly specified in a later section. Slightly rephrasing Equation (1.9) then leads to

$$L_o - \mathbf{K} L_o = L_e, \tag{1.10}$$

respectively,

$$(\mathbf{I} - \mathbf{K}) L_o = L_e. \tag{1.11}$$

Assuming there exists an inverse operator $(\mathbf{I}-\mathbf{K})^{-1}$, where $\mathbf{I}$ is the identity, then the above operator equation can be solved with respect to L_o by

$$L_o = (\mathbf{I} - \mathbf{K})^{-1} L_e. \tag{1.12}$$

[174, Rudin 1998] Section 2.3.3.1.1 $\|\cdot\|$ (860) Now, we know from the theory of infinite series, that an expression of the form $(\mathbf{I}-\mathbf{K})^{-1}$ can be interpreted as the limit of a geometric series, whose terms are composed of the powers of the operator $\mathbf{K}$, where $\|\mathbf{K}\| < 1$ is assumed—we discuss this in more detail in one of the next sections. That is, it holds:

$$\sum_{i=0}^{\infty} \mathbf{K}^i = \frac{1}{(\mathbf{I}-\mathbf{K})} = (\mathbf{I}-\mathbf{K})^{-1}. \tag{1.13}$$

Defining the partial sums:

$$S_n \stackrel{def}{=} \sum_{i=0}^{n} \mathbf{K}^i = \mathbf{I} + \mathbf{K} + \mathbf{K}^2 + \cdots \tag{1.14}$$

and substituting the inverse operator by the partial sums, then Equation (1.12) can be written as:

$$L_o = S_n L_e, \tag{1.15}$$

that is, for sufficiently large n, the sequence of partial sums multiplied by L_e converges to an exact solution of the operator equation (1.12), or in other words converges to a solution of our light transport equation. This procedure of finding a solution method to an integral operator equation underlying a Fredholm integral equation of the 2^{nd} kind is called the *Neumann series approach*. Section 2.3.3.1.1

As a resume, now we can summarize: If it is possible to implement the Neumann series approach into a rendering algorithm, then we have a chance to solve the global illumination problem exactly.

1.2.2 A FINITE ELEMENT APPROACH

Another approach for finding a solution to the global illumination problem valid in a vacuum is based on *finite element methods*. To illustrate this technique, we assume, that the light transport in free space is expressed in terms of another radiometric quantity, *radiosity*, instead of radiance. The associated equation is also denoted as the *radiosity equation*, and is of the form

$$B(\mathbf{s}) \quad = \quad B_e(\mathbf{s}) + \rho_{dh}(\mathbf{s}) \int_{\partial \mathcal{V}} B(\mathbf{s}')\, \mathcal{G}'(\mathbf{s}' \leftrightarrow \mathbf{s}) \mathcal{V}(\mathbf{s}' \leftrightarrow \mathbf{s})\, d\mu^2(\mathbf{s}'). \tag{1.16}$$

In this equation, the functions B_e and B represent the emitted as well as the reflected fraction of light at point $\mathbf{s}$ in all directions. $\rho_{dh}(\mathbf{s})$ is a function, that returns the fraction of light incident at $\mathbf{s}$ which is reflected, and $\mathcal{V}$ and $\mathcal{G}'$ are functions that provide information about the visibility and the geometry of the surface points $\mathbf{s}$ and $\mathbf{s}'$. Equation (1.16) can be considered as a mathematical formulation of light transport in a vacuum, assuming that the object surfaces existing in the scene are all Lambertian reflectors, and the light sources are ideal diffuse emitters.

Radiosity Equation (784)
$\rho_{dh}(\mathbf{s})$ (338)
$\mathcal{V}$ (45)
$\mathcal{G}'$ (129)
Lambertian Reflector (349)

Now, like the vacuum light transport equation, the radiosity equation is also a Fredholm integral equation of the 2^{nd} kind, where the involved functions B_e and B are elements of an infinite-dimensional function space. By partitioning the underlying integration domain in a finite set of so-called *patches*, and restricting these patches furthermore, then, due to functional analytical considerations, the functions B_e and B can be projected onto functions from a finite-dimensional function space constructed over the set of patches. This discretizing procedure leads to a modification in the radiosity equation, namely, the replacement of the integral by a finite sum. As a result, we get a linear system of equations of the type

Linear Integral Equation (127)
Linear Function Space (28)
Section 2.1.1

$$\mathbf{B} = \mathbf{B}_e + \mathbf{M}\,\mathbf{B}, \tag{1.17}$$

where $\mathbf{B}$ and $\mathbf{B}_e$ are vectors, namely the finite-dimensional analogues of B_e respectively B, and $\mathbf{M}$ corresponds to a matrix. Similar to the Neumann series approach, the finite element approach, thus Equation (1.17), also results in an operator equation, but now based on operators defined on finite-dimensional spaces, namely matrices.

Section 1.2.1
Linear Operator Equation (61)

Analogue to our discussion in the foregoing section, even this operator equation can be written as:

$$(\mathbf{I} - \mathbf{M})\,\mathbf{B} = \mathbf{B}_e. \tag{1.18}$$

Assuming, that the matrix $\mathbf{M}$ is invertible, the above operator equation can be solved via

$$\mathbf{B} = (\mathbf{I} - \mathbf{M})^{-1}\,\mathbf{B}_e, \tag{1.19}$$

that is, an approximate solution to the stationary light transport equation in vacuum can also be derived by using one of the well-known direct or iterative solvers for linear systems of equations.

SLTEV (398)
Section 2.3.3.2.4.2

1.3 MONTE CARLO RAY TRACING AND RADIOSITY METHODS FOR SOLVING THE LIGHT TRANSPORT EQUATIONS

Based on the above two approaches from functional analysis, the Neumann series approach and the finite elements method, we now present two classes of rendering algorithms for solving approximations of the global illumination problem: *Monte Carlo rendering* and *radiosity algorithms*.

Chapter 9 Chapter 10

Monte Carlo rendering algorithms are a special class of numerical methods, which can be used to solve the light transport equation by means of probabilistic principles. They all have in common the idea to solve the light transport equation via the probabilistic concept of the *random walk*. Random walks are paths generated over the object surfaces of the scene to be rendered. On these paths, light particles can travel, according to the laws of geometric optics, from light sources to the eye of an observer or a virtual camera, and contribute their amount of light to the final image.

Chapter 9 Section 2.4 Section 9.1 Section 4.2.1

Radiosity approaches find their origins in the 50s of the last century, as scientists are interested in the exchange of radiant heat between object surfaces. Simulations in this area were especially of enormous relevance for the emerging research of universe. Mid-80s, researchers from Cornell University [70, Goral & al. 1984] and the universities in Fukuyama and Hiroshima [139, Nishita & al. 1985] developed, based on these research, the first method for solving the global illumination problem in realistic image synthesis. These procedures are all based on the law of conservation of energy from physics and allow, in their simplest versions, the simulation of diffuse multiple reflections.

Chapter 10

1.3.1 MONTE CARLO PATH TRACING - A PROBABILISTIC APPROACH BASED ON THE NEUMANN SERIES

Monte Carlo path tracing, as we will introduce it in Section 9.1, is in some sense a generalized form of ray tracing. Instead of generating rays in a deterministic way, a path tracing algorithm generates rays in a probabilistic way. In contrast to classic ray tracing, path tracing traces—depending on the surface properties of the scene objects—only a single ray through the scene to be rendered. In its classic form, the algorithm stops the process of a path extension if the length of the current path extends a predefined value, the contribution of the path to the final image is less than a given threshold, if a path hits one of the existing light sources or if it does not hit any scene object, see Figure 1.12. As Monte Carlo path tracing attempts to map the exact physical behavior of light at surfaces, the algorithm can simulate effects such as soft shadows, depth of field, motion blur, caustics, and indirect illumination, see Figure 1.13.

Section 9.1

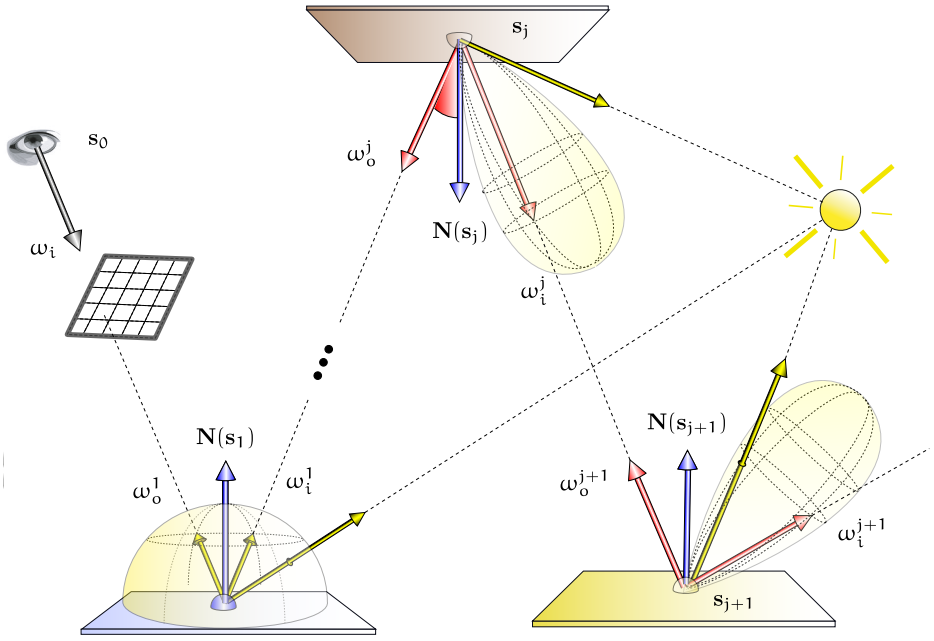


FIGURE 1.12: MONTE CARLO PATH TRACING. A ray starting at the eye is shot through a pixel of the image plane into the scene to be rendered. Monte Carlo path tracing determines the first hit of the ray with any object of scene and generates in dependence of the surface properties another new ray. This ray is then traced recursively through the scene. Note: Since the probability is very small, that a path ends in one of the light sources, for enhancing the image quality, the algorithm generates shadow rays at each hit point in direction to the light sources, this then corresponds to Monte Carlo path tracing with next event estimation.

Apart from all of these advantages, Monte Carlo path tracing has also a significant disadvantage: Its slow convergence to the exact image. Since the light sources within a scene are mostly small comparable with the other objects in a scene, the probability that a path ends in one of the light sources can be very small, that is, a path generated with MCPT will usually not hit a light source. Due to the fact the most traced paths do not contribute to the final image, the quality of an image can only be enhanced via the generation of a large number of paths in the hope that one or more of these hit a light source. These error is noticeable in pictures by noise, in particular if the scene to be rendered consists of many diffuse objects, see Figure 1.14.

Now, we pose the question: How is it possible to interpret the solution of the SLTEV in the sense of an algorithm like path tracing? Can we built a bridge between the light

SLTEV (398)

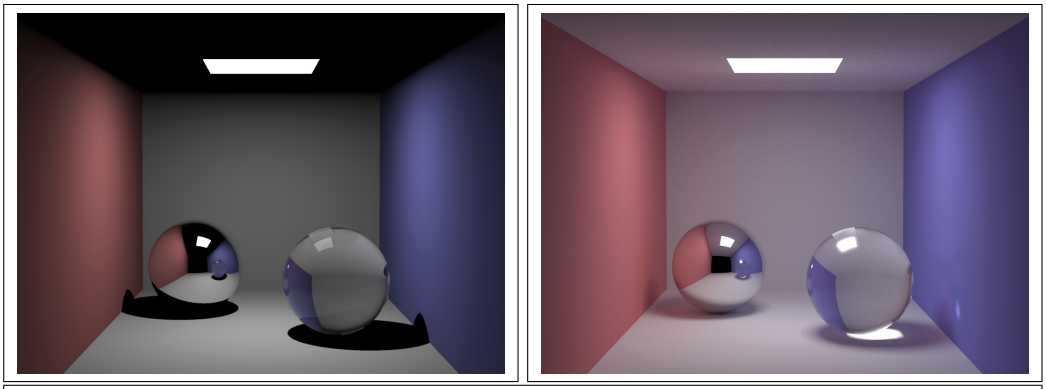


FIGURE 1.13: RAY TRACING AND MONTE CARLO PATH TRACING. Left a ray-traced image of a simple box scene. Right the scene rendered using Monte Carlo path tracing. Note: Unlike ray tracing algorithms, Monte Carlo path tracing can simulate all light paths. Notice the illumination of the ceiling and the caustic below the glass sphere. Image courtesy of Henrik Wann Jensen, UCSD.

transport equation in free space and our simple Monte Carlo path tracing algorithm?

SLTEV (398)

For this, let us consider once more the Neumann series approach for solving the SLTEV from Section (1.2.1). Due to Equation (1.13) an approximate solution is given by:

$$\begin{aligned} L_o &\approx S_n L_e, && (1.20)\\ &= \sum_{i=0}^{n} \mathbf{K}^i L_e && (1.21)\\ &= L_e + \mathbf{K} L_e + \mathbf{K^2} L_e + \mathbf{K}^3 L_e + \cdots + \mathbf{K}^n L_e, && (1.22) \end{aligned}$$

that is, the total amount of light can be computed via the light emitted from light sources in addition to the emitted light that is reflected or transmitted at a surface, and the emitted light that is twice reflected or transmitted at a surface and so on. Obviously, Equation (1.22) can be interpreted as sum of paths starting at a pixel of an image and ending at a light source in a scene, thus, a path of length one, a path of length two, a path of length three and so on. That is, light can flow along these paths from sources to a detector: This process is identical to *Monte Carlo path tracing*.

Section 9.1

Equation (1.22) shows that Monte Carlo path tracing is a global illumination algorithm since we can combine the terms of the sum as follows:

$$L_o = \underbrace{L_e + \mathbf{K} L_e}_{\text{local Illumination}} + \underbrace{\mathbf{K}^2 L_e + \mathbf{K}^3 L_e + \ldots + \mathbf{K}^n L_e}_{\text{global Illumination}}, \quad (1.23)$$

that is, light paths of length ≤ 2 correspond to a local illumination model, and paths of length > 2 are associated with a global illumination model, that is, they simulate global interaction of light with the objects of a scene.

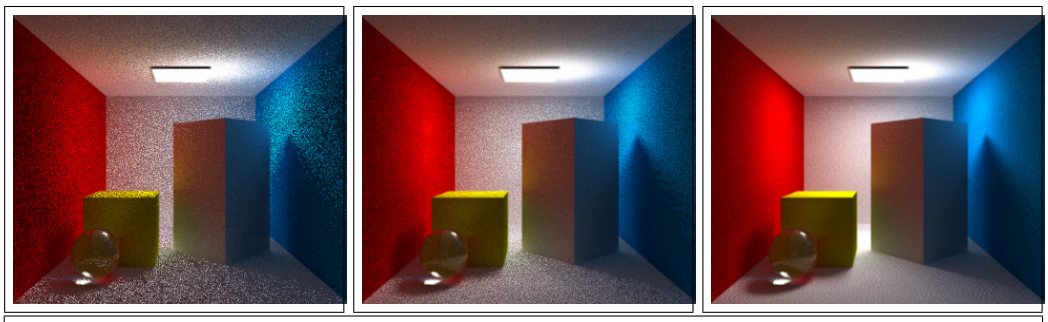


FIGURE 1.14: MONTE CARLO PATH TRACING. Monte Carlo path tracing can simulate full global illumination, but often results in noisy images as seen in these simple box scenes. Generating more than one path per pixel can correct the quality of the images: Left, 5 paths, in the center, 25 paths, and in the right image 125 paths per pixel.

1.3.2 THE RADIOSITY METHOD — A FINITE ELEMENT APPROACH

The *radiosity method* is based on the principle of energy conservation of physics, that is, all light falling on a surface is also reflected, if it is not absorbed by this surface.

Chapter 10

A radiosity algorithm assumes that all surfaces in a scene to be rendered are perfectly diffuse. By discretizing the surfaces into small, simple geometrical patches P_i, such as quadrilaterals, triangular elements, or Voronoi-diagrams, the entire scene is covered by a net of these patches, see Figure 1.15. Under the assumption that the reflectivity and the radiosity over all of these patches is constant, thus $\rho_{dh}(s) = \rho_i$ and $B_i(\mathbf{s}) = B_i$, $\forall \mathbf{s} \in P_i$, the original radiosity equation, a Fredholm integral equation of the 2^{nd} kind, can be transformed into a system of n discrete radiosity equations

Radiosity (264)

Freholm Integral Equation (127)

$$B_i = E_i + \rho_i \sum_{j=1}^{n} F_{ij} B_j \tag{1.24}$$

for $1 \leq i \leq n$. In this equation, the factors F_{ij} are called the *patch-to-patch form factors*. They indicate what fraction of light on patch P_i originates at patch P_j, in other words, they describe how well the patches can see each other. Thus, patches that are far away from each other, or that are oriented at oblique angles to each other, will have smaller form factors than patches that are opposite to each other. If two patches are covered by a third patch, then their form factor will be reduced or zero, depending on whether the occlusion is partial or total.

Form Factor (784)

The form factors represent, together with the reflectivities ρ_i, the coefficients of the linear system from Equation (1.24). Thus, the linear system of equations can also be

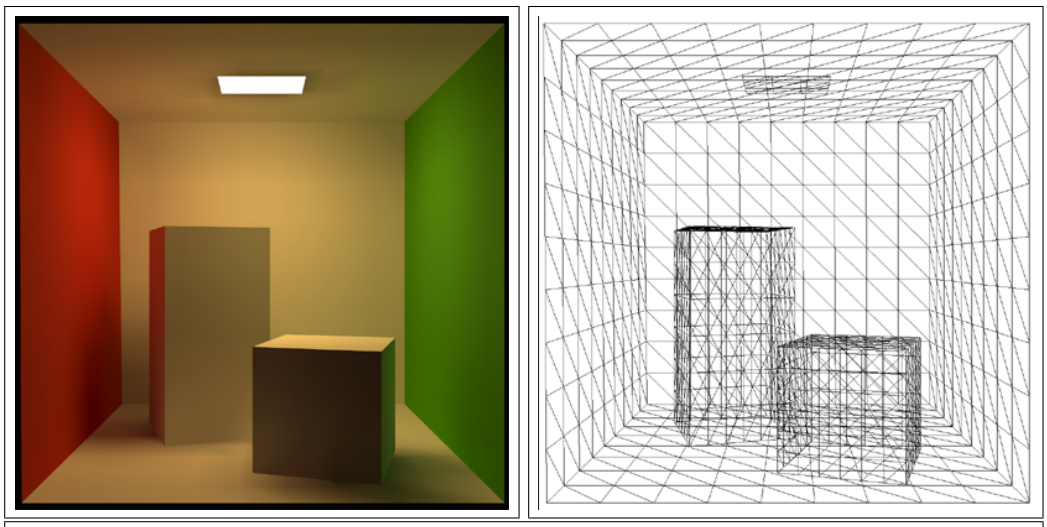


FIGURE 1.15: RADIOSITY. The Cornell box rendered via a radiosity algorithm, where the right image shows the associated partition of the surfaces in a mesh of disjoint small surface patches.

written as

$$(\mathbf{I} - \mathbf{M})\,\mathbf{B} = \mathbf{B}_e \tag{1.25}$$

with

$$(\mathbf{I} - \mathbf{M}) = \begin{pmatrix} (1-\rho_1 F_{11}) & -\rho_1 F_{12} & \cdots & -\rho_1 F_{1n} \\ -\rho_2 F_{21} & (1-\rho_2 F_{22}) & \cdots & -\rho_2 F_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ -\rho_n F_{n1} & -\rho_n F_{n2} & \cdots & (1-\rho_n F_{nn}) \end{pmatrix}. \tag{1.26}$$

Assuming, that the matrix $(\mathbf{I} - \mathbf{M})$ is invertible, the above operator equation can be solved using a direct or an iterative solver such as a Jacobi, Gauss-Seidel, or Southwell iteration method. It then results in the radiosity of each patch, taking into account interaction of light at diffuse surfaces and soft shadows. While ray tracing can be considered as a so-called *image-space based algorithm*, the radiosity method is an *object-space based algorithm*, that computes the light distribution in the entire scene independent of the position of the viewer.

Section 2.3.3.2.4.2

CHAPTER TWO

MATHEMATICAL FOUNDATIONS OF REALISTIC RENDERING

"Wenn uns die Beantwortung eines mathematischen Problems nicht gelingen will, so liegt häufig der Grund darin, dass wir noch nicht den allgemeineren Gesichtspunkt erkannt haben, von dem aus das vorgelegte Problem nur als Glied einer Kette verwandter Probleme erscheint."

DAVID HILBERT (1862 - 1943)

Until today, a great deal of research has been directed toward Monte Carlo and finite element methods for solving the light transport equations, but only little is known about the continuous equations beyond the existence and uniqueness of its solution. But to design more efficient, and elegant algorithms for finding finite-dimensional approximate solutions to the light transport equations a full understood of the continuous global illumination problem is more than helpful.

The common approach to transform the light transport equations, all together Fredholm integral equation of the 2^{nd} whose solutions exist in infinite-dimensional function spaces, occurs through the formalism of functional analysis. Functional analysis is the study of abstract infinite-dimensional function spaces and mappings that operate on these spaces. This elegant mathematical theory provides us with the tools for solving concrete analytical problems firmly anchored in real world, such as differential and linear integral equations. But not only functional analysis but also many other branches of mathematics, such as numerical analysis, measure, and integration theory as well as probability theory play a central role for a full understanding of the global illumination problem.

OVERVIEW OF THIS CHAPTER. The present chapter is divided in four sections. Each of these sections covers a distinct area from mathematics, which is needed for a full understanding of the theoretical basis of realistic rendering algorithms.

The first section is devoted to the principles of *functional analysis*. Mathematics, Section 2.1

in particular functional analysis: With its concepts of the linear space and the linear operator as a mapping on and between linear spaces, functional analysis provides us with the necessary tools for a formal description of the operations defined in realistic rendering processes and their required input parameters. Therefore, the most elementary concepts of functional analysis, where we turn our attention to the application areas of functional analysis with respect to global illumination theory, are presented in this part of the chapter. In Section 2.2 we present fundamental constructs and concepts relating to *measure and integration theory*. They build the necessary fundament for understanding and analyzing rendering processes based on stochastic and finite element methods. With the Lebesgue integral and the concept of the linear operator then we are ready to talk about *integral equations*, where we are primarily interested in integral equations of the so-called *Fredholm type*. So, we will define and discuss, with respect to the integral equations of global illumination, the most important solution methods to this type of integral equations. Finally we will conclude the chapter with a short excursion into the *probability theory*. With the Lebesgue integral at hand, and the necessary concept of the measurable function we will be able to study general probability spaces, especial types of random variables, and estimators. All these tools are necessarily needed to make reasonable statements about the quality of Monte Carlo estimators resulting from rendering algorithms.

Section 2.2
Section 2.3
Section 2.4

2.1 PRINCIPLES OF FUNCTIONAL ANALYSIS

Functional analysis, a melting pot of a large variety of different mathematical disciplines, may be regarded as the cornerstone of each and every kind of analysis aiming at the solution of operator equations and extremal problems in abstract, infinite-dimensional function spaces. This elegant mathematical theory provides us with the tools for solving concrete analytical problems firmly anchored in real world, such as differential equations, variational problems, and of particular interest here, linear integral equations.

For that purpose, the problem to be solved is first reformulated as an operator equation in an abstract infinite-dimensional function space, such as the *Lebesgue space* of square-integrable functions. For these kind of operator equations, functional analysis makes general statements about the existent and uniqueness of solutions. Provided that the general assumptions of the given problem are satisfied, functional analysis, conducted with the help of highly specialized tools, then supplies exact or at least approximate approaches for solving the problem.

Applied to the global illumination problem this means, that we will formulate the general light transport as an operator equation over an abstract, infinite-dimensional function space and then identify the rendering algorithms, based on ray tracing or radiosity, as those solution strategies that have their origin deep in functional analysis.

We will start in this section with a discussion about linear function spaces. The concept of the linear function space is the building block for the study of functional analysis. So, we will present the well-known function spaces of bounded and continuous functions, but we will also derive and study the rather more complex function spaces underlying the global illumination problem in more detail. We consider sequences of functions defined on function spaces, discuss different types of convergence, and investigate their limit behavior. Here, in particular we are interested in complete function spaces, the so-called *Banach* and *Hilbert spaces*. There are Hilbert spaces which fulfill the properties that make them appropriate to function spaces from which we can find solutions to the global illumination problem, exact in theory or approximative for the practical use. Afterwards, we will introduce the reader into the theory of linear operators—closely related to the concept of the linear function space—as those tools, that will allow us to transform the extremely complicated task of solving a linear integral equation into a simple linear operator equation. For that purpose, we will also talk about linear operator equations in more detail and demonstrate solution strategies for linear operator equations as well as its dual analogues, namely: *adjoint equations*.

Section 2.1.1
Section 2.1.2
Section 2.1.3
Section 2.1.4
Section 2.1.5
Section 2.1.6

2.1.1 LINEAR FUNCTION SPACES

Most problems in functional analysis exist in infinite-dimensional function spaces, that is, they can often not be solved exactly. In particular, this holds for problems deeply rooted in practice. Solutions to these problems can only be derived in numerical analysis. This requires to frequently examine the closeness of a numerical solution to its exact counterpart. Thus, we need a measure for estimating the difference between a numerical solution and the associated exact solution. From the Appendix it is known, that such a measure exists for vectors given by the concept of the norm. Therefore, we have to equip a function space with a topological structure, based on the concept of the norm. This then enables us to estimate the quality of an approximate solution of a problem. In functional analysis, such types of function spaces are known as *complete, linear normed function spaces*, often also better known as *Banach spaces*. If the norm underlying these spaces arises from an *inner product* then we call them *Hilbert spaces*.

Norm (860)
Banach Space (35)
Hilbert Space (36)

LINEAR FUNCTION SPACES. The main objective of this book is to find solution methods for solving the stationary light transport equation, also known as SLTE, given by:

SLTE (296)

$$L_i(\mathbf{x}, \omega_i) = \beta(\mathbf{s} \to \mathbf{x})\, \epsilon_b(\mathbf{s}, \omega_o) + \int_{[0, \|\mathbf{s}-\mathbf{x}\|]} \beta(\mathbf{x}' \to \mathbf{x})\, Q_o(\mathbf{x}', \omega_o)\, d\mu(\alpha), \tag{2.1}$$

a Fredholm integral equation of the 2^{nd} kind. As in the case of the computation of the Riemann integral

Linear Integral Equation (127)

$$f(x) = \int e^x dx, \tag{2.2}$$

in a certain sense also a simple integral equation, namely an integral equation of the first kind, whose solution e^x is an element of the set of all differentiable functions. Now, solutions of integral equations are members of sets that are equipped with additional algebraic and topological structures: so-called *function spaces*. So, the concept of the function space is not only required to decide whether there exist solutions to a linear integral equation and whether a solution is unique, but it is also useful for deriving techniques that can be applied to find these solutions—or at least to find approximate solutions of integral equations in practice.

DEFINITION 2.1 (Linear Function Space) *Let $\mathcal{S}$ be the set of all functions f from any set X to a set Y, thus,*

$$\mathcal{S} \stackrel{def}{=} \{f \mid f : X \to Y\}, \tag{2.3}$$

then $\mathcal{S}$ is called a linear function space*, or briefly a* function space*, if $\mathcal{S}$ satisfies the conditions required to a linear space. In other words, there must exist two operations, an addition and a s-multiplication, such that for two functions $f, g \in \mathcal{S}$ apart from $f + g$ also $\alpha \cdot f \in \mathcal{S}$ applies, where $\alpha \in \mathbb{K}$ and $\mathbb{K}$ is either the field $\mathbb{R}$ or $\mathbb{C}$.*

Linear Space (854)

Linear Space (854) Continuous Function (868) Closed Interval (829) EXAMPLE 2.1 (The Linear Function Space $C[a, b]$) *Let us introduce the function space $C([a, b])$, thus the space of all continuous, real valued functions defined over the closed set $[a, b]$. It may readily be seen that $C([a, b])$ satisfies the laws of a linear space. Intuitive, this should be clear, since with two functions $f, g \in C([a, b])$ also their sum $\alpha f + \beta g$ with $\alpha, \beta \in \mathbb{R}$ is continuous and $C([a, b])$ contains with $f = 0$ and $-f = (-1)f$ both, the zero as well as the inverse element of the space.*

It should also be clear, that $1, x, x^2, x^3, \ldots, x^n$ are all continuous functions and that this set of functions is linearly independent no matter how large n is. In accordance with Definition A.16 then the dimension of $C([a, b])$ is infinite.

Linear Independent (857)

Continuous Function (868) Closed Interval (829) EXAMPLE 2.2 *Let us consider the set of all non-negative, real-valued, continuous functions $C^{\geq 0}[0, 1]$ defined on the closed interval $[0, 1]$. Obviously, this set is not a function space, as there exists no inverse element $-f$ to any function $f \in C^{\geq 0}[0, 1]$, thus, $C^{\geq 0}[0, 1]$ is not closed with respect to the s-multiplication.*

Chapter 8 EXAMPLE 2.3 (The Linear Function Space $B[a, b]$) *The classic rendering algorithms simulate many natural light phenomena, such as the glittering play of colors in soap bubbles, the oily top layers of wet surfaces, and the color spectrum of the rainbow, only highly unsatisfactorily. For accurate color rendering in computer graphics any algorithm needs access to the full spectral character of the light sources and surfaces within a given scene. Thus, a rendering method must get enough spectral information to compute final values for output to some display such as an RGB-monitor.*

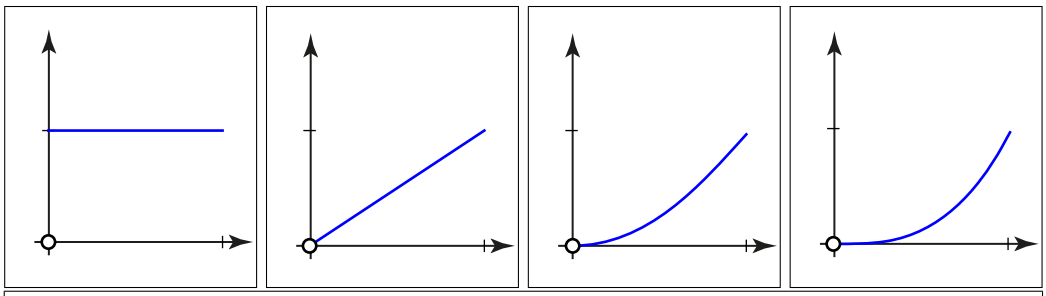


FIGURE 2.1: A LINEAR INDEPENDENT SET OF FUNCTIONS OF SPACE $C[0,1]$. The functions $1, x, x^2, x^3, \ldots$ correspond to an infinite set of linear independent functions of the space of continuous functions defined on the closed interval $[0,1]$.

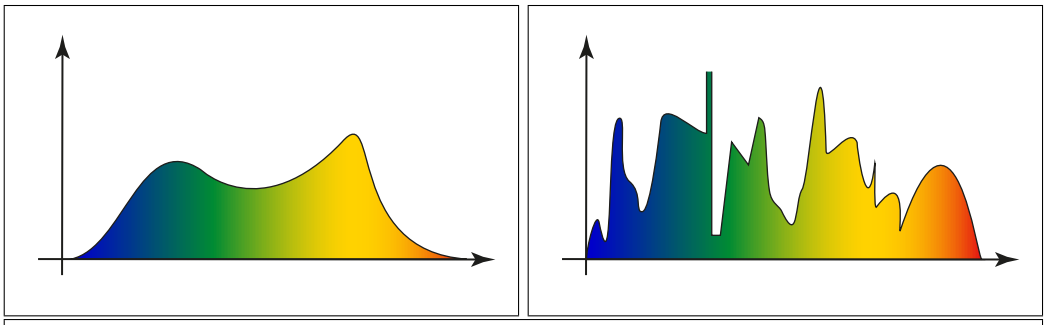


FIGURE 2.2: SPECTRAL POWER DISTRIBUTIONS. Two different spectral power distributions. Left, a smooth, continuous spectral power distribution right, a SPD as it can be associated with a fluorescent light bulb. It is very spiky and has discontinuities at a finite number of locations within Λ.

All these physical data about light and its reflecting properties at surfaces is contained in a spectral power distribution, briefly denoted as a SPD, see Figure 2.2. Except for line spectra, spectral power distributions are commonly continuous and bounded functions defined on the set Λ, the visible spectrum of light, where it holds: $\Lambda = [380\ nm, 780\ nm]$.

Continuous Function (868)

Bounded Function (863)

Let us consider the function space $B([a,b])$, that is, the space of all bounded, real valued functions defined on the closed set $[a,b]$. It may readily be seen that $B([a,b])$ satisfies the laws of a linear space. Intuitive, this should be clear, since with two functions $f, g \in B([a,b])$ also their sum $\alpha f + \beta g$ with $\alpha, \beta \in \mathbb{R}$ is bounded and $B([a,b])$ contains with $f = 0$ and $-f = (-1)f$ both, the zero as well as the inverse element of the space. A finite set of linear independent functions on $B([a,b])$ is then given by

Closed Interval (829)

Linear Independence (857)

$$f_i(x) \stackrel{def}{=} \begin{cases} 1 & x \in B_i \\ 0 & otherwise, \end{cases} \tag{2.4}$$

where $[a, b]$ *is partitioned into a finite subset of disjoint intervals* $\bigcup_{i=1}^{n} B_i, B_i \cap B_j = \emptyset, i \neq j$.

It should be clear, that the functions f_i *are all bounded functions and that this set of functions is linearly independent no matter how large* n *is. In accordance with Definition A.16 then the dimension of* $B([a, b])$ *is infinite-dimensional.*

Linear Independence (857)

REMARK 2.1 *Contrary to vector spaces in linear algebra, function spaces are typically of infinite dimension, that is, it does not exist a finite basis of functions that can be used to describe a function exactly by a linear combination of elements of this basis. There is only the possibility to approximate a function by a set of functions of a given finite basis.*

Basis (857)
Linear Combination (857)

In Section 2.3.3 we present a series of techniques for approximate solving linear integral equations of the Fredholm type. They all make use of the idea of approximating the unknown, usually infinite-dimensional, function within an integral equation by a linear combination of basis functions of a finite-dimensional subspace. As we shall see later, these techniques transform the original given integral equation into a linear system of equations that can be solved directly or iteratively by a corresponding numerical procedure.

Section 2.3.3.2.1
Section 2.3.3.2.2
Section 2.3.3.2.2
Section 2.3.3.2.4.2

SEQUENCES OF FUNCTIONS. Now, let $(S, \|\cdot\|)$ be a linear normed function space. If we declare, as consequence of the norm, a real valued function $\Delta : S \times S \to \mathbb{R}^{\geq 0}$ that satisfies the requirements of a metric, then it is possible to measure the distance between two functions $f, g \in S$. This construct enables us to generate sequences of functions $(f_n)_{n\in\mathbb{N}}, (g_n)_{n\in\mathbb{N}}$, to compare their values and give answers about their limit behavior, see Figure 2.3.

Linear Normed Space (860)
Norm (860)
Metric (866)

DEFINITION 2.2 (Limit of Sequences of Functions) *Suppose* $(f_n)_{n\in\mathbb{N}}$ *is a sequence of functions defined on a common domain* $\mathrm{Dom}(f_n)$ *of a function space* S. *We denote* $(f_n)_{n\in\mathbb{N}}$ *to be* convergent, *if there exists a member* $f \in S$ *for which, given any* $\epsilon > 0$, *a number* N *can be found, such that*

Section 2.2.4

$$\|f_n(x) - f(x)\| < \epsilon, \quad \forall n > N, \tag{2.5}$$

then, we call the function f *the* limit function *of* f_n.

In the appendix is shown, that a linear space can equipped with various norms. But different norms result in different measures of size for a given element of a linear space, that is, the convergence of sequences of functions in a linear function space depends on the chosen norm. Now, one of the most important types of convergence in a function space is the *pointwise convergence*. It is defined as follows:

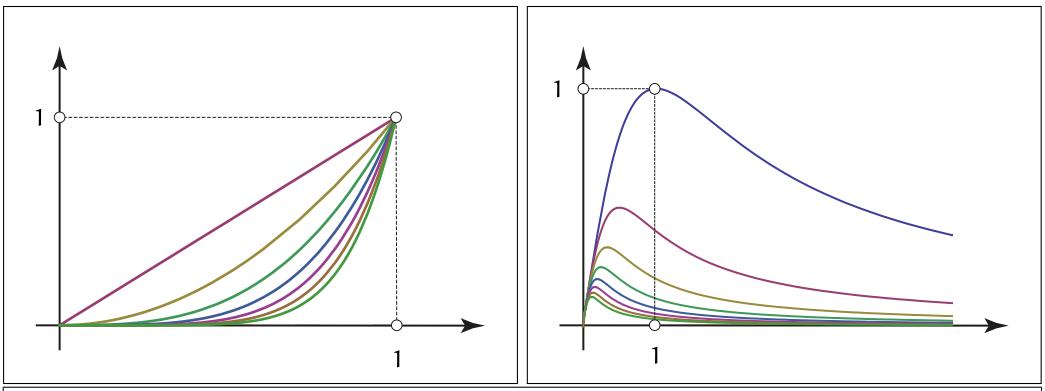


FIGURE 2.3: SEQUENCES OF FUNCTIONS. The first seven elements of the sequences of functions $f_n = x^n, n \geq 0$ and $f_n = \frac{2x}{1+n^2x^2}, n \geq 1$.

DEFINITION 2.3 (Pointwise Convergence) *A sequence* $(f_n)_{n\in\mathbb{N}}$ *of functions of a linear normed function space* $(\mathcal{S}, \|\cdot\|)$ *is said to* converge pointwise *to the limit function* f, *in sign* $f_n \to f$, *if for every* $\epsilon > 0$, *there exists a number* $N(\epsilon, x)$, *depending on* x *and* ϵ, *such that it holds:* Normed Space (860)

$$\|f_n(x) - f(x)\| < \epsilon \tag{2.6}$$

for all $n > N(\epsilon, x)$ *and* $x \in \mathrm{Dom}(f_n)$.

Let us show the concept of pointwise convergence at a famous example: the sequence of power functions defined on the unit interval:

EXAMPLE 2.4 *Let us consider the real-valued sequence of functions,* $(f_n)_{n\in\mathbb{N}_0}$, *with* $f_n(x) = x^n \in C[0,1]$. *With respect to the* supremum norm, *defined by,*

$$\|f_n\|_\infty \stackrel{\mathrm{def}}{=} \sup_{x\in[0,1]} |f_n(x)|, \tag{2.7}$$

the sequence $(f_n)_{n\in\mathbb{N}_0}$ *converges pointwise to the limit function* f, *given by,*

$$f(x) \stackrel{\mathrm{def}}{=} \begin{cases} 0 & \text{if } x \in [0,1) \\ 1 & \text{if } x = 1, \end{cases} \tag{2.8}$$

as it holds:

$$\|f_n(x) - f(x)\|_\infty = \sup_{x\in[0,1)} |f_n(x) - f(x)| \tag{2.9}$$

$$= |x^n| < \epsilon, \tag{2.10}$$

for all $x \in [0,1)$*, and* $\|f_n(1) - 1\|_\infty = 0$.

From this observation we can conclude: f_n *converges pointwise with respect to the supremum norm* $\|f_n\|_\infty$*, towards the limit function* f *that takes the value zero in the half-open interval* $[0,1)$ *and is exactly one for* $x = 1$*. Obviously, the sequence* f_n *converges towards a discontinuous function, that is, the limit function is not a member of the space* $C[0,1]$*, see Figure 2.4.*

Norm (860)

C[0, 1] (28)

The fact, that pointwise convergence does not guarantees the retention of good properties, such as for example continuity of the members of a sequence of functions, is not desirable. Pointwise convergence of a sequence of functions is a too weak property that can lead to problems when integrating a sequence of functions in sense of Riemann.

Discontinuous Function (868)

Riemann Integral (876)

Let us assume, we have a sequence of continuous functions $(f_n)_{n\in\mathbb{N}}$ that converges to a limit function f, such as in the case of a power series, or an infinite series of polynomials that describes a particular function from an infinite-dimensional function space. A good property, that the Riemann integral should fulfill, could be:

$$\int_0^1 f(x)\,dx = \int_0^1 \lim_{n\to\infty} f_n(x)\,dx = \lim_{n\to\infty} \int_0^1 f_n(x)\,dx, \tag{2.11}$$

that is, instead to integrate the limit function—which can be a complicated task—it is also possible to integrate the members of the sequence of functions, and to compute their limit. Thus, for integrating power series, the concept of the Riemann integral makes only sense, if it allows the exchange of the limit and the integral.

Riemann Integral (876)

Now, let us consider the Riemann integral of the sequence of functions f_n given by:

$$f_n(x) = \begin{cases} 1 & \text{if } x \in \{q_1, \ldots, q_n\} \\ 0 & \text{if } x \in [0,1] \setminus \{q_1, \ldots, q_n\}, \end{cases} \tag{2.12}$$

where $\{q_1, \ldots, q_n\}$ are the rational numbers in $[0,1]$.

It is known from calculus that the lower and the upper Riemann-Darboux integral of each function f_n exist and have the same value, that is, each member of the sequence of functions f_n is Riemann-integrable. In Example 2.38 we will see that the limit function of f_n under pointwise convergence, the so-called *Dirichlet function*, is not Riemann-integrable. Obviously, pointwise convergence does not support the exchange of the limit and the integral for the mathematical concept of the Riemann integral. Since this effect is not desirable, we need a stronger convergence type for the Riemann integral that satisfies this requirement: the concept of *uniform convergence*.

Section 2.2.4

DEFINITION 2.4 (Uniform Convergence) *A sequence* $(f_n)_{n\in\mathbb{N}}$ *of functions of a linear normed function space* $(S, \|\cdot\|)$ *is said to* converge uniformly *to the limit function* f*, in sign* $f_n \rightrightarrows f$*, if for every* $\epsilon > 0$*, there exists a number* $N(\epsilon)$*, only depending on* ϵ*, such that it holds:*

Linear Normed Space (860)

$$\|f_n(x) - f(x)\| < \epsilon \tag{2.13}$$

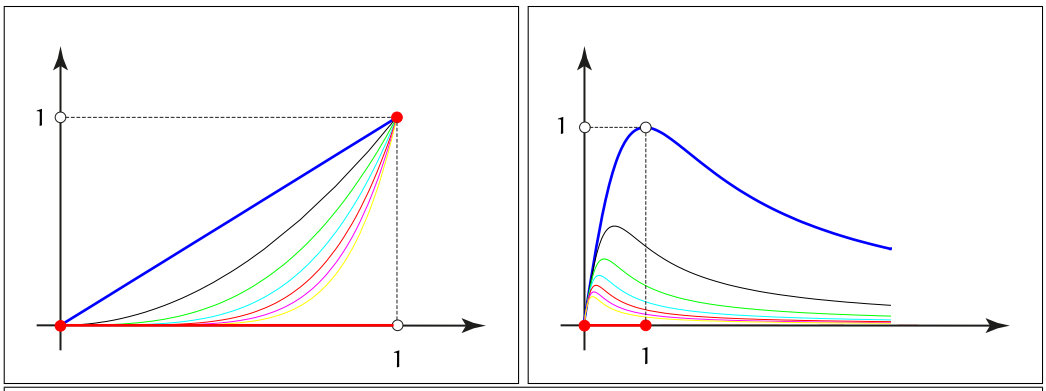


FIGURE 2.4: POINTWISE AND UNIFORM CONVERGENCE. Elements of the sequences of functions $f_n = x^n$ and $g_n = \frac{2x}{1+n^2x^2}$. Obviously, f_n is pointwise convergent to 0 in the interval $[0,1)$, but neither pointwise nor uniformly convergent on the closed interval $[0,1]$. The sequence g_n is uniformly convergent towards 0 on $[0,1]$ with respect to the infinity norm.

for all $n > N(\epsilon)$. *For a uniform convergent sequence of functions we write*

$$f_n \Rightarrow f. \tag{2.14}$$

As we can see from Figure 2.4, uniform convergence of a sequence $(f_n)_{n\in\mathbb{N}}$ of functions has a very simple geometric interpretation: For any given ϵ, all functions f_n lie in a tube of diameter 2ϵ located symmetrically about the limit function f, for n greater than $N(\epsilon)$, which now depends on ϵ, but no more on x.

EXAMPLE 2.5 *Let us consider once more the sequence of functions from Example 2.4. As we have seen,* f_n *converges with respect to the supremum norm pointwise to a limit function.* Supremum Norm (33)

With respect to the supremum norm, *defined by,*

$$\|f_n\|_\infty \stackrel{def}{=} \sup_{x\in[0,1]} |f_n(x)|, \tag{2.15}$$

the sequence of functions is not convergent to the limit function from Equation (2.8) since it holds:

$$\|f_n - f\|_\infty = \sup_{x\in[0,1]} |f_n(x) - f(x)| = 1. \tag{2.16}$$

As this example shows, in infinite-dimensional linear spaces convergence defined by a certain norm can be stronger than by another norm.

REMARK 2.2 *As we will see in the following, the mathematical concept of the uniform convergence of function sequences plays a fundamental role in mathematics, since it transforms important properties of a sequence of functions* $(f_n)_{n\in\mathbb{N}}$*, such as continuity, differentiability, and integrability to the limit function* f*.*

When deriving the Lebesgue integral, thus, the integral concept underlying the theory of linear integral equations, we will see that uniform convergence is already a very strong requirement for functions to be Lebesgue-integrable. There, we will see that the Lebesgue integral allows the exchange of limit and integral under much weaker conditions than uniform convergence of a sequence of functions. This then leads to a broader class of integrable functions, than those attached to the Riemann-integral.

REMARK 2.3 *From the above example we conclude:*

Linear Normed Space (860)

i) It is important to specify which norm is being used when discussing convergence of sequences of functions in linear normed spaces, as convergence with respect to one norm does not necessarily imply convergence with respect to another norm.

ii) It is possible that the limit of a sequence of functions in a linear normed space is not an element of the given space.

As we shall see further below, the behavior of a sequence of functions as described in Example 2.5, where the limit of $(f_n)_{n\in\mathbb{N}}$ is not an element of $\mathcal{S}$, is undesirable. For a number of reasons, we make a strong distinction between spaces in which sequences will converge towards a member of the space and spaces in which we can construct sequences, whose limit is not an element of $\mathcal{S}$.

COMPLETENESS, BANACH AND HILBERT SPACES. From the definition of the limit of a sequence we conclude that convergent sequences are characterized by the fact that the distance of adjacent sequence members—measured in terms of a norm underlying the linear space—decreases if their indices increase. So, a sequence converges normally towards an element within space. Unfortunately, this observation is not applicable to all linear spaces. As demonstrated in functional analysis, in many linear spaces sequences may be constructed, which conform to the condition that the distance between their members gets smaller with increasing index, and thus converge, but whose limit is not an element of the space.

Limit of a Sequence (867)

Transcribed to our problem of finding a solution to the light transport equation this means: Even if we are able to construct a sequence of functions that are all more or less approximate solutions of the SLTE, then it is not guaranteed that the limit of this sequence converges to the exact solution of the SLTE.

For a number of reasons this kind of behavior is not welcome. So a strict differentiation is made in functional analysis between spaces containing sequences behaving like

convergent sequences and converging towards a limit belonging to the linear space $\mathcal{S}$, and spaces containing sequences that converge towards a value not lying in $\mathcal{S}$. This idea necessitates the introduction of the concept of the *completeness* of a linear space, leading to the construction of the most important spaces in functional analysis: *Banach* and *Hilbert spaces*. A tool, that can be used to check the completeness of a linear space is given by the concept of the *Cauchy sequence*. Cauchy sequences have, as we see further below, the properties, that their definition makes no reference to the notion of convergence or that of a limit, since it is possible, that they do not converge. Formally, Cauchy sequences are defined as follows:

DEFINITION 2.5 (Cauchy sequence) *A sequence* $(x_n)_{n\in\mathbb{N}}$ *of a linear normed space* $(\mathcal{S}, \|\cdot\|)$ *is termed a* Cauchy sequence *if, for any given* $\epsilon \in \mathbb{R}, \epsilon > 0$*, there exists a number* $N(\epsilon) \in \mathbb{N}$*, such that* $\|x_n - x_m\| < \epsilon$ *whenever* $m, n > N(\epsilon)$. Linear Normed Space (860)

With the concept of the Cauchy sequence, we are now ready to introduce the fundamental concept of the *Banach space*.

DEFINITION 2.6 (Banach Space) *Let* $(\mathcal{S}, \|\cdot\|)$ *be a linear normed space. If* $\mathcal{S}$ *satisfies the condition that every given Cauchy sequence converges to an element of* $\mathcal{S}$*, then* $(\mathcal{S}, \|\cdot\|)$ *is referred to as a* complete, linear normed space*, also denoted as a* Banach space. Linear Normed Space (860)

EXAMPLE 2.6 (The Complete Space $\mathbb{R}^n$) *Since any Cauchy sequence in* $\mathbb{R}^n$ *converges towards an element of the space,* $(\mathbb{R}^n, \|\cdot\|_2)$ *is a complete linear normed space. This property is fundamental, as it holds only in complete spaces. In incomplete linear spaces, such as* $\mathbb{Q}$*, this property is not valid, as we can construct Cauchy sequences that do not converge to an element of* $\mathbb{Q}$*. For example, if we consider the sequence* $x_{n+1} = \frac{1}{2}\left(x_n + \frac{2}{x_n}\right)$ *then all elements of this sequence are rational numbers, but it holds* $\lim_{n\to\infty} x_n = \sqrt{2} \notin \mathbb{Q}$. Linear Normed Space (860)

EXAMPLE 2.7 (The Banach Space $C([a,b])$) *Let* f *be a function of* $C[a,b]$*. Using the supremum norm* $\|\cdot\|_\infty$ *then it holds:* $C[a,b]$ (28), Supremum Norm (33)

$$\|f\|_\infty \stackrel{def}{=} \sup_{x\in[a,b]} |f(x)| \tag{2.17}$$

$$= \max_{x\in[a,b]} |f(x)|, \quad f \in C([a,b]), \tag{2.18}$$

that is, $(C([a,b]), \|\cdot\|_\infty)$ *becomes a linear, normed space. Obviously, this space is complete, as* $\|\cdot\|_\infty$ *fulfills the conditions required to a norm and every Cauchy sequence* $(f_n)_{n\in\mathbb{N}}$ *converges uniformly towards a limit function* $f \in C([a,b])$*. We leave this proof to the interested reader.* Linear Normed Space (860)

EXAMPLE 2.8 (The Banach Space $B([a,b])$) *Let* f *be a function of* $B[a,b]$*. Using the* $B[a,b]$ (28)

Supremum Norm (33) *supremum norm* $\|\cdot\|_\infty$ *then it holds:*

$$\|f\|_\infty \overset{def}{=} \sup_{x\in[a,b]} |f(x)|, \quad f \in B([a,b]), \tag{2.19}$$

Linear Normed Space (860) *that is,* $(B([a,b]), \|\cdot\|_\infty)$ *becomes a linear, normed space. Similar to the space* $C([a,b])$*, also the space of all bounded functions on the interval* $[a,b]$ *is a complete linear normed space.*

Linear Normed Space (860) REMARK 2.4 *Every finite-dimensional, linear, normed space* $(S, \|\cdot\|)$ *is complete. Concretely, a complete linear normed space is a space that does not have any holes in it.*

Linear Space (854) DEFINITION 2.7 (Hilbert Space) *If however we provide an arbitrary given linear space*
Inner Product (859) S *with an inner product, then every inner product space* $(S, \langle\cdot,\cdot\rangle_S)$ *may be equipped*
Inner Product Space (859) *with a norm via the construction*

$$\|x\| \overset{def}{=} \sqrt{\langle x, x\rangle_S}, \tag{2.20}$$

which makes it to a linear normed space, also referred to as a pre-Hilbert space*. If this pre-Hilbert space also satisfies the condition of completeness, then it is referred*
Completnesse (34) *to as a* Hilbert space*.*

C[a, b] (28) EXAMPLE 2.9 (Extending a Banach Space to a Pre-Hilbert Space) *As known from above,* $(C[a,b]), \|\cdot\|_\infty)$ *is a Banach space, but not a Hilbert space, since the norm* $\|\cdot\|_\infty$ *was not induced by an inner product. However, if we equip* $C([a,b])$ *with the inner product*

$$\langle f, g\rangle_{C([a,b])} \overset{def}{=} \int_{[a,b]} f(x)\, g(x)\, dx, \tag{2.21}$$

Riemann Integral (877) *where the integral is an ordinary Riemann integral, then, due to Relation 2.20 the*
Pre-Hilbert Spce (36) *space* $(C([a,b]), \|\cdot\|)$ *becomes an inner product space, namely a pre-Hilbert space.*

Note: $(C([a,b]), \|\cdot\|)$ *is not a Hilbert space, but as a pre-Hilbert space it provides us with the concept of orthogonality.*

REMARK 2.5 *It is clear that any Hilbert space is also a Banach space. But Hilbert spaces possess—due to the concept of orthogonality—a richer structure than Banach spaces.*

REMARK 2.6 *Completeness is a very important characteristic of linear spaces. It is precisely this type of a linear space that possesses a number of useful properties missing in a non-complete space. So, it enables a number of solution methods for various different analytical problems firmly anchored within real world.*

In the following we are interested in complete, linear normed function spaces,
Lebesgue Integral (105) *where the inner product is defined via the concept of the Lebesgue integral, which is of*

enormous relevance for the treatment of global illumination algorithms. Depending on norms underlying these function spaces, we get the only function space that may be constructed over the Lebesgue integral and which in addition conforms to the conditions required to a Hilbert space. It is just this space that allows the development of radiosity and ray tracing algorithms as approximate solution strategies for the global illumination problem.

Chapter 10
Chapter 8

As we have seen in the Examples 2.1 and 2.3, the complete, linear, normed spaces of continuous respectively bounded functions on the closed set $[a, b] \subset \mathbb{R}$ are of infinite dimension, that is, it is not possible to find a finite set of continuous functions in $C[a, b]$ respectively in $B[a, b]$ that spans the associated function spaces. The best we can achieve is to construct an infinite sequence of continuous respectively bounded functions such, that any function of $C[a, b]$ respectively $B[a, b]$ can be approximated closely by a finite linear combination of these functions. Arbitrary Hilbert spaces and the function spaces lying in our interest are also of this type. Hence, our aim is to extend the idea of a basis to infinite-dimensional spaces, which results in an infinite, but countable set of elements of the spaces.

$C[a, b]$ (28)
$B[a, b]$ (28)
Linear Combination (857)
Countable Set (827)

From an inner product space, a number of important concepts of vector algebra and calculus may be transferred onto the concept of the Hilbert space. Thus, not only the fundamental *Cauchy-Schwartz inequality* holds in any inner product space, but also the concept of *orthogonality*, well-known from the Euclidean space $\mathbb{R}^n$.

Inner Product Space (859)
Cauchy-Schwartz Inequality (859)
Section 2.3.3.2.2

DEFINITION 2.8 (Orthonormal Basis) *Let $\mathcal{S}$ be an infinite-dimensional inner product space. A system $\mathcal{B}_\phi^\infty = \{\phi_i \mid i \geq 1, \langle \phi_i, \phi_j \rangle = 0,\ i \neq j\}$ of maximal orthogonal elements of $\mathcal{S}$ with $\|\phi_i\| = 1$ is called an* orthonormal basis *of $\mathcal{S}$ if there exist no other non-zero element $\phi \in \mathcal{S}$ that is orthogonal to all elements of $\mathcal{B}_\phi^\infty$.*

Inner Product Space (859)
Orthogonality (859)

Now let us assume $(\mathcal{S}, \langle \cdot, \cdot \rangle_\mathcal{S})$ be an infinite-dimensional inner product space. Considering the finite dimensional subspace $\mathcal{U}$ of $\mathcal{S}$, spanned by the orthonormal set $\mathcal{B}_\phi^n = \{\phi_1, \ldots, \phi_n\}$, then any $u \in \mathcal{U}$ can be written as a linear combination of elements of $\mathcal{B}_\phi^n$, thus,

Linear Subspace (855)
Linear Combination (857)

$$u = \sum_{i=1}^{n} \alpha_i \phi_i, \tag{2.22}$$

where α_i are uniquely defined real or complex numbers.

$\mathbb{R}, \mathbb{C}$ (827)

Extending the orthonormal set $\mathcal{B}_\phi^n = \{\phi_1, \ldots, \phi_n\}$ successively by orthonormal members $\phi_{n+1}, \phi_{n+2}, \cdots$ from $\mathcal{S}$ results in a sequence $(\mathcal{B}_\phi^n)_{n \in \mathbb{N}}$ of orthonormal bases for a sequence of finite dimensional subspaces $(\mathcal{U}_n)_{n \in \mathbb{N}}$ of our originally given inner product space $\mathcal{S}$. Since $\mathcal{S}$ is of infinite dimension, a member $f \in \mathcal{S}$ can only be represented by a linear combination of countably infinite orthonormal elements of $\mathcal{S}$, that is,

Inner Product Space (859)
Countable Set (827)

$$f = \sum_{i=1}^{\infty} \alpha_i \phi_i, \tag{2.23}$$

where α_i are uniquely defined real or complex numbers and Expression (2.23) must be interpreted as the limit $n \to \infty$ of the n^{th}-partial sums u_n defined by Equation (2.22). Thus any partial sum u_n is an approximation of f, and this approximation improves, if n increases. That is, contrary to finite dimensional spaces, the best we can do in infinite-dimensional spaces to represent an element of $\mathcal{S}$, is to construct an infinite sequence $(u_n)_{n\in\mathbb{N}}$ of members from $\mathcal{S}$ with the property, that any element of $\mathcal{S}$ can be approximated arbitrarily closely by a finite linear combination of these members. This statement is a conclusion from the *Best Approximation Theorem*, which also provides information about the choice of the coefficients $\alpha_i, i \geq 1$:

Inner Product Space (859) Orthonormal Function (37)

THEOREM 2.1 (Best Approximation Theorem) *Let $\mathcal{S}$ be an inner product space and $\mathcal{B}_\phi^\infty = \{\phi_1, \phi_2, \ldots\}$ be an orthonormal set from $\mathcal{S}$. Furthermore, let f be a member of $\mathcal{S}$ and the sequences u_n and $\tilde{u}_n \in \mathcal{U} \leq \mathcal{S}$ be given by:*

$$u_n = \sum_{i=1}^{n} \alpha_i \phi_i \quad \text{and} \quad \tilde{u}_n = \sum_{i=1}^{n} \langle f, \phi_i \rangle_{\mathcal{S}} \phi_i, \tag{2.24}$$

where $\langle f, \phi_i \rangle_{\mathcal{S}}$ are denoted as the Fourier coefficients *of u_n and $\tilde{u}_n$, and α_i are arbitrary real or complex numbers. Then it holds:*

$$\|f - \tilde{u}_n\| = \inf_{u_n \in \mathcal{U}} \|f - u_n\|. \tag{2.25}$$

PROOF 2.1 *We omit the proof and point to [22, Berezansky & al. 1996].*

Best Approximation Theorem (38) pre-Hilbert Space (36)

EXAMPLE 2.10 (The Linear Function Space $C(a, b)$) *Due to the Best Approximation Theorem any real-valued function f of the pre-Hilbert space $(C([a, b]), \|\cdot\|)$ from Example 2.9 can be approximated by:*

$$f(x) = \sum_{i=1}^{n} \langle f(x), \phi_i(x) \rangle_{C([a,b])} \phi_i(x) \tag{2.26}$$

$$= \sum_{i=1}^{n} \left(\int_a^b f(x)\phi_i(x)\, dx \right) \phi_i(x) \tag{2.27}$$

Orthonormal Basis (37)

with respect to an n-dimensional orthonormal basis $\{\phi_1(x), \ldots, \phi_n(x)\}$ of $\mathcal{U}_n \leq C([a, b])$.

Now, an n-dimensional basis of $\mathcal{U}_n \leq C([a, b])$ can be defined via a set of piecewise continuous and bounded functions ϕ_i, see Example 2.3, given by

$$\phi_i(x) \stackrel{def}{=} \begin{cases} \sqrt{\frac{n}{b-a}} & x \in \left[a + (i-1)\frac{b-a}{n}, a + i\frac{b-a}{n}\right] \\ 0 & \text{otherwise,} \end{cases} \tag{2.28}$$

with $1 \leq i \leq n$. *The orthogonality of the functions can easily be shown by*

$$\begin{aligned} \langle \phi_i(x), \phi_j(x) \rangle &= \int_{a+(i-1)\frac{b-a}{n}}^{a+i\frac{b-a}{n}} \phi_i(x)\phi_j(x)\, d(x) && (2.29) \\ &= \int_{a+(i-1)\frac{b-a}{n}}^{a+i\frac{b-a}{n}} \sqrt{\frac{n}{b-a}} \cdot 0\, d(x) && (2.30) \\ &= 0 && (2.31) \end{aligned}$$

and the unit length of the functions ϕ_i *can be seen from*

$$\begin{aligned} \langle \phi_i(x), \phi_i(x) \rangle &= \int_{a+(i-1)\frac{b-a}{n}}^{a+i\frac{b-a}{n}} \phi_i(x)\phi_i(x)\, d(x) && (2.32) \\ &= \frac{n}{b-a} \int_{a+(i-1)\frac{b-a}{n}}^{a+i\frac{b-a}{n}} d(x) && (2.33) \\ &= 1. && (2.34) \end{aligned}$$

Extending the Best Approximation Theorem to the case where the orthonormal set is infinite then leads to the *Fourier Series Theorem.*

THEOREM 2.2 (Fourier Series Theorem) *Let* $\mathcal{S}$ *be a Hilbert space and let* $\mathcal{B}_\phi^\infty = \{\phi_1, \phi_2, \ldots\}$ *be a countably infinite set of orthonormal elements from* $\mathcal{S}$. *Then any* $f \in \mathcal{S}$ *can be written in form of an infinite series of members of* $(\mathcal{B}_\phi^\infty)$, *that is,*

$$f = \sum_{i=1}^{\infty} \langle f, \phi_i \rangle_{\mathcal{S}}\, \phi_i \tag{2.35}$$

if and only if $\mathcal{B}_\phi^\infty$ *is an orthonormal basis, in other words, if it is a maximal orthonormal set in* $\mathcal{S}$.

Orhtonormal Functions (37)

PROOF 2.2 *We omit the proof and point to [169, Reddy 1998].*

REMARK 2.7 *The Fourier Series Theorem plays an important role when we are interested in representing a* BRDF *by spherical harmonics. So, we will show in Chapter 4, that a BRDF is a function of two directional variables, that makes a statement about how much light is reflected at an opaque surface. Under certain circumstances it is a very common method of evaluating the SLTEV, the stationary light transport equation within a vacuum, by projecting the integrand, i.e. the BRDF together with the radiance function* L_o, *onto spherical harmonics—the spherical analogues of sines and cosines—which forms a smooth orthonormal basis for functions defined on the unit sphere.*

BRDF (320)
Spherical Harmonics (124)
Orthonormal Basis (37)

REMARK 2.8 *From the above statements we may now conclude, that a Hilbert space*

Hilbert Space (36)

corresponds to a Banach space equipped with a norm induced via an inner product. Due to the inner product involved, a Hilbert space possesses properties which do not apply to general Banach spaces.

Banach Space (35)

Thus, a Hilbert space provides the closest analogue to the Euclidean space $\mathbb{R}^n$ among the function spaces, and its geometry is closely modeled to that of $\mathbb{R}^n$. As we have seen, it is possible, by means of the integral, to induce a norm via an inner product, which provides the concept of orthogonality between functions. This gives any Hilbert space many pleasant properties, such as the Pythagorean theorem *as well as the concept of* orthogonal projections*, which plays a vital role in particular in the radiosity procedure of global illumination theory. There, we obtain suitable solutions from finite-dimensional linear spaces, which correspond to the projection of infinite-dimensional function spaces where the structure of the Hilbert space is required.*

EXAMPLE 2.11 (Spectral Power Distributions) *From Example 2.3 we know, that accurate color rendering in computer graphics is based on the concept of the* spectral power distribution*, except of line spectra, a bounded function defined over the visible spectrum of light. Such functions are elements of the function space $B([a,b], \|\cdot\|)$ with $[a,b] \equiv \Lambda$, and $\Lambda = [380\ nm, 780\ nm]$, and the norm is given via the inner product from Equation 2.21. Due to the power of its domain, it should be clear that it is not possible to represent a SPD exactly in any rendering algorithm, that is, a SPD may only be addressable for computation via approximations.*

B[a, b] (28)

Therefore, the infinite-dimensional space of spectral power distributions is projected onto the n-dimensional function space spanned by a set of orthonormal functions $\{\phi_1, \cdots, \phi_n\}$. As shown in Example 2.10, the functions ϕ_i can be chosen as piecewise bounded function, where the visible spectrum is partitioned into a set of n bins.

Orthonormal Basis (37)

Under these circumstances, the Best Approximation Theorem is applicable, and the following clearly applies for a spectral power distribution $S \in B(\Lambda)$:

$$S(\lambda) = \sum_{i=1}^{n} \langle S(\lambda), \phi_i(\lambda) \rangle \phi_i(\lambda) \tag{2.36}$$

with

$$\langle S(\lambda), \phi_i(\lambda) \rangle \stackrel{def}{=} \int_{\Lambda} S(\lambda)\phi_i(\lambda)\, d\lambda, \tag{2.37}$$

where the spectrum of every light source in a scene may be formulated via the n-dimensional vector $(\langle S(\lambda), \phi_1(\lambda)\rangle, \ldots, \langle S(\lambda), \phi_n(\lambda)\rangle)^T \in \mathbb{R}^n$.

Depending on the accuracy of the approximation, any orthonormal system defined over $B(\Lambda)$ is available for the choice of the basis functions $\{\phi_1, \ldots, \phi_n\}$. Thus, apart from the piecewise constant box functions

$$\phi_i(\lambda) \stackrel{def}{=} \begin{cases} 1 & \lambda_i < \lambda < \lambda_{i+1}, 1 \le i \le n \\ 0 & \text{otherwise,} \end{cases} \tag{2.38}$$

the monochromatic functions defined solely for one single sample

$$\phi_i(\lambda) \stackrel{def}{=} \begin{cases} 1 & \lambda = fix, 1 \leq i \leq n \\ 0 & otherwise \end{cases} \tag{2.39}$$

also represent basis functions that may be efficiently calculated, however only for the approximation of a small class of real spectra. According to [153, Peercy 1993] and [206, Sun & al. 1999] among the more adequate, though computationally more expensive, methods are orthonormal systems based on the following system of trigonometric functions

$$\begin{cases} \phi_1(\lambda) \stackrel{def}{=} 1 \\ \phi_2(\lambda) \stackrel{def}{=} \cos\left(2\pi\frac{(\lambda-\lambda_{min})}{(\lambda_{max}-\lambda_{min})}\right) \\ \phi_3(\lambda) \stackrel{def}{=} \sin\left(2\pi\frac{(\lambda-\lambda_{min})}{(\lambda_{max}-\lambda_{min})}\right) \\ \dots \end{cases} \tag{2.40}$$

or orthonormal systems, whose elements may be described via the Gauss functions

$$\phi_i(\lambda) \stackrel{def}{=} e^{-\ln 2(2(\lambda-\lambda_{c,i})w_i)^2} \qquad 1 \leq i \leq n \tag{2.41}$$

where $\lambda_{c,i}$ and w_i denote parameters here not specified further.

2.1.2 THE SCENE MODEL IN RENDERING ALGORITHMS

In view of generating and analyzing realistic rendering procedures the Euclidean space $\mathbb{R}^3$ plays an important role. As we will show in one of the chapters to follow below, realistic rendering algorithms require not only many parameters, but also the description of a virtual scene, defined by a large number of differently-shaped objects in $\mathbb{R}^3$. Such a scene object may be visualized as composed of either simple geometric elementary structures, such as points, lines, and triangles, or simple respectively more or less complex mathematical constructs of differential geometry, such as parallelepipeds, spheres, cones, and tori, as well as special 2-dimensional surfaces, see Figure 2.5.

Chapter 8

We treat the scene, underlying any rendering algorithms, as a union of a finite number of 2-*dimensional surfaces* and 3-*dimensional volumes* within the space $\mathbb{R}^3$. Following this approach, our scenes are modeled as a set of finite volumes $\mathcal{V} = \{\mathcal{V}_1, \dots, \mathcal{V}_n\} \subset \mathbb{R}^3$ and their boundaries $\partial\mathcal{V}$, thus a finite set of surfaces describing the solid objects in the scene. We assume that all of these boundaries are closed and piecewise differentiable, where the space between these objects, an open set denoted by $\mathcal{V}^o = \mathcal{V} \setminus \partial\mathcal{V}$, can be filled with participating media.

Closed Set (864)
Open Set (864)

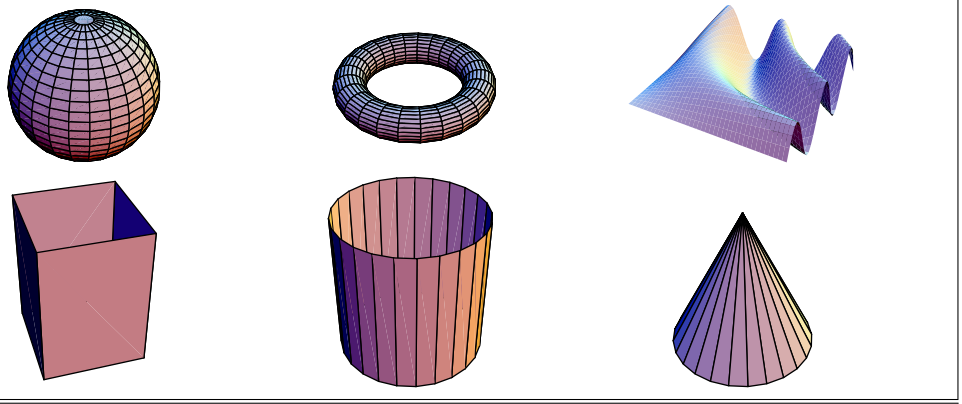


FIGURE 2.5: DIFFERENTLY-SHAPED OBJECTS IN $\mathbb{R}^3$. A sphere, a torus, and a surface as examples for 2-dimensional surfaces and a parallelepiped, a cylinder, and a cone as examples for 2-dimensional surfaces, piecewise differentiable except for their edges and dot singularities.

2.1.3 RAY SPACES AND FUNCTION SPACES ON RAYS

As already informally described in our introductory chapter, a classic ray tracing algorithm uses the theoretical concept of the mathematical ray to simulate the physical model of a light ray, known from ray optics. Such an algorithm fires a ray $\mathbf{r}$ into the underlying scene, where at possible hit points of $\mathbf{r}$ with object surfaces in the environment, the portion of incident illumination is determined. Because this amount of light is not only dependent on the quantity of light that directly arrives from existing light sources, but also of all other surfaces within the scene, a ray tracing procedure also needs to know information about the light field around a hit point. Via the mechanism of recursive ray generation at these intersections points, the procedure then samples so to say, the light distribution in the whole scene via rays and contributes the light traveling along these rays from all reachable points to the illumination at the point to be shaded, see Figure 2.6.

Cartesian Product (829)
Linear Normed Space (860)
Ray Spaces (43)
Function Spaces on Rays (46)
Lebesgue Integral (105)
Incident & Exitant Function (48)
Chapter 5

To describe and analyze this process of light transport formally, it is required to capture the concept of the light ray and of light energy, carried by a ray, in a more mathematical way, namely as Cartesian products of sets and elements from linear normed spaces. These so-called *ray spaces*, defined over all rays starting at points in a scene, are the natural bases of our light transport calculations. They allow the construction of function spaces, which—equipped with a norm based on the Lebesgue integral—can be used to specify the interaction of photon events at object surfaces. So, we will also introduce the concept of the incident and exitant function, of fundamental relevance for the definition of the operator model of light and importance transport as well as for the field of radiometry. Incident and exitant functions also play a central role in the measurement of radiance,

quasi the amount of light, arriving at or leaving from a scene point in a particular direction.

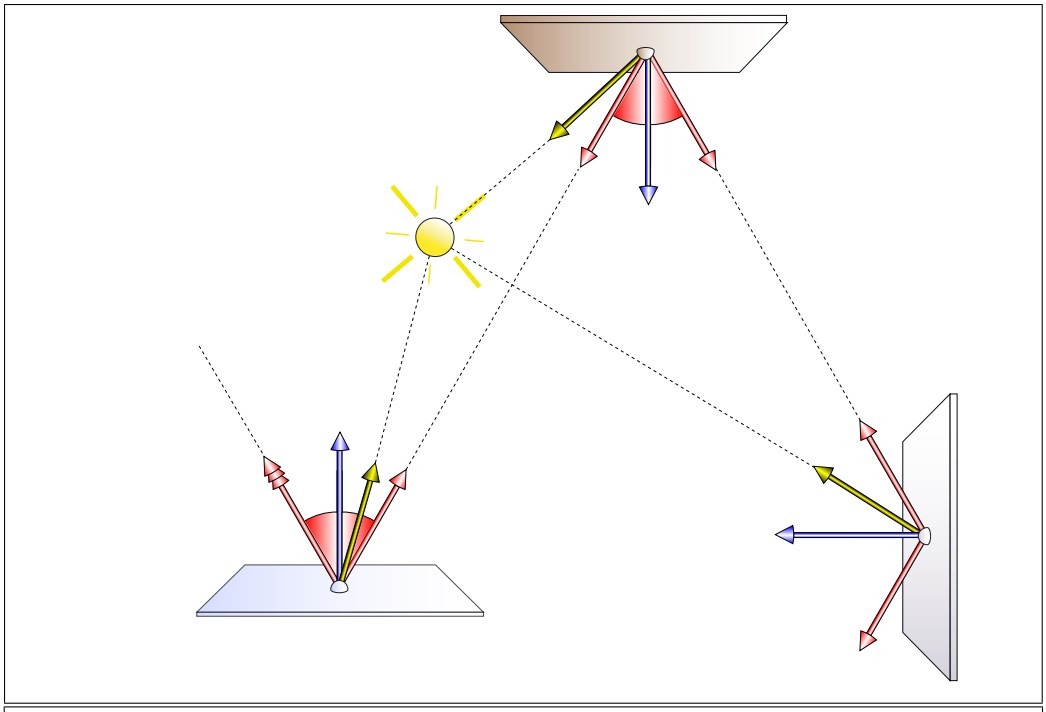


FIGURE 2.6: THE PRINCIPLE OF RAY TRACING BASED ALGORITHMS. At the first hit point of a ray with the closest scene object, the algorithm can generate, depending on the properties of the material of the concerned surface, up to three new types of rays: a reflection ray, a refraction ray, and a shadow ray. The algorithm estimates the incoming light at the intersection point of the primary ray with an object and combines this information to a contribution to the final color of the pixel. The computation of the light contributions of the reflected as well as the refracted ray are taken recursively until a diffuse surface is hit, the ray doesn't intersect an object within the scene, or the intensity of the ray is below a threshold value respectively the recursion depth of ray generation exceeds a predefined value.

THE RAY SPACES $\mathcal{R}^{\partial\mathcal{V}}, \mathcal{R}^{\mathcal{V}^o}$ AND $\mathcal{R}$. Our definition of a mathematical ray, given by $\mathbf{r} = \mathbf{x} + \alpha\omega$, implies the representation of $\mathbf{r}$ based on a Cartesian product of two sets: a set of starting points and a set of directions. As the transport of light in participating media is different from that in a vacuum—in participating media, light traveling between points can also be e.g. absorbed by the media, which is not the case in free space—we must distinguish between a description of light transport in free space and the light transport in participating media. Hence, we want explicitly distinguish between rays starting at object surfaces $\partial\mathcal{V}$ and rays that start at inner points of a volume. For that purpose, we now define three different types of ray spaces: the *ray space over the surfaces* $\partial\mathcal{V} \subset \mathcal{V}$,

$\partial\mathcal{V}$ (41)
$\mathcal{V}$ (41)

the *ray space over inner points of* $\mathcal{V}^o \subset \mathcal{V}$, and the *extended ray space*, defined on
$\mathcal{V}^o$ (41) $\mathcal{V} = \partial\mathcal{V} \cup \mathcal{V}^o$.

DEFINITION 2.9 (The Ray Spaces $\mathcal{R}^{\partial\mathcal{V}}$, $\mathcal{R}^{\mathcal{V}^o}$, and $\mathcal{R}$) *The* ray space over surfaces *of* $\mathcal{V}$,
$\partial\mathcal{V}$ (41) *denoted by* $\mathcal{R}^{\partial\mathcal{V}}$*, is defined as the set of all rays starting at points on surfaces in a given scene and going in any direction, that is,*

$$\mathcal{R}^{\partial\mathcal{V}} \stackrel{\text{def}}{=} \partial\mathcal{V} \times S^2 = \left\{(\mathbf{s}, \omega) |\, \mathbf{s} \in \partial\mathcal{V}, \omega \in S^2\right\}, \tag{2.42}$$

where S^2 *is the unit sphere around a point* $\mathbf{s} \in \partial\mathcal{V}$.

In the case where we consider the light transport in participating media our rays
$\mathcal{V}^o$ (41) *can also start at inner points* $\mathbf{x} \in \mathcal{V}^o$ *of a medium and the associated* ray space over inner points of $\mathcal{V}$*, also referred to as* $\mathcal{R}^{\mathcal{V}^o}$*, is defined by*

$$\mathcal{R}^{\mathcal{V}^o} \stackrel{\text{def}}{=} \mathcal{V}^o \times S^2 = \left\{(\mathbf{x}, \omega) |\, \mathbf{x} \in \mathcal{V}^o, \omega \in S^2\right\}. \tag{2.43}$$

The extended ray space*,* $\mathcal{R}$*, is then defined as the union of this two disjoint sets, that is,*

$$\mathcal{R} \stackrel{\text{def}}{=} \mathcal{R}^{\partial\mathcal{V}} \cup \mathcal{R}^{\mathcal{V}^o} = \left\{(\mathbf{x}, \omega) |\, \mathbf{x} \in \mathcal{V}, \omega \in S^2\right\}. \tag{2.44}$$

In Box A.1, we defined the construct of a direction ω by a line starting at the center of the unit sphere and passing through a point on the unit sphere. This construction now implies that the above ray spaces can also be explained via the following construction, similar to that in [221, Veach 1998]:

$$\mathcal{R}^{\mathcal{V}^o} \stackrel{\text{def}}{=} \mathcal{V}^o \times \mathcal{V}^o \tag{2.45}$$

$$\mathcal{R}^{\partial\mathcal{V}} \stackrel{\text{def}}{=} \partial\mathcal{V} \times \partial\mathcal{V} \tag{2.46}$$

$$\mathcal{R} \stackrel{\text{def}}{=} \mathcal{V} \times \mathcal{V}, \tag{2.47}$$

i.e., a ray $\mathbf{r} \equiv \mathbf{x} \rightarrow \mathbf{x}'$ is given with respect to points within a medium or with respect to points on the boundaries of the medium. This allows to abstract from the definition of a bounding sphere to the restriction of the scene as no light rays may be generated towards infinitely distant points. This representation of a ray is mostly useful when the rendered scene is closed and we are only interested in light transport between elements of $\mathcal{V}^o$ respectively $\partial\mathcal{V}$.

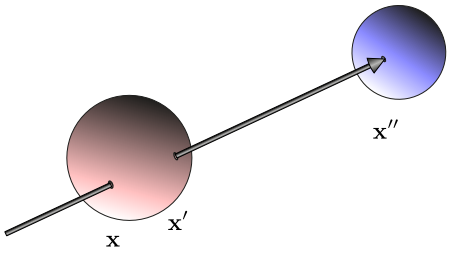
REMARK 2.9 *The construction of a ray space via the Relation (2.46) - (2.47) is of particular interest for the stochastic generation of shadow rays, as we will present it in Chapter 8. As already mentioned, shadow rays are generated from a point on a surface towards a point on a light source or an object surface of special relevance for the problem at hand. In Chapter 8 we will show, that the ray generation between two predetermined or stochastically chosen points on two surfaces is to be preferred to*

the method in which rays are generated via a starting point and a direction ω from S^2 since it can not be guaranteed that ω shows in direction to one of the interesting surfaces.

BOX 2.1 (The Visibility Function $\mathcal{V}$)

Any rendering algorithm based on ray tracing makes use of a special function that provides information about the intersection of a ray with objects in a scene: the *visibility function* $\mathcal{V}$. The *visibility function* $\mathcal{V}$ is defined as a mapping over any of the ray spaces $\partial\mathcal{V}$ or $\mathcal{V}^o$ into the set $\{0, 1\}$, where it holds:

$$\mathcal{V}(\mathbf{x} \leftrightarrow \mathbf{x}') \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } \mathbf{x} \text{ and } \mathbf{x}' \text{ are mutually visible} \\ 0 & \text{otherwise.} \end{cases} \tag{2.48}$$



Used in a rendering algorithm, the visibility function returns information about the set of points that are visible from a given point. More precisely, $\mathcal{V}(\mathbf{x} \leftrightarrow \mathbf{x}')$ decides if point $\mathbf{x}'$ is visible from point $\mathbf{x}$. The visibility function $\mathcal{V}$ is closely linked to another function, frequently used in rendering algorithms, the *ray-casting function* γ.

REMARK 2.10 (The Borel σ-algebras $\mathfrak{B}(\mathcal{R}^{\partial\mathcal{V}})$, $\mathfrak{B}(\mathcal{R}^{\mathcal{V}^o})$ and $\mathfrak{B}(\mathcal{R})$) *When discussing the transport of particles we have often to integrate functions defined on the above ray spaces. As we will show in Section 2.2, this requires the construction of measures that are defined on the Borel σ-algebras over the ray spaces. Now, due to Definition A.25 such σ-algebras are generated by all open subsets of $\mathcal{R}^{\partial\mathcal{V}}$, $\mathcal{R}^{\mathcal{V}^o}$, and $\mathcal{R}$, that is,*

Borel σ-algebra (865)
Measure (79)
σ-algebra (828)
Open Set (864)

$$\mathfrak{B}(\mathcal{R}^{\partial\mathcal{V}}) = \mathfrak{B}(\partial\mathcal{V} \times S^2) \tag{2.49}$$
$$\mathfrak{B}(\mathcal{R}^{\mathcal{V}^o}) = \mathfrak{B}(\mathcal{V}^o \times S^2) \tag{2.50}$$
$$\mathfrak{B}(\mathcal{R}) = \mathfrak{B}(\mathcal{V} \times S^2). \tag{2.51}$$

It will be these σ-algebras on which we construct the throughput measures, which makes it possible to integrate functions defined on the ray spaces.

Throughput Measures (94)

Any ray tracing algorithm needs information about the distribution of light in a given scene. Thus, it would be helpful if one could make a statement about the amount of light incident at a surface point. In [221, Veach 1998], an elegant and highly significant

technique was introduced, to reach this: the construction of a linear space based on one of the ray spaces from above. Such a linear space allows to construct all the tools and techniques needed for ray tracing procedures aimed at formal descriptions of phenomena relating to particle transport in participating media and under vacuum conditions: the *function spaces* $\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}})$, $\mathcal{L}(\mathcal{R}^{\mathcal{V}^o})$, and $\mathcal{L}(\mathcal{R})$ defined over the ray spaces $\mathcal{R}^{\partial\mathcal{V}}$, $\mathcal{R}^{\mathcal{V}^o}$ and $\mathcal{R}$.

Chapter 5

THE FUNCTION SPACES $\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}})$, $\mathcal{L}(\mathcal{R}^{\mathcal{V}^o})$, AND $\mathcal{L}(\mathcal{R})$. Based on the ray spaces from above, we now introduce function spaces that play an important role in our further considerations. Later, we will see, that the elements of these function spaces reflect the light distribution in the given scene and that they simplify the structure of our formulas by hiding the details of the ray representation. Due to [221, Veach 1998], they also emphasize that the representations are superficial decisions that can easily be changed and allow us to define concepts, whose meaning do not depend on how rays are represented.

DEFINITION 2.10 (The Function Spaces $\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}})$, $\mathcal{L}(\mathcal{R}^{\mathcal{V}^o})$, and $\mathcal{L}(\mathcal{R})$) *Let f be a real-valued function defined on one of the ray spaces $\mathcal{R}^{\partial\mathcal{V}}$, $\mathcal{R}^{\mathcal{V}^o}$, or $\mathcal{R}$. Let $\mathcal{R}^\star$ denote any of these ray spaces, then the set of all functions*

$$f : \mathcal{R}^\star \longrightarrow \mathbb{R} \tag{2.52}$$

with

$$\mathbf{r} = (\mathbf{x}, \omega) \longmapsto f(\mathbf{r}), \tag{2.53}$$

is referred to as $\mathcal{L}(\mathcal{R}^\star)$.

Linear Space (854) Function Space (28) Supremum Norm (33) Linear Normed Space (860)

Obviously $\mathcal{L}(\mathcal{R}^\star)$ are linear spaces, more precisely, the function spaces of real-valued functions *defined on one of the ray spaces $\mathcal{R}^\star$. Together with the supremum norm $\|\cdot\|_\infty$ the function spaces $\mathcal{L}(\mathcal{R}^\star)$ are linear normed spaces: the* linear normed function spaces *$(\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}}), \|\cdot\|_\infty)$, $(\mathcal{L}(\mathcal{R}^{\mathcal{V}^o}), \|\cdot\|_\infty)$ as well as $(\mathcal{L}(\mathcal{R}), \|\cdot\|_\infty)$.*

EXAMPLE 2.12 *An interesting example of a function of the function space $\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}})$ is the boundary distance function $d_{\partial\mathcal{V}}$ used in the ray-casting function γ, see Box 2.2.*

Chapter 5

REMARK 2.11 *Relating to $\mathcal{L}(\mathcal{R}^\star)$ as the function spaces defined on our ray spaces from Definition 2.9 we can make the following important statements:*

Linear Space (854)

i) It is easy to see that $\mathcal{L}(\mathcal{R}^\star)$ satisfy the conditions required to a linear space. As a function f from any of these function spaces may take values from $\mathbb{R}$, the spaces $\mathcal{L}(\mathcal{R}^\star)$ are closed with respect to vector addition and scalar multiplication, i.e. with f, also the additive inverse function $-f$ and the identity $f = 0$ exist.

$\mathcal{C}(\cdot)$ (28) Subspace (855)

ii) Obviously the spaces of continuous functions defined on $\mathcal{R}^\star$ and denoted by $C(\mathcal{R}^\star)$ are subspaces of the function spaces $\mathcal{L}(\mathcal{R}^\star)$, that is, $\mathcal{C}(\mathcal{R}^\star) \leq \mathcal{L}(\mathcal{R}^\star)$. Equipped

Supremum Norm (33) Banach Space (35)

with the supremum norm $\|\cdot\|_\infty$*, they become complete linear spaces: the infinite-dimensional Banach spaces* $(\mathcal{C}(\mathcal{R}^\star), \|\cdot\|_\infty)$.

BOX 2.2 (The Ray-casting Function γ)

The *ray-casting function* γ defined by

$$\gamma : \partial\mathcal{V} \times S^2 \longrightarrow \partial\mathcal{V} \tag{2.54}$$

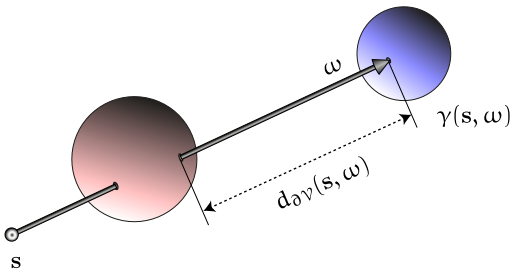
with

$$\gamma(\mathbf{s}, \omega) = \mathbf{s} + d_{\partial\mathcal{V}}(\mathbf{s}, \omega)\,\omega, \tag{2.55}$$

where it holds:

$$d_{\partial\mathcal{V}}(\mathbf{s}, \omega) \stackrel{\text{def}}{=} \inf_{\alpha>0}\{\mathbf{s} + \alpha\,\omega \ \in \partial\mathcal{V}\}, \tag{2.56}$$

returns the nearest point of intersection of a ray starting at $\mathbf{s}$ with an object in the scene in direction ω. It is clear that both functions γ and $d_{\partial\mathcal{V}}$, the *boundary distance function*, are defined on the ray space $\mathcal{R}^{\partial\mathcal{V}}$ but only $d_{\partial\mathcal{V}}$ belongs to $\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}})$.



REMARK 2.12 *Applied to our present still rudimentary knowledge about particle transport phenomena, the above construction of the function spaces* $\mathcal{L}(\mathcal{R}^\star)$ *means that we can use elements from these function spaces to describe processes such as emission, absorption, and scattering of light particles. Those functions then provide information about the number of emitted, scattered, or absorbed particles at all points and directions of the ray spaces.* Chapter 5

Note that, due to the vector space requirements, the range of a function from $\mathcal{L}(\mathcal{R}^\star)$ *has to be expanded to include the negative domain of the real numbers.*

REMARK 2.13 *The ray-casting function is a classical problem of computer graphics. Due to its influence on the rendering time and effort of the procedure it can be regarded as the heart of any rendering method based on the principle of ray tracing. For this reason, as well as in particular the fact, that in ray tracing algorithms routines implementing the ray-casting function must be executed over and over again,*

a large number of techniques have been developed for optimizing the process of ray intersection with scene objects.

REMARK 2.14 *When deriving an operator model for light transport in participating media, we need a function, that returns the closest hit point of a ray, starting at an inner point of a medium, with an object in the scene. For that purpose we have to extend the domain of the ray-casting function* γ *from* $\partial\mathcal{V} \times S^2$ *to* $\mathcal{V} \times S^2$.

BOX 2.3 (The Reversible Ray Space $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$)

In Chapter 5 we discuss mathematical models of light and importance transport. Since, as we will show in Section 5.2, importance flows in the opposite direction to light, we need a function space which allows to describe importance functions. This function space is based on the so-called *reversible ray space*, $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$, for a more detailed description see [221, Veach 1998].

DEFINITION 2.11 (The Reversible Ray Space, $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$) *The reversible ray space,* $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$*, is the space of all reversible rays form* $\mathcal{R}^{\partial\mathcal{V}}$*. It is defined via*

$$\widetilde{\mathcal{R}}^{\partial\mathcal{V}} \stackrel{\text{def}}{=} \{\mathbf{r} \in \mathcal{R}^{\partial\mathcal{V}} \,|\, d_{\partial\mathcal{V}}(\mathbf{r}) < \infty\}, \tag{2.57}$$

where $d_{\partial\mathcal{V}}(\mathbf{r})$ *is the boundary distance function from Box 2.2. Due to this definition, only rays that end at object surfaces are reversible.*

INCIDENT AND EXITANT FUNCTIONS ON $\mathcal{L}(\mathcal{R}^\star)$. Until now, we specified directions as vectors pointing away from some point. But this does not specify necessarily the flow of a quantity, such as light, at which we are interested in. For that purpose, it is useful to have a notation that makes it possible to describe light incident or exitant at surface points. This can be done by classifying functions from $\mathcal{L}(\mathcal{R}^\star)$ into *incident* and *exitant* functions, denoted as $\mathcal{L}_i(\mathcal{R}^\star)$ as well as $\mathcal{L}_o(\mathcal{R}^\star)$.

Chapter 5

As is illustrated in Figure 2.7, a function $f_i(\mathbf{s}, \omega_i) \in \mathcal{L}_i(\mathcal{R}^{\partial\mathcal{V}})$ provides information on the amount of a quantity incident from direction ω_i at surface point $\mathbf{s}$, while $f_o(\mathbf{s}, \omega_o) \in \mathcal{L}_o(\mathcal{R}^{\partial\mathcal{V}})$ describes the amount exitant from $\mathbf{s}$ in the direction ω_o. It is easy to prove that both $\mathcal{L}_i(\mathcal{R}^\star)$ and $\mathcal{L}_o(\mathcal{R}^\star)$ define subspaces of the associated linear spaces $\mathcal{L}(\mathcal{R}^\star)$ and $\mathcal{L}(\mathcal{R}^\star)$, which, equipped with the supremum norm $\|\cdot\|_\infty$, then become Banach spaces.

Section 5.1.1.2

Section 5.1.1.2

Supremum Norm (33)

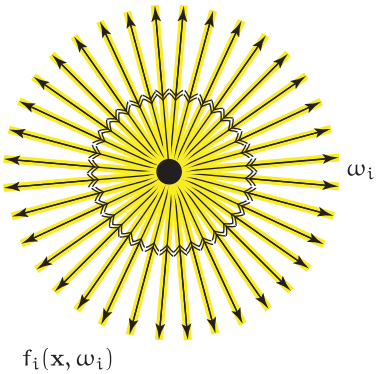
Banach Space (35)

q (282)

σ_a (4.9)

Very nice examples based on the concept of an exitant as well as an incident function are the *volumetric emission function*, q, and the absorption *function*, σ_a. In Chapter 4, when deriving the stationary particle transport equation, we use this functions to describe the emission behavior of light at a surface point as well as the absorption behavior of light particles in participating media.

EXAMPLE 2.13 (Volumetric Emission and Absorption Function) *While the* volumetric emis-



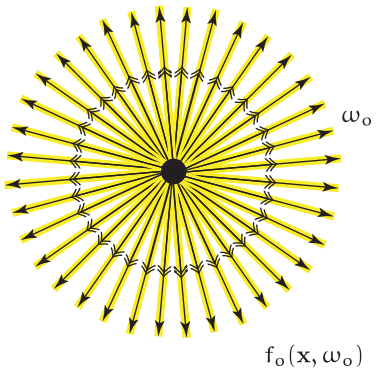


FIGURE 2.7: INCIDENT AND EXITANT FUNCTIONS. For incident and exitant functions, directions are always specified to point away from some point $\mathbf{x}$. The flow of a quantity, specified by the function, indicates the type of the function. An incident function $f_i(\mathbf{x}, \omega_i)$ transports a quantity in direction ω_i towards point $\mathbf{x}$, while an exitant function $f_o(\mathbf{x}, \omega_o)$ transports the quantity from point $\mathbf{x}$ away in direction ω_o.

sion function q *given by:* q (282)

$$q : \mathcal{V}^o \times S^2 \to \mathbb{R} \tag{2.58}$$

with

$$(\mathbf{x}, \omega_o) \to q(\mathbf{x}, \omega_o) \tag{2.59}$$

is defined on the ray space $\mathcal{R}^{\mathcal{V}^o}$ *and returns the number of particles created at a volumetric point* $\mathbf{x}$ *in direction* ω_o*, i.e.* $q \in \mathcal{L}_o(\mathcal{R}^{\mathcal{V}^o})$*, the* absorption function σ_a σ_a (282) *defined by*

$$\sigma_a : \mathcal{V}^o \times S^2 \to \mathbb{R} \tag{2.60}$$

with

$$(\mathbf{x}, \omega_i) \to \sigma_a(\mathbf{x}, \omega_i) \tag{2.61}$$

is an element of $\mathcal{L}_i(\mathcal{R}^{\mathcal{V}^o})$*. It provides the fraction of particles incident at point* $\mathbf{x}$ *from direction* ω_i*, which will be absorbed, see Figure 2.8.*

For rendering images light sources are necessary that illuminate the underlying scene model. Now in real world, light sources exist in a variety of shapes, colors, and sizes. So, we will present and discuss the most important types of light sources and their properties for rendering in Section 4.3. As all light sources can be interpreted as emitters of light

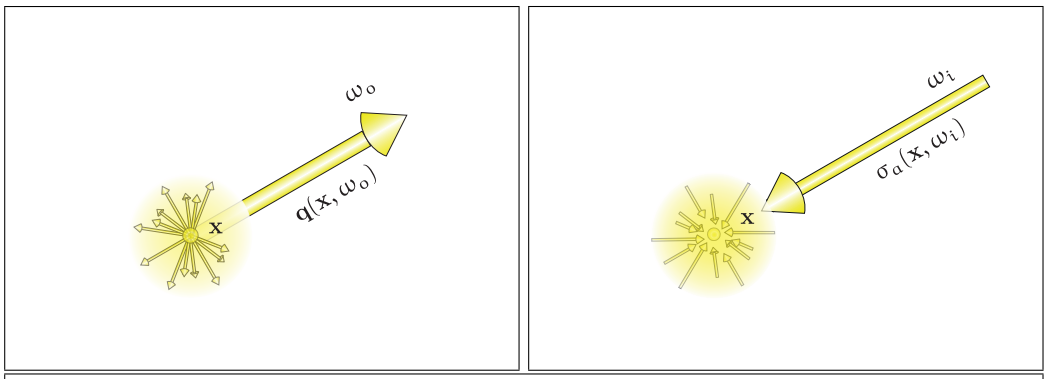


FIGURE 2.8: EMISSION AND ABSORPTION FUNCTION. Emission of particles at point $\mathbf{x}$ described via an emission function q, and the absorption of particles at $\mathbf{x}$ described via the absorption function σ_a.

particles, it lies on the hand to define a light source via the mathematical concept of the exitant function from a corresponding function space. For that purpose, we now define the two most important concepts of light sources in computer graphics: *point light sources* and *area light sources*. Later, in Section 4.3, we will once again take with respect to these types of light sources, talk about their properties and we will introduce additionally a few other types of light sources often used in rendering.

DEFINITION 2.12 (Ideal Point Light Source) *An* ideal point light source *can be considered as a point* $\mathbf{x} \in \mathbb{R}^3$*, that has neither a size nor a shape. The point* $\mathbf{x}$ *then serves as the center of a spherical field of light, where light is uniformly radiated in all directions, see the left image of Figure 2.9.*

Due to this definition, an ideal point light source within a scene can then be described via an exitant function f_o, given by

$$f_o : \mathbb{R}^3 \times S^2 \to \mathbb{R} \tag{2.62}$$

with

$$(\mathbf{x}, \omega_o) \mapsto \begin{cases} C & \mathbf{x} \in \mathbb{R}^3, \omega_o \in S^2(\mathbf{x}) \\ 0 & \text{otherwise,} \end{cases} \tag{2.63}$$

$\mathcal{V}$ (41) where the point $\mathbf{x}$ is located within a subvolume $\mathcal{V}$ of $\mathbb{R}^3$, that represents a vacuum or any participating medium, and ω_o is any direction of the upper hemisphere about $\mathbf{x}$.

Now, except of the sun, which, under certain circumstances can be considered as a point light source, ideal point light emitters does not exist in real world. Light sources in real world have a finite size, some amount of surface area, and takes up some finite amount

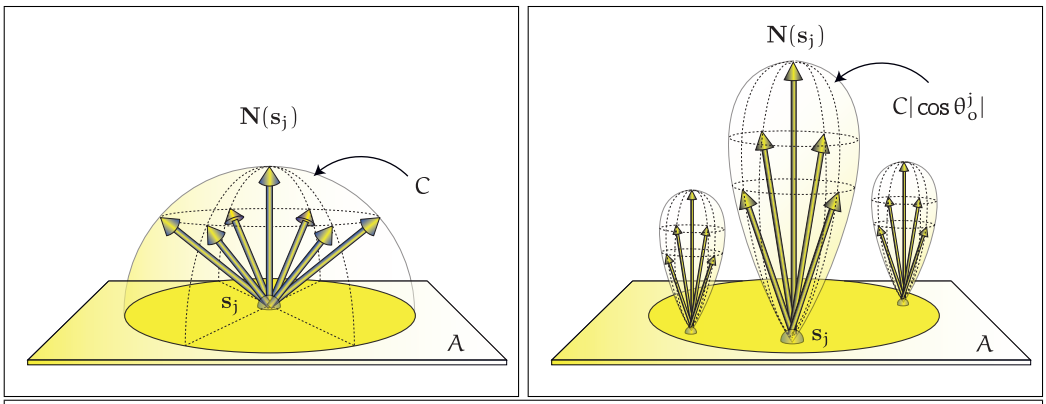


FIGURE 2.9: POINT AND AREA LIGHT SOURCES. Left, a point light source located at a surface patch emitting a constant light amount from point $\mathbf{s}_j \in \partial\mathcal{V}$ in all directions over the upper hemisphere $\mathcal{H}^2_+(\mathbf{s}_j)$. Right, the visualization of an area light source that emits light proportional to a cosine lobe around the surface normal at $\mathbf{s}_j$.

of space. Those light sources can rather be considered as an infinitely large set of points, that emit their light in all directions uniformly or non uniformly.

DEFINITION 2.13 (Area Light Source) *An* area light source *is any finite 2-dimensional surface* ∂V *of* $\mathbb{R}^3$. *Every point of* $\mathbf{s} \in \partial V$ *then serves as an ideal point light source that emits light in all directions of the upper hemisphere* $\mathcal{H}^2_+$ *about* $\mathbf{s}$ *uniformly or non uniformly, see the right image of Figure 2.9.* $\mathcal{V}$ (41) $\mathcal{H}^2_+$ (849)

Due to this definition, an area light source within a scene can then be described via an exitant function $f_o^{\partial\mathcal{V}}$, given by:

$$f_o^{\partial\mathcal{V}} : \partial\mathcal{V} \times \mathcal{H}^2_+ \to \mathbb{R} \tag{2.64}$$

with

$$(\mathbf{s}, \omega_o) \mapsto \begin{cases} f_o^{\partial\mathcal{V}}(\mathbf{s}) & \mathbf{s} \in \partial\mathcal{V}, \omega_o \in \mathcal{H}^2_+(\mathbf{s}) \\ 0 & \text{otherwise,} \end{cases} \tag{2.65}$$

where the point $\mathbf{s}$ is located on $\partial\mathcal{V} \subset \mathbb{R}^3$ that represents the shape of an area light source and ω_o is any direction of the upper hemisphere about $\mathbf{s}$.

EXAMPLE 2.14 (Point and Area Light Sources) *Let* $f_o^{\mathcal{V}^o}$ *be a point light source located at point* $\mathbf{x} \in \mathcal{V}^o$ *within a medium and let* $f_o^{\partial\mathcal{V}}$ *be an area light source represented by a surface* $\partial\mathcal{V}_i$ *of* $\partial\mathcal{V}$. *While the first emits a constant light amount in each direction, the latter emits its constant light energy contribution depending on the emission direction* ω *with respect to the normal at a point* $\mathbf{s} \in \partial\mathcal{V}_i$. *Based on our concept of* $\mathcal{V}^o$ (41)

exitant functions, light sources may be mathematically described by elements of the function spaces $\mathcal{L}_o(\mathcal{R}^{\partial\mathcal{V}})$ *and* $\mathcal{L}_o(\mathcal{R}^{\mathcal{V}^o})$. *While the involved point light source may be simulated via an exitant function* $f_o^{\mathcal{V}^o}$ *defined over* $\mathcal{R}^{\mathcal{V}^o}$ *with* $f_o^{\mathcal{V}^o}(\mathbf{x}, \omega_o) = C$ *for all* $\omega_o \in S^2$, *the area light source may be simulated via an exitant function* $f_o^{\partial\mathcal{V}}$ *defined over* $\mathcal{R}^{\partial\mathcal{V}}$, *with* $f_o^{\partial\mathcal{V}}(\mathbf{s}, \omega_o) = C\langle\omega_o, \mathbf{N}(\mathbf{s})\rangle$, $C \in \mathbb{R}^{\geq 0}$, *that is* $f_o^{\partial\mathcal{V}} \neq 0$ *only for the element* $\partial\mathcal{V}_i$ *of* $\partial\mathcal{V}$.

Obviously, the two light sources are described by two functions $f_o^{\mathcal{V}^o}$ *and* $f_o^{\partial\mathcal{V}}$ *of the function spaces* $(\mathcal{L}_o(\mathcal{R}^{\mathcal{V}^o}), \|\cdot\|_\infty)$ *and* $(\mathcal{L}_o(\mathcal{R}^{\partial\mathcal{V}}, \|\cdot\|_\infty)$. *Due to their definitions over the above ray spaces these functions may be assumed to be bounded as for physical reasons both* $\|f_o^{\mathcal{V}^o}\|_\infty = c < \infty$ *and* $\|f_o^{\partial\mathcal{V}}\|_\infty = c\langle\omega_o, \mathbf{N}(\mathbf{s})\rangle < \infty$ *holds.*

EXAMPLE 2.15 (The Incident and Exitant Hemispheres $\mathcal{H}_i^2$ and $\mathcal{H}_o^2$) *For defining the concepts of the BRDF and the BTDF it is convenient to introduce the constructs of the incident and exitant hemispheres,* $\mathcal{H}_i^2$ *and* $\mathcal{H}_o^2$. *Both constructs are based on the same set of directions, that is,* $\mathcal{H}_i^2 = \mathcal{H}_o^2$, *and can be used to represent the upper as well as the lower hemisphere.*

Section 4.2.2.2 $\mathcal{H}_i^2$ and $\mathcal{H}_o^2$ (849)

Chapter 9 Global Illumination problem (6)

As show in the following, ray tracing procedures based on stochastic principles represent approximate solutions to the global illumination problem. When a ray, generated at an arbitrary point of a sensor—be this the surface of a primitive or real camera or the retina of the human eye—enters a given illuminated scene, the algorithm calculates approximates of the portion of light incident at a finite number of points of the ray spaces $\mathcal{R}^{\partial\mathcal{V}}$ or $\mathcal{R}^{\mathcal{V}^o}$. Generating a $s_x \times s_y$-regular grid onto a sensor, the amount of energy measured in a visibility test then roughly corresponds to the light incident from the environment corresponding to the scene projection onto this pixel array. The sensor then reflects the light distribution in the scene via the vector $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_s)$, $s = s_x \cdot s_y$ of the complete linear normed space $(\mathbb{R}^s, \|\cdot\|_2)$.

Section 4.6

2.1.4 LINEAR OPERATORS AND THEIR ADJOINTS

In order to be able to follow the path suggested by functional analysis for solving a complex mathematical problem it is a good idea to reformulate it as an operator equation in an abstract linear space. In particular, this holds for problems that underlie linear integral equations. In this manner, the problem can then be approached using solution methods developed within functional analysis. As a basis for that serves the functional analytic concept of the linear operator and its adjoint as mappings between abstract linear spaces. These concepts enable us to represent the global illumination problem as simple linear equations based on the well-known concepts of *light* and *importance*. Additionally, operators of this kind play the central role in the theory of finite-element methods, which, for example, is the mathematical foundation of all radiosity algorithms.

Linear Operator Equation (61) Linear Integral Equation (127) Linear Operator (53) Linear Space (854) Chapter 5 Chapter 10

LINEAR OPERATORS. For our further discussions let $\mathcal{S}$ and $\mathcal{T}$ be two linear spaces.

Linear Space (854)

DEFINITION 2.14 (Linear Operator) *A mapping* $\mathbf{T}$ *defined on the domain* $\mathrm{Dom}(\mathbf{T})$ *over the linear space* $\mathcal{S}$ *taking values within* $\mathrm{Ran}(\mathbf{T}) \subseteq \mathcal{T}$*, thus,*

Dom(·) (835)
Ran(·) (835)

$$\mathbf{T} : \mathrm{Dom}(\mathbf{T}) \subseteq \mathcal{S} \quad \longrightarrow \quad \mathrm{Ran}(\mathbf{T}) \subseteq \mathcal{T}, \tag{2.66}$$

is denoted as a linear operator *if it satisfies the* linearity property*:*

$$\mathbf{T}(\alpha x + \beta y) \quad = \quad \alpha(\mathbf{T}x) + \beta(\mathbf{T}y), \tag{2.67}$$

where $\alpha, \beta \in \mathbb{R}$ *and* $x, y \in \mathcal{S}$*. The linear operator* $\mathbf{T}$ *is termed as* degenerated *if* $\dim \mathrm{Im}(\mathbf{T}) < \infty$ *applies.*

EXAMPLE 2.16 (Matrices as Linear Operators Between Finite Dimensional Linear Spaces) *The well-known concept of a* matrix *from linear algebra is a simple example of a linear operator between two finite-dimensional linear spaces.*

Matrix (853)
Linear Space (854)

Let us assume, two finite-dimensional linear spaces V *and* W *are given, with* $\dim V = m$ *and* $\dim W = n$*. Then, we can map an element* $\mathbf{x} \in V$ *to an element* $\mathbf{y} \in W$ *by means of the matrix-vector product, thus,*

$$\mathbf{A} : V \to W \tag{2.68}$$

with

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{2.69}$$

where the components y_i *of* $\mathbf{y}$ *are given by:* $y_i = \sum_{j=1}^{m} a_{ij} x_j$ *for* $1 \leq i \leq n$*. It can easily be shown that* $\mathbf{T}$ *satisfies the linearity property of a linear operator from Equation (2.67).*

EXAMPLE 2.17 (The Differential Operator $\frac{d}{dx}$) *A common known example of a linear operator results from the process of differentiation. Let us recall, the derivative of a polynomial* p_n *of degree* n *is a polynomial of degree* $n-1$*. This implies the construction of a linear operator between the spaces* $\mathcal{P}_n$ *and* $\mathcal{P}_{n-1}$ *by:*

Section A.4
$\mathcal{P}_n$ (855)

$$\left(\frac{d}{dx} p_n\right)(x) = \frac{d}{dx} \sum_{i=0}^{n} \alpha_i x^i = \sum_{i=0}^{n} \alpha_i \frac{d}{dx} x^i = \sum_{i=0}^{n-1} (i+1)\, \alpha_{i+1} x^i. \tag{2.70}$$

As the space of polynomials of degree $n-1$ *is finite-dimensional,* $\frac{d}{dx}$ *is a degenerated linear operator.*

EXAMPLE 2.18 (The Gradient Operator ∇ in $\mathbb{R}^n$) *The* Gradient operator ∇ *is most often applied to a real-valued function* $f(x_1, \ldots, x_n)$*, differentiable at each point* $\mathbf{x}_0 = (x_{0_1}, \ldots, x_{0_n})$*. It is defined by*

$$(\nabla f)(x_1, \ldots, x_n) \stackrel{def}{=} \frac{\partial f}{\partial x_1} \mathbf{e}_1 + \ldots + \frac{\partial f}{\partial x_n} \mathbf{e}_n. \tag{2.71}$$

Obviously, the image of the gradient operator, also briefly denoted as the gradient*, is a vector of the Euclidean space* $\mathbb{R}^n$ *whose components are the partial derivatives of* f*. In Section 4.1.1 we will encounter the gradient operator* ∇ *in combination with the* Gauss Divergence Theorem *that enable us to transform a surface integral into a volume integral.*

Partial Derivative (871)

Gauss Divergence Theorem (283)

EXAMPLE 2.19 (Multiplication and Evaluation Operators) *Two other interesting examples of linear operators, now defined between two infinite-dimensional linear function spaces, are the* multiplication operator *and the* evaluation operator*.*

Let $\mathcal{S}$ *be a function space and* f, g *be two functions from* $\mathcal{S}$*, let furthermore* x *be a point within the domain of* f*. Due to [8, Arvo 1993] the multiplication operator* $\mathbf{M}_g$ *is defined as:*

$$(\mathbf{M}_g f)(x) \stackrel{def}{=} g(x)\, f(x) \tag{2.72}$$

and for the evaluation operator it holds:

$$(\mathbf{E} f)(x) \stackrel{def}{=} f(x). \tag{2.73}$$

Both operators play a central role in our further considerations when deriving a mathematical model of light and importance transport. There, we will need the multiplication operator to express the attenuation of light propagating in participating media.

Chapter 5

EXAMPLE 2.20 (Spectral Power Distributions) *Let us consider the inner product spaces* $B(\Lambda)$ *and* $\mathbb{R}^n$ *with* $\dim(\mathbb{R}^n) = n$*. As we have seen in Example 2.11, every function* $S \in B(\Lambda)$ *can be approximated by a* n*-dimensional vector* $(\langle S(\lambda), \phi_1(\lambda)\rangle, \ldots, \langle S(\lambda), \phi_n(\lambda)\rangle)^T$*. With respect to our definition of a linear operator the Best Approximation Theorem supplies us with a degenerated linear operator* $\mathbf{T}$ *from* $B(\Lambda)$ *to* $\mathbb{R}^n$*, namely:*

$\mathcal{B}(\Lambda)$ (40)

Best-approximation Theorem (38)

Degenerated Linear Operator (53)

$$\mathbf{T} : B(\Lambda) \to span(\phi_1, \ldots, \phi_n) \leq \mathbb{R}^n$$

with

$$\mathbf{T}S = (\langle S, \phi_1\rangle, \ldots, \langle S, \phi_n\rangle)^T \tag{2.74}$$

and

$$\begin{aligned}
\mathbf{T}(\alpha S + \beta S') &= (\langle \alpha S + \beta S', \phi_1\rangle, \ldots, \langle \alpha S + \beta S', \phi_n\rangle) && (2.75)\\
&= (\alpha\langle S, \phi_1\rangle + \beta\langle S', \phi_1\rangle, \ldots, \alpha\langle S, \phi_n\rangle + \beta\langle S', \phi_n\rangle) && (2.76)\\
&= (\alpha\langle S, \phi_1\rangle, \ldots, \alpha\langle S, \phi_n\rangle) + (\beta\langle S', \phi_1\rangle, \ldots, \beta\langle S', \phi_n\rangle) && (2.77)\\
&= \alpha(\mathbf{T}S) + \beta(\mathbf{T}S'), && (2.78)
\end{aligned}$$

where we have used the linearity property of the inner product from Definition A.17 in the second step.

REMARK 2.15 *Due to Definition A.2, the condition that $\mathcal{T}$ is a linear space in the above definition, is not required. For the definition of a linear operator it suffices if the domain of the operator $\mathbf{T}$ is a linear space, the image of $\mathbf{T}$ can also be any arbitrary set $\mathcal{T}$.*

DEFINITION 2.15 (Linear Functional) *Let $\mathbf{T}$ be a linear operator from linear space $\mathcal{T}$ into the Euclidean space $\mathbb{R}$, then $\mathbf{T}$ is called a* linear functional.

REMARK 2.16 *Due to the* Riesz Representation Theorem *the inner product $\langle\cdot,\cdot\rangle$ of an inner product space $\mathcal{T}$ can be interpreted as the action of a functional by defining the functional l for a given $u \in \mathcal{T}$ as:*

$$lv \stackrel{def}{=} \langle u, v\rangle \tag{2.79}$$

for any $v \in \mathcal{T}$, [169, Reddy 1998].

REMARK 2.17 *As we will see in our following discussions, the mathematical concept of the linear functional is fundamental for the field of realistic rendering when measuring the amount of light incident onto a measure device, such as a pixel. Another reason for the importance of this concept is exemplified in the Dirac δ-construct, which plays an important role in many branches of physics and engineering as well as in describing scattering models in global illumination theroy.*

Section 4.6
Dirac δ-Distribution (118)
Section 4.2.2.2

Now, suppose $\mathbf{T}$ is a mapping between the linear normed space $(\mathcal{S}, \|\cdot\|_{\mathcal{S}})$ while the linear space $\mathcal{T}$ is equipped with a norm $\|\cdot\|_{\mathcal{T}}$. Then, $\mathbf{T}$ is called a *bounded linear operator*, if a real number $c > 0$ exists, so that the following applies for all $x \in \mathcal{S}$

Linear Normed Space (860)
Norm (860)

$$\|\mathbf{T}x\|_{\mathcal{T}} \leq c\,\|x\|_{\mathcal{S}}. \tag{2.80}$$

EXAMPLE 2.21 (Converting Spectral Values into RGB-Color Values for Output Devices) *For illustrations on output devices the spectral value S, in Example 2.11 calculated via a rendering procedure and assigned to a pixel, must be converted into a corresponding RGB-color value. This means that in a first step, using the three color-matching functions $\bar{x}, \bar{y}$ and $\bar{z}$, the tristimulus values X, Y and Z must be determined via the relations*

$$X = \int_\Lambda \bar{x}(\lambda)S(\lambda)\,d\lambda \stackrel{(2.36)}{=} \sum_{i=1}^{n}\left(\int_\Lambda \bar{x}(\lambda)\phi_i(\lambda)\,d\lambda\right)\langle S(\lambda), \phi_i(\lambda)\rangle \tag{2.81}$$

$$Y = \int_\Lambda \bar{y}(\lambda)S(\lambda)\,d\lambda \stackrel{(2.36)}{=} \sum_{i=1}^{n}\left(\int_\Lambda \bar{y}(\lambda)\phi_i(\lambda)\,d\lambda\right)\langle S(\lambda), \phi_i(\lambda)\rangle \tag{2.82}$$

$$Z = \int_\Lambda \bar{z}(\lambda)S(\lambda)\,d\lambda \stackrel{(2.36)}{=} \sum_{i=1}^{n}\left(\int_\Lambda \bar{z}(\lambda)\phi_i(\lambda)\,d\lambda\right)\langle S(\lambda), \phi_i(\lambda)\rangle, \tag{2.83}$$

which can be formulated in matrix-vector notation as

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \mathbf{M} \cdot \begin{pmatrix} \langle S(\lambda), \phi_1(\lambda) \rangle \\ \langle S(\lambda), \phi_2(\lambda) \rangle \\ \cdots \\ \langle S(\lambda), \phi_n(\lambda) \rangle \end{pmatrix} \overset{(2.74)}{=} \mathbf{M} \cdot \mathbf{T}S(\lambda) \tag{2.84}$$

with

$$\mathbf{M} = \begin{pmatrix} \langle \bar{x}(\lambda), \phi_1(\lambda) \rangle & \cdots & \langle \bar{x}(\lambda), \phi_n(\lambda) \rangle \\ \langle \bar{y}(\lambda), \phi_1(\lambda) \rangle & \cdots & \langle \bar{y}(\lambda), \phi_n(\lambda) \rangle \\ \langle \bar{z}(\lambda), \phi_1(\lambda) \rangle & \cdots & \langle \bar{z}(\lambda), \phi_n(\lambda) \rangle \end{pmatrix}. \tag{2.85}$$

Inner Product (859) B(Λ) (40) *According to the definition of the inner product* $\langle \cdot, \cdot \rangle$ *of the function space* $B(\Lambda)$*, the matrix* $\mathbf{M}$*, originating from Equations (2.81) - (2.83), then serves, together with a* (3×3)*-matrix* $\widetilde{\mathbf{M}}$*, describing the chromaticities of the output device, as a bounded linear operator defined over the* n*-dimensional function space generated over the* Orthonormal Set (861) *orthonormal system* $\{\phi_1, \ldots, \phi_n\}$ *with values from* $[0,1]^3 \subset [-1,1]^3$*. This implies that the RGB-color value assigned to the pixel is given by*

$$\begin{pmatrix} R_\square \\ G_\square \\ B_\square \end{pmatrix} = \widetilde{\mathbf{M}}\mathbf{M} \cdot \mathbf{T}S(\lambda) \overset{(2.74)}{=} \widetilde{\mathbf{M}}\mathbf{M} \cdot \begin{pmatrix} \langle S(\lambda), \phi_1(\lambda) \rangle \\ \langle S(\lambda), \phi_2(\lambda) \rangle \\ \cdots \\ \langle S(\lambda), \phi_n(\lambda) \rangle \end{pmatrix}. \tag{2.86}$$

Obviously, the set of all bounded linear operators, $\mathbf{L}(\mathcal{S}, \mathcal{T})$ satisfies the conditions required to a linear space. As is shown in functional analysis, it becomes, together with the *operator norm*

$$\|\mathbf{T}\|_{\mathbf{L}} \overset{def}{=} \sup_{x \in \mathcal{S}, x \neq 0} \frac{\|\mathbf{T}x\|_{\mathcal{T}}}{\|x\|_{\mathcal{S}}}, \tag{2.87}$$

a linear normed space: the space $(\mathbf{L}(\mathcal{S}, \mathcal{T}), \|\cdot\|_{\mathbf{L}})$. If we identify the space $\mathcal{T}$ with the real number field $\mathbb{R}$ then $\mathbf{L}(\mathcal{S}, \mathbb{R})$ describes the space of all *bounded linear functionals*. Equipped with the norm

$$\|\mathbf{t}\|_{\mathbf{l}} \overset{def}{=} \sup_{x \in \mathcal{S}, x \neq 0} \frac{|\mathbf{t}x|}{\|x\|_{\mathcal{S}}}, \tag{2.88}$$

it becomes a linear normed space: the *dual space* $(\mathbf{l}(\mathcal{S}, \mathbb{R}), \|\cdot\|_{\mathbf{l}})$.

Let us now consider an interesting example of a bounded linear space, which will serve as a motivation for the formulation of light transport in free space to be sought below in terms of an operator equation defined over the linear normed spaces $(\mathcal{L}_i(\mathcal{R}^{\partial\mathcal{V}}), \|\cdot\|_\infty)$, $(\mathcal{L}_o(\mathcal{R}^{\partial\mathcal{V}}), \|\cdot\|_\infty)$ and $(\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}}), \|\cdot\|_\infty)$ respectivly. $(\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}}), \|\cdot\|_\infty)$ (46)

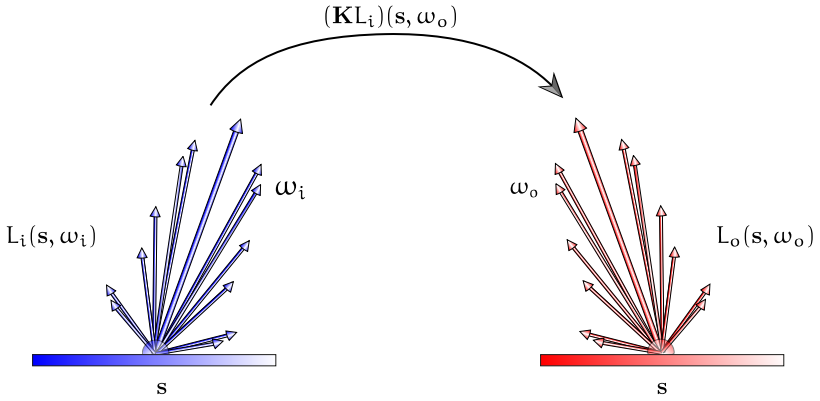


FIGURE 2.10: REFLECTION OPERATOR. The reflection operator $\mathbf{K}$ maps the incident function $\mathrm{L_i}$ onto the exitant function $\mathrm{L_o}$.

EXAMPLE 2.22 *Let us assume the two function spaces* $(\mathcal{L}_\mathrm{i}(\mathcal{R}^{\partial\mathcal{V}}), \|\cdot\|_\infty)$ *and* $(\mathcal{L}_\mathrm{o}(\mathcal{R}^{\partial\mathcal{V}}, \|\cdot\|_\infty)$ *from the last section are given. Then, the linear operator* $\mathbf{K}$*, given by,*

$$\mathbf{K} : \mathcal{L}_\mathrm{i}(\mathcal{R}^{\partial\mathcal{V}}) \quad \longrightarrow \quad \mathcal{L}_\mathrm{o}(\mathcal{R}^{\partial\mathcal{V}}), \tag{2.89}$$

with

$$\mathrm{L_i}(\mathbf{x},\omega_\mathrm{i}) \quad \longmapsto \quad \mathrm{L_o}(\mathbf{x},\omega_\mathrm{o}) = (\mathbf{K}\mathrm{L_i})(\mathbf{x},\omega_\mathrm{o}), \tag{2.90}$$

obviously maps the incident function $\mathrm{L_i}$ *onto an exitant function* $\mathrm{L_o}$*, see Figure 2.10. Using a bounded operator* $\mathbf{K}$*, then it becomes possible to mathematically formulate a physical process describing a finite quantity that enters from a direction* ω_i *incident at a surface point* $\mathbf{x}$ *and exits in a direction* ω_o*. In such a case the operator* $\mathbf{K}$ *could be interpreted as a mathematical formulation of the physical phenomenon of light reflection, or refraction at a surface. Defined between the ray spaces* $\mathcal{L}_\mathrm{i}(\mathcal{R}^{\mathcal{V}^\circ})$ *and* $\mathcal{L}_\mathrm{o}(\mathcal{R}^{\mathcal{V}^\circ})$*,* $\mathbf{K}$ *can be considered as the scattering of light at a small particle in participating media.*

Incident & Exitant Function (48)

Section 5.1.1.1

REMARK 2.18 *The concept of the operator norm, as defined in Equation (2.87), is of great importance when constructing mathematical models for light and importance transport in Chapter 5. As above illustrated at the example of reflection, we will describe the whole light and importance transport in a scene, thus, all possible effects at object surfaces and within participating media, by a single linear operator, the so-called* light transport operator *respectively the* importance transport operator*. Both*

operators will be defined over function spaces based on the ray spaces from Section 2.1.3, and are used to construct operator norms that help to estimate approximate solutions of the global illumination problem.

Section 5.1.2.2

Section 5.1.1.2

An other important property of a linear operator is *compactness*. So, we denote a linear operator $\mathbf{T}$ to be *compact* if the image of a bounded set $B \in \mathcal{S}$ under the mapping $\mathbf{T}$ is compact[1] in $\mathcal{T}$. An important feature of compact operators is their ability to map bounded sets onto sets with additional characteristics—a fact, which will be of great use to us in a number of situations. Another important feature of compact operators is that they may be approximated via degenerated operators. This means that for every $\epsilon > 0$ there exists a sequence of finite-dimensional operators $(\mathbf{T}_n)_{n\in\mathbb{N}}$ that converges towards $\mathbf{T}$, i.e.: $\|\mathbf{T}_n - \mathbf{T}\|_{\mathcal{T}} < \epsilon$.

Bounded Set (862)

Section 5.1.1.2

Section 5.1.2.2

Let us now assume that the spaces we are interested in are Hilbert spaces. The features of these spaces, more specifically, the orthogonality property of the spaces, then allows to construct so-called *projection operators*. These are operators that map elements of an infinite-dimensional Hilbert space $\mathcal{S}$ to elements of finite-dimensional subspaces $\mathcal{U}$ of $\mathcal{S}$.

Hilbert Space (36)

Section 2.3.3

DEFINITION 2.16 (Linear Projection Operator) *Let $\mathbf{T}$ be a linear operator from $\mathcal{S}$ into $\mathcal{S}$, then $\mathbf{T}$ is denoted as a* linear projection operator *if $\forall x \in \mathcal{S}$ the following applies*

$$\mathbf{T}^2 = \mathbf{T}. \tag{2.91}$$

Obviously, a linear projection operator leaves its image unchanged, that is, it formalizes and generalizes the idea of a graphical projection.

In accordance with the Best Approximation Theorem and the definition of inner product spaces, the linear operator $\mathbf{T}$ is termed an *orthogonal projection operator* if it permits the representation of Theorem 2.1 as a finite Fourier series with respect to the orthonormal basis $\{\phi_1, \phi_2, \ldots, \phi_n\}$. Conversely, on the basis of the above made statements, the elements of an infinite-dimensional space may also be approximated as images of compact linear operators or via a sequence of projection operators, as examples of degenerated operators.

Best Approximation Theorem (38)

Inner Product Space (859)

FourierSeries (39)

Orthonormal Basis (37)

EXAMPLE 2.23 (Orthogonal Projection) *Let $\mathcal{P}_n$ be the space of polynomials of degree n defined over a closed set $S \subset \mathbb{R}$. Then, a polynomial $p_n(x) = \sum_{i=0}^{n} \alpha_i x^i$ can be represented via the vector $(\alpha_0, \ldots, \alpha_n)^T \in \mathbb{R}^{n+1}$. The Hilbert space structure of the $n+1$-dimensional Euclidean space is clearly transmitted over the isomorphism $\mathbf{T}$ to $\mathcal{P}_n$, where $\mathbf{T}(\mathbf{e}_i) = x^{i-1}, 1 \leq i \leq n+1$ and $(\mathbf{e}_1, \ldots, \mathbf{e}_{n+1})$ forms an orthonormal basis*

Hilbert Space (36)

Orthonormal Basis (37)

[1] A set B of a linear normed space $(\mathcal{S}, \|\cdot\|)$ is denoted as *compact* if for every sequence $(x_n)_{n\in\mathbb{N}}$ there exist a convergent subsequence $(x'_n)_{n\in\mathbb{N}}$ that converges to a limit from $\mathcal{S}$.

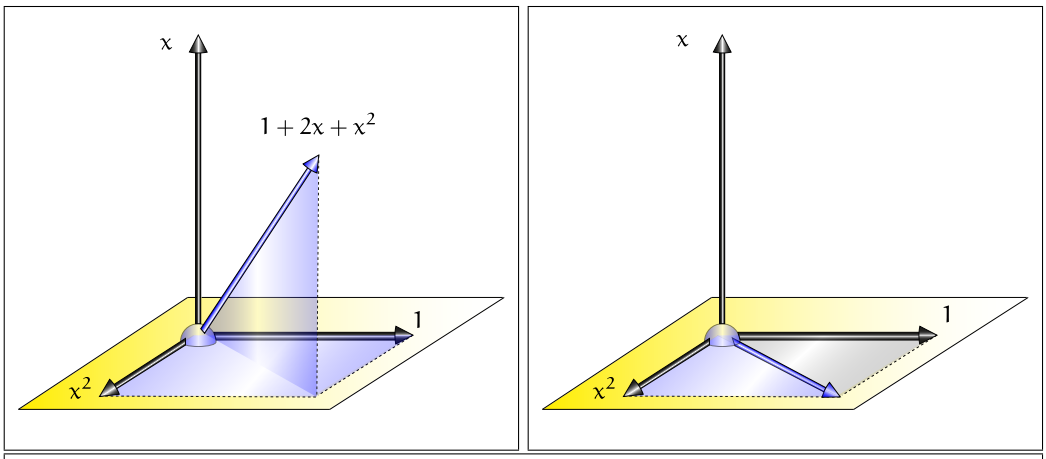


FIGURE 2.11: ORTHOGONAL PROJECTION. The orthogonal projection of the polynomial $1+2x+x^2$ onto the polynomial $1+x^2$ using the orthogonal projection operator $\mathbf{T}$ defined in Example 2.23.

of $\mathbb{R}^{n+1}$. *If we now consider an operator* $\mathbf{Pr}$ *given by:*

$$\mathbf{Pr} : \mathcal{P}_n \longrightarrow \mathcal{P}_{even} \tag{2.92}$$

with

$$p_n(x) \quad \longmapsto \quad (\mathbf{Pr}\, p_n)(x) \stackrel{def}{=} \sum_{i=0}^{\frac{n}{2}} \alpha_{2i} x^{2i}, \tag{2.93}$$

then $\mathbf{Pr}$ *projects a polynomial* $p_n \in \mathcal{P}_n$ *onto the linear subspace of polynomials of even degree.*

Linear Subspace (855)

Let us consider the particular case $(n = 2)$, *then—as shown in Figure 2.11 for the polynomial* x^2+2x+1*—*$\mathbf{Pr}$ *maps a square polynomial, represented by the vector* $(\alpha_0, \alpha_1, \alpha_2) \in \mathbb{R}^3$, *onto the vector* $(\alpha_0, \alpha_2) \in \mathbb{R}^2$, *that is, the even polynomial* $\alpha_0 + \alpha_2 x^2$.

REMARK 2.19 *With respect to finding solutions to the global illumination problem, the orthogonal projection occupies a particular position among projection operators. So, the theoretical foundation of a highly popular rendering method often used in computer graphics, the so-called* radiosity procedure *[36, Cohen and Wallace 1993], [13, Ashdown 1994], [190, Sillion and Puech 1994] and [68, Glassner 1995], is the Best Approximation Theorem and the concept of the linear projection operator from functional analysis. In particular, radiosity algorithms resulting from the* Galerkin method *are based on orthogonal projection methods used for the numerical solution of* linear integral equations.

Global Illumination Problem (6)
Section 2.3.3
Chapter 10
Section 2.3.3.2.2
Linear Integral Equations (126)

ADJOINT OPERATORS. Based on the concept of the linear operator we will now present a technique that proves to be very interesting in understanding light transport algorithms. It will allows us to evaluate measurements in a variety of ways leading to new insights and approaches for solving the light transport equation.

Chapter 5

Hilbert Space (36)

DEFINITION 2.17 (Linear Adjoint Operator) *Let us assume that $\mathcal{S}$ and $\mathcal{T}$ are given Hilbert spaces, and $\mathbf{T} : \mathrm{Dom}(\mathbf{T}) \subset \mathcal{S} \longrightarrow \mathcal{T}$ be a linear operator between $\mathcal{S}$ and $\mathcal{T}$. Then, we call a linear mapping $\mathbf{T}^*$ defined by*

$$\mathbf{T}^* : \mathrm{Dom}(\mathbf{T}^*) \subset \mathcal{T} \longrightarrow \mathcal{S} \tag{2.94}$$

where it holds

$$\langle \mathbf{T}^*\mathbf{x}, \mathbf{y} \rangle_{\mathcal{S}} \stackrel{\mathrm{def}}{=} \langle \mathbf{x}, \mathbf{T}\mathbf{y} \rangle_{\mathcal{T}}, \tag{2.95}$$

for all $\mathbf{x} \in \mathrm{Dom}(\mathbf{T}^*), \mathbf{y} \in \mathrm{Dom}(\mathbf{T})$, the linear adjoint operator *to $\mathbf{T}$. If in particular it holds that*

$$\mathbf{T} = \mathbf{T}^*, \tag{2.96}$$

then the operator $\mathbf{T}$ is termed as self-adjoint.

EXAMPLE 2.24 (The Transpose of a Matrix) *A well-known example of the adjoint of a linear operator is the transpose of a matrix as a mapping between two finite-dimensional linear spaces.*

Let us assume, $\mathbf{A}$ be a linear mapping between the linear spaces $\mathbb{R}^m$ and $\mathbb{R}^n$. Then, the following clearly holds to the operator $\mathbf{A}^\mathsf{T} : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ with $a^\mathsf{T}_{ij} \stackrel{\mathrm{def}}{=} a_{ji}, 1 \leq i \leq m, 1 \leq j \leq n$ for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$

$$\langle \mathbf{A}^\mathsf{T}\mathbf{x}, \mathbf{y} \rangle = (\mathbf{A}^\mathsf{T}\mathbf{x})^\mathsf{T}\mathbf{y} = \mathbf{x}^\mathsf{T}\mathbf{A}^{\mathsf{T}\mathsf{T}}\mathbf{y} = \mathbf{x}^\mathsf{T}(\mathbf{A}\mathbf{y}) = \langle \mathbf{x}, \mathbf{A}\mathbf{y} \rangle, \tag{2.97}$$

i.e. the transpose of a matrix represents an adjoint operator between finite-dimensional linear spaces. In particular, the matrix $\mathbf{A}$ is a self-adjoint linear operator if it is symmetric, that is in our case, if it is a mapping between $\mathbb{R}^m$ or respectively $\mathbb{R}^n$.

Inner Product (859)

Making use of the linearity properties of the inner product, it is straightforward to show that any adjoint operator $\mathbf{T}^*$ satisfies the linearity property of an operator, that is,

$$\begin{aligned}
\langle \mathbf{T}^*(\alpha\mathbf{x} + \beta\mathbf{y}), \mathbf{z} \rangle &= \langle (\alpha\mathbf{x} + \beta\mathbf{y}), \mathbf{T}\mathbf{z} \rangle && (2.98)\\
&= \langle \alpha\mathbf{x}, \mathbf{T}\mathbf{z} \rangle + \langle \beta\mathbf{y}, \mathbf{T}\mathbf{z} \rangle && (2.99)\\
&= \alpha\langle \mathbf{x}, \mathbf{T}\mathbf{z} \rangle + \beta\langle \mathbf{y}, \mathbf{T}\mathbf{z} \rangle && (2.100)\\
&= \alpha\langle \mathbf{T}^*\mathbf{x}, \mathbf{z} \rangle + \beta\langle \mathbf{T}^*\mathbf{y}, \mathbf{z} \rangle. && (2.101)
\end{aligned}$$

Finally, we still mention a few important properties of adjoint linear operators, which play a central role in Chapter 5. Thus it holds for two operators $\mathbf{T}$ and $\mathbf{S}$:

$$\mathbf{I}^* = \mathbf{I} \tag{2.102}$$
$$(\mathbf{T}+\mathbf{S})^* = \mathbf{T}^*+\mathbf{S}^* \tag{2.103}$$
$$(\mathbf{TS})^* = \mathbf{S}^*\mathbf{T}^* \tag{2.104}$$
$$(\mathbf{T}^{-1})^* = (\mathbf{T}^*)^{-1}. \tag{2.105}$$

In a similar way as we proof the linearity of an adjoint operators, we can also proof the Identities (2.102) - (2.105). We omit this work, and let the proof to the interested reader.

2.1.5 LINEAR OPERATOR EQUATIONS

A famous problem in functional analysis, resulting from many applications in practice, is the so-called *fixed-point problem*. It is based on the concept of the linear operator, and can be described by a so-called *linear operator equation* of the type

Linear Operator (53)

$$f(x) = (\mathbf{T}f)(x), \tag{2.106}$$

where $\mathbf{T}$ corresponds to a linear operator on the complete linear normed space $\mathcal{S}$ and f is an element from S.

Complete Linear Space (35)

If the involved operator in Equation (2.106) satisfies the condition that it is contractive, that is, the norm of the operator is less than one, then an elegant technique for solving this equation results from the *Banach Fixed-point Theorem*:

THEOREM 2.3 (Banach Fixed-point Theorem) *Let $\mathcal{S}$ be a Banach space over $\mathbb{K}$ and $\mathcal{D}(\mathbf{T})$ as well as $\mathcal{R}(\mathbf{T})$ non-empty, closed subsets of $\mathcal{S}$, furthermore let $\mathbf{T} : \mathcal{D}(\mathbf{T}) \subseteq \mathcal{S} \longrightarrow \mathcal{R}(\mathbf{T}) \subseteq \mathcal{S}$ be a* contracting linear operator*, i.e.,*

$$\|\mathbf{T}x - \mathbf{T}y\| \leq k\|x-y\|, \tag{2.107}$$

$\forall x, y \in \mathcal{D}(\mathbf{T}), 0 \leq k < 1$, then the operator equation

$$x = \mathbf{T}x \tag{2.108}$$

has a unique solution $x \in \mathcal{S}$ and the iteration

$$x_{n+1} = \mathbf{T}x_n, \qquad n = 0, 1, \ldots \tag{2.109}$$

converges for every $x_0 \in \mathcal{D}(\mathbf{T})$ towards the exact solution x.

PROOF 2.3 *We omit the proof and point to [16, Atkinson & Han 2007].*

Applied to Equation (2.106), the Banach Fixed-point theorem then guarantees the existence of a solution, which can recursively be computed via the limit of a sequence of functions $(f_n)_{n\in\mathbb{N}_0}$ defined on $\mathcal{S}$, and given by:

$$f_{n+1}(x) \stackrel{\text{def}}{=} (\mathbf{T}f_n)(x). \tag{2.110}$$

Cauchy Sequence (35)

This can easily be verified by showing, that the sequence $(f_n)_{n\in\mathbb{N}_0}$ corresponds to a Cauchy sequence constructed over the space $\mathcal{S}$. For that purpose, let us consider the distance between the sequence elements f_{n+m} and f_n, obviously, it holds:

$$\begin{aligned}
\|f_{n+m} - f_m\| &= \left\| \sum_{i=m+1}^{n+m} (f_i - f_{i-1}) \right\| && (2.111)\\
&\stackrel{(2.110)}{=} \left\| \mathbf{T}^m \sum_{i=1}^{n} (f_i - f_{i-1}) \right\| && (2.112)\\
&\stackrel{\Delta\text{-I.E.}}{\leq} \|\mathbf{T}^m\| \sum_{i=1}^{n} \|f_i - f_{i-1}\| && (2.113)\\
&= \|\mathbf{T}^m\| \sum_{i=0}^{n-1} \|\mathbf{T}^i\| \|f_1 - f_0\| && (2.114)\\
&= \|\mathbf{T}\|^m \frac{1}{1-\|\mathbf{T}\|} \|f_1 - f_0\|, && (2.115)
\end{aligned}$$

where we have used the sum formula for the geometric series in the last step under the assumption, that $n \to \infty$.

Since the operator $\mathbf{T}$ was assumed to be contracting, i.e. $\mathbf{T}$ satisfies the condition: $\|\mathbf{T}\| < 1$, the distance of elements of f_n for sufficiently large n goes to zero. This means, that the sequence $(f_n)_{n\in\mathbb{N}_0}$ corresponds to a Cauchy sequence which converges to the exact unique solution of Equation (2.106), namely f. This can easily be verified as follows: From $f_0 \in \mathcal{S}$, we get by induction, that also $f_{n+1}(x) = (\mathbf{T}f_n)(x) \in \mathcal{S}$ for all $n \in \mathbb{N}_0$. Since $\mathcal{S}$ is a closed set, we obtain with $f_n \in \mathcal{S}$ that it also holds: $\mathbf{T}f_n \in \mathcal{S}$. Due to the Banach Fixed-point Theorem, the absolute error of the approximate solution f_n can then be estimated by:

$$\begin{aligned}
\|f_n - f\| &= \|(\mathbf{T}f_{n-1}) - (\mathbf{T}f)\| && (2.116)\\
&\leq \|\mathbf{T}\| \|f_{n-1} - f\| && (2.117)\\
&\leq \|\mathbf{T}^n\| \|f_0 - f\| \stackrel{\|\mathbf{T}\|<1}{\longrightarrow} 0, && (2.118)
\end{aligned}$$

which, with $f_{n+1} = (\mathbf{T}f_n)$, implies that it holds: $f(x) = (\mathbf{T}f)(x)$.

REMARK 2.20 (Iteration Methods) *In mathematics, equations of the form (2.110) are called* iteration methods. *The idea behind an iteration method is to transform the*

given problem into the above fixed-point problem *of form,*

$$f_{n+1}(x) \quad = \quad (\mathbf{T}f_n)(x), \tag{2.119}$$

where $\mathbf{T}$ *is a linear operator, and* $f_n(x)$ *is any arbitrary starting value. Under the corresponding conditions of the Banach Fixed-point Theorem, repeated application of this formula then leads to new, always better approximations to the fixed-point, namely: the solution of the original equation.*

Due to make our formulas more readable, we will in the following also often write the above fixed-point problem in the form

$$f^{(k+1)}(x) \quad = \quad (\mathbf{T}f^{(k)})(x), \qquad k \geq 0, \tag{2.120}$$

that is, we use upper indices instead of lower indices. This notation has its advantage in particular when we consider iteration methods on the n*-dimensional, linear normed space* $(\mathbb{R}^n, \|\cdot\|)$*, where the unknown* f *is a vector* $f = (f_1, \ldots, f_n)$ *consisting of* n *coordinates.*

EXAMPLE 2.25 *When deriving a mathematical formulation for light transport in Chapter 4, we will encounter apart from the* stationary light transport equation in free space *its adjoint counterpart, the* stationary importance transport equation in free space. *Based on the construction of the linear as well as the adjoint operator, both may be written more simply as operator equations of the type*

SLTEV (296)
SITEV (413)
Chapter 5

$$f(x) = g(x) + (\mathbf{T}\,f)(x), \tag{2.121}$$

where $g(x)$ *is a given real-valued function, a so-called source function, and* $\mathbf{T}$ *is a linear operator of the type introduced in the preceding section.*

By defining a linear operator $\widetilde{\mathbf{T}}$ *as:*

$$(\widetilde{\mathbf{T}}f)(x) \stackrel{\text{def}}{=} g(x) + (\mathbf{T}f)(x), \tag{2.122}$$

Equation (2.121) can then be written as an operator equation of the form (2.106), thus:

$$f(x) = (\widetilde{\mathbf{T}}f)(x). \tag{2.123}$$

Choosing a sequence of functions $(f_n)_{n\in\mathbb{N}_0}$ *according to Equation (2.110) by*

$$f_0(x) \quad \equiv \quad 0 \tag{2.124}$$

$$f_{n+1}(x) \quad = \quad \widetilde{\mathbf{T}}f_n(x), \qquad n \geq 0, \tag{2.125}$$

then the sequence $(f_n)_{n\in\mathbb{N}_0}$ *is a Cauchy sequence that, under the condition of the contraction of the integral operator* $\mathbf{T}$*, converges towards the actual solution*

Cauchy Sequence (35)

$$\widetilde{\mathbf{T}} = (\mathbf{I} - \mathbf{T})^{-1} \tag{2.126}$$

of Equation (2.121).

In Section 2.3.3.1.1, we pick up this idea to solve Fredholm integral equations of the 2^{nd} kind. There, we use the fact that the infinite sum of powers of a contracting linear operator $\mathbf{T}$ can be written as the inverse of the operator $(\mathbf{I}-\mathbf{T})$, thus $(\mathbf{I}-\mathbf{T})^{-1}$.

LEMMA 2.1 *Let us consider the sequence*

$$(\mathbf{t}_n)_{n\in\mathbb{N}_0} = \sum_{i=0}^{n} \mathbf{T}^i = \mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \mathbf{T}^3 + \ldots, \tag{2.127}$$

on the linear normed function space $\mathcal{S}$, where $\mathbf{T}$ is a contracting linear operator. Let
Complete Linear Space (35)
us furthermore assume that the space $\mathcal{S}$ is complete, then it holds:

$$\lim_{n\to\infty} \mathbf{t}_n = (\mathbf{I}-\mathbf{T})^{-1}, \tag{2.128}$$

Cauchy Sequence (35)
that is, the Cauchy sequence $\mathbf{t}_n$ converges to an element of $\mathcal{S}$ which can be written as $(\mathbf{I}-\mathbf{T})^{-1}$.

PROOF 2.1 *As is easily seen, it holds:*

$$\begin{aligned}
(\mathbf{I}-\mathbf{T})\,\mathbf{t}_n &= (\mathbf{I}-\mathbf{T}) \sum_{i=0}^{n} \mathbf{T}^i && (2.129)\\
&= \sum_{i=0}^{n} \mathbf{T}^i\,(\mathbf{I}-\mathbf{T}) && (2.130)\\
&= \sum_{i=0}^{n} \mathbf{T}^i - \sum_{i=1}^{n+1} \mathbf{T}^i && (2.131)\\
&= \mathbf{I} + \sum_{i=1}^{n} \mathbf{T}^i - \sum_{i=1}^{n} \mathbf{T}^i - \mathbf{T}^{n+1} && (2.132)\\
&= \mathbf{I} - \mathbf{T}^{n+1}, && (2.133)
\end{aligned}$$

that is,

$$\begin{aligned}
\|(\mathbf{I}-\mathbf{T})\,\mathbf{t}_n - \mathbf{I}\| &\overset{(2.133)}{=} \left\|\mathbf{I} - \mathbf{T}^{n+1} - \mathbf{I}\right\| && (2.134)\\
&= \left\|\mathbf{T}^{n+1}\right\| \overset{\|\mathbf{T}\|<1}{\to} 0 && (2.135)
\end{aligned}$$

for sufficiently large n, thus we get: $\mathbf{t}_n \to (\mathbf{I}-\mathbf{T})^{-1}$.

2.1.6 ADJOINT EQUATIONS

Complete Linear Space (35)
In the following, let f and g be functions defined on the complete linear normed space $\mathcal{S}$, and $\mathbf{T}$ denotes a contracting linear operator on $\mathcal{S}$, where it holds:

$$f(x) = g(x) + (\mathbf{T}f)(x). \tag{2.136}$$

Let us further assume, that the operator $\mathbf{T}$ has an adjoint, $\mathbf{T}^*$. Then, for any arbitrary function $h \in \mathcal{S}$ there exists a function $i \in \mathcal{S}$ satisfying the equation

$$h(x) = i(x) + (\mathbf{T}^* h)(x). \tag{2.137}$$

The Equations (2.136) and (2.137) are said to be adjoints of each other, often they are also denoted as a pair of *adjoint equations*.

Since $\mathbf{T}^*$ is also a linear operator, solutions to adjoint equations are evidently—under the assumption that $\mathbf{T}^*$ is contracting—given by:

$$(\mathbf{I} - \mathbf{T}^*)^{-1} i(x) \overset{(2.102)-(2.105)}{=} \left((\mathbf{I} - \mathbf{T})^{-1}\right)^* i(x). \tag{2.138}$$

Now, an interesting property between the solutions of an operator equation and its associated adjoint equation is hidden in the linear functionals $\langle f(x), i(x)\rangle$ and $\langle g(x), h(x)\rangle$. Linear Functional (55)

Let us assume, we are interested in measuring the function f with respect to any function i. Due to our discussion in Section 2.1.4 this is equivalent to the evaluation of the linear functional $\langle f(x), i(x)\rangle$, that is, it holds: Measurement Equation (416)

$$\begin{aligned}
\langle f(x), i(x)\rangle &\overset{(2.126)}{=} \langle (\mathbf{I} - \mathbf{T})^{-1} g(x), i(x)\rangle && (2.139)\\
&\overset{(2.17)}{=} \langle g(x), \left((\mathbf{I} - \mathbf{T})^{-1}\right)^* i(x)\rangle && (2.140)\\
&\overset{(2.105)}{=} \langle g(x), ((\mathbf{I} - \mathbf{T})^*)^{-1} i(x)\rangle && (2.141)\\
&\overset{(2.138)}{=} \langle g(x), h(x)\rangle. && (2.142)
\end{aligned}$$

Obviously, there are two independent ways to achieve the same effect. Namely, solving the direct problem

$$f(x) = g(x) + (\mathbf{T} f)(x) \tag{2.143}$$

followed by measuring the solution function f with respect to any function i delivers the same result as solving the *dual problem*, that is, the adjoint equation

$$h(x) = i(x) + (\mathbf{T}^* h)(x), \tag{2.144}$$

and measuring its solution h with respect to the source function g of the direct problem.

One question that now arises is: Which way should be follow? Now, the answer to this question depends on the problem which has to be solved. From the mathematical point of view there is no difference in implementing either the direct or the adjoint formulation.

REMARK 2.21 *The above result is fundamental for the field of realistic rendering algorithms. Applied to the stationary light transport, discussed in Chapter 5, the concept of the adjoint operator allows us to evaluate measurements in a variety of ways leading to new insights and approaches for solving the light transport equation.* Chapter 9

Another example for the use of a linear functional is the problem of antialiasing where we are forced to measure the weighted average radiance for each pixel of the image plane.

2.2 A BIT OF MEASURE AND INTEGRATION THEORY

Chapter 9

As will be seen in more detail below, Monte Carlo rendering algorithms are based on measure and probability theoretical approaches for solving the global illumination equation approximatively. Therefore, and particularly due to Kolmogorov's interpretation of the probability of an event as the measure[2] of a certain set, fundamental knowledge of the concepts and methods relating to general measure and integration theory are necessary for the understanding and analyzing rendering processes where stochastic methods are used.

Broadly speaking, measure and integration theory concerns with the theoretical basis of measurements of the content of intervals, surfaces, and volumes. Here, of particular interest is the measurement of complex sets in higher dimensional spaces, where we will utilize already known results of elementary geometry from earlier days of mathematics. So, to assign to such a region a specific volume, a useful approach is to represent this region as a disjoint union of elementary sets and to define the sum of the volumes of theses elementary sets as the content of the considered region. As we will see, in such a case it is not only sufficient to restrict the decomposition of a complex region into a finite number of elementary sets, but we also has to account for countably infinite representations. Then, measure theory shows that a measure with natural properties may be assigned, if not to all, then at least to all open and closed sets of the given region. The concept of the *Lebesgue integral*, which results from this procedure, is obtained by means of a natural approximation process, in which the classical *Riemann integral* is expanded to form the modern Lebesgue integral, generally recognized as the basis of the modern functional analytical theory of differential and integral equations.

Countable Set (827)
Open Set (864)
Riemann Integral (876)

On the basis of the Lebesgue integral, a strictly mathematical treatment of many complex problems of physics, including the questions of interest here on global illumination, is possible via the construction of function spaces based on sets of Lebesgue-integrable functions. Due to the beautiful limiting properties of the Lebesgue integral these function spaces represent complete metric spaces in which the fundamental Cauchy convergence criterion holds, which is not valid for the classical Riemann integral.

Function Space (28)
Metric Space (866)
Cauchy Sequence (35)

The goal of this section is the introduction of the Lebesgue integral and the associated function spaces underlying the global illumination problem. For that purpose, we need some background from measure theory, that is, we will first present the concept of the *outer Lebesgue measure* as the intuitive tool for measuring intervals, and then classify, via *Carathéodory's measurability criterion*, the subsets of $\mathbb{R}$ to which we can assign a

Section 2.2.1

[2] Modern measure theory goes back to the discovery of the σ-additivity of the elementary geometrical length by *Émile Borel*, on the basis of which *Henry Lebesgue* shows in his *thèse*, that it may be continued to form a *measure* on a certain σ-algebra of subsets of $\mathbb{R}$, which he refers to as *measurable sets*. Henry Lebesgue's most important contribution lies in his foundation of the *Lebesgue integral*, an integral concept with a flexibility highly superior to its predecessor, the *Riemann integral*.

measure. Because the construction of measurable sets via Carathéodory's measurability criterion is hard, we show, at the example of the construction of the *Borel measurable sets* on $\mathbb{R}$, that the measurability of a set is closely related to the concept of the σ-algebra of open sets on $\mathbb{R}$.

Afterwards, we will devote to the concept of the *general measure* as a real-valued, σ-additive, and non-negative set function defined on any σ-algebra over a base set $\mathcal{R}$. We will present the most important properties of a measure, and introduce the concept of the *discrete measure*, of fundamental importance for *discrete probability spaces*. We also present the concept of the *product measure* as a set function defined on σ-algebras over Cartesian products. With the help of examples, we present the most important measures needed for an understanding of the concepts of global illumination theory. These will comprise the *Dirac measure* and the *counting measure*, the *Lebesgue measure on* $\mathbb{R}^n$, the *solid angle measures*, as well as the *throughput measures*, as the measures actually underlying all the mathematical formulations to the global illumination problem. Section 2.2.2

The topic of Section 2.2.3 will be the construct of the *measurable function*. It is needed not only for the derivation of the Lebesgue integral, but also serves as the basis for the definition of the probabilistic theoretical concept of the *random variable*. After defining measurable functions, we investigate some well known functions with respect to their measurability and introduce the ν-almost everywhere property. We also investigate different types of convergence of sequences of measurable functions that are of fundamental importance for the *limit theorems* of probability theory. Section 2.2.3

The derivation of the *Lebesgue integral* as the fundamental concept of measure and integration theory is given in Section 2.2.4. Based on the Lebesgue integral, we introduce the construct of the Lebesgue spaces, denoted by $\mathcal{L}^p(\mathcal{R}, \mu)$, that is, normed functions spaces defined on a general set $\mathcal{R}$, whose norm is given via the Lebesgue integral. In this context, we also construct the *Lebesgue spaces* $\mathcal{L}^p(\mathcal{R}^{\partial\mathcal{V}}, \zeta^{\perp})$, $\mathcal{L}^p(\mathcal{R}^{\mathcal{V}^o}, \zeta)$ and $\mathcal{L}^p(\mathcal{R}, \zeta)$, defined on the ray spaces that we will use in our following discussions as the basis for generating particle distributions in a scene to be rendered. Here, we are interested in the cases $p = 1$ and, in particular $p = 2$, thus, the spaces that serve as the basis of all rendering procedures based on the principle of *radiosity* and representing the only Hilbert space that may be constructed over the ray spaces known from Section 2.1.3. Afterwards, we discuss the functional analytical construct of the *Fourier Transform*—a highly useful tool which allows to analyze the efficiency of patterns, resulting from different sampling processes in Monte Carlo and quasi-Monte Carlo procedures. We finish this subsection with the *Fubini-Tonelli Theorem*, that makes statements about the evaluation of the Lebesgue integral on multi-dimensional integration domains. Section 2.2.4

In the last part of this section, we present the *Dirac δ-distribution*, an important mathematical construct based on the notation of the Lebesgue integral, and we will show, how this concept can be used to describe the ideal specular reflection and ideal refraction of light at object surfaces in an elegant mathematical way. After this we show that some Lebesgue square-integrable functions can be represented as a infinite series of so-called Section 2.2.5

spherical harmonic functions, a technique that proves to be useful for representing *bidirectional reflectance distribution functions*.

2.2.1 AN INTUITIVE APPROACH TO THE LEBESGUE MEASURE ON $\mathbb{R}$

Even if probability theory requires to consider general measure spaces we will introduce in this section the concept of a measure rather via a more intuitive approach. Contrary to the common procedure of introducing a measure by specifying a set function defined on a σ-algebra, our approach is based on the length of a decomposition of an arbitrary set as a collection of points and intervals. We expect that this approach demonstrates how this important measure, the *Lebesgue measure on the real line*, arises quite naturally from considerations of the lengths of sets of real numbers and leads to a theory of integration which greatly extends that of Riemann.

Set Function (837)
σ-algebra (828)
Lebesgue Measure (71)
Riemann Integral (876)

MOTIVATION FOR DEFINING THE LEBESGUE MEASURE. If we are interested in evaluating the area between the graph of the Dirichlet function and the x-axis in any closed interval $[a, b] \subset \mathbb{R}$ then we have a problem, as the Dirichlet function is not Riemann-integrable. Since the Dirichlet function takes non-zero values only on intervals of type $[a, b] \cap \mathbb{Q}$, where it equals one, the area under the graph should be very closely linked to the length of the interval $[a, b] \cap \mathbb{Q}$. Because the sets $\mathbb{Q}$ and $\mathbb{R} \setminus \mathbb{Q}$ are so different from intervals, one has to ask: How should we measure the lengths of such more general sets? An elegant way of defining the length of such a set could be to describe the set by intervals and to measure the lengths of these intervals.

Dirichlet Function (836)
Riemann Integrable (876)

It is well-known that the length or the *measure*, as we will say in the future, of an interval $[a, b] \subset \mathbb{R}, a \neq b$, equals to $b - a$, that is: The closed interval $[0, 1]$ has obviously the measure one, and this should be ok. If we now consider half-open or open intervals what should be the measure of any of these intervals? Now, because these sets are subsets of the closed interval $[a, b]$ their measure should surely not exceed that of $[a, b]$, in mathematics one says, the desired measure should be *monotonic*. Similarly, it would be make sense, if we define the measure of an interval, where at least one of the end points lies at infinity, intuitively to be infinite. Obviously, we can conclude that the measure, which we are trying to develop, should be a monotonic function, that maps an arbitrary subset of $\mathbb{R}$ to the non-negative extended real numbers, we also speak of a so-called *monotonic set function*.

Closed Interval (829)
Open Interval (829)
Set Function (837)

Further, it should be irrelevant for our measure whether we consider the interval $[a, b] \subset \mathbb{R}$ or the intervall $[c + a, c + b]$ for $c \in \mathbb{R}$: Both intervals should have the same measure, i.e. our measure should be *translation invariant*.

Since any interval $[a, b]$ can be decomposed in two disjoint intervals $\left[a, \frac{b-a}{2}\right)$ and $\left[\frac{b-a}{2}, b\right]$, we conclude from the decomposition $[a, b] = \left[a, \frac{b-a}{2}\right) \cup \left[\frac{b-a}{2}, b\right]$ as well as from

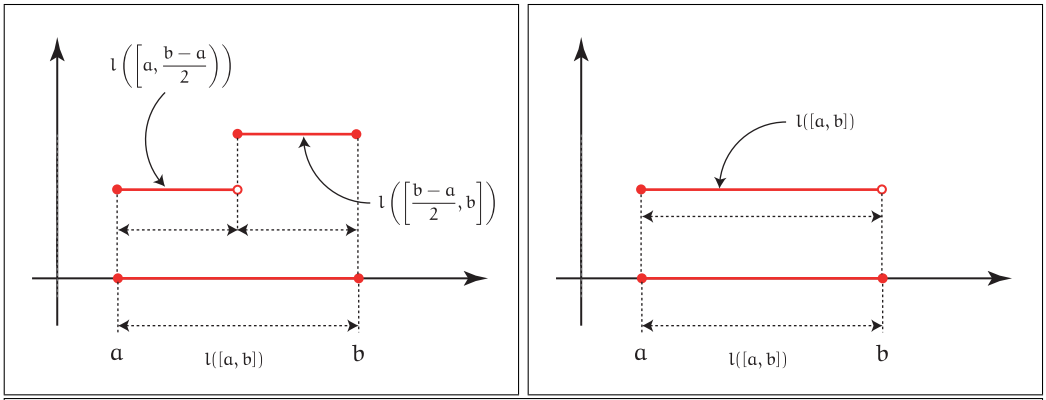


FIGURE 2.12: COMPOSITION OF INTERVALS. Left, the composition of the closed interval $[a,b]$ into a half-open interval $\left[a, \frac{b-a}{2}\right)$ and a closed interval $\left[\frac{b-a}{2}, b\right]$ of the same length. The image on the right-hand side visualizes the decomposition of a closed interval $[a,b]$ into a half-open interval $[a,b)$ with the same length as the interval $[a,b]$ and the point b with length zero.

the lengths of the intervals $[a,b]$ and $\left[\frac{b-a}{2}, b\right]$, thus, $l([a,b]) = b-a$ and $l\left[\frac{b-a}{2}, b\right] = b - \frac{b-a}{2}$, that the measure of the half-open interval $\left[a, \frac{b-a}{2}\right)$ and the measure of the closed interval $\left[\frac{b-a}{2}, b\right]$ must be the same, see Figure 2.12. This means that our measure should be *additive*. This additivity then implies that the measure of a single point set is zero, as it holds: $[a,b] = [a,b) \cup \{b\}$, that is, the measures of a closed, half-open, or an open interval with same end points must be equal, see Figure 2.12. Obviously, the additivity appears to be a reasonable requirement to a measure because points are dimensionless and consequently should have the measure zero.

l (836)

Before we extend this idea to more general sets let us first consider the length of a finite point set. Now, a finite point set is not an interval. Because it consists of a finite number of points, where every single point has measure zero, the measure of a finite point set should also be zero. The same argumentation should hold for countably infinite point sets. Since any countably infinite set consist of single points of measure zero, the measure of such a set should also be zero.

Section 2.2.2

Countably Infinite Set (827)

Expressing the closed interval $[a,b]$ as an uncountable infinite union of points, that is,

Uncountable Set (827)

$$[a,b] = \bigcup_{a \in [a,b]} \{a\}, \tag{2.145}$$

then we have a problem. The measure of this interval is $b-a \neq 0$, but the measure of the expression on the right-hand side is—as the sum of an uncountable infinite number of points of measure zero—again zero. Obviously, here we have a discrepancy, since the whole is not equal to the sum of its parts. In this context, let us consider the closed

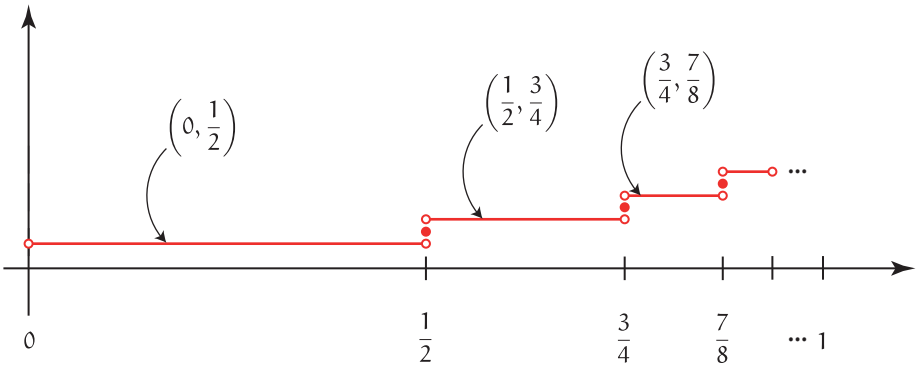


FIGURE 2.13: A COUNTABLE UNION OF OPEN INTERVALS AND POINTS. The closed interval $[0,1]$ and its composition into a countable union of open intervals and points.

interval $[0,1]$. Evidently, this interval can be decomposed into a countable union of open intervals and points, namely,

$$[0,1] = \{0\} \cup \left(0, \frac{1}{2}\right) \cup \left\{\frac{1}{2}\right\} \cup \left(\frac{1}{2}, \frac{3}{4}\right) \cup \left\{\frac{3}{4}\right\} \cup \left(\frac{3}{4}, \frac{7}{8}\right) \cup \left\{\frac{7}{8}\right\} \cup \ldots \quad (2.146)$$

$$= \bigcup_{i=1}^{\infty} \left(\frac{2^{i-1}-1}{2^{i-1}}, \frac{2^i-1}{2^i}\right) \cup \bigcup_{i=0}^{\infty} \left\{\frac{2^i-1}{2^i}\right\}, \quad (2.147)$$

see Figure 2.13.

As it is easily seen, the sum of the lengths of the individual intervals on the right-hand side results in the geometric series

$$\sum_{i=1}^{\infty} \frac{1}{2^i} = \frac{1}{1-\frac{1}{2}} - 1 = 1, \quad (2.148)$$

and the measure of the points $\frac{2^i-1}{2^i}$ is zero, i.e. the whole is the sum of its parts again.

Obviously it appears, that, only in the case where a set is decomposable into a countable union of disjoint sets, our desired measure should have the property to be additive, exactly spoken, it should be σ-*additive* or *countably additive*.

Now, the above considerations also suggest that countable sets should have measure zero and, if it is possible to decompose a set into a finite number of disjoint sets, the measure of such a set should be the sum of the measures of the corresponding pieces. $\mathbb{Q}$ (827) Applied to the set of rational numbers, this means, that the measure of $\mathbb{Q}$ is zero and that the measure of any interval $[a,b] \subset \mathbb{R}, a \neq b$ is different from zero as any interval of this type consist of an uncountable number of points.

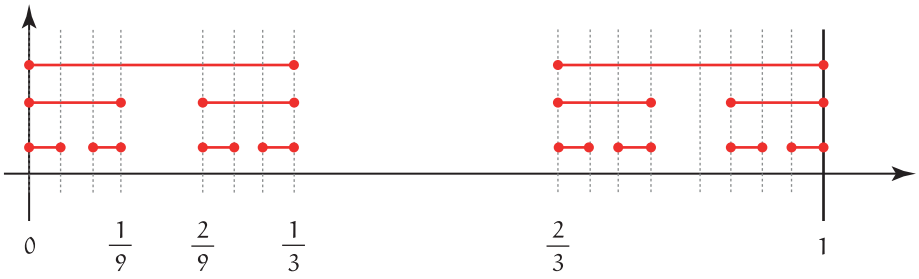


FIGURE 2.14: THE CANTOR SET ON $[0,1]$. The Cantor ternary set is created by repeatedly deleting the open middle thirds of a closed interval $[0,1]$. One starts by deleting the open middle third $\left(\frac{1}{3},\frac{2}{3}\right)$ from the interval $[0,1]$ resulting in two intervals $\left[0,\frac{1}{3}\right]\cup\left[\frac{2}{3},1\right]$. Next, the open middle third of each of these remaining intervals is deleted. This process is continued ad infinitum.

In summary, it is fair to say that any countable infinite set is a so-called *null set*, that is, a set with measure zero. Evidently, null sets appear to be closely related to countable sets. This is certainly not surprising because any proper interval is uncountably infinite and the points of any countable subset, compared with an interval, are quite sparsely distributed, hence making no real contribution to its length.

REMARK 2.22 *The concept of the null set is fundamental when discussing direct and indirect illumination of a surface point, and here in particular if we consider point light sources. As point light sources can be interpreted as points of a set, a set of point light sources has always measure zero.* Section 4.4.2.2

Finally, let us consider the interval $[0,1] \subset \mathbb{R}$. The great german mathematician Cantor showed that it is possible to remove a countable number of disjoint intervals from $[0,1]$, whose total measure is one. What remains is the so-called *Cantor set*, see Figure 2.14, an uncountable infinite set of measure zero. The Cantor set is an example of an uncountable set, in which the points are sufficiently sparsely distributed although the Cantor set is uncountably infinite with measure zero. Examples like the Cantor set require a careful and mathematically based approach to define the concept of a measure. Superficial phrases like *should be*, *the whole is the sum of its parts* etc. must be replaced with a careful analysis, and this is what we want to do. But our discussions above has been fruitful because we now know what conditions our desired measure should satisfy.

THE LEBESGUE MEASURE ON $\mathbb{R}$. Motivated by the foregoing discussion, now, we are interested in constructing a function μ, also called a *measure*, defined on all subsets of $\mathbb{R}$ that assigns any of these sets a non-negative real number, see Figure 2.15. Additionally, Set Function (837)

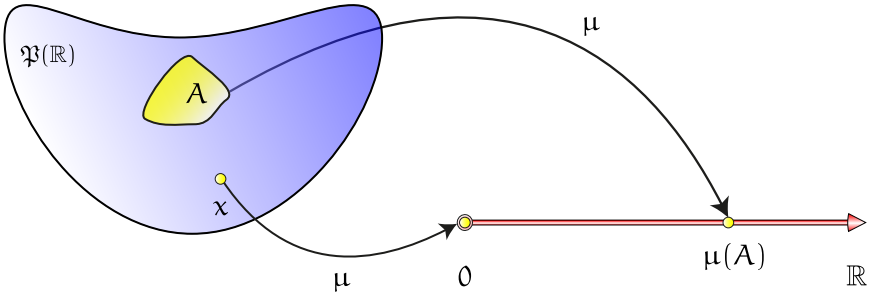


FIGURE 2.15: THE GENERAL MEASURE PROBLEM. Defined on the power set of $\mathbb{R}$, a measure μ should map any subset of $\mathbb{R}$ to a real number.

we expect that this function satisfies as many of the following conditions as possible for all sets $A, B \in \mathfrak{P}(\mathbb{R})$:

$\mathfrak{P}(\mathbb{R})$ (828)

i) First, we wish that any subset A of the real line is measurable, additionally,

ii) the measure of a set $A \subseteq \mathbb{R}$ must always be non-negative, i.e. we require: $0 \leq \mu(A) \leq \infty$.

iii) The measure of a set $A \subseteq B$ should also not exceed the measure of the set B, i.e. μ must be monotonic.

iv) Because the empty set contains no elements, its measure has to be zero.

v) Since a point $a \in \mathbb{R}$ is dimensionless, we also require that its measure $\mu(\{a\})$ is zero.

vi) The measure of an interval $[a, b]$ should corresponds to its length, namely $b - a$, which also implies, that

vii) the measure should be translation invariant, that is, $\mu([c+a, c+b]) = \mu([a, b]), c \in \mathbb{R}$ for an interval $[a, b] \subset \mathbb{R}$.

viii) Last but not least, we require that the measure μ should be σ-additive, i.e. it should hold: $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$.

Section 2.4.2 REMARK 2.23 (Drawing a Random Number) *In one of the next sections of this chapter we will show that the probability of drawing a random number from any interval* Example 2.70 $[a, b]$ *requires the concept of a measure defined over this interval. So for example* Random Variable (168) *it is desirable, that the probability of drawing a random number is always a non-*

negative real number and the probability of choosing the random number from equal length subintervals $[\alpha, \beta] \subset [a, b]$ *should always be the same, namely the length of the interval. That is, the concept of probability must be* translation invariant, *since the probability of choosing a number from* $\left[0, \frac{1}{10}\right] \subset [0, 1]$ *should be the same as the probability of choosing a number from* $\left[\frac{n-1}{10}, \frac{n}{10}\right]$ *for* $n = 1, \ldots, 10$.

In the following discussion we show how it is possible to construct a measure, that satisfies as many of the conditions (i) through (viii) as possible. For this, we utilize that any subset of the real numbers can always be covered by a countable number of intervals from $\mathbb{R}$. The key to our idea of the measure is put in the simple concept of the null set, which tells us what we can ignore when measuring a set.

Cover (865)
Null Set (80)

DEFINITION 2.18 (The Outer Lebesgue Measure on $\mathbb{R}$) *Let* $\mathfrak{P}(\mathbb{R})$ *be the set of all subsets over the base set* $\mathbb{R}$. *The outer Lebesgue measure* μ^* *is a function defined on* $\mathfrak{P}(\mathbb{R})$ *that maps any set* A *from* $\mathfrak{P}(\mathbb{R})$ *to a real number, that is,*

$\mathfrak{P}(\mathbb{R})$ (828)
Set Function (837)
Infimum (862)

$$\mu^* : \mathfrak{P}(\mathbb{R}) \longrightarrow \overline{\mathbb{R}} \stackrel{def}{=} \mathbb{R} \cup \pm\infty$$

$$A \mapsto \mu^*(A) \stackrel{def}{=} \inf \left\{ \sum_{i=1}^{\infty} l(I_i) \middle| A \subset \bigcup_{i=1}^{\infty} I_i \right\}, \tag{2.149}$$

where $\bigcup_{i=1}^{\infty} I_i$ *is a cover of* A *by open intervals and* $l(I_i)$ *denotes the length of the interval* I_i, *see Figure 2.16.*

Cover (865)
Open Interval (829)
l (836)

Evidently, the outer Lebesgue measure tries to make a statement about the minimal length of all possible open covers of A. We now ask: Is the outer Lebesgue measure our desired measure?

[30, Burk 1998]

From measure theory it is known that μ^* satisfies all of our intuitive conditions except for the last: the σ-additivity, since it holds:

σ-additivity (72)

THEOREM 2.4 *The outer Lebesgue measure is* countably subadditive, *briefly denoted as* σ-subadditive, *i.e. for any sequence of sets* $(A_i)_{i \in \mathbb{N}}$ *with* $A_i \in \mathfrak{P}(\mathbb{R})$ *it holds:*

$$\mu^* \left(\bigcup_{i=1}^{\infty} A_i \right) \leq \sum_{i=1}^{\infty} \mu^*(A_i). \tag{2.150}$$

PROOF 2.4 *We omit the proof and point to [22, Berezansky & al. 1996].*

What we really want to reach is to ensure that for a countably infinite collection of disjoint sets the Inequality (2.150) should become an equality. This is a natural requirement, since a decomposition of a set into a finite number of disjoint subsets ought not to alter its measure. If we can show that for all subsets $A, B \in \mathfrak{P}(\mathbb{R})$ the relation

$$\mu^*(A \cup B) = \mu^*(A) + \mu^*(B) \tag{2.151}$$

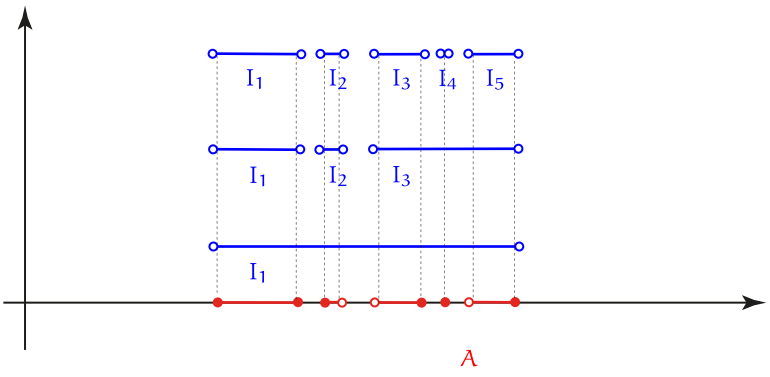


FIGURE 2.16: DEFINITION OF THE OUTER LEBESGUE MEASURE ON $\mathbb{R}$. Visualization of a set A on the real line. Shown are three covers of A by open intervals: Buttom, $A \subset I_1$, in the middle $A \subset I_1 \cup I_2 \cup I_3$, and above $A \subset I_1 \cup I_2 \cup I_3 \cup I_4 \cup I_5$.

holds for any arbitrary $A \cap B = \emptyset$, then it can be shown by induction that μ^* is σ-additive. This means: μ^* could be our desired measure.

σ-additivity (72)

Unfortunately, as one can suspect, countable subadditivity is by far the best we can get for the outer Lebesgue measure. In 1905, the italian mathematician Vitali gave a famous example of a set of real numbers that could not be decomposed in an additive fashion, for details see [30, Burk 1998]. That is, with the concept of outer Lebesgue measure we were so close to our desired measure, what should we do now?

In measure theory it is shown that the conditions (ii) through (vii) from above seem to be indispensable, i.e. if we stay with the outer Lebesgue measure, the only thing we have to do is to modify the conditions (i) or (viii), or both. Now, Vitali showed, that if one adheres to condition (viii), then the general measure problem is not solvable in $\mathbb{R}^n, n \geq 1$ [54, Elstrodt 1996]. So, it doesn't make sense to define μ^* on the whole power set $\mathfrak{P}(\mathbb{R})$. This means: To demand that every subset of $\mathbb{R}$ could be assigned a real number as its measure is too strong. The way we must go is to restrict ourselves to those sets that can be decomposed in an additive fashion with respect to the outer Lebesgue measure. But how do we filter out such sets? And even if we have determined these sets, is such a collection large enough to build a useful theory of integration?

$\mathfrak{P}(\mathbb{R})$ (828)

The crux to reach this goal is *Carathéodory's measurability criterion*, which will turn out to be the key idea of the abstract concept of a measure. It will be used to generalize the concept of length to a large class of subsets of $\mathbb{R}$, while it gives a special role to those sets, which split every other set additively.

DEFINITION 2.19 (Carathéodory's Measurability Criterion) *A set* $A \subseteq \mathbb{R}$ *is called* Lebesgue-

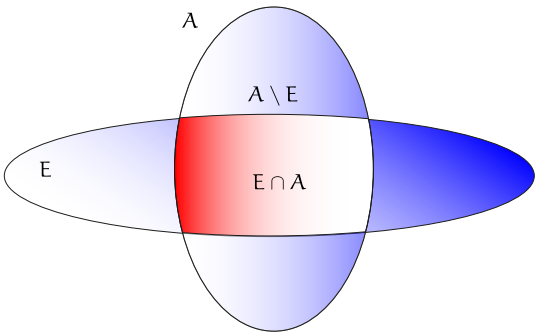


FIGURE 2.17: CARATHÉODORY'S MEASURABILITY CRITERION. A set A is Lebesgue-measurable, if A interacts in an additive fashion with every subset E of $\mathbb{R}$.

measurable *if for each* $E \subseteq \mathbb{R}$ *it holds:*

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap \overline{A}), \tag{2.152}$$

i.e. a Lebesgue-measurable set A *interacts in an additive fashion with every subset of* $\mathbb{R}$*, in other words:* A *splits every subset of* $\mathbb{R}$ *additively, see Figure 2.17.*

Now, let us show how we can use this criterion to filter out the Lebesgue-measurable sets from $\mathfrak{P}(\mathbb{R})$. For that purpose let $A \in \mathfrak{P}(\mathbb{R})$. Then, we test every subset $E \in \mathfrak{P}(\mathbb{R})$, whether it fulfills Carathéodory's measurability criterion. If we answer yes then the set A is Lebesgue-measurable. If there exists only a single set E, such that $\mu^*(E) \neq \mu^*(E \cap A) + \mu^*(E \cap \overline{A})$, then we discard A and say A is a Lebesgue non-measurable set of real numbers, see Figure 2.18. That is: the set A is Lebesgue-measurable, if it splits every subset of $\mathbb{R}$ additively relative to μ^*

$\mathfrak{P}(\mathbb{R})$ (828)

Apparently, the empty set, $\emptyset$, and the set of all real numbers, $\mathbb{R}$, are Lebesgue-measurable sets but does it make sense to develop an integration theory on a collection of only two sets? Certainly, not! So it is time to lay our focus to the concept of the σ-algebra and to show why it is so important in measure theory.

σ-alegebra (828)

To define our desirable measure, we are interested in a large σ-algebra consisting of Lebesgue-measurable sets that contains many more elements than $\emptyset$ and $\mathbb{R}$. Indeed, this will be reached via *Carathéodory's Theorem*, which provides us with one of the two most important results about the Lebesgue measure.

THEOREM 2.5 (Carathéodory Theorem) *Let* μ^* *be the outer Lebesgue measure as defined in Equation (2.149) and* $\mathfrak{M}$ *denote the collection of sets* $A \subseteq \mathbb{R}$ *that satisfies Carathéodory's measurability criterion from Equation (2.152), then it holds:*

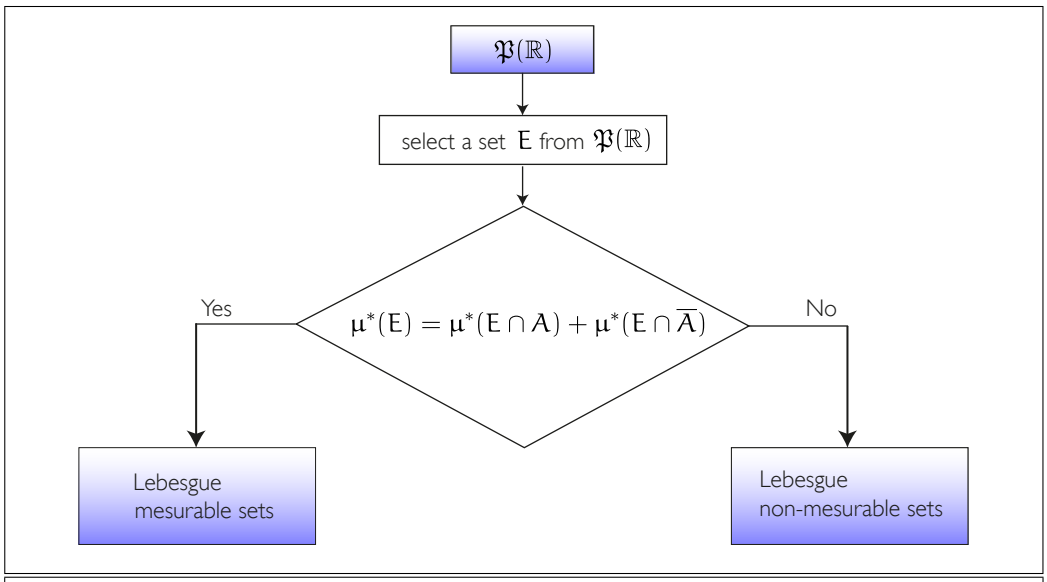


FIGURE 2.18: LEBESGUE-MEASURABLE SETS. Due to Vitali's example, we know, that the left container is not empty and the sets $\mathbb{R}$ and $\emptyset$ are Lebesgue-measurable sets.

i) $\mathfrak{M}$ is a σ-algebra and

σ-alegebra (828) *ii) the restriction μ of the outer Lebesgue measure μ^* to $\mathfrak{M}$, thus,*

$$\mu \stackrel{def}{=} \mu^*|_{\mathfrak{M}} \tag{2.153}$$

is a measure. *We call the collection $\mathfrak{M}$ the* Lebesgue-measurable sets *and μ is referred to as the* Lebesgue measure *on $\mathbb{R}$.*

PROOF 2.5 *We omit the proof and point to [30, Burk 1998].*

As already mentioned above, $\emptyset$ and $\mathbb{R}$ are surely in $\mathfrak{M}$. How does it looks like with intervals? Do intervals satisfy Carathéodory's measurability criterion? Yes, intervals are ok, they are Lebesgue-measurable [30, Burk 1998]. This fact relates the outer Lebesgue measure to the length of an interval. It contains the crux of the theory, since it demonstrates, that the formal definition of the outer Lebesgue measure, which is applicable to all subsets of the real numbers, coincides with the intuitive idea of the length of intervals, which was the start of our thought process. But are we limited to intervals?

Carathéodory Criterion (74)
μ^* (73)

As one can easily see, our definition of $\mathfrak{M}$ does not directly lend itself to verification, that a particular subset of $\mathbb{R}$ belongs to $\mathfrak{M}$. Evidently, it is also hard to show that $\mathfrak{M}$ is closed under various set theoretical operations. Therefore, we will present another

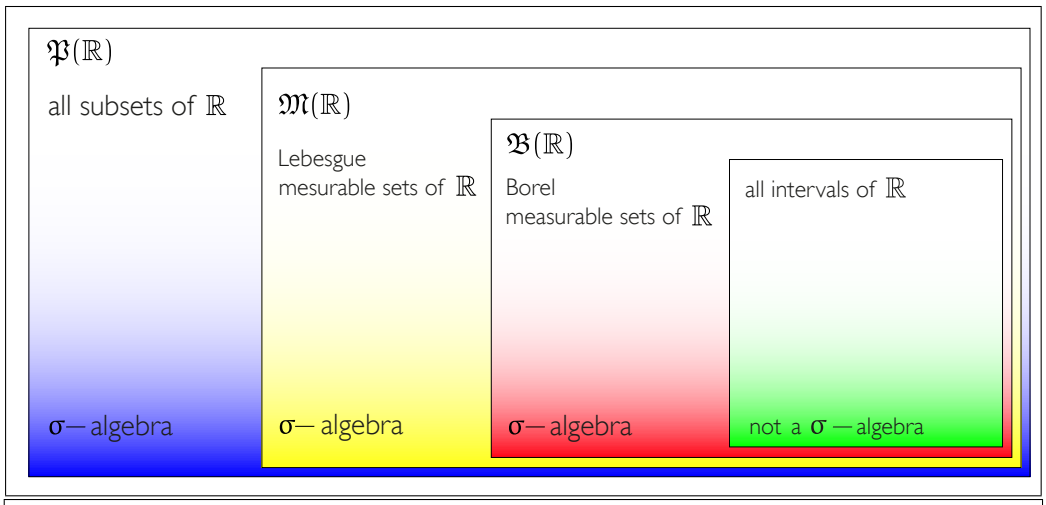


FIGURE 2.19: HIERARCHY OF MEASURABLE SETS IN $\mathbb{R}$.

construction, which shows more directly how open sets and the structure of a σ-algebra lies at the heart of many of the concepts we have developed. Because a σ-algebra is closed with respect to the set theoretical operations of countable unions, countable intersections, complements etc. we can also measure countable unions of countable intersections of complements etc. of intervals. These are very complicated sets. To simplify this, we utilize the concept of the *Borel* σ-algebra.

Countability (827)

Borel σ-algebra (865)

$\mathfrak{B}(\mathbb{R})$ be the Borel σ-algebra of real numbers, and $\mathfrak{M}(\mathbb{R})$ be the Lebesgue-measurable sets on $\mathbb{R}$, then measure theory says, that every Borel set is also Lebesgue-measurable, that is,

Borel σ-algebra (865)

Lebesgue-measurable Set (75)

$$\mathfrak{B}(\mathbb{R}) \subset \mathfrak{M}(\mathbb{R}). \tag{2.154}$$

It is just the σ-algebra of Borel-measurable sets, that plays a key role in our further discussion. We will encounter them not only when introducing the solid angle concepts and the Lebesgue integral but also when discussing spherical harmonics and many other important constructs.

Section 2.4

Solid Angle Measures (84)

Spherical Harmonics (124)

We will now take a closer look at the above proposed approach for the construction of the *Lebesgue measure* on $\mathbb{R}$, the fundamental concept of integration theory. Additionally, it allows us to develop many of the important measures used in global illumination and probability theory.

Section 2.2.2

EXAMPLE 2.26 (The Lebesgue Measure of Bounded Sets on $\mathbb{R}$) *Let $\mathfrak{B}(\mathbb{R})$ be the Borel σ-algebra generated by all half-open intervals of type $[a, b) \subset \mathbb{R}$. We define the outer*

Borel σ-algebra (865)

Intervals (829)

Lebesgue measure μ^* *on the* σ*-algebra of Lebesgue-measurable sets* $\mathfrak{M}(\mathbb{R})$ *by:*

$$\mu^*(A) \stackrel{def}{=} \inf\left\{\sum_{i=1}^{\infty} l(I_i) \middle| A \subset \bigcup_{i=1}^{\infty} I_i\right\}, \tag{2.155}$$

where $l(I) \stackrel{def}{=} (b-a)$ *is called the length of the half-open interval* $I = [a,b)$ *and* $\mu \stackrel{def}{=} \mu^*\big|_{\mathfrak{M}(\mathbb{R})}$ *is the restriction of the outer Lebesgue measure to the* σ*-algebra* $\mathfrak{M}(\mathbb{R})$.

l (836)

Clearly the σ*-algebra* $\mathfrak{B}(\mathbb{R})$ *contains not only the intervals* $[\alpha,\beta) \subseteq [a,b)$*, but also point sets of the form* $B = \{x\}$ *and* $B = \cup_{i=1}^{\infty}\{x_i\}$ *respectively. This is due to the fact, that it is always possible to take an arbitrary half-open interval* $[\alpha,\beta)$ *that contains* x *as a cover*

$$0 \leq \mu^*(\{x\}) \overset{(2.155)}{\leq} \inf_{x\in[\alpha,\beta)} l([\alpha,\beta)) = \inf_{x\in[\alpha,\beta)} (\beta-\alpha) = 0 \tag{2.156}$$

σ-additivity (72)

and the σ*-additivity of the outer measure, resulting in*

$$\mu^*\left(\bigcup_{i=1}^{\infty}\{x_i\}\right) = \sum_{i=1}^{\infty}\mu^*(\{x_i\}) = 0. \tag{2.157}$$

That is, the sets $B = \{x\}$ *and* $B = \cup_{i=1}^{\infty}\{x_i\}$ *are sets with measure zero and thus belong to the* σ*-algebra* $\mathfrak{M}(\mathbb{R})$.

Lebesgue-measurable Set (75)

It is readily to seen, that closed and open intervals of $\mathbb{R}$*—thus intervals of type* $[a,b]$ *and* (a,b)*—are also measurable. They may be easily constructed by the union of the measurable sets* $[a,b)$ *and* $\{b\}$ *respectively the difference between the interval* $[a,b)$ *and the measurable set* $\{a\}$*, i.e.* $[a,b] = [a,b) \cup \{b\}$ *and* $(a,b) = [a,b) \setminus \{a\}$ *respectively.*

Open Set (864)

Countability (827)

Closed Set (864)

Thus, not only open intervals but every kind of open sets of $\mathbb{R}$ *that may be regarded as a countable union of open intervals are Lebesgue-measurable. If we embed a closed set* F *in an open interval* (α,β)*, then the set* $G = (\alpha,\beta) \setminus F$ *is certainly open and* F*—as the difference between the open sets* (α,β) *and* G*—is Lebesgue-measurable. Finally, apart from these open and closed sets the* σ*-algebra* $\mathfrak{M}([a,b))$ *contains all sets of the type* G_o *(countable intersection of open sets),* F_a *(countable union of closed sets) and* G_{oa} *and* F_{oa} *(countable union of* G_o *sets and countable intersection of* F_o*) etc..*

2.2.2 GENERAL MEASURES

An alternative, rather abstract and theoretical more challenging approach to measure theory than those we have presented in the previous section is to start with the definition of a measure as a set function defined on a σ-algebra which requires the above properties as axioms. We go this way now because it is the common approach to construct *general*

probability measures, which will be of great interest to us.

Set Function (837)

σ-algebra (828) GENERAL MEASURES. Apart from the Lebesgue measure, the central measure concept of probability theory, we need further, rather more complex types of measures for developing new probabilistic solution approaches to the global illumination problem. So, we will now introduce the general concept of a measure, namely: as a mapping that assigns a real number to any set of a σ-algebra.

Section 5.4

DEFINITION 2.20 (Measure) *Let $\mathfrak{R}$ be a σ-algebra generated on any base set $\mathcal{R}$. Let us assume that a real-valued and* non-negative set function *ν is given on the σ-algebra $\mathfrak{R}$, i.e.*

σ-algebra (828)
Set Function (837)

$$\nu : \mathfrak{R} \longrightarrow \overline{\mathbb{R}} \stackrel{\text{def}}{=} \mathbb{R} \cup \pm\infty \tag{2.158}$$

$$\mathfrak{R} \ni B \longmapsto \nu(B) \in \overline{\mathbb{R}}, \tag{2.159}$$

then ν is called a measure, *if ν satisfies for all $B \in \mathfrak{R}$ not only the conditions*

$$\nu(B) \geq 0, \quad \nu(\emptyset) = 0, \tag{2.160}$$

but also the σ-additivity, *so that the following holds to $B_1, B_2, \ldots \in \widetilde{\mathfrak{R}}$ with $B_j \cap B_k = \emptyset, (j \neq k)$ and $\cup_{j=1}^{\infty} B_j \in \widetilde{\mathfrak{R}}$:*

$$\nu\left(\bigcup_{j=1}^{\infty} B_j\right) = \sum_{j=1}^{\infty} \nu(B_j), \tag{2.161}$$

see Figure 2.20.

A first example for a simple, yet no less important measure, defined on a general set, is the *Dirac measure*. Apart of its important role in probability theory, the Dirac measure can also be used as a possibility to define the concept of the *Dirac δ-distribution*, which we use to describe physical phenomena such as ideal reflection and ideal refraction of particles at surfaces in an exact mathematical way.

EXAMPLE 2.27 (The Dirac Measure δ_x) *Let us assume a base set $\mathcal{R}$ and the σ-algebra $\mathfrak{R}$ generated on $\mathcal{R}$ are given, then the* Dirac measure *δ_x is defined for $x \in \mathcal{R}$ and $B \in \mathfrak{R}$ as follows:*

σ-algebra (828)

$$\delta_x(B) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{otherwise.} \end{cases} \tag{2.162}$$

As is easily seen, the set function δ_x satisfies the above conditions required to a measure, since it obviously applies not only that $\delta_x(B) \geq 0, \forall B \in \mathfrak{R}$ and $\delta_x(\emptyset) = 0$,

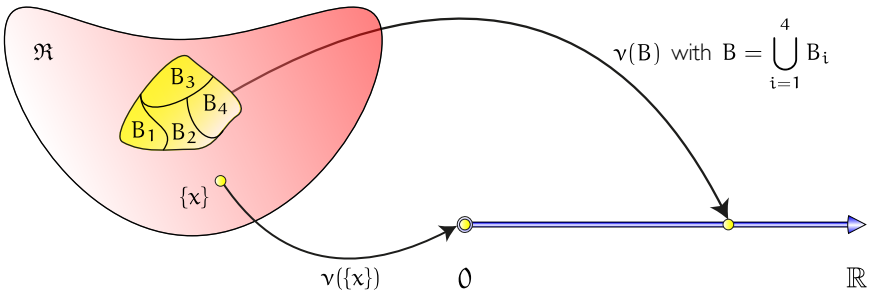


FIGURE 2.20: THE CONCEPT OF THE MEASURE, DEFINED AS A SET FUNCTION ON A σ-ALGEBRA. Defined on the σ-algebra $\mathfrak{R}$, a measure ν assigns a real number to any subset of $\mathfrak{R}$.

but also $\delta_x(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \delta_x(B_i)$, *as* x *can only be an element of one of the disjoint sets* $B_i, i \geq 1$.

Since it satisfies the normalization property $\delta_x(\mathcal{R}) = 1$, *the Dirac measure is a probability measure, which in terms of probability can be interpreted as the almost sure outcome* x *in the* sample space $\mathcal{R}$.

Probability Measure (80)

PROPERTIES OF MEASURES. Now a measure given over a σ-algebra possesses a number of characteristic features. Additionally, it serves as the basis of many concepts required in measure theory encountered in particular in the definition of the *measure* and the *measurable space*.

Let us suppose a σ-algebra $\mathfrak{R}$ is given over the base set $\mathcal{R}$. Then, the tupel $(\mathcal{R}, \mathfrak{R})$ is called a *measurable space* and the subsets B, with $B \in \mathfrak{R}$, are called *measurable sets* or $\mathfrak{R}$-*measurable* for short. If we equip the measurable space $(\mathcal{R}, \mathfrak{R})$ with a measure ν constructed on the σ-algebra $\mathfrak{R}$, then the triple $(\mathcal{R}, \mathfrak{R}, \nu)$ is called a *measure space*. If in particular $\nu(\mathcal{R}) = 1$ holds, we denote the measure ν as a *probability measure* or a *probability distribution* and the measure space $(\mathcal{R}, \mathfrak{R}, \nu)$ is called a *probability space* in probability theory. The measure ν is termed as a *finite measure* if $\forall B \in \mathfrak{R}$ it holds: $\nu(B) < \infty$. In the case that ν takes also infinite values and a non-descending sequence $B_1 \subseteq B_2 \subseteq B_3 \ldots \in \mathcal{R}$ exists with $\nu(B_i) < \infty$, $\cup_{i=1}^{\infty} B_i = \mathcal{R}$, we speak of a σ-*finite measure*. In particular, a set $B \in \mathfrak{R}$ with $\nu(B) = 0$ is called a *null set*. In such a case ν is termed *complete* when every subset of a null set is measurable, thus, if it is an element of $\mathfrak{R}$. Finally, the measure ν is denoted as *absolutely continuous* with respect to the measure ν', denoted as $\nu \ll \nu'$, if ν' is a measure on $\mathfrak{R}$ and for every null set $B \in \mathfrak{R}$ with $\nu'(B) = 0$ it holds: $\nu(B) = 0$.

σ-algebra A.1
Section 2.2.4
Section 2.4.1
Section 2.4.2

All hitherto presented measures were based on σ-algebras generated over uncountable base sets. However, a category of measures of great importance for the concerns of probability theory may also be defined over σ-algebras based only on finite or countably infinite sets: the class of *discrete measures*.

Uncountable Set (827)
Countable Set (827)

DEFINITION 2.21 (Discrete Measure) *Let $(\mathbf{x}_n)_{n\in\mathbb{N}}$ be a fixed sequence of distinct points of a base set $\mathcal{R}$ and $(\nu_n)_{n\in\mathbb{N}}$ an associated sequence of non-negative numbers. Then, the real-valued set function ν, defined by:*

Discrete Probability Space (163)
Set Function (837)

$$\nu(B) \stackrel{\text{def}}{=} \sum_{\{j|x_j\in B\}} \nu_j, \tag{2.163}$$

where $B \in \mathfrak{P}(\mathcal{R})$ and $\nu(\emptyset) \stackrel{\text{def}}{=} 0$, is denoted as a discrete measure *on $\mathfrak{P}(\mathcal{R})$.*

$\mathfrak{P}(\mathcal{R})$ (828)

EXAMPLE 2.28 (The Counting Measure #) *Let us define a set function # on the σ-algebra $\mathfrak{P}(\mathbb{N}_0)$, thus the set of all subsets of $\mathbb{N}_0$, by:*

$$\#(B) \stackrel{\text{def}}{=} \begin{cases} n & \text{if } |B| = n \\ \infty & \text{if } |B| = \infty, \end{cases} \tag{2.164}$$

where $B \in \mathfrak{P}(\mathbb{N}_0)$. Then, the sequences $(\mathbf{x}_n)_{n\in\mathbb{N}_0}$ with $x_n \in \mathbb{N}_0$ and $(\nu_n)_{n\in\mathbb{N}_0} = n$ clearly satisfies the requirements to a measure. Obviously, # represents a discrete and, in particular, a σ-finite measure: the counting measure *on $\mathbb{N}_0$. The counting measure describes the size of a set by the number of its elements.*

The concept of discrepancy*, a measure for the derivation of a set of points from an ideal uniform distribution, is based on the counting measure. It plays an important role in developing and analyzing quasi-Monte Carlo algorithms for solving integrals and integral equations.*

Discrepancy (621)
Uniform Distribution (180)
Chapter 7

From our introductory chapter, we know that the stationary vacuum light transport equation describes the total amount of light, that comes from everywhere at a surface point and is reflected along a particular viewing direction. Mathematically this means that we have to integrate a function over the entire unit sphere, one of the hemispheres, or all existing surfaces in the scene to be rendered. For that purpose, we need measures, which must be defined on product spaces constructed over S^2, $\mathcal{H}^2_+$, $\mathcal{H}^2_-$, as well as $\mathbb{R}^2$. The measure theoretical concept allowing us to integrate functions defined on such complex spaces is the concept of the *product measure*.

SLTEV (398)

DEFINITION 2.22 (Product Measure and Product Measure Space) *Let $(\mathcal{R}', \mathfrak{R}', \nu')$ and $(\mathcal{R}'', \mathfrak{R}'', \nu'')$ be two measure spaces. Furthermore we assume that $\mathcal{R} = \mathcal{R}' \times \mathcal{R}''$ and $\mathfrak{R}$ corresponds to the σ-algebra $\mathfrak{R}' \times \mathfrak{R}''$ generated over the Cartesian product*

Solid Angle Measures (84)
Measure Space (80)
σ-algebra (828)

$$\left\{ B' \times B'' \middle| B' \in \mathfrak{R}', B'' \in \mathfrak{R}'' \right\}, \tag{2.165}$$

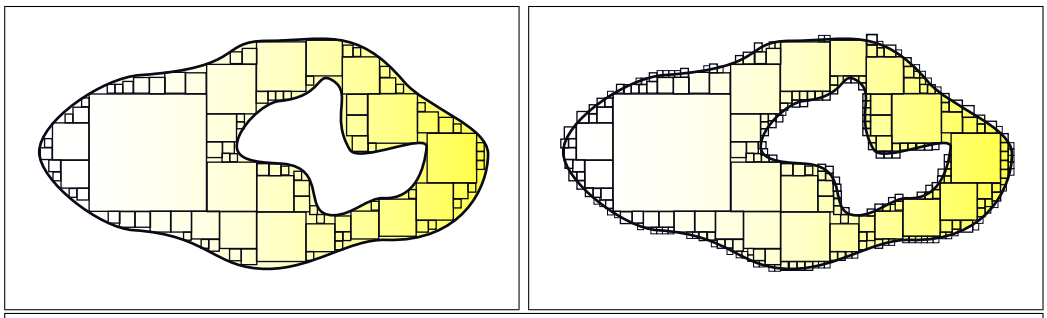


FIGURE 2.21: THE CONSTRUCTION OF THE LEBESGUE AREA MEASURE. Left, the inner Lebesgue measure $\mu_*^2(\mathbf{B})$ defined as the supremum over all subsets $\bigcup_{i=1}^{n}[\alpha_i,\beta_i)^2$ of the set $\mathbf{B}\subset\mathbb{R}^2$. On the right-hand side, the outer Lebesgue measure $\mu^{*2}(\mathbf{B})$ defined as the infimum over all given covers $\bigcup_{i=1}^{n}[\alpha_i,\beta_i)^2$ of $\mathbf{B}$.

then the product measure ν *is defined as:*

$$\nu(B'\times B'') \stackrel{\text{def}}{=} \nu'(B')\,\nu''(B'') \tag{2.166}$$

and the tupel $(\mathcal{R},\mathfrak{R},\nu)$ *is called the* product measure space.

EXAMPLE 2.29 (The Lebesgue Area Measure μ^2) *Due to Definition 2.22 the Lebesgue area measure* μ^2 *can now be defined via the common Lebesgue measure from Theorem 2.5 by:*

$$\mu^2(B) = \mu(B')\,\mu(B'') \tag{2.167}$$

Lebesgue Measurable Set (75) *where* B' *and* B'' *are Lebesgue measurable sets of* $\mathbb{R}$ *and* $B = B'\times B''$.

$\mathfrak{B}(\mathbb{R}^n)$ (865) EXAMPLE 2.30 (The Lebesgue Measure of a Bounded Set on $\mathbb{R}^n$) *Let* $\mathfrak{B}(\mathbb{R}^n)$ *be the Borel* σ*-algebra generated by all half-open intervals of type* $[\mathbf{a},\mathbf{b})\subset\mathbb{R}^n, n\geq 3$. *If we define*

Set Function (837) *a real-valued, non-negative, and* σ*-additive set function* μ^n *on* $\mathfrak{B}(\mathbb{R}^n)$ *by:*

$$\mu^n(A) \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} \text{vol}(\mathbf{I}_i) \tag{2.168}$$

for $A\subset\bigcup_{i=1}^{\infty}\mathbf{I}_i$, *where* $\text{vol}(\mathbf{I}) \stackrel{\text{def}}{=} \Pi_{i=1}^{n}(b_i-a_i)$ *is called the* volume *of the half-open interval* $\mathbf{I}=[\mathbf{a},\mathbf{b})=[a_1,b_1)\times\ldots\times[a_n,b_n)$, *then* μ^n *satisfies the conditions required for a measure.* μ^n *is denoted as the* Lebesgue measure *on* $\mathbb{R}^n$. *For an illustration of the case* $n=2$, *see Figure 2.21.*

BRDF (320) SOLID ANGLES. In Section 4.2.2, we will introduce the concept of the BRDF as a function that measures how light reflects off a surface when viewed under various viewing positions.

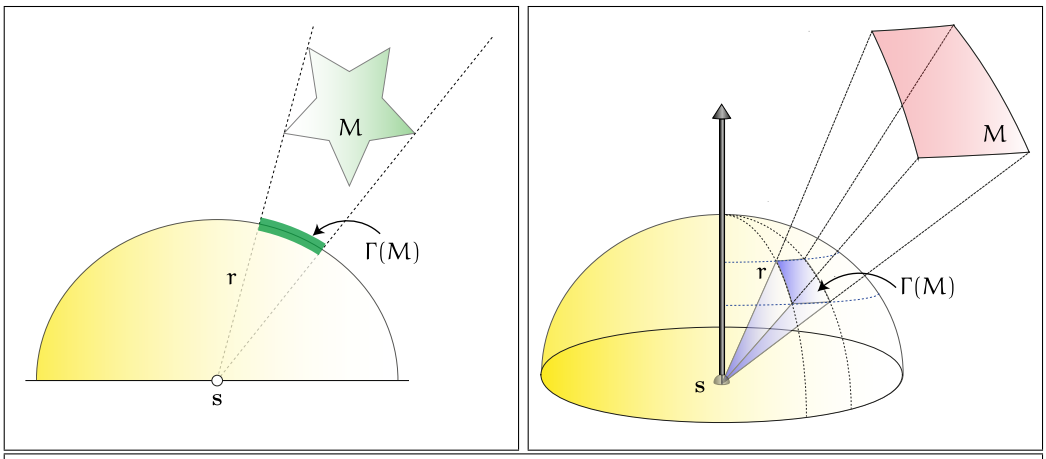


FIGURE 2.22: THE DEFINITION OF ANGLE AND SOLID ANGLE. The angle subtended by a 2-dimensional shape in the plane is defined as the length of its central projection onto the unit circle around the center $\mathbf{s}$. Equivalently, the solid angle subtended by a 3-dimensional shape in $\mathbb{R}^3$ is defined as the area of its central projection onto the unit sphere around its midpoint $\mathbf{s}$.

For that, we must have a good understanding of how much light arrives at or leaves a surface patch from a particular direction. Now, in *radiometry* light arriving or leaving a surface is measured in terms of *flux* through an area, that is, it makes no sense to consider light with respect a single direction, instead we should speak of light arriving at or passing a surface through a small region of directions. This then allows to determine the amount of light incident or exitant at a small surface element by taking into account the amount of light passing through a cross-sectional area surrounding a direction, the so-called *solid angle*.

Chapter 3 Radiant Flux (249)

DEFINITION 2.23 (The Construct of Solid Angle) *The* solid angle $\Gamma(\mathrm{M})$ *subtended by an object* M *viewed from point* $\mathbf{s} \in \mathbb{R}^3$ *is the radial projection* $\mathrm{M}^{\perp}$ *of* M *onto a sphere* $\mathrm{S}^2_{\mathrm{r}}(\mathbf{s})$ *with radius* r *centered at* $\mathbf{s}$*, see Figure 2.22. The size of the solid angle is the ratio of the area of* $\mathrm{M}^{\perp}$ *to the squared radius of the sphere, that is,*

$$\Gamma(\mathrm{M}) \stackrel{\mathrm{def}}{=} \frac{\mu^2(\mathrm{M}^{\perp})}{\mathrm{r}^2}. \tag{2.169}$$

Due to this definition, the solid angle $\Gamma(\mathrm{M})$ subtended by an object M can be interpreted as the continuous set of hit points of rays starting at the center of a sphere of radius r with the surface of the sphere.

EXAMPLE 2.31 (The Solid Angle Subtended by a Small Surface Patch) *Let us consider a small surface patch* A *whose normal at point* $\mathbf{s}'$ *is given by* $\mathbf{N}(\mathbf{s}')$ *then it holds for*

the solid angle subtended by A:

$$\Gamma(A) \stackrel{def}{=} \frac{\mu^2(A^\perp)}{r^2} \tag{2.170}$$

$$= \frac{\mu^2(A)\langle \omega_i, \mathbf{N}(\mathbf{s}')\rangle}{r^2} \tag{2.171}$$

$$= \frac{\mu^2(A)|\cos\omega_i|}{r^2}, \tag{2.172}$$

where ω_i is a direction starting at the center of a sphere of radius r, passing through point $\mathbf{s}'$ on A and $\cos\omega_i$ is the cosine of the angle between direction ω_i and the patch normal $\mathbf{N}$.

REMARK 2.24 *Since a solid angle is defined via an object or a 2-dimensional surface patch and a point in space, we have also to specify both, the object or the considered surface patch and the point whenever we will use the concept of the solid angle in words. So, we will often speak of the solid angle subtended by a surface patch, the lens of a virtual camera, or a pixel of the image plane as seen from a point in space or at an object surface.*

REMARK 2.25 *Evidently, our definition of the solid angle is an extension of the angle in two dimensions, where the angle subtended by an object is defined as the arc length of the central projection of the object onto the unit circle. Although solid angles are dimensionless, they are expressed in* squared radians, *or briefly* steradians, sr, *as both, the width and the length of the rectangular patch are measured in radians. Thus, the solid angle covered by an entire hemisphere is 2π sr, that covered by the unit sphere is 4π sr.*

To compute the solid angle subtended by a 2-dimensional surface or an object M in $\mathbb{R}^3$, we must first project M radially onto the hemisphere or the unit sphere, and then compute the solid angle of the projection of M. Thus, it should be clear, that the solid angle subtended by two different objects in shape can be the same, see Figure 2.23.

THE SOLID ANGLE MEASURES σ AND $\sigma^\perp$. As already mentioned above, it is convenient to integrate functions that measure light arriving at or leaving a surface point over areas surrounding a direction, that is, over regions on spheres, and here in particular the unit sphere or the lower and upper hemisphere. Thus, for integrating functions e.g. over the unit sphere, we are interested in a measure defined on $\mathfrak{B}(S^2)$, i.e. the Borel σ-algebra constructed over S^2.

$\mathfrak{B}(\cdot)$ (865) Measure (79)

For that purpose, let $\mathfrak{B}(S^2)$ be the Borel σ-algebra defined on the unit sphere. Now, our goal is to construct a measure σ on $\mathfrak{B}(S^2)$, but how should we define this measure? Now, let us consider the rectangle $\mathbf{B} = [0,\pi] \times [0,2\pi) \subset \mathbb{R}^2$ and the mapping

$$\mathbf{T} : \mathbb{R}^2 \to \mathbb{R}^3 \tag{2.173}$$

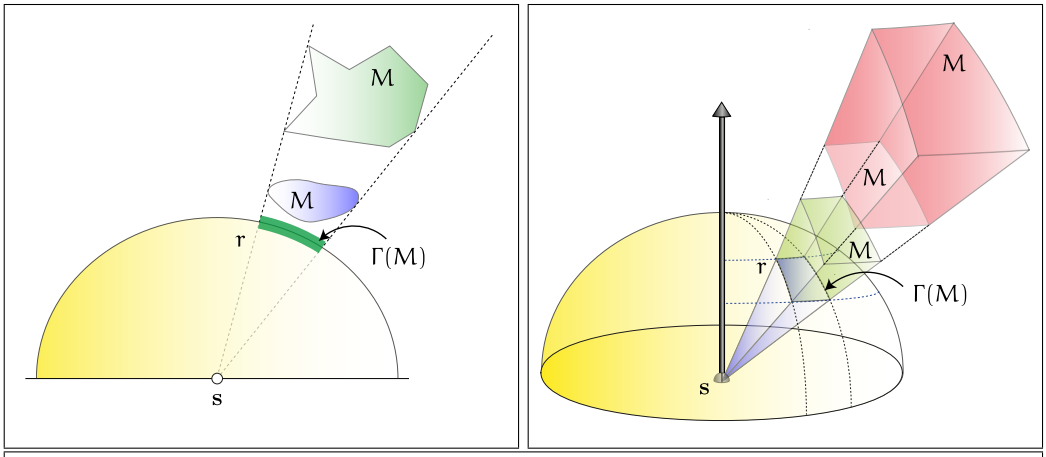


FIGURE 2.23: ANGLE AND SOLID ANGLE OF TWO DIFFERENT OBJECTS. Two different objects can occupy the same angle respectively the same solid angle.

with

$$\mathbf{T}(\theta,\phi) = \begin{pmatrix} x(\theta,\phi) \\ y(\theta,\phi) \\ z(\theta,\phi) \end{pmatrix} = \begin{pmatrix} r\sin\theta\cos\phi \\ r\sin\theta\sin\phi \\ r\cos\theta \end{pmatrix} \tag{2.174}$$

and $r > 0, r \in \mathbb{R}$, which assigns any point $(\theta,\phi) \in [0,\pi] \times [0,2\pi)$ a point on the sphere with radius r around the origin $\mathbf{0}$. Obviously, $\mathbf{T}$ maps points from $[0,\pi] \times [0,2\pi)$ onto a 2-dimensional surface over the unit sphere in $\mathbb{R}^3$.

Let us now consider an open cover $\cup_{i=1}^{\infty}\mathbf{I}_i$ of $\mathbf{B}$ by open intervals $\mathbf{I}_i = [\theta_i,\theta_{i+1}) \times [\phi_i,\phi_{i+1})$. The image of the rectangle $\mathbf{I}_i = [\theta_i,\theta_{i+1}) \times [\phi_i,\phi_{i+1})$ on the unit sphere is then given via the region enclosed by the four curve segments $c(\theta_i,\phi)$, $c(\theta_{i+1},\phi)$ as well as $c(\theta,\phi_i)$ and $c(\theta,\phi_{i+1})$, see Figure 2.24. This region can be approximated by a parallelogram $\mathbf{T}(\mathbf{I}_i)$ attached at point $(\theta_i,\phi_i,\mathbf{T}(\theta_i,\phi_i))^T$. Obviously, this parallelogram lies in the tangent plane

Open Cover (865)

$$\mathbf{t}(\theta,\phi) = \mathbf{T}(\theta_i,\phi_i) + \frac{\partial\mathbf{T}(\theta_i,\phi_i)}{\partial\theta}(\theta-\theta_i) + \frac{\partial\mathbf{T}(\theta_i,\phi_i)}{\partial\phi}(\phi-\phi_i) \tag{2.175}$$

where

$$\frac{\partial\mathbf{T}(\theta_i,\phi_i)}{\partial\theta} = \begin{pmatrix} \frac{\partial x(\theta_i,\phi_i)}{\partial\theta} \\ \frac{\partial y(\theta_i,\phi_i)}{\partial\theta} \\ \frac{\partial z(\theta_i,\phi_i)}{\partial\theta} \end{pmatrix} = \begin{pmatrix} r\cos\theta_i\cos\phi \\ r\cos\theta_i\sin\phi_i \\ -r\sin\theta_i \end{pmatrix} \tag{2.176}$$

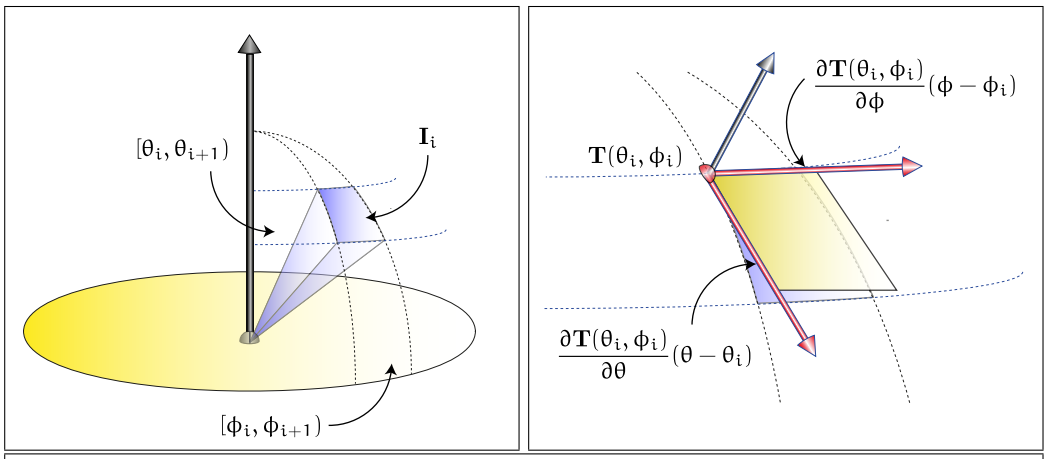


FIGURE 2.24: APPROXIMATING A REGION ON A SPHERE BY A PARALLELOGRAM. On the left side, a region bounded by the spherical interval $[\theta_i, \theta_{i+1})$ and $[\phi_i, \phi_{i+1})$. The area of this region can be a approximateted by the area of the parallelogram at point $\mathbf{T}(\theta_i, \phi_i)$ spanned by the vectors $\frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial \theta}(\theta - \theta_i$ and $\frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial \phi}(\phi - \phi_i)$.

and

$$\frac{\partial \mathbf{T}(\theta_i, \phi_i)}{\partial \phi} = \begin{pmatrix} \frac{\partial x(\theta_i,\phi_i)}{\partial \phi} \\ \frac{\partial y(\theta_i,\phi_i)}{\partial \phi} \\ \frac{\partial z(\theta_i,\phi_i)}{\partial \phi} \end{pmatrix} = \begin{pmatrix} -r \sin\theta_i \sin\phi_i \\ r \sin\theta_i \cos\phi_i \\ 0 \end{pmatrix} \tag{2.177}$$

are the tangent vectors at point $(\theta_i, \phi_i, \mathbf{T}(\theta_i, \phi_i)^\top$ in directions θ and ϕ, see Figure 2.24.

From Figure 2.24 it should also be clear, that the parallelogram $\mathbf{T}(\mathbf{I}_i)$ is spanned by vector $\frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial \theta}(\theta - \theta_i)$ in direction θ and by $\frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial \phi}(\phi - \phi_i)$ in direction ϕ. That is, due to Relation (A.54) the area of $\mathbf{T}(\mathbf{I}_i)$ is given by the norm of the cross product of the above spanning vectors, thus,

$$\left\| \frac{\partial \mathbf{T}(\theta_i, \phi_i)}{\partial \theta}(\theta - \theta_i) \times \frac{\partial \mathbf{T}(\theta_i, \phi_i)}{\partial \phi}(\phi - \phi_i) \right\|_2 . \tag{2.178}$$

We now define the solid angle measure of the parallelogram $\mathbf{T}(\mathbf{I}_i)$ via its area, that

is,

$$\begin{aligned}
\sigma(\mathbf{T}(\mathbf{I}_i)) &= \left\| \frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial\theta}(\theta-\theta_i) \times \frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial\phi}(\phi-\phi_i) \right\|_2 && (2.179)\\
&= \left\| \frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial\theta} \times \frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial\phi} \right\|_2 \mu([\theta-\theta_i))\mu([\phi-\phi_i)) && (2.180)\\
&= \left\| \frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial\theta} \times \frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial\phi} \right\|_2 \mu^2(\mathbf{I}_i) && (2.181)\\
&= \left\| \begin{pmatrix} r^2\sin^2\theta_i\cos\phi_i \\ r^2\sin^2\theta_i\sin\phi_i \\ r^2\sin\theta_i\cos\theta_i \end{pmatrix} \right\|_2 \mu^2(\mathbf{I}_i) && (2.182)\\
&= r^2\sin\theta_i\,\mu^2(\mathbf{I}_i). && (2.183)
\end{aligned}$$

This construction then implies the following definition of the *solid angle measure*:

DEFINITION 2.24 (The Solid Angle Measure σ and the Differential Solid Angle $d\sigma$) *Let $\mathfrak{B}(S^2)$ be the Borel σ-algebra constructed over the unit sphere. The* solid angle measure *σ is a function defined on $\mathfrak{B}(S^2)$, which maps any set A from $\mathfrak{B}(S^2)$ to a real number, that is,*

$\mathfrak{B}(\cdot)$ (865)
Set Function (837)

$$\sigma : \mathfrak{B}(S^2) \longrightarrow \mathbb{R}$$

with

$$A \mapsto \sigma(A) \quad \stackrel{def}{=} \quad \mu^2\left(\cup_{i=1}^{\infty} \sin\theta_i \mathbf{I}_i\right) = \sum_{i=1}^{\infty} \sin\theta_i\,\mu^2(\mathbf{I}_i), \tag{2.184}$$

where A is the image of the open cover $\cup_{i=1}^{\infty}\mathbf{I}_i \subseteq [0,\pi]\times[0,2\pi)$ under the mapping $\mathbf{T}$ from Equation (2.173).

Open Cover (865)

Considering only differential quantities, then the differential solid angle *$d\sigma$ is given by:*

$$d\sigma(\omega) \tag{2.185}$$

or equivalently:

$$d\sigma(\theta,\phi) = \sin\theta\, d\mu(\theta) d\mu(\phi). \tag{2.186}$$

Another important measure, needed to define important radiometric quantities, is the *projected solid angle measure*, that is, the area of the projection of the solid angle measure of a set $A \in \mathfrak{B}(S^2)$ on a disk with radius r, see Figure 2.25.

Chapter 3

The projected solid angle measure can then be defined via the area of the parallelogram spanned by the orthogonal projection of the vectors $\frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial\theta}(\theta-\theta_i)$ in direction θ

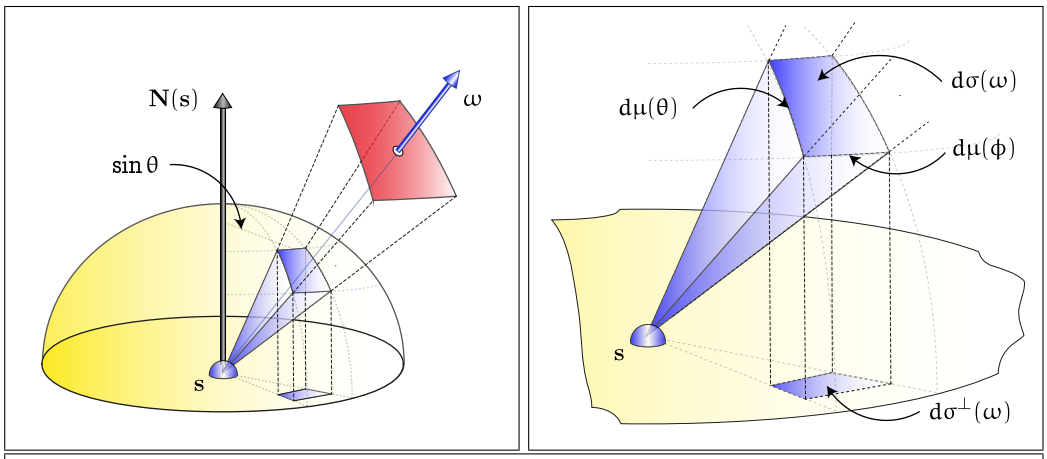


FIGURE 2.25: GEOMETRY FOR CONSTRUCTING THE PROJECTED SOLID ANGLE MEASURE.

and $\frac{\partial \mathbf{T}(\theta_i,\phi_i)}{\partial \phi}(\phi - \phi_i)$ in direction ϕ. Due to the projection a factor of $|\cos\theta_i|$ is attached at the above spanning vector in direction θ. This then leads to:

$$\begin{aligned}
\sigma^{\perp}(\mathbf{T}(\mathbf{I}_i)) &= \left\| \cos\theta_i \begin{pmatrix} r^2 \sin^2\theta_i \cos\phi_i \\ r^2 \sin^2\theta_i \sin\phi_i \\ r^2 \sin\theta_i \cos\theta_i \end{pmatrix} \right\|_2 \mu^2(\mathbf{I}_i) && (2.187)\\
&= |\cos\theta_i|\ \sigma(\mathbf{T}(\mathbf{I}_i)) && (2.188)\\
&= r^2 \sin\theta_i\, |\cos\theta_i|\ \mu^2(\mathbf{I}_i). && (2.189)
\end{aligned}$$

This implies the following definition of the *projected solid angle measure*:

DEFINITION 2.25 (The Projected Solid Angle Measure $\sigma^{\perp}$ and the Differential Solid Angle $d\sigma^{\perp}$) *Let* $\mathfrak{B}(S^2)$ *be the Borel* σ*-algebra constructed over the unit sphere. The* projected solid angle measure $\sigma^{\perp}$ *is a function defined on* $\mathfrak{B}(S^2)$*, which maps any set* A *from* $\mathfrak{B}(S^2)$ *to a real number, that is,*

$\mathfrak{B}(\cdot)$ (865) Set Function (837)

$$\sigma^{\perp} : \mathfrak{B}(S^2) \longrightarrow \mathbb{R}$$

with

$$A \mapsto \sigma^{\perp}(A) \quad \stackrel{\text{def}}{=} \quad \mu^2(\cup_{i=1}^{\infty} \sin\theta_i\, |\cos\theta_i|\, \mathbf{I}_i) = \sum_{i=1}^{\infty} \sin\theta_i\, |\cos\theta_i|\ \mu^2(\mathbf{I}_i), \qquad (2.190)$$

Open Cover (865) *where* A *is the image of the open cover* $\cup_{i=1}^{\infty} \mathbf{I}_i \subseteq [0,\pi] \times [0,2\pi)$ *under the mapping* $\mathbf{T}$ *from Equation (2.173).*

Considering only differential quantities, then the differential projected solid angle $d\sigma^\perp$ *is given by:*

$$d\sigma^\perp(\omega) = |\cos\theta|\ d\sigma(\omega) \tag{2.191}$$

or equivalently:

$$d\sigma^\perp(\theta,\phi) = \sin\theta\,|\cos\theta|\ d\mu(\theta)d\mu(\phi). \tag{2.192}$$

REMARK 2.26 *To circumvent the problem of the orientation of the surface normal in our formulas—in some rendering systems the surface normal is always assumed to point outside the surface, others, like pbrt, do not assume that the surface normal lies on the same side as the incident direction ω_i—we have added an absolute value to the cosine term, hidden in the projected solid angle.*

Let us now show how these new concepts can be applied to problems from global illumination theory.

EXAMPLE 2.32 (Integration with Respect to the Projected Solid Angle Measure) *Based on the definition of the projected solid angle measure and the projected differential solid angle measure we have now a better understanding of integrating functions with respect to the solid angle measures.*

Let us consider the stationary light transport equation in a vacuum presented in Equation (4.390). It was introduced in the form that integration was done with respect to the projected solid angle measure, thus,

$$L_o(s,\omega_o) = L_e(s,\omega_o) + \int_{S^2(s)} f_s(s,\omega_i \to \omega_o) L_i(s,\omega_i)\ d\sigma^\perp_s(\omega_i). \tag{2.193}$$

Expressing a direction ω, due to Definition A.7, by (θ,ϕ) and applying the definition of the differential projected solid angle from Equation (2.25), then the integration in the SLTEV can be done with the help of the Lebesgue area measure μ^2 with respect to the variable θ_i and ϕ_i by:

$$\begin{aligned} L_o(s,\theta_o,\phi_o) &= L_e(s,\theta_o,\phi_o) + \\ &\int_{[0,\pi]\times[0,2\pi)} f_s(s,(\theta_i,\phi_i) \to (\theta_o,\phi_o)) L_i(s,\theta_i,\phi_i)\ \sin\theta_i\,|\cos\theta_i|\ d\mu(\theta_i)d\mu(\phi_i). \end{aligned} \tag{2.194}$$

TRANSFORMING THE SOLID ANGLE MEASURES TO THE LEBESGUE AREA MEASURE. In Section 9.1.2 we will see, that, in many cases more efficient rendering routines can be written if we could integrate the light transport equations over all visible surfaces instead over the unit sphere. Now, changing the integration domain is connected with a change in the integration

Lebesgue Area Measure (82)

measure followed by a variable transformation, resulting in a new representation of the light transport equations, the so-called *area formulations* also denoted as the 3-*point formulations* of the light transport equations.

∂𝒱 (41) σ⊥ (88) γ (47)

Changing the integration domain from S^2 to $\partial\mathcal{V} \times \partial\mathcal{V}$ requires replacing the projected solid angle measure $\sigma^\perp$ by the Lebesgue area measure μ^2. From Figure 2.26 it can be seen, that this can be done as follows: Via the ray-casting function γ we have to find the point on the closest visible surface along a ray starting at point $\mathbf{s}_i$ and pointing in direction ω_i^i, we call this point $\mathbf{s}_j$. Afterwards, we project the infinitesimal patch $d\mu^2(\mathbf{s}_j)$ onto the unit sphere resulting in the differential solid angle $d\sigma_{\mathbf{s}_i}(\omega_i^i)$ subtended by $d\mu^2(\mathbf{s}_j)$ as seen from $\mathbf{s}_i$. Expressed in terms of the differential surface patch $d\mu^2(\mathbf{s}_j)$ then it holds:

$$d\sigma_{\mathbf{s}_i}(\omega_i^i) = \frac{d\mu^2(\mathbf{s}_j)\left|\left\langle \omega_o^j, \mathbf{N}(\mathbf{s}_j)\right\rangle\right|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2} \tag{2.195}$$

$$= \frac{d\mu^2(\mathbf{s}_j)\left|\cos\theta_o^j\right|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2}, \tag{2.196}$$

where $\left\langle \omega_o^j, \mathbf{N}(\mathbf{s}_j)\right\rangle$ is the cosine-foreshortening due to the radial projection of $d\mu^2(\mathbf{s}_j)$ onto the unit sphere around $\mathbf{s}_i$ and the distance of the two points $\mathbf{s}_j$ and $\mathbf{s}_i$ in the denominator is used to account for the projected area fraction of the patch onto the unit sphere.

Using this result, then the differential projected solid angle measure can be expressed in terms of the Lebesgue area measure as

$$d\sigma_{\mathbf{s}_i}^\perp(\omega_i^i) \overset{(2.191)}{=} \left|\langle \omega_i^i, \mathbf{N}(\mathbf{s}_i)\rangle\right| d\sigma_s(\omega_i^i) \tag{2.197}$$

$$\overset{(2.196)}{=} \left|\langle \omega_i^i, \mathbf{N}(\mathbf{s}_i)\rangle\right| \frac{d\mu^2(\mathbf{s}_j)\left|\left\langle \omega_o^j, \mathbf{N}(\mathbf{s}_j)\right\rangle\right|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2} \tag{2.198}$$

$$= \frac{\left|\cos\theta_i^i \cos\theta_o^j\right|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2} d\mu^2(\mathbf{s}_j), \tag{2.199}$$

where θ_i^i and θ_o^j are the angles between the directions ω_i^i at $\mathbf{s}_i$ and ω_o^j at $\mathbf{s}_j$ and the surface normals at $\mathbf{s}_i$ and $\mathbf{s}^j$. Note: When deriving equivalent formulations of the light transport equation in free space in Section 4.4.2, we will use this technique over and over again.

EXAMPLE 2.33 (The Classical Differential-to-Differential-Area Form Factor) *In Chapter 10 when deriving the radiosity equation, we will encounter the concept of the* form factor*, as the proportion of light leaving a surface patch that is received by another patch. As we show in this and the next example, the concept of the form factor has a lot in common with our concept of the projected solid angle. Thus, Relation (2.199)* 𝒱 (45) *resembles—except of the missing terms* $\frac{1}{\pi}$ *and the visibility function* $\mathcal{V}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)$*—the*

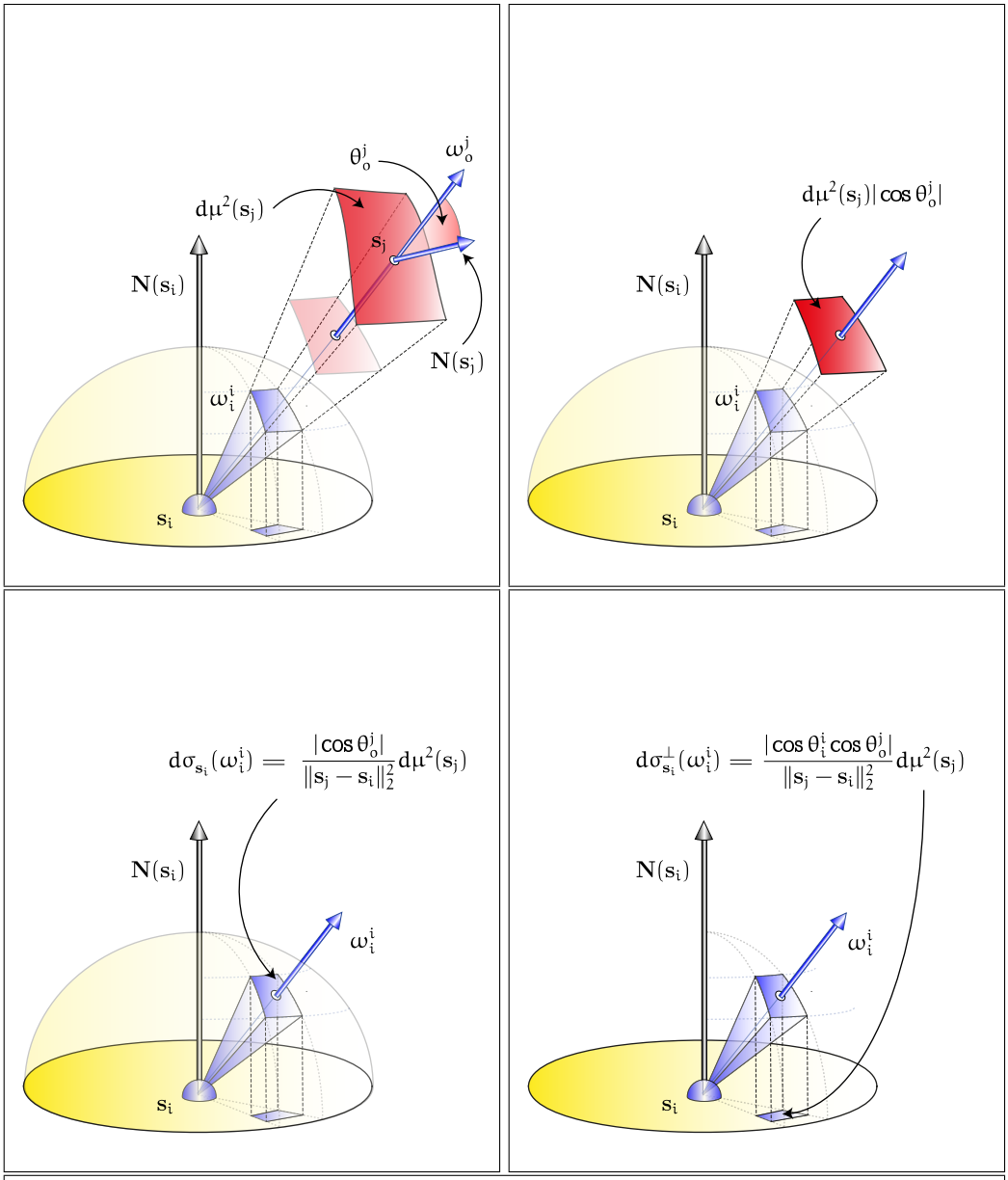


FIGURE 2.26: TRANSFORMING THE PROJECTED SOLID ANGLE MEASURE TO THE LEBESGUE AREA MEASURE. The transformation of the projected solid angle area measure into the Lebesgue area measure can be visualized in a three step procedure. The radial projection of the patch onto the sphere around s_i going through the surface point s_j, followed by a projection in direction ω_i^i onto the unit sphere, followed the orthogonal projection of this new generated patch onto the base of the unit sphere.

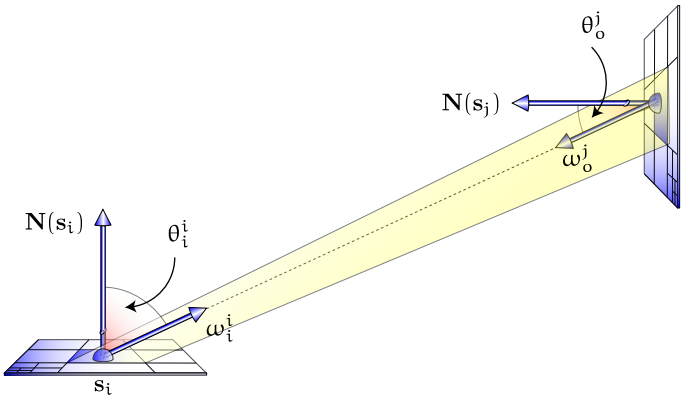


FIGURE 2.27: GEOMETRY FOR COMPUTING THE CLASSICAL DIFFERENTIAL-TO-DIFFERENTIAL FORM FACTOR.

so-called classical differential-to-differential-area form factor *between differential areas around the points* $\mathbf{s}_i$ *and* $\mathbf{s}_j$*, denoted as* $F_{\mathbf{s}_i\mathbf{s}_j}$*. It corresponds to the proportion of light leaving the differential area around* $\mathbf{s}_i$ *that is received by the differential area around point* $\mathbf{s}_j$ *and is defined as:*

$$F_{\mathbf{s}_i\mathbf{s}_j} = \frac{1}{\pi}\frac{\left|\cos\theta_i^i \cos\theta_o^j\right|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2}\mathcal{V}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\mu^2(\mathbf{s}_j), \tag{2.200}$$

see Figure 2.27. Except of the missing terms $\frac{1}{\pi}$ *and* $\mathcal{V}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)$*, Equation (2.200) is identically to Equation (2.199).*

Obviously, the fraction of light leaving the differential area around $\mathbf{s}_i$ *that arrives at the differential area around* $\mathbf{s}_j$ *is proportional to the differential solid angle subtended by the differential area around* $\mathbf{s}_j$ *as seen from* $\mathbf{s}_i$*. The form factor thus depends inversely on the square of the distance between the areas and on the cosines of the angles between the surface normals* $\mathbf{N}(\mathbf{s}_i)$ *and* $\mathbf{N}(\mathbf{s}_j)$ *as well as the in and outgoing directions* ω_i^i *and* ω_o^j*. We will discuss the concept of the classical form factor in more detail in Section 10.1.3.*

In the same context, let us also consider the following situation:

EXAMPLE 2.34 (The Nusselt Analog) *Let* $\mathcal{H}_+^2$ *be the upper hemisphere about surface point* $\mathbf{s}_i$*. Furthermore, let us assume that in our scene there exists a series of opaque surface patches* P_j *such as light sources, see Figure 2.28. Now, the solid angle of patch* P_j *as seen from the center of the hemisphere is defined as the covered surface*

of its radial projection, that is,

$$\Gamma(P_j) \overset{(2.196)}{=} \int_{P_j^\perp} d\sigma_{\mathbf{s}_i}(\omega_i) \tag{2.201}$$

$$= \int_{P_j} \frac{|\cos\theta_o^j|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2}\, d\mu^2(\mathbf{s}_j). \tag{2.202}$$

Due to Equation (2.199), the orthogonal projection of the solid angle subtended by P_j *as seen from* $\mathbf{s}_i$ *onto the base of the hemisphere can then be computed via*

$$\Gamma(P_j)^\perp = \int_{P_j^\perp} d\sigma^\perp_{\mathbf{s}_i}(\omega_i) \tag{2.203}$$

$$= \int_{P_j^\perp} \left|\cos\theta_i^i\right| \, d\sigma_{\mathbf{s}_i}(\omega_i) \tag{2.204}$$

$$= \int_{P_j} \frac{\left|\cos\theta_i^i \cos\theta_o^j\right|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2}\, d\mu^2(\mathbf{s}_j). \tag{2.205}$$

Equipped with the visibility function, $\mathcal{V}$*, the fraction of the base area covered by this projection is then defined as the* unoccluded differential-to-finite-area form factor, *denoted by* $F_{\mathbf{s}_i P_j}$*, i.e.* $\mathcal{V}$ (45)

$$F_{\mathbf{s}_i P_j} \overset{def}{=} \frac{1}{\pi} \int_{P_j^\perp} \left|\cos\theta_i^i\right| \, \mathcal{V}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\sigma_{\mathbf{s}_i}(\omega_i) \tag{2.206}$$

$$= \int_{P_j} \frac{\left|\cos\theta_i^i \cos\theta_o^j\right|}{\pi\|\mathbf{s}_j - \mathbf{s}_i\|_2^2}\, \mathcal{V}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\mu^2(\mathbf{s}_j). \tag{2.207}$$

We call this construction the Nusselt analog. *From the Nusselt analog follows, that two surface patches with the same solid angle share the same unoccluded differential-to-finite-area form factor, that is, unoccluded differential-to-finite-area form factors are only depending on their projection onto the upper hemisphere* $\mathcal{H}^2_+$.

The Nusselt analog can be considered as the basis for the hemicube form factor algorithm, *introduces in Section 10.1.3.2.2, where surface patches are projected onto the planar faces of a half cube instead onto the hemisphere.*

REMARK 2.27 (Direct Illumination) *The above measure transformation must also be applied when computing the direct illumination at a scene point via generating shadow rays. As we know, shadow rays are fired from a surface point in direction to the light sources for computing the light that directly illuminates the point of interest. Commonly such rays can be generated by choosing points on the unit sphere around a surface point and shooting a ray from the center of the sphere through this point*

Direct Illumination (617)

Shadow Ray (14)

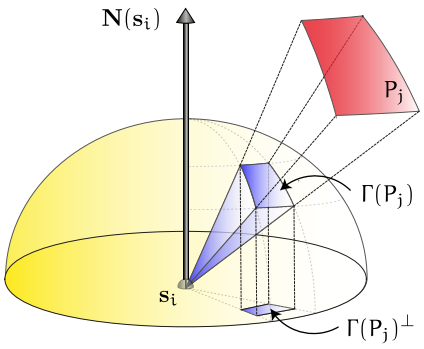


FIGURE 2.28: THE NUSSELT ANALOG. The form factor from the differential area around s_i to patch P_j is proportional to the area of the orthogonal projection of solid angle $\Gamma(P_j)$.

on the sphere. But if the light sources are small and far away, it should be clear, that the probability that such a ray hits a light source is very small. A better strategy to generate shadow rays would be to make use of the knowledge about the position, orientation, and the shape of the light sources. Such a strategy requires the above mentioned transformation of a spherical integral into a surface integral via the measure transformation from Equation (2.199).

THE THROUGHPUT MEASURES. We know, that in cases where we abstract from participating media, rays will start only at object surfaces, that is, at boundaries $\partial\mathcal{V}$ of the involved participating media. Based on the Lebesgue measure μ^n, $n = 2,3$ and the solid angle measures σ and $\sigma^\perp$, now the concept of the product measure implies the construction of measures $\zeta^\perp, \zeta^o$ and ζ on the Borel σ-algebras $\mathfrak{B}(\mathcal{R}^{\partial\mathcal{V}})$, $\mathfrak{B}(\mathcal{R}^{\mathcal{V}^o})$, and $\mathfrak{B}(\mathcal{R})$ generated over all open subsets of the ray spaces $\mathcal{R}^{\partial\mathcal{V}}$, $\mathcal{R}^{\mathcal{V}^o}$, as well as $\mathcal{R}$, the so-called *throughput measures*. They can be used to measure the light-carrying capacity of a ray in a vacuum and in participating media.

μ^n (82)
Solid Angle Measures (84)
Ray Spaces (44)

DEFINITION 2.26 (The Throughput Measures $\zeta^\perp, \zeta^o$, and ζ) *The* throughput measures ζ *and* $\zeta^\perp$ *are simply defined via the products of the Lebesgue measures* μ^2 *respectively* μ^3 *and the inner volumes* $\mathcal{V}^o$ *respectively the boundaries* $\partial\mathcal{V}$ *combined with the solid angle measures* σ *and* $\sigma^\perp$*, thus,*

$$\zeta^\perp(B_{\partial\mathcal{V}} \times B_{S^2}) \stackrel{def}{=} \mu^2(B_{\partial\mathcal{V}})\,\sigma^\perp(B_{S^2}) \tag{2.208}$$

$$\zeta^o(B_{\mathcal{V}^o} \times B_{S^2}) \stackrel{def}{=} \mu^3(B_{\mathcal{V}^o})\,\sigma(B_{S^2}) \tag{2.209}$$

with $B_{\partial\mathcal{V}} \times B_{S^2} \in \mathfrak{B}(\mathcal{R}^{\partial\mathcal{V}})$*, when considering ray functions defined in a vacuum, re-*

Chapter 5
μ^s (82)
$\partial\mathcal{V}, \mathcal{V}^o$ (41)
Solid Angle Measures (84)

spectively $B_{\mathcal{V}^o} \times B_{S^2} \in \mathfrak{B}(\mathcal{R}^{\mathcal{V}^o})$ *in the case where we consider ray functions in participating media. Based on the throughput measures, then the triples* $(\mathcal{R}^{\partial\mathcal{V}}, \mathfrak{B}(\mathcal{R}^{\partial\mathcal{V}}), \zeta^{\perp})$ *and* $(\mathcal{R}^{\mathcal{V}^o}, \mathfrak{B}(\mathcal{R}^{\mathcal{V}^o}), \zeta)$ *define measure spaces.* Measure Space (80)

Last but not least, we define the extended troughput measure, ζ, *via the measures* $\zeta^{\perp}$ *and* ζ^o *by:*

$$\zeta : \mathfrak{B}(\mathcal{R}) \to \overline{\mathbb{R}} \tag{2.210}$$

by

$$B \mapsto \zeta(B) \stackrel{def}{=} \begin{cases} \zeta^{\perp}(B) & \text{if } B \in \mathfrak{B}(\mathcal{R}^{\partial\mathcal{V}}) \\ \zeta^o(B) & \text{if } B \in \mathfrak{B}(\mathcal{R}^{\mathcal{V}^o}). \end{cases} \tag{2.211}$$

REMARK 2.28 *With the throughput measures defined on the ray spaces* $\mathcal{R}^\star$ *and the above constructed measures, now we have the means to derive and understand a number of constructs of great importance to us here, which are based on measure theoretical concepts and used for the analysis of rendering algorithms based on probabilistic methods. Thus, the significance of the solid angle measure and projected solid angle measure will become clear in the definition of the light transport equation and its adjoint counterpart, the importance equation. There, the throughput measures play a central role when we formulate the phenomena, involved in light transport, as integral operators on the function spaces* $\mathcal{L}(\mathcal{R}^\star)$. Ray Spaces (44) Chapter 5 Linear Integral Operator (61) $\mathcal{L}(\mathcal{R}^\star)$ (46)

Additionally, in the chapter where we present the approaches of the quasi-Monte Carlo methods, a new measure will be defined via the counting and the Lebesgue measure, namely the concept of the discrepancy *of a point set. This measure will be of great importance when we analyze quasi randomly chosen point sets to make statements about the quality of their distribution. Last but not least, with the* continuous path measure *we will present another important new measure. The continuous path measure will enable us to represent the global illumination equation not as an integral equation but as a simple integral. This integral, also called a* path integral, *must be solved over the* σ*-algebra of all paths generated in a scene to be rendered. We will do this via the probability theoretical model of the* Markov chain *and the construction of so-called* random walks. Chapter 7 Discrepancy (622) Section 7.2 Continuous Path Measure (461) Section 5.4 Section 2.4.7.2

REMARK 2.29 *From the view of measure theory, it will suffice to discuss for the analysis and construction of global illumination algorithms the most important* σ*-algebras in measure theory: the Borel* σ*-algebras, and here in particular the Borel* σ*-algebras constructed over the ray spaces* $\mathcal{R}^\star$. Borel σ-algebra (865)

2.2.3 MEASURABLE FUNCTIONS

In Appendix A.5 it is shown that there is a crucial difference in the approaches for computing the Riemann and the Lebesgue integral. The Riemann integral is based on the

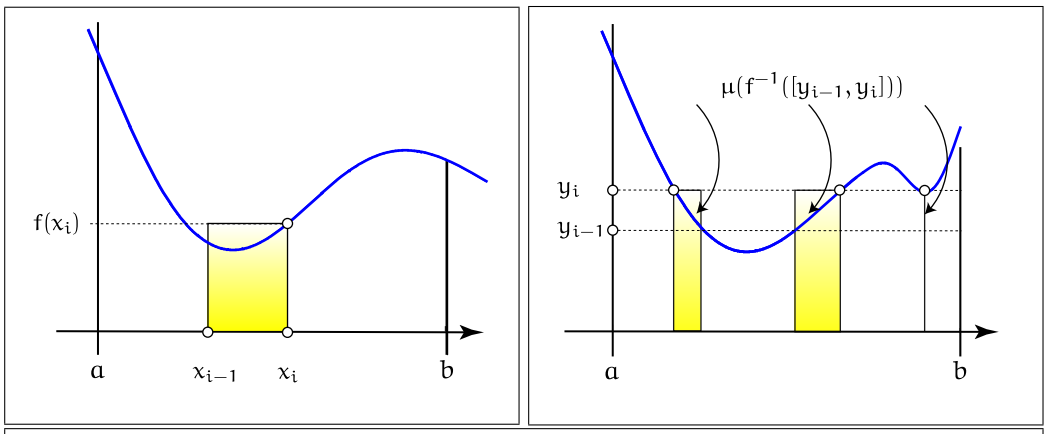


FIGURE 2.29: THE IDEA BEHIND THE RIEMANN AND THE LEBESGUE INTEGRAL. The Riemann integral is based on a decomposition of the integration domain $[a, b]$ via the construction of rectangles of area $\sup_{x \in I_i} f(x) \cdot (x_i - x_{i-1})$, while the Lebesgue integral is build over rectangles constructed via the pre-image of a decomposition of the range of f, i.e., $y_i \cdot \mu(f^{-1}([y_{i-1}, y_i]))$.

evaluation of sums of the form

$$\mathcal{I}_u \stackrel{def}{=} \sum_{i=1}^{n} \sup_{x \in I_i} f(x)(x_i - x_{i-1}) \tag{2.212}$$

or

$$\mathcal{I}_o \stackrel{def}{=} \sum_{i=1}^{n} \inf_{x \in I_i} f(x)(x_i - x_{i-1}) \tag{2.213}$$

for approximating the area between the graph of the function f and the real axis over the partition of the integration domain into a disjoint union $\bigcup_{i=1}^{n} I_i$ of small subintervals $I_i = [x_{i-1}, x_i]$, see Figure 2.29.

Henri Lebesgue recognized that in the Riemann approach the shape of the graph of f plays no role, where arbitrary partitions of the integration domain are constructed without any reference to the graph of the function f. This can thoroughly lead to bad approximation results. To utilize the information hidden in the shape of the graph of f, a better approach is to decompose the range of f into a disjoint union $\bigcup_{i=1}^{n} I_i$ of small intervals $I_i = [y_{i-1}, y_i]$.

The approximation of the area between the graph f and the real axis over the inte- Section A.5
gration domain via sums of the form

$$\mathcal{I} \stackrel{def}{=} \sum_{i=1}^{n} y_{i-1} \mu(f^{-1}([y_{i-1}, y_i])) \tag{2.214}$$

or

$$\mathcal{I} \stackrel{\text{def}}{=} \sum_{i=1}^{n} y_i \mu(f^{-1}([y_{i-1}, y_i])) \tag{2.215}$$

then requires knowledge about the nature of the pre-images of the intervals, $[y_{i-1}, y_i]$ which are commonly complex sets, and not necessarily intervals, see Figure 2.29.

Here, then we need the concept of the *measurable function*, which makes statements about the measurability of these sets. If all these sets are measurable, the measure μ can be applied to the pre-image of $[y_{i-1}, y_i]$ to evaluate the sum for approximating the Lebesgue integral of f. Thus, one can say that the Lebesgue integral is in some sense not a *blind, but an informed integration technique*, such as an *importance sampling strategy*. This is expressed in the following statement by Henri Lebesgue: Section 6.6.2

On peut dire encore qu'avec le procédé de Riemann ... on opéra it ... comme le ferait un comerçant sans méthode qui compterait pièces et billets au hasard de l'ordre où ils lui tomberait sous la main; tandis nous opérons le comerçant méthodique qui dit:

i) j'ai $m(E_1)$ *pièces de* 1 *couronne valent* $1 \cdot m(E_1)$,

ii) j'ai $m(E_2)$ *pièces de* 2 *couronnes valent* $2 \cdot m(E_2)$,

iii) j'ai $m(E_3)$ *pièces de* 5 *couronnes valent* $5 \cdot m(E_3)$,

etc., j'ai donc en tout: $S = 1 \cdot m(E_1) + 2 \cdot m(E_2) + 5 \cdot m(E_3) + \cdot$. *Les deux procèdés conduiront, certes, le commerçant au même résultat parce que, si riche q'il soit, il n'a qu'un nombre fini de billets à computer; mais pour nous, qui avons à additionner une infinité d'indivisibles, la différence enter les deuux façons de faire est capital*, [54, Elstrodt 1996].

Now, the concept of the measurable function is not only needed for the derivation of the Lebesgue integral, it also serves as the basis for the definition of the probabilistic theoretical construct of the *random variable*. Additionally, measurable functions, which are different at *not to many locations*, can be considered as equivalent functions. As we will see, this property leads to new and beautiful limit notions, which play a relevant role when considering sequences of measurable functions for the derivation of the Lebesgue integral in Section 2.2.4 and when introducing the *limit theorems* of probability theory in Section 2.4.6.

With the concept of the σ-algebra and the concept of a measure we have already studied two immensely important concepts of measure theory. But for the definition of a new integral notion, urgently required for integration and probability theory, we still need a structure-preserving mapping between measurable spaces: the above mentioned concept of the *measurable function*.

σ-algebra (828)
Measure (79)
Measurable Space (80)

DEFINITION 2.27 *Let* $(\mathcal{R}, \mathfrak{R})$ *and* $(\mathcal{R}', \mathfrak{R}')$ *be two measurable spaces. A function* f *given by:*

$$f : (\mathcal{R}, \mathfrak{R}) \longrightarrow (\mathcal{R}', \mathfrak{R}') \tag{2.216}$$

Measurable Space (80) *with*

$$\mathcal{R} \ni x \mapsto f(x) \in \mathcal{R}' \tag{2.217}$$

is called $\mathfrak{R}$-$\mathfrak{R}'$-measurable, *also briefly denoted as* measurable, *if the pre-image of an* $\mathfrak{R}'$*-measurable set is* $\mathfrak{R}$*-measurable, that is, the following must hold for all* $B \in \mathfrak{R}'$*:*

$$f^{-1}(B) = \{x \in \mathcal{R} \,|\, f(x) \in B\} \in \mathfrak{R}. \tag{2.218}$$

REMARK 2.30 (Borel or Lebesgue-measurable Functions) *A function* f *is termed* Borel-measurable *if it is* $\mathfrak{B}(\mathcal{R})$-$\mathfrak{B}(\mathcal{R}')$*-measurable. In the case of a real-valued Borel-measurable function* f*, we only have to check the measurability of the sets*

$$\{f < a\} \stackrel{def}{=} f^{-1}((-\infty, a)) \tag{2.219}$$

$$= \{x \in \mathbb{R} \,|\, f(x) < a\} \tag{2.220}$$

for any $a \in \mathbb{R}$*. Obviously, this holds since* $f^{-1}((-\infty, a))$ *can be written as set difference of the two measurable sets,* $\mathbb{R}$ *and* $f^{-1}([a, \infty))$*, namely:*

$$f^{-1}((-\infty, a)) = f^{-1}(\mathbb{R} \setminus [a, \infty)) \tag{2.221}$$

$$= \mathcal{R} \setminus f^{-1}([a, \infty)). \tag{2.222}$$

This statement on the measurability of a real-valued Borel-measurable function remains valid even if we replace the set $\{f < a\}$ *by* $\{f \leq a\}, \{f > a\}$*, or* $\{f \geq a\}$*.*

If the measurability of the function f *holds on the measurable spaces* $(\mathcal{R}, \mathfrak{M}(\mathcal{R}))$ *and* $(\mathcal{R}', \mathfrak{B}(\mathcal{R}))$*, then* f *is called* Lebesgue-measurable.

As the notion of the measurable function is characterized in terms of measurable sets, measurable sets and measurable functions are closely related. For an illustration of the concept of the measurable function, see Figure 2.30.

EXAMPLE 2.35 (Two Simple Borel-measurable Functions) *Let* $f : \mathbb{R} \longrightarrow \mathbb{R}$ *with* $x \mapsto f(x) = c, c \in \mathbb{R}$ *be a constant, real-valued function defined on the measurable space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$*. Then the following clearly holds:*

Measurable Space (80)

$$\{f < a\} = \begin{cases} \mathbb{R} & \text{if } a > c \\ \emptyset & \text{otherwise,} \end{cases} \tag{2.223}$$

which implies that the measurability of f *follows from the measurability of* $\mathbb{R}$ *and* $\emptyset$*,* Measurable Set (80) *which are* $\mathfrak{B}(\mathbb{R})$*-measurable sets.*

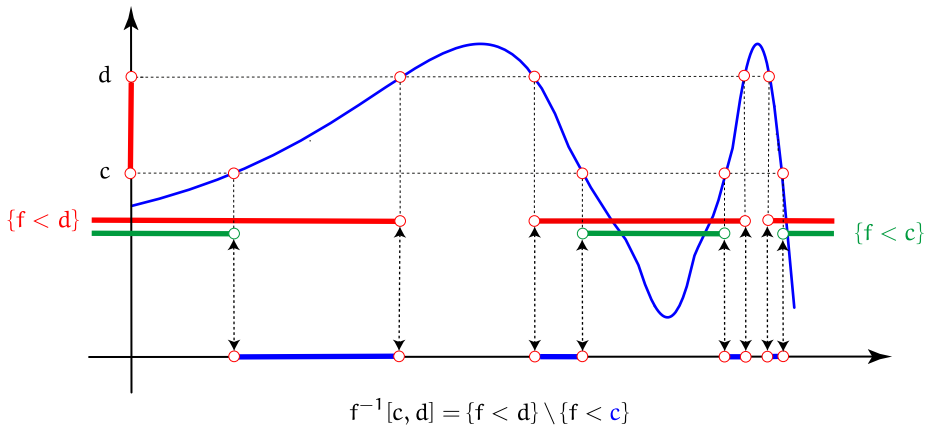


FIGURE 2.30: MEASURABILITY OF REAL-VALUED FUNCTIONS. For the measurability of a real-valued function defined on the space $\mathcal{R}$ it is sufficient to show that the pre-image $f^{-1}([c,d))$ of any half-open interval is a measurable set. Whether a function is measurable depends on the measure on $\mathfrak{R}$, and, in particular, it only depends on the σ-algebra of measurable sets in $\mathfrak{R}$. In fact, practically any function that can be described is measurable.

Considering a further example, the characteristic function χ_B,

$$\chi_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{otherwise,} \end{cases} \tag{2.224}$$

with $B \in \mathbb{R}$. *The measurability of* χ_B *follows from the measurability of the sets* $\mathbb{R}, B,$ *and* $\emptyset$ *as subsets of* $\mathfrak{B}(\mathbb{R})$ *since it holds:* χ_B (839)

$$\{\chi_B < a\} = \begin{cases} \mathbb{R} & \text{if } a > 1 \\ \overline{B} & \text{if } 0 < a \leq 1 \\ \emptyset & \text{if } a \leq 0, \end{cases} \tag{2.225}$$

see Figure 2.31. Which means that the two-valued function χ_B *is measurable if the base set* $\mathbb{R}$, *the complement of* B, *and the empty-set are all measurable sets.* Measurable Set (80)

REMARK 2.31 (The Visibility Function, a Measurable Function) *As the most important consequence of the measurability of the characteristic function and the measurability of simple functions defined over measurable sets it immediately follows that the visibility function* $\mathcal{V}$ *given by:* Measurable Set (80) $\mathcal{V}$ (45)

$$\mathcal{V}(\mathbf{x}' \leftrightarrow \mathbf{x}) \equiv \chi_B(\mathbf{x}, \mathbf{x}') \tag{2.226}$$

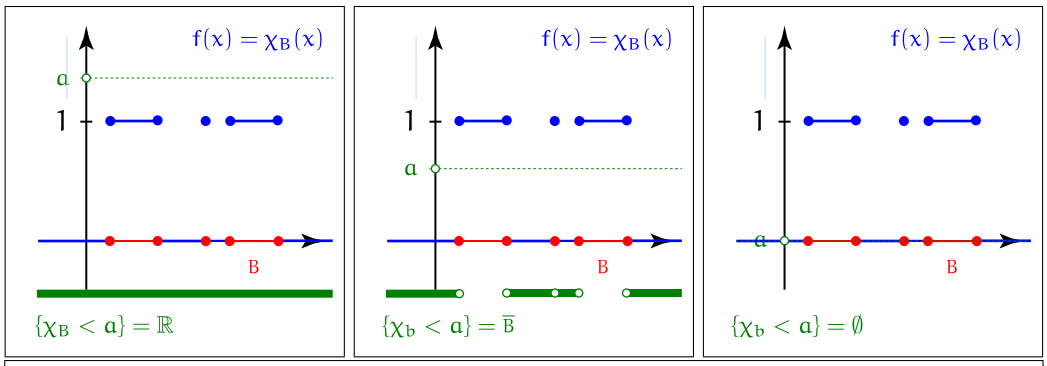


FIGURE 2.31: MEASURABILLITY OF THE CHARACTERISTIC FUNCTION, χ_B. The left image shows that the set $\{\chi_B < a\}$ corresponds to the measurable set $\mathbb{R}$. The image in the center identifies $\{\chi_B < a\}$ as the set of points which are contained in the complement of the measurable set B. The right image shows that there are no points of $\mathbb{R}$ whose image is smaller than zero.

with $B = \{(\mathbf{x}, \mathbf{x}') \in \mathcal{V} \times \mathcal{V} \mid \mathbf{x}$ *and* $\mathbf{x}'$ *are mutually visible*$\}$ *is measurable if and only if* B *is measurable. The proof immediately follows from the fact that for a measurable set* B *the relation*

$$\{\mathcal{V} < a\} = \begin{cases} \mathcal{V} \times \mathcal{V} & \text{if } a > 1 \\ \overline{B} & \text{if } 0 < a \leq 1 \\ \emptyset & \text{if } a \leq 0 \end{cases} \tag{2.227}$$

holds.

In the following discussions we are mainly interested in the class of real-valued measurable functions. Due to [22, Berezansky & al. 1996], this class of measurable functions is a linear space—we leave the proof to the interested reader. Additionally it holds for two measurable functions f, g, that the product $f \cdot g$, $|f|$, $\frac{f}{g}$, as well as $\max\{f, g\}$, and $\min\{f, g\}$ are measurable functions. The real- or complex-valued functions defined over a measurable set, taking only a finite number of values out of $\mathbb{R}$ or $\mathbb{C}$ may also be assumed to be measurable as they represent a linear combination of simple functions. As every measurable function f, defined on the measurable space $(\mathcal{R}, \mathfrak{R})$, permits a partition into its *positive* and *negative parts*, see Figure 2.32, thus,

Linear Space (854)
Linear Combination (857)
Simple Functions (839)

$$f = f^+ - f^- \tag{2.228}$$

with

$$f^+ \stackrel{def}{=} \max_{x \in \mathcal{R}} \{f(x), 0\} \qquad \text{and} \qquad f^- \stackrel{def}{=} \max_{x \in \mathcal{R}} \{-f(x), 0\}, \tag{2.229}$$

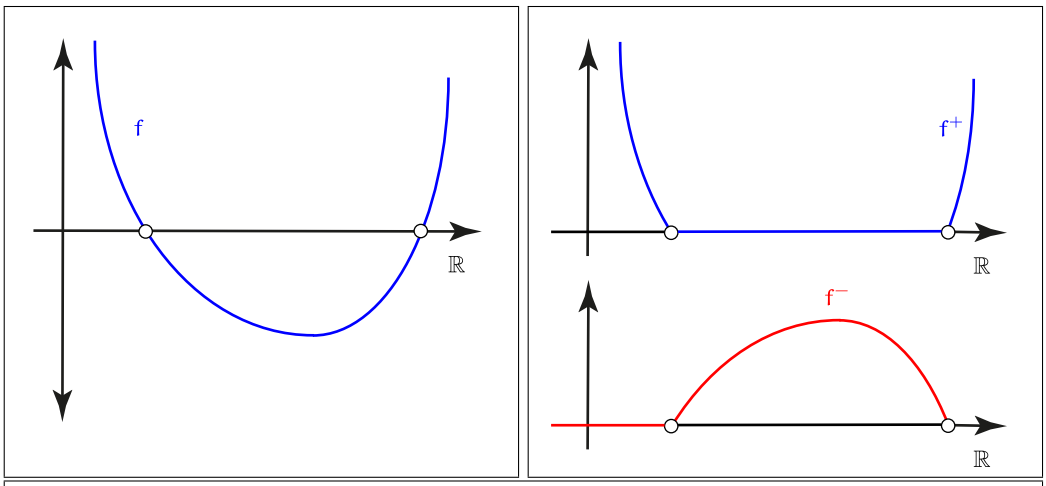


FIGURE 2.32: POSITIVE AND NEGATIVE PART OF A FUNCTION. A real-valued function f and its decomposition into a positive part f^+ and a negative part f^-.

the positive and negative parts of a function are also ultimately found to be measurable as they are defined via the maximum of the measurable functions f and 0. In fact, practically any function that can be described is measurable. But on the other side it is also possible to construct non-measurable functions.

Recall from the calculation of the Lebesgue measure of a closed interval $[a, b] \subset \mathbb{R}$, then it holds: $\mu([a, b]) = \mu([a, b)) + \mu(\{b\}) = \mu([a, b))$, i.e. the null set $\{b\}$ does not contribute to the calculation of the measure $\mu([a, b])$. A similar situation will now be encountered for calculating the pre-image measures of measurable functions that take on certain values over null sets of their domain. This leads us to the concept of the *equivalence of functions* on the basis of the ν-*almost everywhere property*.

μ (75)

Null Set (80)

DEFINITION 2.28 (ν-almost Everywhere Property) *Let $(\mathcal{R}, \mathfrak{R}, \nu)$ be a measure space. We call a property a* ν-almost everywhere property, *if it holds on the set $A \setminus N, A \in \mathfrak{R}$ with $\nu(N) = 0$. That is, a ν-almost everywhere property holds on a set except for a null set.*

Measure Space (80)

Null Set (80)

Let us consider a real-valued, measurable function defined on $(\mathcal{R}, \mathfrak{R})$ that ν-almost everywhere takes on the value 0, thus $\nu(\{x | f(x) \neq 0\}) = 0$. This in effect means that the measure zero is assigned to a set that is not identical to zero. This situation goes counter to the natural, elementary geometric features of a measure as a synonym of the concept of length or volume. We solve such a problem by using the concept of the ν-almost everywhere property in which two measurable functions f and g are treated as *equivalent*, i.e. $f \sim g$, if $\nu(\{f \neq g\}) = 0$. The values of f and g coincide except for a

null set, whereby two equivalent functions are assigned the same measure. So, it can be shown that $\sim$ satisfies the conditions required to an *equivalence relation*, and that any function equivalent to a measurable function is therefore measurable itself. If $\mathcal{M}(\mathcal{R})$ is called the set of all measurable functions on the measurable space $(\mathcal{R}, \mathfrak{R})$, then the factor set $\mathcal{M}\big|_{\sim}$ consists of the equivalence classes $[f] = \{h \in \mathcal{M}(\mathcal{R}) \big| h \sim f\}$ of equivalent measurable functions, where a representative of the class $[f]$ is given via any function f taken from the class [31, Capiński & al. 2000].

Let us illustrate this abstract concept by means of a famous example, the Dirichlet function.

EXAMPLE 2.36 (Once again the Dirichlet Function) *We know from Relation (A.8) that the Dirichlet function takes the value one only at rational points of the measure space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mu)$. *Since* $\mu(\mathbb{Q}) = 0$, *the Dirichlet function is* μ*-almost everywhere zero, thus:* $D \sim 0$.

EXAMPLE 2.37 (The Function Space $\mathcal{L}^\infty(\mathcal{R})$) *Let* $\mathcal{L}^\infty(\mathcal{R})$ *be the set of all functions measurable on the base set* $\mathcal{R}$, *which, apart from a null set, take on finite values on their domain, that is,*

$$\mathcal{L}^\infty(\mathcal{R}) \stackrel{\text{def}}{=} \left\{ f \,|\, |f(x)| \leq c \quad \nu\text{-almost everywhere on } \mathcal{R},\ c \in \mathbb{R}^{>0} \right\}, \tag{2.230}$$

then $\mathcal{L}^\infty(\mathcal{R})$ *clearly satisfies the conditions requested to a linear space as with* $f, g \in \mathcal{L}^\infty(\mathcal{R})$ *also* $\alpha f + \beta g \in \mathcal{L}^\infty(\mathcal{R})$ *holds for* $\alpha, \beta \in \mathbb{R}$.

Now strictly speaking, an element $f \in \mathcal{L}^\infty(\mathcal{R})$ *does not correspond to a measurable function but rather to an equivalence class of measurable functions. This implies that the concept of the supremum is of no interest and must be replaced by the essential supremum, i.e. the largest lower bound of upper bounds which holds* ν*-almost everywhere for* f. *This may be expressed formally as:*

$$\operatorname*{ess\,sup}_{x \in \mathcal{R}} |f(x)| \stackrel{\text{def}}{=} \inf \left\{ c \,|\, |f(x)| \leq c \quad \nu\text{-almost everywhere on } \mathcal{R} \right\}. \tag{2.231}$$

Via the essential supremum, we now define a norm on $\mathcal{L}^\infty(\mathcal{R})$ *by*

$$\|f\|_{\mathcal{L}^\infty} \stackrel{\text{def}}{=} \operatorname*{ess\,sup}_{x \in \mathcal{R}} |f(x)|, \tag{2.232}$$

which $(\mathcal{L}^\infty(\mathcal{R}), \|\cdot\|_{\mathcal{L}^\infty})$ *arises to a linear normed space.*

Let us now show how the ν*-almost everywhere property can be coupled to the ray space* $\mathcal{R}^{\partial\mathcal{V}}$, *to obtain a linear normed space* $(\mathcal{L}^\infty(\mathcal{R}^{\partial\mathcal{V}}), \|\cdot\|_{\mathcal{L}^\infty})$ *which, apart from the natural reflection and transmission behavior, also covers the ideal specular reflection and the ideal refraction properties of materials. Defining*

$$\mathcal{L}^\infty(\mathcal{R}^{\partial\mathcal{V}}) \stackrel{\text{def}}{=} \left\{ L \,|\, |L(s, \omega)| \leq c \quad \nu\text{-almost everywhere on } \mathcal{R}^{\partial\mathcal{V}},\ c \in \mathbb{R}^{>0} \right\}, \tag{2.233}$$

with

$$\|L\|_{\mathcal{L}^\infty} \stackrel{def}{=} ess\sup_{s\in\partial\mathcal{V}} ess\sup_{\omega\in S^2} |L(s,\omega)|, \tag{2.234}$$

then $(\mathcal{L}^\infty(\mathcal{R}^{\partial\mathcal{V}}), \|\cdot\|_{\mathcal{L}^\infty})$ *not only contains ray functions with finite values on* $\mathcal{R}^{\partial\mathcal{V}}$*, but also a number of important unbounded functions from the field of global illumination. These functions can take on infinite values, as the space, due to the* ν*-almost everywhere property, ignores unbounded function values assumed on null sets. As a consequence, incident and excitant* ν*-almost everywhere unbounded functions may thus be composed with bounded functions from* $\mathcal{L}^\infty(\mathcal{R}^{\partial\mathcal{V}})$ *via the elementary vector space operations.*

Incident & Exitant Function (48)

REMARK 2.32 *Strictly speaking, in all discussions relating to equivalent functions a difference should be made between the equivalent class* $[f]$ *and a representative* f*. Nevertheless, provided the precise nature of the involved spaces is kept in mind, in measure theory it is general practice to always refer to the elements of a space as functions, so that in effect no explicit difference between representatives and classes is made.*

Contrary to the construction of the *Riemann integral* in calculus, the mathematical concept of convergence of sequences of functions plays a fundamental role in measure and integration theory. Coupled with the ν-almost everywhere property, we need this concept not only for the definition of the *Lebesgue integral*, but also for expressing a number of important convergence statements in probability theory. These culminate in the *Weak and Strong Laws of the Large Numbers* and the *Central Limit Theorem*.

Riemann Integral (876)
Sequence of Functions (30)
Lebesgue Integral (105)
Section 2.4.6

DEFINITION 2.29 (ν-almost Everywhere Convergence and Convergence in Measure) *Let us assume* $(f_n)_{n\in\mathbb{N}}$ *be a sequence of measurable functions defined on the measure space* $(\mathcal{R}, \mathfrak{R}, \nu)$*. The sequence* f_n *is called* ν-almost everywhere convergent *to the limit function* f*, if, except for a null set,* f_n *converges point-wise to the limit function* f*. This means that there exists a set* $N \subset \mathcal{R}$ *with* $\nu(N) = 0$*, so that* $\forall x \in \mathcal{R}\setminus N$ *the following holds:*

Measure Space (80)
Pointwise Convergence (31)

$$f_n(x) \xrightarrow{a.e.} f(x). \tag{2.235}$$

Otherwise, we say that f_n converges in measure ν *to* f*, if it applies:*

$$\lim_{n\to\infty} \nu\left(\{|f_n - f| \geq \tau\}\right) = 0 \tag{2.236}$$

for any $\tau > 0$*. To denote convergence in measure we write* $f_n \xrightarrow{\nu} f$*.*

On the basis of these convergence types one obtains the measurability of the limit function of a sequence of functions $(f_n)_{n\in\mathbb{N}}$ that are ν-almost everywhere finite and con-

Limit Function (31)

verges to f in measure. Additionally, we obtain also the measurability of f under the condition of uniform and point-wise convergence of the sequence of functions $(f_n)_{n\in\mathbb{N}}$, a result of fundamental importance for the definition of the Lebesgue integral.

Uniform Convergence (32)

2.2.4 THE LEBESGUE INTEGRAL AND THE $\mathcal{L}^p$ SPACES

Why we have decided to introduce the Lebesgue integral? Would it not be satisfying for our concerns to work with the easier and well-known concept of the Riemann integral? Then we would have spared us the study of measure, and measurable functions from the previous section!

Now, as we will see, the class of Lebesgue-integrable functions is considerably richer than the class of Riemann-integrable functions, that is, the Lebesgue integral extends the concept of integrability to a class of special functions, such as those with uncountable infinite discontinuities. But do we need such functions? No, our goal will not be to be able to integrate exotic functions like the Dirichlet function.

Dirichlet Function (836)

From calculus it is known, that the Riemann integral and the limit can be exchanged when considering uniformly convergent sequences of functions. But uniform convergence of a sequence of functions is already a strong condition. In many applications of functional analysis we need limit processes in spaces of integrable functions, which allow us to exchange the concepts of the integral and the limit of a sequence of functions under much more general conditions then uniform convergence. Such statements are required in particular in functional analysis to find methods for solving differential and integral equations given over complex function spaces, the so-called *Lebesgue spaces*. The Lebesgue integral underlies all those function spaces since the associated convergence theorems require much weaker conditions from sequences of functions. Additionally, the Lebesgue integral allows to consider measures that have no densities, such as the Dirac measure. Today, the Lebesgue integral is the integral concept of modern mathematics. Its generality and beauty makes them to an indispensable tool for functional analysis and probability theory, which are completely build on the concept of the Lebesgue integral.

Based on the mathematical concepts of the σ-algebra, the measure, and the measurable function, we are now in the position to introduce the *Lebesgue integral* on general function spaces. As already mentioned above, it is the integral notion of modern mathematics and may be regarded as the base for the theory of integral equations in functional analysis as well as the most important function and probability spaces in integration and probability theory.

σ-algebra (828)
Measure (79)
Function Space (28)
Section 2.4

The classical procedure for deriving the Lebesgue integral of measurable functions, carried out on the measure space $(\mathcal{R}, \mathfrak{R}, \mu)$ equipped with a finite measure, normally consists of three phases:

Measure Space (80)

1) In the first step a definition of the Lebesgue integral is given for simple functions.

Simple Function (839)

2) After that, a μ-almost everywhere non-negative bounded function is approximated by means of a sequence of nondecreasing simple, measurable functions that converge towards f, where the Lebesgue integral is defined in terms of the resulting limit. Sequence of Functions (30) μ-almost everywhere (101)

3) In the last step, a μ-almost everywhere bounded measurable function f may be partitioned into its positive as well as its negative part, i.e. $f = f^+ - f^-$, and the associated Lebesgue integrals are computed like in step 2. The difference of these two Lebesgue integrals then results in the integral of f. μ-almost everywhere (101) f^+, f^- (100)

THE CLASSICAL PROCEDURE FOR DERIVING THE LEBESGUE INTEGRAL. In the following let $(\mathcal{R}, \mathfrak{R}, \mu)$ be a measure space with a finite measure μ. Measure Space (80) Measure (79)

DEFINITION 2.30 (The Lebesgue Integral of Real-valued Functions) *The Lebesgue integral of f is then defined by:*

i) If f represents a simple, non-negative measurable function, then the following applies to the Lebesgue integral *of f over a measurable set $B = \cup_{j=1}^{m} B_j \in \mathfrak{R}$:* Simple Function (839) Measurable Function (98) Measurable Set (80)

$$\int_B f(x)\,d\mu(x) \stackrel{def}{=} \int_{\mathcal{R}} \chi_B(x)\, f(x)\,d\mu(x) \tag{2.237}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \mu(B_i \cap B_j). \tag{2.238}$$

ii) Assuming $(f_n)_{n\in\mathbb{N}}$ be a nondecreasing sequence of simple, non-negative, bounded, and measurable functions with values in $\mathbb{R}$, which converges pointwise towards the limit function f, then the Lebesgue integral of f is defined as: Bounded Function (863) Pointwise Convergence (31)

$$\int_B f(x)d\mu(x) \stackrel{def}{=} \lim_{n\to\infty} \int_B f_n(x)\,d\mu(x) \tag{2.239}$$

$$= \lim_{n\to\infty} \int_{\mathcal{R}} \chi_B(x)\, f_n(x)\,d\mu(x) \tag{2.240}$$

$$= \lim_{n\to\infty} \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_{n_i} \mu(B_i \cap B_j), \tag{2.241}$$

where α_{n_i} are the coefficients of the simple function f_n.

iii) If we now consider a real-valued, μ-almost everywhere finite, and measurable function f defined on the measure space $(\mathcal{R}, \mathfrak{R}, \mu)$, then f may be written as the sum of its positive and its negative part. If in such a case the Lebesgue integral constructed over the positive and negative parts exists, then the Lebesgue integral also exists for f, and we obtain: μ-almost everywhere (101) f^+, f^- (100)

$$\int_B f(x)d\mu(x) \stackrel{def}{=} \int_B f^+(x)\,d\mu(x) - \int_B f^-(x)\,d\mu(x)$$

$$= \int_{\mathcal{R}} \chi_B(x)\, f^+(x)\,d\mu(x) - \int_{\mathcal{R}} \chi_B(x)\, f^-(x)\,d\mu(x). \tag{2.242}$$

The function f *is referred to as* Lebesgue-integrable *with respect to the measure* μ *and* f *is assigned to the set of* Lebesgue-integrable functions *over* $\mathcal{R}$*, denoted as* $\mathcal{L}^1(\mathcal{R}, \mathfrak{R}, \mu)$*, or briefly* $\mathcal{L}^1(\mathcal{R}, \mu)$.

Let us demonstrate the power of the Lebesgue integral by means of the *Dirichlet function*, one of the most famous examples of a function which is non-Riemann but Lebesgue-integrable.

EXAMPLE 2.38 (The Dirichlet Function, a non Riemann, but Lebesgue-integrable Function) *As any subinterval of* $[0,1]$ *contains rational as well as irrational numbers, we have for the Dirichlet function, that the superior Riemann integral* $\overline{\int}_0^1 D(x)\,dx$ *over the closed interval* $[0,1]$ *is equal to one, while we get the value zero for the inferior Riemann integral* $\underline{\int}_0^1 D(x)\,dx$*, that is, the Dirichlet function is not Riemann-integrable on the unit interval.*

Riemann Integral (877)
D (836)

Let us now consider the Lebesgue integral of D*, where the unit interval is partitioned into a disjoint union of the sets* $([0,1] \cap \mathbb{Q})$ *and* $([0,1] \cap (\mathbb{R} \setminus \mathbb{Q}))$*, thus the set of rational points and the set of irrational points in* $[0,1]$*. From the countability of* $\mathbb{Q}$ *and* $[0,1] \cap \mathbb{Q}$*, we conclude:*

Countability (827)

$$\int_{[0,1]} D(x)\,d\mu(x) \overset{(2.238)}{=} 1 \cdot \mu([0,1] \cap \mathbb{Q}) + 0 \cdot \mu([0,1] \cap (\mathbb{R} \setminus \mathbb{Q})) \tag{2.243}$$

$$\overset{\mu([0,1]\cap\mathbb{Q})=0}{=} 0 \cdot \mu([0,1] \cap (\mathbb{R} \setminus \mathbb{Q})) \tag{2.244}$$

$$= 0. \tag{2.245}$$

That is, the Dirichlet function is not Riemann-integrable, but it is in fact Lebesgue-integrable with Lebesgue measure zero. It is an example of a measurable function with uncountably infinite discontinuities, which shows the power of the Lebesgue integral.

Measurable Function (98)

Let us now establish some useful properties of the Lebesgue integral, which serve as base of many new concepts needed for discussing the global illumination problem.

Measure Space (80)

LEMMA 2.2 *Let* f *and* g *be measurable functions from the measure space* $(\mathcal{R}, \mathfrak{R}, \mu)$*, the Lebesgue integral then satisfies the following conditions:*

i) If $f \geq 0$ *on a set of measure zero, we have:*

$$\int_{\mathcal{R}} f(x)\,d\mu(x) \geq 0. \quad \textit{(Non-negativity)} \tag{2.246}$$

ii) If $f \geq g$ *on a set of measure zero, then it holds:*

$$\int_{\mathcal{R}} f(x)\,d\mu(x) \geq \int_{\mathcal{R}} g(x)\,d\mu(x). \quad \textit{(Monotonicity)} \tag{2.247}$$

iii) Let $\mathcal{R} = \bigcup_{i=1}^{\infty} A_i$ be a countably additive partition of the base set $\mathcal{R}$, with disjoint set A_i, then it holds

$$\int_{\bigcup_{i=1}^{\infty} A_i} f(x)\, d\mu(x) = \sum_{i=1}^{\infty} \int_{A_i} f(x)\, d\mu(x). \quad \textit{(Countable additivity)} \tag{2.248}$$

PROOF 2.2 *The proof of this lemma is based on the definition of the Lebesgue integral. It is very easy and can be found in [22, Berezansky & al. 1996]. Hence we omit the proof, and leave it to the interested reader as an exercise.*

REMARK 2.33 *As we will see in Section 10.1.2, the countable additivity of the Lebesgue integral, that is property* iii) *from Lemma 2.2, is fundamental for the derivation of the classical discrete radiosity equation since it allows the decomposition of the integration domain, thus the scene to be rendered, into a finite set of 2-dimensional surfaces. The countable additivity of the Lebesgue integral plays also a central role when discussing rendering algorithms with respect to direct and indirect illumination at surface points in the Chapters 8 and 9.*

Discrete Radiosity Equation (784)

THE LEBESGUE FUNCTION SPACES. Until now, we have treated points of measure spaces as our interesting objects. We will now alter our point of view lying the focus on integrable functions as points in function spaces and here in particular on function spaces that are the basis for all of our rendering procedures.

Measure Space (80)
Section 2.1.1
Section 2.1.3

DEFINITION 2.31 (The Lebesgue Spaces $\mathcal{L}^p(\mathcal{R}, \mu)$) *Let $(\mathcal{R}, \mathfrak{R}, \mu)$ be a measure space and $\mathcal{L}^p(\mathcal{R}, \mu)$, defined by:*

Measure Space (80)

$$\mathcal{L}^p(\mathcal{R}, \mu) \stackrel{def}{=} \left\{ f \in \mathcal{L}(\mathcal{R}, \mu) \middle| \textit{ with } \int_{\mathcal{R}} |f(x)|^p\, d\mu(x) < \infty \right\},\ p \geq 1, p \neq \infty, \tag{2.249}$$

be the set of all measurable functions which are p-Lebesgue-integrable. *Then it may be shown, via the* Minkovski inequality for integrals[3]*, that $\mathcal{L}^p(\mathcal{R}, \mu)$ satisfies the prerequisites for a linear space. With the norm*

Measurable function (98)
Linear Space (854)
$\|\cdot\|$ (860)

$$\|f\|_{\mathcal{L}^p} \stackrel{def}{=} \left(\int_{\mathcal{R}} |f(x)|^p\, d\mu(x) \right)^{\frac{1}{p}},\ p \geq 1, p \neq \infty, \tag{2.251}$$

$\mathcal{L}^p(\mathcal{R}, \mu)$ becomes a linear normed space. $(\mathcal{L}^p(\mathcal{R}, \mu), \|\cdot\|_{\mathcal{L}^p})$ is referred to as the Lebesgue space *with respect to the Lebesgue measure, also briefly denoted as the $\mathcal{L}^p$-*space*, where $\|f\|_{\mathcal{L}^p}$ is called the $\mathcal{L}^p$-*norm*. This statement can easily be proofed by*

Linear Normed Space (860)

[3] Assuming $f, g \in \mathcal{L}(\mathcal{R}, \mu)^p$ and $1 \leq p < \infty$, then the following applies:

$$\left(\int_{\mathcal{R}} |f \pm g|^p\, d\mu(x) \right)^{\frac{1}{p}} \leq \left(\int_{\mathcal{R}} |f|^p\, d\mu(x) \right)^{\frac{1}{p}} + \left(\int_{\mathcal{R}} |g|^p\, d\mu(x) \right)^{\frac{1}{p}}. \tag{2.250}$$

applying the Minkovski inequality and the axioms satisfying a linear space—we leave the proof to the interested reader as an exercise.

Now, let us explain the new concept of the Lebesgue function space by means of the ray spaces introduced in Section 2.1.3. These new function spaces are the fundamental building block for the derivation of operator models for the light and importance transport in a vacuum and in participating media, which we will present in Chapter 5.

Ray Spaces (44)

EXAMPLE 2.39 (The Lebesgue Spaces $\mathcal{L}^1(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp)$ and $\mathcal{L}^1(\mathcal{R}^{\mathcal{V}^o}, \zeta)$) *Let us recall the function spaces $\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}})$ and $\mathcal{L}(\mathcal{R}^{\mathcal{V}^o})$ defined over the ray space $\mathcal{R}^{\partial\mathcal{V}}$ and $\mathcal{R}^{\mathcal{V}^o}$. From a physical perspective it may be said, that in global illumination theory the focus is placed only on those functions of $\mathcal{R}^{\partial\mathcal{V}}$ and $\mathcal{R}^{\mathcal{V}^o}$ that may in a certain sense be measurable. In particular we are interested in those functions that have a finite measure.*

Ray Spaces (44)

Throughput Measures (94)

Lebesgue Measure (75)

Solid Angle Measures (87)

Further above, it was shown that the throughput-measures are defined on the ray spaces $\mathcal{R}^{\partial\mathcal{V}}$ and $\mathcal{R}^{\mathcal{V}^o}$ via the ordinary Lebesgue measure and the solid angle measures. Based on this construct, we are now able to define the spaces that contain functions that could be used to describe the particle distribution in a scene to be rendered. These spaces then play a central role in analyzing the properties of light transport operators.

Chapter 5

$\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}})$ (46)

Obviously, the space of ray functions $\mathcal{L}(\mathcal{R}^{\partial\mathcal{V}})$ in free space can be equipped with the throughput measure $\zeta^\perp$. This leads us to a first Lebesgue space on $\mathcal{R}^{\partial\mathcal{V}}$: The space $\mathcal{L}^1(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp)$, it is defined as:

$$\mathcal{L}^1(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp) \stackrel{\text{def}}{=} \left\{ f \in \mathcal{L}(\mathcal{R}^{\partial\mathcal{V}}) \,\middle|\, \|f\|_{\mathcal{L}^1} < \infty \right\}, \tag{2.252}$$

where the $\mathcal{L}^1$-norm is given via the throughput measure $\zeta^\perp$, namely by:

$$\|f\|_{\mathcal{L}^1} \stackrel{\text{def}}{=} \int_{\partial\mathcal{V}} \int_{S^2(\mathbf{s})} |f(\mathbf{r})| \, d\zeta^\perp \mathbf{r} \tag{2.253}$$

$$\stackrel{(2.208)}{=} \int_{\partial\mathcal{V}} \int_{S^2(\mathbf{s})} |f(\mathbf{s}, \omega)| \, d\mu^2(\mathbf{s}) \, d\sigma^\perp_{\mathbf{s}}(\omega) \tag{2.254}$$

with $\mathbf{r} = (\mathbf{s}, \omega)$.

The space $\mathcal{L}^1(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp)$ contains all functions that integrated over the unit sphere at all boundary points of a volume $\mathcal{V}$ yield a finite value.

$\mathcal{L}(\mathcal{R}^{\mathcal{V}^o})$ (46)

Analog to the above construction, the space of ray functions $\mathcal{L}(\mathcal{R}^{\mathcal{V}^o})$ in participating media can be equipped with the throughput measure ζ^o. This leads us to a first Lebesgue space on $\mathcal{R}^{\mathcal{V}^o}$: The space $\mathcal{L}^1(\mathcal{R}^{\mathcal{V}^o}, \zeta^o)$ is defined as:

$$\mathcal{L}^1(\mathcal{R}^{\mathcal{V}^o}, \zeta^o) \stackrel{\text{def}}{=} \left\{ f \in \mathcal{L}(\mathcal{R}^{\mathcal{V}^o}) \,\middle|\, \|f\|_{\mathcal{L}^1} < \infty \right\}, \tag{2.255}$$

where the $\mathcal{L}^1$-norm is given via the throughput measure ζ^o by:

$$\|f\|_{\mathcal{L}^1} \overset{def}{=} \int_{\mathcal{V}^o}\int_{S^2(\mathbf{x})} |f(\mathbf{r})|\,d\zeta^o\mathbf{r} \tag{2.256}$$

$$\overset{(2.209)}{=} \int_{\mathcal{V}^o}\int_{S^2(\mathbf{x})} |f(\mathbf{x},\omega)|\,d\mu^3(\mathbf{x})\,d\sigma_{\mathbf{x}}(\omega). \tag{2.257}$$

In Section 3.3 we will see, that functions of $\mathcal{L}^1(\mathcal{R}^{\partial\mathcal{V}},\zeta^\perp)$ as well as $\mathcal{L}^1(\mathcal{R}^{\mathcal{V}^o},\zeta^o)$ can be used to describe the radiometric quantity radiant power.

Note: The difference between the two above constructed spaces only lies in the integration with respect to the solid angle measures. At inner points of a medium, where only hypothetical surfaces exist, we have to integrate with respect to the solid angle measure σ while at the boundaries of a volume we have to integrate with respect to the projected solid angle measure $\sigma^\perp$. σ (84)

REMARK 2.34 *In the case, where we consider only functions defined on object surfaces with directions over one of the hemispheres, $\mathcal{L}^1(\mathcal{R}^{\partial\mathcal{V}},\zeta^\perp)$ can be defined as:*

$$\mathcal{L}^1(\mathcal{R}^{\partial\mathcal{V}},\zeta^\perp) \overset{def}{=} \left\{ f \in \mathcal{L}(\mathcal{R}^{\partial\mathcal{V}}) \,\middle|\, \|f\|_{\mathcal{L}^1} < \infty \right\}, \tag{2.258}$$

where the integration domain within the $\mathcal{L}^1$-norm is defined over $\partial\mathcal{V}\times\mathcal{H}^2_+$ or $\partial\mathcal{V}\times\mathcal{H}^2_-$.

REMARK 2.35 *Although the $\mathcal{L}^1$-function spaces from Example 2.39 are complete linear spaces, and thus are candidates for function spaces on which solutions to the light transport equations could exist, we do not use these spaces as the functions spaces in realistic rendering. In Example 2.41, we will show that the specification of the global illumination problem requires the restriction of the $\mathcal{L}^1$-spaces to smaller linear spaces. Namely, if we endow the $\mathcal{L}^1$-spaces with corresponding inner products, based on the Lebesgue integral, then we get functions spaces, the so-called $\mathcal{L}^2$-spaces, which are Hilbert spaces. These spaces will be the function spaces on which we will search for solutions to the global illumination problem.*

Due to the definition of the Lebesgue integral[4], the function spaces $\mathcal{L}^p(\mathcal{R},\mu)$ also satisfy the condition of completeness which implies that they are Banach spaces. In the particular case that the set $\mathcal{R}$ is bounded, then the following applies for $p \geq q, p, q \in \mathbb{R}$: Banach Space (35) Bounded Set (862)

$$\mathcal{L}^p(\mathcal{R},\mu) \subset \mathcal{L}^q(\mathcal{R},\mu), \tag{2.259}$$

that is, together with the space $\mathcal{L}^\infty(\mathcal{R},\mu)$ we obtain $\mathcal{L}^\infty(\mathcal{R},\mu)$ (102)

[4]Contrary to $\mathcal{L}^p(\mathcal{R},\mu)$ the space of Riemann p-integrable functions with $1 \leq p < \infty$ is not complete, as it is possible to construct Cauchy sequences of Riemann p-integrable functions whose limit functions are not Riemann-integrable themselves.

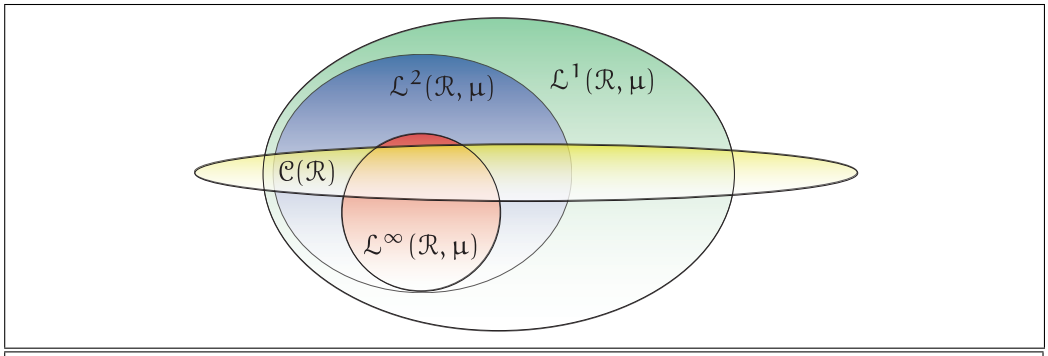


FIGURE 2.33: $\mathcal{L}^p$-SPACE: The relationship between the $\mathcal{L}^p$-spaces and the space of continuous functions.

$$\mathcal{L}^\infty(\mathcal{R},\mu) \subset \cdots \subset \mathcal{L}^p(\mathcal{R},\mu) \subset \cdots \subset \mathcal{L}^1(\mathcal{R},\mu). \tag{2.260}$$

From this subset relation, we conclude, that with increasing p, the size of the space $\mathcal{L}^p(\mathcal{R},\mu)$ will be smaller.

Continuous Function (869)

REMARK 2.36 *In this context, it is also interesting to note, that the space* $C(\mathcal{R})$ *of continuous functions is not a subset of any of the* $\mathcal{L}^p$*-spaces, see Figure 2.33. As an example let us consider the function* x^{-1}*, which indeed belongs to* $C((0,1))$ *but is unbounded on* $[0,1)$*. On the other hand it is obvious, that the space of bounded continuous functions is a subset of* $\mathcal{L}^\infty(\mathcal{R},\mu)$*.*

Unbounded Function (863)

One special case of a $\mathcal{L}^p$-space of great importance for many applications is the case $(p = 2)$. A number of features applying in the Euclidian space and used for the solution of many problems may be transferred to this case via the construction of an inner product.

Inner Product (859)

DEFINITION 2.32 (The Lebesgue Space $\mathcal{L}^2(\mathcal{R},\mu)$) *Suppose the Lebesgue space* $\mathcal{L}^2(\mathcal{R},\mu)$ *is equipped with the inner product*

Inner Product (859)

$$\langle f, g\rangle \stackrel{\text{def}}{=} \int_{\mathcal{R}} f(x)\, g(x)\, d\mu(x), \tag{2.261}$$

$\forall f, g \in \mathcal{L}^2(\mathcal{R},\mu)$*. If we define a norm* $\|\cdot\|_{\mathcal{L}^2}$ *based on this inner product by:*

$$\|f\|_{\mathcal{L}^2} \stackrel{\text{def}}{=} \langle f, f\rangle = \int_{\mathcal{R}} |f(x)|^2\, d\mu(x), \tag{2.262}$$

Hilbert Space (36)

then the Lebesgue $\mathcal{L}^2(\mathcal{R},\mu)$ *becomes a Hilbert space, in fact the only* $\mathcal{L}^p$*-space that is simultaneously also a Hilbert space.*

EXAMPLE 2.40 (The Lebesgue Space $(S^2, \mathfrak{B}(S^2), \sigma)$) *In the following, again and again, we need to integrate functions that are defined on the unit sphere. Such a function should be an element of the Lebesgue space $(S^2, \mathfrak{B}(S^2), \sigma)$, where $\mathfrak{B}(S^2)$ is the Borel σ-algebra over the unit sphere and σ is the solid angle measure.*

Spherical Harmonics (124)
Solid Angle Measure (84)

REMARK 2.37 *The inner product on a $\mathcal{L}^2$-space makes it possible to apply the concept of orthogonality to functions of $\mathcal{L}^2(\mathcal{R}, \mu)$. Thus, square-integrable functions, defined on infinite dimensional spaces, can be projected onto approximate functions in finite dimensional subspaces, a point of fundamental importance for developing rendering methods based on radiosity procedures. Additionally, a number of phenomena whose integral representations are complex may be formulated more simply as inner products.*

Orthogonality (859)
Chapter 10
Chapter 5

We will now construct three important Lebesgue spaces based on the ray spaces $\mathcal{R}^{\partial\mathcal{V}}$, $\mathcal{R}^{\mathcal{V}^o}$ and $\mathcal{R}$ as well as the throughput measures ζ, $\zeta^\perp$ and $\overline{\zeta}$, which we will use in particular in Chapter 5.

Ray Spaces (44)
Throughput Measures (94)

EXAMPLE 2.41 (The Lebesgue Spaces $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp)$, $\mathcal{L}^2(\mathcal{R}^{\mathcal{V}^o}, \zeta)$ and $\mathcal{L}^2(\mathcal{R}, \overline{\zeta})$) *According to Definition 2.32, the space of square Lebesgue-integrable ray functions $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp)$ in free space is defined as:*

$$\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp) \stackrel{\text{def}}{=} \left\{ f \in \mathcal{L}(\mathcal{R}^{\partial\mathcal{V}}) \,\middle|\, \|f\|_{\mathcal{L}^2} < \infty \right\}, \tag{2.263}$$

and the $\mathcal{L}^2$-norm is given by the inner product

$$\begin{aligned} \|f\|_{\mathcal{L}^2} \stackrel{\text{def}}{=} \langle f, f \rangle &= \int_{\partial\mathcal{V} \times S^2(\mathbf{s})} |f(\mathbf{r})|^2 \, d\zeta^\perp(\mathbf{r}) && (2.264) \\ &\stackrel{(2.208)}{=} \int_{\partial\mathcal{V}} \int_{S^2(\mathbf{s})} |f(\mathbf{s}, \omega)|^2 \, d\mu^2(\mathbf{s}) \, d\sigma^\perp_{\mathbf{s}}(\omega). && (2.265) \end{aligned}$$

Thus, $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp)$ contains all functions, which are square-integrated over the unit sphere and at all surface points $\partial\mathcal{V}$.

Analogously, we can define the space $\mathcal{L}^2(\mathcal{R}^{\mathcal{V}^o}, \zeta^o)$ as the space of square Lebesgue-integrable ray functions in participating media by:

$$\mathcal{L}^2(\mathcal{R}^{\mathcal{V}^o}, \zeta^o) \stackrel{\text{def}}{=} \left\{ f \in \mathcal{L}(\mathcal{R}^{\mathcal{V}^o}) \,\middle|\, \|f\|_{\mathcal{L}^2} < \infty \right\}, \tag{2.266}$$

where the $\mathcal{L}^2$-norm is given via the throughput measure ζ^o by:

$$\begin{aligned} \|f\|_{\mathcal{L}^2} \stackrel{\text{def}}{=} \langle f, f \rangle &= \int_{\mathcal{V}^o \times S^2(\mathbf{x})} |f(\mathbf{r})|^2 \, d\zeta^o(\mathbf{r}) && (2.267) \\ &\stackrel{(2.209)}{=} \int_{\mathcal{V}^o} \int_{S^2(\mathbf{x})} |f(\mathbf{x}, \omega)|^2 \, d\mu^3(\mathbf{x}) \, d\sigma_{\mathbf{x}}(\omega). && (2.268) \end{aligned}$$

Last but not least, we define the Lebesgue space $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ *by:*

$$\mathcal{L}^2(\mathcal{R},\overline{\zeta}) \stackrel{\text{def}}{=} \left\{ f \in \mathcal{L}(\mathcal{R}) \,\middle|\, \|f\|_{\mathcal{L}^2} < \infty,\ f = f^{\partial\mathcal{V}} + f^{\mathcal{V}^o} \right\}, \tag{2.269}$$

with $f^{\partial\mathcal{V}}(\mathbf{x},\omega) \equiv 0$ *for* $\mathbf{x} \in \mathcal{R} \setminus \partial\mathcal{V}$, $f^{\mathcal{V}^o} \equiv 0$ *for* $\mathbf{x} \in \mathcal{R} \setminus \mathcal{V}^o$ *where* $\mathcal{V} = \mathcal{V}^o \cup \partial\mathcal{V}$.

REMARK 2.38 *In the case, where we consider only functions defined on points at object surfaces with directions over one of the hemispheres,* $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}},\zeta^{\perp})$ *can also be defined as:*

$$\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}},\zeta^{\perp}) \stackrel{\text{def}}{=} \left\{ f \in \mathcal{L}(\mathcal{R}^{\partial\mathcal{V}}) \,\middle|\, \|f\|_{\mathcal{L}^2} < \infty \right\}, \tag{2.270}$$

where the integration domain within the $\mathcal{L}^2$*-norm was defined over* $\partial\mathcal{V} \times \mathcal{H}^2_+$ *or* $\partial\mathcal{V} \times \mathcal{H}^2_-$.

REMARK 2.39 *The spaces* $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}},\zeta^{\perp})$, $\mathcal{L}^2(\mathcal{R}^{\mathcal{V}^o},\zeta)$ *and* $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$*, generated over the ray spaces* $\mathcal{R}^{\partial\mathcal{V}}$, $\mathcal{R}^{\mathcal{V}^o}$ *and* $\mathcal{R}$ *are the fundamental function spaces within which we try to develop methods for finding approximate solutions for the global illumination problem.*

As these spaces are Hilbert spaces, they are complete linear normed spaces, that fulfill the Cauchy-convergence criterion, that is, the Banach fixed-point theorem is applicable to solve operator equations via iteration methods.

Measurement Equation (416)

So, in Section 4.6 it will be possible to formulate the measurement equation, which has to be evaluated for solving the global illumination problem, simply as the inner product of two functions from the above Lebesgue spaces. The construction of these Lebesgue spaces also allows to formulate the light transport as well as the importance transport via an interpretation equivalent to the natural perspective.

Chapter 5

Section 2.4.4

In addition, the $\mathcal{L}^2$*-spaces play a central role in the variance analysis of random variables and random vectors—as the calculation of the variance of a random variable requires that the random variable is a square-Lebesgue-integrable function—and represents the basis of finite element algorithms for solving the global illumination problem.*

Chapter 10

REMARK 2.40 *In functional analysis it is shown—by constructing exotic sequences of functions, such as in Relation 2.12—that the space of* square-Riemann-integrable functions, $\mathcal{R}^2(\mathcal{R})$, *given by:*

$$\mathcal{R}^2(\mathcal{R}) \stackrel{\text{def}}{=} \left\{ f \,\middle|\, \int_{\mathcal{R}} |f(x)|^2 \, dx \right\}, \tag{2.271}$$

where the integral must be interpreted in the sense of Riemann, is not a complete function space. So, it is possible that a Cauchy-sequence of square-Riemann-integrable functions f_n *can converge to a limit function* f *that is not square-Riemann-integrable function.*

Since the Riemann integral only leads to pre-Hilbert spaces—where the Cauchy-convergence criterion is not valid—it can not be used as the concept of integration in the theory of integral equations. Here it is often only possible to approximate the exact solution of an integral equation via a sequence of functions from the underlying function space where the exact solution lives. So, the theory of integral equations can only be handled in a strict mathematical sense if the underlying function spaces have at least Banach or Hilbert space structure. This is also the main reason, why we musten decide to use the Lebesgue integral as the concept of integration in realistic rendering.

EXAMPLE 2.42 (Fourier Series Representations) *Let us recall the* Fourier Series Theorem *based on a Hilbert Space $\mathcal{S}$ spanned by a countably infinite set of orthonormal functions $\mathcal{B}_\phi^\infty = \{\phi_0, \phi_1, \ldots\}$. The space of square-Lebesgue-integrable functions $\mathcal{L}^2([-\pi,\pi], \mathfrak{B}([-\pi,\pi]), \mu)$ on the closed interval $[-\pi,\pi]$ with the inner product $\langle\cdot,\cdot\rangle_\mathcal{S}$ defined by:*

Fourier Series Theorem (39)
Orthonormal Function (861)

$$\langle f(x), g(x)\rangle_\mathcal{S} \stackrel{\text{def}}{=} \int_{[-\pi,\pi]} f(x)\, g(x) d\mu(x) \tag{2.272}$$

then satisfies the requirements to a Hilbert space. Using the orthonormal basis $\mathcal{B}_\phi^\infty = \{\phi_0, \phi_1, \ldots\}$ given by:

Orthonormal Basis (37)

$$\phi_0(x) = \frac{1}{\sqrt{2\pi}},\ \phi_{2n}(x) = \frac{1}{\sqrt{\pi}}\cos(nx),\ \phi_{2n+1}(x) = \frac{1}{\sqrt{\pi}}\sin(nx) \tag{2.273}$$

then any function $f \in \mathcal{L}^2([-\pi,\pi], \mathfrak{B}([-\pi,\pi]), \mu)$ may be represented by an infinite series of sines and cosines, namely as:

$$f(x) = \sum_{i=0}^{\infty} \langle f, \phi_i\rangle_\mathcal{S}\, \phi_i. \tag{2.274}$$

The proof of this statement is leaved to the interested reader as an exercise.

EXAMPLE 2.43 (Fourier Transform) *Let $\mathfrak{B}(\mathbb{R}^s)$ be the Borel σ-algebra over the open subsets of $\mathbb{R}^s$ and f be a complex-valued, Lebesgue-integrable function of $\mathcal{L}\big(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s)\big)$, then the linear operator*

Section 7.3
Borel σ-Algebra (865)
Linear Operator (53)

$$\mathcal{F} : \mathcal{L}\big(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s)\big) \to \mathcal{L}\big(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s)\big) \tag{2.275}$$

with

$$f \to \mathcal{F}(f) \tag{2.276}$$

is referred to as the Fourier transform, *while the operator $\mathcal{F}^{-1}$ inverse to $\mathcal{F}$, is called the* inverse Fourier transform. *The image $\widehat{f} \stackrel{\text{def}}{=} \mathcal{F}f$ of the function f under the operator $\mathcal{F}$ defined as:*

$$\widehat{f}(\mathbf{t}) : \mathbb{R}^s \to \mathbb{C} \tag{2.277}$$

$$\mathbf{t} \to \widehat{f}(\mathbf{t}) = \frac{1}{(2\pi)^{\frac{s}{2}}} \int_{\mathbb{R}^s} e^{-i\,\langle \mathbf{t},\mathbf{x}\rangle_{\mathbb{R}^s}} f(\mathbf{x}) d\mu^s(\mathbf{x}) \tag{2.278}$$

is called the Fourier transform *of* f. *The inverse operator, i.e.* $\widehat{f^{-1}} \stackrel{def}{=} \mathcal{F}^{-1}f$, *defined as:*

$$\widehat{f^{-1}} : \mathbb{R}^s \to \mathbb{C} \tag{2.279}$$

$$x \to \widehat{f^{-1}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{s}{2}}} \int_{\mathbb{R}^s} e^{i\,\langle \mathbf{x},\mathbf{t}\rangle_{\mathbb{R}^s}} f(\mathbf{t}) d\mu^s(\mathbf{t}) \tag{2.280}$$

is referred to as the inverse Fourier transform *of* f *[54, Elstrodt 1996].*

Now, as a linear operator the Fourier transform satisfies certain conditions, where the most important and most powerful of them is given by the convolution

$$\mathcal{F}(f \star g)(\mathbf{t}) = \mathcal{F}(f)(\mathbf{t}) \cdot \mathcal{F}(g)(\mathbf{t}), \tag{2.281}$$

defined as:

$$f(\mathbf{x}) \star g(\mathbf{x}) \stackrel{def}{=} \int_{\mathbb{R}^s} f(\mathbf{y})\, g(\mathbf{y}-\mathbf{x})\, d\mu^s(\mathbf{y}). \tag{2.282}$$

With the Fourier transform, the Fourier analysis provides us a tool for analyzing the efficiency of patterns resulting from various different sampling processes in Monte Carlo and quasi-Monte Carlo procedures. If f *corresponds to a pattern on a pixel or the information relating to an image defined over a pixel, then a single point* $\widehat{f}(\mathbf{t}_1, \mathbf{t}_2)$ *of the Fourier transform, referred to in this case as the* Fourier spectrum, *indicates how much information of the spatial frequency* $(\mathbf{t}_1, \mathbf{t}_2)$ *is contained in the image. This implies that a point in the Fourier space contains information on the entire image.*

Chapter 6 Chapter 7

Section 7.3

In a later section further below, we will generate the Fourier spectra of the most frequently used sampling techniques in Monte Carlo algorithms. They provide us with statements on the efficiency and the use of these techniques in rendering procedures for the solution of the global illumination problem.

As already mentioned, the Lebesgue integral is the integral concept required not only in global illumination but also in probability theory. There, we are often faced with the task of drawing samples from probability spaces composed of $\mathbb{R}^2$ or $\mathbb{R}^3$, and the upper as well as the lower hemisphere or the unit sphere. Mathematically, this corresponds to integrating a measurable function over domains which are Cartesian products of different sets. The *Theorem of Fubini-Tonelli* now yields not only a simple method for calculating such integrals, but also delivers a method which allows to define distributions over *conditional probability spaces*. Moreover, it is of high practical use in the analysis of rendering pro-

Section 2.4 Section 6.5 Measurable Function (98) Section 2.4.5

cedures based on Monte Carlo methods for solving the measurement equation underlying the global illumination problem. Defined as the integral over $\partial\mathcal{V}$ and S^2, the theorem of Fubini-Tonelli suggests to solve this integral by generating a large number of rays emitted from a point on $\partial\mathcal{V}$ in a iterative manner and carrying out the integration over the surface of the sensor.

$\partial\mathcal{V}$ (41)

THEOREM 2.6 (Theorem of Fubini-Tonelli) *Let $(\mathcal{R}, \mathfrak{R}, \nu)$ and $(\mathcal{R}', \mathfrak{R}', \nu')$ be measure spaces, additionally let us suppose, that f is a measurable function defined on $\mathcal{R} \times \mathcal{R}'$ with values in $\mathbb{R}$. Under the condition, that the iterated integrals*

Measure Space (80)
Measurable Function (98)

$$f'(x') = \int_{\mathcal{R}''} f(x', x'') \, d\nu''(x'') \tag{2.283}$$

and

$$f''(x'') = \int_{\mathcal{R}'} f(x', x'') \, d\nu'(x') \tag{2.284}$$

can be computed, the following holds:

$$\begin{aligned} \int_{\mathcal{R}' \times \mathcal{R}''} f(x', x'') \, d(\nu' \otimes \nu'')(x', x'') &= \int_{\mathcal{R}'} \left(\int_{\mathcal{R}''} f(x', x'') \, d\nu''(x'') \right) d\nu'(x') && (2.285) \\ &= \int_{\mathcal{R}''} \left(\int_{\mathcal{R}'} f(x', x'') \, d\nu'(x') \right) d\nu''(x''). && (2.286) \end{aligned}$$

PROOF 2.6 *For a proof, see [31, Capiński & Kopp 2000].*

The Theorem of Fubini-Tonelli establishes a connection between a multiple integral and iterated integrals such that it allows to compute a multiple integral using iterated integrals. Additionally, it allows to change the order of integration in iterated integrals. Let us demonstrate this technique by means of two simple examples which we will use in the following again and again.

EXAMPLE 2.44 (The Area of the Unit Sphere) *From our discussion about the solid angle measure, it should be clear, that the area of the unit sphere is given by:*

Solid Angle Measure (87)

$$\int_{S^2} d\sigma(\omega). \tag{2.287}$$

Parameterizing the sphere in spherical coordinates and transforming the solid angle measure via

Spherical Coordinates (832)

$$d\sigma(\theta, \phi) \overset{(2.186)}{=} \sin\theta \, d\mu(\theta) d\mu(\phi) \tag{2.288}$$

to the Lebesgue measure on $[0, 2\pi) \times [0, \pi]$ yields:

μ^2 (82)

$$\int_{S^2} d\sigma(\omega) \overset{(2.186)}{=} \int_{[0,2\pi)\times[0,\pi]} \sin\theta \, d\mu(\theta) \, d\mu(\phi) \tag{2.289}$$

$$= \int_{[0,2\pi)} \left(\int_{[0,\pi]} \sin\theta \, d\mu(\theta) \right) d\mu(\phi) \tag{2.290}$$

$$\overset{(2.6)}{=} \int_{[0,2\pi)} \left(-\cos\theta \Big|_0^{\pi} \right) d\mu(\phi) \tag{2.291}$$

$$= 2 \int_{[0,2\pi)} d\mu(\phi) \tag{2.292}$$

$$= 4\pi. \tag{2.293}$$

EXAMPLE 2.45 (Projected Area of the Hemisphere) *The projected area of the upper hemisphere on a surface can be written as the Lebesgue integral based on the projected solid angle measure,* $\sigma^\perp$ (88)*, thus,*

$$\int_{\mathcal{H}^2_+} d\sigma^\perp(\omega). \tag{2.294}$$

Spherical Coordinates (A.1)

Parameterizing the hemisphere in spherical coordinates and transforming the solid angle measure via

$$d\sigma^\perp(\theta,\phi) \overset{(2.186)}{=} \sin\theta \, |\cos\theta| \, d\mu(\theta) d\mu(\phi) \tag{2.295}$$

to the Lebesgue measure on $[0,2\pi) \times \left[0, \frac{\pi}{2}\right]$ *yields:*

$$\int_{\mathcal{H}^2_+} d\sigma^\perp(\omega) \overset{(2.186)}{=} \int_{[0,2\pi)\times[0,\frac{\pi}{2}]} \sin\theta \, |\cos\theta| \, d\mu(\theta) \, d\mu(\phi) \tag{2.296}$$

$$= \int_{[0,2\pi)} \left(\int_{[0,\frac{\pi}{2}]} \sin\theta \, |\cos\theta| \, d\mu(\theta) \right) d\mu(\phi) \tag{2.297}$$

$$\overset{(2.6)}{=} \int_{[0,2\pi]} \left(-\frac{\cos^2\theta}{2} \Big|_0^{\frac{\pi}{2}} \right) d\mu(\phi) \tag{2.298}$$

$$= \frac{1}{2} \int_{[0,2\pi)} d\mu(\phi) \tag{2.299}$$

$$= \pi. \tag{2.300}$$

Section 1.1.3

REMARK 2.41 *Let us recall once more the naive principle of ray tracing from our introductory chapter—a detailed discussion follows in Section 8.3—then ray tracing may be considered as a rough implementation of the Theorem of Fubini-Tonelli. So, the shading of a pixel corresponds to the inner integration carried out over the unit sphere, while the outer integration, carried out on the object surfaces, corresponds to the construction of an image on the image plane.*

When dealing with probability theory and random variables in Section 6.5.1, we will be confronted again and again with the following problem: To a given random variable X with known probability density function p_X and a function T we are seeking the density function of the random variable $Y = T(X)$ in terms of p_X. The key to this problem lies in the *Theorem of Transformation* for the Lebesgue integral.

THEOREM 2.7 (Theorem of Transformation) *Let $\mathbf{U}$ be an open subset in $\mathbb{R}^s$ and $\mathbf{T}^{-1} : \mathbf{U} \to \mathbb{R}^s$ an injective differentiable function with continuous partial derivatives, where the Jacobian for every $\mathbf{y} \in \mathbf{U}$ is nonzero. Then, for any real-valued, continuous function f it holds:*

Injectivity (840)
Jacobian (872)

$$\int_{\mathbf{U}} f(\mathbf{y}) d\mu^s(\mathbf{y}) = \int_{\mathbf{T}^{-1}(\mathbf{U})} f(\mathbf{T}^{-1}(\mathbf{y})) \det(J_{\mathbf{T}^{-1}}(\mathbf{y})) |d\mu^s(\mathbf{y}). \tag{2.301}$$

Usually, the theorem of transformation is not formulated as we did it above, but it is this formulation, which we make use in all of our discussion when describing the probability density function of a random variable $\mathbf{Y}$ in terms of an already known PDF of given random $\mathbf{X}$.

PROOF 2.7 *For a proof, see [174, Rudin 1998].*

2.2.5 THE LEBESGUE INTEGRAL IN GLOBAL ILLUMINATION THEORY

To conclude this short trip into measure and integration theory, we now present a few useful concepts for our further discussions. Based on the concept of the Lebesgue integral, they enable us to represent a number of functions of relevance for global illumination theory in an elegant and more simple way, of great use for a number of perspectives to follow below.

Lebesgue Integral (105)

THE DIRAC δ-DISTRIBUTION FOR GENERAL MEASURES. The *Dirac δ-distribution*, also briefly denoted as the *Dirac delta* or the *delta function*, is a mathematical construct introduced by theoretical physicist Paul Dirac. It is commonly defined to be zero everywhere on the real line with an infinitely large spike at position $\mathbf{x}_0$ such, that its total integral is 1, i.e.:

D0*) $\delta_\mu(x) = 0$ for $x \neq x_0$

D1*) $\delta_\mu(x) \to \infty$ at $x = x_0$ and

D2*) $\int_{]-\infty,\infty[} \delta_\mu(x)\, d\mu(x) = 1$,

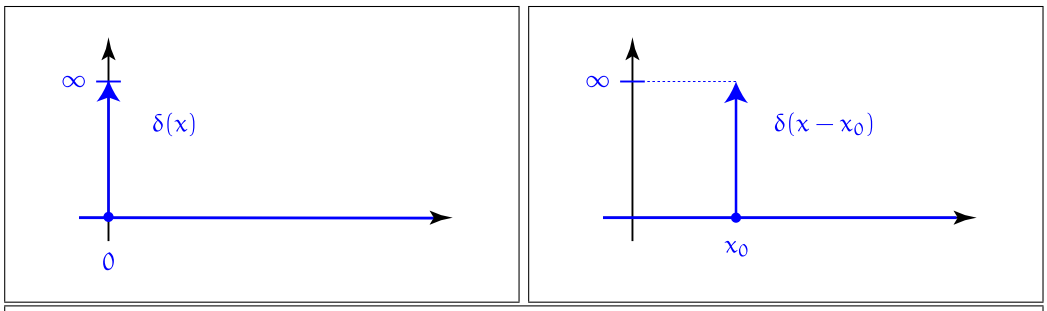


FIGURE 2.34: THE DIRAC DELTA. An attempt to visualize the Dirac δ-distribution.

see Figure 2.34.

Considering these properties, then it is clear, that the Dirac delta is not strictly a function defined on $\mathbb{R}$. Due to the fact, that δ is almost-everywhere zero on the real line, the value of the Lebesgue integral of δ over $]-\infty,\infty[$ should give zero. However, it is not possible to construct a function in the usual sense having the properties above described. Thus more correctly, the Dirac delta is defined as a bounded linear functional on the space of continuous functions $C^n(A)$, where A is a domain on $\mathbb{R}^n$.

Linear Functional (55)

Continuous Function (869)

DEFINITION 2.33 (The Dirac Delta Distribution) *Let* $C(A)$ *be the space of continuous functions on* A*. Then the mapping* δ_μ

$$\delta_\mu : C(A) \to \mathbb{R}$$

defined by

$$\begin{aligned} \delta_\mu f(\mathbf{x}) &\stackrel{\text{def}}{=} \int_A \delta_\mu(\mathbf{x}-\mathbf{x}_0)\, f(\mathbf{x})\, d\mu^n(\mathbf{x}) \\ &= \int_A \delta_\mu(x_1-x_{0_1})\cdots\delta_\mu(x_n-x_{0_n})\, f(x_1,\ldots,x_n)\, d\mu(x_1)\cdots d\mu(x_n) \\ &\stackrel{\text{def}}{=} f(\mathbf{x}_0) \end{aligned} \tag{2.302}$$

is called the Dirac δ-distribution *with respect to the measure* μ*, or referred as the* Dirac delta*, in the literature often also called the* Dirac delta function*. Based on this definition, the Dirac delta is a special kind of operator, namely a* bounded linear functional*, also known as a* distribution*, that acts on a continuous function and delivers the value of that function at position* $\mathbf{x}_0$*.*

Linear Functional (55)

C([a, b]) (38)

Unit Sphere (849)

EXAMPLE 2.46 *Let us consider the space* $C(S^2)$ *of all continuous functions defined on the unit sphere. Then, the corresponding Dirac delta can be constructed as follows:*

$$\delta_\sigma : C([0,\pi]\times[0,2\pi)) \to \mathbb{R} \tag{2.303}$$

with

$$\delta_\sigma f(\omega) \stackrel{def}{=} \int_{S^2} \delta(\omega - \omega_0)\, f(\omega)\, d\sigma(\omega) \tag{2.304}$$

$$\stackrel{def}{=} f(\omega_0). \tag{2.305}$$

REMARK 2.42 *The usage of the integral notation in the above definition is only symbolic. The Dirac δ-distribution is not defined as an integral with respect to the Lebesgue measure, but it is the evaluation of the functional δ at* $f(x)$*. As it produces the value of the continuous function* f *at point* x_0*, the Dirac δ-distribution has a sampling property.*

Another more strict method to define the Dirac delta could be to define it as the Lebesgue integral with respect to the Dirac measure from Example 2.27, thus,

$$\int_A f(x - x_0) d\delta_{x_0}(x) = f(x_0). \tag{2.306}$$

Due to the fact, that in Definition 2.33 there is nothing to be integrated in the usual sense, the integral notation in the above definition should be thought of as purely symbolic. In a certain sense the Dirac delta nullifies integration, replacing the integral by the integrand at a certain point of the integration area. We mention it once more, the Dirac delta is, for reasons executed above, *not* a function, but it is a bounded linear functional or a distribution, which can be manipulated as a function.

Linear Functional (55)

As we will see in our further discussions, the concept of the Dirac δ-distribution is a useful tool for a mathematical formulation many light phenomena in global illumination theory. But when deriving formulas for these light phenomena, we have to be careful, in particular, if the integration measure is different from the measure used for the definition of δ-distribution, since it is possible to get meaningless results. Therefore, we will now present several properties that make the usage of the Dirac-δ construct easier for us:

$$\text{(D0)} \quad \int_{]-\infty,\infty[^s} \delta_\mu(x - x_0)\, d\mu^n(x) = 1 \tag{2.307}$$

$$\text{(D1)} \quad \int_{]-\infty,\infty[^s} \delta_\mu(x - x_0)\, f(x)\, d\mu^n(x) = f(x_0) \tag{2.308}$$

$$\text{(D2)} \quad \delta_\mu(g(x) - g(x_0)) = \frac{1}{|g'(x_0)|}\delta_\mu(x - x_0) \tag{2.309}$$

$$\text{(D3)} \quad \delta_\mu(g(x)) = \sum_{\{x_i | g(x_i)=0\}} \frac{1}{|g'(x_i)|}\delta_\mu(x - x_i) \tag{2.310}$$

$$\text{(D4)} \quad \delta_\mu(x - x_0)\, f(x) = \frac{d\nu}{d\mu}(x_0)\, \delta_\nu(x_0 - x) \tag{2.311}$$

$$\text{(D5)} \quad \delta_\mu(x - x_0)\, f(x) = \delta_\mu(x_0 - x)\, f(x) = f(x_0). \tag{2.312}$$

REMARK 2.43 *The above identities are formulated for the multidimensional case. It is clear that they also hold in one dimension where we only have to use the Lebesgue integral on $\mathbb{R}$ symbolically. We also mention that* (D2) *is a special case of* (D3). *Here, we must sum over all arguments where the function f takes the value zero. In particular, property* (D4) *will be very useful when deriving formulas for the reflection of light at ideal specular surfaces. In the following, we give a short insight into the work with the Dirac δ-distribution.*

Lebesgue Integral (105)

EXAMPLE 2.47 *Let us consider the continuous function $g(x) = x^2 - a^2$ with $a \in \mathbb{R}$. Obviously, this function is zero at $x_{1,2} = \pm a$, that is, for the functional $\delta(g(x))\,x$ we have:*

$$\int_{]-\infty,\infty[} \delta_\mu(x^2 - a^2)\,x\,d\mu(x) \overset{(D3)}{=} \int_{]-\infty,\infty[} \frac{1}{2|a|}(\delta_\mu(x-a) + \delta_\mu(x+a))\,x\,d\mu(x) \tag{2.313}$$

$$= \frac{1}{2|a|}\int_{]-\infty,\infty[} \delta_\mu(x-a)\,x\,d\mu(x) + \frac{1}{2|a|}\int_{]-\infty,\infty[} \delta_\mu(x+a)\,x\,d\mu(x) \tag{2.314}$$

$$\overset{(D1)}{=} \frac{1}{2|a|}a - \frac{1}{2|a|}a. \tag{2.315}$$

EXAMPLE 2.48 *For the practical use, formulas in the directional variable ω have often to be expressed in terms of polar coordinates (θ, ϕ). With respect to Example 2.46, where the Dirac δ-distribution was introduced with respect to the solid angle measure σ, we can then formulate an equivalent expression for the δ-distribution with respect to the Lebesgue measure μ:*

σ (87)

$$\delta_\sigma f(\omega) \overset{def}{=} \int_{S^2} \delta(\omega - \omega_0)\,f(\omega)\,d\sigma(\omega) \tag{2.316}$$

$$= \int_{S^2} \delta(\omega - \omega_0)\,f(\omega)\,\frac{d\sigma(\omega)}{d\mu(\theta)d\mu(\phi)}\,d\mu(\theta)d\mu(\phi) \tag{2.317}$$

$$\overset{(2.186)}{=} \int_{[0,2\pi)}\int_{[0,\pi]} \delta((\theta,\phi) - (\theta_0,\phi_0))\,f(\theta,\phi)\,\sin\theta\,d\mu(\theta)d\mu(\phi) \tag{2.318}$$

$$\overset{(2.311)}{=} \sin\theta_0\,f(\theta_0,\phi_0). \tag{2.319}$$

Let us now present the power and elegance of the Dirac delta construct by means of an example from global illumination: the *ideal specular reflection*. As we will see in one of the next chapters, a surface is called *ideal specular reflective* if a light ray, incoming from a single direction, is reflected into the mirrored direction. This physical effect can mathematically be formulated with the help of the Dirac delta.

Section 4.2.1.2

EXAMPLE 2.49 (A Simple Reflection Model) *In order to illustrate the elegance of the Dirac delta construct let us consider two functions L_o and L_i defined on the Lebesgue*

Incident & Exitant Function (48)

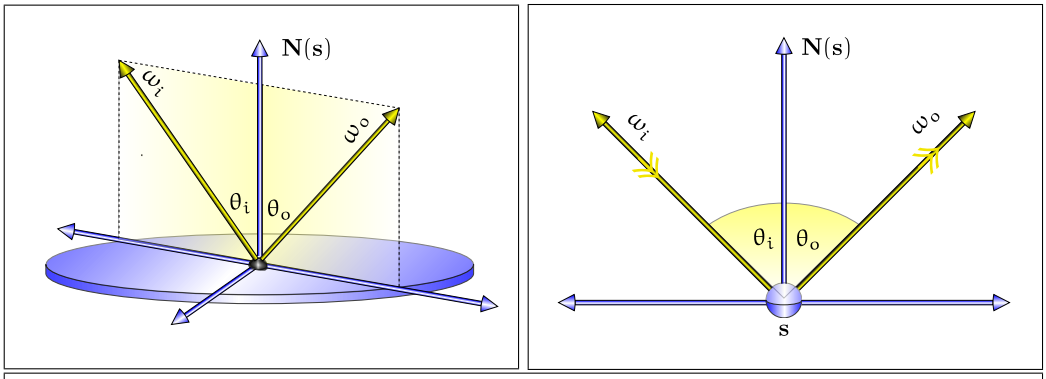


FIGURE 2.35: A SIMPLE REFLECTION MODEL. Given an incident direction ω_i at point s that makes an angle θ_i with the surface normal at s then the reflected direction ω_o makes an angle $\theta_o = \theta_i$ with $\mathbf{N}(s)$ and lies within the plane spanned by the vectors ω_i and $\mathbf{N}(s)$.

space $\mathcal{L}^2(\mathcal{R}, \overline{\zeta})$. *Suppose furthermore that the function* L_i *returns the amount of light that arrives in direction* ω_i *at a surface point* s, *and that* L_o *provides information about the light that leaves that point in direction* ω_o.

$\mathcal{L}^2(\mathcal{R}, \overline{\zeta})$ (111)

Let us assume that the reflection of a light ray at a surface is loss-less, then the entire amount of incident light L_i *from direction* ω_i *is reflected into a single outgoing direction* ω_o. *Due to the* Law of Reflection *it holds obviously:*

Law of Reflection (300)

$$L_o(s, \omega_o) = L_i(s, \omega_i), \tag{2.320}$$

which, with $\omega_i = (\theta_i, \phi_i)$, *as well as* $\omega_o = (\theta_o, \phi_o)$, *i.e.,*

Spherical Coordinates (832)

$$\theta_o = \theta_i \quad \textit{and} \quad \phi_o = \phi_i \pm \pi, \tag{2.321}$$

can be written as:

$$\begin{aligned} L_o(s, \theta_o, \phi_o) &= L_i(s, \theta_i, \phi_i \pm \pi) && (2.322)\\ &= L_i(s, \theta_o, \phi_o \pm \pi), && (2.323) \end{aligned}$$

for a visualization, see also Figure 2.35.

This fact can now be modeled by an infinitely large spike function *in direction of* ω_o, *thus a Dirac* δ*-distribution. Using the Dirac delta construct—now symbolically written with respect to the Lebesgue integral based on the solid angle measure* σ *and transformed into spherical coordinates, see Example 2.46—then the radiance in*

σ (87)

direction ω_o *can be represented as:*

$$L_i(s, \theta_o, \phi_o \pm \pi) \tag{2.324}$$

$$\stackrel{def}{=} \int_{[0,2\pi)} \int_{[0,\pi]} \delta_\mu((\theta_i, \phi_i) - (\theta_o, \phi_o \pm \pi)) L_i(s, \theta_i, \phi_i))\, d\mu(\theta_i) d\mu(\phi_i) \tag{2.325}$$

$$\stackrel{(2.302)}{=} \int_{[0,2\pi)} \int_{[0,\pi]} \delta_\mu(\theta_i - \theta_o) \delta(\phi_i - (\phi_o \pm \pi))\, L_i(s, \theta_i, \phi_i))\, d\mu(\theta_i) d\mu(\phi_i). \tag{2.326}$$

Slightly reformulated, this relation can be represented as

$$L_i(s, \theta_o, \phi_o \pm \pi) \stackrel{def}{=} \int_{[0,2\pi)} \int_{[0,\pi]} \frac{\delta_\mu(\theta_i - \theta_o)}{\sin\theta_i |\cos\theta_i|} \delta(\phi_i - (\phi_o \pm \pi)) \tag{2.327}$$
$$L_i(s, \theta_i, \phi_i))\, \sin\theta_i |\cos\theta_i|\, d\mu(\theta_i) d\mu(\phi_i).$$

$f_r^\vee$ (325) *Replacing the Dirac* δ*-distribution in Equation (2.327) by the function* $f_r^\vee$*, the so-called* ideal specular BRDF*, defined by:*

$$f_r^\vee(s, (\theta_i, \phi_i) \longrightarrow (\theta_o, \phi_o)) \stackrel{def}{=} \frac{\delta_\mu(\theta_i - \theta_o)\, \delta_\mu(\phi_i - (\phi_o \pm \pi))}{\sin\theta_i\, |\cos\theta_i|}, \tag{2.328}$$

then we obtain a version of the light transport equation which describes the physical Section 4.2.1.2 *effect of* light reflection *at an* ideal specular surface*, well-known to every globillumer, namely:*

$$L_o(s, \theta_o, \phi_o) \tag{2.329}$$
$$= \int_{\mathcal{H}^2_+(s)} f_r^\vee(s, (\theta_i, \phi_i) \longrightarrow (\theta_o, \phi_o))\, L_i(s, \theta_i, \phi_i)\, \sin\theta_i\, |\cos\theta_i|\, d\mu(\theta_i) \mu(\phi_i).$$

Using property (D2) *of the Dirac* δ*-distribution from Equation (2.311), then* $f_r^\vee$ *in angular form can also be written as:*

$$f_r^\vee(s, (\theta_i, \phi_i) \longrightarrow (\theta_o, \phi_o)) = \frac{\delta_\mu(\cos\theta_i - \cos\theta_o)\, \delta_\mu(\phi_i - (\phi_o \pm \pi))}{|\cos\theta_i|} \tag{2.330}$$

$$= 2\delta_\mu(\sin^2\theta_i - \sin^2\theta_o)\, \delta_\mu(\phi_i - (\phi_o \pm \pi)). \tag{2.331}$$

We leave the detailed derivation of the representations (2.330) and (2.331) to the interested reader as an easy exercise.

REMARK 2.44 (Different Mathematical Formulations of Reflection at Ideal Specular Surfaces) *When deriving the so-called specular BRDF in Section 4.2.2.2 we will rely on a variant of Equation (2.329), where the Lebesgue measure* μ *is replaced by the* $\sigma^\perp$ (88) *projected solid angle measure* $\sigma^\perp$*.*

The desired relationship between L_i *and* L_o *for a perfect mirror is then given by:*

$$L_o(s, \omega_o) = L_i(M_N(\omega_o)) \tag{2.332}$$

which, using the Dirac δ-distribution $\delta_{\sigma^\perp}$ can easily be written as:

$$L_o(s,\omega_o) = \int_{S^2(s)} \delta_{\sigma^\perp}(\omega_i - M(\mathbf{N}(\omega_o)) L_i(s,\omega_i)\, d\sigma_s^\perp(\omega_i). \tag{2.333}$$

The ideal specular BRDF $f_r^\vee$—expressed in terms of incident and exitant directions, ω_i and ω_o—can then be defined via:

$$f_r^\vee(s,\omega_i \longrightarrow \omega_o) \stackrel{def}{=} \delta_{\sigma^\perp}(\omega_i - M(\mathbf{N}(\omega_o)). \tag{2.334}$$

In the literature, Equation (2.329) is also often expressed in terms of the solid angle measure σ instead of the projected solid angle measure $\sigma^\perp$. Using property (D4), then the ideal specular BRDF $f_r^\vee$ can also easily be expressed in terms of the the Dirac δ-distribution δ_σ with respect to the solid angle measure, namely: $\sigma^\perp$ (88)

$$\begin{aligned}
\delta_{\sigma^\perp}(\omega_i - M(\mathbf{N}(\omega_o)) &\stackrel{(2.311)}{=} \frac{d\sigma}{d\sigma^\perp}(M(\mathbf{N}(\omega_o))\delta_\sigma(\omega_i - M(\mathbf{N}(\omega_o)) && (2.335)\\
&\stackrel{(2.192)}{=} \frac{\delta_\sigma(\omega_i - M(\mathbf{N}(\omega_o))}{|\langle \omega_i, \mathbf{N}(s)\rangle|} && (2.336)\\
&= \frac{\delta_\sigma(\omega_i - M(\mathbf{N}(\omega_o))}{|\cos\theta_i|}. && (2.337)
\end{aligned}$$

This implies that the ideal specular BRDF $f_r^\vee$—used in Equation (2.329), which is given in the form

$$L_o(s,\omega_o) = \int_{S^2(s)} f_r^\vee(s,\omega_i \to \omega_o)\, \cos\theta_i\, d\sigma_s(\omega_i), \tag{2.338}$$

must be defined as:

$$f_r^\vee(s,\omega_i \longrightarrow \omega_o) \stackrel{def}{=} \frac{\delta_\sigma(\omega_i - M(\mathbf{N}(\omega_o))}{|\langle \omega_i, \mathbf{N}(s)\rangle||}. \tag{2.339}$$

In the above definition of the Dirac delta construct, the Lebesgue integral was not used in its traditional sense, since nothing has to be integrated. In the following discussion we now show how the Lebesgue integral can be used in its proper sense, namely as the modern integral concept in mathematics. Lebesgue Integral (105)

SPHERICAL HARMONICS. As a consequence of the Fourier Series Theorem it is known that a function given on a Hilbert space admits the representation as an infinite Fourier series. In analogy to this Fourier series expansion, there exist for any Lebesgue square-integrable function, defined on the unit sphere S^2, a representation as an infinite series of so-called *spherical harmonic functions*. This means, that the spherical harmonic functions form an orthonormal basis in the Lebesgue space $\mathcal{L}^2(S^2, \mathfrak{B}(S^2))$, thus the function space of square-integrable functions defined over the unit sphere. Fourier Series Theorem (39) Hilbert space (36) Orthonormal Basis (37) $\mathcal{L}^2(S^2, \mathfrak{B}(S^2))$ (111)

DEFINITION 2.34 *Let us consider the* spherical harmonic basis functions $Y_{l,m}(\omega)$, *which are defined as:*

$$Y_{l,m}(\theta,\phi) \stackrel{def}{=} \begin{cases} N_{l,m}\, P_{l,m}(\cos\theta)\cos(m\phi) & \text{falls } m > 0 \\ N_{l,0}\, \frac{P_{l,0}(\cos\theta)}{\sqrt{2}} & \text{falls } m = 0 \\ N_{l,m}\, P_{l,|m|}(\cos\theta)\sin(|m|\phi) & \text{falls } m < 0, \end{cases} \tag{2.340}$$

with the normalization constants $N_{l,m}$*:*

$$N_{l,m} \stackrel{def}{=} \sqrt{\frac{2l+1(l-|m|)!}{2\pi(l+|m|)!}}, \tag{2.341}$$

where the recursively defined associated Legendre polynomials $P_{l,m}$ *[235, Weisstein 2003] corresponds to:*

$$\begin{aligned} P_{0,0}(\cos\theta) &\stackrel{def}{=} 1 \\ P_{m,m}(\cos\theta) &\stackrel{def}{=} (1-2m)\sqrt{1-\cos^2\theta}\, P_{m-1,m-1}(\cos\theta) \\ P_{m+1,m}(\cos\theta) &\stackrel{def}{=} \cos\theta(2m+1)P_{m,m}(\cos\theta) \\ P_{l,m}(\cos\theta) &\stackrel{def}{=} \cos\theta\left(\frac{2l-1}{l-m}\right)P_{l-1,m}(\cos\theta) - \left(\frac{l+m-1}{l-m}\right)P_{l-2,m}(\cos\theta). \end{aligned} \tag{2.342}$$

Then a function $f \in \mathcal{L}^2(S^2, \mathfrak{B}(S^2))$ *can be written as*

$$f(\omega) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} C_{l,m}Y_{l,m}(\omega), \tag{2.343}$$

whereby $Y_{l,m}$ *are the* orthonormal, spherical harmonic basis functions *from above and the coefficients* $C_{l,m}$ *are given according to Equation (2.1) by:*

Fourier Coefficients (38)

$$C_{l,m} \stackrel{def}{=} \int_{S^2} f(\omega)Y_{l,m}(\omega)\, d\sigma(\omega). \tag{2.344}$$

EXAMPLE 2.50 *Obviously, the first four spherical harmonics are given as:*

$$Y_{0,0}(\omega) = \sqrt{\frac{1}{4\pi}}, \qquad Y_{1,0}(\omega) = \sqrt{\tfrac{3}{4\pi}}\cos\theta, \qquad Y_{1,\pm1}(\omega) = \mp\sqrt{\frac{3}{8\pi}}e^{\pm i\phi}\sin\theta. \tag{2.345}$$

Fourier Transform (113)

The spherical harmonics, for a visualization of the first basis functions see Figure 2.36, can be considered as the analogue to the Fourier Transform. While the Fourier Transform works over the unit circle for one-dimensional functions, spherical harmonics work over the unit sphere for two-dimensional functions. They can be used to approximate any mathematical function defined on the unit sphere, where the approximation gets better as more basis functions are used in the series expansion.

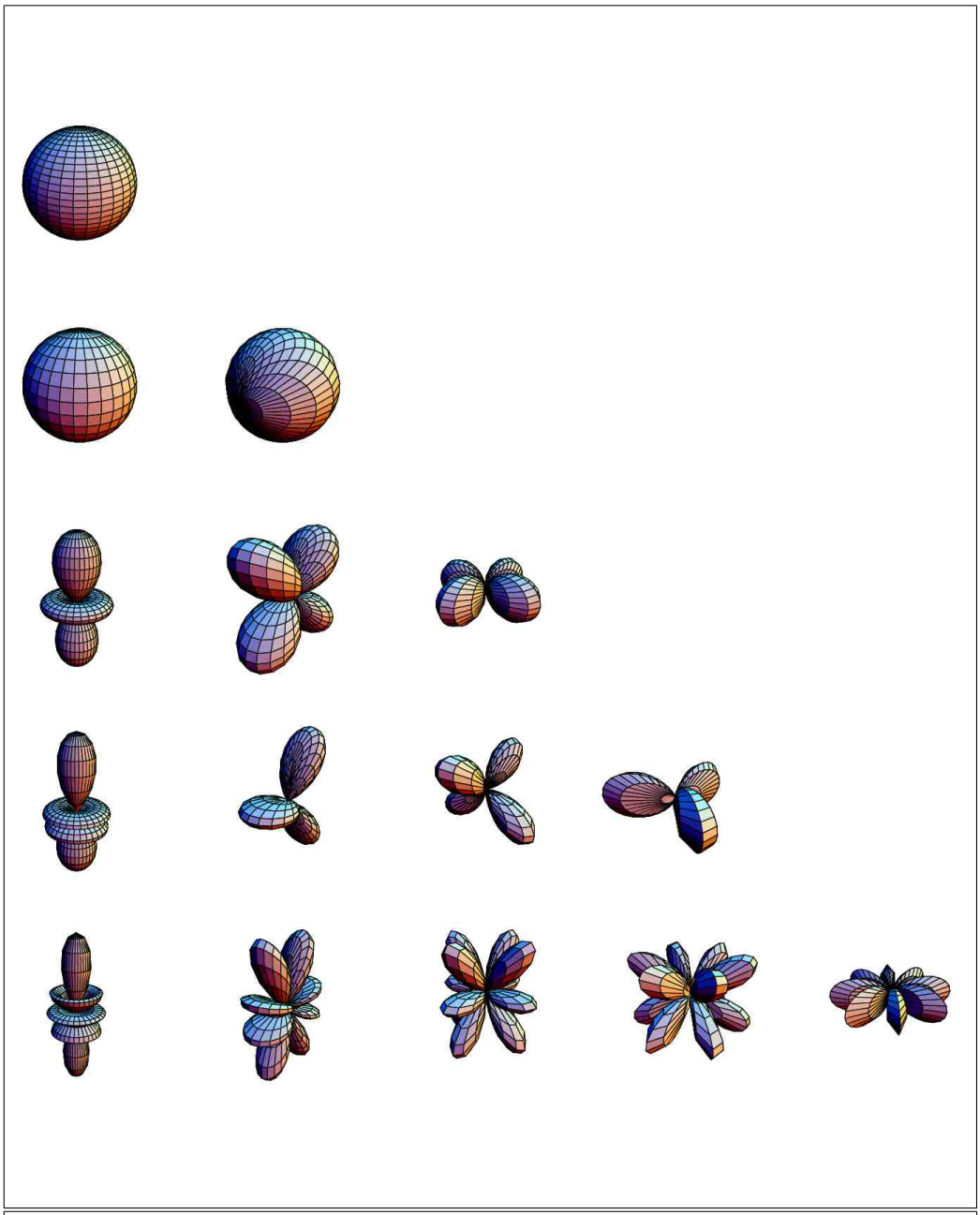


FIGURE 2.36: SPHERICAL HARMONICS. The first few spherical harmonic basis functions $Y_{l,m}(\omega)$ for $0 \leq l \leq 6, 0 \leq m \leq l$.

BRDF Representation by SH (345)

As we will see later, spherical harmonics play an important role in global illumination theory, in particular for the representation of bidirectional reflectance distribution functions. There, a very common method is to project the BRDF on spherical harmonics, where, if it is smooth and simple, the BRDF can be represented as a series consisting of only very few spherical harmonic basis functions. This results in a cost-efficient variant for computing the values of a BRDF. Another area where SHs play a central role is *pre-computed radiance transfer*, a technique for rendering a scene in real time using complex light interactions, see Example 4.1.

2.3 LINEAR INTEGRAL EQUATIONS

Light Transport Equation (17)
Exitant Function (48)

In our introductory chapter we have introduced the light transport equation in a vacuum, expressed in exitant radiance, it has the form

$$L_o(\mathbf{s}, \omega_o) = L_e(\mathbf{s}, \omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_o(\gamma(\mathbf{s}, \omega_i), -\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i). \quad (2.346)$$

Obviously, it specifies a function L_o expressed in terms of an integral that contains this function. In analogy to the definition of a differential equation—thus, an equation contains apart from the unknown function also the derivative of the unknown function—an equation of the above type, where the unknown function also appears under one or more integrals, is referred to as an *integral equation*.

The intention of the present section is to make the reader familiar with just this mathematical construct. Motivated by the light transport equation in free space from above, we want to introduce the reader specifically into the world of Fredholm type integral equations and their solution methods. Finding solutions to this type of integral equations is the main task of this book.

For that purpose, first, we have to define the construct of the *Fredholm integral equation of the* 2^{nd} *kind*. Based on the functional analytical concept of the linear operator, then we will show how it is possible to transform a linear integral equation into a linear integral operator equation as well as into its dual analogue, an adjoint integral equation. Afterwards, we talk about solution approaches for linear operator equations underlying Fredholm integral equations of the 2^{nd} kind. Here, we will distinguish between two different fundamentally approaches:

Section 2.3.1
Section 2.3.2
Section 2.3.3

- *analytical approaches*, resulting from functional analytical methods for finding the exact solution of linear operator equations, and
- *numerical approaches*, also derived from functional analysis, but with the focus on computing approximate solutions to linear operator equations.

CLASSIFICATION OF LINEAR INTEGRAL EQUATIONS. Let us start with introducing the general concept of the integral equation. As already mentioned above, any equation where the unknown function also appears under one or more integrals, is referred to as an *integral equation*. Formally, a linear integral equation is defined as:

DEFINITION 2.35 (Linear Integral Equation) *Let* $f, g \in \mathcal{L}^2(\mathcal{R}, \mu)$ *be real-valued functions, where* f *is linear and* λ *in general is a complex number. Let furthermore* $k \in \mathcal{L}^2(\mathcal{R} \times \mathcal{R}, \mu \times \mu)$ *be a real-valued function of two variables, called the* kernel of integration, *then an equation of the form*

$\mathcal{L}^2(\mathcal{R}, \mu)$ (110)

$$f(x) = g(x) + \lambda \int_{\mathcal{R}} k(x, y) f(y) \, d\mu(y) \tag{2.347}$$

is called a linear integral equation.

Depending on their external form, linear integral equations are classified in equations of the 1^{st}, 2^{nd}, and 3^{rd} *kind*. So, Equation (2.347) is called an *integral equation of the* 1^{st} *kind* if the unknown function f appears only on the right-hand side under the integral. It is denoted as an *integral equation of the* 2^{nd} or 3^{rd} *kind*, depending on whether f outside of the integral is weighted by a constant or a function [56, Engl 1997]. If the upper integration boundary is fixed, the equation is referred to as a *Fredholm equation*, if it is dependent on the variable x, it is called a *Volterra equation*. Is the *source function* $g(x)$, often also called the driving function, equals zero, then we talk of a *homogeneous*, otherwise of an *inhomogeneous* integral equation.

REMARK 2.45 (Integro-differential Equation) *In Chapter 4, we will derive a general formula for particle transport from which all of the integral equations of global illumination theory can be deduced. This equation contains the unknown function not only under and outside of the integral but it also contains the derivative of the unknown function. Such an equation is called a* Fredholm integro-differential equation *[234, Wazwaz 1997].*

Particle Transport Equation (294)

FREDHOLM INTEGRAL EQUATIONS OF THE 2^{nd} KIND. As we shall see in Section 4.6, the measurement equation will tell us something about the precise color of a pixel in an image. As the evaluation of the measurement equation requires the computation of radiance incident at a pixel, the primary job of any realistic rendering algorithm is to find the solution of the light transport equation. Therefore, we focus exclusively on the class of integral equations that describes the light and importance transport in form of a linear, inhomogeneous, Fredholm integral equation of the 2^{nd} kind.

Measurement Equation (416)
Radiance (250)
Chapter 9
Chapter 5

DEFINITION 2.36 (Fredholm Integral Equation of the 2^{nd} Kind) *A* Fredholm integral equation of the 2^{nd} kind *has the form*

$$f(x) = g(x) + \int_{\mathcal{R}} k(x, y) f(y) \, d\mu(y), \tag{2.348}$$

$\mathcal{L}^2(\mathcal{R},\mu)$ (110) *where* f, g *are elements of* $\mathcal{L}^2(\mathcal{R},\mu)$ *and the kernel* $k \in \mathcal{L}^2(\mathcal{R}\times\mathcal{R}, \mu\times\mu)$.

SLTEV (398) Comparing this form with the stationary light transport equation in free space expressed in exitant radiance, namely,

$$L_o(s,\omega_o) = L_e(s,\omega_o) + \int_{S^2(s)} f_s(s,\omega_i \to \omega_o)\, L_o(\gamma(s,\omega_i), -\omega_i)\, d\sigma_s^{\perp}(\omega_i), \quad (2.349)$$

then we note that the SLTEV is not conform to our standard definition of a Fredholm integral equation of the 2^{nd} kind from above. While we integrate in Equation (2.348) over the entire domain of the unknown function f, the integration in Equation (2.349) goes only over S^2 instead of $\mathcal{R}^{\partial\mathcal{V}} = \partial\mathcal{V}\times S^2$, which would correspond to the domain of the unknown function L_o in the light transport equation in a vacuum. Another difference comes from the occurence of the ray casting function γ in the first argument of the exitant radiance function L_o under the integral, while the unknown function f in Equation (2.348) is not dependent on another function but only on the variable x.

S^2 (849) $\mathcal{R}^{\partial\mathcal{V}}$ (44) γ (47) Exitant Function (48)

However, it is always possible to transform an equation, equivalent to the light transport equation in free space from Relation (2.349), into the standard form of a Fredholm integral equation of the 2^{nd} kind. To this, first we must remove the ray casting function by changing the variable of integration. Then, we have to transform the original integration over the smaller subset of the domain of L_o to an integration over the whole domain $\partial\mathcal{V}$ of the unknown function.

γ (47) $\partial\mathcal{V}$ (41)

Let us show this procedure in the following example based on the SLTEV from Equation 2.346.

EXAMPLE 2.51 (Transforming the Light Transport Equation into a Fredholm Integral Equation) *For sake of simplicity we rewrite the SLTEV from Equation (2.349) in terms of a surface point* s_j *with incident and exitant directions* ω_i^j *and* ω_o^j*, thus,*

$$L_o(s_j,\omega_o^j) = L_e(s_j,\omega_o^j) + \int_{S^2(s_j)} f_s(s_j,\omega_i^j \to \omega_o^j)\, L_o(\gamma(s_j,\omega_i^j), -\omega_i^j)\, d\sigma_{s_j}^{\perp}(\omega_i^j). \quad (2.350)$$

Exitant Function (48) *Then we define the exitant radiance from point* s_j *outgoing in direction* ω_o^j *as a function of the points* s_j *and* s_{j-1} *by*

$$L(s_j \to s_{j-1}) \equiv L_o(s_j,\omega_o^j), \quad (2.351)$$

where $s_{j-1} = \gamma(s_j,\omega_o^j)$ *is the hit point of the ray starting in* s_j *with a surface of a scene, see Figure 2.37. Based on this idea, then we can write the kernel* f_s *at point* s_j *as:*

$$f_s(s_{j+1} \to s_j \to s_{j-1}) \equiv f_s(s_j,\omega_i^j \to \omega_o^j) \quad (2.352)$$

with $s_{j+1} = \gamma(s_j,\omega_i^j)$.

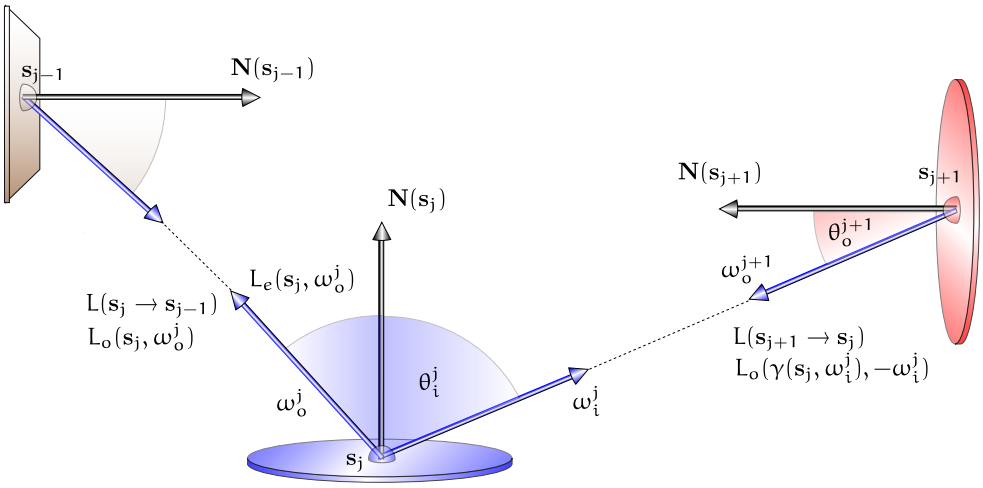


FIGURE 2.37: DEFINITION OF EXITANT POINT TO POINT RADIANCE. The radiance $L_o(\mathbf{s}_j, \omega_o^j)$ exitant from point $\mathbf{s}_j$ in direction ω_o^j can also be expressed in terms of radiance transfer $L(\mathbf{s}_j \to \mathbf{s}_{j-1})$ between two points $\mathbf{s}_j$ and $\mathbf{s}_{j-1}$ where $\mathbf{s}_{j-1} = \gamma(\mathbf{s}_j, \omega_o^j)$.

$\sigma^\perp$ (89) μ^2 (82) *To finish this procedure, we have to transform the projected solid angle measure* $\sigma^\perp$ *to the Lebesgue area measure* μ^2 *defined on object surfaces of* $\partial\mathcal{V}$. *Via the measure transformation from Equation (2.199) and the visibility function* $\mathcal{V}$, *then we can define the so-called* geometry term, $\mathcal{G}$, *by:*

$$\mathcal{G}(\mathbf{s}_{j+1} \leftrightarrow \mathbf{s}_j) \stackrel{\text{def}}{=} \mathcal{V}(\mathbf{s}_{j+1} \leftrightarrow \mathbf{s}_j) \frac{\left|\cos\theta_o^{j+1} \cos\theta_i^j\right|}{\|\mathbf{s}_{j+1} - \mathbf{s}_j\|_2^2}, \tag{2.353}$$

where $\mathcal{V}(\mathbf{s}_{j+1} \leftrightarrow \mathbf{s}_j)$ *makes a statement about the visibility of* $\mathbf{s}_{j+1}$ *and* $\mathbf{s}_j$, $\cos\theta_o^{j+1}$ *and* $\cos\theta_i^j$ *represent the angles between the corresponding surface normals at point* $\mathbf{s}_{j+1}$ *and* $\mathbf{s}_j$, *and the exitant, respectively, the incident directions* ω_o^{j+1} *and* ω_i^j *and* $\|\mathbf{s}_{j+1} - \mathbf{s}_j\|_2^2$ *is the distance between* $\mathbf{s}_{j+1}$ *and* $\mathbf{s}_j$. *Linked with the Lebesgue area measure* μ^2, *the original integration measure* $\sigma^\perp$ *can then be replaced by the product of Equation (2.353) and the Lebesgue area measure* μ^2. *That is, we get the SLTEV in a version equivalent to its spherical form from Equation (2.349), namely the SLTEV in 3-point form*

$\mathcal{V}$ (45) μ^2 (82) SLTEV (398) 3-point form SLTEV (402)

$$L(\mathbf{s}_j \to \mathbf{s}_{j-1}) = L_e(\mathbf{s}_j \to \mathbf{s}_{j-1}) + \int_{\partial\mathcal{V}} f_s(\mathbf{s}_{j+1} \to \mathbf{s}_j \to \mathbf{s}_{j-1})\, \mathcal{G}(\mathbf{s}_{j+1} \leftrightarrow \mathbf{s}_j)\, L(\mathbf{s}_{j+1} \to \mathbf{s}_j)\, d\mu^2(\mathbf{s}_{j+1}). \tag{2.354}$$

Obviously, defining the kernel of an integral equation via the product of the

BSDF (371)

BSDF, f_s, and the geometry term, $\mathcal{G}$, this representation of the SLTEV now matches the form of a Fredholm integral equation of the 2^{nd} kind introduced in Definition 2.36.

2.3.1 LINEAR INTEGRAL OPERATOR EQUATIONS

As already noted above, the main problem in global illumination theory lies in the construction of efficient algorithms for solving the *light transport equation* as well as the *importance equation*, both Fredholm integral equations of the 2^{nd} kind. While integral equations, together with differential equations, represent the most important mathematical models of real processes encountered in physics and technology, in mathematics they are also of great interest. Here they serve as the basis of many statements formulated in functional analysis and the inspiration of numerous definitions encountered in this field.

Chapter 5
Fredholm Integral Equation (127)

Fredholm Integral Equation (127)

Let us take a closer look at the Fredholm type integral equation

$$f(x) = g(x) + \int_{\mathcal{R}} k(x,y)\, f(y)\, d\mu(y) \tag{2.355}$$

$\mathcal{L}^2(\mathcal{R},\mu)$ (110)
Linear Operator (53)

with k of $\mathcal{L}^2(\mathcal{R}\times\mathcal{R}, \mu\times\mu)$ and $f \in \mathcal{L}^2(\mathcal{R},\mu)$. It may be seen, that, due to the integration with respect to the 2^{nd} argument, the integral in Equation (2.355) represents a function of $\mathcal{L}^2(\mathcal{R},\mu)$ depending on the variable x. On the basis of the concept of the linear operator, the integral term then may also be regarded as a kind of linear operator:

DEFINITION 2.37 (Linear Integral Operator) *A* linear integral operator $\mathbf{K}$ *is a linear mapping defined on the Lebesgue space* $\mathcal{L}^2(\mathcal{R},\mu)$

$\mathcal{L}^2(\mathcal{R},\mu)$ (110)

$$\mathbf{K} : \mathcal{L}^2(\mathcal{R},\mu) \to \mathcal{L}^2(\mathcal{R},\mu)$$

with

$$(\mathbf{K}f)(x) \stackrel{def}{=} \int_{\mathcal{R}} k(x,y)\, f(y)\, d\mu(y), \tag{2.356}$$

where the kernel k *of the integral operator* $\mathbf{K}$ *is an element of the space* $\mathcal{L}^2(\mathcal{R}\times\mathcal{R}, \mu\times\mu)$.

$\mathcal{L}^2(\mathcal{R}\times\mathcal{R}, \mu\times\mu)$ (110)

Lebesgue Integral (105)
Linear Bounded Operator (55)
Linear Normed Space (860)

Obviously, the linearity of the Lebesgue integral identifies $\mathbf{K}$ as a linear operator between the Lebesgue spaces $\mathcal{L}^2(\mathcal{R},\mu)$. Under the condition of the boundedness of $\mathbf{K}$, this allows the definition of a linear normed space of bounded integral operators $(\mathbf{L}(\mathcal{L}^2,\mathcal{L}^2), \|\cdot\|_{\mathbf{L}})$ equipped with the operator norm from Relation (2.87).

Linear Bounded Operator (55)

A Fredholm equation of the 2^{nd} kind defined via a bounded integral operator from

ear Operator Equation (61) $(\mathbf{L}(\mathcal{L}^2, \mathcal{L}^2), \|\cdot\|_{\mathbf{L}})$, can then easily be written in form of an operator equation, that is,

$$f(x) = g(x) + (\mathbf{K}f)(x) \tag{2.357}$$
$$= (\mathbf{K}'f)(x) \tag{2.358}$$

with

$$(\mathbf{K}'f)(x) \stackrel{\text{def}}{=} g(x) + \int_{\mathcal{R}} k(x,y)\, f(y)\, d\mu(y). \tag{2.359}$$

As this type of equation, a *linear integral operator equation*, has the form of an operator equation—well-known to us from Section 2.1.5—it can be solved, under certain conditions, according to the Banach Fixed-point Theorem, by means of an appropriate iteration procedure as proposed in Lemma 2.1. Banach Fixed-point Theorem (61)

REMARK 2.46 *Integral operator equations of type defined in Equation (2.357) are usually called* operator equations or equations of the 2^{nd} kind. *Often, we also denote this type of an operator equation more specific as an* integral operator equation *or an* integral operator equation of the 2^{nd} kind.

REMARK 2.47 *Although sigularities—an integral equation is singular, if its domain is infinite, the integrand is unbounded somewhere in the domain, the kernel is discontinous, or a combination of some or all of these occurs—play an important role in light transport equations, we will focus our discussions on nonsingular integral equations.*

2.3.2 ADJOINT INTEGRAL EQUATIONS

Let us consider once more the Fredholm type integral equation

$$f(x) = g(x) + \int_{\mathcal{R}} k(x,y)\, f(y)\, d\mu(y), \tag{2.360}$$

which we can transform with the help of the linear integral operator $\mathbf{K}$ from Equation (2.356) into a linear operator equation of type Fredholm Integral Equation (127)

$$f(x) = g(x) + (\mathbf{K}f)(x). \tag{2.361}$$

Obviously, the linear integral operator $\mathbf{K}$ has an adjoint, $\mathbf{K}^*$, given by: Adjoint Operator (60)

$$\mathbf{K}^*(x) \stackrel{\text{def}}{=} \int_{\mathcal{R}} k(y,x)\, f(y)\, d\mu(y). \tag{2.362}$$

This can easily be checked by using Theorem 2.6:

$$\langle f, \mathbf{K}g\rangle \overset{(2.261)}{=} \int_{\mathcal{R}} f(x)\,(\mathbf{K}g)(x)\,d\mu(x) \tag{2.363}$$

$$\overset{(2.356)}{=} \int_{\mathcal{R}} f(x)\left(\int_{\mathcal{R}} k(x,y)\,g(y)\,d\mu(y)\right) d\mu(x) \tag{2.364}$$

$$\overset{(2.356)}{=} \int_{\mathcal{R}} \underbrace{\left(\int_{\mathcal{R}} k(x,y)\,f(x)\,d\mu(x)\right)}_{\mathbf{K}^* f} g(y)\,d\mu(y) \tag{2.365}$$

$$\overset{(2.261)}{=} \langle \mathbf{K}^* f, g\rangle. \tag{2.366}$$

Via the adjoint $\mathbf{K}^*$, we can then introduce the *adjoint* to the Fredholm integral equation of the 2^{nd} kind from (2.360), namely,

$$h(x) = i(x) + \int_{\mathcal{R}} k(y,x)\,h(y)\,d\mu(y), \tag{2.367}$$

$\mathcal{L}^2(\mathcal{R},\mu)$ (110) where the kernel k is an element of the space $\mathcal{L}^2(\mathcal{R}\times\mathcal{R}, \mu\times\mu)$, the unknown function $h \in \mathcal{L}^2(\mathcal{R},\mu)$, and the source function is given by $i \in \mathcal{L}^2(\mathcal{R},\mu)$. It may be seen, that due to the integration with respect to the 1^{th} argument, the integral in Equation (2.367) represents a function of $\mathcal{L}^2(\mathcal{R},\mu)$ depending on the variable x.

EXAMPLE 2.52 *Let us consider the following integral equation*

$$f(x) = 1 + \int_{(0,1]} \ln(xy)\,f(y)d\mu(y). \tag{2.368}$$

The associated operator equation is then given by

$$f(x) = g(x) + (\mathbf{K}f)(x) \tag{2.369}$$

with $g(x) = 1$, *and* $\mathbf{K}(x,y) = \ln(xy)$. *Due to Equation (2.362) the adjoint operator* $\mathbf{K}^*$ *is given by*

$$\mathbf{K}^*(y,x) = \ln(yx), \tag{2.370}$$

that is, an adjoint integral equation can be written as:

$$h(x) \;=\; 1 + \int_{(0,1]} \ln(yx)h(y)d\mu(y). \tag{2.371}$$

From our discussion in Section 2.1.5 it is known, that there exists a solution to Equation (2.360) if it holds: $\|\mathbf{K}\| < 1$, where this solution is given via $(\mathbf{I}-\mathbf{K})^{-1}$. This means, if the adjoint integral operator $\mathbf{K}^*$ is contracting, then $(\mathbf{I}-\mathbf{K}^*)^{-1}$ can also be Adjoint Operator Equation (65) considered as the solution, h, of the adjoint integral operator equation

$$h(x) = i(x) + (\mathbf{K}^* h)(y) \tag{2.372}$$

associated with the integral equation from (2.367).

Now, the fundamental result from Section 2.1.6 for a pair of adjoint equations

$$f(x) = g(x) + (\mathbf{T} f)(x) \tag{2.373}$$

and

$$h(x) = i(x) + (\mathbf{T}^* h)(x) \tag{2.374}$$

was the identity

$$\langle f(x), i(x) \rangle = \langle g(x), h(x) \rangle. \tag{2.375}$$

Applied to the pair (f, h) of adjoints of the Fredholm integral equations of the 2^{nd} kind, it is just this type of an inner product that has to be solved for image synthesis, namely,

$$\begin{aligned} \langle f(x), i(x) \rangle &= \left\langle (\mathbf{I} - \mathbf{T})^{-1} g(x), i(x) \right\rangle && (2.376) \\ &= \left\langle g(x), \left((\mathbf{I} - \mathbf{T})^{-1} \right)^* i(x) \right\rangle && (2.377) \\ &= \left\langle g(x), ((\mathbf{I} - \mathbf{T})^*)^{-1} i(x) \right\rangle && (2.378) \\ &= \langle g(x), h(x) \rangle. && (2.379) \end{aligned}$$

REMARK 2.48 *We have already introduced an inner product of the above type in the introductory chapter, namely: the global illumination problem. In Chapter 5, we will discuss this inner product, that can be used to compute the color of a pixel on the image plane, or the amount of light striking a surface in a scene to be rendered, in more detail.*

2.3.3 ANALYTICAL APPROACHES AND NUMERICAL METHODS FOR SOLVING INTEGRAL OPERATOR EQUATIONS OF THE 2^{nd} KIND

Only very few integral equations derived from practical applications are found to be exactly solvable in practice. In particular this applies to the present case of interest to us here, in which solutions are sought for the integral equations underlying the global illumination problem. Now, there are two fundamental approaches for solving linear integral operator equations:

Global Illumination Problem (6)

- *analytical approaches*, resulting from functional analytical methods for computing the exact solutions of linear operator equations, and

- *numerical approaches*, also derived from functional analysis, but with the focus on finding approximate solutions to linear operator equations.

While analytical approaches—often also called *infinite-dimensional* approaches—attempt to solve an integral operator equation exactly in an infinite dimensional function space, numerical approaches consider an operator equation as living in a finite dimensional space. So, they are also capable to deliver only finite dimensional, approximate solutions of the problem.

Section 2.3.3.2 Under numerical approaches for solving Fredholm type integral equations, we understand deterministic algorithms that are based on *quadrature*, *finite basis* and *projection methods*. In particular, they are based on the concept of the *finite element method*. Analytical solution approaches, as we will presente them in the next section, play no central role in numerical algorithms. Rather, they serve as basis of probabilistic algorithms for solving Fredholm type integral equations, which will be discussed in Chapter 6.

2.3.3.1 ANALYTICAL APPROACHES FOR SOLVING INTEGRAL OPERATOR EQUATIONS OF THE 2^{nd} KIND

From the multitude of analytical procedures for solving integral operator equations, we pick out two important deterministic ideas from functional analysis:

Section 2.3.3.1.1
- the *Neumann series approach*, and

Section 2.3.3.1.1
- the *method of successive substitution*.

Both techniques are based on different approaches, but lead to the same result, namely, the infinite-dimensional solution of an operator equation underlying a Fredholm integral equation of the 2^{nd} kind. In addition, they guarantee, provided that certain conditions are valid, the existence of approximate solutions techniques for linear integral equations that are not solvable in practice.

While the Neumann series approach is based on the representation of the solution of an operator equation as an infinite series whose powers are build over a contracting operator, the method of successive substitution results in a recursive sequence, constructed over a contracting operator, where the convergence of the sequence is guaranteed by the Banach Fixed-point Theorem.

Banach Fixed-point Theorem (61)

REMARK 2.49 *Although these both analytical approaches are not directly applicable in procedures for solving Fredholm equations, yet they play an import role in the process of finding the unknown function* f *around them an integral equation is build. To see the real power of these approaches, we must be patient up to the introduction of probabilistic methods for solving integral equations.*

Section 6.7

2.3.3.1.1 THE NEUMANN SERIES APPROACH

As the kernel of an integral equation can be of an arbitrary complex nature—the kernels of integral equations, deeply rooted in real world, can have infinitely many discontinuities—often, we have no chance of finding an exact and general valid solution. But for theoretical reasons it is useful to know such a solution, since it can help to draw conclusions to an approximate solution for the given problem. So, our goal is to find the exact solution of the integral operator equation

Integral Kernel (127)

Integral Operator Equation (130)

$$f(x) = g(x) + \mathbf{K}f(x), \tag{2.380}$$

which can also be written as:

$$\begin{aligned} g(x) &= f(x) - \mathbf{K}f(x) && (2.381)\\ &= (\mathbf{I} - \mathbf{K})f(x), && (2.382) \end{aligned}$$

where $\mathbf{K}$ is a linear bounded integral operator on $\mathcal{L}^2(\mathcal{R}, \mu)$ with $\|\mathbf{K}\| < 1$.

Linear Bounded Operator (55)

$\mathcal{L}^2(\mathcal{R}, \mu)$ (110)

Operator Norm (56)

From Lemma 2.1 in Section 2.1.5 it is known, that $\|\mathbf{K}\| < 1$ implies the existence of the inverse operator $(\mathbf{I} - \mathbf{K})^{-1}$ to $(\mathbf{I} - \mathbf{K})$, which is given by the *Neumann series*, $\mathbf{M}$, that is,

$$\begin{aligned} \mathbf{M} &\stackrel{def}{=} (\mathbf{I} - \mathbf{K})^{-1} && (2.383)\\ &= \mathbf{I} + \mathbf{K} + \mathbf{K}^2 + \cdots && (2.384)\\ &= \sum_{i=0}^{\infty} \mathbf{K}^i. && (2.385) \end{aligned}$$

Multiplying both sides of Equation (2.382) by the inverse operator $(\mathbf{I} - \mathbf{K})^{-1}$ then yields:

$$\begin{aligned} f(x) &= (\mathbf{I} - \mathbf{K})^{-1} g(x) && (2.386)\\ &\stackrel{(2.385)}{=} \sum_{i=0}^{\infty} \mathbf{K}^i \, g(x). && (2.387) \end{aligned}$$

Thus, a solution to the above integral operator equation may be represented in form of the Neumann series, whose terms are composed of the powers of the operator $\mathbf{K}$ and the source function g. So, we can conclude that, under the condition, that the inverse integral operator $(\mathbf{I} - \mathbf{K})^{-1}$ exists in the form of a Neumann series, a solution to the operator equation from (2.380), and thus also for the associated integral equation, may be found easilys by means of the source function, $g(x)$, and the powers of the integral operator $\mathbf{K}$.

Integral Operator Equation (131)

Integral Operator (130)

Linear Operator Equation (61)

Let us check the correctness of this idea by means of a simple example.

EXAMPLE 2.53 *Given be the integral equation*

$$f(x) = 1 + \frac{1}{4}\int_{[0,\frac{\pi}{2}]} \cos x\, f(y)\, d\mu(y). \tag{2.388}$$

With $\lambda = \frac{1}{4}, g(x) = 1$, *and* $\mathbf{K} = \lambda \cos x$, *the associated integral operator equation then looks like this:*

$$f(x) = 1 + \mathbf{K}f(x), \tag{2.389}$$

which implies:

$$\begin{aligned}
f(x) &\overset{(2.387)}{=} \sum_{i=0}^{\infty} \mathbf{K}^i && (2.390)\\
&= 1 + \frac{1}{4}\int_{[0,\frac{\pi}{2}]} \cos x\, d\mu(y) + \frac{1}{16}\int_{[0,\frac{\pi}{2}]}\int_{[0,\frac{\pi}{2}]} \cos x \cos x_1\, d\mu(y)d\mu(x_1) + \dots \\
&= 1 + \cos x\Bigg(\underbrace{\frac{1}{4}\int_{[0,\frac{1}{2}]} d\mu(y)}_{\frac{\pi}{2}} + \underbrace{\frac{1}{16}\int_{[0,\frac{\pi}{2}]}\int_{[0,\frac{\pi}{2}]} \cos x_1\, d\mu(y)d\mu(x_1)}_{\frac{\pi}{2}} + \\
&\qquad \underbrace{\frac{1}{64}\int_{[0,\frac{\pi}{2}]}\int_{[0,\frac{\pi}{2}]}\int_{[0,\frac{\pi}{2}]} \cos x_1 \cos x_2\, d\mu(y)d\mu(x_1)d\mu(x_2))}_{\frac{\pi}{2}} + \dots\Bigg) && (2.391)\\
&= 1 + \frac{\pi}{2}\cos x \underbrace{\sum_{i=1}^{\infty} \frac{1}{4^i}}_{\frac{1}{3}} && (2.392)\\
&= 1 + \frac{\pi}{6}\cos x. && (2.393)
\end{aligned}$$

Replacing $f(x)$ *in Equation (2.388) by* $1 + \frac{\pi}{6}\cos x$ *then we get:*

$$\begin{aligned}
1 + \frac{\pi}{6}\cos x &= 1 + \frac{1}{4}\int_{[0,\frac{\pi}{2}]} \cos x \left(1 + \frac{\pi}{6}\cos y\right) d\mu(y) && (2.394)\\
&= 1 + \frac{1}{4}\int_{[0,\frac{\pi}{2}]} \cos x\, d\mu(y) + \frac{\pi}{24}\int_{[0,\frac{\pi}{2}]} \cos x \cos y\, d\mu(y) && (2.395)\\
&= 1 + \frac{\pi}{8}\cos x + \frac{\pi}{24}\cos x && (2.396)\\
&= 1 + \frac{\pi}{6}\cos x. && (2.397)
\end{aligned}$$

Since any computer is a deterministic machine, the calculation of the value of a Neumann series by means of a computer must be done via a numerical procedure. Such a

method approximates the value of the series by using only a finite number of infinite series terms is accounted for. For estimating the absolute error of this approximation from the true value of the Neumann series let us define a sequence of linear operators $\mathbf{M}_n$ by:

$$\mathbf{M}_n \stackrel{def}{=} \sum_{i=0}^{n} \mathbf{K}^i, \tag{2.398}$$

that is, $\mathbf{M}_n$ returns the value of the Neumann series truncated after $n+1$ terms. The absolute error of the approximation can now be estimated as:

$$\begin{aligned}
\|\mathbf{M} - \mathbf{M}_n\| &= \left\| \sum_{i=0}^{\infty} \mathbf{K}^i - \sum_{i=0}^{n} \mathbf{K}^i \right\| && (2.399)\\
&= \left\| \sum_{i=n+1}^{\infty} \mathbf{K}^i \right\| && (2.400)\\
&\stackrel{\Delta-\text{inequality}}{\leq} \sum_{i=n+1}^{\infty} \|\mathbf{K}^i\| && (2.401)\\
&\stackrel{\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|}{\leq} \sum_{i=n+1}^{\infty} \|\mathbf{K}\|^i && (2.402)\\
&= \sum_{i=0}^{\infty} \|\mathbf{K}\|^i - \sum_{i=0}^{n} \|\mathbf{K}\|^i && (2.403)\\
&= \frac{1}{1-\|\mathbf{K}\|} - \frac{\|\mathbf{K}\|^{n+1}-1}{\|\mathbf{K}\|-1} && (2.404)\\
&= \frac{\|\mathbf{K}\|^{n+1}}{1-\|\mathbf{K}\|}. && (2.405)
\end{aligned}$$

Here, the first inequality is based on the triangle inequality of the operator norm, the second follows from $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$ with respect to two operators $\mathbf{A}$ and $\mathbf{B}$, and we have used the formula $\sum_{i=0}^{n} q^i = \frac{1-q^{n+1}}{1-q}$ for a finite geometric series to estimate the sum $\sum_{i=0}^{n} \|\mathbf{K}\|^i$. Operator Norm (56)

While an estimation for Equation (2.405), measured in the operator norm, supplies a statement on the quality of the operator $\mathbf{M}_n$ for approximating the exact operator $\mathbf{M}$, it says nothing about the quality of an approximated solution of Equation (2.387). For that purpose, let us define an approximate solution of the Neumann series by:

$$f_n \stackrel{def}{=} \mathbf{M}_n g, \tag{2.406}$$

then a statement on the quality of this approximate solution follows directly from the

residual from Relation (2.405), namely:

$$\begin{aligned} \|f - f_n\| &= \|\mathbf{M}g - \mathbf{M}_n g\| && (2.407)\\ &= \|(\mathbf{M} - \mathbf{M}_n)\, g\| && (2.408)\\ &\leq \|\mathbf{M} - \mathbf{M}_n\|\|g\| && (2.409)\\ &\overset{(2.405)}{\leq} \frac{\|\mathbf{K}\|^{n+1}}{1 - \|\mathbf{K}\|}\|g\|. && (2.410) \end{aligned}$$

This implies, that for sufficiently large n, the sequence $(f_n)_{n\in\mathbb{N}_0}$ will converge to the exact solution of Equation (2.380). That is, with the Neumann series approach we have a first numerical method for solving Fredholm integral equations of the 2^{nd} kind.

Section 2.3

2.3.3.1.2 THE METHOD OF SUCCESSIVE SUBSTITUTION

Let us now pursue the approach of *successive substitution* where the solution of the integral equation is formulated as an infinite series of single and multiple integrals. As the unknown function f is replaced by the known source function g the evaluation of several multiple integrals is possible and easily computable.

Applied to the associated integral operator equation, the unknown function f on the right-hand side of Equation (2.380) is replaced by $g + \mathbf{K}f$. This, then implies

$$\begin{aligned} f(x) &= g(x) + \mathbf{K}f(x) && (2.411)\\ &= g(x) + \mathbf{K}\underbrace{(g(x) + \mathbf{K}f(x)}_{f(x)} && (2.412)\\ &= g(x) + \mathbf{K}\underbrace{(g(x) + \mathbf{K}\underbrace{(g(x) + \mathbf{K}f(x))}_{f(x)})}_{f(x)} && (2.413)\\ &= \cdots. && (2.414) \end{aligned}$$

Obvioulsy, the function f may then be written in form of the following recurrence equation

$$\begin{aligned} f_0(x) &= g(x)\\ f_{n+1}(x) &= \mathbf{K}f_n(x) + g(x), \qquad n \geq 0. && (2.415) \end{aligned}$$

Banach Fixed-point Theorem (61)

Based on the Banach Fixed-point Theorem for contracting linear operators the following then applies for the distance of members of the sequence $(f_n)_{n\in\mathbb{N}_0}$ for sufficiently large n:

$$\begin{aligned} \|f_{n+m} - f_m\| &\overset{(2.115)}{\leq} \|\mathbf{K}\|^n \frac{1}{(1 - \|\mathbf{K}\|)}\|\mathbf{K}g\| && (2.416)\\ &\leq \frac{\|\mathbf{K}\|^{n+1}}{1 - \|\mathbf{K}\|}\|g\|. && (2.417) \end{aligned}$$

Cauchy Sequence (35)

Obviously, the sequence $(f_n)_{n\in\mathbb{N}_0}$ is a Cauchy sequence, which, under the condition of the contraction of the integral operator $\mathbf{K}$, due to Equation (2.118), converges towards the exact solution of a Fredholm integral equation of the 2^{nd} kind.

2.3.3.2 NUMERICAL METHODS FOR SOLVING INTEGRAL OPERATOR EQUATIONS OF THE 2^{nd} KIND

The current section gives a brief overview of the most important numerical procedures for solving linear operator equations underlying Fredholm integral equations of the 2^{nd} kind. Since it is not possible here to present all existing numerical methods for solving Fredholm type integral equations, our discussion will focus only on the following categories of solution methods:

Fredholm Integral Equation (127)

- *quadrature methods* as well as
- *finite basis*, and *projection methods*.

While quadrature methods replace the process of integration inside a Fredholm integral equation by a finite weighted sum of the involved kernel at predetermined points of the integration domain, finite basis and projection methods try to find an approximate solution of the integral equation in a finite-dimensional subspace, for example in the space, $\mathcal{P}_n$, of polynomials of degree $n-1$.

$\mathcal{P}_n$ (855)

Following [10, Arvo 1995], we present in the next two sections for each of the above mentioned methods a single representative, where we are interested in particular in those numerical solution methods which promise to be useful for solving the global illumination problem.

Global Illumination Problem (6)

2.3.3.2.1 QUADRATURE METHOD

The *quadrature method* utilizes the similarity between the kernel of an integral operator and its finite-dimensional analog, the matrix. The idea behind the quadrature method is to approximate the integral operator $\mathbf{K}$ in

Linear Integral Operator (130)
Matrix (853)

$$f(x) = g(x) + \mathbf{K}f(x) \tag{2.418}$$

by means of a quadrature rule which leads to a linear system of equations. For that, first we choose points $x_1, x_2, \ldots, x_n$ and wish to evaluate the function f at these points.

Replacing the operator $\mathbf{K}$ in Equation (2.418) by a quadrature rule, $\mathbf{Q}$, of the form

$$\mathbf{Q}f(x) = \sum_{i=1}^{n} w_i(x)\, k(x, t_i)\, f(t_i), \tag{2.419}$$

where $w_i(x)$ are weight functions, then leads to an approximation of f at the chosen points $x_1, \ldots, x_n$, namely,

$$f(x_j) \overset{(2.422)}{\approx} g(x_j) + \sum_{i=1}^{n} w_i(x_j)\, k(x_j, t_i)\, f(x_i) \tag{2.420}$$

$$\overset{t_i = x_i}{\approx} g(x_j) + \sum_{i=1}^{n} w_i(x_j)\, k(x_j, x_i)\, f(x_i), \tag{2.421}$$

where we have evaluated the kernel $k(x, t_i)$ at the points $t_i = x_i$ for $x = x_j$.

Now, the operator equation from (2.418) can be expressed in terms of the unknowns $y_j = f(x_j)$ and $y_i = f(x_i)$ by:

$$y_j = f(x_j) \approx g(x_j) + \sum_{i=1}^{n} w_i(x_j)\, k(x_j, x_i)\, y_i, \quad j = 1, 2, \ldots, n. \tag{2.422}$$

Applying this strategy for computing Equation (2.418) at all points $x_1, x_2, \ldots, x_n$, then we obtain a linear system of n equations, in the unknowns $y_1, y_2, \ldots, y_n$, namely:

$$y_j = g(x_j) + \sum_{i=1}^{n} w_i(x_j)\, k(x_j, x_i)\, y_i, \qquad 1 \le i, j \le n, \tag{2.423}$$

Linear Operator Equation (61) which can also be written as a linear operator equation

$$\mathbf{y} = \mathbf{g} + \mathbf{WK}\mathbf{y}. \tag{2.424}$$

Here $\mathbf{y}$ and $\mathbf{g}$ are n-dimensional vectors representing the unknowns as well as the given source function $\mathbf{g}$ at points x_j and $\mathbf{WK}$ is a $n \times n$ matrix whose coefficients are
Integral Kernel (127) given by the products of the weights and the integral kernel at points x_j and x_i, thus $w_i(x_j), k(x_j, x_i)_{1 \le i,j \le n}$. Slightly rephrased, then this operator equation has the form:

$$\begin{aligned} \mathbf{y} - \mathbf{WK}\,\mathbf{y} &= \mathbf{g} && (2.425)\\ (\mathbf{I} - \mathbf{WK})\,\mathbf{y} &= \mathbf{g} && (2.426)\\ \mathbf{y} &= (\mathbf{I} - \mathbf{WK})^{-1}\mathbf{g}, && (2.427) \end{aligned}$$

Linear Operator Equation (61) which can now be written in matrix-vector notation as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 - w_{11}k_{11} & -w_{21}k_{12} & \ldots & -w_{n1}k_{1n} \\ -w_{12}k_{21} & 1 - w_{22}k_{22} & \ldots & -w_{n2}k_{2n} \\ \vdots & \vdots & & \vdots \\ -w_{1n}k_{n1} & -w_{2n}k_{n2} & \ldots & 1 - w_{nn}k_{nn} \end{pmatrix}^{-1} \cdot \begin{pmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_n) \end{pmatrix}. \tag{2.428}$$

So, an approximate solution to a linear Fredholm integral equation of the 2^{nd} kind may also be found via the solution of a linear system of equations in a finite number of
Chapter 10 variables.

ear Integral Operator (130)

REMARK 2.50 *The above method of approximating an integral operator via its finite-dimensional analog, namely a matrix, is also known in the literature as the* Nyström method. *It has the great advantage that it avoids the process of integration for calculating the coefficients of the linear system of equations.*

Although Nyström procedures are very efficient algorithms for solving integral equations, due to the nature of the integral kernels involved in the global illumination equation and the associated quadrature rules, not much use will be made of them here. Instead, we will take recourse to the Neumann series approach and the procedures presented below which are build on finite basis methods.

2.3.3.2.2 FINITE BASIS AND PROJECTION METHODS

Finite basis and projection methods pursue the idea of approximating an infinite dimensional function space by a finite-dimensional subspace $\mathcal{U}_n$. The goal is to find a function $\widehat{f} \in \mathcal{U}_n$ that is in some sense a good approximation of f.

Section 2.1.1

Subspace (855)

Let us show this approach for finding approximate solutions to linear integral equations by means of two different methods:

i) the *collocation method* and

ii) the *Galerkin method*,

where we assume, that $\mathcal{U}_n$ is a finite-dimensional subspace of $\mathcal{L}^2(\mathcal{R}, \mu)$ spanned by a finite set of basis functions $\{\phi_1, \phi_2, \ldots, \phi_n\}$.

$\mathcal{L}^2(\mathcal{R}, \mu)$ (110)

Basis (857)

THE COLLOCATION METHOD. Given be the integral operator equation of the 2^{nd} kind, thus an equation of the form

$$f(x) = g(x) + \mathbf{K} f(x), \tag{2.429}$$

which also can be written as:

Linear Operator Equation (61)

$$g(x) = (\mathbf{I} - \mathbf{K}) f(x). \tag{2.430}$$

Now, the *collocation method* suggests to approximate the right-hand side of Equation (2.430) by a function $\widehat{f} \in \mathcal{U}_n$ that can be expressed in terms of the basis function $\phi_1, \ldots, \phi_n$ namely,

$$\widehat{f}(x) = (\mathbf{I} - \mathbf{K}) \sum_{i=1}^{n} \alpha_i \phi_i(x) \tag{2.431}$$

with $\alpha_i \in \mathbb{R}, 1 \leq i \leq n$. In addition, we require, that $\widehat{f}$ agrees with g at n so-called *collocation points* $x_1, x_2, \ldots, x_n \in \mathcal{R}$, that is,

$$\widehat{f}(x_j) \stackrel{def}{=} g(x_j). \tag{2.432}$$

Under these conditions, evaluated at the collocation point x_j, Equation (2.430) can now be written as:

$$g(x_j) \overset{(2.431)}{=} (\mathbf{I}-\mathbf{K}) \sum_{i=1}^{n} \alpha_i \, \phi_i(x_j) \tag{2.433}$$

$$= \sum_{i=1}^{n} \alpha_i \underbrace{(\mathbf{I}-\mathbf{K})\phi_i(x_j)}_{\widehat{\phi}_i(x_j)} \tag{2.434}$$

$$= \sum_{i=1}^{n} \alpha_i \widehat{\phi}_i(x_j). \tag{2.435}$$

Since Relation (2.432) must hold at n points, we obtain a linear system of n equations in the unknowns α_i, that is,

$$g(x_j) = \sum_{i=1}^{n} \alpha_i \, \widehat{\phi}_i(x_j), \tag{2.436}$$

Linear Operator Equation (61)

which can be rephrased in form of an operator equation, thus,

$$\mathbf{g} = \widehat{\mathbf{\Phi}} \, \boldsymbol{\alpha}. \tag{2.437}$$

Here $\mathbf{g}$ is an n-dimensional vector representing the values of function g at the collocation points, $\boldsymbol{\alpha}$ corresponds to the vector of unknowns, and $\widehat{\mathbf{\Phi}}$ is a $n \times n$ matrix, whose coefficients are given by the basis function $\widehat{\phi}_i$ evaluated at the collocation points. If $\widehat{\mathbf{\Phi}}$ is invertible, then we obtain:

$$\widehat{\mathbf{\Phi}}^{-1} \mathbf{g} = \boldsymbol{\alpha} \tag{2.438}$$

written in matrix form, thus,

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \widehat{\phi}_1(x_1) & \dots & \widehat{\phi}_n(x_1) \\ \vdots & & \vdots \\ \widehat{\phi}_1(x_n) & \dots & \widehat{\phi}_n(x_n) \end{pmatrix}^{-1} \cdot \begin{pmatrix} g(x_1) \\ \vdots \\ g(x_n) \end{pmatrix}. \tag{2.439}$$

This implies, that a solution of Equation (2.436) delivers the coefficients $\alpha_1, \alpha_2, \dots, \alpha_n$ of the basis functions $\phi_1, \dots, \phi_n$ for approximating the unknown f, and thus, an approximate solution of the operator equation from Equation (2.430).

REMARK 2.51 *As we will see in Chapter 10, in contrast to the Nyström method, the collocation method is often used in global illumination. The problem that comes with collocation is the evaluation of the coefficients $\widehat{\phi}_i(x_j)$. As they are the images of the basis functions ϕ_i under the operator $\mathbf{I}-\mathbf{K}$, where $\mathbf{K}$ represents a complex evaluation of an integral. So, the computation of the coefficients $\widehat{\phi}_i(x_j)$ can be extremely difficult for the most basis function.*

olm Integral Equation (127)

Let us now apply the collocation method to a Fredholm integral equation of the 2^{nd} kind, which describes the exchange of radiant energy between surfaces, namely, the radiative transfer in the absence of a participating medium.

EXAMPLE 2.54 (A Simple Radiosity Approach) *Let us consider the following Fredholm integral equation of the* 2^{nd} *kind* Fredholm Integral Equation (127)

$$L_o(s,\omega_o) = L_e(s,\omega_o) + \int_{\mathcal{H}_i^2(s)} f_r(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i), \tag{2.440}$$

where the functions L_o *and* L_e *are exitant radiance functions and the integral kernel is given by the BRDF* f_r*. In this context,* L_o *and* L_e *can be interpreted as the radiance outgoing as well as emitted at a surface point* s *in direction* ω_o *and* f_r *returns the amount of exitant light after reflected at point* s*.*

We will show how it is possible to solve this integral equation by means of the collocation method, for more details see [8, Arvo 1993].

Now, in the collocation method the matrix coefficients $\widehat{\phi}_i(x_j)$ *are defined due to Equation (2.434) by:*

$$\widehat{\phi}_i(x_j) = (\mathbf{I}-\mathbf{K})\phi_i(x_j) \tag{2.441}$$

$$= \phi_i(x_j) - (\mathbf{K}\phi_i)(x_j). \tag{2.442}$$

That is, for generating the linear system from Relation (2.436) there are three operations that need to be performed: the evaluations of the basis function at the collocation points, thus, $\phi_i(x_j)$ *and the computation of the images of the basis function under the operator* $\mathbf{I}-\mathbf{K}$*, that is,* $(\mathbf{K}\phi_i)(x_j)$*. Based on these values then the source function* $g(x_j)$ *has to evaluated.*

Applied to our integral equation from above, now we break the set of surfaces $\partial\mathcal{V}$ *into* n *disjoint patches* $P_1,\ldots,P_n$*, such as a grid or a Voronoi diagram. For further simplification, we also assume that these patches are pure diffuse reflectors, that is, the BRDF* f_r *is constant, thus,* $f_r = C$*. The choice of a set of rays* $(s_j,\omega_o^j), 1 \le j \le n$ *as our collocation points* x_j *and the choice of basis functions* ϕ_i *given by:* Voronoi Diagram (577) Diffuse Reflector (339)

$$\phi_i(s_j,\omega_o^j) \stackrel{def}{=} \begin{cases} 1 & \text{if } s_j \in P_i \\ 0 & \text{otherwise,} \end{cases} \tag{2.443}$$

then implies, that the finite-dimensional subspace $\mathcal{U}_n$ *is spanned by the basis* $\phi_1,\ldots,\phi_n$*,* Subspace (855)
where only point s_j *lies within the support of function* ϕ_i*.*

Under these conditions, the radiance at point s_j *in direction to* ω_o^j *due to a single reflection of the radiant energy emitted according to the basis function* ϕ_i *can be described by:* Radiance (250)

$$(\mathbf{K}\phi_i)(s_j,\omega_o^j) = \int_{\mathcal{H}_i^2(s_j)} f_r(s_j,\omega_i \to \omega_o^j)\, \phi_i(s_j,\omega_o^j)\, d\sigma_{s_j}^2(\omega_o^j) \tag{2.444}$$

$$= C\int_{\mathcal{H}_i^2(s_j)} \phi_i(s_j,\omega_o^j)\, d\sigma_{s_j}^2(\omega_o^j) \tag{2.445}$$

We will encounter this equation in Chapter 10, but in a different form, where the integration goes over the set of all patches existing within the scene, not over the upper hemisphere about a surface point.

THE GALERKIN METHOD. With the so-called *Galerkin method* we now present a projection method that in the field of computer graphics serves as the basis of numerous algorithms for solving a particular version of the global illumination equation.

Chapter 10

In Section 2.1.4 we introduced the concept of orthogonality under the condition, that a linear space is a Hilbert space. This concept allows to project functions given in infinite-dimensional linear spaces onto functions defined in finite-dimensional spaces. Therefore, we will now consider a projection operator $\mathbf{Pr}_n$ defined on the Hilbert-space $\mathcal{L}^2(\mathcal{R},\mu)$ with values in a subspace $\mathcal{U}_n \leq \mathcal{L}^2(\mathcal{R},\mu)$. The aim of the Galerkin-procedure is the construction of an approximation $\widehat{f} \in \mathcal{U}_n$ of Equation (2.430), for which the following applies:

Hilbert Space (36)
Projection Operator (58)
$\mathcal{L}^2(\mathcal{R},\mu)$ (110)

$$\left(\widehat{f}(x) - g(x)\right) \perp \mathcal{U}_n \qquad \Leftrightarrow \qquad \left((\mathbf{I}-\mathbf{K})f_n(x) - g(x)\right) \perp \mathcal{U}_n. \tag{2.446}$$

This implies that the *residual function* r defined via

$$r(x) \stackrel{def}{=} (\mathbf{I}-\mathbf{K})f_n(x) - g(x), \tag{2.447}$$

lies in the orthogonal complement of $\mathcal{U}_n$. The following then clearly applies to the image of the residual function under the projection operator $\mathbf{Pr}_n$ onto $\mathcal{U}_n$:

$$\mathbf{Pr}_n\left((\mathbf{I}-\mathbf{K})f_n(x) - g(x)\right) = 0, \tag{2.448}$$

Projection Operator (58)

or with $\mathbf{Pr}_n f_n = f_n$

$$(\mathbf{I}-\mathbf{Pr}_n\mathbf{K})f_n(x) = \mathbf{Pr}_n g(x). \tag{2.449}$$

Basis (857)
Subspace (855)

Since the set of functions $\{\phi_1,\ldots,\phi_n\}$ forms a basis of the subspace $\mathcal{U}_n$, due to the orthogonality condition from (2.446), Equation (2.446) may also be written as:

$$\left\langle\left((\mathbf{I}-\mathbf{K})f_n(x) - g(x)\right), \phi_j(x)\right\rangle = 0, \qquad j = 1,2,\ldots,n \tag{2.450}$$

Inner Product (859)

or, making use of the bilinearity of the inner product, equivalently to:

$$\left\langle\Big(\underbrace{(\mathbf{I}-\mathbf{K})f_n(x)}_{\widehat{f}_n(x)}, \phi_j(x)\right\rangle = \left\langle g(x), \phi_j(x)\right\rangle. \tag{2.451}$$

Defining f_n via a linear combination of the basis function $\phi_1,\ldots,\phi_n$ as an element of the finite-dimensional function sub space $\mathcal{U}_n \leq \mathcal{R}$, thus,

$$f_n(x) \stackrel{def}{=} \sum_{i=1}^{n} \alpha_i \phi_i(x), \tag{2.452}$$

then leads for $j = 1, 2, \ldots, n$ to:

$$\left\langle (\mathbf{I} - \mathbf{K}) \sum_{i=1}^{n} \alpha_i \phi_i(x), \phi_j(x) \right\rangle = \langle g(x), \phi_j(x) \rangle . \tag{2.453}$$

Slightly reformulated we get:

$$\sum_{i=1}^{n} \alpha_i \Big\langle \underbrace{(\mathbf{I} - \mathbf{K})\phi_i(x)}_{\widehat{\phi}_i(x)}, \phi_j(x) \Big\rangle = \big\langle g(x), \phi_j(x) \big\rangle . \tag{2.454}$$

Let $\boldsymbol{\alpha}$ be the n-dimensional vector of unknowns and $\mathbf{g\Phi}$ be the vector of the inner products of the functions g and ϕ_j at point x, then with $\langle \widehat{\phi}_i, \phi_j \rangle_{1 \le i,j \le n}$ as the coefficients of the matrix, $\widehat{\mathbf{\Phi}}\mathbf{\Phi}$, the above equations can be written as a linear operator equation, that is,

$$\widehat{\mathbf{\Phi}}\mathbf{\Phi} \cdot \alpha = \mathbf{g\,\Phi} \quad \Rightarrow \quad \boldsymbol{\alpha} = \left(\widehat{\mathbf{\Phi}}\mathbf{\Phi} \right)^{-1} \cdot \mathbf{g\,\Phi}, \tag{2.455}$$

or expressed in matrix notation as:

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \langle \widehat{\phi}_1(x), \phi_1(x) \rangle & \ldots & \langle \widehat{\phi}_n(x), \phi_1(x) \rangle \\ \vdots & & \vdots \\ \langle \widehat{\phi}_1(x), \phi_n(x) \rangle & \ldots & \langle \widehat{\phi}_n(x), \phi_n(x) \rangle \end{pmatrix}^{-1} \cdot \begin{pmatrix} \langle g(x), \phi_1(x) \rangle \\ \vdots \\ \langle g(x), \phi_n(x) \rangle \end{pmatrix} . \tag{2.456}$$

Similar to the collocation method, the scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$ used as the coefficients of the basis functions for approximating f may be obtained as the solution of the linear system of equations associated with the Galerkin method.

REMARK 2.52 *As we will see in Chapter 10, the Galerkin method is the numerical procedure normally used for approximating the* radiosity equation, *like the light transport equation, a Fredholm integral equation of the* 2^{nd} *kind.* Radiosity Equation (19)

Like the collocation method, so even the Galerkin approach suffers from the complexity of the evaluation of the matrix coefficients $\langle \widehat{\phi}_i(x), \phi_j(x) \rangle_{1 \le i,j \le n}$, *that involves to compute the image of the basis functions* ϕ_i *under the operator* $\mathbf{I} - \mathbf{K}$.

REMARK 2.53 *The above described finite basis and projection methods applied to Fredholm integral equations of the* 2^{nd} *kind give rise to a number of interesting rendering procedures used in computer graphics, the so-called* radiosity procedures. *Restricting the global illumination equation in a way like we did it in Example 2.54, a large number of efficient procedures based on projection methods may be developed for the solution of specific variants of the global illumination equation [13, Ashdown 1994], [190, Sillion and Puech 1994], [36, Cohen and Wallace 1993] and [68, Glassner 1995].* Chapter 10

REMARK 2.54 *Unfortunately, due to the complex geometry of the environment and the potential discontinuity of the involved kernels, integral equations of the above type may not be solved analytically. In the search for a solution we will therefore be forced to submit the problem underlying this equation to certain restrictions or, alternatively, to fall back upon methods based on probabilistic approaches. Both alternatives will be presented, the first leading us to the radiosity procedure presented briefly in the above example. Regarding procedures for the solution of integrals and integral equations based on probabilistic approaches, the focus in this work will be especially on Monte Carlo methods of applied mathematics.*

Chapter 10

Chapter 6

2.3.3.2.3 THE FINITE ELEMENT METHOD FOR SOLVING FREDHOLM INTEGRAL EQUATIONS OF THE SECOND KIND

Section 2.3.3.2.2

The problem that comes with the Galerkin method is the determination of a suitable set of basis functions $\{\phi_1, \ldots, \phi_n\}$ for the space $\mathcal{U}_n$. In practice, this can be extremely difficult in particular in cases, where the domain of the integration does not have a simple shape. But here, the *finite element method*, also briefly denoted by *FEM*, can help. It circumvents this problem by choosing basis functions that are piecewise polynomials and that are nonzero only on a relatively small part of the integration domain. So, finite element methods can handle domains of fairly arbitrary shapes.

For that, the finite element method partitions the integration domain in a finite set of subdomains. All these subdomains have a finite area—not an infinitesimally small or large area—this gives the method also its name. In a second step, a FEM algorithm then constructs corresponding basis functions on these subdomains. Applied to integral equations, the unknown function is approximated by a finite linear combination of the chosen basis function—as we did it in the Galerkin method. This then leads to a linear system of equations, that can be solved via one of the methods presented in the next section.

Section 2.3.3.2.2

Let us now describe the two main steps of a finite element method and show how it works for the Galerkin method for solving linear, inhomogeneous integral equations of the 2^{nd} kind.

Fredholm Integral Equation (127)

REMARK 2.55 *For a better understanding of the discussion in the following it could be a good idea to interpret the integration domain $\mathcal{R}$ of an integral equation as a bounded subset of $\mathbb{R}$ or the two-dimensional Euclidean space $\mathbb{R}^2$. So, we will do this in all the figures within this section.*

Fredholm Integral Equation (127)

THE FINITE ELEMENT MESH. Given be the Fredholm type integral equation

$$f(x) = g(x) + \int_{\mathcal{R}} k(x, y)\, f(y)\, d\mu(y) \tag{2.457}$$

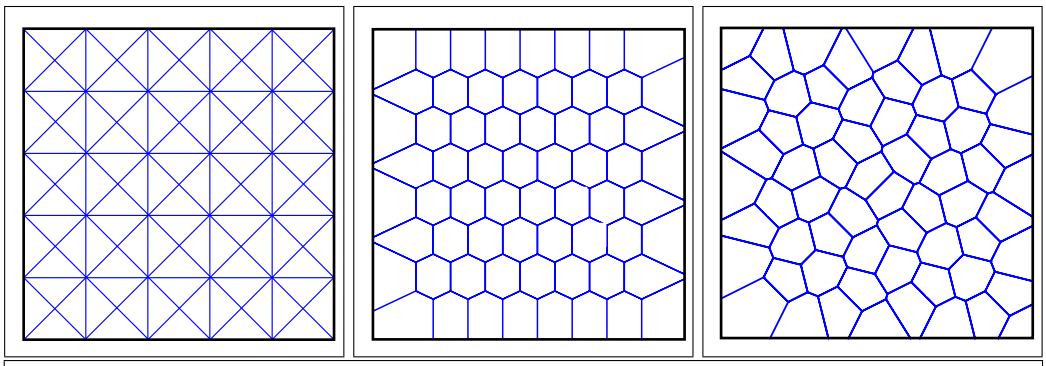


FIGURE 2.38: A SQUARE AND ITS SUBDIVISION INTO FINITE ELEMENT MESHES. The square $[\mathbf{a}, \mathbf{b}] \subset \mathbb{R}^2$ is partitioned into a finite set of disjoint triangles or other polygonal shapes.

$\mathcal{L}^2(\mathcal{R}, \mu)$ (110) with k of $\mathcal{L}^2(\mathcal{R} \times \mathcal{R}, \mu \times \mu)$ and $f \in \mathcal{L}^2(\mathcal{R}, \mu)$.

Initially, any FEM algorithm partitions the domain $\mathcal{R}$ into a finite number m of disjoint subdomains $P_1, P_2, \ldots, P_m$, also called a *finite element mesh* or a set of *patches*. It is required that these patches satisfy the following conditions:

$$P_i \cap P_k = \emptyset \quad \text{for} \quad i \neq k, \quad \text{and} \quad \bigcup_{k=1}^{m} \overline{P}_k = \overline{\mathcal{R}}, \tag{2.458}$$

that is, the open subdomains P_k are disjoint, and the closure $\overline{\mathcal{R}}$ can be written as a finite union of the closed subsets $\overline{P}_k$, thus, the open subdomains together with their boundaries. Closed Set (864) Open Set (864)

EXAMPLE 2.55 *Let us assume $\mathcal{R}$ be the square $[\mathbf{a}, \mathbf{b}] \subset \mathbb{R}^2$, then the domain can be partitioned into triangles, as shown in Figure 2.38, quadrilaterals, or other polygonal shapes.*

After subdividing the domain, we choose points within the subdomains $\overline{P}_k$, so-called *nodes* or *nodal points*, that play a central role in the finite element method. As nodal points we identify at least all vertices of a subdomain. But to improve the desirable approximation, we can choose further points within a subdomain as nodal points, such as the midpoints or any other point of a subdomain, see Figure 2.39. Numbered with $1, 2, \ldots, N$, then we have a set of locations $\mathcal{N}_n \stackrel{\text{def}}{=} x_1, x_2, \ldots, x_n$ within the domain $\mathcal{R}$.

THE CHOICE OF THE GLOBAL BASIS FUNCTIONS. Based on the construct of the finite element mesh, we are now ready to describe how the basis functions $N_i, 1 \leq i \leq n$, can be constructed. They have to satisfy the following conditions:

i) All basis functions N_i, $1 \leq i \leq n$, should at least be bounded—a stronger require- $B(\cdot)$ (28)

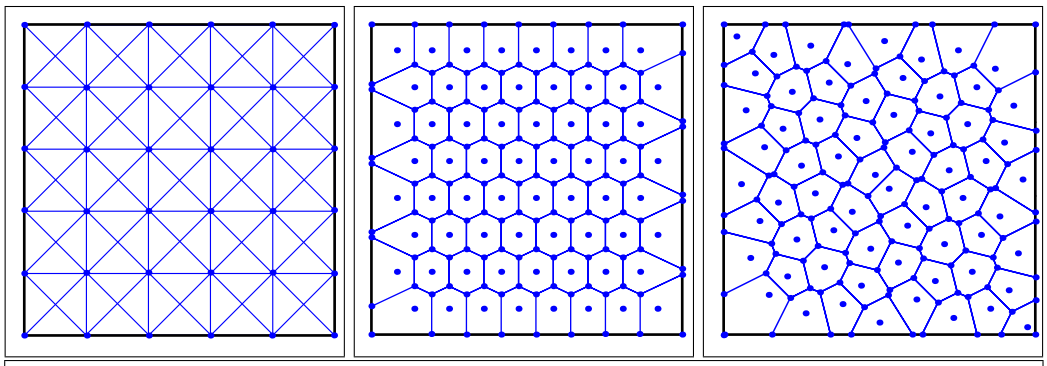


FIGURE 2.39: A SQUARE WITH ASSOCIATED FINITE ELEMENT MESHES. The left image shows a mesh, where nodal points are only chosen at the vertices of the patches. The images in the middle and on the right-hand side are Voronoi diagrams associated with the square $[\mathbf{a}, \mathbf{b}]$, here we have additionally chosen points within the patches as nodal points.

ment could be continuity—that is, $C(\cdot)$ (28)

$$N_i \in B(\overline{\mathcal{R}}) \quad \text{or perhaps even} \quad N_i \in C(\overline{\mathcal{R}}). \tag{2.459}$$

ii) Since the functions N_i should represent a basis of the n-dimensional subspace $\mathcal{U}_n$ of $\mathcal{R}$, such a basis consists of n functions N_i, one for each node, where each function is non-zero only on those patches P_k that are connected with node i. This means that for the restriction of N_i to P_k, in sign $N_i^{P_k}$, it must hold:

$$N_i(x)\big|_{P_k} \stackrel{\text{def}}{=} N_i^{P_k}(x) \equiv 0 \quad \text{if } x \notin \overline{P}_k. \tag{2.460}$$

iii) Furthermore, we require, that the basis functions N_i are equal to 1 only at node i, at all other nodes they take the value zero, thus,

$$N_i(x_j) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{2.461}$$

iv) The restriction of N_i to P_k, namely the function $N_i^{P_k}$, is a polynomial of degree at most l, that is,

$$N_i\big|_{P_k} \equiv N_i^{P_k}, \qquad N_i^{P_k} \in \mathcal{P}_l(P_k), \tag{2.462}$$

$\mathcal{P}_l(\cdot)$ (855) where $\mathcal{P}_l(P_k)$ is the space of polynomials of degree at most of l on P_k.

From the above conditions, it should be clear, that the restriction $N_i^{P_i}$ fulfills the condition:

$$N_i^{P_i}(x_j) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \tag{2.463}$$

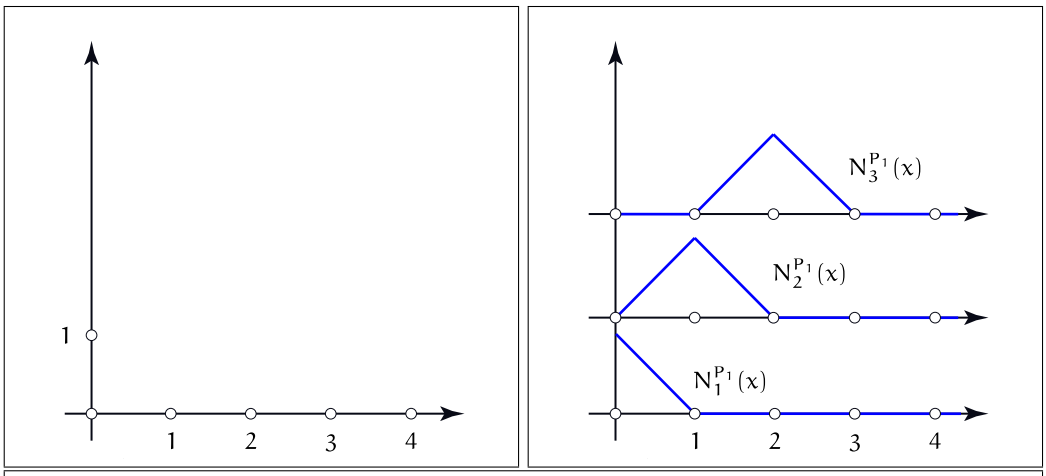


FIGURE 2.40: LOCAL BASIS FUNCTIONS. The interval $[0,4]$ is partitioned into 4 subintervals of equal size. The set of nodal points, $\mathcal{N}_5 = \{0,1,2,3,4\}$, then implies five local basis functions $N_i^{P_i}, 1 \le i \le 5$. The first three linear, local basis functions are plotted in the right image.

for all $x_i, x_j \in P_k$. We call $N_i^{P_k}$ a *local basis function*, see Figure 2.40.

A *global basis function* N_i can then be patched together from local basis functions $N_i^{P_k}$ associated with node i and P_k at the neighboring patches of node i. That is, the basis functions N_i are piecewise polynomials that have small support, in that they are nonzero only on a small region of the integration domain. Obviously, the number and placement of the nodal nodes are depending on the degree of the polynomials used. Due to the limited support of the basis functions, an approximation using the basis functions N_i is determined by summing only the basis functions, whose support overlaps the element.

EXAMPLE 2.56 (The One-dimensional Linear Basis Functions) *In one dimension, the linear basis function are given by:*

$$N_i(x) \stackrel{\text{def}}{=} \begin{cases} \frac{x-x_{i-1}}{x_i-x_{i-1}} & \text{for } x_{i-1} < x < x_i \\ \frac{x_{i+1}-x}{x_{i+1}-x_i} & \text{for } x_i < x < x_{i+1} \\ 0 & \text{otherwise.} \end{cases} \tag{2.464}$$

Let us assume, that the domain $[a,b]$ is partitioned into 4 elements $P_1,\ldots,P_4$ each of Lebesgue measure $\mu(P_i) = \frac{(b-a)}{4}$. The set of nodal points is given by $\mathcal{N}_5 = \{x_1,\ldots,x_5\}$ with $x_i = a+i\cdot\mu(P_i)$ for $1 \le i \le 5$, see Figure 2.40. Applied to the interval $[0,4]$, the linear basis functions N_i, defined on the nodal points $\mathcal{N}_5 = \{0,1,2,3,4\}$ are then given by:

$$N_1(x) \stackrel{\text{def}}{=} \begin{cases} 1-x & \text{for } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases} \tag{2.465}$$

$$N_2(x) \stackrel{\text{def}}{=} \begin{cases} x & \text{for } 0 < x < 1 \\ 2 - x & \text{for } 1 < x < 2 \\ 0 & \text{otherwise,} \end{cases} \tag{2.466}$$

etc., see Figure 2.40.

REMARK 2.56 *Apart from polynomials, finite element methods can also be constructed using other types of polynomials, such as Legendre and Jacobi polynomials, or also non-polynomial functions, see [29, Brenner & Scott 1994].*

DEFINITION 2.38 (The Finite Element) *Due to Ciarlet, [29, Brenner & Scott 1994], a* finite element *is defined as the tuple* $(\mathcal{R}, \mathcal{P}_l, \mathcal{N}_n)$*, where* $\mathcal{R}$ *is the domain with piecewise smooth boundary,* $\mathcal{P}_l$ *is the* l*-dimensional space of polynomials on* $\mathcal{R}$*, and* $\mathcal{N}_n = \{x_1, x_2, \ldots, x_n\}$ *is the set of nodal points given on* $\mathcal{R}$*.*

EXAMPLE 2.57 (The One-dimensional Lagrange Element) *i) Let us assume* $\mathcal{R}$ *be the unit interval* $[0, 1]$*, the set of nodal points* $\mathcal{N}_2$ *is given by* $\{0, 1\}$*, using linear polynomials from* $\mathcal{P}_1$*. Obviously, the nodal nodes are located at the boundaries of the interval, so we get* $N_1(x) = 1 - x$ *and* $N_2(x) = x$*.*

ii) If we assume $\mathcal{R} = [-1, 1]$*, then the set of nodal points* $\mathcal{N}_{k+1}$ *is given by* $\{-1, -1 + \frac{2}{k}, \ldots, 1 - \frac{2}{k}, 1\}$*, using polynomials* $\mathcal{P}_k$*. The local basis functions now correspond to the Lagrange polynomials of degree* k*, given by:*

$$N_j(x) \stackrel{\text{def}}{=} \frac{(x - x_1) \ldots (x - x_{j-1})(x - x_{j+1}) \ldots (x - x_{k+1})}{(x_j - x_1) \ldots (x_j - x_{j-1})(x_j - x_{j+1}) \ldots (x_j - x_{k+1})}. \tag{2.467}$$

Thus, for $k = 2$*, we have:* $\mathcal{N}_3 = \{-1, 0, 1\}$ *with* $N_1(x) = \frac{1}{2}x(x-1), N_2(x) = 1 - x^2$ *and* $N_3(x) = \frac{1}{2}x(x-1)$*.*

We leave the verification of the constructed finite elements to the interested reader as an exercise.

EXAMPLE 2.58 (The Two-dimensional Linear Basis Functions) *In two dimensions, the linear basis functions are given by:*

$$N_i(x) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{at node } x = x_i \\ [0, 1] & \text{within adjacent elements} \\ 0 & \text{at all other nodal nodes} \\ 0 & \text{outside adjacent elements.} \end{cases} \tag{2.468}$$

Obviously, this construction ensures that the basis functions N_i *satisfy the conditions from Equation (2.459) to (2.462).*

THE GALERKIN METHOD BASED ON A FINITE ELEMENT APPROACH. For the Galerkin method, based on a finite element approach, the function f_n from Equation (2.450) can

be written as a linear combination of global basis functions N_i defined on $\mathcal{R}$. That is, approximating f_n by:

$$f_n(x) = \sum_{i=1}^{n} \alpha_i N_i(x), \tag{2.469}$$

then Equation (2.451) has the form:

$$\langle (\mathbf{I} - \mathbf{K}) \sum_{i=1}^{n} \alpha_i N_i(x), N_j(x) \rangle \quad = \quad \langle g(x), N_j(x) \rangle, \tag{2.470}$$

thus

$$\sum_{i=1}^{n} \alpha_i \langle \underbrace{(\mathbf{I} - \mathbf{K}) N_i(x)}_{\widehat{N}_i(x)}, N_j(x) \rangle \quad = \quad \langle g(x), N_j(x) \rangle. \tag{2.471}$$

If $\boldsymbol{\alpha}$ represents the n-dimensional vector of unknowns, the term $\mathbf{gN}$ stands for the vector of inner products of the functions g and N_j at node x, and $\langle \widehat{N}_j, N_i \rangle_{1 \le i,j \le n}$ are the coefficients $(a_{ij})_{1 \le i,j \le n}$of the linear operator $\widehat{\mathbf{N}}\mathbf{N}$, then we can rewrite the above equation in form of an operator equation, that is: Linear Operator Equation (61)

$$\widehat{\mathbf{N}}\mathbf{N} \cdot \alpha = \mathbf{bN} \quad \Rightarrow \quad \boldsymbol{\alpha} = \left(\widehat{\mathbf{N}}\mathbf{N} \right)^{-1} \cdot \mathbf{g}, \tag{2.472}$$

which can be formulated in matrix notation as

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \langle \widehat{N}_1(x), N_1(x) \rangle & \dots & \langle \widehat{N}_n(x), N_1(x) \rangle \\ \vdots & & \vdots \\ \langle \widehat{N}_1(x), N_n(x) \rangle & \dots & \langle \widehat{N}_n(x), N_n(x) \rangle \end{pmatrix}^{-1} \cdot \begin{pmatrix} \langle g(x), N_1(x) \rangle \\ \vdots \\ \langle g(x), N_n(x) \rangle \end{pmatrix}. \tag{2.473}$$

2.3.3.2.4 SOLUTION METHODS FOR LINEAR SYSTEMS OF EQUATIONS

Whether quadrature or finite basis and projection methods, applied to Fredholm integral equations of the 2^{nd} kind, they all lead to a linear system of equations of the form

$$\mathbf{Ax} = \mathbf{b}, \tag{2.474}$$

where $\mathbf{A}$ is a $(n \times n)$-matrix, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{x}$ is an n-dimensional vector of unknowns.

Now, in mathematics, there exists are large field of algorithms for solving linear systems which can be partitioned into two classes:

- *direct solution procedures*, and Section 2.3.3.2.4.1
- *iterative methods*. Section 2.3.3.2.4.2

The idea behind direct methods is to determine the inverse $\mathbf{A}^{-1}$ of the matrix $\mathbf{A}$, if it exist. If $\mathbf{A}^{-1}$ is known, then the solution of the above system can easily be computed by multiplying the right hand side of Equation (2.474) with the inverse matrix $\mathbf{A}^{-1}$, leading to:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{2.475}$$

In contrast to direct solution procedures, iterative methods starts with an initial guess of the solution, any vector $\mathbf{x}^{(0)} \in \mathbb{R}^n$, and successively improve it until the computed iterate is as accurate as desired. So, an iteration method generates a sequence $\left(\mathbf{x}^{(n)}\right)_{n\in\mathbb{N}_0}$ of approximate solutions, which should converge to the desired, exact solution. Thus, in practice, an iteration procedure can be terminated if the error between the exact solution and the approximation is smaller than a pre-given bound.

2.3.3.2.4.1 DIRECT METHODS FOR SOLVING LINEAR SYSTEM OF EQUATIONS

Let us assume that the matrix $\mathbf{A}$ of the linear system of equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.476}$$

is invertible, that is, there exists $\mathbf{A}^{-1}$. Then, the first idea for solving this system is to find the analytic solution

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{2.477}$$

A well-known method for computing the vector $\mathbf{x}$ is the *Gaussian elimination procedure*, know from school mathematics. But there are still other direct methods, mostly variants of the Gaussian elimination, such as the *Gauss-Jordan algorithm*, the *Cholesky procedure*, or the *procedure by Crout*, see [202, Stoer & Bulirsch 1979] or [179, Schmeißer & Schirmeier 1976]. But all of these methods suffer from that they are prohibitively expensive when applied to large matrices. Since direct solution methods, such as the Gaussian elimination, require $O(n^3)$ operations to solve a linear system of equations or to compute the inverse of the associated matrix—where n is the number of unknowns in the system—these methods are not suitable for solving large systems of equations. Applied to large problems, direct methods are prohibitively expensive and should only be used for solving linear system whose associated matrices are sparse. Here, numerical methods, as we will discus them in the next section, have been proved to be more efficient solvers.

2.3.3.2.4.2 ITERATIVE METHODS FOR SOLVING LINEAR SYSTEMS OF EQUATIONS

The idea behind iterative methods for solving linear systems of equations is to express the operator equation

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{2.478}$$

Fixed-point Problem (62) as a fixed-point problem, similar to our discussion in Section 2.1.5. That is, we transform a linear system of equations into an equation of the form,

$$\mathbf{x}^{(k+1)} = \mathbf{T}\mathbf{x}^{(k)}, \qquad k \geq 0, \tag{2.479}$$

where $\mathbf{T}$ is a linear operator on $\mathbb{R}^n$, and $\mathbf{x}^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)})^T \in \mathbb{R}^n$ is any arbitrary starting value. New values can then be computed by repeated application of Formula (2.479).

Under the condition that the linear operator $\mathbf{T}$—now describing a linear mapping defined between the finite-dimensional space $\mathbb{R}^n$—is contracting, that is, if it holds, $\|\mathbf{T}\| < 1$, the Banach Fixed-point Theorem guarantees the convergence of the sequence $\mathbf{x}^{(k+1)}$ Banach Fixed-point Theorem (61) towards the right solution of Equation (2.478), for a detailed discussion see Section 2.1.5.

As the theoretical foundations for the convergence of iteration methods are already given by the Banach fixed-point Theorem, the question that now arises: How can we find appropriate iteration methods for solving linear systems of equations?

THE CLASSICAL ITERATION METHODS. For the following discussion, let $\mathbf{A}$ be a non-singular, that is, an invertible, $(n \times n)$-matrix with coefficients from $\mathbb{R}$, and let $\mathbf{b}$ be a n-dimensional vector from $\mathbb{R}^n$. The linear system of equations of type

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{2.480}$$

can then be transformed into an iteration method by using any arbitrary invertible matrix $\mathbf{B}$ via

$$\begin{aligned} \mathbf{B}\mathbf{x} + (\mathbf{A} - \mathbf{B})\mathbf{x} &= \mathbf{b} && (2.481)\\ \mathbf{B}\mathbf{x} &= (\mathbf{B} - \mathbf{A})\mathbf{x} + \mathbf{b} && (2.482)\\ \mathbf{x} &= \mathbf{B}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{x} + \mathbf{B}^{-1}\mathbf{b} && (2.483)\\ \mathbf{x} &= (\mathbf{I} - \mathbf{B}^{-1}\mathbf{A})\mathbf{x} + \mathbf{B}^{-1}\mathbf{b}. && (2.484) \end{aligned}$$

Choosing

$$\mathbf{T}\mathbf{x} \stackrel{def}{=} (\mathbf{I} - \mathbf{B}^{-1}\mathbf{A})\mathbf{x} + \mathbf{B}^{-1}\mathbf{b}, \tag{2.485}$$

then Equation (2.484) can be considered as a recurrence equation of the form (2.479) and for solving the equation, we can construct the following iteration method:

$$\begin{aligned} \mathbf{x}^{(k+1)} &\stackrel{def}{=} \mathbf{T}\mathbf{x}^{(k)} && (2.486)\\ &= (\mathbf{I} - \mathbf{B}^{-1}\mathbf{A})\mathbf{x}^{(k)} + \mathbf{B}^{-1}\mathbf{b}, \qquad k \geq 0 && (2.487) \end{aligned}$$

where $\mathbf{x}^{(0)} = (x_1^0, \ldots, x_n^0)$ is any arbitrary so-called *starting vector*.

REMARK 2.57 *From numerical analysis it is also known that the starting vector* $\mathbf{x}^{(0)}$ *can influence the rate of convergence, that is, a starting vector close to the final solution will require fewer iterations. If there is no information available for the choice on an initial guess,* $\mathbf{x}^{(0)}$ *can easily be chosen as a vector of zeros.*

REMARK 2.58 *As we have seen at the beginning of this section, the choice of any arbitrary invertible matrix* $\mathbf{B}$ *leads to an iteration method. Such an iteration method becomes even more efficient, the better it satisfies the following conditions:*

i) The system from (2.487) can easily be evaluated for computing $\mathbf{x}^{(k+1)}$*, and*

ii) the norm $\|\mathbf{A}\|$ *of the operator* $\mathbf{A}$ *is small.*

REMARK 2.59 (Used Notation in Iteration Methods) *All classical iteration methods generate sequences of vectors from* $\mathbb{R}^n$*. Since they generate a new iterate* $\mathbf{x}^{(k+1)} \in \mathbb{R}^n$ *from an already computed vector* $\mathbf{x}^{(k)}$ *due to the Formula (2.487) in one iteration—we also often say in an iteration cycle—such an iteration requires the computation of all components of* $\mathbf{x}^{(k+1)}$*. So, a complete iteration step of an iteration method consists of* n *steps for computing the single components of the vector* $\mathbf{x}^{(k+1)}$*.*

Now, iterative methods start with a guess $\mathbf{x}^{(0)}$ for the solution. By repeated, preferably inexpensive, and efficient computation of the following members of the sequence $\mathbf{x}^{(k)}$, the method then drives the original guess to a better approximate. If this approximate is close to the exact solution, then we call the procedure convergent. But how can we specify an approximate, if the exact solution is unknown?

Now, due to

$$\mathbf{A}\mathbf{x}^{(k)} - \mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{x}^{(k)} - \mathbf{x}), \tag{2.488}$$

the error $\mathbf{e}^{(k)}$ between the exact solution $\mathbf{x}$ and the result from the k^{th}-step of an iteration method, $\mathbf{x}^{(k)}$, can be specified as:

$$\mathbf{e}^{(k)} \stackrel{def}{=} \mathbf{x}^{(k)} - \mathbf{x}. \tag{2.489}$$

Since the exact solution $\mathbf{x}$ is not known, the error $\mathbf{e}^{(k)}$ cannot be quantified directly. To make a statement about the quality of the approximate $\mathbf{x}^{(k)}$, we have to use another measure, the residual $\mathbf{r}^{(k)}$, already known from the discussion of the Galerkin method. The residual $\mathbf{r}^{(k)}$ is defined by:

Residual (144)

$$\begin{aligned} \mathbf{r}^{(k)} &\stackrel{def}{=} \mathbf{A}\mathbf{e}^{(k)} && (2.490)\\ &= \mathbf{A}\mathbf{x}^{(k)} - \mathbf{A}\mathbf{x} && (2.491)\\ &= \mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}. && (2.492) \end{aligned}$$

It should be clear, that, if the residual is zero, then the approximate $\mathbf{x}^{(k)}$ is correct and the error is zero. Unlike the error, the residual is a quantity that can be directly

measured. Thus, the essence behind any iterative method is to compute a more accurate iterate $\mathbf{x}^{(k+1)}$ and residual $\mathbf{r}^{(k+1)}$, and to replace the last computed approximate $\mathbf{x}^{(k)}$ and residual $\mathbf{r}^{(k)}$ by these new values.

We will now introduce the most important iteration methods resulting from the varying choice of the matrix $\mathbf{B}$. The idea behind these methods is, that at each step of the procedure one component of the residual vector will be forced to be zero, that is, if the i^{th} component of the iterate $\mathbf{x}^{(k)}$ will be changed then we expect that this change leads to $r_i^{(k+1)} = 0$. Even if this requirement can lead to an increase of other components of $\mathbf{r}^{(k)}$, we hope that a complete iteration cycle leads to a better approximate $\mathbf{x}^{(k+1)}$. Since adjusting a component of an approximate so that the associated residual goes to zero is called *relaxing* the component, iterative methods of this type are also often called *relaxation methods*.

For the following, let us assume, that the matrix $\mathbf{A}$ can be decomposed in the following way

$$\mathbf{A} = \mathbf{D} - \mathbf{U} - \mathbf{L}, \tag{2.493}$$

where

$$\mathbf{D} = (a_{ii})_{1\le i\le n} = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & a_{nn} \end{pmatrix} \tag{2.494}$$

is a diagonal matrix consisting of the diagonal elements of $\mathbf{A}$,

$$\mathbf{L} = (-a_{ij})_{1\le j<i\le n} = \begin{pmatrix} 0 & \dots & \dots & 0 \\ -a_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & 0 \\ -a_{n1} & \dots & -a_{nn-1} & 0 \end{pmatrix} \tag{2.495}$$

is a lower triangle matrix, consisting of the reverse elements below the diagonal of $\mathbf{A}$, and

$$\mathbf{U} = (-a_{ij})_{1\le i<j\le n} = \begin{pmatrix} 0 & -a_{12} & \dots & -a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -a_{n-1n} \\ 0 & \dots & \dots & 0 \end{pmatrix} \tag{2.496}$$

are the reverse elements of the remaining upper triangle matrix.

THE JACOBI ITERATION. The *Jacobi iteration method* is based on the choice

$$\mathbf{B} = \mathbf{D}, \tag{2.497}$$

that is, we set:

$$\begin{aligned}
\mathbf{x} &= (\mathbf{I}-\mathbf{D}^{-1}\mathbf{A})\mathbf{x}+\mathbf{D}^{-1}\mathbf{b} && (2.498)\\
&= (\mathbf{I}-\mathbf{D}^{-1}(\mathbf{D}-\mathbf{U}-\mathbf{L}))\mathbf{x}+\mathbf{D}^{-1}\mathbf{b} && (2.499)\\
&= \mathbf{D}^{-1}(\mathbf{U}+\mathbf{L})\mathbf{x}+\mathbf{D}^{-1}\mathbf{b} && (2.500)\\
&= \mathbf{D}^{-1}(\mathbf{b}+(\mathbf{U}+\mathbf{L})\mathbf{x}). && (2.501)
\end{aligned}$$

As $\mathbf{D}$ is a diagonal matrix, the coefficients of $\mathbf{D}^{-1}$ are the reciprocals of $\mathbf{D}$ and the i^{th} component of $\mathbf{x}$ can easily be computed by multiplying the vectors $(\mathbf{b}+(\mathbf{U}+\mathbf{L})\mathbf{x})$ with $\frac{1}{a_{ii}}$. Then, the i^{th} component of the new approximate $\mathbf{x}^{(k+1)}$ corresponds to

$$x_i{}^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^{(k)}\right), \tag{2.502}$$

for $k \geq 0$ and $1 \leq i \leq n$.

This means: Computing one element of the new iterate $\mathbf{x}^{(k+1)}$ from the previous approximate $\mathbf{x}^{(k)}$ requires the evaluation of Formula (2.502). Since the iterate $\mathbf{x}^{(k)}$ is a n-dimensional vector, n steps of the Jacobi method has to be performed to get the new iterate $\mathbf{x}^{(k+1)}$. As the components of a new iterate do not depend on each other, they can not be computed simultaneously, that is, apart from storage requirements for the new iterate $\mathbf{x}^{(k+1)}$, we also need memory for storing $\mathbf{x}^{(k)}$. The Jacobi method does not always converge, but it is guaranteed to converge under the condition that the matrix is strictly diagonally dominant, i.e. $|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|$ for $1 \leq i \leq n$, but the convergence rate may be very slow.

Expressed in terms of the residual vector $\mathbf{r}^{(k+1)}$, then the Jacobi iteration can also be written as:

$$\begin{aligned}
x_i{}^{(k+1)} &= \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1\\ i\neq j}}^{n} a_{ij}x_j^{(k)}\right) && (2.503)\\
&= \frac{1}{a_{ii}}\left(a_{ii}x_i^{(k)} + \underbrace{b_i - \sum_{j=1}^{n} a_{ij}x_j^{(k)}}_{r_i^{(k)}}\right) && (2.504)\\
&= x_i^{(k)} + \frac{r_i^{(k)}}{a_{ii}} \qquad k \geq 0, 1 \leq i \leq n. && (2.505)
\end{aligned}$$

The Jacobi iteration, see Figure 2.41, works as follows: First, a starting vector $\mathbf{x}^{(0)} = \mathbf{0}$ is created. Then, if the error in the solution is not low enough, the residual vector $\mathbf{r}^{(k)}$

JACOBI ITERATION {

$\forall\ x_i^{(0)} \in \mathbf{x}^{(0)}$ do {

$x_i^{(0)} = 0$

}

while (not converged) {

$\mathbf{r}^{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}$

$\forall\ x_i^{(k+1)} \in \mathbf{x}^{(k+1)}$ do {

$x_i^{(k+1)} = x_i^{(k)} + \frac{r_i^{(k)}}{a_{ii}}$

}

}

FIGURE 2.41: JACOBI ITERATION.

must be computed and for each component $x_i^{(k)}$, the correction factor $\frac{r_i^{(k)}}{a_{ii}}$ is added to $x_i^{(k)}$. This brings the associated residual component $r_i^{(k+1)}$ to zero, as it holds:

$$r_i^{(k+1)} = b_i - \sum_{j=1}^{n} a_{ij} x_i^{(k+1)} \tag{2.506}$$

$$\overset{(2.505)}{=} b_i - \sum_{\substack{j=1 \\ i \neq j}}^{n} a_{ij} x_i^{(k)} - a_{ii} \underbrace{\left(x_i^{(k)} + \frac{r_i^{(k)}}{a_{ii}} \right)}_{x_i^{(k+1)}} \tag{2.507}$$

$$= \underbrace{b_i - \sum_{j=1}^{n} a_{ij} x_i^{(k)}}_{r_i^{(k)}} - r_i^{(k)} = 0, \tag{2.508}$$

where we have assumed that only the the i^{th} component of the iterate $\mathbf{x}^{(k+1)}$ was used for computing the residual-component $r_i^{(k+1)}$. Afterwards, the method returns to the top and test for convergence again.

REMARK 2.60 *Note: The Jacobi algorithm is, strictly speaking, not a relaxation method, as always only the first component of the residual $\mathbf{r}^{(k+1)}$ would actually result in zeroing when using all new computed components of an iterate $\mathbf{x}^{(k+1)}$. We leave the proof for this statement to the interested reader as an exercise.*

Let us now analyze the Jacobi iteration: Computing the residual vector $\mathbf{r}^{(k)}$ requires

taking the dot product of the guess $\mathbf{x}^{(k)}$ with the matrix $\mathbf{A}$ for each component of $\mathbf{r}^{(k)}$ and a subtraction with the associated element of the vector $\mathbf{b}$. That is, a complete iteration of the Jacobi iteration has costs $O(n^2)$.

REMARK 2.61 *In Section 10.1.4.2 we will pick up the Jacoby iteration and discuss the method when applied to the discrete radiosity equation. There we will reveal its essential similarity to the Neumann series described in Section 1.2.1. Thus, we can conclude that this simple algorithm will converge to the correct solution.*

THE GAUSS-SEIDEL ITERATION. Often, the Jacobi iteration only converges slowly to a desired solution, which is also the reason, why it is seldom used for solving linear systems of equations. But the Jacobi method gives us a good understanding of how and why iterative methods work. A more promising method for solving a linear system of equations is the *Gauss-Seidel iteration*. It is based on the choice

$$\mathbf{B} = \mathbf{D} - \mathbf{U}, \tag{2.509}$$

that is, we get:

$$\begin{aligned}
\mathbf{x} &= (\mathbf{I} - (\mathbf{D} - \mathbf{U})^{-1}\mathbf{A})\mathbf{x} + (\mathbf{D} - \mathbf{U})^{-1}\mathbf{b} && (2.510)\\
&= (\mathbf{I} - (\mathbf{D} - \mathbf{U})^{-1}(\mathbf{D} - \mathbf{U} - \mathbf{L}))\mathbf{x} + (\mathbf{D} - \mathbf{U})^{-1}\mathbf{b} && (2.511)\\
&= (\mathbf{D} - \mathbf{U})^{-1}\mathbf{L}\mathbf{x} + (\mathbf{D} - \mathbf{U})^{-1}\mathbf{b} && (2.512)\\
&= (\mathbf{D} - \mathbf{U})^{-1}(\mathbf{L}\mathbf{x} + \mathbf{b}). && (2.513)
\end{aligned}$$

The resulting iteration formula is then given by

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right), \qquad k \geq 0, 1 \leq i \leq n. \tag{2.514}$$

Obviously, the Gauss-Seidel method uses for generating the new iterate $\mathbf{x}^{(k+1)}$ already computed components of $\mathbf{x}^{(k+1)}$, while the Jacobi method generates the new iterate $\mathbf{x}^{(k+1)}$ solely on the basis of the iterate $\mathbf{x}^{(k)}$, computed in the previous iteration cycle.

Compared with the Jacobi iteration, the Gauss-Seidel method does not require duplicate storage for storing the vector $\mathbf{x}^{(k+1)}$, since the components of $\mathbf{x}^{(k+1)}$ can be overwritten if they are computed. As each component depends on previous ones, the iterate $\mathbf{x}^{(k+1)}$ has to be computed successively. Although the Gauss-Seidel iteration does not always converge, it is guaranteed to converge under conditions that are somewhat weaker—if matrix is symmetric and positive definite—than those for the Jacobi method. The Gauss-Seidel iteration provides a true relaxation method that converges about twice as fast as the Jacobi iteration scheme.

Let us now consider the i^{th} component of the residual in iteration cycle k, obviously, then it holds:

$$r_i^{(k\to k+1)} = b_i - \sum_{j=1}^{i} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)}. \tag{2.515}$$

Using this formula in Equation (2.514), we get for the residual $r_i^{(k\to k+1)}$:

$$x_i^{(k+1)} \overset{(2.514)}{=} \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) \tag{2.516}$$

$$= \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + a_{ii} x_i^{(k+1)} \right) \tag{2.517}$$

$$= x_i^{(k+1)} + \frac{r_i^{(k\to k+1)}}{a_{ii}}, \tag{2.518}$$

that is, the i^{th}-component of the residual $\mathbf{r}^{(k\to k+1)}$ will be zero.

If all residuals $r_i^{(k\to k+1)}$ are updated after an iteration, we can compute the iterate $\mathbf{x}^{(k+1)}$ via:

$$x_i^{(k+1)} = x_i^{(k)} + \frac{r_i^{(k\to k+1)}}{a_{ii}}, \qquad \text{for} \quad 1 \le i \le n. \tag{2.519}$$

Like the Jacobi iteration, a complete iteration cycle of the Gauss-Seidel method, see Figure 2.42, has even costs $O(n^2)$. This can easily be shown by analyzing the computation of the residual vector or the construction of the iterate $\mathbf{x}^{(k+1)}$. We leave this simple task to the interested reader as an exercise.

CONVERGENCE OF THE CLASSICAL ITERATION METHODS. Obviously, the classical iteration methods deliver for every starting vector $\mathbf{x}^{(0)}$ a sequence $(\mathbf{x}^{(k)})_{k\in\mathbb{N}_0}$ of vectors from $\mathbb{R}^n$. We denote an iteration method as *convergent*, if this sequence converges, for all starting vectors $\mathbf{x}^{(0)}$, towards the exact solution $\mathbf{A}^{-1}\mathbf{b}$ of the system.

As already mentioned in the introductory paragraph of this section, all iteration methods of the form (2.487) are convergent, if the operator $(\mathbf{I} - \mathbf{B}^{-1}\mathbf{A})$ is contracting, that is, if it holds:

$$\|\mathbf{I} - \mathbf{B}^{-1}\mathbf{A}\| < 1. \tag{2.520}$$

Since we are moving within the finite dimensional space $\mathbb{R}^n$, the operator $(\mathbf{I}-\mathbf{B}^{-1}\mathbf{A})$ is a $(n \times n)$-matrix, that is, the matrix is contracting if its spectral radius $\rho(\mathbf{I} - \mathbf{B}^{-1}\mathbf{A})$, i.e. its largest eigenvalue, is smaller than one.

Now, numerical analysis supports with two fundamental lemmata, valid only for contracting linear operators between finite dimensional linear spaces, that make a statement about the convergence of iterative methods:

GAUSS-SEIDEL ITERATION {

$\forall\ x_i^{(0)} \in \mathbf{x}^{(0)}$ do {

$x_i^{(0)} = 0$

}

while (not converged) {

$\forall\ x_i^{(k+1)} \in \mathbf{x}^{(k+1)}$ do {

$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{k)} \right)$

}

}

FIGURE 2.42: GAUSS-SEIDEL ITERATION.

LEMMA 2.3 (The Strong Row Sum Criterion) *Let* $\mathbf{A}$ *be a* $n \times n$*-matrix on* $\mathbb{R}^n$. *The classical iteration methods are convergent for all matrices* $\mathbf{A}$ *with*

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \tag{2.521}$$

for $1 \leq i \leq s$.

PROOF 2.3 *For a proof of this lemma, see [203, Stoer & Bulirsch 1978].*

REMARK 2.62 (The Strong Column Sum Criterion) *There exists also a* Strong Column Sum Criterion *that guarantees the convergence of the matrix* $\mathbf{A}$. *It can be formulated as*

$$|a_{ii}| > \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}| \tag{2.522}$$

for $1 \leq k \leq s$. *For a proof of this statement, see [203, Stoer & Bulirsch 1978].*

OVER AND UNDER RELAXATION. For accelerating the convergence behavior of an iterative method, you can also reformulate the fixed-point form in dependence of a parameter α, in such a way that the norm of the operator $\mathbf{T}$ becomes smaller than with the classical iteration methods, described above. So, the so-called *Jacobi relaxation method* is given

by:

$$x_i^{(k+1)} = (1-\alpha)x_i^{(k)} + \frac{\alpha}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ i \neq j}}^{n} a_{ij}x_j^{(k)}\right), \tag{2.523}$$

and the *Gauss-Seidel relaxation method* has the form:

$$x_i^{(k+1)} = (1-\alpha)x_i^{(k)} + \frac{\alpha}{a_{ii}}\left(b_i - \sum_{j=1}^{I-1} a_{ij}x_i^{(k+1)} - \sum_{j=I+1}^{n} a_{ij}x_i^{k)}\right) \tag{2.524}$$

for $k \geq 0$ and $1 \leq i \leq n$.

Depending on the choice of parameter α, we speak of *over relaxation* in the case, where α is chosen greater than 1, while *under relaxation* corresponds a value of α less than 1.

2.4 THE MOST IMPORTANT CONCEPTS FROM PROBABILITY THEORY

As a branch of higher mathematics, probability theory deals with the description and analysis of random processes and the features of mathematical structures conceived for them. Although early research in the field of probability were not based on an axiomatic theory in its present sense, a number of features, congruent with the intuitive concepts of probability, were already valid for the *Laplacian concept of probability*, which is based on finite sets.

If one limits to the classical Laplace concept of probability, defined over a finitely or countably infinite base set Ω, then we have a problem when modeling natural random experiments, whose results represent real numbers that are contained in a given interval. Applied to the task of solving the light transport equation, this means, that even elementary processes such as the stochastic sampling of a point within a light source for generating a shadow ray—a process encountered in every Monte Carlo rendering algorithm—cannot be modeled via the classical Laplacian concept of probability. In order to develop a sufficiently versatile theory with respect to these considerations, it is therefore necessary to incorporate uncountable base sets Ω into considerations of probability theory.

Section 6.5

If one tries to take this naïve path of classical probability theory, in which probability measures are defined on the whole power set of Ω, one soon encounters difficulties,[5] which

[5] If one chooses the half-open interval $[0,1) \in \mathbb{R}$ as the basic set Ω, then there exist no probability distribution that assigns the sets $A \in \mathfrak{P}(\Omega)$ and $A + h \stackrel{\text{def}}{=} \{a + h | a \in A, a + h \in [0,1)\} \in \mathfrak{P}(\Omega)$ the same probabilities.

can only be overcome by restricting the domain of probability measures. In so doing, great care must be taken to ensure the existence of a sufficiently large amount of probability distributions with additional particular characteristics to prevent the loss of too many important events.

One way of avoiding these problems is to adopt a measure-theoretical approach based on the concept of the σ-algebra, as developed by Kolmogorov in 1933. For developing new and analyzing already existing concepts of probability theory in a reasonable and feasible manner, we must therefore ourselves first acquaint with the fundamental concepts
Section 2.2 of general measure and integration theory.

Chapter 6 Consequently, before discussing Monte Carlo algorithms—based on probability theoretical concepts for finding solutions of particular integrals—we precede an outline of the most important concepts and constructions of probability theory, which we most frequently encounter in the analysis of Monte Carlo algorithms. Thereby, the emphasis lies on techniques and tools, relevant to the study of global illumination algorithms.

Section 2.4.1 The present section is structured as follows: First, we introduce the concept of the *probability space* as the fundamental building block of probability theory. Afterwards, Section 2.4.2 we talk about *random variables* and *random vectors*, the central point around which Section 2.4.3 all probability theory is turning. Considered as measurable functions from a probability space to a measurable space over $\mathbb{R}$ or $\mathbb{R}^n$, random variables and random vectors imply measures on, in particular probability measures, that describe the probability distribution of a random variable. We will also present the concepts of the *probability density* and the *cumulative distribution function* in the discrete and the continuous case. Based on the construct of the random variable, we then dedicate ourselves to the notions of the *expected value* and the *variance of a random variable*, that make statements about the Section 2.4.4 location of a probability distribution. With the concept of the *conditional probability*, Section 2.4.5 we present a tool that allows to determine the probability of a complex event via the Section 2.4.6 probabilities of simple events. Afterwards, we discuss the most important *limit theorems* of probability theory that makes statements about the limit behavior of a large number of random variables and we will present, with the concepts of *discrete-time Markov chain* Section 2.4.7 and the *discrete-time Markov process*, the most important types of *stochastic processes*.

2.4.1 PROBABILITY SPACES

Supported by a number of examples, in the preceding section an attempt was made, to introduce the most important concepts of measure and integration theory. Now, these concepts are not only useful for deriving solution methods for Fredholm integral equations of the 2^{nd} kind, but they also play a fundamental role in probability theory. Without the Measure (79) concept of the measure and the σ-algebra it is not possible to introduce the mathematical σ-algebra (828) model of the probability space, the most fundamental notion in probability theory. It is

the basic structure for the description and analysis of a random process and can be seen as a mathematical model of a real-world situation where randomness plays the central role.

Since all Monte Carlo methods for solving the light transport equation need the concepts of the probability space, the random variable, and the probability distribution, we now focus on introducing the notion of the *probability space*. Chapter 6

Based on the concept of the measure, we introduce in this section the construct of the *probability space* as a particular measure space, defined on a base set Ω. Furthermore, we present a technique, which makes it possible to construct a probability space from a given abstract measure space. This is the central idea of Monte Carlo integration, where the domain of an integral is interpreted as the base set of an appropriate probability space.

DEFINITION 2.39 (Probability Space) *Let Ω be any arbitrary set, $\mathfrak{F}(\Omega)$ a σ-algebra of subsets of Ω, and $\mathbb{P}$ a measure on $\mathfrak{F}(\Omega)$ such that*

σ-algebra (828)
Measure (79)

$$\mathbb{P}(\Omega) = 1, \tag{2.525}$$

then we call the triple $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ a probability space *with* probability measure, *in the following referred to as $\mathbb{P}$. The base set Ω is also called the* sample set *and the elements of $\mathfrak{F}(\Omega)$ are also denoted as the* events *of Ω.*

Measure Space (80)

REMARK 2.63 (Probability Distribution) *Note: In the following, we use the notions of the probability measure and the probability distribution synonymously.*

Dependent on the cardinality of the sample space—finite or countably infinite as well as uncountably infinite—we distinguish between two types of probability spaces: *discrete probability spaces* and *continuous probability spaces*.

Countable Set (827)
Uncountabe Set (827)

DEFINITION 2.40 (Discrete Probability Space) *Let Ω be a finite or countably infinite base set and let $\mathfrak{P}(\Omega)$ be the power set associated with it. Furthermore, let $(\omega_n)_{n\in\mathbb{N}}$ be a finite or infinite sequence of elements of Ω and let $(p_n)_{n\in\mathbb{N}}$ be a corresponding sequence of non-negative numbers of $\mathbb{R}$. Defining a discrete measure $\mathbb{P}$ via:*

Countable Set (827)
$\mathfrak{P}(\Omega)$ (828)
Discrete Measure (81)

$$\mathbb{P}(B) \overset{\text{def}}{=} \mathbb{P}\left(\bigcup_{\{\omega_n \in B | n\in\mathbb{N}\}} \{\omega_n\}\right) \tag{2.526}$$

$$= \sum_{\{\omega_n \in B | n\in\mathbb{N}\}} \mathbb{P}(\{\omega_n\}) \tag{2.527}$$

$$\overset{\mathbb{P}(\{\omega_n\})=p_n}{=} \sum_{\{\omega_n \in B | n\in\mathbb{N}\}} p_n, \tag{2.528}$$

where $\mathbb{P}(\Omega) = 1$ and $B \in \mathfrak{P}(\Omega)$, then the measure space $(\Omega, \mathfrak{P}(\Omega), \mathbb{P})$ is referred to as a discrete probability space *with probability measure $\mathbb{P}$.*

Measure Space (80)

Let us demonstrate the modeling of discrete probability spaces by means of two simple random experiments.

EXAMPLE 2.59 (Random Experiment of Flipping a Coin) *The sample space of this random experiment is given by* $\Omega = \{0, 1\}$*, that is, the set* Ω *consists of two elements,* $\omega_1 = 0$ *and* $\omega_2 = 1$*. Obviously, we can choose the power set* $\mathfrak{P}(\Omega)$*, thus the set of all subsets of* Ω*, as the* σ*-algebra of the associated measure space. Defining a measure* $\mathbb{P}$ *via:*

σ-algebra (828)

$$\mathbb{P}(B) = \sum_{\omega_n \in B} \mathbb{P}(\omega_n), \quad B \in \mathfrak{P}(\Omega), n = 1, 2 \tag{2.529}$$

where $\mathbb{P}(\omega_1) = \mathbb{P}(\omega_2) = p_1 = p_2 = \frac{1}{2}$*, then the measure space* $(\Omega, \mathfrak{P}(\Omega), \mathbb{P})$ *becomes a discrete probability space. You can see this easily, since* $\mathbb{P}$ *satisfies the requirements to a measure, and additionally fulfills the normalization property, thus,*

Measure Space (80)

Measure (79)

$$\begin{aligned} \mathbb{P}(\Omega) &= \mathbb{P}(\{\omega_1, \omega_2\}) && (2.530)\\ &= \mathbb{P}(\{\omega_1\}) + \mathbb{P}(\{\omega_2\}) && (2.531)\\ &= p_1 + p_2 = 1. && (2.532) \end{aligned}$$

EXAMPLE 2.60 (Random Experiment of Flipping a Coin s-times) *For simulating the random experiment of flipping a coin s-times, we choose a discrete probability space* $(\Omega, \mathfrak{P}(\Omega), \mathbb{P})$*, where* $\Omega = \{0, 1\}^s$*,* $\mathfrak{P}(\Omega)$ *is the power set of* Ω *and the probability distribution* $\mathbb{P}$ *is given by:*

$$\begin{aligned} \mathbb{P}(B) &\stackrel{def}{=} \mathbb{P}\left(\bigcup_{\omega_i \in B} \{\omega_i\}\right) && (2.533)\\ &= \sum_{\omega_i \in B} \mathbb{P}(\{\omega_i\}) && (2.534)\\ &= \sum_{\omega_i \in B} p_i, && (2.535) \end{aligned}$$

with $\mathbb{P}(\omega_i) \equiv p_i = \frac{1}{2^s} \geq 0$ *for* $1 \leq i \leq 2^s$ *and* $B \in \mathfrak{P}(\Omega)$*.*

Since the sample space Ω *consist of* 2^s *elements, where the probability of an elementary element is* $\frac{1}{2^s}$*, the measure* $\mathbb{P}$ *satisfies the normalization property of a probability measure in other words:* $(\Omega, \mathfrak{P}(\Omega), \mathbb{P})$ *is a discrete probability space.*

EXAMPLE 2.61 (Modeling a Discrete Probability Space via the Dirac Measure) *The Dirac measure, introduced in Example 2.27, can be used to model a discrete probability space in a convenient way. To show this, we suppose that a discrete probability space* $(\Omega, \mathfrak{P}(\Omega), \mathbb{P})$ *is given with* $\Omega = \{\omega_1, \omega_2, \ldots\}$ *and* $\mathbb{P}(\omega_i) = p_i, i \geq 1$*. Then, the associated probability measure* $\mathbb{P}$ *can be written as an infinite sum of the Dirac measure* δ_{ω_i}*, that is,*

$$\mathbb{P}(B) = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}(B) \tag{2.536}$$

for any $B \in \mathfrak{P}(\Omega)$.

Probability spaces, whose underlying probability measure is a Dirac measure, play a central role for the theoretical analysis of Monte Carlo rendering algorithms, as we will introduce them in Chapter 9. Since all these algorithms also have to cover lighting situations under idealized scene conditions, they also take into account light interaction at perfectly smooth surfaces, where we have ideal specular reflection and refraction. That is, sampling techniques have to be developed which guarantee to sample the ideal reflected or the ideal refracted ray for a given incoming ray. This can then be done via the usage of a Dirac measure.

In a later chapter we encounter again and again the random experiment of sampling points or directions from sets that are subsets of $\mathbb{R}^s$. Now, the cardinality of these sets is uncountably infinite. As we have seen when introducing the concept of the measure, in cases where we would like to define a measure on an uncountably infinite base set, the measure of a countably infinite set is zero. That would mean that we choose a point or a direction from a subset of $\mathbb{R}^s$ with probability zero, which would not be a good idea. Since the process of sampling of points and directions over the unit sphere or subsets thereof may be regarded as one of the cornerstones of every Monte Carlo rendering procedure we need an extended model of a probability space: the concept of the *continuous probability space*.

Chapter 6
Uncountability (827)
Countability (827)

DEFINITION 2.41 (Continuous Probability Space) *Let Ω be an uncountably infinite set, $\mathfrak{F}(\Omega)$ a σ-algebra of subsets of Ω, and $\mathbb{P}$ a measure on $\mathfrak{F}(\Omega)$. Then, the measure space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ is called a* continuous probability space *if it holds:*

Uncountable Set (827)
σ-algebra (828)
Measure Space (80)

$$\mathbb{P}(\Omega) = 1. \tag{2.537}$$

As in the discrete case, let us also demonstrate the modeling of a continuous probability space using a very important example, namely the random experiment of drawing a number from the unit interval $I = [0, 1]$.

EXAMPLE 2.62 (Stochastic Experiment of Drawing a Number from $[0, 1]$) *Obviously, the set $[0, 1]$ is uncountably infinite with Lebesgue measure $\mu([0, 1]) = 1$, i.e. the probability space $([0, 1], \mathfrak{B}([0, 1]), \mu)$ precisely describes what we mean by this experiment. Since all subintervals of $[0, 1]$ with the same length have equal measure, the probability measure $\mathbb{P}$ is spread uniformly over $[0, 1]$, that is: The probability of drawing a number from $\left(0, \frac{1}{4}\right]$ is the same as that of drawing a number from $\left(\frac{1}{4}, \frac{1}{2}\right]$, namely $\frac{1}{4}$, thus, all intervals of the same length are equally probable, see Figure 2.43.*

Contrary to the previous examples, where it was possible to ask for the probability of an elementary event such as a head or a number, which are elements of the sample space, this makes no sense in a continuous probability space, because all countably infinite, measurable sets are sets of measure zero.

Lebesgue Measure (75)
$\mathfrak{B}([0, 1])$ (865)
Measurable Set (80)

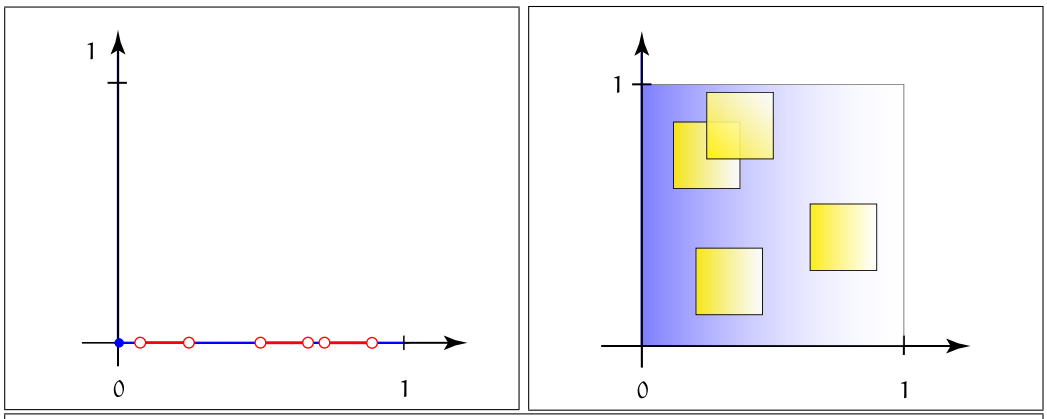


FIGURE 2.43: DRAWING RANDOM NUMBERS FROM $[0,1]$ OR $[0,1]^2$. All shown intervals have the same length, so, the probability of drawing a random number from one of these intervals is the same independent on the location of an interval. The same is also true for 2-dimensional intervals. They have the same area, thus, the probability of drawing a random number from one of these regions is the same independent on the location of such a 2-dimensional interval.

Lebesgue Measure (75) $\mathfrak{B}([0,1])$ (865)

EXAMPLE 2.63 (The Canonical Probability Space $([0,1]^s, \mathfrak{B}([0,1]^s), \mu^s)$) *Via the Lebesgue measure* μ^s *the measurable space* $([0,1]^s, \mathfrak{B}([0,1]^s))$ *can be extended to the so-called* canonical probability space $([0,1]^s, \mathfrak{B}([0,1]^s), \mu^s)$ *since it holds:*

$$\mu^s\left([0,1]^s\right) = \prod_{i=1}^{s} \mu([0,1]) = 1. \tag{2.538}$$

Section 6.5 $\mathfrak{B}([a,b]^s)$ (865)

EXAMPLE 2.64 (The Probability Space $([a,b]^s, \mathfrak{B}([a,b]^s), \mu^s)$) *When discussing sampling methods we will encounter again and again the problem of sampling a random number, not from the* s*-dimensional unit interval, but from the interval* $[a,b]^s$*. To sample from such an interval, a probability space* $([a,b]^s, \mathfrak{B}([a,b]^s), \mathbb{P})$ *is required, where the probability distribution is given by:*

$$\mathbb{P}(\mathbf{B}) \stackrel{def}{=} \frac{\mu^s(\mathbf{B})}{\mu^s([a,b]^s)} \tag{2.539}$$

$$= \frac{\prod_{i=1}^{s} \mu(B_i)}{\prod_{i=1}^{s} \mu([a_i,b_i])}, \tag{2.540}$$

with $\mathbf{B} = B_1 \times \ldots \times B_s \in \mathfrak{B}([a,b]^s)$*. It is straightforward to show, that* $\mathbb{P}$ *is a probability measure.*

We will now present a technique which plays an important role in the theory of Monte Carlo integration. Based on measure theoretical concepts it allows to construct a

Chapter 6

probability space from a given measure space. This makes it possible to substitute the measure underlying an integral by a probability measure and to represent an integral as the stochastic expected value of a random variable defined on a probability space.

Expected Value (196)

CONSTRUCTING A PROBABILITY SPACE FROM A GIVEN MEASURE SPACE. Let $(\Omega, \mathfrak{F}(\Omega), \nu)$ be a measure space over an uncountable base set Ω and let q be any measurable, non-negative function defined on Ω. Now, measure theory says that it is possible to generate a new measure $\mathbb{P}$ by:

Measurable Space (80)
Uncountable Set (827)
Measurable Function (98)
Measure (79)

$$\mathbb{P}(B) = \int_B q(x)\, d\nu(x), \quad B \in \mathfrak{F}(\Omega). \tag{2.541}$$

To extend this new generated measure to a probability measure, the measure $\mathbb{P}$ must be normalized, i.e. it must hold $\mathbb{P}(\Omega) = 1$. This can be obtained by multiplying the right-hand side of Equation (2.541) with a normalization factor given by:

$$\frac{1}{\int_\Omega q(x)\, d\nu(x)}. \tag{2.542}$$

Our original measure space $(\Omega, \mathfrak{F}(\Omega), \nu)$ then becomes a probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ with probability measure $\mathbb{P}$ defined by

Measure Space (80)

$$\mathbb{P}(B) \stackrel{def}{=} \frac{1}{\int_\Omega q(x)\, d\nu(x)} \int_B q(x)\, d\nu(x). \tag{2.543}$$

Let us show by means of a simple example how this technique of constructing a probability space works for a given measure space.

EXAMPLE 2.65 *Given be the measure space* $([0,\pi], \mathfrak{B}([0,\pi]), \mu)$*, where* $\mathfrak{B}([0,\pi])$ *is the Borel* σ*-algebra over* $[0,\pi]$*,* μ *is the Lebesgue measure on* $\mathbb{R}$*, and* $q(x) = x^2$ *is a non-negative, measurable function defined on the interval* $[0,\pi]$*. To construct the associated probability space* $([0,\pi], \mathfrak{B}([0,\pi]), \mathbb{P})$*, we choose the normalization factor as:*

$\mathfrak{B}(\cdot)$ (865)
μ (75)
Measurable Function (98)

$$\frac{1}{\int_{[0,\pi]} x^2\, d\mu(x)} = \frac{1}{\frac{1}{3}x^3\big|_0^\pi} = \frac{3}{\pi^3}, \tag{2.544}$$

and define the probability measure $\mathbb{P}$ *via:*

$$\mathbb{P}([a,b]) \stackrel{def}{=} \frac{3}{\pi^3} \int_{[a,b]} x^2\, d\mu(x) = \frac{3}{\pi^3}\frac{1}{3}x^3\big|_a^b = \frac{b^3 - a^3}{\pi^3} \tag{2.545}$$

with $[a,b] \in \mathfrak{B}([0,\pi])$*.*

Keep your eyes open for this technique; you will see it used more and more in Monte Carlo integration.

Section 6.2

2.4.2 RANDOM VARIABLES AND DISTRIBUTION FUNCTIONS

The model of the probability space, as introduced in the previous section, is still always very abstract and without any relation to practical problems. The question that now arises: How can we fill this theoretical concept with life?

For that, we will map the base set Ω of a probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ into the Euclidean space $\mathbb{R}$, in such a way, that we can also define events and probabilities in $\mathbb{R}$, this space is well known to us. Under the condition that these mappings are structure-preserving, we can them define as *random variables* which assign any event from $\mathfrak{F}(\Omega)$ a real number. So, random variables can be used to describe absolute or relative frequencies, lengths, or weights and so on. As they allow to transfer the concept of probability to the Borel sets on $\mathbb{R}$, the construct of the random variable extends the Euclidean space $\mathbb{R}$ to a probability space, where the whole mathematical framework of differential and integral calculus is available. The construct of the *random variable* is the central point around which all probability theory is turning. Thus, without any qualification we can say that probability theory is the study of random variables as well as that of functions of random variables. Also, let us introduce them now.

Based on the probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ we introduce in this section the concept of the *random variable* as a measurable function into the measurable space $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. We illustrate the difference between *discrete* and *continuous random variables* and show that it is possible to describe the probability distribution of a random variable by means of a so-called *cumulative distribution function*. With uniform sampling of points from the unit interval and directions from the hemisphere we already presented first examples of a series of *sampling techniques* used in Monte Carlo rendering algorithms.

Random Variable (168)
Measurable Space (80)
CDF (171)
Section 6.5

RANDOM VARIABLES AND FUNCTIONS OF RANDOM VARIABLES. Let us consider once more the random experiment of flipping a coin s-times from Example 2.60. It can be modeled by the probability space $(\{0,1\}^s, \mathfrak{P}(\{0,1\}^s), \mathbb{P})$. The outcomes of this random experiment can now be seen from different angles. So for example, we can be interested in the number of heads that are flipped, thus $0, 1, 2, \ldots, s$ or perhaps in the outcome of only the first, second or third flip, thus 0 and 1. This implies that a new sample space, the set $\{0, 1, 2, \ldots\}$ or $\{0, 1\}$, now forms the basis of our random experiment. The question that arises is: How can we deduce from our original probability measure $\mathbb{P}$ a probability measure for the new sample space? The idea behind it is the concept of the *random variable*.

Probability Space (163)

Probability Space (163)
Measurable Function (98)

DEFINITION 2.42 (Random Variable) *Let $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ be a probability space and X a measurable function*

$$X : (\Omega, \mathfrak{F}(\Omega)) \longrightarrow (\mathbb{R}, \mathfrak{B}(\mathbb{R})) \tag{2.546}$$

with

$$\Omega \ni \omega \longmapsto X(\omega) \in \mathbb{R}, \tag{2.547}$$

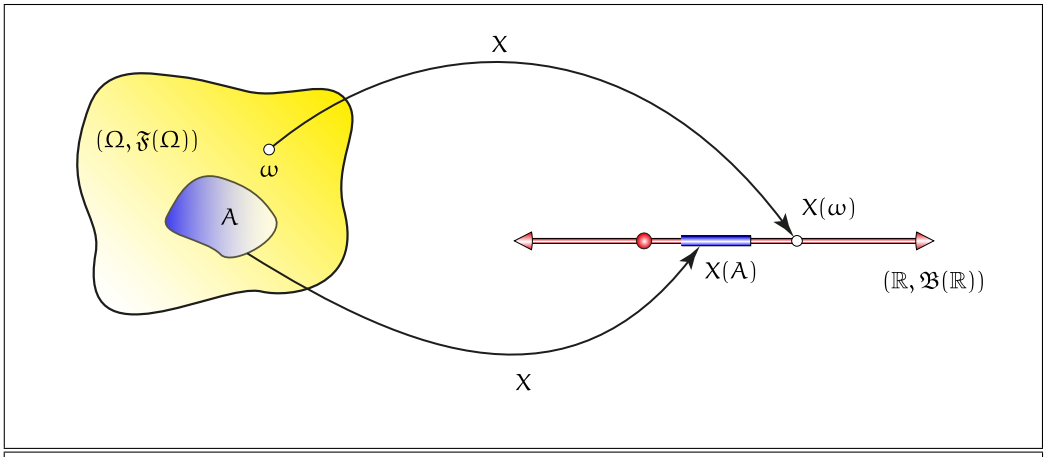


FIGURE 2.44: DEFINITION OF A RANDOM VARIABLE. Defined as a measurable function between the measurable spaces $(\Omega, \mathfrak{F}(\Omega))$ and $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, a random variable X is a mapping that assigns any element of $\mathfrak{F}(\Omega)$ an element of the σ-algebra $\mathfrak{B}(\mathbb{R})$.

then X *is called a* random variable, *see Figure 2.44.*

REMARK 2.64 (Discrete and Continuous Random Variables) *In probability theory it is distinguished between discrete and continuous random variables. Thus, a random variable* X *is called a* discrete random variable, *if its sample space* Ω *is finite or countably infinite, if* Ω *is uncountably infinite,* X *is called a* continuous random variable.

Countable Set (827)
Uncountable Set (827)

REMARK 2.65 (Functions of Random Variables) *Since the composition of measurable functions is also measurable, it is clear that with* X *the composition* $f \circ X$, *where* $f : \mathbb{R} \to \mathbb{R}$ *is a measurable function, will also be a random variable on* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$.

Measurable Function (98)

CUMULATIVE DISTRIBUTION FUNCTION OF A RANDOM VARIABLE. Now, as a measurable function, a random variable can imply a measure onto the Borel σ-algebra $\mathfrak{B}(\mathbb{R})$, the so-called *image measure* $\mathbb{P}_X$. The image measure is defined as follows:

Borel σ-algebra (865)
Measure (79)

DEFINITION 2.43 (The Image Measure of a Random Variable) *Let* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *be a probability space and let* X *be a random variable defined on* $(\Omega, \mathfrak{F}(\Omega))$ *with values in* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. *Then, the random variable* X *implies a measure* $\mathbb{P}_X$*:*

Probability Space (163)

$$\mathbb{P}_X : (\mathbb{R}, \mathfrak{B}(\mathbb{R})) \to [0, 1] \tag{2.548}$$

the so-called image measure. *It is defined via the probability measure* $\mathbb{P}$ *by:*

$$\begin{aligned} \mathbb{P}_X(B) &\stackrel{def}{=} \left(\mathbb{P} \circ X^{-1}\right)(B), && (2.549)\\ &= \mathbb{P}\left(X^{-1}(B)\right) && (2.550)\\ &= \mathbb{P}\left(\{\omega \in \Omega \,|\, X(\omega) \in B\}\right) && (2.551) \end{aligned}$$

for any $B \in \mathfrak{B}(\mathbb{R})$.

$\mathbb{P}_X$ *is also denoted as the* probability distribution of the random variable X. *It completely characterizes the random variable* X *in the sense that it provides the probabilities of all events from* $\mathfrak{B}(\mathbb{R})$. *It should also be clear that with the image measure* $\mathbb{P}_X$ *the triple* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mathbb{P}_X)$ *becomes a probability space.*

Event (163)

REMARK 2.66 *Note:* $\mathbb{P}_X$ *is a probability distribution over the observation space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ *instead of* $(\Omega, \mathfrak{F}(\Omega))$.

REMARK 2.67 *From Remark 2.65 it is known that the composition of a random variable* X *with a measurable function* f *is also a random variable. Thus, a non-negative function* f *defined on the measurable space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ *can be used to construct a probability distribution for the random variable* $f \circ X$*, if we define:*

Measurable Function (98)
Measurable Space (80)

$$\begin{aligned} \mathbb{P}_{f\circ X}(B) &\stackrel{def}{=} \left(\mathbb{P} \circ (f \circ X)^{-1}\right)(B) && (2.552)\\ &= \mathbb{P}\left(X^{-1}\left(f^{-1}(B)\right)\right) && (2.553)\\ &= \mathbb{P}\{\omega \in \Omega \,|\, f(X(\omega)) \in B\} && (2.554) \end{aligned}$$

with $B \in \mathfrak{B}(\mathbb{R})$.

As we know from the previous section to build a probability measure on a probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ we have to assign a probability to all events from $\mathfrak{F}(\Omega)$. In the case of a finite or a countably infinite base set Ω, this makes no problems, since we have to determine the probabilities for all elementary events from $\mathfrak{F}(\Omega)$. But in the case where Ω is an interval or an uncountably infinite set, we have even to determine the probabilities of all elements of $\mathfrak{F}(\Omega)$. Now, since this set is uncountably infinite, this is an impossible task. But via the probability space $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mathbb{P}_X)$, induced by a random variable X this task can easily be accomplished. Here it is sufficient to specify a function, the *cumulative distribution function of a random variable*, that describes the probabilities of all of these events. That is, if we know this function, then we know the probability of any interval and also the probability of any Borel set. As already mentioned above, the introduction of the concept of the random variable allows us to leave the abstract probability space and to move ourself within the well-known Euclidean space $\mathbb{R}$.

Based on the image measure $\mathbb{P}_X$, we now introduce the concept of the *cumulative distribution function of a random variable*, that describes the probability distribution of a random variable, or in the language of measure theory, the probability measure on the σ-algebra $\mathfrak{B}(\mathbb{R})$.

DEFINITION 2.44 (Cumulative Distribution Function) *Let* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *be a probability space and let* X *be a random variable defined on* $(\Omega, \mathfrak{F}(\Omega))$ *with values in* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. *The non-descending, right continuous function* F_X

Probability Space (163)

$\mathfrak{B}(\mathbb{R})$ (865)

$$F_X : \mathbb{R} \longrightarrow [0, 1] \tag{2.555}$$

defined by:

$$F_X(x) \stackrel{def}{=} \mathbb{P}_X((-\infty, x]) \tag{2.556}$$

$$\stackrel{(2.551)}{=} \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) \leq x\}) \tag{2.557}$$

with the properties

$$\lim_{x \to -\infty} F_X(x) = 0 \qquad \textit{and} \qquad \lim_{x \to \infty} F_X(x) = 1 \tag{2.558}$$

is denoted as the cumulative distribution function of X *or briefly the* distribution function *of* X, *abbreviated also simply the* CDF *of* X. *Obviously,* F_X *describes the probability distribution of the random variable* X, *i.e.* $F_X(x)$ *is interpreted as the probability that* X *takes on a value less than or equal to* x.

As seen from the definition above, the cumulative distribution function is based on the image measure induced by the involved random variable. Because this measure can be discrete or continuous, we need to define also two different types of CDFs. A cumulative distribution function for a discrete random variable, as well as a cumulative distribution function for the continuous case of a random variable.

Image Measure (170)

CUMULATIVE DISTRIBUTION FUNCTION OF A DISCRETE RANDOM VARIABLE. For a discrete random variable, the above introduced image measure is defined as a measurable function from a discrete probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ to the measurable space $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. This means that a subset $A \in \mathfrak{F}(\Omega)$ is mapped to a null set $A' \in \mathfrak{B}(\mathbb{R})$ with measure $\mathbb{P}_X(A') = 0$. To handle this drawback, we adapt the above presented image measure to the discrete case of a random variable by introducing a so-called *probability mass function*, which determines completely the properties of a discrete random variable.

Measurable Space (80)

Null Set (80)

DEFINITION 2.45 (Probability Mass Function) *Let* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *be a discrete probability space and let* X *be a discrete random variable defined on* $(\Omega, \mathfrak{F}(\Omega))$ *with values in the measurable space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, *where* Ω *is a finite or countably infinite set. Then, the* probability mass function p_X *of the random variable* X *is defined as:*

Probability Space (163)

Measurable Space (80)

$$p_X(x) \stackrel{def}{=} \mathbb{P}_X(X = x) \tag{2.559}$$

for all $x \in \mathrm{Im}(X)$. *Since* p_X *is defined via the probability measure* $\mathbb{P}_X$, *a probability mass function* p_X *satisfies the following conditions*

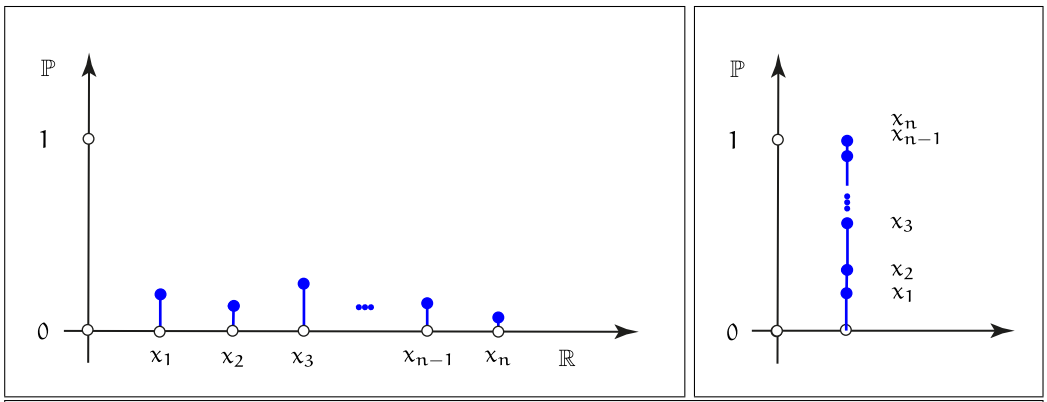


FIGURE 2.45: PROBABILITY MASS FUNCTION. Based on the image measure $\mathbb{P}_X$, a PMF satisfies the non-negativity property of a measure, i.e. $p_X(x) \geq 0, \forall x \in \mathsf{Im}(X)$, illustrated in the left image, and the normalization property of a probability measure, $\sum_{x \in \mathsf{Im}(X)} p_X(x) = 1$, shown in the right image.

i) $p_X(x) \geq 0$

ii) $\sum_{x \in \mathrm{Im}(X)} p_X(x) = 1$

for all x *from the image area of* X*, see Figure 2.45.*

Image Measure (170)

REMARK 2.68 *Due to the definition of the image measure* $\mathbb{P}_X$*, a probability mass function can also be expressed in terms of the probability measure* $\mathbb{P}$*, that is, we can also use the formula*

$$p_X(x) \stackrel{\mathrm{def}}{=} \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x\}) \tag{2.560}$$

for computing the probability distribution of a discrete random variable. This means, that a PMF can be evaluated on two different ways: via the probability measure $\mathbb{P}$ *from* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *or via the image measure* $\mathbb{P}_X$ *of the probability space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mathbb{P}_X)$*.*

Let us now illustrate the concept of a probability mass function by means of a simple example.

EXAMPLE 2.66 (Probability Mass Function Induced by a Discrete Random Variable) *Let us consider once more our random experiment of flipping a coin* s*-times, where we assume* $s = 3$*. If we are only interested in the number of heads, then the associated random variable must be defined as:*

$$X : (\{0,1\}^3, \mathfrak{P}(\{0,1\}^3)) \longrightarrow (\mathbb{R}, \mathfrak{B}(\mathbb{R})) \tag{2.561}$$

with

$$\Omega \ni \omega \longmapsto X(\omega) \stackrel{\text{def}}{=} \#1's \text{ in } \omega, \tag{2.562}$$

i.e. we count the number of one's in an elementary event.

Obviously, the discrete random variable X is a measurable function as it holds Measurable Function (98)

$$\{X < a\} = \begin{cases} \emptyset & \text{if } a \leq 0 \\ \{000\} & \text{if } 0 < a \leq 1 \\ \{001, 010, 100\} & \text{if } 1 < a \leq 2 \\ \{011, 101, 110\} & \text{if } 2 < a \leq 3 \\ \Omega & \text{if } 3 < a. \end{cases} \tag{2.563}$$

Due to Relation (2.559), the probability $\mathbb{P}_X(X = 1)$ *can be interpreted as the occurrence of exactly a single one, whereas this event can be computed with respect to the probability measure of* $(\{0,1\}^3, \mathfrak{P}(\{0,1\}^3), \mathbb{P})$*, that is,*

$$\mathbb{P}_X(X = 1) \stackrel{(2.549)}{=} \left(\mathbb{P} \circ X^{-1}\right)(X = 1) \tag{2.564}$$

$$= \mathbb{P}\left(X^{-1}(X = 1)\right). \tag{2.565}$$

Now, $X^{-1}(X = 1)$ *is the set of elementary events from* $\{0,1\}^3$ *which are mapped by X onto the outcomes* 0 *and* 1*. Mathematically, this can be expressed as:*

$$X^{-1}(X = 1) = \{\omega \,|\, X(\omega) = 1\} = \{001, 010, 100\}. \tag{2.566}$$

Combining these results leads to:

$$\mathbb{P}_X(X = 1) = \mathbb{P}(\{001, 010, 100\}) \tag{2.567}$$

$$= \mathbb{P}\left(X^{-1}(X = 1)\right) \tag{2.568}$$

$$= \underbrace{\mathbb{P}(\{001, 010, 100\})}_{\frac{3}{8}}. \tag{2.569}$$

Let us now show, by means of another interesting example from computer graphics, how we can construct a probability mass function.

EXAMPLE 2.67 *Assume, we have to render a scene, which is illuminated by more than a single light source. Then, a rendering algorithms can choose samples from all of these light sources. In particular, it must take more samples from light sources that contribute more light to the scene than others, that is, we must construct a probability distribution over all existing light sources depending on the power of the light sources.* Chapter 8 Radiant Power (249)

This situation can now be modeled by a probability space $(\Omega, \mathfrak{P}(\Omega), \mathbb{P})$*, whereas* $\Omega = \{\omega_1, \ldots, \omega_n\}$ *denotes the set of our light sources, each equipped with a fixed*

power of light, $\Phi(i), 1 \leq i \leq n$. *We then construct a random variable* X *defined on* $(\Omega, \mathfrak{P}(\Omega))$ *via*

$$X(\omega_i) = i, \qquad 1 \leq i \leq n, \tag{2.570}$$

with

$$\begin{aligned} p_X(i) &= \mathbb{P}_X(X = i) && (2.571)\\ &\stackrel{def}{=} \frac{\Phi(X(\omega_i))}{\sum_{k=1}^{n} \Phi(X(\omega_k))}. && (2.572) \end{aligned}$$

It should be clear that $p_X(i)$ *corresponds to the probability for sampling from light source* ω_i. *Provided that all light sources contribute the same amount of power to the scene then it holds:*

Section 9.1.2

$$\begin{aligned} p_X(i) &= \mathbb{P}_X(X = i) && (2.573)\\ &= \frac{\Phi(X(\omega_i))}{\sum_{k=1}^{n} \Phi(X(\omega_k))} = \frac{1}{n}, && (2.574) \end{aligned}$$

that is, all light sources are drawn with the same probability.

Based on the probability mass function, we are now ready to define the *cumulative distribution function* of a discrete random variable. It makes a statement about the probability that the value of a random variable is less than or equal to some given real number.

DEFINITION 2.46 (Cumulative Distribution Function of a Discrete Random Variable) *Let* X *be a discrete random variable on the discrete probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *with probability mass function* p_X. *Due to Definition 2.44 the* cumulative distribution function F_X *of the* discrete random variable X *can be written as:*

Probability Space (163)

$$\begin{aligned} F_X(x) &\stackrel{def}{=} \mathbb{P}_X(X \leq x) && (2.575)\\ &= \sum_{\xi \leq x} \mathbb{P}_X(X = \xi) && (2.576)\\ &\stackrel{(2.559)}{=} \sum_{\xi \leq x} p_X(\xi). && (2.577) \end{aligned}$$

See Figure 2.46 for an illustration of the CDF of a discrete random variable.

REMARK 2.69 *Due to the definition of the image measure* $\mathbb{P}_X$, *the discrete cumulative distribution function* F_X *can also be expressed in terms of the probability measure* $\mathbb{P}$, *that is, we can also use the formula*

Image Measure (170)

$$\begin{aligned} F_X(x) &\stackrel{(2.557)}{=} \sum_{\xi \leq x} \mathbb{P}\left(\{\omega \in \Omega \,|\, X(\omega) = \xi\}\right) && (2.578)\\ &= \mathbb{P}\left(\{\omega \in \Omega \,|\, X(\omega) \leq x\}\right) && (2.579) \end{aligned}$$

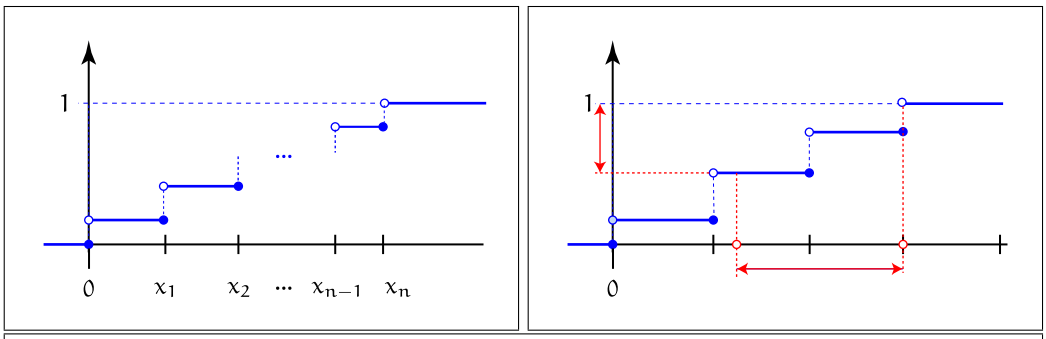


FIGURE 2.46: CUMULATIVE DISTRIBUTION FUNCTION OF A DISCRETE RANDOM VARIABLE. As you can easily see from the image, the illustrated CDF satisfies the condition requires to a CDF from Definition 2.44. Thus, the image range of the function lies within the interval $[0,1]$ and you can detect the stepwise increasing of function values from the graph. The function also satisfies the limiting properties required to a CDF, namely $F_X(x) \to 0$ for sufficiently small values and $F_X(x) \to 1$ for sufficiently large arguments. Obviously, F_X is a monotonically increasing step function. It has its jumps at those ordinates x_i which are the values of the random variable X. At these positions, F_X is right continuous. The magnitudes of the jumps are given by the probability distributions of the events taking the values of x_i. Due to this observation it is possible to compute the probability that the random variable X lies between two values, e.g., 1 and 3 by computing $\mathbb{P}_X(1.25 < X \leq 3) = F_X(3) - F_X(1.25) = \frac{1}{2}$.

with $x \in \mathbb{R}$. This means, that the CDF of a discrete random variable can be evaluated on two different ways: via the probability measure $\mathbb{P}$ from $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ or via the image measure $\mathbb{P}_X$ of the probability space $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mathbb{P}_X)$.

EXAMPLE 2.68 (Cumulative Distribution Function of a Discrete Random Variable) *Let us consider once more our random experiment of flipping a coin 3-times. Based on the probability space $(\{0,1\}^3, \mathfrak{P}(\{0,1\}^3), \mathbb{P})$, e.g. $F_X(1.7)$ can be interpreted as the probability that no more than a single one occurs, that is,*

Probability Space (163)

$$\begin{aligned}
F_X(1.7) &\overset{(2.575)}{=} \sum_{\xi \leq 1.7} \mathbb{P}_X(X = \xi) && (2.580)\\
&\overset{\xi \in \{0,1\}}{=} \mathbb{P}_X(X = 0) + \mathbb{P}_X(X = 1) && (2.581)\\
&\overset{\xi \in \{0,1\}}{=} \mathbb{P}(\{\omega \,|\, X(\omega) = 0\}) + \mathbb{P}(\{\omega \,|\, X(\omega) = 1\}) && (2.582)\\
&= \mathbb{P}(\{000\}) + \mathbb{P}(\{001, 010, 100\}) && (2.583)\\
&\overset{\mathbb{P}(\{\omega\}) = \frac{1}{8}}{=} \frac{1}{2}. && (2.584)
\end{aligned}$$

CUMULATIVE DISTRIBUTION FUNCTION OF A CONTINUOUS RANDOM VARIABLE. From measure theory we know that any finite or countably infinite set is a null set, i.e. it has measure

Countable Set (827)

zero. Transferred to probability spaces, constructed over uncountably infinite base sets, this means that events, which can occur surely, would be assigned the probability zero. This is one of the reasons why we cannot define the distribution function of a continuous random variable in the same way as we did it with the cumulative distribution function of a discrete random variable. Here, we need a further mathematical concept that enables us to define such a CDF: the construct of the *probability density function*.

Null Set (80) Uncountable Set (827) CDF (174)

DEFINITION 2.47 (Probability Density Function) *Let $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ be a probability space, X a continuous random variable defined on $(\Omega, \mathfrak{F}(\Omega))$, and $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ a measurable space, with two measures $\mathbb{P}_X$ and μ, where μ is the Lebesgue measure on the real axis. As a consequence of the* Radon-Nikodým Theorem[6]*, we can construct a non-negative, measurable function p_X via the* Radon-Nikodým derivative *with respect to the measures $\mathbb{P}_X$ and μ, by:*

Probability Space (163) Measurable Space (80) Measurable Function (98)

$$p_X = \frac{d\mathbb{P}_X}{d\mu}. \tag{2.585}$$

The function p_X is called the probability density function *with respect to the random variable X, most often also simply denoted as the* density*, or the* PDF *of the random variable X. Since p_X is defined via the probability measure $\mathbb{P}_X$, a probability density function p_X satisfies the following conditions*

Image Measure (170)

i) $p_X(x) \geq 0$

ii) $\int_{(-\infty,\infty)} p_X(x)\, d\mu(x) = 1$

for all $x \in (-\infty, \infty)$, see Figure 2.47

REMARK 2.70 *As a consequence from Definition 2.47 we can derive the following useful relation*

$$\int_B p_X(x)\, d\mu(x) = \int_B d\mathbb{P}_X(x) = \mathbb{P}_X(B) \qquad \forall B \in \mathfrak{B}(\mathbb{R}). \tag{2.586}$$

We will now present two simple probability density functions, which play an import role in our further considerations, and which we use in the following again and again.

EXAMPLE 2.69 (Uniformly Distributed Random Variables on $\mathbb{R}$) *We are interested in uniformly distributed random numbers from a finite interval of the real number axis. How we have to define the associated PDF?*

[6] Assuming $(\mathcal{R}, \mathfrak{R}, \nu)$ to be a measure space with a σ-finite measure ν and ν' an absolute continuous measure ν, then there exist an—apart from a zero set—unique determined integrable function $f : \mathcal{R} \longrightarrow \overline{\mathbb{R}}$ with $\nu'(B) = \int_B f(x) d\nu(x)$, $\forall B \in \mathfrak{R}$.

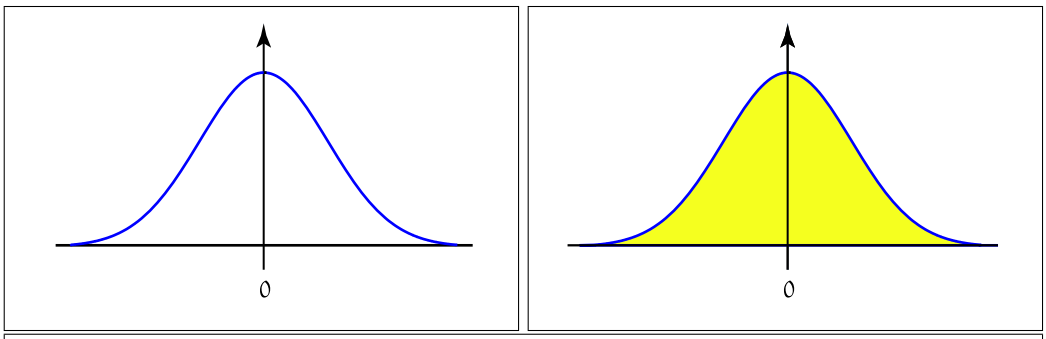


FIGURE 2.47: PROBABILITY DENSITY FUNCTION. Based on the image measure $\mathbb{P}_X$, a PDF satisfies the non-negativity property of a measure, i.e. $p_X(x) \geq 0, \forall x \in \text{Im}(X)$, illustrated in the left image, and the normalization property of a probability measure, $\int_{(-\infty,\infty)} p_X(x)\, d\mu(x) = 1$, shown in the right image.

For that purpose, let $([0,1], \mathfrak{B}([0,1]), \mu)$ *be the canonical probability space, where the probability measure* $\mathbb{P}$ *corresponds to the Lebesgue measure* μ.

i) A random variable X *defined by* $X(\omega) = \omega$ *maps the base set* $[0,1]$ *of the probability space* $([0,1], \mathfrak{B}([0,1]), \mu)$ *onto* $[0,1] \in \mathfrak{B}([0,1])$. *Then, the image measure on the measurable space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ *induced by* X *should be defined as:*

Measurable Space (80)

$$\mathbb{P}_X(B) \overset{(2.549)}{=} \mathbb{P}\underbrace{\left(X^{-1}(B)\right)}_{B\cap[0,1]} \tag{2.587}$$

$$\overset{\mathbb{P}=\mu}{=} \mu(B \cap [0,1]) \tag{2.588}$$

$$= \int_{B\cap[0,1]} d\mu(x) \tag{2.589}$$

for $B \in \mathfrak{B}(\mathbb{R})$. *Due to the normalization property of a PDF, we then get:*

$$\mathbb{P}_X((-\infty,\infty)) \overset{(2.586)}{=} \int_{(-\infty,\infty)} p_X(x) d\mu(x) \tag{2.590}$$

$$= \int_{(-\infty,\infty)\cap[0,1]} d\mu(x) = 1, \tag{2.591}$$

from which we conclude that the associated probability density function must be defined as:

$$p_X(x) = \begin{cases} 1 & : \text{ for } x \in [0,1] \\ 0 & : \text{ otherwise.} \end{cases} \tag{2.592}$$

ii) Let us now consider the random variable $Y \equiv f(X) = a + X(b-a)$ *with* $a, b \in \mathbb{R}$, *then the image of the base set* $[0,1]$ *of the probability space* $([0,1], \mathfrak{B}([0,1]), \mu)$ *under*

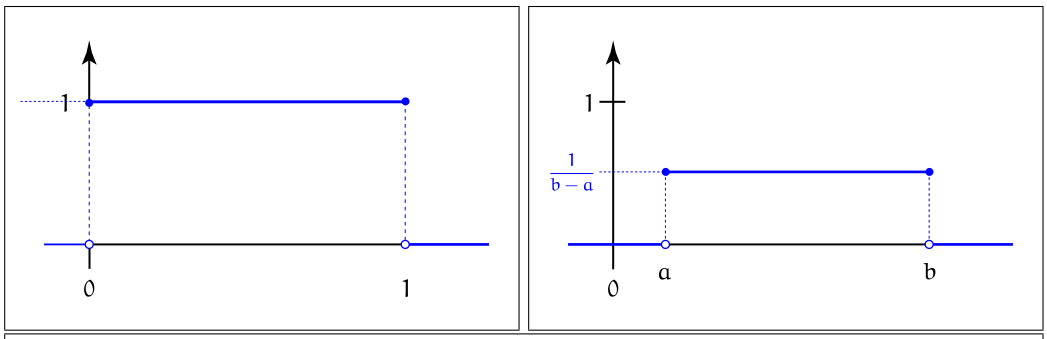


FIGURE 2.48: PROBABILITY DENSITY FUNCTIONS FOR UNIFORM SAMPLING FINITE INTERVALS FROM $\mathbb{R}$. Left, we have a PDF $p_X(x) = 1$ associated with the canonical probability space $([0,1], \mathfrak{B}([0,1]), \mu)$, right, the PDF $p_Y(y) = \frac{1}{b-a}$ associated with the canonical probability space $([a,b], \mathfrak{B}([a,b]), \mu)$.

the mapping Y *is the interval* $[a,b] \in \mathfrak{B}([a,b])$. *With the same arguments from above,*
Image Measure (170) *the image measure* $\mathbb{P}_Y$ *on the probability space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mathbb{P}_Y)$ *must be given by:*

$$\begin{aligned} \mathbb{P}_Y(B) &\overset{def}{=} && \mu(B \cap [f(0), f(1)]) && (2.593)\\ &\underset{f(X)=a+X(b-a)}{=} && \mu(B \cap [a,b]) && (2.594)\\ &= && \int_{B\cap[a,b]} d\mu(y) && (2.595) \end{aligned}$$

for $B \in \mathfrak{B}(\mathbb{R})$, *that is, the associated probability density function* p_Y *is given by:*

$$p_Y(y) = \begin{cases} \frac{1}{b-a} & : \text{ for } y \in [a,b] \\ 0 & : \text{ otherwise} \end{cases} \tag{2.596}$$

as it holds:

$$\begin{aligned} \int_{(-\infty,\infty)} p_Y(y) d\mu(y) &\overset{(2.596)}{=} && \int_{[a,b]} p_Y(y) d\mu(y) && (2.597)\\ &= && \frac{1}{b-a} \int_{[a,b]} d\mu(y) && (2.598)\\ &= && \frac{1}{b-a} y\Big|_a^b = 1. && (2.599) \end{aligned}$$

For an illustration of p_X *and* p_Y, *see Figure 2.48.*

REMARK 2.71 (A Method for Constructing Probability Density Functions) *Let* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mu)$
Measure Space (80) *be a measure space and* q *a measurable, non-negative function defined on a subset* B
Measurable Function (98) *of* $\mathbb{R}$. *Then, the function* p *defined by:*

$$p(x) \overset{\text{def}}{=} \frac{q(x)}{\int_{(-\infty,\infty)} q(x)\, d\mu(x)} \tag{2.600}$$

is a probability density function with respect to the Lebesgue measure μ. Measure (79)

DEFINITION 2.48 (Cumulative Distribution Function of a Continuous Random Variable) *Let us assume* X *be a continuous random variable defined on the continuous probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *with associated probability density function* p_X. *Due to Definition 2.44 the following holds to the* cumulative distribution function F_X *of the random variable* X Probability Space (163)

$$\begin{aligned} F_X(x) &\overset{(2.575)}{=} \mathbb{P}_X((-\infty, x]) && (2.601)\\ &= \int_{(-\infty,x]} d\mathbb{P}_X(\xi) && (2.602)\\ &\overset{(2.585)}{=} \int_{(-\infty,x]} p_X(\xi)\, d\mu(\xi), && (2.603) \end{aligned}$$

where μ *is the Lebesgue integral on* $\mathbb{R}$. *That is, if the measure* $\mathbb{P}$ *is* μ*-differentiable with Radon-Nikodým derivative* p_X, *then* F_X *can be computed by integrating* p_X *with respect to measure* μ *over the interval* $(-\infty, x]$.

For an illustration of a CDF associated with a continuous random variable, see Figure 2.49

REMARK 2.72 *Due to the definition of the image measure* $\mathbb{P}_X$, *the continuous cumulative distribution function* F_X *can also be expressed in terms of the probability measure* $\mathbb{P}$, *that is, we can also use the formula* Image Measure (170)

$$\begin{aligned} F_X(x) &\overset{(2.575)}{=} \mathbb{P}_X((-\infty, x]) && (2.604)\\ &\overset{(2.551)}{=} \mathbb{P}\left(X^{-1}((-\infty, x])\right) && (2.605)\\ &\overset{(2.557)}{=} \int_{\omega \leq X^{-1}(-\infty,x]} d\mathbb{P}(\omega). && (2.606) \end{aligned}$$

with $x \in \mathbb{R}$. *This means, that the CDF of a continuous random variable can be evaluated on two different ways: via the probability measure* $\mathbb{P}$ *from* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *or via the image measure* $\mathbb{P}_X$ *of the probability space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mathbb{P}_X)$.

Based on the definition of the distribution function F_X we can now define the probability that the value of a continuous random variable X lies between two points $a, b \in \mathbb{R}$

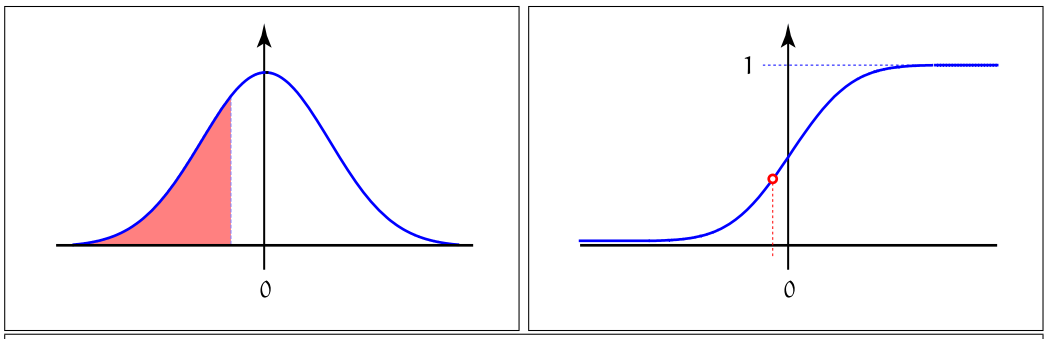


FIGURE 2.49: CUMULATIVE DISTRIBUTION FUNCTION OF A CONTINUOUS RANDOM VARIABLE. A PDF with associated CDF. As you can easily see from the image, the illustrated CDF satisfies the condition required to a CDF from Definition 2.44. So, the image range of the function corresponds to the interval $[0,1]$ and you can detect the continuous increasing of function values from the graph. The function also satisfies the limiting properties required to a CDF, namely $F_X(x) \to 0$ for sufficiently small values and $F_X(x) \to 1$ for sufficiently large argumets.

by:

$$\begin{aligned}
\text{prob}(a < X \leq b) &\overset{\text{def}}{=} \mathbb{P}_X((a,b]) && (2.607)\\
&= \int_{(a,b]} d\mathbb{P}_X(x) && (2.608)\\
&= \int_{(-\infty,b]} d\mathbb{P}_X(x) - \int_{(-\infty,a]} d\mathbb{P}_X(x) && (2.609)\\
&= F_X(b) - F_X(a). && (2.610)
\end{aligned}$$

Now, we demonstrate these newly defined concepts with the help of a simple example: the uniform distribution of a random variable on the interval $[a,b]$.

Section 6.5 EXAMPLE 2.70 (CDF of a Uniformly Distributed Random Variable on $[a,b]$) *Let X be a random variable defined on the probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$, whose values are uniformly distributed within the interval $[a,b] \in \mathfrak{B}(\mathbb{R})$. Since the values of X are uniformly distributed within $[a,b]$, we call a random variable of this type a* uniformly distributed random variable.

As we know from part ii) of Example 2.69, the probability density function of a uniformly distributed random variable X on the interval $[a,b]$ with $a,b \in \mathbb{R}$ is given by:

$$p_X(x) = \begin{cases} \frac{1}{b-a} & : \text{ for } x \in [a,b] \\ 0 & : \text{ otherwise.} \end{cases} \qquad (2.611)$$

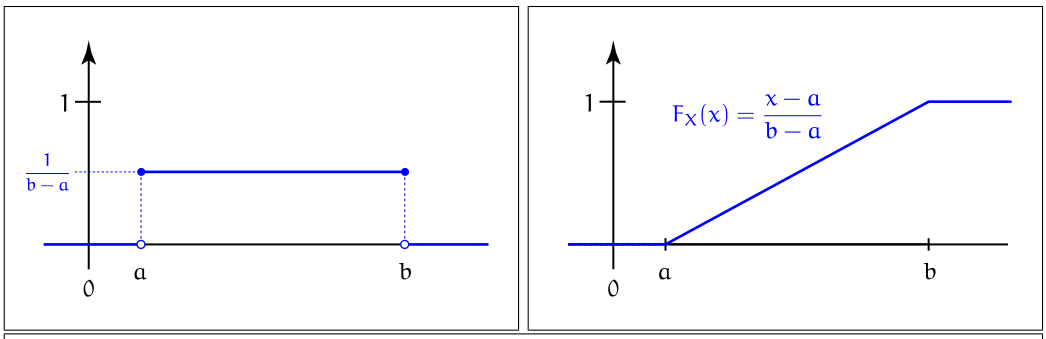


FIGURE 2.50: UNIFORM DISTRIBUTION ON $[a, b]$. Based on the probability space $([a, b], \mathfrak{B}([a, b]), \mathbb{P})$, on the left-hand side the plot of the PDF $p_X(x) = \frac{1}{b-a}$ and besides the corresponding CDF $F_X(x) = \frac{x-a}{b-a}$.

In terms of the probability density function p_X the associated cumulative distribution function F_X can then be written as:

$$\begin{aligned} F_X(x) &= \int_{[a,x]} p_X(\xi)\, d\mu(\xi) && (2.612)\\ &= \frac{1}{b-a}\int_{[a,x]} d\mu(\xi) && (2.613)\\ &= \frac{x-a}{b-a}, && (2.614) \end{aligned}$$

see Figure 2.50.

Based on this CDF, the probability that a point is drawn in an interval $(\alpha, \beta] \subset [a, b]$ can be computed via

$$\begin{aligned} \text{prob}\,(\alpha < X \leq \beta) &= \frac{1}{b-a}\int_{(\alpha,\beta]} d\mu(x) && (2.615)\\ &= \frac{\beta-\alpha}{b-a} && (2.616)\\ &= \frac{\mu([\alpha,\beta])}{\mu([a,b])}. && (2.617) \end{aligned}$$

We will obtain the same result if we choose the following approach

$$(\alpha, \beta] = (-\infty, \beta] \setminus (-\infty, \alpha] \quad (2.618)$$

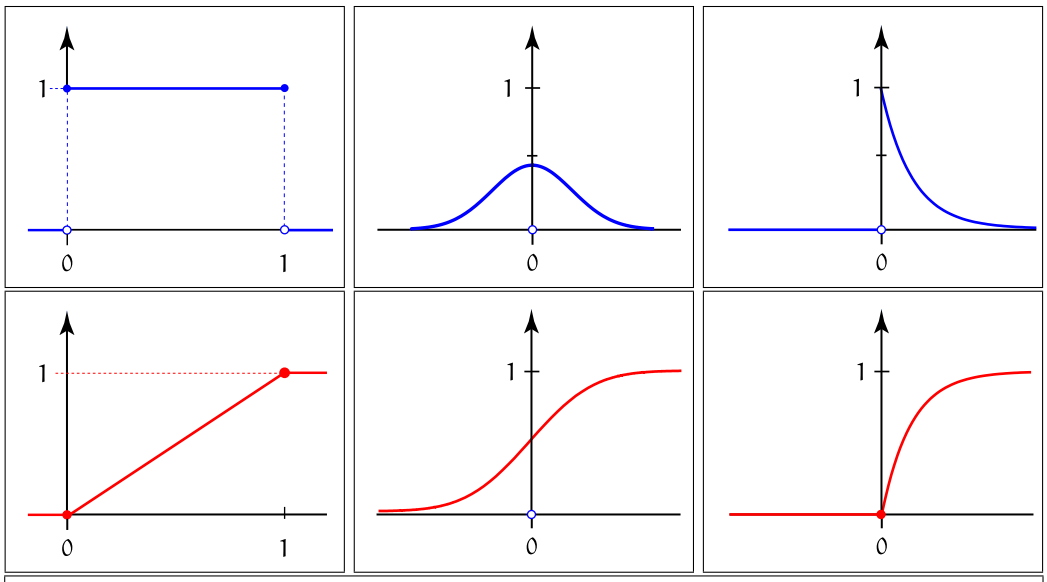


FIGURE 2.51: IMPORTANT PDFS AND CDFS. Left, PDF and CDF of a uniformly distributed random variable, in the center, the PDF and CDF of a normally distributed random variable, and right, the PDF and CDF of an exponentially distributed random variable.

and then compute the corresponding probability via

$$\begin{aligned}
\text{prob}\,(\alpha < X \leq \beta) &= \mathbb{P}_X(\alpha, \beta) && (2.619)\\
&= \mathbb{P}_X((-\infty, \beta]) - \mathbb{P}_X((-\infty, \alpha]) && (2.620)\\
&= F_X(\beta) - F_X(\alpha) && (2.621)\\
&= \frac{\beta}{b-a} - \frac{\alpha}{b-a} && (2.622)\\
&= \frac{\beta - \alpha}{(b-a} && (2.623)\\
&= \frac{\mu([\alpha, \beta])}{\mu([a, b])}. && (2.624)
\end{aligned}$$

REMARK 2.73 (PDF as the Derivative of the CDF, Univariate Case) *Setting our focus once more on the Relations (2.604) - (2.603), then, due to the Fundamental Theorem of Calculus, we get for the representation of the density* p_X*:*

[174, Rudin 1998]

$$p_X(x) = \frac{dF_X(x)}{d\mu(x)}, \tag{2.625}$$

i.e. the probability density function p_X *is the derivative of the cumulative distribution function of a continuous random variable with respect to the Lebesgue measure.*

2.4.3 RANDOM VECTORS AND DISTRIBUTION FUNCTIONS

As we will see, Monte Carlo integration is the preferred method for integrating functions over high-dimensional domains. It requires sampling strategies on probability spaces which are composed as products of already well-known probability spaces, such as for example: the drawing of random numbers from the unit square or the unit cube, the hemisphere or the unit sphere. Hence, it makes sense to extend the concepts of the random variable and the distribution function in particular to product probability spaces.

Chapter 6
Section 6.5
Probability Space (163)
Random Variable (168)
Product Measure Space (81)

Based on the *probability space* $(\Omega, \mathfrak{F}, \mathbb{P})$ we continue our excursion into probability theory. Thus, we introduce the construct of the *random vector* as a *measurable function* of the probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ into the measurable product space $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$. Interpreting the existence and uniqueness of a measurable function defined on Ω as a random vector, we get *the s-dimensional probability density* as a consequence of the *Radon-Nikodým Theorem*. We then define the concept of the *probability distribution function* of a random vector, for the discrete and the continuous case, and show how it can be used to choose random points from the unit interval $[0, 1]^s$, as well as the unit sphere and the upper hemisphere.

Probability Space (163)
Measurable Function (98)
Probability Density Function (189)
Probability Distribution (163)

RANDOM VECTORS. In analogy to the concept of the random variable from the previous section, we present now the concept of the random vector as a measurable function defined on a probability space with values in the s-dimensional Euclidean space $\mathbb{R}^s$.

DEFINITION 2.49 (Random Vector) *Let $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ be a probability space and let X_i be measurable functions*

Probability Space (163)
Measurable Function (98)

$$X_i : (\Omega, \mathfrak{F}(\Omega)) \longrightarrow (\mathbb{R}, \mathfrak{B}(\mathbb{R})) \tag{2.626}$$

with

$$\Omega \ni \omega \longmapsto X_i(\omega) \in \mathbb{R}, \tag{2.627}$$

then $\mathbf{X} = (X_1, \ldots, X_s)$ is called a multivariate *or an* s-dimensional random variable *in the following often also simply denoted as a* random vector.

REMARK 2.74 *As is easily seen, a random vector can also be interpreted as a measurable function*

Measurable Function (98)

$$\mathbf{X} : (\Omega, \mathfrak{F}(\Omega)) \longrightarrow (\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s)) \tag{2.628}$$

with

$$\Omega \ni \omega \longmapsto \mathbf{X}(\omega) = (X_1(\omega), \ldots, X_s(\omega)) \in \mathbb{R}^s \tag{2.629}$$

defined on the probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$.

THE CUMULATIVE DISTRIBUTION FUNCTION OF A RANDOM VECTOR. In accordance with the definition of the image measure of a random variable, now, we are able to define the image measure $\mathbb{P}_{\mathbf{X}}$ induced by the random vector $\mathbf{X}$.

Image Measure (170)

DEFINITION 2.50 (The Joint Image Measure $\mathbb{P}_{\mathbf{X}}$ of a Random Vector $\mathbf{X}$) *Let $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ be a probability space and let $\mathbf{X}$ be a random vector from $(\Omega, \mathfrak{F}(\Omega))$ into $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$. Then, $\mathbf{X}$ implies a measure $\mathbb{P}_{\mathbf{X}}$:*

Probability Space (163)

$$\mathbb{P}_{\mathbf{X}} : (\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s)) \to [0, 1] \tag{2.630}$$

on the measurable space $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$, which we denote as the joint image measure *of the random vector $\mathbf{X}$. Based on the concept of a probability measure, $\mathbb{P}_{\mathbf{X}}$ is defined by:*

Measurable Space (80)
Probability Measure (80)

$$\begin{aligned} \mathbb{P}_{\mathbf{X}}(\mathbf{B}) &\stackrel{\text{def}}{=} \left(\mathbb{P} \circ \mathbf{X}^{-1}\right)(\mathbf{B}), && (2.631)\\ &= \mathbb{P}\left(\{\omega \in \Omega \,|\, \mathbf{X}(\omega) \in \mathbf{B}\}\right) && (2.632) \end{aligned}$$

for any $\mathbf{B} \in \mathfrak{B}(\mathbb{R}^s)$. $\mathbb{P}_{\mathbf{X}}$ is also denoted as the joint probability distribution of the random vector $\mathbf{X}$.

REMARK 2.75 *Note, that $\mathbb{P}_X$ is a probability distribution over the observation space $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$ instead of $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$.*

REMARK 2.76 *In the following, we will often use the notion of the random vector and the random variable synonymously if it is clear from context which of the two concepts should be used.*

Based on the joint image measure of a random vector, now, we are able to construct the cumulative distribution function of a random vector in a similar way as we did it for a random variable.

DEFINITION 2.51 (Joint Cumulative Distribution Function of a Random Vector) *Let us assume, that $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ is a probability space and $\mathbf{X} = (X_1, \ldots, X_n)$ is a measurable function, thus a random vector from $(\Omega, \mathfrak{F}(\Omega))$ to $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$. A non-descending, right continuous function $F_{\mathbf{X}}$:*

Probability Space (163)
Measurable Function (98)
Continuous Function (869)

$$F_{\mathbf{X}} : \mathbb{R}^s \longrightarrow [0, 1] \tag{2.633}$$

defined by

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &\stackrel{\text{def}}{=} \mathbb{P}_{\mathbf{X}}((-\infty, \mathbf{x}]) && (2.634)\\ &= \mathbb{P}(\{\omega \in \Omega \;|\, \mathbf{X}(\omega) \leq \mathbf{x}\}) && (2.635) \end{aligned}$$

with $(-\infty, \mathbf{x}] = (-\infty, x]^s$ and the properties

$$\lim_{x_1 \to -\infty, \ldots, x_s \to -\infty} F_{\mathbf{X}}(x_1, \ldots, x_s) = 0 \tag{2.636}$$

as well as

$$\lim_{x_1 \to \infty, \ldots, x_s \to \infty} F_{\mathbf{X}}(x_1, \ldots, x_s) = 1 \tag{2.637}$$

is denoted as the joint cumulative distribution function *of* $\mathbf{X}$*, or briefly the joint distribution function of* $\mathbf{X}$*, abbreviated also simply as the* joint CDF *of* $\mathbf{X}$*. Obviously,* $F_{\mathbf{X}}(\mathbf{x})$ *denotes the probability that the random variable* $\mathbf{X}$ *takes on a value which will be less than or equal to* $\mathbf{x}$*.*

THE CUMULATIVE DISTRIBUTION FUNCTION OF A DISCRETE RANDOM VECTOR. As in the univariate case, we also distinguish in the multivariate case between two different types of random vectors: *discrete* and *continuous random vectors.*

DEFINITION 2.52 (Joint Probability Mass Function of a Discrete Random Vector) *Let us assume* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *be a discrete probability space and* $\mathbf{X}$ *a discrete random vector defined on* $(\Omega, \mathfrak{F}(\Omega))$ *into the measurable product space* $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$*. Then, we define the* joint probability mass function*,* $p_{\mathbf{X}}$*, of the random vector* $\mathbf{X}$ *by:*

Discrete Probability Space (163) Discrete Random Vector (185) Measurable Space (80)

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &\equiv p_{X_1,\ldots,X_s}(x_1, \ldots x_s) && (2.638)\\ &\stackrel{def}{=} \mathbb{P}_{\mathbf{X}}(\mathbf{X} = \mathbf{x}) && (2.639)\\ &\equiv \mathbb{P}_{X_1,\ldots,X_s}(X_1 = x_1, \ldots, X_s = x_s) && (2.640) \end{aligned}$$

for all $\mathbf{x} \in \mathrm{Im}(\mathbf{X})$*. Since* $p_{\mathbf{X}}$ *is defined via the probability measure* $\mathbb{P}$*, a joint probability mass function* $p_{\mathbf{X}}$ *satisfies the following conditions:*

i) $p_{\mathbf{X}}(\mathbf{x}) \geq 0$,

ii) $\sum_{\mathbf{x} \in \mathrm{Im}(\mathbf{X})} p_{\mathbf{X}}(\mathbf{x}) \equiv \sum_{x_1 \in \mathrm{Im}(X_1)} \cdots \sum_{x_s \in \mathrm{Im}(X_s)} p_{X_1,\ldots,X_s}(x_1, \ldots x_s) = 1$

for all $\mathbf{x} = (x_1, \ldots, x_s)$ *from the image area of* $\mathbf{X}$*, see Figure 2.52.*

REMARK 2.77 *Due to the definition of the image measure* $\mathbb{P}_{\mathbf{X}}$*, a joint probability mass function can also be expressed in terms of the probability measure* $\mathbb{P}$*, that is, we can also use the formula*

Image Measure (170)

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &\stackrel{def}{=} \mathbb{P}(\{\omega \in \Omega \,|\, \mathbf{X}(\omega) = \mathbf{x}\}) && (2.641)\\ &= \mathbb{P}(\{\omega \in \Omega \,|\, X_1(\omega) = x_1, \ldots, X_s(\omega) = x_s\}. && (2.642) \end{aligned}$$

for computing the probability distribution of a discrete random variable. This means, that a joint PMF can be evaluated on two different ways: via the probability measure $\mathbb{P}$ *from* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *or via the image measure* $\mathbb{P}_{\mathbf{X}}$ *of the probability space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mathbb{P}_{\mathbf{X}})$*.*

Now, let us illustrate the concept of a joint probability mass function by means of a simple example.

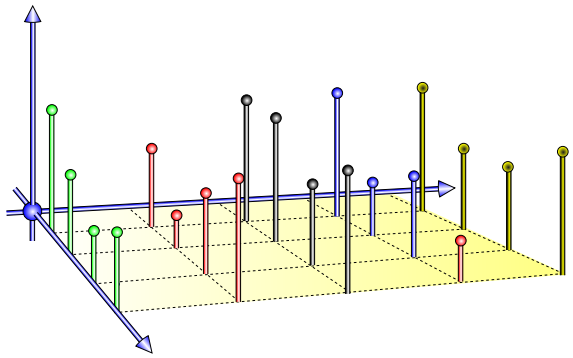


FIGURE 2.52: JOINT PROBABILITY MASS FUNCTION. Based on the image measure $\mathbb{P}_{\mathbf{X}}$, a joint PMF satisfies the non-negativity property of a measure, i.e., $p_{\mathbf{X}}(\mathbf{x}) \geq 0, \forall \mathbf{x} \in \mathrm{Im}(\mathbf{X})$ and the normalization property of a probability measure, $\sum_{\mathbf{x} \in \mathrm{Im}(\mathbf{X})} p_{\mathbf{X}}(\mathbf{x}) = 1$.

EXAMPLE 2.71 (A Joint Probability Mass Function Induced by a Discrete Random Vector) *With respect to our example of flipping a coin 3-times, let us suppose that we are interested in the number of ones from the first two tosses and the number of ones from the third toss. We can model this experiment by a random vector* $\mathbf{X} = (X_1, X_2)$, *where* X_1 *takes on the values* $\Omega_1 = \{0, 1, 2\}$ *and the image range of* X_2 *is the set* $\omega_2 = \{0, 1\}$, *that is:* $\mathbf{X} = (X_1, X_2)$ *is a measurable function from probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *with* $\Omega = \{0, 1\}^3$ *onto* $(\mathbb{R}^2, \mathfrak{B}(\mathbb{R}^2))$, *as it holds:*

Measurable Function (98)

$$\{X_1 < a_1, X_2 < a_2\} = \begin{cases} \emptyset & \text{if } a_1 \leq 0 \\ \emptyset & \text{if } a_2 \leq 0 \\ \{000\} & \text{if } a_1 \leq 1 \text{ and } a_2 \leq 1 \\ \{00\} \times \Omega_2 & \text{if } a_1 \leq 1 \text{ and } 1 < a_2 \\ \{000\}, \{010\}, \{100\} & \text{if } a_1 \leq 2 \text{ and } a_2 \leq 1 \\ \{00\} \times \Omega_2, \{01\} \times \Omega_2, \{10\} \times \Omega_2 & \text{if } a_1 \leq 2 \text{ and } 1 < a_2 \\ \Omega_1 \times \{0\} & \text{if } 2 < a_1 \text{ and } a_2 \leq 1 \\ \Omega_1 \times \Omega_2 & \text{if } 2 < a_1 \text{ and } 1 < a_2. \end{cases} \tag{2.643}$$

Product Measure (81)

Due to the definition of a product measure, the joint image measure $\mathbb{P}_{\mathbf{X}}(X_1 = 1, X_2 = 0)$—*thus, the probability that exactly a single one occurs in the two first tosses*

and no one occurs in the third toss—has the form

$$\begin{aligned}
\mathbb{P}_{\mathbf{X}}(\mathbf{X}=\mathbf{x}) &= \mathbb{P}_{X_1,X_2}(X_1=1,X_2=0) && (2.644)\\
&= \mathbb{P}(\{\omega \,|\, X_1(\omega)=1 \wedge X_2(\omega)=0\}) && (2.645)\\
&= \mathbb{P}(\{010\},\{100\}) && (2.646)\\
&= \frac{1}{4}, && (2.647)
\end{aligned}$$

where it holds $(\mathbf{X}=\mathbf{x}) = (X_1=1, X_2=0)$.

Based on the probability mass function, we are now ready, to define the cumulative distribution function of a discrete random vector. It makes a statement about the probability, that the value of a random vector is less than or equal to some real number.

DEFINITION 2.53 (Joint Cumulative Distribution Function of a Discrete Random Vector)
Let $\mathbf{X}$ *be a discrete random vector on the discrete probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *with joint probability mass function* $p_{\mathbf{X}}$. *Due to Definition 2.51, the* joint cumulative distribution function $F_{\mathbf{X}}$ *of the* discrete random vector $\mathbf{X}$ *can be written as:*

Discrete Random Variable (168)
Discrete Probability Space (163)

$$\begin{aligned}
F_{\mathbf{X}}(\mathbf{x}) &\overset{def}{=} \mathbb{P}_{\mathbf{X}}(\mathbf{X}\leq\mathbf{x}) && (2.648)\\
&= \sum_{\xi\leq\mathbf{x}} \mathbb{P}_{\mathbf{X}}(\mathbf{X}=\xi) && (2.649)\\
&\equiv \sum_{\xi_1\leq x_1} \cdots \sum_{\xi_s\leq x_s} \mathbb{P}_{X_1,\ldots,X_s}(X_1=\xi_1,\ldots,X_n=\xi_s) && (2.650)\\
&\overset{(2.641)}{=} \sum_{\xi\leq\mathbf{x}} p_{\mathbf{X}}(\xi) && (2.651)\\
&\equiv \sum_{\xi_1\leq x_1} \cdots \sum_{\xi_s\leq x_s} p_{X_1,\ldots,X_s}(\xi_1,\ldots,\xi_s). && (2.652)
\end{aligned}$$

REMARK 2.78 *Due to the definition of the image measure* $\mathbb{P}_{\mathbf{X}}$, *the discrete cumulative distribution function,* $F_{\mathbf{X}}$, *can also be expressed in terms of the probability measure* $\mathbb{P}$, *that is, we can also use the formula*

Image Measure (170)

$$\begin{aligned}
F_{\mathbf{X}}(\mathbf{x}) &= \sum_{\xi_1\leq x_1} \cdots \sum_{\xi_s\leq x_s} \mathbb{P}(\{\omega\in\Omega \,|\, X_1(\omega)=\xi_1,\ldots,X_s(\omega)=\xi_s\}) && (2.653)\\
&\equiv \sum_{\xi\leq\mathbf{x}} \mathbb{P}\left(\{\omega\in\Omega \,|\, \mathbf{X}(\omega)=\xi\}\right) && (2.654)\\
&= \mathbb{P}\left(\{\omega\in\Omega \,|\, \mathbf{X}(\omega)\leq\mathbf{x}\}\right). && (2.655)
\end{aligned}$$

with $\mathbf{x}\in\mathbb{R}^s$. *This means, that the CDF of a discrete random vector can be evaluated on two different ways: via the probability measure* $\mathbb{P}$ *from* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *or via the image measure* $\mathbb{P}_{\mathbf{X}}$ *of the probability space* $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s), \mathbb{P}_{\mathbf{X}})$.

EXAMPLE 2.72 (Joint Cumulative Distribution Function of a Discrete Random Vector)
Joint Image Measure (184)
Discrete Random Variable (168)
Based on our considerations with respect to the joint image measure of the discrete random vector $\mathbb{P}_{\mathbf{X}}$ from Example 2.71, the value of the joint cumulative distribution function $F_{X_1,X_2}(1.2, 1.7)$ can be interpreted as the probability that no more than a single one occurs in the first two tosses. In particular $F_{\mathbf{X}}(1.2, 1.7)$ can be computed via:

$$\begin{aligned}
F_{X_1,X_2}(1.2,1.7) &= \sum_{x_1\le 1.2}\sum_{x_2\le 1.7}\mathbb{P}_{X_1,X_2}(X_1=x_1,X_2=x_2) && (2.656)\\
&= \mathbb{P}_{X_1,X_2}(X_1=0,X_2=0)+\mathbb{P}_{X_1,X_2}(X_1=0,X_2=1)+ && (2.657)\\
&\quad \mathbb{P}_{X_1,X_2}(X_1=1,X_2=0)+\mathbb{P}_{X_1,X_2}(X_1=1,X_2=1) && (2.658)\\
&= \mathbb{P}(\{\omega\,|\,X_1(\omega)=0,X_2(\omega)=0\})+ && (2.659)\\
&\quad \mathbb{P}(\{\omega\,|\,X_1(\omega)=0,X_2(\omega)=1\})+ && (2.660)\\
&\quad \mathbb{P}(\{\omega\,|\,X_1(\omega)=1,X_2(\omega)=0\})+ && (2.661)\\
&\quad \mathbb{P}(\{\omega\,|\,X_1(\omega)=1,X_2(\omega)=1\}) && (2.662)\\
&= \mathbb{P}(\{000\})+\mathbb{P}(\{001\})+\mathbb{P}(\{010,100\})+\mathbb{P}(\{011,101\}) && (2.663)\\
&\overset{\mathbb{P}(\{\omega\})=\frac{1}{8}}{=} \frac{3}{4}. && (2.664)
\end{aligned}$$

Often, we have to sample from complex joint probability mass functions. If it is possible to isolate a variable in such a complex probability function, then we get samples distributed according to much a more simpler probability function. Linked with the concept of conditional probability, this technique plays an important role in many sampling procedures. The concept behind this idea is the construct of the marginal probability mass function.

REMARK 2.79 (The Marginal Probability Mass Function) *Let $\mathbf{X}=(X_1,\dots,X_s)$ be a discrete s-dimensional random vector and $p_{\mathbf{x}}$ its associated joint probability mass function. Then we call the function $p_{X_i}, 1\le i\le s$, given by:*

$$x_i \mapsto p_{X_i}(x_i) \overset{def}{=} \sum_{\xi_1\le\infty}\dots\sum_{\xi_{i-1}\le\infty}\sum_{\xi_{i+1}\le\infty}\dots\sum_{\xi_s\le\infty} p_{X_1,\dots,X_s}(\xi_1,\dots,x_i,\dots\xi_s), \quad (2.665)$$

the marginal probability mass function, *also briefly denoted as the* marginal PMF.

For our discussion, the above general case of the marginal probability mass function of an s-dimensional random vector is not of particularly interest. We are rather more interested in the case $s=2$. Then, the associated marginal PMFs, p_{X_1} and p_{X_2}, of a random vector $\mathbf{X}=(X_1,X_2)$ are given by the formulas:

$$p_{X_1}(x_1) = \sum_{\xi_2\le\infty} p_{X_1,X_2}(x_1,\xi_2) \quad (2.666)$$

and

$$p_{X_2}(x_2) = \sum_{\xi_1 \leq \infty} p_{X_1,X_2}(\xi_1, x_2). \tag{2.667}$$

THE CUMULATIVE DISTRIBUTION FUNCTION OF A CONTINUOUS RANDOM VECTOR. Due to the same reasons as in the univariate case, we cannot define a cumulative distribution function for a continuous random variable in the same way as we did it with the cumulative distribution function for a discrete random variable.

Section 2.4.2
CDF (171)

DEFINITION 2.54 (Joint Probability Density Function of a Continuous Random Vector) *Let us assume $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ be a continuous probability space, $\mathbf{X}$ a random vector defined on $(\Omega, \mathfrak{F}(\Omega))$ with values in the measurable space $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$ equipped with two measures $\mathbb{P}_X$ and μ^s, where μ^s is the s-dimensional Lebesgue measure. As a consequence of the Radon-Nikodým Theorem we can construct a non-negative, measurable function $p_{\mathbf{X}}$ via the Radon-Nikodým derivative with respect to the measures $\mathbb{P}_{\mathbf{X}}$ and μ^s, by:*

Probability Space (165)
Continuous Random Vector (185)
Measurable Space (80)
Radon-Nikodým Theorem (176)
Radon-Nikodým Derivative (176)

$$p_{\mathbf{X}} = \frac{d\mathbb{P}_{\mathbf{X}}}{d\mu^s}. \tag{2.668}$$

The function $p_{\mathbf{X}}$ is called the probability density function *with respect to the random vector $\mathbf{X}$, most often also simply called the* density *or the* PDF *of the random vector $\mathbf{X}$. Since $p_{\mathbf{X}}$ is defined via the probability measure, $\mathbb{P}_{\mathbf{X}}$, a probability density function, $p_{\mathbf{X}}$, satisfies the following conditions:*

Image Measure (170)

i) $p_{\mathbf{X}}(\mathbf{x}) \geq 0$

ii) $\int_{(-\infty,\infty)^s} p_{\mathbf{X}}(\mathbf{x}) d\mu^s(\mathbf{x}) = 1$

for all $\mathbf{x} \in (-\infty, \infty)^s$.

REMARK 2.80 *As a consequence from Definition 2.54 we can derive the following useful relation:*

$$\int_{\mathbf{B}} p_{\mathbf{X}}(\mathbf{x})\, d\mu^s(\mathbf{x}) = \int_{\mathbf{B}} d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}_{\mathbf{X}}(\mathbf{B}) \qquad \forall \mathbf{B} \in \mathfrak{B}(\mathbb{R}^s). \tag{2.669}$$

We will now present a few probability density functions, which we will use in the following again and again.

EXAMPLE 2.73 (Uniformly Distributed Random Vectors on $\mathbb{R}^s$) *We are interested in uniformly distributed random numbers from a finite s-dimensional interval over $\mathbb{R}^s$. How we have to define the associated PDF?*

For that purpose, let $([0,1]^s, \mathfrak{B}([0,1^s]), \mu^s)$ be the canonical probability space, where the probability measure $\mathbb{P}$ corresponds to the Lebesgue measure μ^s.

Canonical Probability Space (166)
μ^s (82)

i) Let us consider the random variable $\mathbf{X}(\omega) = \omega, \forall\, \omega \in [0,1]^s$*, which maps the base set* $[0,1]^s$ *of the probability space* $([0,1]^s, \mathfrak{B}([0,1^s]), \mu^s)$ *onto* $[0,1]^s \in \mathfrak{B}(\mathbb{R}^s)$*. Then, the image measure on the measurable space* $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$ *induced by* $\mathbf{X}$ *should be defined as:*

$$\begin{aligned}
\mathbb{P}_{\mathbf{X}}(\mathbf{B}) &\overset{(2.631)}{=} \mathbb{P}\underbrace{\left(\mathbf{X}^{-1}(\mathbf{B})\right)}_{\mathbf{B}\cap[0,1]^s} && (2.670)\\
&\overset{\mathbb{P}=\mu^s}{=} \mu^s(\mathbf{B}\cap[0,1]^s) && (2.671)\\
&= \int_{\mathbf{B}\cap[0,1]^s} d\mu^s(\mathbf{x}) && (2.672)
\end{aligned}$$

for $\mathbf{B} \in \mathfrak{B}(\mathbb{R}^s)$*. Due to the normalization property of a PDF, we then get:*

$$\mathbb{P}_{\mathbf{X}}((-\infty,\infty)^s) = \int_{(-\infty,\infty)^s} p_{\mathbf{X}}(\mathbf{x}) d\mu^s(\mathbf{x}) = \int_{(-\infty,\infty)^s\cap[0,1]^s} d\mu^s(\mathbf{x}) = 1, \tag{2.673}$$

from which we conclude that the associated probability density function must defined as:

$$p_{\mathbf{X}}(\mathbf{x}) = \begin{cases} 1 & : \quad \textit{for } \mathbf{x} \in [0,1]^s \\ 0 & : \quad \textit{otherwise.} \end{cases} \tag{2.674}$$

ii) Let us now consider the random variable $\mathbf{Y} \equiv f(\mathbf{X}) = \mathbf{a} + \mathbf{X}(\mathbf{b}-\mathbf{a})$ *with* $\mathbf{a},\mathbf{b} \in \mathbb{R}^s$*, then the image of the base set* $[0,1]^s$ *under the mapping* $\mathbf{Y}$ *is the interval* $[\mathbf{a},\mathbf{b}] \in \mathfrak{B}([\mathbf{a},\mathbf{b}])$*. With the same arguments as above, the joint image measure* $\mathbb{P}_{\mathbf{Y}}$ *on the probability space* $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s), \mathbb{P}_{\mathbf{Y}})$ *must be given by:*

Joint Image Measure (184)

$$\begin{aligned}
\mathbb{P}_{\mathbf{Y}}(B) &\overset{def}{=} \mu^s(B\cap[f(\mathbf{0}), f(\mathbf{1})]) && (2.675)\\
&= \int_{B\cap[\mathbf{a},\mathbf{b}]} d\mu^s(\mathbf{y}) && (2.676)
\end{aligned}$$

for $B \in \mathfrak{B}(\mathbb{R}^s)$*. Then, the associated probability density function* $p_{\mathbf{Y}}$ *is given by:*

$$p_{\mathbf{Y}}(\mathbf{y}) = \begin{cases} \frac{1}{\prod_{i=1}^s (b_i - a_i)} & : \quad \textit{for } \mathbf{y} \in [\mathbf{a},\mathbf{b}] \\ 0 & : \quad \textit{otherwise} \end{cases} \tag{2.677}$$

as it holds:

$$\begin{aligned}
\int_{(-\infty,\infty)^s} p_{\mathbf{Y}}(\mathbf{y}) d\mu^s(\mathbf{y}) &= \int_{[\mathbf{a},\mathbf{b}]} p_{\mathbf{Y}}(\mathbf{y}) d\mu^s(\mathbf{y}) && (2.678)\\
&= \frac{1}{\prod_{i=1}^s (b_i - a_i)} \int_{[\mathbf{a},\mathbf{b}]} d\mu^s(\mathbf{y}) && (2.679)\\
&= \frac{1}{\prod_{i=1}^s (b_i - a_i)} \mathbf{y}\Big|_{\mathbf{a}}^{\mathbf{b}} = 1. && (2.680)
\end{aligned}$$

REMARK 2.81 (A Method for Deriving Probability Density Functions, Multivariate Case) *Let $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s), \mu)$ be a measure space and q be a measurable, non-negative function defined on a subset $\mathbf{B}$ of $\mathbb{R}^s$, then the function p defined by:*

$$p(\mathbf{x}) \stackrel{\text{def}}{=} \frac{q(\mathbf{x})}{\int_{(-\infty,\infty)^s} q(\mathbf{x})\, d\mu^s(\mathbf{x})} \tag{2.681}$$

is a probability density function with respect to the Lebesgue measure μ^s. μ^s (82)

DEFINITION 2.55 (Joint Cumulative Distribution Function of a Continuous Random Vector) *Let us assume $\mathbf{X}$ be a continuous random vector on the probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$* Continuous Random Vector (185) *with probability density function $p_{\mathbf{X}}$. Due to Definition 2.51, the associated* joint cumulative distribution function $F_{\mathbf{X}}$ *of the* continuous random vector $\mathbf{X}$ *can be written as:*

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &\overset{(2.556)}{=} \mathbb{P}_{\mathbf{X}}((-\infty, \mathbf{x}]) && (2.682)\\ &= \int_{(-\infty,\mathbf{x}]} d\mathbb{P}_{\mathbf{X}}(\xi) && (2.683)\\ &\overset{(2.668)}{=} \int_{(-\infty,\mathbf{x}]} p_{\mathbf{X}}(\xi)\, d\mu^s(\xi) && (2.684)\\ &\equiv \int_{(-\infty,x_s]} \dots \int_{(-\infty,x_1]} p_{X_1,\dots,X_s}(\xi_1,\dots,\xi_s)\, d\mu(\xi_1) \dots d\mu(\xi_s), && (2.685) \end{aligned}$$

where $(-\infty, \mathbf{x}] \stackrel{\text{def}}{=} \times_{i=1}^{s} (-\infty, x_i]$. That is, if the image measure $\mathbb{P}_{\mathbf{X}}$ is μ-differentiable with Radon-Nikodým derivative $p_{\mathbf{X}}$, then $F_{\mathbf{X}}$ can be computed by integrating the probability density function over the s-dimensional volume $(-\infty, \mathbf{x}]$. Radon-Nikodým Derivative (176)

Based on the definiton of the distribution function $F_{\mathbf{X}}$ we can now define the probability that the random vector $\mathbf{X}$ lies between $\mathbf{a}$ and $\mathbf{b}$ with $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ by:

$$\begin{aligned} \text{prob}(\mathbf{a} < \mathbf{X} \leq \mathbf{b}) &\stackrel{\text{def}}{=} \mathbb{P}_{\mathbf{X}}((\mathbf{a}, \mathbf{b}]) && (2.686)\\ &= \int_{(\mathbf{a},\mathbf{b}]} d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) && (2.687)\\ &= \int_{(-\infty,\mathbf{b}]} d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) - \int_{(-\infty,\mathbf{a}]} d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) && (2.688)\\ &= F_{\mathbf{X}}(\mathbf{b}) - F_{\mathbf{X}}(\mathbf{a}). && (2.689) \end{aligned}$$

Now, we will demonstrate these newly defined concepts with the help of a few simple examples: the uniform distribution of a random vector on an s-dimensional interval $[\mathbf{a}, \mathbf{b}]$ and the uniform distribution on the unit circle as well as on the hemisphere.

EXAMPLE 2.74 (CDF of a Random Vector, Uniformly Distributed on $([\mathbf{a}, \mathbf{b}])$ *Let $\mathbf{X}$ be a random vector defined on the probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$, whose values are uniformly distributed within the interval $[\mathbf{a}, \mathbf{b}] \in \mathfrak{B}(\mathbb{R})$. The probability density function*

for such a uniformly distributed random vector $\mathbf{X}$ *on the interval* $[\mathbf{a},\mathbf{b}]$ *with* $\mathbf{a},\mathbf{b}\in\mathbb{R}^s$ *is given by:*

$$p_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \frac{1}{\prod_{i=1}^{s}(b_i-a_i)} & : \quad \text{for } \mathbf{x}\in[\mathbf{a},\mathbf{b}] \\ 0 & : \quad \text{otherwise.} \end{cases} \tag{2.690}$$

In terms of the probability density function $p_{\mathbf{X}}$ *the distribution function* $F_{\mathbf{X}}$ *can then be written as:*

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= \int_{[\mathbf{a},\mathbf{x}]} p_{\mathbf{X}}(\xi)\,d\mu^s(\xi) && (2.691)\\ &= \frac{1}{\prod_{i=1}^{s}(b_i-a_i)}\int_{[a_s,x_s]}\dots\int_{[a_1,x_1]} d\mu(\xi_1)\dots d\mu(\xi_s) && (2.692)\\ &= \frac{\prod_{i=1}^{s}(x_i-a_i)}{\prod_{i=1}^{s}(b_i-a_i)}. && (2.693) \end{aligned}$$

In the case (s=2), i.e. where there is a uniform distribution on a pixel area or a rectangular surface patch, $[\alpha_1,\beta_1]\times[\alpha_2,\beta_2]$*, as required in Monte Carlo rendering algorithms, the associated probability density function is given by:*

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(b_1-a_1)(b_2-a_2)} \tag{2.694}$$

and the cumulative distribution function can be written as:

$$F_{\mathbf{X}}(\mathbf{x}) = \frac{(x_1-a_1)(x_2-a_2)}{(b_1-a_1)(b_2-a_2)}. \tag{2.695}$$

Then, the probability that a point is drawn in an interval $[\alpha,\beta]\subset[\mathbf{a},\mathbf{b}]$ *can be computed via:*

$$\begin{aligned} \text{prob}\,(\alpha<\mathbf{X}\le\beta) &= \frac{1}{(b_1-a_1)(b_2-a_2)}\int_{[\alpha,\beta]} d\mu^2(\mathbf{x}) && (2.696)\\ &= \frac{1}{(b_1-a_1)(b_2-a_2)}\int_{[\alpha_2,\beta_2]}\int_{[\alpha_1,\beta_1]} d\mu(x_1)d\mu(x_2) && (2.697)\\ &= \frac{(\beta_1-\alpha_1)(\beta_2-\alpha_2)}{(b_1-a_1)(b_2-a_2)} && (2.698)\\ &= \frac{\mu^2([\alpha,\beta])}{\mu^2([\mathbf{a},\mathbf{b}])}. && (2.699) \end{aligned}$$

EXAMPLE 2.75 (A Random Vector, Uniformly Distributed on the Unit Circle) *A PDF for uniformly sampling a random vector* R,Θ *within the unit circle is given by:*

Section 6.5

$$p_{R,\Theta}(r,\theta) = \begin{cases} \frac{1}{\pi}r & : \quad \text{for } r\in[0,1] \text{ and } \theta\in[0,2\pi) \\ 0 & : \quad \text{otherwise,} \end{cases} \tag{2.700}$$

where we assume that the unit circle is represented in polar coordinates. Then, the cumulative distribution function $F_{\Theta,\Phi}$ *can be written as:*

$$\begin{aligned} F_{R,\Theta}(r,\theta) &= \frac{1}{\pi}\int_{[0,\theta]}\int_{[0,r]} \xi_1 \, d\mu(\xi_1) d\mu(\xi_2) && (2.701)\\ &= \frac{\theta}{\pi}\int_{[0,r]} \xi_1 \, d\mu(\xi_1) && (2.702)\\ &= \frac{\theta\, r^2}{2\pi}. && (2.703) \end{aligned}$$

Now, let us derive another important probability distribution function, which recourse will be taken repeatedly further below in sampling strategies for finding approximate solutions to the global illumination equation by means of probabilistic methods: the uniform distribution on the hemisphere.

Chapter (6)

EXAMPLE 2.76 (Random Vectors, Uniformly Distributed on the Upper Hemisphere or the Unit Sphere) *Let us consider the probability space* $(\mathcal{H}^2_+, \mathfrak{B}(\mathcal{H}^2_+), \sigma)$*, where our base set is given by the set of points on the upper hemisphere,* $\mathfrak{B}(\mathcal{H}^2_+)$ *is as usual the Borel* σ*-algebra of subsets of* $\mathcal{H}^2_+$*, and* σ *is the solid angle measure defined on* $\mathfrak{B}(\mathcal{H}^2_+)$*. Let* Θ, Φ *be a random vector on* $(\mathcal{H}^2_+, \mathfrak{B}(\mathcal{H}^2_+), \sigma)$*, due to Relation (2.700), the uniform density* $p_{\Theta,\Phi}$ *is given by:*

Borel-σ Algebra (865)

Solid Angle Measure (87)

$$p_{\Theta,\Phi}(\theta,\phi) = \frac{1}{2\pi}\sin\theta, \tag{2.704}$$

with $0 \le \theta \le \frac{\pi}{2}, 0 \le \phi < 2\pi$*, since it holds:*

$$\begin{aligned} \int_{\mathcal{H}^2_+} p_\omega(\omega)\, d\sigma(\omega) &\overset{(2.186)}{=} \frac{1}{2\pi}\int_{[0,\frac{\pi}{2}]\times[0,2\pi)} \sin(\theta)\, d\mu^2(\theta,\phi) && (2.705)\\ &= 1. && (2.706) \end{aligned}$$

Then, the corresponding cumulative distribution function $F_{\Theta,\Phi}$ *can be written as:*

$$\begin{aligned} F_{\Theta,\Phi}(\theta,\phi) &\overset{(2.685)}{=} \frac{1}{2\pi}\int_{[0,\phi]}\left(\int_{[0,\theta]} \sin(\xi_1) d\mu(\xi_1)\right) d\mu(\xi_2) && (2.707)\\ &= \frac{1}{2\pi}\phi(1-\cos\theta). && (2.708) \end{aligned}$$

As it is easily seen, the cumulative distribution function of the random vector (Θ,Φ) *uniformly distributed on the unit sphere, with probability density function* $p_{\Theta,\Phi}(\omega) = \frac{1}{4\pi}$*, is then given by:*

$$F_{\Theta,\Phi}(\theta,\phi) = \frac{1}{4\pi}\phi(1-\cos(\theta)). \tag{2.709}$$

Note the difference of the PDFs and CDFs—expressed in terms of directions or angles—for sampling of directions over one of the hemispheres and the unit sphere:

$$p_\omega(\omega) = \frac{1}{2\pi} \quad or \quad p_{\Theta,\Phi}(\theta,\phi) = \frac{1}{2\pi}\sin\theta \tag{2.710}$$

$$F_\omega(\omega) = \frac{\omega}{2\pi} \quad or \quad F_{\Theta,\Phi}(\theta,\phi) = \frac{1}{2\pi}\phi(1-\cos\theta) \tag{2.711}$$

and

$$p_\omega(\omega) = \frac{1}{4\pi} \quad or \quad p_{\Theta,\Phi}(\theta,\phi) = \frac{1}{4\pi}\sin\theta \tag{2.712}$$

$$F_\omega(\omega) = \frac{\omega}{4\pi} \quad or \quad F_{\Theta,\Phi}(\theta,\phi) = \frac{1}{4\pi}\phi(1-\cos\theta). \tag{2.713}$$

EXAMPLE 2.77 *It is known from the previous example, that the cumulative distribution function of a uniformly distributed random vector defined on the upper hemisphere is* $F_{\Theta,\Phi}(\theta,\phi) = \frac{1}{2\pi}\phi(1-\cos(\theta))$. *The partial derivatives of this function are then given by:*

$$\frac{\partial}{\partial\theta}F_{\Theta,\Phi}(\theta,\phi) = \frac{1}{2\pi}\phi\sin(\theta) \quad and \quad \frac{\partial^2}{\partial\theta\,\partial\phi}F_{\Theta,\Phi}(\theta,\phi) = \frac{1}{2\pi}\sin(\theta), \tag{2.714}$$

thus

$$p_{\Theta,\Phi}(\theta,\phi) = \frac{\partial^2}{\partial\theta\,\partial\phi}F_{\Theta,\Phi}(\theta,\phi). \tag{2.715}$$

REMARK 2.82 (The PDF as the Derivative of the CDF, Multivariate Case) *Setting our focus once more on the Relations (2.682) - (2.685), then, due to the Fundamental Theorem of Calculus, we get for the representation of the density* $p_{\mathbf{X}}$*:*

[174, Rudin 1998]

$$p_{\mathbf{X}}(x_1,\ldots,x_s) = \frac{\partial^s F_{\mathbf{X}}(x_1,\ldots,x_s)}{\partial\mu(x_1)\ldots\partial\mu(x_s)}, \tag{2.716}$$

i.e. the probability density function $p_{\mathbf{X}}$ *is the* s*-dimensional derivative of the cumulative distribution function of an* s*-dimensional random vector.*

For solving multidimensional integrals, in Monte Carlo integration we often have to sample from complex joint probability densities. If it is possible to isolate a variable in such a complex joint probability density, then we get often samples distributed according to such a density in a simple way. Linked with the concept of conditional probability, this technique plays an important role in many sampling procedures in Monte Carlo integration. The concept behind this idea is the construct of the *marginal density*.

Conditional Probability (205)

DEFINITION 2.56 (Marginal Density) *Let* $(\Omega,\mathfrak{F}(\Omega),\mathbb{P})$ *be a probability space,* $\mathbf{X}$ *and* $\mathbf{Y}$ *random variables defined on* $(\Omega,\mathfrak{F}(\Omega))$ *with values in* $(\mathbb{R},\mathfrak{B}(\mathbb{R}))$ *or* $(\mathbb{R}^s,\mathfrak{B}(\mathbb{R}^s))$*, and let*

Probability Space (163)

$p_{\mathbf{X},\mathbf{Y}}$ be the joint probability density function of $\mathbf{X}, \mathbf{Y}$. Then, the marginal densities $p_{\mathbf{X}}$ and $p_{\mathbf{Y}}$ of $\mathbf{X}, \mathbf{Y}$ are defined as:

$$p_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbb{R}^s} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\, d\mu^s(\mathbf{y}) \tag{2.717}$$

$$p_{\mathbf{Y}}(\mathbf{y}) = \int_{\mathbb{R}^s} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\, d\mu^s(\mathbf{x}), \tag{2.718}$$

that is, the marginal distribution of a random variable $\mathbf{X}$ is simply the probability distribution of $\mathbf{X}$ averaged over information about $\mathbf{Y}$ and vice versa.

Obviously, the marginal densities determine the distributions of the random variables $\mathbf{X}$ and $\mathbf{Y}$ since it holds:

$$\begin{aligned}
\mathbb{P}_{\mathbf{X}}(\mathbf{B}) &= \int_{\mathbf{B}\times(-\infty,\infty)^s} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\, d\mu^s(\mathbf{y}) d\mu^s(\mathbf{x}) && (2.719)\\
&= \int_{\mathbf{B}} \left(\int_{(-\infty,\infty)^s} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\, d\mu^s(\mathbf{y}) \right) d\mu^s(\mathbf{x}) && (2.720)\\
&\overset{(2.717)}{=} \int_{\mathbf{B}} p_{\mathbf{X}}(\mathbf{x})\, d\mu^s(\mathbf{x}) && (2.721)
\end{aligned}$$

and due to the Theorem of Fubini-Tonelli we obtain: Theorem of Fubini-Tonelli (115)

$$\begin{aligned}
\mathbb{P}_{\mathbf{Y}}(\mathbf{B}) &= \int_{(-\infty,\infty)^s\times\mathbf{B}} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\, d\mu_1^s(\mathbf{y}) d\mu_2^s(\mathbf{x}) && (2.722)\\
&= \int_{(-\infty,\infty)^s} \left(\int_{\mathbf{B}} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\, d\mu_1^s(\mathbf{y}) \right) d\mu_2^s(\mathbf{x}) && (2.723)\\
&= \int_{\mathbf{B}} \left(\int_{(-\infty,\infty)^s} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\, d\mu_1^s(\mathbf{x}) \right) d\mu_2^s(\mathbf{y}) && (2.724)\\
&\overset{(2.717)}{=} \int_{\mathbf{B}} p_{\mathbf{Y}}(\mathbf{y})\, d\mu^s(\mathbf{y}). && (2.725)
\end{aligned}$$

EXAMPLE 2.78 *In one of the following Chapters, we will see that a pinhole camera creates images where everything is in perfect focus, while a thin lens camera model makes images with depth-of-field effects. In such a camera the pinhole is replaced with a disk-shaped thin lens, which has certain idealized behavior. To simulate depth-of-field effects a rendering algorithm has to generate rays passing through the area of the lens. For that purpose, we need a method for generating uniformly distributed samples within the unit circle, which then can be transformed on the camera lens.* Depth of Field (686)

Representing the unit circle in porlar coordinates, according to Relation (2.700) a PDF for sampling uniformly on the unit disk is given by: Polar Coordinates (832)

$$p_{R,\Theta}(r,\theta) = \frac{1}{\pi} r, \tag{2.726}$$

where $r \in [0,1]$ *and* $\theta \in [0, 2\pi)$. *In order to sample* r *and* θ, *we compute first the marginal density*

Marginal Density Function (

$$p_R(r) = \frac{1}{\pi} \int_{[0,2\pi)} r \, d\mu(\theta) = \frac{r}{\pi}\theta \Big|_0^{2\pi} = 2r. \tag{2.727}$$

Integrating $p_R(r)$ *over* $[0, R]$, *then leads to the CDF*

$$F_R(R) = 2 \int_{[0,R]} r \, d\mu(r) = 2 \frac{1}{2} r^2 \Big|_0^R = R^2. \tag{2.728}$$

REMARK 2.83 *It should be clear that the concept of the marginal density is not restricted to the case of two random variables* $\mathbf{X}$ *and* $\mathbf{Y}$. *By applying the Theorem of Fubini-Tonelli we get corresponding formulas easily.*

Theorem of Fubini-Tonelli (115)

2.4.4 EXPECTED VALUE AND VARIANCE OF A RANDOM VARIABLE

Section 2.4.2

As has been shown in a few of our examples, it is not the probability measure itself that is of interest, but rather a number that may be used to describe particular aspects of a probability measure. Thus, an important aspect of a probability measure is its *location*, that is, the location of the associated probability distribution. It can be described by the *expected value* of a random variable associated with a corresponding probability measure.

Probability Measure (80)

Expected Value (196)

Another interesting concept in connection with a random variable is the *variance* of a random variable. While the expected value may be defined using a kind of *mean value* and makes a connection between the result of the random experiment and its mathematical realization, the variance describes how far a set of values of a random variable are spread out from each other. It describes the extent of the deviation of individual values of the random variable from its appropriate mean value.

Variance (201)

Lebesgue Integral (105)

By means of the definition of the Lebesgue-integral we will now construct these two very important concepts of probability theory. Since they permit statements on the quality of certain probabilistic models, they play a central role in our derivation of stochastic methods for solving linear integral equations.

Chapter 6

THE EXPECTED VALUE OF A RANDOM VARIABLE. Let us start with the definition of the *expected value* of a random variable defined on a given probability space. It is the key idea behind any stochastic algorithm for solving integrals and integral equations.

DEFINITION 2.57 (The Expected Value of a Random Variable) *Let* X *be a discrete random variable defined on the discrete probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *with finite or countably*

Discrete RV (168)

Discrete Probability Space (163)

infinite image range $\mathrm{Im}(X)$. *Then, the* expected value *of* X *is defined as*

$$\mathrm{E}(X) \overset{def}{=} \sum_{x \in \mathrm{Im}(X)} x\, \mathbb{P}(\{\omega \in \Omega \,|\, X(\omega) = x\}) \tag{2.729}$$

$$\overset{\mathbb{P}_X = \mathbb{P} \circ X^{-1}}{=} \sum_{x \in \mathrm{Im}(X)} x\, \mathbb{P}_X(X = x) \tag{2.730}$$

$$\overset{(2.559)}{=} \sum_{x \in \mathrm{Im}(X)} x\, p_X(x). \tag{2.731}$$

In the case where we consider a continuous random variable, defined on the continuous probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *with associated probability density function* p_X, *the expected value of* X *is defined as*

Continuous RV (168)

Probability Density Function (176)

$$\mathrm{E}(X) \overset{def}{=} \int_\Omega X(\omega)\, d\mathbb{P}(\omega) \tag{2.732}$$

$$\overset{\mathbb{P} = \mathbb{P}_X \circ X}{=} \int_\Omega X(\omega)\, d(\mathbb{P}_X \circ X)(\omega) \tag{2.733}$$

$$\overset{X(\omega) = x}{=} \int_{\mathbb{R}} x\, d\mathbb{P}_X(x) \tag{2.734}$$

$$\overset{d\mathbb{P}_X = p_X(x)\, d\mu(x)}{=} \int_{\mathbb{R}} x\, p_X(x)\, d\mu(x), \tag{2.735}$$

where $\mathbb{P}_X$ *is the image measure on the measurable space* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ *induced by* X *and* μ *corresponds to the Lebesgue measure defined on the Borel set* $\mathfrak{B}(\mathbb{R})$.

Measurable Space (80)

Let us clarify the concept of the expected value of a discrete as well as a continuous random variable by means of two simple examples.

EXAMPLE 2.79 (The Expected Value of a Discrete Random Variable) *Let us consider once more our random experiment of flipping a coin three times, where we are interested in the number of one's that can occur. The sample space of this random experiment is the set* $\{0, 1, 2, 3\}$. *Now, a random variable* X *that models this experiment has to map the base set* $\{0, 1\}^3$ *onto the set of outcomes* $\{0, 1, 2, 3\}$. *Based on the definition of a discrete random variable from Equation (2.730), the expected value of* X *can be calculated via the outcomes* x_i *of the random variable and the probabilities of the outcomes* $\mathbb{P}_X(X = x_i)$. *That is, the expected value of the random variable* X *is given by*

$$\begin{aligned} \mathrm{E}(X) &= \sum_{i=1}^{4} x_i\, \mathbb{P}_X(X = x_i) && (2.736)\\ &= \sum_{i=1}^{4} (i-1)\, \mathbb{P}_X(X = i-1) && (2.737)\\ &= 0\, \mathbb{P}_X(X = 0) + 1\, \mathbb{P}_X(X = 1) + 2\, \mathbb{P}_X(X = 2) + 3\, \mathbb{P}_X(X = 3). && (2.738) \end{aligned}$$

With the probability distribution of the random variable X from Example 2.66, thus

$$\mathbb{P}_X(X=0) = \mathbb{P}_X(X=3) = \tfrac{1}{8} \tag{2.739}$$

$$\mathbb{P}_X(X=1) = \mathbb{P}_X(X=2) = \tfrac{3}{8} \tag{2.740}$$

the expected value of the random variable X is

$$E(X) = 1\,\frac{3}{8} + 2\,\frac{3}{8} + 3\,\frac{1}{8} = \frac{3}{2}. \tag{2.741}$$

EXAMPLE 2.80 (The Expected Value of a Continuous Random Variable) *Let us now derive the expected value of a continuous random variable, where we consider the random experiment of drawing a random number from $[0,1]$, as we did it in Example 2.69. There, we defined a random variable X on the canonical probability space $([0,1], \mathfrak{B}([0,1]), \mu)$ where our probability measure corresponds to the Lebesgue measure. The random variable X then maps $\omega \in [0,1]$ onto $X(\omega) = \omega \in [0,1]$. The probability density function associated with X is,*

$$p_X(x) = \begin{cases} 1 & : \quad \text{for } x \in [0,1] \\ 0 & : \quad \text{otherwise,} \end{cases} \tag{2.742}$$

that is, due to the above definition the expected value of X is given by

$$E(X) \overset{(2.735)}{=} \int_{\mathbb{R}} x\, p_X(x)\, d\mu(x) \tag{2.743}$$

$$= \int_{[0,1]} x\, d\mu(x) \tag{2.744}$$

$$= x^2\Big|_0^1 = \frac{1}{2}. \tag{2.745}$$

Based on the definition of the expected value of a random variable, we are now able to define the expectation of a random vector.

Random Vector (183) Probability Space (163)

DEFINITION 2.58 (The Expected Value of a Random Vector) *Let $\mathbf{X}$ be a discrete or a continuous random vector defined on a corresponding probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$. Then, the expected value of the s-dimensional random vector $\mathbf{X}$ is simply given by the vector of the expected values of the component random variables X_i. Mathematically, this can be expressed as*

$$E(\mathbf{X}) = E(X_1, \ldots, X_s) = (E(X_1), \ldots, E(X_s)). \tag{2.746}$$

THE EXPECTED VALUE OF A FUNCTION OF RANDOM VARIABLES. From our discussion about random variables and random vectors in the previous sections, it is known, that measurable functions of random variables or random vectors are also random variables. This fact, now implies the definition of the expected value of a function of random variables or random vectors.

REMARK 2.84 *Since a random variable can be considered as a random vector containing only a single component, we introduce the concept of the expected value of a function of random variables via the concept of the random vector.*

DEFINITION 2.59 (The Expected Value of a Function of Random Vectors) *Let* $\mathbf{X}$ *be a discrete or a continuous random vector defined on a corresponding probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$*, furthermore, let* g *be a measurable function defined on the measurable space* $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$ *with values in* $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$*. In the case that* $\mathbf{X}$ *is a discrete random vector, the expected value of* $g(\mathbf{X})$ *is defined as*

Random Vector (183) Probability Space (163) Measurable Function (98)

$$\mathrm{E}(g(\mathbf{X})) \overset{\mathrm{def}}{=} \sum_{i \geq 1} g(\mathbf{x}_i)\, \mathbb{P}_{\mathbf{X}}(\mathbf{X} = \mathbf{x}_i) \tag{2.747}$$

$$\equiv \sum_{i_s \geq 1} \cdots \sum_{i_1 \geq 1} g(x_{i_1}, \ldots, x_{i_s})\, \mathbb{P}_{X_1,\ldots,X_s}(X_1 = x_{i_1}, \ldots, X_s = x_{i_s}), \tag{2.748}$$

where we also often express the expected value in terms of the joint probability mass function $p_{\mathbf{X}}$*, that is,*

Probability Mass Function (185)

$$\mathrm{E}(g(\mathbf{X})) \overset{\mathrm{def}}{=} \sum_{i_s \geq 1} \cdots \sum_{i_1 \geq 1} g(x_{i_1}, \ldots, x_{i_s})\, p_{X_1,\ldots,X_s}(x_{i_1}, \ldots, x_{i_s}). \tag{2.749}$$

In the case that we consider a continuous random vector with associated joint probability density function $p_{\mathbf{X}}$*, the expected value of* $\mathbf{X}$ *is defined as*

Continuous RV (185) Probability Density Function (189)

$$\mathrm{E}(g(\mathbf{X})) \overset{\mathrm{def}}{=} \int_{\Omega} g(\mathbf{X}(\omega))\, \mathrm{d}\mathbb{P}(\omega) \tag{2.750}$$

$$\overset{\mathbb{P}=\mathbb{P}_{\mathbf{X}} \circ \mathbf{X}}{=} \int_{\mathbb{R}^s} g(\mathbf{x})\, \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}) \tag{2.751}$$

$$\overset{\mathrm{d}\mathbb{P}_{\mathbf{X}} = p_{\mathbf{X}}(\mathbf{X})\, \mathrm{d}\mu(\mathbf{x})}{=} \int_{\mathbb{R}^s} g(\mathbf{x})\, p_{\mathbf{X}}(\mathbf{x})\, \mathrm{d}\mu(\mathbf{x}), \tag{2.752}$$

where $\mathbb{P}_{\mathbf{X}}$ *is the image measure on* $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$ *induced by* $\mathbf{X}$*.*

Image Measure (184)

EXAMPLE 2.81 (The Expected Value of a Function of a Continuous Random Variable) *If we are interested in drawing a random variable from the interval* $[a, b]$*, where* $X(\omega) = \omega(b - a) + a$*, thus,* $g(X) = (b - a)X + a$ *, then the associated expected value is given by*

$$\mathrm{E}(X) \overset{(2.735)}{=} \int_{\mathbb{R}} ((b - a)x + a) \frac{1}{b - a}\, \mathrm{d}\mu(x) \tag{2.753}$$

$$= \int_{[a,b]} x + \frac{a}{b - a}\, \mathrm{d}\mu(x) \tag{2.754}$$

$$= \left(\frac{1}{2} x^2 + \frac{a}{b - a} x \right) \Big|_a^b \tag{2.755}$$

$$= \frac{1}{2} \left(b^2 - a^2 \right) + a. \tag{2.756}$$

Now, let us take a look at *Russian Roulette*, a technique based on the principle of the expected value of a continuous random variable. Used as a variance reduction technique, Russian roulette plays an important role in many Monte Carlo rendering algorithms.

Section 6.6
Section 9.1

EXAMPLE 2.82 (Russian Roulette) *Since light can be reflected infinitely often within a scene, we have no chance of simulating this behavior in a rendering algorithm exactly. To solve this problem we need a technique that controls the generation of paths with respect to its length.*

Section 9.1

Now, simply cutting off a path introduces a bias in our images since this light path might be potentially very important for the final color of a pixel. Otherwise most of the contributions of a path to the final image are very small but equally expensive to evaluate. A technique that can help us to get this problem under control is Russian roulette*, introduced in [11, Arvo & Kirk 1990].*

Bias (507)

Let us assume, that we have to compute a quantity F *that is a sum of infinitely many terms* F_i *thus,*

$$F = \sum_{i=1}^{\infty} F_i, \tag{2.757}$$

such as described above in the case of an infinitely long light path. Let us furthermore assume that U *is a continuous random variable, uniformly distributed on the interval* $[0, 1]$.

Continuous RV (168)
Uniform Distribution (180)

The idea behind Russian roulette is to skip most of the terms F_i *with small contributions to the final value of* F*, and to compensate this by an appropriate weighting of the remaining terms. For that purpose, we define additional random variables* $f_{\alpha_i}(U)$ *by*

$$f_{\alpha_i}(U) = \begin{cases} \frac{1}{\alpha_i} F_i & : \text{ if } U \leq \alpha_i, \\ 0 & : \text{ else}, \end{cases} \tag{2.758}$$

where $\alpha_i \in [0, 1]$.

Then, the random variable U *evaluates* F_i *with the probability* α_i *and weighting* $\frac{1}{\alpha_i}$ *and it discards the evaluation of* F_i *with probability* $1 - \alpha_i$*. With respect to the expected value of* f_{α_i} *the following clearly applies*

$$E(f_{\alpha_i}(U)) \overset{(2.730)}{=} \alpha_i \cdot \frac{1}{\alpha_i} F_i + (1 - \alpha_i) \cdot 0 \tag{2.759}$$

$$= F_i. \tag{2.760}$$

Obviously, using Russian roulette allows to skip the computation of terms whose value is very low but not necessarily zero, while it guarantees the computation of the correct value on the average.

As will be seen further below, Russian roulette is a highly useful technique when it is applied in procedures for simulating global illumination effects in a scene to be rendered.

Section 9.1

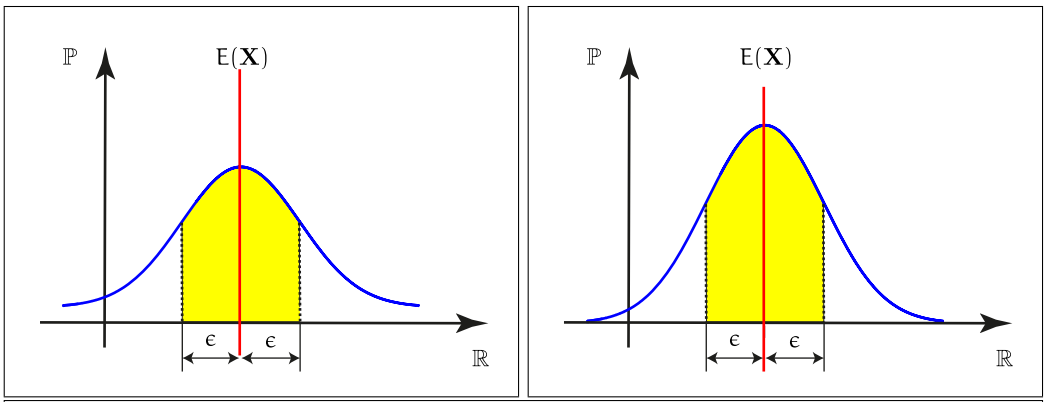


FIGURE 2.53: THE VARIANCE OF RANDOM VARIABLES. Shown are the densities of two normally distributed random variables with same expected value. On the left, we have higher variance than in the right image, since the probability that values of the corresponding random variable are close to its expected value, is—due to the pdf illustrated in the left image—lower than in the right image.

THE VARIANCE OF A RANDOM VARIABLE. Based on the probabilistic concept of the expected value of a random variable, we will now introduce the stochastic construct of the *variance* of a random variable. It plays a central role for the development of efficient probabilistic algorithms for solving the light transport equation.

DEFINITION 2.60 (The Variance and the Central Moments of a Random Variable or a Random Vector) *Let* $\mathbf{X}$ *be a random variable or a random vector defined on a probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$. *The* variance *or, alternatively, the* 2^{nd} moment *of* $\mathbf{X}$ *is defined as* Probability Space (163)

$$\mathrm{Var}(\mathbf{X}) \stackrel{def}{=} \mathrm{E}((\mathbf{X} - \mathrm{E}(\mathbf{X}))^2). \tag{2.761}$$

Generally, we define the central moment of order n *of a random variable or a random vector by*

$$m^n(\mathbf{X}) \stackrel{def}{=} \mathrm{E}((\mathbf{X} - \mathrm{E}(\mathbf{X}))^n), \tag{2.762}$$

i.e the variance of a random variable or random vector corresponds to the 2^{nd} *central moment of* $\mathbf{X}$, *see Figure 2.53.*

EXAMPLE 2.83 (Russian Roulette, Continued) *Let us now make a statement about the variance of the random variable* $f_{\alpha_i}(u)$ *from the foregoing example. Due to Example*

2.82 it holds with $E(f_{\alpha_i}(U)) = F_i$

$$\begin{aligned} \mathrm{Var}\left(f_{\alpha_i}(U)\right) &= E\left(\left(f_{\alpha_i}(U) - E\left(f_{\alpha_i}(U)\right)\right)^2\right) && (2.763)\\ &\overset{E(f_{\alpha_i}(U))=F_i}{=} E\left(\left(f_{\alpha_i}(U) - F_i\right)^2\right) && (2.764)\\ &= E\left(f^2_{\alpha_i}(U) - 2f_{\alpha_i}(U)F_i + F_i^2\right). && (2.765) \end{aligned}$$

Measurable Function (98)
Linear Space (100)

Now, from our discussion about measurable functions is known that the class of measurable functions defines a linear space, that is, the sum of two measurable functions is measurable again. Based on this fact, the term $f^2_{\alpha_i}(U) - 2f_{\alpha_i}(U)F_i + F_i^2$ *also corresponds to a random variable, which is defined by*

$$f^2_{\alpha_i}(U) - 2f_{\alpha_i}(U)F_i + F_i^2 = \begin{cases} \left(\frac{1}{\alpha_i}F_i\right)^2 - 2\frac{1}{\alpha_i}F_i^2 + F_i^2 & : \quad \textit{if } U \leq \alpha_i, \\ F_i^2 & : \quad \textit{else.} \end{cases} \tag{2.766}$$

For the expected value of this random variable then it holds

$$\begin{aligned} & E\left(f^2_{\alpha_i}(U) - 2f_{\alpha_i}(U)F_i + F_i^2\right) && (2.767)\\ &= \alpha_i \cdot \left(\left(\frac{1}{\alpha_i}F_i\right)^2 - 2\frac{1}{\alpha_i}F_i^2 + F_i^2\right) + (1-\alpha_i)\cdot F_i^2 && (2.768)\\ &= \frac{1}{\alpha_i}F_i^2 - F_i^2. && (2.769) \end{aligned}$$

Combining the Formula (2.765) with Relation (2.769) leads to

$$\mathrm{Var}(f_{\alpha_i}(U)) = F_i^2\left(\frac{1}{\alpha_i} - 1\right), \tag{2.770}$$

which means: If the parameter α_i *is very large, the evaluation of* F_i *will continue many times and the approximation will be more accurate. If* α_i *is small, the evaluation of* F_i *will stop soon, but the result will have a higher variance, see Figure 2.54*

PROPERTIES OF RANDOM VARIABLES AND RANDOM VECTORS. We finish this section with some remarks on properties of random variables and random vectors.

Measurable Function (98)
Lebesgue Integral (105)

From our discussion about measurable functions as well as the linearity of the Lebesgue integral, we get obviously in addition to

$$\begin{aligned} E(a\mathbf{X}) &= a\,E(\mathbf{X}) && (2.771)\\ \mathrm{Var}(a\mathbf{X}) &= a^2\,\mathrm{Var}(\mathbf{X}) && (2.772) \end{aligned}$$

also the following useful identity

$$E\left(\sum_{i=1}^{N}\mathbf{X}_i\right) = \sum_{i=1}^{N}E(\mathbf{X}_i). \tag{2.773}$$

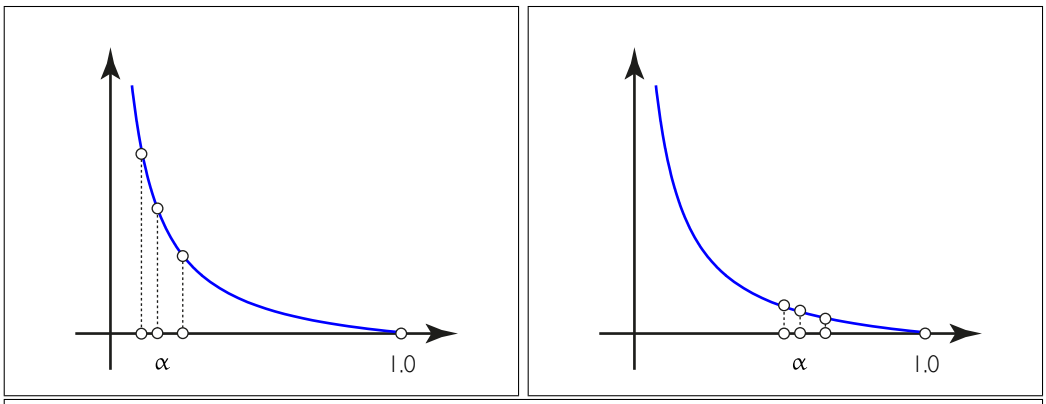


FIGURE 2.54: THE CHOICE OF THE PARAMETER α IN RUSSIAN ROULETTE. The plots of the function $f(\alpha) = \frac{1}{\alpha} - 1$ shows that values of α close to 0 lead to large function values, see the left image, and values close to 1 lead to high function values of $f(\alpha) = \frac{1}{\alpha} - 1$, see the right image.

Furthermore, the variance of a random variable or a random vector $\mathbf{X}$ may also be expressed in the following way, highly useful for many calculations:

$$\begin{aligned} \mathrm{Var}(\mathbf{X}) &= \mathrm{E}\left((\mathbf{X} - \mathrm{E}(\mathbf{X}))^2\right) && (2.774)\\ &= \mathrm{E}\left(\mathbf{X}^2 - 2\mathrm{E}(\mathbf{X}) + \mathrm{E}^2(\mathbf{X})\right) && (2.775)\\ &\overset{(2.771)}{=} \mathrm{E}(\mathbf{X}^2) - \mathrm{E}^2(\mathbf{X}). && (2.776) \end{aligned}$$

EXAMPLE 2.84 (Russian Roulette once More) *The variance of the random variable* $f_{\alpha_i}(U)$ *from Example 2.82 can be computed more easily—as we did it in the preceding example—by using Relation (2.776)*

$$\begin{aligned} \mathrm{Var}(f_{\alpha_i}(U)) &= \mathrm{E}\left(f_{\alpha_i}^2(U)\right) - \mathrm{E}^2\left(f_{\alpha_i}(U)\right) && (2.777)\\ &= \alpha_i \left(\frac{1}{\alpha_i} F_i\right)^2 + (1 - \alpha_i) \cdot 0 - F_i^2 && (2.778)\\ &= F_i^2 \left(\frac{1}{\alpha_i} - 1\right). && (2.779) \end{aligned}$$

Another interesting quantity which arises in the connection of two random variables or random vectors is the *covariance*. As we will see, the covariance is a quantity that makes a statement about the linear connection of two random variables which have the same distribution. It can be calculated via the variance of random variables.

DEFINITION 2.61 (Covariance of Random Variables or Random Vectors) *Let* $\mathbf{X}, \mathbf{Y}$ *be two random variables or random vectors defined on a probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$. Probability Space (163)

Then, the covariance *of* $\mathbf{X}$ *and* $\mathbf{Y}$ *is defined as*

$$\mathrm{Cov}(\mathbf{X},\mathbf{Y}) \stackrel{\mathrm{def}}{=} \mathrm{E}\left((\mathbf{X}-\mathrm{E}(\mathbf{X}))(\mathbf{Y}-\mathrm{E}(\mathbf{Y}))\right). \tag{2.780}$$

An alternative formula for computing the covariance of $\mathbf{X}$ *and* $\mathbf{Y}$ *is given by*

$$\mathrm{Cov}(\mathbf{X},\mathbf{Y}) \stackrel{\mathrm{def}}{=} \mathrm{E}(\mathbf{XY}) - \mathrm{E}(\mathbf{X})\mathrm{E}(\mathbf{Y}). \tag{2.781}$$

Based on Definition 2.61, the variance of the sum of two random variables or random vectors $\mathbf{X}$ and $\mathbf{Y}$ can now also be written as the sum of the single variances and the covariance

$$\begin{aligned}
\mathrm{Var}(\mathbf{X}+\mathbf{Y}) &\stackrel{(2.776)}{=} \mathrm{E}((\mathbf{X}+\mathbf{Y})^2) - \mathrm{E}^2(\mathbf{X}+\mathbf{Y}) && (2.782)\\
&= \mathrm{E}(\mathbf{X}^2 + 2\mathbf{XY} + \mathbf{Y}^2) - && (2.783)\\
&\quad (\mathrm{E}^2(\mathbf{X}) + 2\mathrm{E}(\mathbf{X})\mathrm{E}(\mathbf{Y}) + \mathrm{E}^2(\mathbf{Y}))\\
&= \mathrm{E}(\mathbf{X}^2) - \mathrm{E}^2(\mathbf{X}) + \mathrm{E}(\mathbf{Y}^2) - \mathrm{E}^2(\mathbf{Y}) + && (2.784)\\
&\quad 2(\mathrm{E}(\mathbf{XY}) - \mathrm{E}(\mathbf{X})\mathrm{E}(\mathbf{Y}))\\
&\stackrel{(2.781),(2.776)}{=} \mathrm{Var}(\mathbf{X}) + \mathrm{Var}(\mathbf{Y}) + 2\mathrm{Cov}(\mathbf{X},\mathbf{Y}). && (2.785)
\end{aligned}$$

A similar formula holds for the sum of n random variables or random vectors $\mathbf{X}_1,\ldots,\mathbf{X}_n$, it is given by

$$\mathrm{Var}\left(\sum_{i=1}^{n}\mathbf{X}_i\right) = \sum_{i=1}^{n}\mathrm{Var}(\mathbf{X}_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathrm{Cov}(\mathbf{X}_i,\mathbf{X}_j). \tag{2.786}$$

A furthermore important property of a random variable or a random vector is *independence*.

DEFINITION 2.62 (Independence of a Random Variable or a Random Vector) *Two random variables or random vectors* $\mathbf{X}$ *and* $\mathbf{Y}$ *are called* independent, *if the outcome of* $\mathbf{X}$ *does not influence the outcome of* $\mathbf{Y}$. *Mathematically, the independence of* $\mathbf{X}$ *and* $\mathbf{Y}$ *can be expressed as*

$$\mathbb{P}(\mathbf{XY}) = \mathbb{P}(\mathbf{X})\cdot\mathbb{P}(\mathbf{Y}). \tag{2.787}$$

Based on Definition 2.62 we can now easily derive the following formulae for independent random variables or random vectors $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_N$

$$\mathbb{P}\left(\prod_{i=1}^{N}\mathbf{X}_i\right) = \prod_{i=1}^{N}\mathbb{P}(\mathbf{X}_i), \tag{2.788}$$

$$\mathrm{E}\left(\prod_{i=1}^{N}\mathbf{X}_i\right) = \prod_{i=1}^{N}\mathrm{E}(\mathbf{X}_i), \tag{2.789}$$

and with

$$\mathrm{Cov}(\mathbf{X}_i, \mathbf{X}_j) = \mathrm{E}(\mathbf{X}_i)\mathrm{E}(\mathbf{X}_j) - \mathrm{E}(\mathbf{X}_i)\mathrm{E}(\mathbf{X}_j) = 0 \tag{2.790}$$

for the variance of n independent random variables or random vectors also

$$\mathrm{Var}\left(\sum_{i=1}^{N} \mathbf{X}_i\right) = \sum_{i=1}^{N} \mathrm{Var}(\mathbf{X}_i). \tag{2.791}$$

2.4.5 CONDITIONAL PROBABILITY

One important task of probability theory is to develop procedures with which probabilities of complex events may be derived from probabilities of simple events. When some partial information about a random experiment is available, via the concept of *conditional probability* and the knowledge of the occurrence of events, information on the probability of the occurrence of other events may be simply achieved.

DEFINITION 2.63 (Conditional Probability) *Let* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *be a discrete probability space and* A, B *be two events from the* $\mathfrak{F}(\Omega)$*, then* Probability Space (163)

$$\mathbb{P}(B|A) \equiv \mathbb{P}_{\cdot|A}(B) \quad \stackrel{\mathrm{def}}{=} \quad \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \tag{2.792}$$

is the probability of some event B*, given the occurrence of an other event* A*, also briefly denoted as the* conditional probability *of* B *given* A*. Since this definition provides us with a measure on the* σ*-algebra* $\mathfrak{F}(\Omega)$*, the so-called* conditional probability measure $\mathbb{P}_{\cdot|A}$*, the triple* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P}_{\cdot|A})$ *becomes a probability space, more precisely: a* conditional probability space. Measure (79) σ-algebra (828)

Now, the concept of the conditional probability can also be extended to continuous probability spaces. For that purpose, let us assume $(\Omega, \mathfrak{F}(\Omega), \mathbb{P}_{\cdot|A})$ *be a continuous probability space with conditional probability measure* $\mathbb{P}_{\cdot|B}$*, then the* conditional probability *of* B*, given* A *is defined as follows:*

$$\mathbb{P}(B|A) \quad \stackrel{\mathrm{def}}{=} \quad \int_B \mathrm{d}\mathbb{P}_{\cdot|A}. \tag{2.793}$$

In the following, we illustrate the concept of conditional probability with the help of an example which we use as a spring board for the introduction of two stochastic models of great interest to the analysis of Monte Carlo rendering algorithms: *discrete-time Markov chains* and *discrete-time Markov processes*. Section 2.4.7.1 Section 2.4.7.2

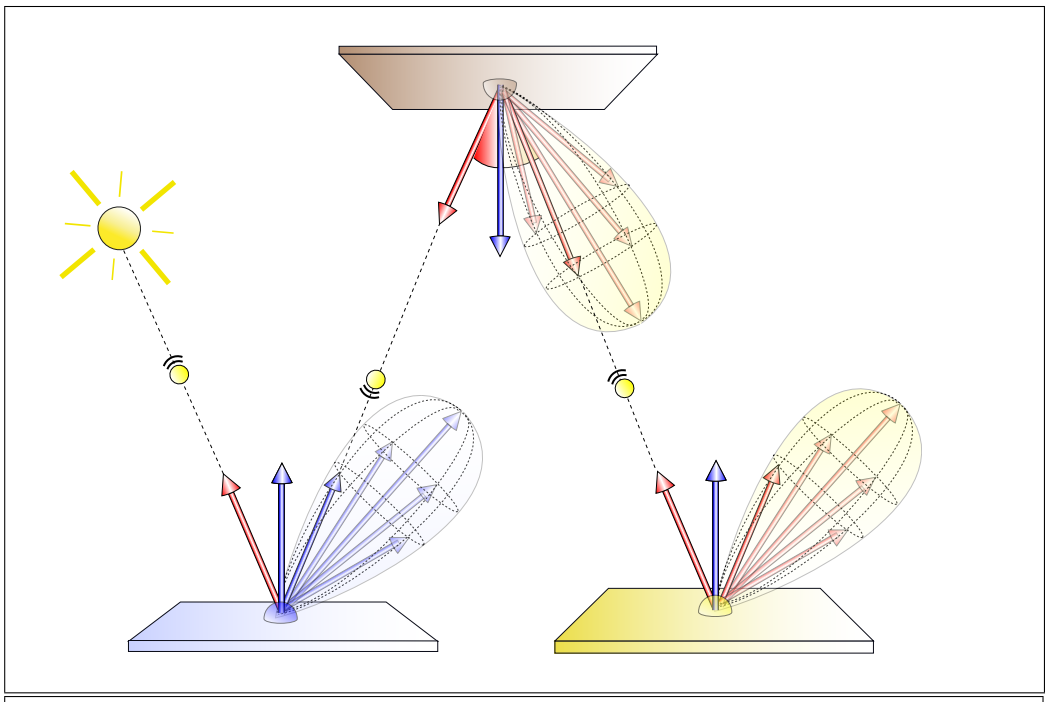


FIGURE 2.55: TRAVEL OF A PHOTON THROUGH A SCENE. At each hit point of a photon with a gloss object within the scene, the reflected direction of the photon is depending on its incoming direction. Thus, we can simulate the travel of a photon through a scene via the concept of a Markov process over the object surfaces of the environment.

EXAMPLE 2.85 (Markov Property) *Let us consider the travel of a photon through a scene consisting of objects with gloss surfaces, see Figure 2.55. Due to the surface properties of the objects, the outgoing direction of the photon after its collision with an object can not be exactly determined. For this, the outgoing direction should be chosen via the outcome of a random experiment, simulating the reflection behavior at an object surface.*

Now, the trip of a photon through the scene can take a long time, theoretically, it can bounce back and force infinitely between the objects. Thus, we need infinitely many random variables to simulate this process. Because the reflected direction of a photon at a gloss surface is not only dependent on the material and the surface properties of the object that has been hit, but also on the incident direction of the photon, the travel can be modeled via the probabilistic concept of the discrete-time Markov process.

Random Variable (168)
Glossy Reflection (304)
DT Markov Process (236)

As we will see in Section 2.4.7.2 in more detail, a discrete-time Markov process, $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$, is a sequence of not necessarily independent random variables defined over

a particular probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$. *We say that a sequence* $\mathbf{X}_n$ *of random variables has the* Markov property*, if it holds:*

$$\mathbb{P}(\mathbf{X}_{n+1} = i_{n+1} | \mathbf{X}_n = i_n, \ldots, \mathbf{X}_0 = i_0) = \mathbb{P}(\mathbf{X}_{n+1} = i_{n+1} | \mathbf{X}_n = i_n), \tag{2.794}$$

where $(i_n)_{n \in \mathbb{N}_0}$ *is a sequence of values, that can be interpreted as the state set of* $\mathbf{X}_n$. *Contrary to the general case, in which the probability of* $\mathbf{X}_{n+1}$ *being in state* j *depends on all other random variables* $\mathbf{X}_i, 0 \leq i \leq n$*, the Markov property says that the probability of* $\mathbf{X}_{n+1}$ *being in state* j *depends only on the state of the random variable* $\mathbf{X}_n$.

It can easily be proved by induction that all probabilities are completely fixed by a so-called initial distribution $\mathbb{P}(\mathbf{X} = i_0)$*, that is, the probability that the sequence starts in state* i_0*, and the conditional distributions* $\mathbb{P}(\mathbf{X}_{n+1} = i_{n+1} | \mathbf{X}_n = i_n)$. *Thus, for* $n = 0$ *we conclude from Equation (2.792):*

$$\mathbb{P}(\mathbf{X}_1 = i_1, \mathbf{X}_0 = i_0) = \mathbb{P}(\mathbf{X}_1 = i_1 | \mathbf{X}_0 = i_0) \, \mathbb{P}(\mathbf{X}_0 = i_0). \tag{2.795}$$

Due to the definition of the conditional probability, the Markov property, and the induction hypothesis

$$\mathbb{P}(\mathbf{X}_n = i_n, \ldots, \mathbf{X}_0 = i_0) = \mathbb{P}(\mathbf{X}_0 = i_0) \prod_{j=1}^{n} \mathbb{P}(\mathbf{X}_j = i_j | \mathbf{X}_{j-1} = i_{j-1}), \tag{2.796}$$

we then get:

$$\begin{aligned}
&\mathbb{P}(\mathbf{X}_{n+1} = i_{n+1}, \ldots, \mathbf{X}_0 = i_0) \\
&\overset{(2.792)}{=} \mathbb{P}(\mathbf{X}_{n+1} = i_{n+1} | \mathbf{X}_n = i_n, \ldots, \mathbf{X}_0 = i_0) \, \mathbb{P}(\mathbf{X}_n = i_n, \ldots, \mathbf{X}_0 = i_0) \\
&\overset{(2.794)}{=} \mathbb{P}(\mathbf{X}_{n+1} = i_{n+1} | \mathbf{X}_n = i_n) \, \mathbb{P}(\mathbf{X}_n = i_n, \ldots, \mathbf{X}_0 = i_0) \\
&\overset{(2.794),(2.796)}{=} \mathbb{P}(\mathbf{X}_0 = i_0) \prod_{j=1}^{n+1} \mathbb{P}(\mathbf{X}_j = i_j | \mathbf{X}_{j-1} = i_{j-1}).
\end{aligned} \tag{2.797}$$

CONDITIONAL CUMULATIVE DISTRIBUTION FUNCTION OF A RANDOM VARIABLE. According to the definition of the CDF of a random variable $\mathbf{X}$, we now also define the conditional cumulative distribution function of a random variable $\mathbf{Y}$ given $\mathbf{X}$. Let us begin with the discrete case.

For defining the conditional CDF, it is useful to introduce the concept of the *conditional probability mass function*, it is defined as follows:

DEFINITION 2.64 (Conditional Probability Mass Function) *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be discrete random variables or random vectors on a product probability space* $(\Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}, \mathfrak{F}_{\mathbf{X}}(\Omega_{\mathbf{X}}) \times \mathfrak{F}_{\mathbf{Y}}(\Omega_{\mathbf{Y}}), \mathbb{P}_{\mathbf{X}} \times \mathbb{P}_{\mathbf{Y}})$ *with finite or countably infinite image range. Then, the* conditional

Product Measure Space (81)
Countably Infinite Set (827)

probability mass function, $p_{\mathbf{Y}|\mathbf{X}}$, *of the random variable* $\mathbf{Y}$ *given* $\mathbf{X} = \mathbf{x}$ *is defined as:*

$$p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X} = \mathbf{x}) \stackrel{def}{=} \mathbb{P}_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) \tag{2.798}$$

$$= \frac{\mathbb{P}_{\mathbf{X},\mathbf{Y}}(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})}{\mathbb{P}_{\mathbf{X}}(\mathbf{X} = \mathbf{x})} \tag{2.799}$$

$$= \frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})}, \tag{2.800}$$

Joint PMF (185) Marignal PMF (188) *where* $p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})$ *is the joint probability mass function and* $p_{\mathbf{X}}(\mathbf{x})$ *is the associated marginal probability mass function of* $\mathbf{X}$.

DEFINITION 2.65 (Discrete Conditional Cumulative Distribution Function) *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be discrete random variables or random vectors on a product probability space* $(\Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}, \mathfrak{F}_{\mathbf{X}}(\Omega_{\mathbf{X}}) \times \mathfrak{F}_{\mathbf{Y}}(\Omega_{\mathbf{Y}}), \mathbb{P}_{\mathbf{X}} \times \mathbb{P}_{\mathbf{Y}})$ *with finite or countably infinite image range. Due to Definition 2.46 the* discrete conditional cumulative distribution function, $F_{\mathbf{Y}|\mathbf{X}}$, *of the discrete random variable* $\mathbf{Y}$ *given* $\mathbf{X} = \mathbf{x}$ *can be written as:*

Product Measure Space (81) Countably Infinite Set (827)

$$F_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X} = \mathbf{x}) \stackrel{def}{=} \mathbb{P}_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} = \mathbf{x}) \tag{2.801}$$

$$= \sum_{\mathbf{y}_i \leq \mathbf{y}} \mathbb{P}_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y} = \mathbf{y}_i|\mathbf{X} = \mathbf{x}) \tag{2.802}$$

$$\stackrel{(2.798)}{=} \sum_{\mathbf{y}_i \leq \mathbf{y}} p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X} = \mathbf{x}) \tag{2.803}$$

$$\stackrel{(2.800)}{=} \sum_{\mathbf{y}_i \leq \mathbf{y}} \frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}_i)}{p_{\mathbf{X}}(\mathbf{x})}. \tag{2.804}$$

In the following, let $\mathbf{X}$ and $\mathbf{Y}$ be continuous random variables or random vectors on a product probability space $(\Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}, \mathfrak{F}_{\mathbf{X}}(\Omega_{\mathbf{X}}) \times \mathfrak{F}_{\mathbf{Y}}(\Omega_{\mathbf{Y}}), \mathbb{P}_{\mathbf{X}} \times \mathbb{P}_{\mathbf{Y}})$ with joint probability density function $p_{\mathbf{X},\mathbf{Y}}$ and marginal density $p_{\mathbf{X}}$. Using the definition of the conditional probability from Relation (2.792), then we get:

Product Measure Space (81) Probability Density Function (189)

$$\mathbb{P}_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} \leq \mathbf{x}) \stackrel{def}{=} \frac{\mathbb{P}_{\mathbf{X},\mathbf{Y}}(\mathbf{X} \leq \mathbf{x}, \mathbf{Y} \leq \mathbf{y})}{\mathbb{P}_{\mathbf{X}}(\mathbf{X} \leq \mathbf{x})} \tag{2.805}$$

$$= \frac{\int_{(-\infty,\mathbf{x}]\times(-\infty,\mathbf{y}]} \mathrm{d}\left(\mathbb{P}_{\mathbf{X}} \times \mathbb{P}_{\mathbf{Y}}\right)(\mathbf{x}, \mathbf{y})}{\int_{(-\infty,\mathbf{x}]} \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x})} \tag{2.806}$$

$$= \frac{\int_{(-\infty,\mathbf{x}]\times(-\infty,\mathbf{y}]} p_{\mathbf{X},\mathbf{Y}}(\xi, \eta)\, \mathrm{d}\mu^s(\xi, \eta)}{\int_{(-\infty,\mathbf{x}]} p_{\mathbf{X}}(\xi)\, \mathrm{d}\mu^s(\xi)} \tag{2.807}$$

$$= \int_{(-\infty,\mathbf{y}]} \underbrace{\frac{\int_{(-\infty,\mathbf{x}]} p_{\mathbf{X},\mathbf{Y}}(\xi, \eta)\, \mathrm{d}\mu^s(\xi)}{\int_{(-\infty,\mathbf{x}]} p_{\mathbf{X}}(\xi)\, \mathrm{d}\mu^s(\xi)}}_{p_{\mathbf{Y}|\mathbf{X}}(\eta|\mathbf{X}\leq\mathbf{x})} \mathrm{d}\mu^s(\eta) \tag{2.808}$$

$$= \int_{(-\infty,\mathbf{y}]} p_{\mathbf{Y}|\mathbf{X}}(\eta|\mathbf{X} \leq \mathbf{x}) \mathrm{d}\mu^s(\eta). \tag{2.809}$$

This now implies to define a conditional probability density function $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X} \leq \mathbf{x})$ of $\mathbf{Y}$ given $\mathbf{X}$ via the fraction of the joint probability density $p_{\mathbf{X},\mathbf{Y}}$ and the marginal density $p_{\mathbf{X}}$ integrated with respect to $\mathbf{x}$.

DEFINITION 2.66 (Conditional Probability Density Function) *Let $\mathbf{X}$ and $\mathbf{Y}$ be random variables or random vectors on a product probability space $(\Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}, \mathfrak{F}_{\mathbf{X}}(\Omega_{\mathbf{X}}) \times \mathfrak{F}_{\mathbf{Y}}(\Omega_{\mathbf{Y}}), \mathbb{P}_{\mathbf{X}} \times \mathbb{P}_{\mathbf{Y}})$ with joint probability density function $p_{\mathbf{X},\mathbf{Y}}$ and marginal density $p_{\mathbf{X}}$. Then, the* conditional probability density function *of $\mathbf{Y}$, given that $\mathbf{X} \leq \mathbf{x}$, is defined by:*

Product Measure Space (81)
Probability Density Function (189)

$$p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X} \leq \mathbf{x}) \stackrel{\text{def}}{=} \frac{\int_{(-\infty,\mathbf{x}]} p_{\mathbf{X},\mathbf{Y}}(\xi,\mathbf{y})\, d\mu^s(\xi)}{\int_{(-\infty,\mathbf{x}]} p_{\mathbf{X}}(\xi)\, d\mu^s(\xi)}. \tag{2.810}$$

For the special case $\mathbf{X} = \mathbf{x}$, Relation (2.810) makes no sense, since the denominator would be zero, thus, we define the conditional probability density function *of $\mathbf{Y}$, given $\mathbf{X} = \mathbf{x}$ via:*

$$p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X} = \mathbf{x}) \stackrel{\text{def}}{=} \frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})}. \tag{2.811}$$

REMARK 2.85 *i) For the sake of simplicity in our formulas, in the following we will mostly use $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ for describing the conditional densities $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X} \leq \mathbf{x})$ and $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X} = \mathbf{x})$.*

ii) On the basis of the Relations (2.810) and (2.811) we then obtain the following two important identities for the joint probability density function $p_{\mathbf{X},\mathbf{Y}}$:

$$\begin{aligned} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) &= p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) && (2.812)\\ &= p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) p_{\mathbf{X}}(\mathbf{x}). && (2.813) \end{aligned}$$

DEFINITION 2.67 (Continuous Conditional Probability Distribution) *Let $\mathbf{X}$ and $\mathbf{Y}$ be random variables or random vectors on a product probability space $(\Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}, \mathfrak{F}_{\mathbf{X}}(\Omega_{\mathbf{X}}) \times \mathfrak{F}_{\mathbf{Y}}(\Omega_{\mathbf{Y}}), \mathbb{P}_{\mathbf{X}} \times \mathbb{P}_{\mathbf{Y}})$ with joint probability density function $p_{\mathbf{X},\mathbf{Y}}$ and marginal density $p_{\mathbf{X}}$. Then, the* conditional probability distribution *of $\mathbf{Y}$ given $\mathbf{X}$, is defined by:*

Product Measure Space (81)
Probability Density Function (189)

$$\begin{aligned} F_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) &\stackrel{\text{def}}{=} \mathbb{P}_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y} \leq \mathbf{y}|\mathbf{X} \leq \mathbf{x}) && (2.814)\\ &= \int_{(-\infty,\mathbf{y}]} p_{\mathbf{Y}|\mathbf{X}}(\eta|\mathbf{X} \leq \mathbf{x}) d\mu^s(\eta) && (2.815)\\ &= \int_{(-\infty,\mathbf{y}]} \underbrace{\frac{\int_{(-\infty,\mathbf{x}]} p_{\mathbf{X},\mathbf{Y}}(\xi,\eta)\, d\mu^s(\xi)}{\int_{(-\infty,\mathbf{x}]} p_{\mathbf{X}}(\xi)\, d\mu^s(\xi)}}_{p_{\mathbf{Y}|\mathbf{X}}(\eta|\mathbf{X} \leq \mathbf{x})} d\mu^s(\eta). && (2.816) \end{aligned}$$

When we know the probability of an event B, given the occurrence of an event A, often, we are interested in computing so-called *inverse probabilities*, such as the conditional

probability $P(A|B)$. This can be done via *Bayes' Theorem* involving the so-called *prior* or *unconditional probabilities* of A and B. With the help of Definition 2.63 we can simply derive *Bayes' theorem*:

THEOREM 2.8 (Bayes' Theorem) *Let* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *be a probability space and* A, B *be two events from* $\mathfrak{F}(\Omega)$*, then it holds:*

$$\frac{\mathbb{P}(B|A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A|B)}{\mathbb{P}(A)}. \tag{2.817}$$

PROOF 2.8 *Bayes' theorem can easily be proved via*

$$\mathbb{P}(B|A)\mathbb{P}(A) \overset{(2.792)}{=} \mathbb{P}(A \cap B) \overset{(2.792)}{=} \mathbb{P}(A|B)\mathbb{P}(B). \tag{2.818}$$

The key idea behind Bayes' theorem is that the probability of an event A given an event B depends not only on the relationship between events A and B but also on the marginal probability of occurrence of each event. The Bayes' theorem plays an important role for understanding and analyzing *acceptance-rejection sampling*.

Section 6.5.2

Section 6.6.1

Another important concept, which we will use in our further discussions, is the *conditional expectation* of a random variable or a random vector. It is based on the concept of the *conditional probability density*. We will use the results, derived above, to introduce the concept of the conditional expected value of a continuous random variable.

DEFINITION 2.68 (Conditional Expected Value of a Continuous Random Variable or a Random Vector) *The* conditional expected value of a continuous random variable *or a* continuous random vector $G = (g(\mathbf{X}, \mathbf{Y}))$ *where* $\mathbf{X}$ *and* $\mathbf{Y}$ *are also random variables or random vectors on* $(\Omega_1 \times \Omega_2, \mathfrak{F}_1(\Omega_1) \times \mathfrak{F}_2(\Omega_2), \mathbb{P}_1 \times \mathbb{P}_2)$ *is defined as:*

Continuous RV (168)

Product Measure Sace (81)

$$\begin{aligned} E_{\mathbf{Y}}(G) \equiv E(G|\mathbf{X}) &\overset{def}{=} \int_{\Omega_2} g(\mathbf{x}, \mathbf{y}) p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \, d\mu_2^s(\mathbf{y}) && (2.819)\\ &\overset{(2.813)}{=} \int_{\Omega_2} g(\mathbf{x}, \mathbf{y}) \frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{p_{\mathbf{X}}(\mathbf{x})} \, d\mu_2^s(\mathbf{y}) && (2.820)\\ &= \frac{\int_{\Omega_2} g(\mathbf{x}, \mathbf{y}) p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) \, d\mu_2^s(\mathbf{y})}{\int_{\Omega_2} p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) \, d\mu_2^s(\mathbf{y})}. && (2.821) \end{aligned}$$

We conclude this section with a lemma that allows to express the variance of a random variable or a random vector $G = (g(\mathbf{X}, \mathbf{Y}))$ in terms of its conditional expectation. We use the result of this lemma in Section 6.6.1.

LEMMA 2.4 *Let* $\mathbf{X}, \mathbf{Y}$ *and* $G = (g(\mathbf{X}, \mathbf{Y}))$ *be random variables or random vectors on* $(\Omega_1 \times \Omega_2, \mathfrak{F}_1(\Omega_1) \times \mathfrak{F}_2(\Omega_2), \mathbb{P}_1 \times \mathbb{P}_2)$*, then it holds:*

Continuous RV (168)

Product Measure Sace (81)

$$Var(G) = E_{\mathbf{X}}(Var_{\mathbf{Y}}(G)) + Var_{\mathbf{X}}(E_{\mathbf{Y}}(G)). \tag{2.822}$$

PROOF 2.4 *Due to the Relation (2.776) the variance of* $G(\mathbf{X},\mathbf{Y})$ *can be written as:*

$$
\begin{aligned}
&Var(G)\\
&= \int_{\Omega_1\times\Omega_2} (g(\mathbf{x},\mathbf{y}))^2\, p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\, d\mu^2(\mathbf{x},\mathbf{y}) - \qquad (2.823)\\
&\quad \left(\int_{\Omega_1\times\Omega_2} g(\mathbf{x},\mathbf{y})\, p_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})\, d\mu^2(\mathbf{x},\mathbf{y})\right)^2\\
&\overset{(2.810),(2.811)}{=} \int_{\Omega_1} p_{\mathbf{X}}(\mathbf{x}) \left(\int_{\Omega_2} (g(\mathbf{x},\mathbf{y}))^2\, p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, d\mu(\mathbf{y})\right) d\mu(\mathbf{x}) - \qquad (2.824)\\
&\quad \left(\int_{\Omega_1} p_{\mathbf{X}}(\mathbf{x}) \left(\int_{\Omega_2} g(\mathbf{x},\mathbf{y})\, p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\, d\mu(\mathbf{y})\right) d\mu(\mathbf{x})\right)^2\\
&\overset{(2.819)}{=} \int_{\Omega_1} p_{\mathbf{X}}(\mathbf{x}) E_{\mathbf{Y}}\left(G^2\right) d\mu(\mathbf{x}) - \left(\int_{\Omega_1} p_{\mathbf{X}}(\mathbf{x}) E_{\mathbf{Y}}(G) d\mu(\mathbf{x})\right)^2 \qquad (2.825)\\
&\overset{(2.821)}{=} E_{\mathbf{X}}\left(E_{\mathbf{Y}}\left(G^2\right)\right) - E^2_{\mathbf{X}}(E_{\mathbf{Y}}(G)) \qquad (2.826)\\
&= E_{\mathbf{X}}\left(E_{\mathbf{Y}}\left(G^2\right)\right) - E_{\mathbf{X}}\left(E_{\mathbf{Y}}\left(G^2\right)\right) + E_{\mathbf{X}}\left(E_{\mathbf{Y}}\left(G^2\right)\right) - E^2_{\mathbf{X}}\left(E_{\mathbf{Y}}(G)\right) \qquad (2.827)\\
&= E_{\mathbf{X}}\left(E_{\mathbf{Y}}\left(G^2\right) - E_{\mathbf{Y}}\left(G^2\right)\right) + E_{\mathbf{X}}\left(E_{\mathbf{Y}}\left(G^2\right)\right) - E^2_{\mathbf{X}}\left(E_{\mathbf{Y}}(G)\right) \qquad (2.828)\\
&\overset{(2.776)}{=} E_{\mathbf{X}}\left(Var_{\mathbf{Y}}(G)\right) + Var_{\mathbf{X}}(E_{\mathbf{Y}}(G)). \qquad (2.829)
\end{aligned}
$$

REMARK 2.86 *The variance of the random variable* $G(\mathbf{X},\mathbf{Y})$ *allows the representation of* G *as the sum of a conditional expected value and a conditional variance. Relation (2.829) proves very useful further below in the discussion of* sampling techniques, *particularly when analyzing* use of expected values, *and* stratified sampling, *two so-called variance reduction techniques used in Monte Carlo integration.* Section 6.5 Section 6.6.1 Section 6.6.4

REMARK 2.87 *For simplifying our formulas, in the following we will often neglect the index in the notation of the conditional probability density function. Therefore, we will often write* $p(\mathbf{y}|\mathbf{x})$ *instead of* $p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})$ *respectively* $p(\mathbf{x}|\mathbf{y})$ *instead of* $p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$.

2.4.6 THE LAWS OF LARGE NUMBERS AND THE CENTRAL LIMIT THEOREM

In preparation for the definition of the Lebesgue integral we presented a series of types of convergence: the pointwise, and the uniform convergence, already introduced in Section 2.1.1, as well as the ν-almost everywhere convergence and the convergence according to measure. With respect to random variables, all these types of convergence formalize in different ways the convergence behavior of a sequence of random variables towards a particular random variable. This then allows to consider the concept of the random variable from different perspectives. As a random variable is a measurable function, but can also be interpreted as a realization of a random experiment endowed with a distribution function, Lebesgue Integral (105) a.e. Convergence (103) Convergence in Measure (103) Random Variable (168)

we can couple the idea of convergence to the notion of pointwise convergence of a sequence of measurable functions, but also to properties of random variables such as the expected value, moments, or other quantities.

Now, in all of our previous discussions, the distributions of a sequence of random variables were always known, which will not be the case when discussing the realistic modeling of practical applications. Often, then only very little information—such as the expected value or other moments—is available about the random variables involved. Here, it would be useful, if at least information about the behavior of a large sum of random variables would be available. The limit theorems of probability theory will show, that via the behavior of the mean value of a large number of random variables, relatively detailed, approximate statements may be derived about the distribution of the random variables and the occurring probabilities. But before we will devote our interest to the limit theorems of probability theory, let us start with the famous *Chebychev inequality*.

Chebychev Inequality (212)

Measurable Function (98)
Measure Space (80)
Lebesgue Integral (105)

THE LAWS OF LARGE NUMBERS. Let us assume f represents a non-negative, real-valued, measurable function on the measure space $(\Omega, \mathfrak{F}, \mu)$, then, with $0 < p < \infty$ as well as $0 < \epsilon < \infty$ and the features of the Lebesgue integral, the following clearly applies:

$$\frac{1}{\epsilon^p} \int_{\mathcal{R}} f^p(x)\, d\mu(x) \overset{\mathcal{R} \supset \{x|f(x) \geq \epsilon\}}{\geq} \frac{1}{\epsilon^p} \int_{\{x|f(x)\geq\epsilon\}} f^p(x)\, d\mu(x) \tag{2.830}$$

$$\overset{f(x)\geq\epsilon}{\geq} \frac{1}{\epsilon^p} \int_{\{x|f(\omega)\geq\epsilon\}} \epsilon^p\, d\mu(x) \tag{2.831}$$

$$= \mu\{x \,|\, f(x) \geq \epsilon\}. \tag{2.832}$$

If we reformulate this relation as follows:

$$\mu\{x|f(x) \geq \epsilon\} \leq \frac{1}{\epsilon^p} \int_{\mathcal{R}} f^p(x)\, d\mu(x), \tag{2.833}$$

then we obtain one of the most fundamental inequalities of measure and integration theory: the *Chebyshev Inequality*. In probability theory, the Chebyshev Inequality is mostly used in the following version:

Probability Space (163)
Random Variable (168)
Expected Value (196)
Variance (201)

THEOREM 2.9 (The Chebyshev Inequality) *Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be any probability space, $\mathbf{X}$ a random variable or a random vector defined on $(\Omega, \mathfrak{F}, \mathbb{P})$, and $0 < \epsilon < \infty$, then it holds*

$$\text{prob}\Big\{ \big|\mathbf{X} - E(\mathbf{X})\big| \geq \epsilon \Big\} \leq \frac{Var(\mathbf{X})}{\epsilon^2}. \tag{2.834}$$

Measurable Function (98)

PROOF 2.9 *Replacing the real-valued, measurable function f from Equation (2.833) by*

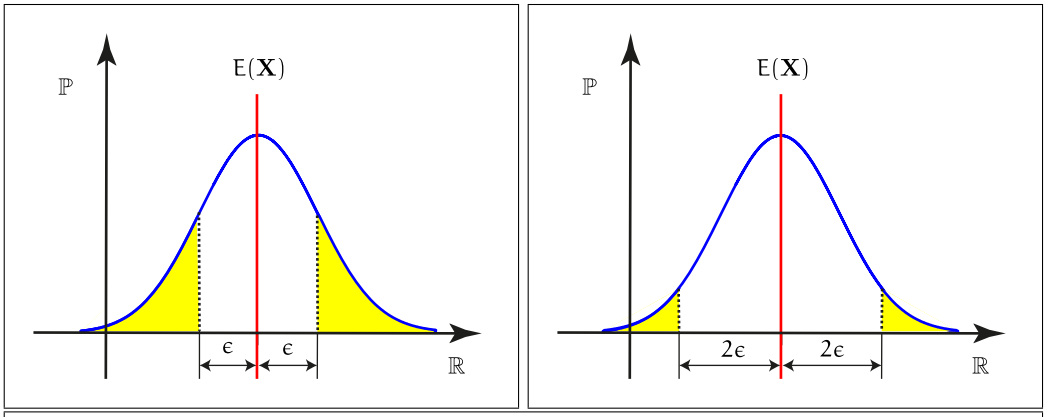


FIGURE 2.56: THE CHEBYSHEV INEQUALITY. The values of a random variable with finite variance are close to the expected value of the random variable. The probability that a value of the random variable lies outside the ϵ-strip in the left image is greater than the probability that it lies outside the double ϵ-strip in the right image.

the random variable $\mathbf{X} - \mathrm{E}(\mathbf{X})$ *and the Lebesgue measure* μ *by the probability measure* $\mathbb{P}$*, then we obtain for the case* $p = 2$*:*

$$\begin{aligned} \mathrm{prob}\left\{\left|\mathbf{X} - \mathrm{E}(\mathbf{X})\right| \geq \epsilon\right\} &\overset{(2.607)}{=} \mathbb{P}\{\omega \,|\, |\mathbf{X}(\omega) - \mathrm{E}(\mathbf{X})| \geq \epsilon\} && (2.835)\\ &\overset{(2.833)}{\leq} \frac{1}{\epsilon^2}\int_{\Omega} (\mathbf{X}(\omega) - \mathrm{E}(\mathbf{X}))^2 \, \mathrm{d}\mathbb{P}(\omega) && (2.836)\\ &\overset{(2.761)}{=} \frac{\mathrm{Var}(\mathbf{X})}{\epsilon^2}. && (2.837) \end{aligned}$$

Evendently, the Chebyshev inequality delivers a tool for estimating the probability of a non-negative random variable by means of the 2^{nd} moments of this random variable. It states that the values of the random variable $\mathbf{X}$, with finite variance, are close to the mean value, or a little more precisely: If we choose $\epsilon = k\,\sigma, k > 0$, where σ is the *standard deviation*, which is defined as the square root of the 2^{nd} central moment of $\mathbf{X}$, thus:

Moment of a RV (201)
Variance (201)

$$\sigma \overset{def}{=} \sqrt{\mathrm{Var}(\mathbf{X})}, \tag{2.838}$$

then the Chebyshev Inequality can be formulated as:

$$\mathrm{prob}\left\{\left|\mathbf{X} - \mathrm{E}(\mathbf{X})\right| \geq k\,\sigma\right\} \leq \frac{1}{k^2}. \tag{2.839}$$

This means, that for $k = \sqrt{2}$, half of the values of $\mathbf{X}$ are no more than $\sqrt{2}$ standard deviations away from the mean. Following this, a first statement on the asymptotical behavior of a sequence of random variables may be made via Chebyshev's inequality from Equation (2.834) which in turn leads to *Chebyshev's Weak Law of Large Numbers*.

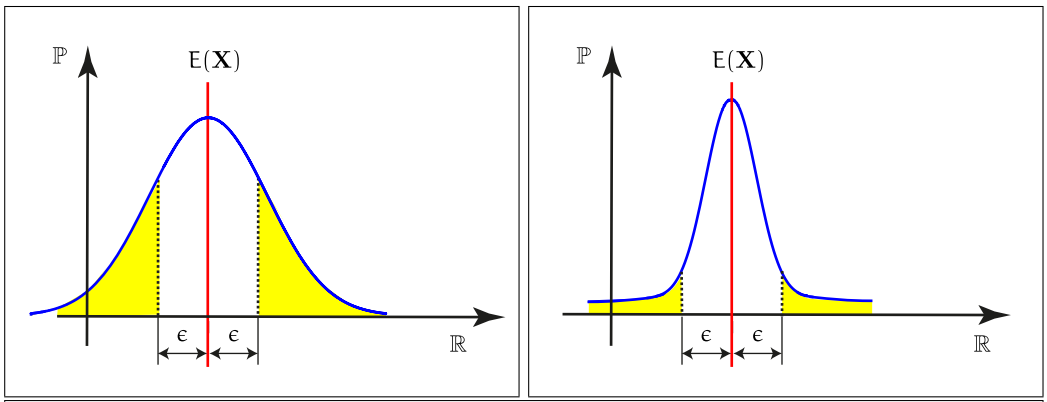


FIGURE 2.57: CHEBYSHEV'S WEAK LAW OF LARGE NUMBERS. Left, the PDF p_{n_1} of $\sum_{i=1}^{n_1} \mathbf{X}_i$, right, the PDF p_{n_2} of $\sum_{i=1}^{n_2} \mathbf{X}_i$ with $n_1 \ll n_2$. The probability of sampling a value of the random variable close to its expected value is more likely with p_{n_2} than with p_{n_1}. The WLLN states that, for a given ϵ, the area under the graph of the PDF tends to 0 as n grows without limit. In other words, outside the ϵ-strip, the *tails* of the PDF become negligible when n goes to infinity.

THEOREM 2.10 (Chebyshev's Weak Law of Large Numbers, WLLN) *Let* $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ *representing independent random variables or random vectors, not all necessarily uniformly distributed, with* $E(\mathbf{X}_i) = E(\mathbf{X})$ *and* $Var(\mathbf{X}_i) \leq Var(\mathbf{X}) < \infty$ *for* $1 \leq i \leq n$, *then Equation (2.834) implies:*

Independent RV (204)
Uniform Distribution (180)

$$\text{prob}\left\{\left|\frac{\mathbf{X}_1 + \ldots + \mathbf{X}_n}{n} - E(\mathbf{X})\right| \geq \epsilon\right\} \leq \frac{Var(\mathbf{X})}{n\epsilon^2}, \tag{2.840}$$

thus:

$$\frac{\mathbf{X}_1 + \ldots + \mathbf{X}_n}{n} \xrightarrow{\mathbb{P}} E(\mathbf{X}), \tag{2.841}$$

Convergence In Measure (103)

i.e. the random variable $\frac{\mathbf{X}_1+\ldots+\mathbf{X}_n}{n}$ *converges in probability towards the expected value* $E(\mathbf{X})$.

Linearity of EV (202)

PROOF 2.10 *The proof is very easy. From linearity property of the expected value we obtain:*

$$E\left(\frac{\sum_{i=1}^{n} \mathbf{X}_i}{n}\right) \overset{(2.773)}{=} \frac{1}{n}\sum_{i=1}^{n} E(\mathbf{X}_i) \tag{2.842}$$

$$\overset{E(\mathbf{X}_i)=E(\mathbf{X})}{=} E(\mathbf{X}). \tag{2.843}$$

Since the random variables are independent, we get due to Relation (2.791):

$$\mathrm{Var}\left(\frac{\sum_{i=1}^{n} \mathbf{X}_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(\mathbf{X}_i) \tag{2.844}$$

$$\overset{\mathrm{Var}(\mathbf{X}_i)\leq \mathrm{Var}(\mathbf{X})}{\leq} \frac{1}{n}\mathrm{Var}(\mathbf{X}). \tag{2.845}$$

Applying Chebyshev's inequality then leads to:

$$\mathrm{prob}\left\{\left|\frac{\mathbf{X}_1 + \ldots + \mathbf{X}_n}{n} - \mathrm{E}(\mathbf{X})\right| \geq \epsilon\right\} \leq \frac{\mathrm{Var}\left(\frac{\sum_{i=1}^{n} \mathbf{X}_i}{n}\right)}{\epsilon^2} \tag{2.846}$$

$$\overset{(2.845)}{\leq} \frac{\mathrm{Var}(\mathbf{X})}{n\epsilon^2}. \tag{2.847}$$

REMARK 2.88 *As a consequence of Chebyshev's inequality the Weak Law of Large Numbers says that the average of a set of random variables converges in probability towards the expected value of the random variables. Thus, if we choose a margin, no matter how small, then there will be a very high probability that with a sufficiently large sample size the average of observations will be close to its expected value.*

Convergence In Measure (103)
Expected Value (196)

As we will see later in a subsequent chapter of our book, often in many practical situations it is not ever possible to compute the value of an integral directly. Then, the Weak Law of Large Numbers is the basis of a probabilistic integration method which gives an approximate solution by random sampling of points, the so-called *Monte Carlo integration*. Let us illustrate this technique by means of a simple example.

Chapter 6
Section 6.5

EXAMPLE 2.86 *Let $f \in \mathcal{L}^2(\mathbb{R}, \mu)$ be an integrable real-valued function defined on the unit interval $[0,1]$. Let furthermore $X, X_1, \ldots, X_n$ be uniformly distributed random variables on $[0,1]$. Due to Chebyshev's Weak Law of Large Numbers it can be shown that the random variable*

Integrable Function (105)
Uniform Distribution (180)

$$S_n \overset{\mathrm{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f(X_i) \tag{2.848}$$

converges in probability towards the integral

Convergence In Measure (103)

$$\mathcal{I} \overset{\mathrm{def}}{=} \int_{[0,1]} f(x)\, \mathrm{d}\mu(x). \tag{2.849}$$

Now, the random variables $X, X_1, \ldots, X_n$ are uniformly distributed on $[0,1]$, that is, for the random variable $f(X)$ it holds:

Uniform Distribution (180)

$$\mathrm{E}(f(X)) \overset{(2.735)}{=} \int_{[0,1]} f(x)\, \mathrm{d}\mathbb{P}_X(x) \tag{2.850}$$

$$\overset{\mathrm{d}\mathbb{P}_X = \mathrm{d}\mathbb{P}_U}{=} \int_{[0,1]} f(u)\, \mathrm{d}\mathbb{P}_U(u), \tag{2.851}$$

where U is as usual our standardized on $[0,1]$ uniformly distributed random variable. PDF (176) *With the probability density function $p_U \equiv 1$ this then leads to:*

$$E(f(X)) = \int_{[0,1]} f(u)\, p_U(u)\, d\mu(u) \tag{2.852}$$

$$\overset{p_U \equiv 1}{=} \int_{[0,1]} f(u)\, d\mu(u). \tag{2.853}$$

*Since $X, X_1, \ldots, X_n$ fulfills the requirements of Chebyshev's Weak Law of Large Numbers—$E(X) = E(X_1) = \ldots = E(X_n)$ and the variance of f, as a Lebesgue square-*Integrable Function (105) *integrable function, is obviously finite—the random variable S_n then provides the desired result*

$$S_n \xrightarrow{\mathbb{P}} \int_{[0,1]} f(x)\, d\mu(x). \tag{2.854}$$

If in Chebyshev's weak law of large numbers one limits oneself to identically distributed random numbers, then one obtains— without the variance condition from WLLN and under as weak conditions as possible—for a large general class of random variables *Kolmogorov's Strong Law of Large Numbers.*

THEOREM 2.11 (Kolmogorov's Strong Law of Large Numbers) *Let us assume $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$* Uniformly Distributed RV (180) *representing independent, identically distributed random variables or random vectors with $E(\mathbf{X}_i) = E(\mathbf{X})$, then the following holds:*

$$\frac{\mathbf{X}_1 + \ldots + \mathbf{X}_n}{n} \xrightarrow{a.e.} E(\mathbf{X}). \tag{2.855}$$

PROOF 2.11 *The proof of Kolmogorov's Strong Law of Large Numbers is more complex than that of Chebyshev's Weak Law of Large Numbers. Because we need only the result of this theorem, we omit the proof and refer the interested reader to [15, Ash & Doléans-Dade 2000].*

With respect to the analysis of algorithms, Kolmogorov's Strong Law of Large Numbers implies that the arithmetic mean of a sequence of random variables converges towards Expected Value (196) the expected value of the random variables not only according to measure, but in fact $\mathbb{P}$-Convergence a.e. (103) almost everywhere except for a set of measure zero. This in turn may be interpreted, Measure (79) that every observation sequence gained on the basis of stochastically independent random variables ultimately leads to the expected value via an appropriate construction of a mean value.

THE CENTRAL LIMIT THEOREM, 1D CASE. As demonstrated in probability theory, Kolmogorov's Strong and Chebyshev's Weak Law of Large Numbers supply information on

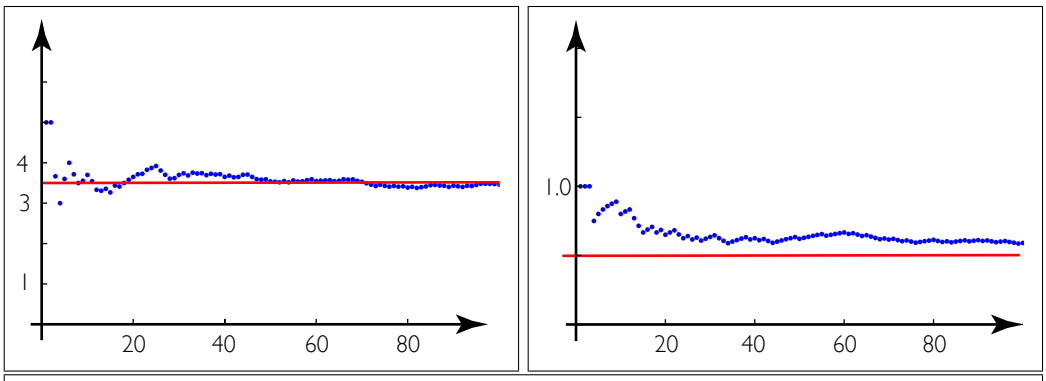


FIGURE 2.58: ILLUSTRATIONS OF KOLMOGOROV'S STRONG LAW OF LARGE NUMBERS. The left image visualize the rolling of an idealized die, the right image the flipping of an idealized coin. From both images you can see: if the number of rolls or flips increases, the average of the values of all results approaches to the value 3.5 for the die and 0.5 for the coin, and these are the expected values of the associated random variables.

the convergence of the expected value of random variables but they say nothing about their assigned distributions.

Such approximations are, however, of particularly great importance as the precise distribution of the arithmetic mean of random variables is not easly to calculate. Regarding this point, a fundamental statement is provided by the *Central Limit Theorem* according to which the arithmetic mean of n independent random variables with arbitrary distributions is at least approximately normally distributed, if their variances are finite. If we limit ourselves to one-dimensional, identically distributed random variables, then the statement of the Central Limit Theorem may be formulated as follows:

Random Variable (168)

THEOREM 2.12 (The Central Limit Theorem, 1D Case) *If* $X, X_1, \ldots, X_n$ *are independent and identically distributed one-dimensional random variables with expected value* $E(X_i)$ *and finite variance* $Var(X_i) < \infty$. *The following then applies for the random variable:*

Random Variable (168)

$$S_n \stackrel{def}{=} \frac{\sum_{i=1}^{n} (X_i - E(X_i))}{\sqrt{n}\,\sigma} \tag{2.856}$$

as n *goes to* ∞

$$\text{prob}\,(S_n \leq x) \longrightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, d\mu(t), \tag{2.857}$$

that is, the standardized random variable S_n *from above is asymptotically* normal distributed, *see Figure 2.59.*

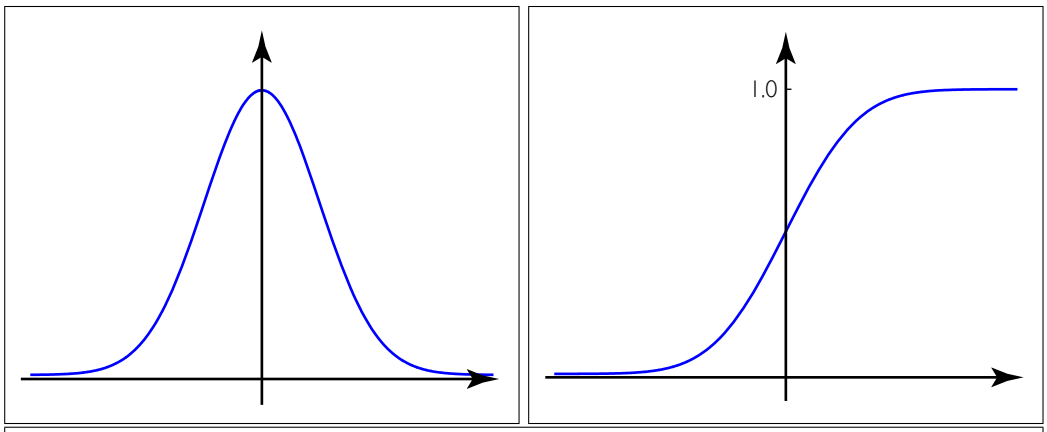


FIGURE 2.59: 1D NORMAL DISTRIBUTION. PDF and CDF of the 1D normal distribution $\mathcal{N}(0,1)$, thus with expected value $\mu = 0$ and standard deviation $\sigma = 1$. The left figure confirms, what mathematically can be shown: 68.27% of all values are located in an interval of standard deviation $\pm\sigma$ around the expected value, 95.45% of all values are located in the interval $[-2\sigma, 2\sigma]$, and 99.73% of all values are located $[-3\sigma, 3\sigma]$.

PROOF 2.12 *For a proof, see [15, Ash & Doléans-Dade 2000] or [84, Hesse 2003].*

REMARK 2.89 *In probability theory, the central limit theorem can be considered as a set of weak-convergence statements. Common to all of these statements is that the sum of a large number of independent and identically distributed random variables will tend to be normally distributed. Since many processes in real world can be considered as the average result of many unknown random processes, the CLT also justifies their normal distributed character.*

2.4.7 STOCHASTIC PROCESSES

Random Variable (168)

As we have seen in the previous sections, random variables are highly useful in modeling static problems in which randomness plays an important role. When discussing sequences of random experiments we also observed that the outcome for each random experiment was not influenced by the outcomes of previous experiments. The reason for this was that all these random experiments are based on independent distributed random variables.

Independent Distributed RV (168)

Recalling Example 2.85, where we have analyzed the travel of a photon through a scene consisting of gloss objects. As the scattered direction of the photon at a surface not only depends on the material and surface properties of the object that has been hit, but also on the incoming direction, modeling the interaction of a photon at a surface via a random experiment also requires to account for the outcome of the previous experiment. An adequate description of such events requires the construction of particular types of

random variables, so-called *stochastic processes*. They deliver information how a process might evolve under time.

In this section, we first present the probabilistic model of the *stochastic process* and talk about a classification of stochastic processes. Via the Markov property, introduced in Section 2.4.5, we then define a great class of stochastic processes: *Markov chains* and *Markov processes*. So, we will develop the most important properties of Markov chains and Markov processes and shortly discuss their use in algorithms for solving the light transport problem.

Section 2.4.7.1
Section 2.4.7.2

From now on, we free ourselves from the way of thinking static and focus our attention on the temporal development of a random experiment. Because of its complexity and due to the fact that it opposed to the dynamic way of thinking, the set Ω will completely occur in the background. Our central concept will not be a random experiment, but the concept of the stochastic process.

DEFINITION 2.69 (Stochastic Process) *A* stochastic process *is a family of random variables* $(\mathbf{X}_t)_{t\in T}$ *defined on the probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$*, where* T *is any arbitrary index set. The set of all possible values* $\mathbf{X}_t(\omega)$*, which the random variable* $\mathbf{X}_t$ *can take on, is denoted as the* state space *or the* state set S *of a stochastic process. In particular,* $\mathbf{X}_t(\omega)$ *is interpreted as the* state of the process at time t.

Random Variable (168)
Probability Space (163)

Let us present a first simple example of a stochastic process known from probability courses in school: *The Bernoulli chain*.

EXAMPLE 2.87 (Bernoulli Chain) *A* Bernoulli chain *is an example of a stochastic process. It can be modeled by a finite or countably infinite sequence* $(X_n)_{n\in\mathbb{N}_0}$ *of independent and identically distributed random variables* X_n *each with two outcomes* $\{0, 1\}$*, where it holds:* $\mathbb{P}(X_n = 0) = p$ *and* $\mathbb{P}(X_n = 1) = 1 - p$.

Usually, stochastic processes are classified by its index set and its state space. So we speak of *processes*, if the state space S is uncountably infinite; in the case, where the state space S is finite or countably infinite, a stochastic process is usually called a *chain*. Via the type of the index set T, we then further specify a process or a chain. If T is uncountably infinite, a process is called a *continuous-time process* and a chain is called a *continuous-time chain*, while in the case where T is finite or countably infinite we denote a process a *discrete-time process* and a chain is called a *discrete-time chain*.

Countable Set (827)
Uncountable Set (827)

EXAMPLE 2.88 (Bernoulli Chain, continued) *Obviously, a* Bernoulli chain *is a discrete-time, discrete-state stochastic process.*

EXAMPLE 2.89 *Let* T *and* S *be two finite or countably infinite sets, then the sequence of random variables* $(\mathbf{X}_t)_{t\in T}$ *represents a discrete-time chain. In the case where*

T *is uncountably infinite, we speak of a continuous-time chain. If both sets are uncountably infinite,* $(\mathbf{X}_t)_{t\in T}$ *is a continuous-time process.*

In the following example, we will introduce the concept of the *random walk*. As we will see later in our book, the concept of the random walk is the mathematical foundation of the most-promising ray based rendering algorithms.

Section 9

EXAMPLE 2.90 (Random Walk) *Let* $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ *be a discrete-time, continuous-state stochastic process, where* $\mathbf{X}_n = (X_{1_n}, \ldots, X_{s_n}) \in A \subseteq \mathbb{R}^s$ *are independent and identically distributed* s*-dimensional random variables, then the random variable* $\mathbf{S}_n$ *defined by:*

$$\mathbf{S}_n \stackrel{def}{=} \sum_{i=0}^{n} \mathbf{X}_i \tag{2.858}$$

is called a random walk *in* A*. An example of a 2-dimensional random walk is shown in the left image of Figure 2.60.*

Let us assume $A = \mathbb{Z}^2$*, then a random walk in* $\mathbb{Z}^2$ *can be simulated via a sequence of independent and identically distributed, random variables* $(X_n)_{n\in\mathbb{N}_0}$ *taking the four outcomes* $\{0, 1, 2, 3\}$*, where it holds:*

$$\mathbb{P}(X = i) = \frac{1}{4}, \quad i \in \{0, 1, 2, 3\}. \tag{2.859}$$

If the random walk is at point $\mathbf{S}_{n-1} = (p, q) \in \mathbb{Z}^2$ *at time* $n-1$*, then it will be continued at time* n *as follows:*

$$S_n = S_{n-1} + \begin{cases} (p-1, q) & \text{if} \quad X = 0 \\ (p+1, q) & \text{if} \quad X = 1 \\ (p, q-1) & \text{if} \quad X = 2 \\ (p, q+1) & \text{if} \quad X = 3. \end{cases} \tag{2.860}$$

For an illustration of a random walk in $\mathbb{Z}^2$*, see the right image in Figure 2.60.*

REMARK 2.90 *In the following, we do not always explicitly specify the type of a stochastic process or a chain, and we use again and again the synonyms* process*, for a discrete or a continuous-time process, as well as* chain*, for a discrete or a continuous-time chain.*

REMARK 2.91 *The processes, which we will discuss and analyze in the following, are all either discrete-time processes or discrete-time chains, as their index set is finite or countably infinite. That is, the family of random variables associated with a discrete-time chain are discrete random variables with values in the measurable space* $(S, \mathfrak{P}(S))$*, and the random variables associated with a discrete-time process are continuous with values in the measurable space* $(S, \mathfrak{B}(S))$.

$\mathfrak{P}(\cdot)$ (828)
Measurable Space (80)

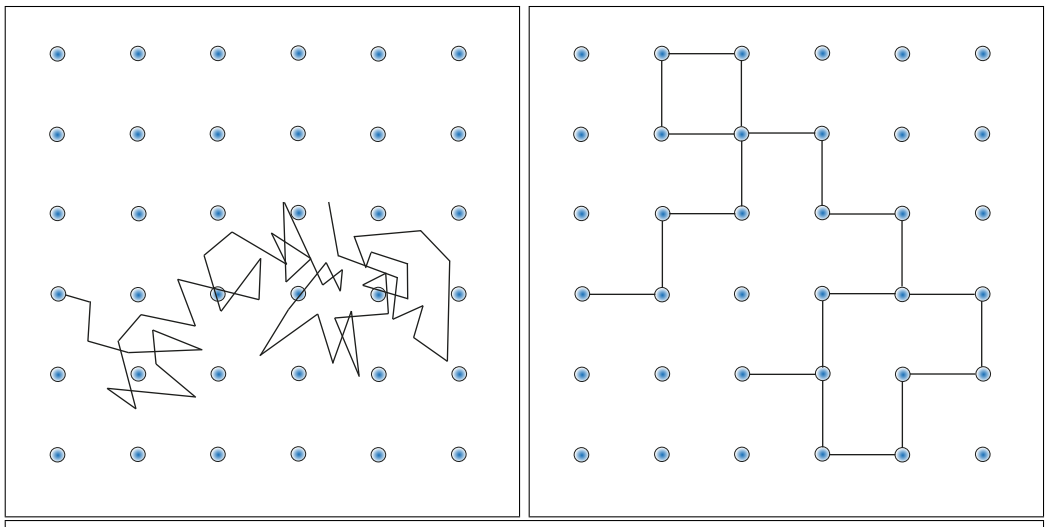


FIGURE 2.60: RANDOM WALKS. Left, a continuous random walk is shown. At state $n-1$ the random walk is continued in x- and y-direction via a uniformly distributed, 2-dimensional random variable $\mathbf{X} = (X, Y) \in [0,1]^2$. On the right-hand side a discrete random walk is shown. At each state, the process makes a step in one of the directions left, right, top, or bottom.

Let us close this introductory section on stochastic processes with a few examples of great interest for us. In the first example, we show how the travel of an abstract particle through a small scene can be modeled via a discrete-time random walk.

EXAMPLE 2.91 (Discrete Travel of a Photon Through a Small Scene) *Let us consider the discrete travel of a photon through a small scene consisting of four diffuse surfaces, where a photon can be reflected at each surface only in one of three possible directions, see Figure 2.61. Let us assume, the photon can start at any of the surfaces* $\{s_1, s_2, s_3, s_4\}$ *and makes a transition from* s_i *to* s_j*, denoted by* $s_i \to s_j$*, for* $1 \leq i, j \leq 4, i \neq j$*. Since all surfaces are diffuse, we can simulate the travel of the photon through the scene by a stochastic process of independent and identically distributed random variables* $(X_n)_{n \in \mathbb{N}_0}$*, with* $X_n \in \{s_1, s_2, s_3, s_4\}$*. Obviously, the probability distribution of the random variables is given by:* $\mathbb{P}(s_i \to s_j) = \frac{1}{3}$ *for* $i \neq j$ *and* $\mathbb{P}(s_i \to s_j) = 0$ *for* $i = j$*, where* $X_n = s_i$ *and* $X_{n+1} = s_j$ *is assumed.*

In the next example, we consider a stochastic process where the involved random variables are identically distributed, but does not take on the same values.

EXAMPLE 2.92 (Transport Paths in a Scene) *Let us consider a scene with diffuse and specular objects, labeled by* D *and* S*, which are illuminated by light sources* L*. There*

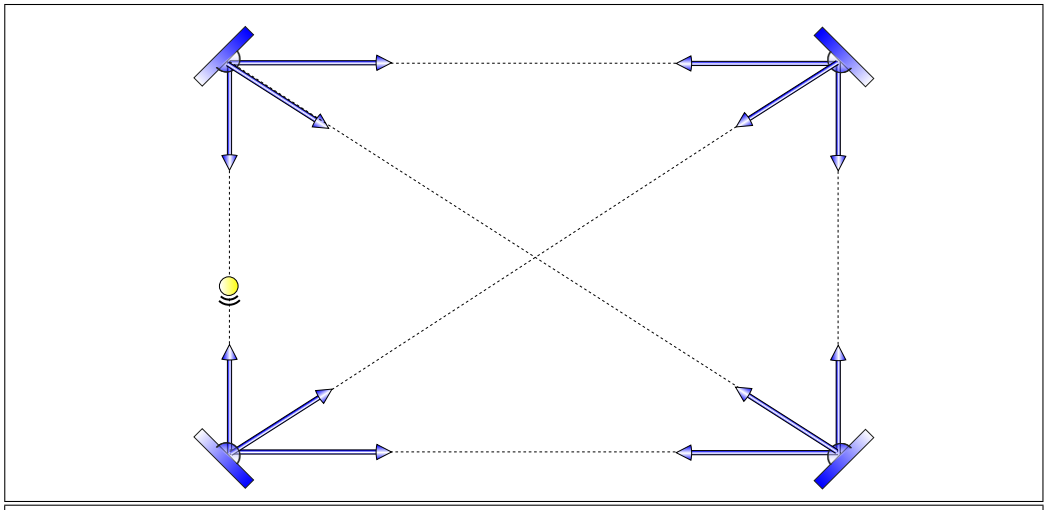


FIGURE 2.61: DISCRETE TRAVEL OF A PHOTON THROUGH A SMALL SCENE. At any state of the scene, a photon can make three transitions to one of another state. Since all surfaces in the scene are diffuse, the probability of a transition is given by $\frac{1}{3}$.

is also an observer E *within the scene. We are interested in all possible paths passing over the objects starting at a light source and ending at the eye of an observer—we will discuss the concept of the light and eye path in more detail in Section 8.1. Obviously, there exist a path from a light source to the eye, this is the single path that has length one. Path of lengths two can be constructed from* L*, via a diffuse or specular object to the eye. The graph in Figure 2.62 represents all possible paths between* L *and* E*.*

Now, the random experiment of constructing of a path between a light source and the eye *can be described by a stochastic process* $(X_n)_{n\in\mathbb{N}_0}$ *with finite state space* $S = \{L, D, S, E\}$*, where* L *corresponds to the starting point of the chain and* E *is the end state of the chain. The chain enters in states* D, S *or* E *with probability* $\frac{1}{3}$*. Transitions from* L *to* L*, or* E *to* E *occurs with probability zero.*

In the following example, now we will present a stochastic process based approach for solving linear systems of equations. In Section 6.7.3, we will see that this approach can easily be extended for finding appropriate approximative solutions to linear integral equations.

Section 2.3

EXAMPLE 2.93 (The Expected Value of a Sequence of Random Variables as Solution to a Linear System of Equations) *Let us consider a system of linear equations, in vector*

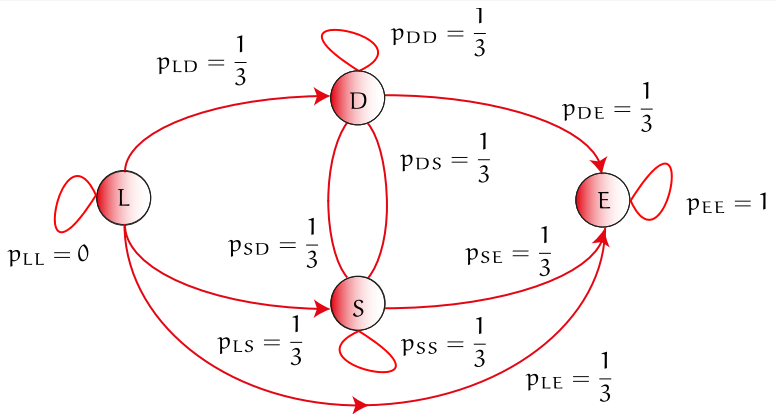


FIGURE 2.62: SIMULATION OF TRANSPORT PATHS VIA A STOCHASTIC PROCESS. The random experiment *constructing of a path between a light source and the eye* can be simulated via a stochastic process based on the finite state space $S = \{L, D, S, E\}$, where L is the starting point of the chain and E the end state of the chain. The chain starts with probability $p_0 = 1$ in state L. It can loop in the states D and S with probability $p_{DD} = p_{SS} = \frac{1}{3}$.Transition from L to L respectively E and E are not allowed.

form written as:

$$\mathbf{M}\,\mathbf{x} = \mathbf{b}, \tag{2.861}$$

where the column vectors $\mathbf{x}, \mathbf{b}$ *are from* $\mathbb{R}^n$ *and* $(m_{ij})_{1 \le i,j \le n}$ *is a* $n \times n$*-dimensional regular matrix with coefficients from* $\mathbb{R}$*.*

Replacing the matrix $\mathbf{M}$ *in Relation (2.861) by* $(\mathbf{I} - \mathbf{A})$*, where* $\mathbf{I}$ *is the identity, then the system from above may be written similarly to an operator equation as:* Linear Operator Equation (61)

$$\mathbf{x} = \mathbf{A}\,\mathbf{x} + \mathbf{b}. \tag{2.862}$$

Under the condition that the operator $\mathbf{A}$ *contracts, i.e.* $\|\mathbf{A}\| < 1$*, due to Equation (2.387) there exists the inverse to* $(\mathbf{I} - \mathbf{A})$ *in form of a Neumann series. The exact solution of Equation (2.861) is then given by:* Operator Norm (56) Neumann series (135)

$$\mathbf{x} = \mathbf{M}^{-1}\,\mathbf{b} = (\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{b} = \sum_{i=0}^{\infty} \mathbf{A}^i\,\mathbf{b}. \tag{2.863}$$

Now, on a computer it is not possible to compute an exact solution of the Neumann series, which is why we search for an approximation $\tilde{\mathbf{x}}$ *of Equation (2.863). After* $m+1$ *steps such an approximation can be obtained via:*

$$\tilde{\mathbf{x}}^{(m+1)} = \sum_{i=0}^{m} \mathbf{A}^i\,\mathbf{b}, \tag{2.864}$$

with $\widetilde{\mathbf{x}}^{(0)} = 0$ *and* $\mathbf{A}^0 = \mathbf{I}$. *Then, the* j^{th} *component of* $\widetilde{\mathbf{x}}^{(m+1)}$ *is equal to:*

$$\begin{aligned}\widetilde{x}_j^{(m+1)} &= b_j + \sum_{i_1=1}^{n} a_{ji_1} b_{i_1} + \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} a_{ji_1} a_{i_1 i_2} b_{i_2} \\ &\quad + \ldots + \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_m=1}^{n} a_{ji_1} a_{i_1 i_2} \ldots a_{i_{m-1} i_m} b_{i_m}\end{aligned} \tag{2.865}$$

or in closed form:

$$\widetilde{x}_j^{(m+1)} = b_j + \sum_{k=1}^{m}\left(\sum_{i_1=1}^{n}\cdots\sum_{i_k=1}^{n}\left(a_{ji_1}\prod_{l=1}^{k-1} a_{i_l i_{l+1}}\right) b_{i_k}\right).$$

Obviously, the iterate $\widetilde{x}_j^{(m+1)}$ *can also be expressed as:*

$$\begin{aligned}\widetilde{x}_j^{(m+1)} &= b_j + \sum_{i_1=1}^{s}\frac{a_{ji_1}}{p_{ji_1}} b_{i_1} p_{ji_1} + \\ &\quad \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\frac{a_{ji_1} a_{i_1 i_2}}{p_{ji_1 i_2}} b_{i_2} p_{ji_1 i_2} + \ldots + \\ &\quad \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_m=1}^{n}\frac{a_{ji_1} a_{i_1 i_2}\ldots a_{i_{m-1} i_m}}{p_{ji_1 i_2 \ldots i_m}} b_{i_m} p_{ji_1 i_2 \ldots i_m}\end{aligned} \tag{2.866}$$

$$\begin{aligned}&= b_j + \\ &\quad \sum_{k=1}^{m}\left(\sum_{i_1=1}^{n}\cdots\sum_{i_k=1}^{n}\left(\frac{a_{ji_1}\prod_{l=1}^{k-1} a_{i_l i_{l+1}}}{p_{ji_1\ldots i_k}}\right)\cdot b_{i_k}\cdot p_{ji_1\ldots i_k}\right).\end{aligned} \tag{2.867}$$

Under the condition, that $p_{ji_1}, p_{ji_1 i_2}, \ldots, p_{ji_1\ldots i_m}$ *are positive real numbers, the single terms in the representation of* $\widetilde{x}_j^{(m+1)}$ *can be interpreted as the expected values of discrete random variables* $\mathbf{X}_{ji_1}, \ldots, \mathbf{X}_{ji_1\ldots i_m}$ *where it holds:*

$$\mathbf{X}_{ji_1} \in \left\{\frac{a_{ji_1}}{p_{ji_1}} b_{i_1} \,\middle|\, 1 \le i_1 \le n\right\} \tag{2.868}$$

$$\vdots$$

$$\mathbf{X}_{ji_1\ldots i_m} \in \left\{\frac{a_{ji_1} a_{i_1 i_2}\ldots a_{i_{m-1} i_m}}{p_{ji_1\ldots i_m}} b_{i_m} \,\middle|\, 1 \le i_1, \ldots, i_m \le n\right\}. \tag{2.869}$$

Now, with $\mathbf{X}_{ji_1}, \ldots, \mathbf{X}_{ji_1\ldots i_m}$, *obviously also the sum of these random variables, given by:*

$$\mathbf{X} \stackrel{def}{=} b_j + \sum_{k=1}^{m}\mathbf{X}_{ji_1\ldots i_k}, \tag{2.870}$$

is also a random variable.

Choosing probability distributions, induced by the associated probability measures, via:

$$\begin{aligned}
\mathbb{P}\left(\mathbf{X}_{ji_1} = \frac{a_{ji_1}}{p_{ji_1}} b_{i_1}\right) &= p_{ji_1} \\
&\vdots \qquad (2.871)\\
\mathbb{P}\left(\mathbf{X}_{ji_1\ldots i_m} = \frac{a_{ji_1} a_{i_1 i_2} \ldots a_{i_{m-1} i_m}}{p_{ji_1\ldots i_m}} b_{i_m}\right) &= p_{ji_1\ldots i_m} \qquad (2.872)
\end{aligned}$$

then $\mathbf{X}_{ji_1}, \ldots, \mathbf{X}_{ji_1\ldots i_m}$ are independent distributed random variables and the expected value of the sum $\mathbf{X}$ of these random variables corresponds to the iterate $\widetilde{x}_j^{(m+1)}$. This can easily be shown as follows: For computing the expected value of $\mathbf{X}$, we have to multiply each value of $\mathbf{X}_{ji_1\ldots i_k}$, $k \geq 1$ by its probability $p_{ji_1\ldots i_k}$, that is, the expected value of the random variable $\mathbf{X}$ is then given by:

Expected Value (197)

$$\begin{aligned}
E(\mathbf{X}) &= E\left(b_j + \sum_{k=1}^{m} \mathbf{X}_{ji_1\ldots i_k}\right) && (2.873)\\
&= b_j + E\left(\mathbf{X}_{ji_1}\right) + \ldots + E\left(\mathbf{X}_{ji_1\ldots i_m}\right) && (2.874)\\
&= b_j + \sum_{i_1=1}^{s} \frac{a_{ji_1}}{p_{ji_1}} b_{i_1} p_{ji_1} + \\
&\quad \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \frac{a_{ji_1} a_{i_1 i_2}}{p_{ji_1 i_2}} b_{i_2} p_{ji_1 i_2} + \ldots + && (2.875)\\
&\quad \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_m=1}^{n} \frac{a_{ji_1} a_{i_1 i_2} \ldots a_{i_{m-1} i_m}}{p_{ji_1\ldots i_m}} b_{i_m} p_{ji_1\ldots i_m} \\
&= b_j + \sum_{i_1=1}^{n} a_{ji_1} b_{i_1} + \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} a_{ji_1} a_{i_1 i_2} b_{i_2} && (2.876)\\
&\quad + \ldots + \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_m=1}^{n} a_{ji_1} a_{i_1 i_2} \ldots a_{i_{m-1} i_m} b_{i_m} && (2.877)\\
&= b_j + \sum_{k=1}^{m}\left(\sum_{i_1=1}^{n}\cdots\sum_{i_k=1}^{n}\left(a_{ji_1}\prod_{l=1}^{k-1} a_{i_l i_{l+1}}\right) b_{i_k}\right). && (2.878)
\end{aligned}$$

Since the last equation corresponds to the same formula that we have derived when computing the $m+1^{th}$ approximation of the j^{th} component of the solution $\mathbf{x}$ of $\mathbf{Mx} = \mathbf{b}$, we obtain:

$$\widetilde{x}_j^{(m+1)} = E(\mathbf{X}). \qquad (2.879)$$

Obviously, this example shows that the solution of a linear system of equations can be interpreted as the expected value of a sum of independent distributed, discrete random variables which are not identically distributed. That is, a sufficiently long series of trials based on an independent random experiment—associated with the above stochastic process—leads to the solution of Equation (2.861). But how can we model such a random experiment? Before we can answer this question we have to talk about Markov chains and Markov processes.

2.4.7.1 DISCRETE-TIME MARKOV CHAINS

In the following we assume that a family of random variables $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$, defined on the probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$, is given with values in a finite or countably infinite state space S.

Probability Space (163)
Countable Set (827)

DEFINITION 2.70 (Discrete-time Markov Chain) *Let $i_0, \ldots, i_{n-1}, i, j$ be a sequence of $n+2$ points from a finite or countably infinite state space S. Furthermore, let us assume that the sequence of discrete random variables $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ satisfies the Markov property, thus,*

Markov Property (207)

$$\mathbb{P}(\mathbf{X}_{n+1} = j | \mathbf{X}_n = i, \mathbf{X}_{n-1} = i_{n-1}, \ldots, \mathbf{X}_0 = i_0) = \mathbb{P}(\mathbf{X}_{n+1} = j | \mathbf{X}_n = i). \quad (2.880)$$

Then, the sequence of random variables $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ is called a discrete-time Markov chain*, or simply a* Markov chain*, in computer graphics one also often speaks of a* random walk*. In this case $\mathbf{X}_n$ is interpreted as the state of the Markov chain at time n, and we say, the chain is in state i at time n, if it holds $\mathbf{X}_n = i$.*

If we refer to $\mathbf{X}_{n+1}$ as the future, $\mathbf{X}_{n-1}, \ldots, \mathbf{X}_0$ as the past and $\mathbf{X}_n$ as the present, then, due to the Markov property, the probability that a Markov chain will be in the future in a particular state depends solely on the present state and its index n, not however, on any states adopted in the past. The characteristic of a Markov chain is its property, that with knowledge about the prehistory of the chain, the same statements about the future development are possible as with knowledge about the whole prehistory of the process.

The probabilities from Equation (2.880) are denoted as the *transition probabilities*, because they provide information of all the potentially possible transitions of $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ from state i into state j, see Figure 2.64. If all these probabilities are independent from time point n, we denote them as

$$p_{ij} \stackrel{def}{=} \mathbb{P}(\mathbf{X}_{n+1} = j | \mathbf{X}_n = i), \quad (2.881)$$

and we say that the Markov chain is *homogeneous*.

REMARK 2.92 *Unless stated otherwise, we assume in the following that all stochastic processes in which we are interested are homogeneous.*

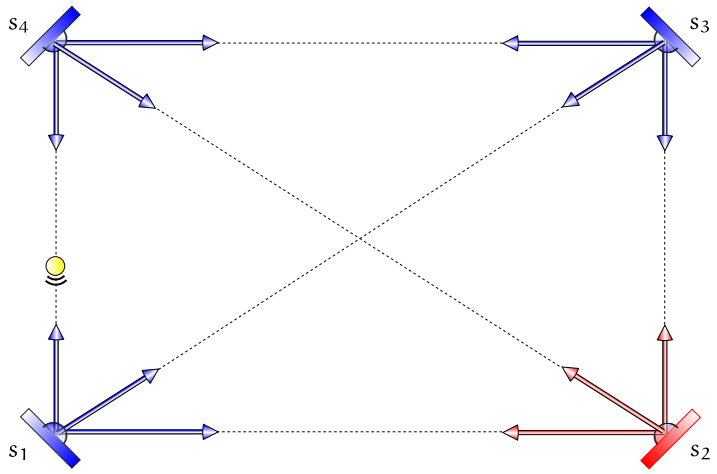


FIGURE 2.63: DISCRETE TRAVEL OF A PHOTON THROUGH A SMALL SCENE. At any state of the scene, a photon can make three transitions to one of the other state. At the three diffuse surfaces, the probability of a transition is given by $\frac{1}{3}$, but at the gloss surface, the interaction of the photon is predetermined by the incident direction.

EXAMPLE 2.94 (Discrete Travel of a Photon Through a Small Scene, Continued) *Let us consider the discrete travel of a photon through the small scene from Example 2.91. Now, we will assume, that one of the surfaces, namely the surface S_2, is perfectly specular, see Figure 2.63. Obviously, the reflection at the specular surface is dependent on the incident direction of the photon at the surface, that is, the reflection at the specular surface can not be simulated via an independent random variable, as we did it in Example 2.91. Here, we have to simulate the travel of the photon via a discrete Markov chain.*

The state space S can now be chosen as the set of all directions within Figure 2.63, that is:

$$S = \{\omega_{ij} \mid \text{there is an edge between } s_i \text{ and } s_j \text{ for } i \neq j\}, \tag{2.882}$$

where for the associated probability distribution at diffuse surfaces it holds:

$$\mathbb{P}(\omega_{jk}|\omega_{ij}) = \frac{1}{3}, \tag{2.883}$$

and—due to the Law of Reflection— the probability distribution at the specular surface is given by

Law of Reflection (300)

$$\mathbb{P}(\omega_{23}|\omega_{12}) = 1 \tag{2.884}$$

$$\mathbb{P}(\omega_{24}|\omega_{42}) = 1 \tag{2.885}$$

$$\mathbb{P}(\omega_{21}|\omega_{32}) = 1, \tag{2.886}$$

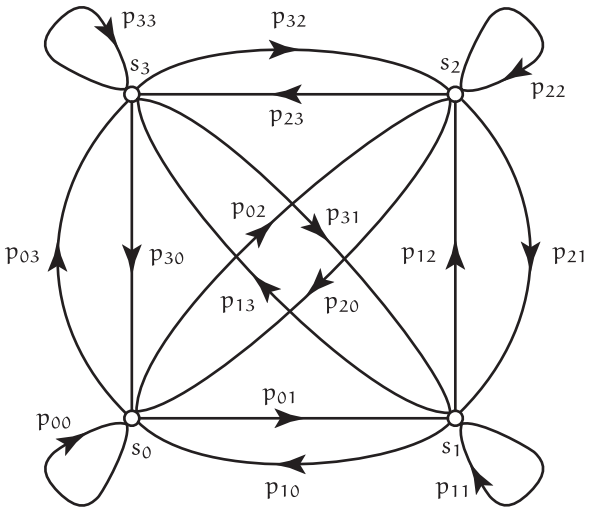


FIGURE 2.64: VISUALIZATION OF A MARKOV CHAIN. A discrete-time Markov chain generated over the state space $S = \{s_0, s_1, s_2, s_3\}$ with associated transition probabilities $p_{ij} = \frac{1}{4}, 0 \leq i, j \leq 3$.

all other probabilities, such as the impossible transition from state ω_{13} into state ω_{14}, have to be zero.

Via the transition probabilities p_{ij} and the *initial distribution* p_0 of the random variable $\mathbf{X}_0$, given by

$$p_0 \stackrel{\text{def}}{=} \mathbb{P}(\mathbf{X}_0 = i_0), \tag{2.887}$$

then all common distributions of a Markov chain are fixed, since it holds

$$\mathbb{P}(\mathbf{X}_n = i_n, \ldots, \mathbf{X}_0 = i_0) \stackrel{(2.797)}{=} \mathbb{P}(\mathbf{X}_0 = i_0) \left(\prod_{k=1}^{n} \mathbb{P}\left(\mathbf{X}_k = i_k \middle| \mathbf{X}_{k-1} = i_{k-1}\right) \right) \tag{2.888}$$

$$= p_0 \prod_{k=1}^{n} p_{i_{k-1} i_k}. \tag{2.889}$$

EXAMPLE 2.95 (Transport Paths in a Scene, Continued) *Since the chain in Example 2.92 starts at the eye, the initial distribution is given by*

$$p_0 = \mathbb{P}(\mathbf{X}_0 = E) = 1. \tag{2.890}$$

The transition probabilities are given by

$$p_{ij} = \mathbb{P}(\mathbf{X}_1 = j | \mathbf{X}_0 = i) = \frac{1}{3} \tag{2.891}$$

for $i, j \in \{L, D, S, E\}$ *with* $i \neq j$ *and*

$$p_{LL} = p_{EE} = 0 \quad \text{and} \quad p_{DD} = p_{SS} = \frac{1}{3}. \tag{2.892}$$

Based on these results, the probability that the process is in state E *after* n*-steps—thus, a path of length* n *was generated—corresponds to*

$$\begin{aligned} \mathbb{P}(\mathbf{X}_3 = E, \ldots, \mathbf{X}_0 = L) &= \mathbb{P}(\mathbf{X}_0 = L) \cdot \left(\prod_{k=1}^{n-1} \mathbb{P}\left(\mathbf{X}_n = i_k \middle| \mathbf{X}_{k-1} = i_{k-1}\right) \right) \cdot \\ &\quad \mathbb{P}\left(\mathbf{X}_n = E \middle| \mathbf{X}_{n-1} = i_{n-1}\right) && (2.893) \\ &= 1 \cdot \prod_{k=1}^{n} \frac{1}{3} && (2.894) \\ &= \frac{1}{3^n}, && (2.895) \end{aligned}$$

with $i_k \in \{D, S\}$.

To get a better overview of a system, whose temporal development can be described by a Markov chain, we summarize the transition probabilities p_{ij} in a so-called *transition matrix* $\mathbf{M} = (p_{ij})_{i,j \in S}$ of the form

Matrix (853)

$$\mathbf{M} \stackrel{\text{def}}{=} \begin{Vmatrix} p_{00} & p_{01} & p_{02} & \cdots & p_{0i} & p_{0i+1} & \cdots \\ p_{10} & p_{11} & p_{12} & \cdots & p_{1i} & p_{1i+1} & \cdots \\ p_{20} & p_{21} & p_{22} & \cdots & p_{2i} & p_{2i+1} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{Vmatrix}. \tag{2.896}$$

In the special case, where the state space S of the corresponding Markov chain is finite, the matrix $\mathbf{M}$ is of finite dimension, otherwise it is infinitely dimensional. Since $\mathbf{M}$ must satisfy the conditions

$$p_{ij} \geq 0, \qquad \sum_{j \in S} p_{ij} = 1, \qquad i \in S, \tag{2.897}$$

we call $\mathbf{M}$ a *stochastic matrix*.

EXAMPLE 2.96 (Discrete Travel of a Photon Through a Small Scene, Continued) *The stochastic matrix* $\mathbf{M}$*, associated with the Markov chain from Example 2.94, is obvi-*

ously given by

$$\mathbf{M} \stackrel{\text{def}}{=} \begin{Vmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \end{Vmatrix}. \tag{2.898}$$

Obviously, $\mathbf{M}$ *can easily be derived from the transition diagram of the Markov chain from Figure 2.63 by setting a coefficient* p_{ij} *with the transition probability from state* i *into state* j.

REMARK 2.93 *Any Markov chain, represented in form of a stochastic matrix* $\mathbf{M}$*, can also easily be visualized by a corresponding directed graph* $G = (N, V)$*, where the node set* N *is given by the row-numbers* $1, 2, \ldots$ *and the edge set* V *is specified via the transition probabilities, that is, there is an edge in* G *from node* i *to node* j*, if* $p_{ij} \neq 0$.

Based on the transition probabilities, we are now able to define the probabilities that a Markov chain in state i will be in state j after n additional steps, that is,

$$p_{ij}^n \stackrel{\text{def}}{=} \mathbb{P}(\mathbf{X}_{n+m} = j \big| \mathbf{X}_m = i), \tag{2.899}$$

for all $n, m \in \mathbb{N}_0, m \geq 1, i, j \in S$, which we call the n*-step transition probabilities* p_{ij}^n. With $p_{ij}^1 \stackrel{\text{def}}{=} p_{ij}$, they can be written recursively as:

$$p_{ij}^{n+m} = \sum_{k \in S} p_{ik}^n p_{kj}^m \tag{2.900}$$

and are referred to as the *Chapman-Kolmogorov equations*. Here the transition from state i in state j in $n + m$ transitions can be done by going in n steps from state i to a state k and then from k to state j in m additional steps, where we must sum over all intermediate states k. In matrix notation formulated, the Chapman-Kolmogorov equations can be written as the product of the n^{th} and the m^{th} power of the transition matrix $\mathbf{M}$, also

$$\mathbf{M}^{n+m} = \mathbf{M}^n \cdot \mathbf{M}^m. \tag{2.901}$$

EXAMPLE 2.97 (Transport Paths in a Scene, Continued) *The stochastic matrix* $\mathbf{M}$*, associated with the Markov chain from Example 2.92, is obviously given by*

$$\mathbf{M} \stackrel{\text{def}}{=} \begin{Vmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 & 0 \end{Vmatrix}. \tag{2.902}$$

Based on the matrix $\mathbf{M}$*, the* n*-step transition probabilities* p_{ij}^n *are the coefficient of the matrix:*

$$\mathbf{M}^n \stackrel{\text{def}}{=} \begin{Vmatrix} 0 & \frac{2^{n-1}}{3^n} & \frac{2^{n-1}}{3^n} & \frac{2^{n-1}}{3^n} \\ 0 & \frac{2^{n-1}}{3^n} & \frac{2^{n-1}}{3^n} & \frac{2^{n-1}}{3^n} \\ 0 & \frac{2^{n-1}}{3^n} & \frac{2^{n-1}}{3^n} & \frac{2^{n-1}}{3^n} \\ 0 & 0 & 0 & 0 \end{Vmatrix}. \tag{2.903}$$

Thus, the transition probability $p_{04}^4 = \frac{8}{81}$ *corresponds to the probability for generating all 8 possible paths* $L(D|S)^3E$ *between* L *and* E. Section 8.1

Let us consider once more Example 2.93 where we have shown, that the expected value of a stochastic process can be interpreted as solution of a linear system of equations.

EXAMPLE 2.98 (Solving a Linear System of Equations by Simulating a Discrete-time Markov Chain, Continued) *Equation (2.863) has shown that the approximate solution* $\widetilde{\mathbf{x}}$ *to the linear system of equations from (2.861) can be written as:*

$$\widetilde{\mathbf{x}}^{(m+1)} = \sum_{i=0}^{m} \mathbf{A}^i \, \mathbf{b}, \tag{2.904}$$

with $\widetilde{\mathbf{x}}^{(0)} = 0$ *and* $\mathbf{A}^0 = \mathbf{I}$*, and the* j^{th} *component of* $\widetilde{\mathbf{x}}^{(m+1)}$ *was given by:*

$$\begin{aligned} \widetilde{x}_j^{(m+1)} \quad = \quad & b_j + \sum_{i_1=1}^{n} a_{ji_1} b_{i_1} + \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} a_{ji_1} a_{i_1 i_2} b_{i_2} \\ & + \ldots + \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \ldots \sum_{i_m=1}^{n} a_{ji_1} a_{i_1 i_2} \ldots a_{i_{m-1} i_m} b_{i_m}. \end{aligned} \tag{2.905}$$

Let us now consider a discrete-time Markov chain $(\mathbf{X}_n)_{n \in \mathbb{N}_0}$ *with initial distribution* $\mathbf{p}_0 \stackrel{\text{def}}{=} (0, \ldots, \underbrace{1}_{p_j}, \ldots, 0)$ *and transition probabilities* $(p_{ij})_{1 \le i,j \le n}$*:*

$$\sum_{j=1}^{n} p_{ij} = 1, \qquad p_{ij} > 0 \quad \text{if } a_{ij} \neq 0, \quad i, j = 1, \ldots, n. \tag{2.906}$$

We model the chain $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ *by a random walk* $j = i_0 \to i_1 \to i_2 \to \ldots \to i_m$ *of length* m *starting at state* $j = i_0$ *that passes through a sequence of states* $i_1, i_2, \ldots, i_m$. *With such a random walk, then we associate a random variable* $\mathbf{X}_m$, *given by:*

$$\mathbf{X}_m \stackrel{\text{def}}{=} \mathbf{X}_{ji_0} + \sum_{k=1}^{m} \mathbf{X}_{ji_1\ldots i_k}, \tag{2.907}$$

where for $\mathbf{X}_{ji_1\ldots i_m}$ *for* $i_0 \le i_j \le i_m$ *it holds:*

$$\mathbf{X}_{ji_0} \in \left\{ \frac{b_j}{p_j} \right\} \tag{2.908}$$

$$\mathbf{X}_{ji_1} \in \left\{ \frac{a_{ji_1}}{p_j\, p_{ji_1}} b_{i_1} \,\middle|\, 1 \le i_1 \le n \right\} \tag{2.909}$$

$$\vdots$$

$$\mathbf{X}_{ji_1\ldots i_m} \in \left\{ \frac{a_{ji_1} a_{i_1 i_2} \ldots a_{i_{m-1} i_m}}{p_j\, p_{ji_1} p_{i_1 i_2} \cdots p_{i_{m-1} i_m}} b_{i_m} \,\middle|\, 1 \le i_1, \ldots, i_m \le n \right\}, \tag{2.910}$$

and the random variables $\mathbf{X}_{ji_0}, \mathbf{X}_{ji_1}, \mathbf{X}_{ji_1 i_2}, \ldots, \mathbf{X}_{ji_1\ldots i_m}$ *are distributed according to*

$$\mathbb{P}\left(\mathbf{X}_{ji_0} = \frac{b_j}{p_j} \right) = p_j \tag{2.911}$$

$$\mathbb{P}\left(\mathbf{X}_{ji_1} = \frac{a_{ji_1}}{p_j\, p_{ji_1}} b_{i_1} \right) = p_j\, p_{ji_1}$$

$$\vdots \tag{2.912}$$

$$\mathbb{P}\left(\mathbf{X}_{ji_1\ldots i_m} = \frac{a_{ji_1} a_{i_1 i_2} \ldots a_{i_{m-1} i_m}}{p_j\, p_{ji_1} p_{i_1 i_2} \cdots p_{i_{m-1} i_m}} b_{i_m} \right) = p_j\, p_{ji_1} p_{i_1 i_2} \cdots p_{i_{m-1} i_m}. \tag{2.913}$$

Since the the random variable $\mathbf{X}_m$ *is defined along the path* $j = i_0 \to i_1 \to i_2 \to \ldots \to i_m$ *we get for the expected value of* $\mathbf{X}_m$*:*

$$E(\mathbf{X}_m) = E\left(\mathbf{X}_{ji_0} + \sum_{k=1}^{m} \mathbf{X}_{ji_1\ldots i_k} \right) \tag{2.914}$$

$$= E(\mathbf{X}_{ji_0}) + E(\mathbf{X}_{ji_1}) + \ldots + E(\mathbf{X}_{ji_1\ldots i_m}) \tag{2.915}$$

$$= \frac{b_j}{p_j} p_j + \sum_{i_1=1}^{s} \frac{a_{ji_1}}{p_j\, p_{ji_1}} b_{i_1} p_j\, p_{ji_1} + \sum_{i_1=1}^{n} \sum_{i_2=1}^{s} \frac{a_{ji_1} a_{i_1 i_2}}{p_j\, p_{ji_1} p_{i_1 i_2}} b_{i_2} p_j\, p_{ji_1} p_{i_1 i_2} + \ldots + \sum_{i_1=1}^{n} \sum_{i_2=1}^{s} \cdots \sum_{i_m=1}^{s} \frac{a_{ji_1} a_{i_1 i_2} \ldots a_{i_{m-1} i_m}}{p_j\, p_{ji_1} p_{i_1 i_2} \cdots p_{i_{m-1} i_m}} b_{i_m} p_j\, p_{ji_1} p_{i_1 i_2} \cdots p_{i_{m-1} i_m} \tag{2.916}$$

$$= \tilde{x}_j^{(m+1)}. \tag{2.917}$$

But this corresponds to the same formula that we have derived when computing the $m+1^{th}$ *approximation of the* j^{th} *component of the solution* $\mathbf{x}$ *of* $\mathbf{M}\mathbf{x} = \mathbf{b}$*, namely,*

$$\tilde{x}_j^{(m+1)} = E(\mathbf{X}_m). \tag{2.918}$$

Now, the SLLN then implies, that the solution of a linear system of equations can be computed via simulating a large number N *of independent and identically distributed random walks* $j^{(k)} = i_0^{(k)} \to i_1^{(k)} \to i_2^{(k)} \to \ldots \to i_m^{(k)}$ *of length* m *with* $1 \leq k \leq N$*, namely via computing the mean*

SLLN (216)

$$\begin{aligned} \tilde{x}_j^{(m+1)} &\approx \frac{1}{N}\sum_{k=1}^{N} \mathbf{X}_m^{(k)} && (2.919)\\ &= \frac{1}{N}\sum_{k=1}^{N}\left(\mathbf{X}_{ji_0}^{(k)} + \sum_{l=1}^{m} \mathbf{X}_{ji_1\ldots i_l}^{(k)}\right). && (2.920) \end{aligned}$$

Keep your eyes open for this technique; you will see it used more and more in Monte Carlo integration.

2.4.7.2 DISCRETE-TIME MARKOV PROCESSES

Due to the structure of their underlying state spaces, many processes encountered in everyday life cannot be represented via the model of the discrete Markov chain. Such processes require uncountable sets as state spaces, in particular, subsets of the s-dimensional Euclidian space $\mathbb{R}^s$. As an example, let us take a short outlook at *Monte Carlo path tracing*, a rendering method based on stochastic principles for computing an approximative solution to the SLTEV, the light transport equation in a vacuum.

Section 9.1

SLTEV (398)

EXAMPLE 2.99 (A First Short Look at pure-Monte Carlo Path Tracing, pMCPT) *Let us consider the Cornell box from Figure 2.65 consisting of slightly gloss surfaces, two specular spheres illuminated by a single area light source. Pure-Monte Carlo path tracing, for a detailed discussion see Section 9.1, then works as follows: Starting at the eye the algorithm shoots a ray randomly into the scene. At the hit point of the ray with any of the surfaces, pMCPT generates, depending on the material and the surface properties but also depending on the direction of the incoming ray—for glossy reflection, a new ray near the mirrored direction and for specular reflection a new ray in the mirrored direction—see Figure 2.66. The algorithm repeats this little random experiment until a light source is hit, the length of the path generated by the process exceeds a predefined value, or the ray leaves the scene.*

Glossy Reflection (304)

Since the generation of a new outgoing ray at gloss or specular surfaces also depends on the direction of the incoming ray, this process of path generation over the scene objects can not really be modeled by repeated evaluation of independent

Random Variable (168)

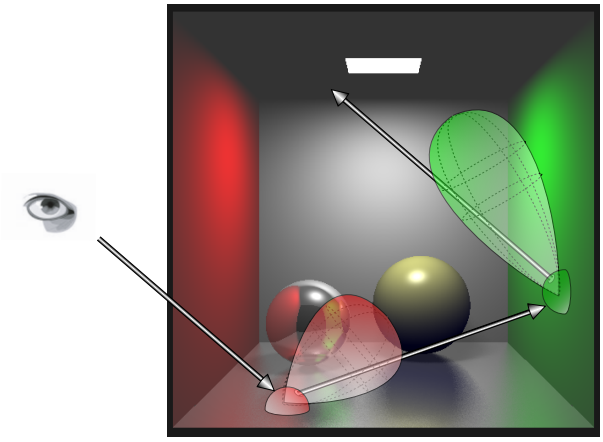


FIGURE 2.65: A FIRST SHORT LOOK AT PURE-MONTE CARLO PATH TRACING. At each hit point of a ray with an object of the scene, a random variable generates, depending on the material and the surface properties, but also depending on the direction of the incoming ray, a new ray. The algorithm then traces this ray recursively until a light source is hit, the current path length exceeds a predefined value, or the ray leaves the scene. Image courtesy of Zack Waters.

random variables with values from the measurable space $(S, \mathfrak{P}(S))$*, where* S *is the set of all existing surfaces within the scene. Since, the associated sample space is a continuous state space—obviously, the lobe of directions around the mirrored direction of the incident direction is uncountably infinite—a discrete-time Markov process can be used for sampling the outgoing direction.*

To express matters in a simplified manner, a *discrete Markov process* $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ can be defined as a discrete Markov chain over the continuous state space $S \subseteq \mathbb{R}^s$. In analogy to the results obtained in the preceding section, a Markov process may then also be represented via a random walk. However, in such a case the uncountability of the state space S requires a corresponding modification of the definitions and features describing a Markov process.

Measure Space (80)

DEFINITION 2.71 (Transition Kernel) *Let us assume,* $(\Omega, \mathfrak{F})$ *and* $(\Omega', \mathfrak{F}')$ *be two measure spaces. Then, we call a function* $\mathcal{K}$*, defined by*

$$\mathcal{K} : \Omega \times \mathfrak{F}' \longrightarrow \mathbb{R}^{\geq 0} \tag{2.921}$$

with

i) $\forall\, \omega \in \Omega, \quad \mathcal{K}(\omega, \cdot)$ *is a measure on* $(\Omega', \mathfrak{F}')$

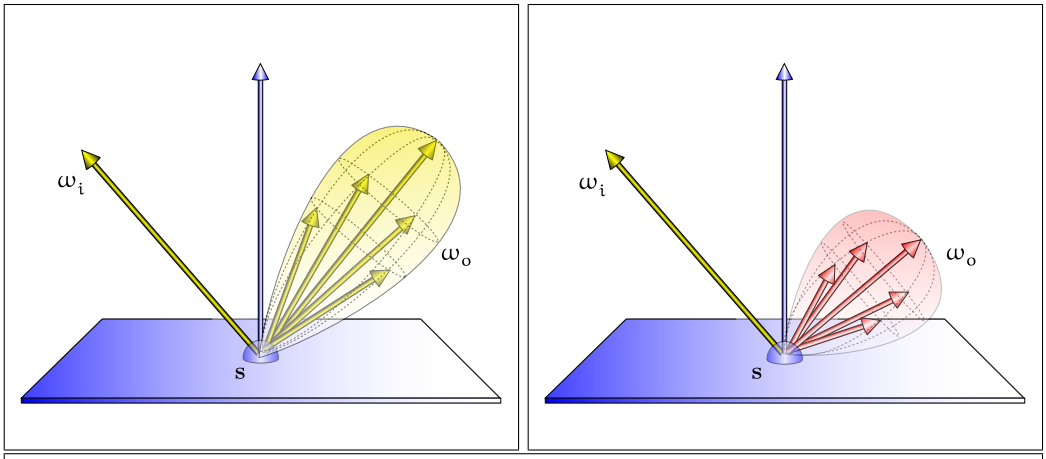


FIGURE 2.66: SAMPLING A DIRECTION IN MONTE CARLO PATH TRACING. At gloss surfaces, Monte Carlo path tracing samples the outgoing ray depending on the incident ray near the mirrored direction. At high gloss surfaces, the outgoing ray is rather chosen near the mirrored ray, for slightly gloss surfaces the lobe of possible exitant directions is larger, and the outgoing ray can leave the hit point further away from the mirrored direction of the incident ray.

ii) $\forall B \in \mathfrak{F}' \quad \mathcal{K}(\cdot, B)$ *is a* Ω-Ω'*-measurable function,*

a transition kernel. *In the case where* $\mathcal{K}(\omega, \cdot)$ *is a probability measure,* $\mathcal{K}$ *is also referred to as a* Markov Kernel *and one writes in the case of integrating a function* $f(\omega')$ *with respect to the measure* $\mathcal{K}(\omega, \cdot)$ *for* $B \in \mathfrak{F}'$ *simply*

$$\int_B f(\omega')\,\mathcal{K}(\omega, d\omega'). \tag{2.922}$$

Due to the above definition, a Markov kernel is a mapping that assigns any element of the base set Ω a measure and at the same time assigns any measurable set form $\mathfrak{F}$ a Ω-Ω'-measurable function. Thus, the Markov kernel $\mathcal{K}(\omega, A)$ corresponds to the conditional probability of a random variable $\mathbf{X}$, namely:

$$\mathcal{K}(\omega, B) \quad = \quad \mathbb{P}(\mathbf{X}_{n+1} \in B | \mathbf{X}_n = \mathbf{x}), \tag{2.923}$$

thus the probability that $\mathbf{X}_{n+1} \in B$ given $\mathbf{X}_n(\omega) = \mathbf{x}$.

In the case where Ω and Ω' are finite or countably infinite sets, the Markov kernel $\mathcal{K}(\omega, B)$ corresponds the transition matrix $\mathbf{M}$ of a discrete-time Markov chain since it

holds:

$$\begin{aligned}\mathcal{K}(\omega_i, B) &= \mathbb{P}(\mathbf{X}_{n+1} \in B | \mathbf{X}_n = \mathbf{x}_i) && (2.924)\\ &= \sum_{\{\omega_j \in B | \mathbf{X}(\omega_j) = \mathbf{x}_j\}} \mathbb{P}(\mathbf{X}_{n+1} = \mathbf{x}_j | \mathbf{X}_n = \mathbf{x}_i) && (2.925)\\ &= \sum_{\omega_j \in B} p_{ij}, && (2.926)\end{aligned}$$

where it holds $\sum_{\omega_j \in \Omega} p_{ij} = 1$. Obviously, $\mathcal{K}(\omega_i, \Omega)$ then represents the i-th row of a stochastic matrix $\mathbf{M}$ with coefficients $(p_{ij})_{1 \le i,j}$. M (229)

If Ω and Ω' are uncountably infinite sets, then the Markov kernel $\mathcal{K}(\mathbf{x}, B)$ corresponds to a conditional probability density $p_{\mathbf{X}_{n+1}|\mathbf{X}_n}(\mathbf{x}_{n+1}|\mathbf{x}_n)$ of the transition $\mathcal{K}(\mathbf{x}, B)$. This can be written as

$$\begin{aligned}\mathbb{P}(\mathbf{X}_{n+1} \in B | \mathbf{X}_n = \mathbf{x}_n) &= \int_B \mathcal{K}(\mathbf{x}_n, d\mathbf{x}_{n+1}) && (2.927)\\ &= \int_B p(\mathbf{x}_{n+1}|\mathbf{x}_n) d\mu(\mathbf{x}_{n+1}), && (2.928)\end{aligned}$$

thus, the probability to get from a particular $\mathbf{x}_n$ to B. Note: To simplify our equations, we do not mention the involved random variables in the index of the density function p.

DEFINITION 2.72 (Discrete Markov Process) *Let $\mathcal{K}$ be a Markov kernel constructed over the measure spaces $(\Omega, \mathfrak{F})$ and $(\Omega', \mathfrak{F}')$, then the stochastic process $(\mathbf{X}_n)_{n \in \mathbb{N}_0}$ defined over the probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ is referred to as a* discrete Markov process, *if it satisfies the* Markov property*:*

$$\begin{aligned}\mathbb{P}(\mathbf{X}_{n+1} \in B | \mathbf{x}_n, \mathbf{x}_{n-1}, \dots, \mathbf{x}_0) &= \mathbb{P}(\mathbf{X}_{n+1} \in B | \mathbf{x}_n)\\ &= \int_B \mathcal{K}(\mathbf{x}_n, d\mathbf{x}_{n+1}) && (2.929)\\ &= \int_B p(\mathbf{x}_{n+1}|\mathbf{x}_n) d\mu(\mathbf{x}_{n+1}), && (2.930)\end{aligned}$$

where $p(\mathbf{x}_{n+1}|\mathbf{x}_n)$ is a conditional probability density function.

As known from the last section, the construction of a Markov-process is also exhaustively defined by the Markov kernel, i.e. the probability of the transition from state $\mathbf{x}_n \in B_n$ into state $\mathbf{X}_{n+1} \in B_{n+1}$ and the initial distribution $\mathcal{K}(\mathbf{x}_0, d\mathbf{x}_1)$, see Figure 2.4.7.2. Mathematically, this can be expressed in terms of

$$\begin{aligned}&\mathbb{P}(\mathbf{X}_{n+1} \in B_{n+1}, \dots, \mathbf{X}_0 \in B_0)\\ &= \int_{B_{n+1}} \dots \int_{B_1} p(\mathbf{x}_0)\, \mathcal{K}(\mathbf{x}_0, d\mathbf{x}_1) \cdots \mathcal{K}(\mathbf{x}_{n-1}, d\mathbf{x}_n)\, \mathcal{K}(\mathbf{x}_n, d\mathbf{x}_{n+1}) && (2.931)\\ &= \int_{B_{n+1}} \dots \int_{B_1} p(\mathbf{x}_0)\, p(\mathbf{x}_1|\mathbf{x}_0) \cdots p(\mathbf{x}_{n+1}|\mathbf{x}_n) d\mu(\mathbf{x}_1) \dots d\mu(\mathbf{x}_{n+1}). && (2.932)\end{aligned}$$

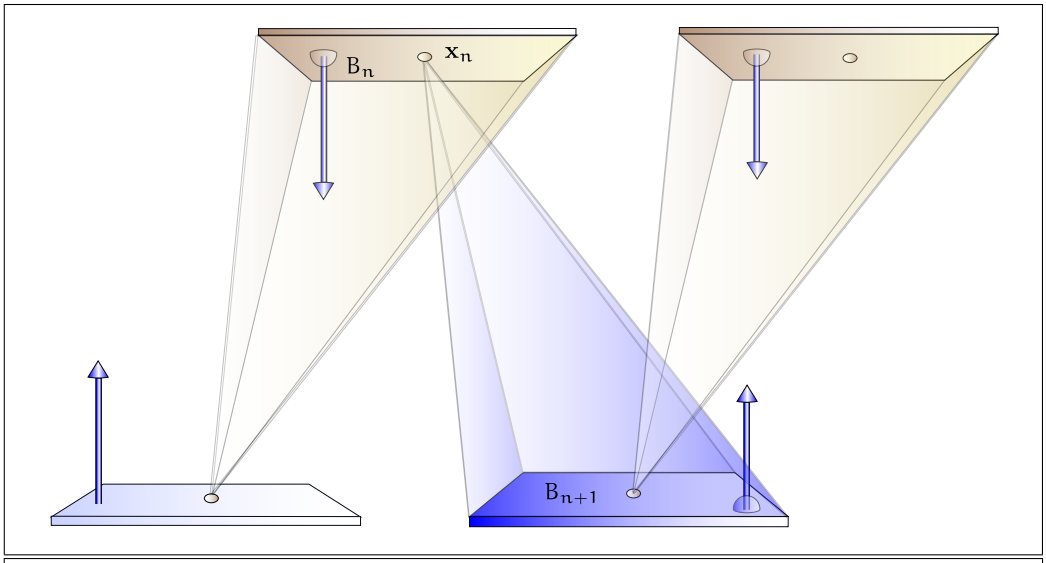


FIGURE 2.67: TRANSITIONS IN A MARKOV PROCESS. The representation of a possible transition from the continuous state space $\mathbf{B}_i$ into the set $\mathbf{B}_{i+1}$.

Due to Equation (2.931) we can define the *Chapman-Kolmogorov equations* for all $n, m \in \mathbb{N}$ with $B \in \mathfrak{F}'$ in the following form

$$\mathcal{K}^{m+n}(\mathbf{x}, B) \overset{def}{=} \int_{\Omega} \mathcal{K}^{n}(\mathbf{y}, B)\, \mathcal{K}^{m}(\mathbf{x}, d\mathbf{y}). \tag{2.933}$$

The Chapman-Kolmogorov equations state that a transition from $\mathbf{x}$ into the set B in $m+n$ steps has to follows in n steps after passing through any $\mathbf{y}$ in the m steps. To conclude this section, let us take a look at a simple example of a random walk generated over the model of the discrete Markov process.

EXAMPLE 2.100 (Continuous Travel of a Photon Through a Small Scene) *A beam of photons emitted from a light source is, depending on the material characteristics, either reflected, transmitted, or absorbed at the surfaces of the objects existing in a given scene. On its trip through the scene such a ray generates a path until it either leaves the environment or all the photons contained within the beam have been absorbed, see Figure 2.68.*

Such a path can be created via a discrete Markov process $(\mathbf{X}_n)_{n \in \mathbb{N}_0}$*, based on the Markov kernel,* $\mathcal{K}$*, given on* $(\partial\mathcal{V} \times \partial\mathcal{V}, \mathfrak{B}(\partial\mathcal{V} \times \partial\mathcal{V}))$ *by* $\partial\mathcal{V}$ (41)

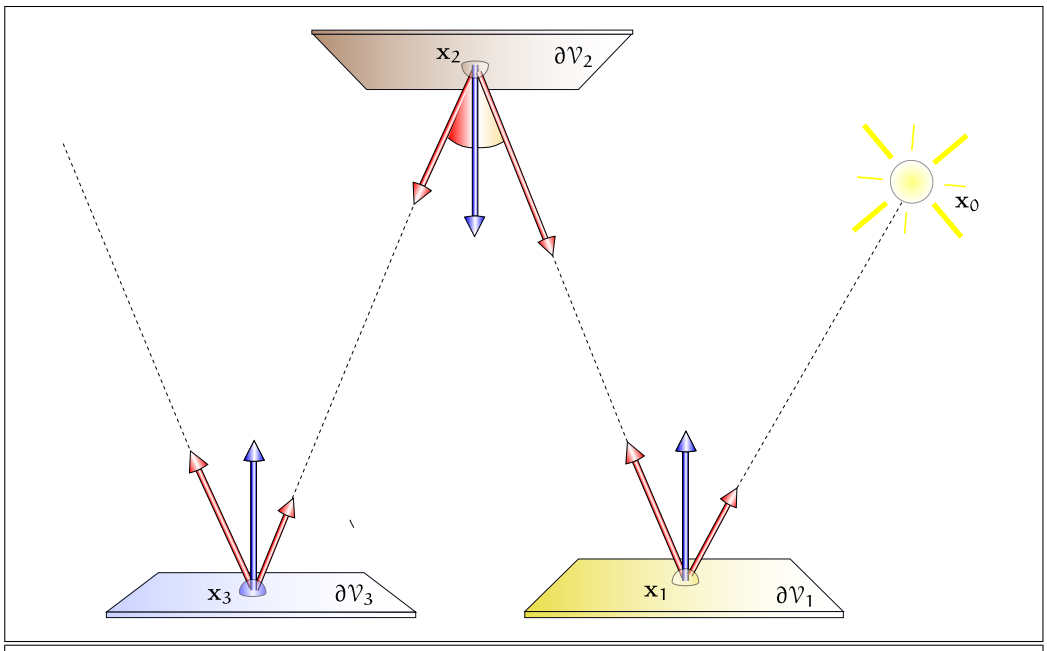


FIGURE 2.68: A CONTINUOUS RANDOM WALK ASSOCIATED WITH A MARKOV PROCESS. The state space is given by the set $\partial\mathcal{V} = \bigcup_{i=0}^{3} \partial\mathcal{V}_i$ of 2-dimensional surfaces within the scene. The random walk in the above figure starts at state $\mathbf{x}_0$ and goes over the surfaces $\partial\mathcal{V}_1, \partial\mathcal{V}_2$ and $\partial\mathcal{V}_3$.

$$\mathcal{K}(\mathbf{x}_n, B) = \int_B \mathcal{K}(\mathbf{x}_n, d\mathbf{x}_{n+1}) \tag{2.934}$$

$$= \int_B p(\mathbf{x}_{n+1}|\mathbf{x}_n) d\mu(\mathbf{x}_{n+1}), \tag{2.935}$$

with $B \in \mathfrak{B}(\partial\mathcal{V} \times \partial\mathcal{V}))$*, thus, the probability to arrive at* A *when starting in point* $\mathbf{x}$*.*

Assuming the probability of starting in point $\mathbf{x}_0$ *is given by* $p(\mathbf{x}_0)$ *and using the Chapman-Kolmogorov equations then implies that the path* $\mathbf{x}_0\mathbf{x}_1\mathbf{x}_2\mathbf{x}_3$*, constructed in Figure 2.68, requires the evaluation of the integral*

$$\int_{\partial\mathcal{V}_3} \int_{\partial\mathcal{V}_2} \int_{\partial\mathcal{V}_1} p(\mathbf{x}_0)\, \mathcal{K}(\mathbf{x}_0, d\mathbf{x}_1) \mathcal{K}(\mathbf{x}_1, d\mathbf{x}_2) \mathcal{K}(\mathbf{x}_2, d\mathbf{x}_3) \tag{2.936}$$

respectively

$$\int_{\partial\mathcal{V}_3} \int_{\partial\mathcal{V}_2} \int_{\partial\mathcal{V}_1} p(\mathbf{x}_0)\, p(\mathbf{x}_1|\mathbf{x}_0)\, p(\mathbf{x}_2|\mathbf{x}_1)\, p(\mathbf{x}_3|\mathbf{x}_2)\, d\mu(\mathbf{x}_3) d\mu(\mathbf{x}_2) d\mu(\mathbf{x}_1). \tag{2.937}$$

In Section 6.7.3, we will generalize this approach to an efficient solution method for Fredholm integral equations of the 2^{nd} *kind.*

2.5 REFERENCE LITERATURE AND FURTHER READING

Because the mathematics behind ray tracing and radiosity algorithms are the central theme of this book, we begin our literature survey of this chapter with a list of sources and text books, which were of great help for us when writing this and the following chapters.

The first chapter was written with the intention to cover all the mathematics for understanding the global illumination equations and already known solution approaches in a compact manner. Thus, we introduce the fundamental notions of functional analysis, as well as measure, integration and probability theory. The most important definitions and constructs to be presented here may be found in many textbook on the subject, where this chapter is mainly based on [213, Taylor & Lay 1986], [22, Berezansky & al. 1996], [123, Lebedev & al. 2003], [53, Edwards & al. 1965] and [239, Yoshida 1980]. Apart from these works, emphasizing the theoretical and mathematical nature of functional analysis and its historical development, [241, Zeidler 1995], [114, Kreyszig 1978], [176, Rynne & al. 2008] and here, inparticular, [169, Reddy 1998] served also as excellent reference books on functional analytic concepts, this time from a more practical point of view. They provide readers, who are unfamiliar or less familiar with the methods of functional analysis, with a succinct and yet comprehensive outline of this area of activity. With the help of practical examples, taken mostly from physics, the authors also demonstrate the fundamental importance of functional analytical concepts and methods in the search for particular solution methods for integral equations. Our constructs of the ray spaces $\mathcal{R}^{\partial\mathcal{V}}$ and $\mathcal{R}^{\mathcal{V}^o}$ are based on the ray space $\mathcal{R}$, introduced in [221, Veach 1998].

For our excourse into measure and integration theory [54, Elstrodt 1996] and [22, Berezansky & al. 1996] were very helpful. For the discussion of specific questions with respect to integration theory, we recommend [111, König 2000]. A brief but concise introduction to the Lebesgue integral of a less general nature is provided in [32, Chae 1995], [30, Burk 1998] and [31, Capinski & Kopp 2000]. [30, Burk 1998] also serves us as basis for the intuitive approach to the Lebesgue integral on $\mathbb{R}$.

A part of the presentation of the theoretical bases and the resulting numerical methods for the solution of integral equations can also be found in [161, Pipkin 1991], [72, Hackbusch 1995], [56, Engl 1997] and [113, Kress 1999]. A less theoretically oriented description of numerical solution procedures for integral equations written from the viewpoint of the user is found in [46, Drabek & Kufner 1996]. Short but excellent introductions to the problems of linear operators and integral equations including, in particular, tips and applications relating to the problems of global illumination are [7, Arvo 1991], [12, Arvo & al. 1994], [8, Arvo 1993]. Together with [68, Glassner 1995], consulted particularly for numerical methods for the solution of Fredholm integral equations, especially the light transport and the radiosity equation, these works formed the theoretical basis of our discussions.

The definitions of the most important concepts and methods of probability theory, which will be used repeatedly in the development and particularly in the analysis of

rendering methods based on Monte Carlo procedures can be looked foremost on the works [15, Ash & Doléans-Dade 2000], [86, Hoffmann-Jørgensen 1994], [214, Taylor 1997], [157, Pfanzagl 1991], [126, Mathar & Pfeifer 1990] and [31, Capiński & Kopp 2000]. [31, Capiński & Kopp 2000] is excellent in regard that it introduces the measure theoretical concepts in connection with their meaning in probabiltiy theory. All these books focus largely on the measurement theoretical approach to probability theory. In addition, [84, Hesse 2003], [132, Miller & Miller 1999], [162, Pitman 1999] and [171, Ross 2000] will provide us with a swift, more application-oriented introduction to probability calculation without touching upon measure theory. The discussions on the definitions of the probabilistic theoretical model of the Markov chain is based on [172, Rubinstein 1981] and [84, Hesse 2003], while the analogues relating to Markov processes is taken from [65, Gilks & 1996] and [170, Robert & Casella 1999]. More about the ergodic theory of Markov chains can be found in [130, Meyn & Tweedie 1993] or [204, Stroock 2005].

CHAPTER THREE

RADIOMETRY AND A LITTLE BIT OF PHOTOMETRY

Light is a form of radiation within the electromagnetic spectrum that can be transferred through space via emission, reflection, and absorption processes. As electromagnetic radiation, it can be interpreted both as a sinusoidal wave—consisting of an electrical, $\overrightarrow{\mathbf{E}}$, and a magnetic field component, $\overrightarrow{\mathbf{B}}$, perpendicular to each other and the propagation of the wave, $\overrightarrow{\mathbf{k}}$, see Figure 3.1—as well as a flow of particles, called *photons*, carrying a certain form of energy. This is known as *particle-wave duality*.

The electromagnetic spectrum, see Figure 3.1, shows the distribution of electromagnetic radiation within space. It is expressed in terms of frequency or wavelength and runs from meter-sized radio waves down to picometer scale γ-rays. In this spectrum, the visible light, perceived as different colors by our eyes, occupies only a small range, namely the range from [380 nm, 780 nm]. Since it varies from person to person, the spectral range of visible light cannot exactly be determined. Apart from the sun—it emits light waves composed by the superposition of wavelengths of the entire visible spectrum at roughly the same intensities—light can also be produced by many other natural or technical sources. In every case, electrons are excited via thermal, quantum, and other effects, leave their energy state and emit photons at certain wavelengths before they fallback to their normal states.

Due to its nature as wave, electromagnetic radiation can also interact among itself and form *interference* patterns. Additionally, it can also be *polarized*, that is, the coupled electric and magnetic field—usually perpendicular to each other and to the propagation direction of the wave—are correlated. This results in so-called polarization-effects seen as the blue color of the sky. Electromagnetic radiation can also be *coherent* in phase at all points in space and time if the wavefront of light stays. For a precise understanding and a detailed analysis of all these light phenomena the wave model of light, based on the *Maxwell* and *Kirchhof equations* is required [44, Ditchburn], [80, Hecht 2001]. However, in the following discussions we neglect all wave specific properties of light, and consider light in terms of photons, traveling along rays through space.

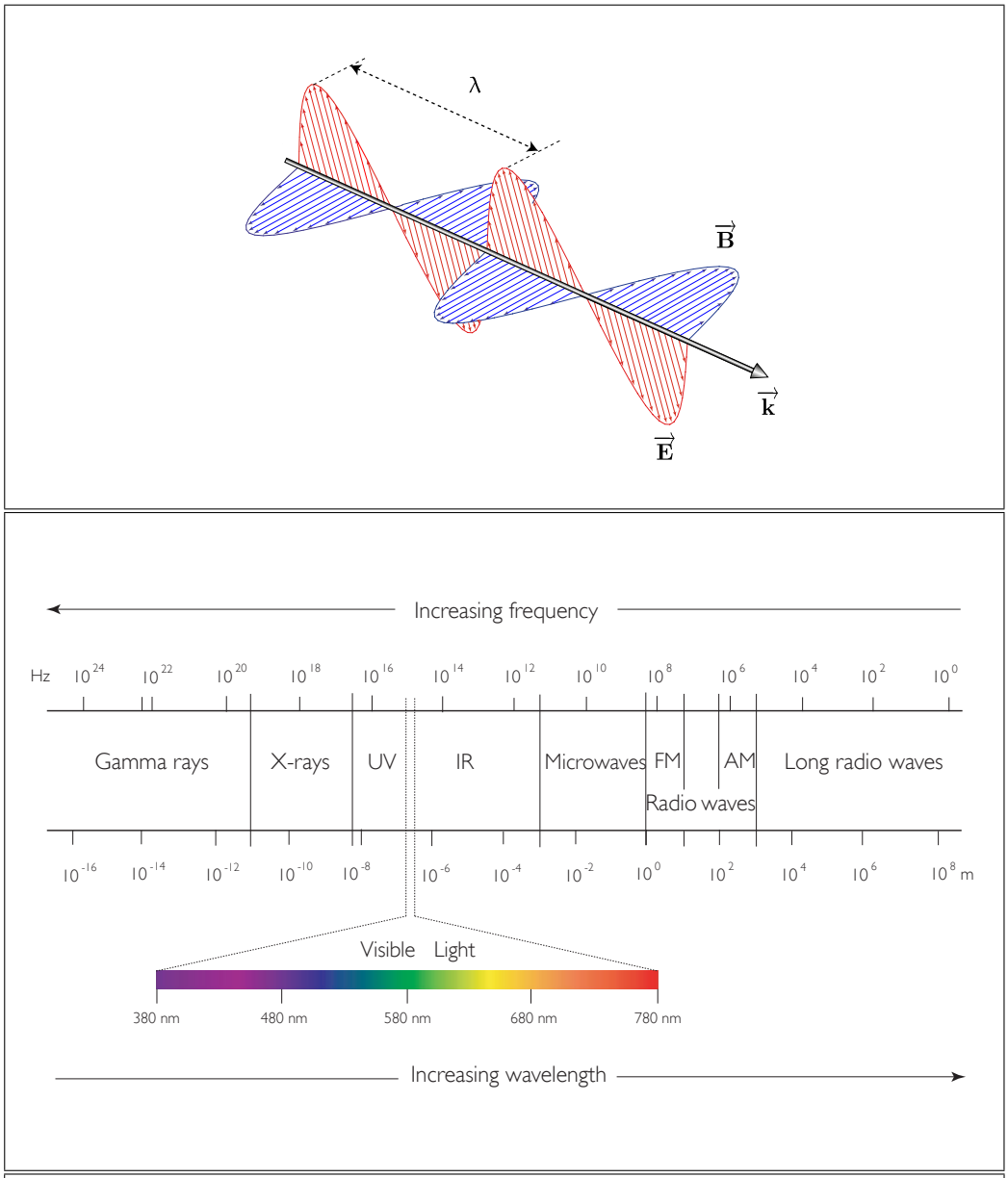


FIGURE 3.1: LIGHT AS AN ELECTROMAGNETIC WAVE AND AND THE ELECTROMAGNETIC SPECTRUM. The upper image shows the behavior of light as a form of electromagnetic radiation of wavelength λ, composed of electric and magnetic field components, $\vec{\mathbf{E}}$ and $\vec{\mathbf{B}}$, perpendicular to each other and to the direction of propagation $\vec{\mathbf{k}}$. The image below represents the electromagnetic spectrum it shows the distribution of types of radiation.

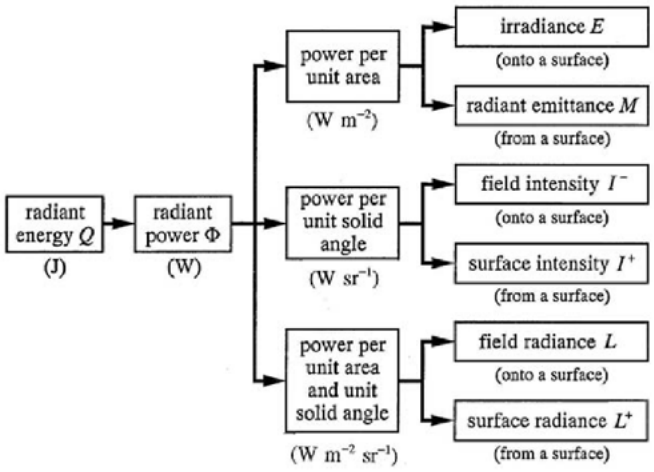


FIGURE 3.2: HIERARCHY OF RADIOMETRIC QUANTITIES. The image is a copy from the *Ocean Optics Book* by [28, Boss & all. 2011].

As our eyes are sensitive to light, all rendering techniques must simulate the physical concept of light and have to compute the distribution of photons in a scene to be rendered. That is, it is useful—apart from the well-know phenomena of light, such as reflection and refraction of a light beam at a surface—also to study the physical quantities that characterize light distribution at points and directions in a scene.

Radiometry provides us with this framework for analyzing and understanding the phenomena occurring when particles are transported through either a vacuum or a participating medium. In radiometry the concepts, terminology, and mathematical relationships are defined to describe and measure electromagnetic radiation and its interaction with the matter, see Figure 3.2. As it is the science that measures electromagnetic radiation, the most important quantities in this field play also a fundamental role for understanding the principles of realistic rendering.

Because we are concerned with light that is eventually perceived by human eyes, we should also account for quantities from the psycho-physical field of measuring the visual sensation in humans caused by electromagnetic radiation. The required visual response or perception by an observer can then be taken into account in a rendering procedure using the principles of photometry as a post-process such as *tone mapping*. This is done in *photometry*, the field of optics that deals with the quantification of the perception of light energy.

OVERVIEW OF THIS CHAPTER. In this chapter we introduce a series of fundamental *radiometric quantities*, which will serve as basic tools for deriving a mathematical formulation Section 3.1 Section 3.2

Section 3.3 of light transport in the next chapter. For that purpose, we introduce the theoretical
Section 3.4 constructs of the *particle space* and *particle space density*. Based on the physical con-
Section 3.5 cept of the *photon*, as a particle that carries energy, we then derive the radiometric basic
Section 3.6 quantities *radiant energy* and *radiant power*, as well as *radiance*, *irradiance*, *radiosity*, and *radiant intensity* that are used in the context of computer graphics. We conclude
Section 3.7 the chapter with some remarks about photometry.

3.1 ABSTRACT PARTICLES VS PHOTONS

Before we will focus our interest on the complex model of the *light quantum* we restrict our considerations to the simpler model of an *abstract particle*, where we assume that

i) all particles are so small and numerous that their statistical distribution can be treated as a continuum, and

ii) at any point in time a particle is completely characterized by its position $\mathbf{x}$, velocity v, and a few internal states, [10, Arvo 1995].

PARTICLE SPACE. Based on the above assumptions, we can now interpret a particle as an element of a set Ξ, given by:

$$\Xi \stackrel{\text{def}}{=} \mathbb{R}^3 \times S^2, \tag{3.1}$$

that is, particles are specified by its position in space and its direction of motion, where we have also assumed, that all particles move with constant velocity.

Measure (79) On Ξ, we can then define a measure, the so-called *particle space measure* ξ, based on
μ^3 (82) the Lebesgue measure μ^3 on $\mathbb{R}^3$ and the solid angle measure σ given over the unit sphere
σ (84) S^2 via the concept of the product measure by:

$$\xi \stackrel{\text{def}}{=} \mu^3 \times \sigma. \tag{3.2}$$

Measure Space (80) With these definitions, the triple $(\Xi, \mathfrak{B}(\Xi \times S^2), \xi)$ then declares a measure space, the so-called *particle space*.

COUNTING PARTICLES. Based on the concept of the particle space $(\Xi, \mathfrak{B}(\Xi \times S^2), \xi)$, now we want define radiometric quantities by counting the number of particles that are moving within a given subset from $\mathfrak{B}(\Xi \times S^2)$.

It is known from physics, that the concept of flow of any kind of particles is always specified in connection with a real or hypothetical surface. In detail, the flux of particles through any surface is defined as the number of particles flowing across a real or hypothetical surface or region in space per unit time.

As particles can cross such a surface at different points coming from different directions, let us firstly consider the flow of particles perpendicular through a differential surface $d\mu^2(\mathbf{x})$ around a point $\mathbf{x}$ in time $d\mu(t)$, as illustrated in Figure 3.3. Assuming all particles are moving with the same velocity—that is, same size and same direction—then, all these particles are contained within a tube with base area $\mu^2(\mathbf{x})$ and length $d\mu(s) = v\, d\mu(t)$, where v is the velocity of the particles, see the lower image in Figure 3.3. As all particles outside this tube are not fast enough to reach the surface in time $d\mu(t)$, the flow of particles across this infinitesimal surface patch can be computed by multiplying the volume $d\mu^2(\mathbf{x})\, d\mu(s) = d\mu^2(\mathbf{x})\, v\, d\mu(t)$ by a *particle space density* n.

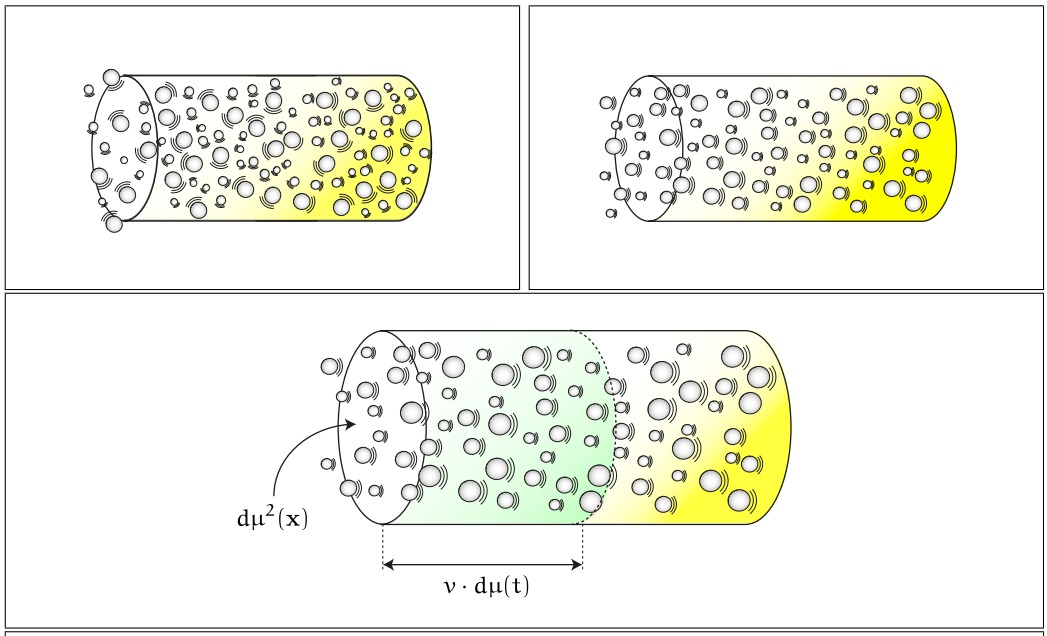


FIGURE 3.3: FLOW OF PARTICLES THROUGH A REAL OR HYPOTHETICAL SURFACE, I. The upper left image shows how particles cross a surfaces at different positions coming from different directions. We restrict our discussion to the flow of particles perpendicular to the surface area $d\mu^2(\mathbf{x})$, upper right image. The green part of the tube in the lower image contains all those particles that cross the surface $d\mu^2(\mathbf{x})$ in time interval $d\mu(t)$.

Assuming $n(\mathbf{x}, t)$ delivers the distribution of particles at point $\mathbf{x}$ at time t and all particles are moving perpendicular in direction towards $d\mu^2(\mathbf{x})$, then we count

$$n(\mathbf{x}, t) \underbrace{d\mu(s)}_{v\, d\mu(t)} d\mu^2(\mathbf{x}) = n(\mathbf{x}, t)\, v\, d\mu(t)\, d\mu^2(\mathbf{x}) \tag{3.3}$$

particles, that cross the infinitesimal patch $d\mu^2(\mathbf{x})$ in time $d\mu(t)$.

Now, the flow of particles across the base area must not necessarily be perpendicular. Particles can cross this surface also in directions, that are different from the direction of the surface normal $\mathbf{N}(\mathbf{x})$ at point $\mathbf{x}$. This observation then implies that the number of particles flowing through a surface patch is also dependent on the orientation of the surface with respect to their flow. Then, all these particles are contained in the tube built by the differential base area $\langle \mathbf{N}(\mathbf{x}), \omega \rangle d\mu^2(\mathbf{x})$ and its length $v\, d\mu(t)$, where the base area of this tube is foreshortened by the factor of $\cos\theta = \langle \mathbf{N}(\mathbf{x}), \omega \rangle$, see Figure 3.4. So, the number of particles, $\mathcal{N}$, crossing the differential surface $d\mu^2(\mathbf{x})$ is given by

$\langle \cdot, \cdot \rangle$ (845)

$$d^2\mathcal{N}(\mathbf{x}, t) = n(\mathbf{x}, t) \cos\theta \, d\mu^2(\mathbf{x}) \underbrace{d\mu(s)}_{v\, d\mu(t)} \tag{3.4}$$

$$= v\, n(\mathbf{x}, t)\, d\mu(t) \cos\theta \, d\mu^2(\mathbf{x}). \tag{3.5}$$

Let us now generalize our process of counting particles to account for also particles moving in different directions over the hemisphere. This implies that the particle space density function n must also be a function of the directional variable ω. In its functionality as a density function, $n(\mathbf{x}, \omega, t)$, measured in units $\frac{1}{m^3 \cdot sr \cdot s}$, delivers the number of particles per unit volume, per unit solid angle, per unit of time. Now, the particles that pass through the differential area $d\mu^2(\mathbf{x})$ in directions within a differential solid angle $d\sigma(\omega)$ around the surface normal $\mathbf{N}(\mathbf{x})$ in time $d\mu(t)$ are contained within the volume $\cos\theta \, d\mu^2(\mathbf{x})\, v\, d\mu(t) d\sigma(\omega)$. If $\mathbf{x}$ is a point within this differential volume, then the number of particles contained is given by

$$d^3\mathcal{N}(\mathbf{x}, t) = \underbrace{v\, n(\mathbf{x}, \omega, t)}_{\Phi_P(\mathbf{x}, \omega, t)} \cos\theta \, d\mu^2(\mathbf{x}) \, d\sigma(\omega) \, d\mu(t) \tag{3.6}$$

$$= \Phi_P(\mathbf{x}, \omega, t) \cos\theta \, d\mu^2(\mathbf{x}) \, d\sigma(\omega) \, d\mu(t), \tag{3.7}$$

where we call $\Phi_P(\mathbf{x}, \omega, t)$ the *particle flux* through point $\mathbf{x}$ in direction ω at time t.

PHOTONS. We no longer want to hold on the theoretical concept of the abstract particle. Multiplying our particles by the physical quantity of energy, we give them a physical meaning and can speak of *photons*, all equipped with a fixed velocity, namely the speed of light, denoted by c. Although a photon is coupled with both a frequency and a wavelength, it is not a wave in the sense of classical mechanics. Depending on measurements performed, a photon can be interpreted as both a particle or a wave. As it is sufficient for our discussion to envision that light consists of numerous localized packets of electromagnetic energy, we also abstract from the wave nature of light and only consider photons.

So, let us interpret a photon as a mathematical point carrying an amount of energy moving in space. The energy of a photon is, related to its frequency or wavelength, given

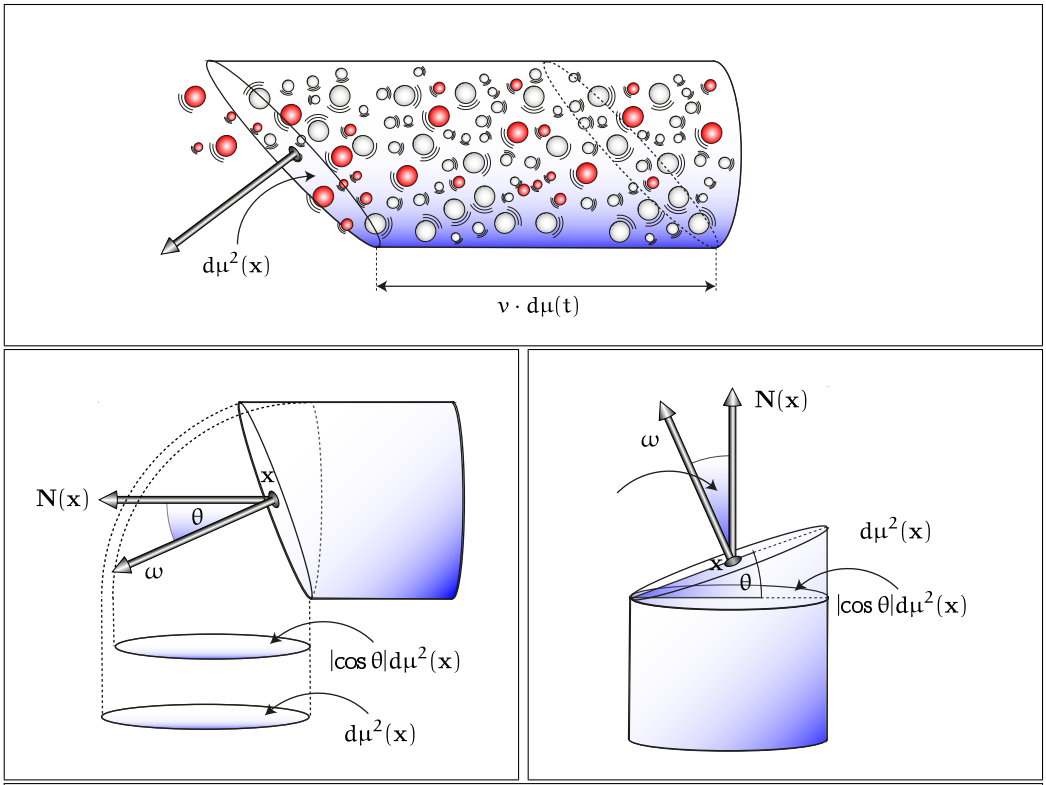


FIGURE 3.4: FLOW OF PARTICLES THROUGH A REAL OR HYPOTHETICAL SURFACE, II. The total number of particles flowing through the surface patch $d\mu^2(x)$ depends on both, the area of the surface and its orientation relative to the flow of the particles. Obviously, the flow is maximal, if the surface is perpendicular to the direction of flow. No particles flow across a surface when it is oriented parallel to the flow. The flow across a surface depends on the cosine between the surface normal and the direction of flow.

by *Einstein's relation*

$$E(\lambda) = h\nu \tag{3.8}$$

$$= \frac{hc}{\lambda}, \qquad [W \cdot s] \tag{3.9}$$

where $c = 2.99792458 \cdot 10^8 \frac{m}{s}$ denotes the velocity of light within free space, $\lambda \in [0, \infty)$ is the wavelength of a photon, $\nu \in [0, \infty)$ corresponds to its frequency, and $h = 6.625 \cdot 10^{-34}$ Js is *Planck's constant*.

REMARK 3.1 *It should be clear that Equation (3.8) implies an inverse relationship between the frequency of a photon and the wavelength of light. That is, light consisting*

of high energy photons has a low wavelength, while light consisting of low energy photons has a large wavelength.

As the speed of a photon compared with the valid parameters in our environments is infinitely large we restrict our interest to the *stationary* or *steady state distribution* of light quanta.

COUNTING OF PHOTONS. Assuming that the energy equilibrium in the environment is reached almost instantaneously, each volume in space then contains only a fixed number of photons per direction. That is, the particle space density can be considered as constant in time, which is why we can abstract from the temporal variable in the particle space density function $n(\mathbf{x}, \omega, t)$ and may write $n(\mathbf{x}, \omega)$ in the future. The number of photons within the differential volume formed by $d\mu^2(\mathbf{x})$, $d\sigma(\omega)$ and $c\, d\mu(t)$ is then given by

$$d^3\mathcal{N}(\mathbf{x}, \omega\, t) = \underbrace{c\, n(\mathbf{x}, \omega)\, \cos\theta d\mu^2(\mathbf{x})\, d\sigma(\omega)}_{d^2\Phi(\mathbf{x},\omega,t)}\, d\mu(t) \tag{3.10}$$

$$= d^2\Phi(\mathbf{x}, \omega, t)\, d\mu(t) \tag{3.11}$$

where $\Phi(\mathbf{x}, \omega, t)$ is called the *flux of photons* through point $\mathbf{x}$ in direction ω at time t.

3.2 RADIANT POWER

From the previous section it is known that the number of photons contained in a differential volume built by the differential base area $\cos\theta\, d\mu^2(\mathbf{x})$, the differential solid angle $d\sigma(\omega)$, and $c d\mu(t)$ is given by

$$d^3\mathcal{N}(\mathbf{x}, \omega\, t) = c\, n(\mathbf{x}, \omega)\, \cos\theta\, d\mu^2(\mathbf{x})\, d\sigma(\omega) d\mu(t). \tag{3.12}$$

Now, except of monochromatic light, light consist of photons of the different wavelength. This means, that we also have to account for the wavelength of photons in the particle density function n. For simplifying our discussion, here we want to make use of the *RGB-model*, known from computer graphics, that is, we differ between three types of photons: Photons from the red, green, and blue spectral band. This implies, that we have also to consider $d^3\mathcal{N}(\mathbf{x}, \omega\, t)$ for each spectral band. This then has the advantage, that we can assume that all photons contained in one of the volumes to be considered have the same wavelength or frequency. That is, multiplying the above expression by the energy $h\nu$ of a photon, we get the *radiant energy*, Q, carried by this volume, that is,

$$d^3Q(\mathbf{x}, \omega, t) = \underbrace{h\nu\, c\, n(\mathbf{x}, \omega)\, \cos\theta\, d\mu^2(\mathbf{x})\, d\sigma(\omega)}_{d^2\Phi(\mathbf{x},\omega)}\, d\mu(t). \quad [J] \tag{3.13}$$

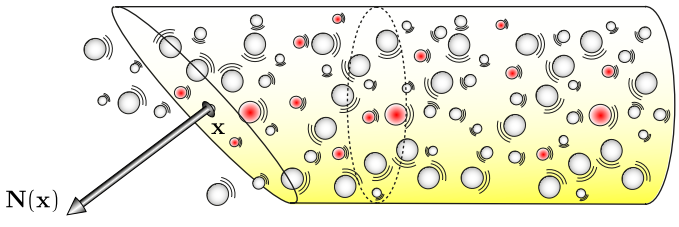


FIGURE 3.5: RADIANT POWER. Radiant Power is defined as the energy of particles passing a surface during a fix time interval. Many of the white particles bounce against the boundary of the tube, so, they don't pass through the whole within the tube.

We are now interested at the time-rate at which radiant energy passes through, emerges from, or hits a real or hypothetical surface. Since this flow of particles can vary over the surface and its direction, it is a function of position and direction. In radiometry this quantity is called *radiant energy* often also denoted as *flux*:

DEFINITION 3.1 (Radiant Power or Flux, Φ) *Let* $Q(\mathbf{x}, \omega, t)$ *be the energy carried by a flow of photons of the same frequency incident at, passing through, or emerging from point* $\mathbf{x}$ *on a real or hypothetical surface at time* t *in direction* ω*, see Figure 3.5. The* radiant power, Φ*, also called* radiant flux*, or simply denoted as* flux *is defined as the radiant energy incident at, passing through, or emerging from point* $\mathbf{x}$ *per unit of time, that is:*

$$\Phi(\mathbf{x}, \omega) = \frac{dQ(\mathbf{x}, \omega, t)}{d\mu(t)} \qquad [W] \equiv \left[\frac{J}{s}\right]. \tag{3.14}$$

REMARK 3.2 *Usually, radiant power is, like radiant energy, a time dependent quantity. As we are mostly concerned with systems in equilibrium, the particle density* N *does not change with time, that is, we omit the variable* t *in our formulas.*

3.3 RADIANCE

Based on the concept of radiant power we now define the most important radiometric quantity: *Radiance*. Due to the fact that the human eye is sensitive to radiance it is this quantity that any global illumination algorithm must compute.

Radiant Power (249)
Sensor Sensitivity (263)

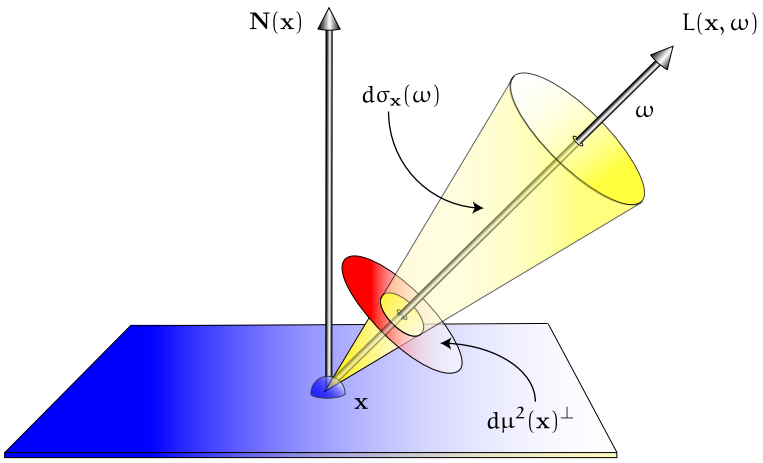


FIGURE 3.6: THE DEFINITION OF RADIANCE. The radiant power carried by photons incident at, passing through, or emerging from point $\mathbf{x}$ on a real or hypothetical surface in a given direction ω is denoted as radiance. Physically, radiance is defined as flux per unit solid angle $d\sigma_{\mathbf{x}}(\omega)$ per unit projected area $d\mu^2(\mathbf{x})^{\perp}$.

DEFINITION 3.2 (Radiance, L) *Let $\Phi(\mathbf{x}, \omega)$ be the radiant power carried by photons incident at, passing through, or emerging from point $\mathbf{x}$ on a real or hypothetical surface in direction ω. Then,* radiance, *denoted by* L, *is defined as the radiant power per unit projected area, per unit solid angle, that is,*

$$L(\mathbf{x}, \omega) \stackrel{\text{def}}{=} \frac{d^2\Phi(\mathbf{x}, \omega)}{d\mu^2(\mathbf{x})^{\perp}\, d\sigma_{\mathbf{x}}(\omega)} \tag{3.15}$$

$$\stackrel{(3.13)}{=} \hbar\nu\, c\, n(\mathbf{x}, \omega), \tag{3.16}$$

see Figure 3.6.

Obviously, radiance is a five-dimensional quantity that varies with position and direction and is measured in units of $\left[\frac{W}{m^2 \cdot sr}\right]$.

Since radiance is a function of position as well as a direction from a surface, it is important when speaking of radiance to specify the surface, the observation point, and the direction from it.

REMARK 3.3 *Relation (3.16) allows to represent the particle space density in terms of radiance as well as in terms or particle flux, that is, the number of photons at point*

$\mathbf{x}$ *flying in direction* ω*, thus:*

$$n(\mathbf{x},\omega) = \frac{1}{c}\frac{1}{\hbar\nu}L(\mathbf{x},\omega) \tag{3.17}$$

$$\overset{(3.10)}{=} \frac{1}{c}\Phi_P(\mathbf{x},\omega). \tag{3.18}$$

This relationship plays an important role, since it builds a bridge between the particle transport equation expressed in terms of number of particles, and the light transport equation based on radiance.

Particle Transport Equation (286)
Light Transport Equation (296)

REMARK 3.4 (Radiant Power.) *By integrating radiance over a finite surface area* A *and a finite solid angle* Γ *we can simply compute the radiant power incident at, passing through, or emerging from a real or hypothetical surface* A *in directions* $\omega \in \Gamma$*, namely:*

Φ (249)

$$\Phi \overset{(3.15)}{=} \int_A \int_\Gamma L_i(\mathbf{x},\omega)\, d\sigma^{\perp}_{\mathbf{x}}(\omega_i)\, d\mu^2(\mathbf{x}). \tag{3.19}$$

Note, we will use this relation very often in our following discussions, in particular if we choose Γ *as one of the hemispheres or the unit sphere.*

Depending on the direction of flow, we distinguish between *incident radiance*, L_i, and *exitant radiance*, L_o, where L_i denotes the radiance arriving at a surface and L_o expresses the radiance leaving a surface patch. From our discussion about incident and exitant ray functions it is known that the functions $L_i(\mathbf{x},\omega_i)$ and $L_o(\mathbf{x},\omega_o)$ are elements of the function space defined over $\mathbb{R}^3 \times S^2$. As L_i and L_o measure different photon events just before their arrival at and just after their departure from $\mathbf{x} \in \mathbb{R}^3$, we must also strictly distinguish between these two functions. Due to absorption, emission, and scattering, as well as reflection, or refraction processes at points $\mathbf{x} \in \mathbb{R}^3$, in participating media it generally holds:

Incident, Exitant Functions (48)

$$L_o(\mathbf{x},\omega_o) \neq L_i(\mathbf{x},\omega_i). \tag{3.20}$$

Only in the case where we consider light in a vacuum we have, as we will see in Theorem 3.3, the identity:

$$L_i(\mathbf{s}_i,\omega_i^i) = L_o(\mathbf{s}_j,\omega_o^j), \tag{3.21}$$

where $\mathbf{s}_i$ and $\mathbf{s}_j$ are two not occluded points on different surfaces and it holds: $\omega_i^i = -\omega_o^j$.

Let us now apply the concepts of radiant power and radiance as well as their relationship to an example from computer graphics. In this example we express the radiometric quantity radiance in terms of radiant power emitted from an area light source used in a conventional ray tracer.

Radiant Power (249)

EXAMPLE 3.1 (Radiant Power of an Area Light Source) *Let us recall Example 2.14 where we introduced the model of the area light source. Assuming, that an area light source emanates a number* $n(s_j, \omega_o^j) = n$ *photons, all with the same frequency* ν *in direction* ω_o *depending on angle* $\langle \mathbf{N}(s_j), \omega_o^j \rangle = |\cos\theta_o^j|$ *between the outgoing direction and the surface normal* $\mathbf{N}(s_j)$ *at point* s_j *on the light source, see the right images in Figure 3.7. The discussion from above then implies that the exitant radiance* $L_e(s_j, \omega_o^j)$ *can be written as:*

Radiance (250)

$$L_e(s_j, \omega_o^j) \overset{(3.15)}{=} C\, n(s_j, \omega_o^j)\, \langle \mathbf{N}(s_j), \omega_o^j \rangle \tag{3.22}$$

$$\overset{n(s_j,\omega_o^j)=n}{=} C\, n\, \langle \mathbf{N}(s_j), \omega_o^j \rangle, \tag{3.23}$$

with $C = c\hbar\nu$. *Due to Relation (3.19) this light source emits a flux of size:*

$$\Phi = C\, n \int_{\text{☼}} \int_{\mathcal{H}^2_o(s_j)} \langle \mathbf{N}(s_j), \omega_o^j \rangle \, d\sigma^{\perp}_{s_j}(\omega_o)\, d\mu^2(s_j) \tag{3.24}$$

$$= C\, n\, \mu^2(\text{☼}) \int_{\mathcal{H}^2_o(s_j)} \langle \mathbf{N}(s_j), \omega_o^j \rangle \, d\sigma^{\perp}_{s_j}(\omega_o^j) \tag{3.25}$$

$$\overset{|\cos\theta_o^j| = \langle \mathbf{N}(s_j), \omega_o^j \rangle}{=} C\, n\, A \int_0^{2\pi} \int_0^{\frac{\pi}{2}} \cos^2\theta_o^j \, \sin\theta_o^j \, d\theta_o^j \, d\phi_o^j \tag{3.26}$$

$$= C\, n\, A\, 2\pi \int_0^{\frac{\pi}{2}} \cos^2\theta_o^j \, \sin\theta_o^j \, d\theta_o^j \tag{3.27}$$

$$= \frac{2}{3} C\, n\, A\, \pi \tag{3.28}$$

where we use $\mu^2(\text{☼}) = A$.

Lambertian Emitter (349)

Assuming the light source is a diffuse emitter, such as a Lambertian emitter, *then we obtain with* $L_e(s_j, \omega_o^j) = C\, n$:

$$\Phi = C\, n \int_A \int_{\mathcal{H}^2_o(s_j)} d\sigma^{\perp}_{s_j}(\omega_o^j)\, d\mu^2(s_j) \tag{3.29}$$

$$\overset{(2.192)}{=} C\, n\, A \int_{[0,2\pi)} \int_{[0,\frac{\pi}{2}]} \left|\cos\theta_o^j\right| \, \sin\theta_o^j \, d\theta_o^j \, d\phi_o^j \tag{3.30}$$

$$= \underbrace{C\, n}_{L_e(s_j, \omega_o^j)} \; A\, \pi. \tag{3.31}$$

Radiant Flux (249)

From this result we can now conclude: If constant radiance emitting sources in a scene are specified via the exitant flux then the radiance $L_e(s_j, \omega_o^j)$ *emitted from a point* s_j *in direction* ω_o^j *can simply be determined by the quotient of the radiant power and the product of the area* A *of the surface as well as the constant* π, *thus:*

$$L_e(s_j, \omega_o^j) = \frac{\Phi}{A\, \pi}. \tag{3.32}$$

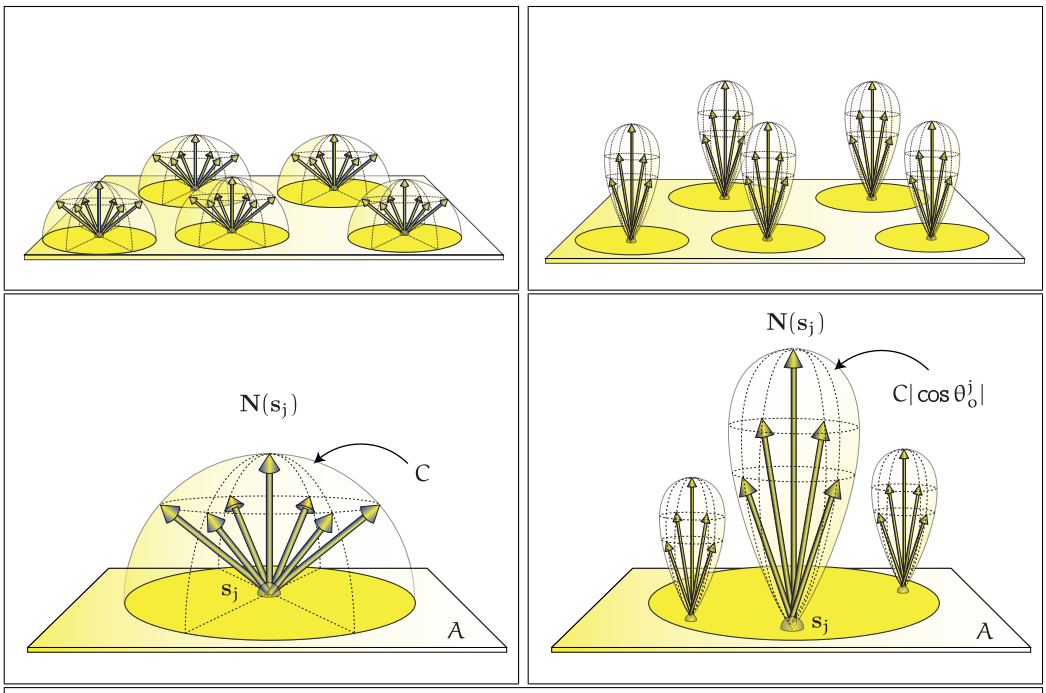


FIGURE 3.7: AREA LIGHT SOURCES. We can represent an area light source as a set of same type of point light sources, for example, as a set of Lambertian emitters, or as a set of point light sources that emit radiance according to a cosine term.

In Section 3.5, we define a new radiometric quantity, radiosity, B, *by the quotient*

$$\frac{d\Phi(\mathbf{x})}{d\mu^2(\mathbf{x})}, \tag{3.33}$$

where $\Phi(\mathbf{x})$ *is the flux exitant at point* $\mathbf{x}$ *on a real or hypothetical surface in all directions over the hemisphere* $\mathcal{H}^2_o$ *at* $\mathbf{x}$. *Based on this definition, radiance and radiosity can be used interchangeably for characterizing the light leaving diffuse surfaces.*

Radiosity (264)

RADIANCE INVARIANCE IN A VACUUM. *Radiance invariance* states that radiance in direction of a light ray remains constant if it propagates along the ray. This very important property of radiance is only valid in a vacuum, where no losses due to absorption and scattering, and no gains due to emission can occur. It is a consequence of the law of conservation of energy from physics, which says that no more energy can leave a point on a surface than arrives at this point. Mathematically, radiance invariance can be expressed as follows:

Conservation of Energy (332)

THEOREM 3.1 (Radiance Invariance) *Let* $\mathbf{s}_i$ *and* $\mathbf{s}_j$ *be two points on different surfaces*

within a scene, where $\mathbf{s}_i$ is visible from $\mathbf{s}_j$. Under vacuum conditions then it holds:

$$L_i(\mathbf{s}_i, \omega_i^i) = L_o(\gamma(\mathbf{s}_i, \omega_i^i), -\omega_i^i) = L_o(\mathbf{s}_j, \omega_o^j), \tag{3.34}$$

γ (47) *where $\mathbf{s}_j = \gamma(\mathbf{s}_i, \omega_i^i)$ and $\omega_i^i = -\omega_o^j$.*

PROOF 3.1 *Let us consider Figure 3.9, where the geometry of two differential surface patches around the points $\mathbf{s}_i$ and $\mathbf{s}_j$ is shown. Due to Definition 3.2, the differential*
Flux (249) *flux leaving patch $d\mu^2(\mathbf{s}_j)$, and arriving at a differential patch $d\mu^2(\mathbf{s}_i)$ is given by*

$$\begin{aligned} d^2\Phi_o &= L_o(\mathbf{s}_j, \omega_o^j)\, d\mu^2(\mathbf{s}_j)\, d\sigma^{\perp}_{\mathbf{s}_j}(\omega_o^j) && (3.35)\\ &= L_o(\mathbf{s}_j, \omega_o^j)\, d\mu^2(\mathbf{s}_j) \left|\cos\theta_o^j\right| d\sigma_{\mathbf{s}_j}(\omega_o^j), && (3.36) \end{aligned}$$

dσs (87) *where $d\sigma_{\mathbf{s}_j}(\omega_o^j)$ denotes the differential solid angle subtended by the differential surface patch $d\mu^2(\mathbf{s}_i)$ seen from point $\mathbf{s}_j$ in direction ω_o^j. In a similar way, the incident flux at point $\mathbf{s}_i$ can be expressed in terms of incident radiance $L_i(\mathbf{s}_i, \omega_i^i)$ at the differential surface $d\mu^2(\mathbf{s}_i)$ coming from directions of the differential solid angle $d\sigma_{\mathbf{s}_i}(\omega_i^i)$, that is,*

$$\begin{aligned} d^2\Phi_i &= L_i(\mathbf{s}_i, \omega_i^i)\, d\mu^2(\mathbf{s}_i)\, d\sigma^{\perp}_{\mathbf{s}_i}(\omega_i^i) && (3.37)\\ &= L_i(\mathbf{s}_i, \omega_i^i)\, d\mu^2(\mathbf{s}_i) \left|\cos\theta_i^i\right| d\sigma_{\mathbf{s}_i}(\omega_i^i). && (3.38) \end{aligned}$$

Transforming the differential solid angles $d\sigma_{\mathbf{s}_j}(\omega_o^j)$ and $d\sigma_{\mathbf{s}_i}(\omega_i^i)$ to the corresponding differential Lebesgue areas leads to:

$$d\sigma_{\mathbf{s}_j}(\omega_o^j) \overset{(2.196)}{=} \frac{d\mu^2(\mathbf{s}_i) \left|\cos\theta_i^i\right|}{\|\mathbf{s}_i - \mathbf{s}_j\|_2^2} \tag{3.39}$$

$$d\sigma_{\mathbf{s}_i}(\omega_i^i) \overset{(2.196)}{=} \frac{d\mu^2(\mathbf{s}_j) \left|\cos\theta_o^j\right|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2}. \tag{3.40}$$

Using these relations in the above expressions for the corresponding differential fluxes $d^2\Phi_o$ and $d^2\Phi_i$ then we get:

$$d^2\Phi_o = L_o(\mathbf{s}_j, \omega_o^j)\, d\mu^2(\mathbf{s}_j) \left|\cos\theta_o^j\right| \frac{d\mu^2(\mathbf{s}_i) \left|\cos\theta_i^i\right|}{\|\mathbf{s}_i - \mathbf{s}_j\|_2^2} \tag{3.41}$$

as well as

$$d^2\Phi_i = L_i(\mathbf{s}_i, \omega_i^i)\, d\mu^2(\mathbf{s}_i) \left|\cos\theta_i^i\right| \frac{d\mu^2(\mathbf{s}_j) \left|\cos\theta_o^j\right|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2}. \tag{3.42}$$

Since we consider the radiance transport under vacuum conditions, there is neither an energy loss due to absorption or out-scattering nor an energy gain due to

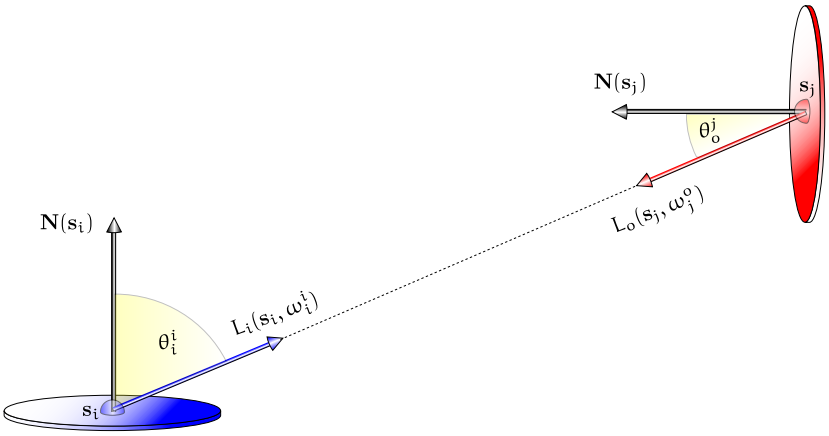


FIGURE 3.8: RADIANCE INVARIANCE IN A VACUUM. Radiance does not attenuate within a vacuum with distance. Since most sensors, such as the human eye or photographic cameras, are sensitive to radiance, the principle of radiance invariance explains why a given surface produces the same visual impression due to its emitted radiance at all viewing distances.

emission or in-scattering processes. According to the law of conservation of energy, all energy leaving patch $d\mu^2(\mathbf{s}_j)$ *in direction to* $d\mu^2(\mathbf{s}_i)$ *must also arrive at* $d\mu^2(\mathbf{s}_i)$*, that is,*

$$d^2\Phi_o = d^2\Phi_i \tag{3.43}$$

or equivalently

Conservation of Energy (332)

$$\frac{L_o(\mathbf{s}_j, \omega_o^j)\, d\mu^2(\mathbf{s}_j) \left|\cos\theta_o^j\right| \frac{d\mu^2(\mathbf{s}_i)\left|\cos\theta_i^i\right|}{\|\mathbf{s}_i - \mathbf{s}_j\|_2^2}}{L_i(\mathbf{s}_i, \omega_i^i)\, d\mu^2(\mathbf{s}_i) \left|\cos\theta_i^i\right| \frac{d\mu^2(\mathbf{s}_j)\left|\cos\theta_o^j\right|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2}} = 1,$$

thus

$$L_o(\mathbf{s}_j, \omega_o^j) = L_i(\mathbf{s}_i, \omega_i^i). \tag{3.44}$$

Obviously from the Theorem of Radiance Invariance, it follows: If the incident or exitant radiance at all surfaces within a scene is known, then the radiance distribution for the whole scene is also known. This property of radiance is one of the reasons why almost all global illumination algorithms work with radiance instead of flux, irradiance, or radiant intensity. Since radiance in a vacuum does not attenuate with distance, we can also conclude that the radiance transfer is:

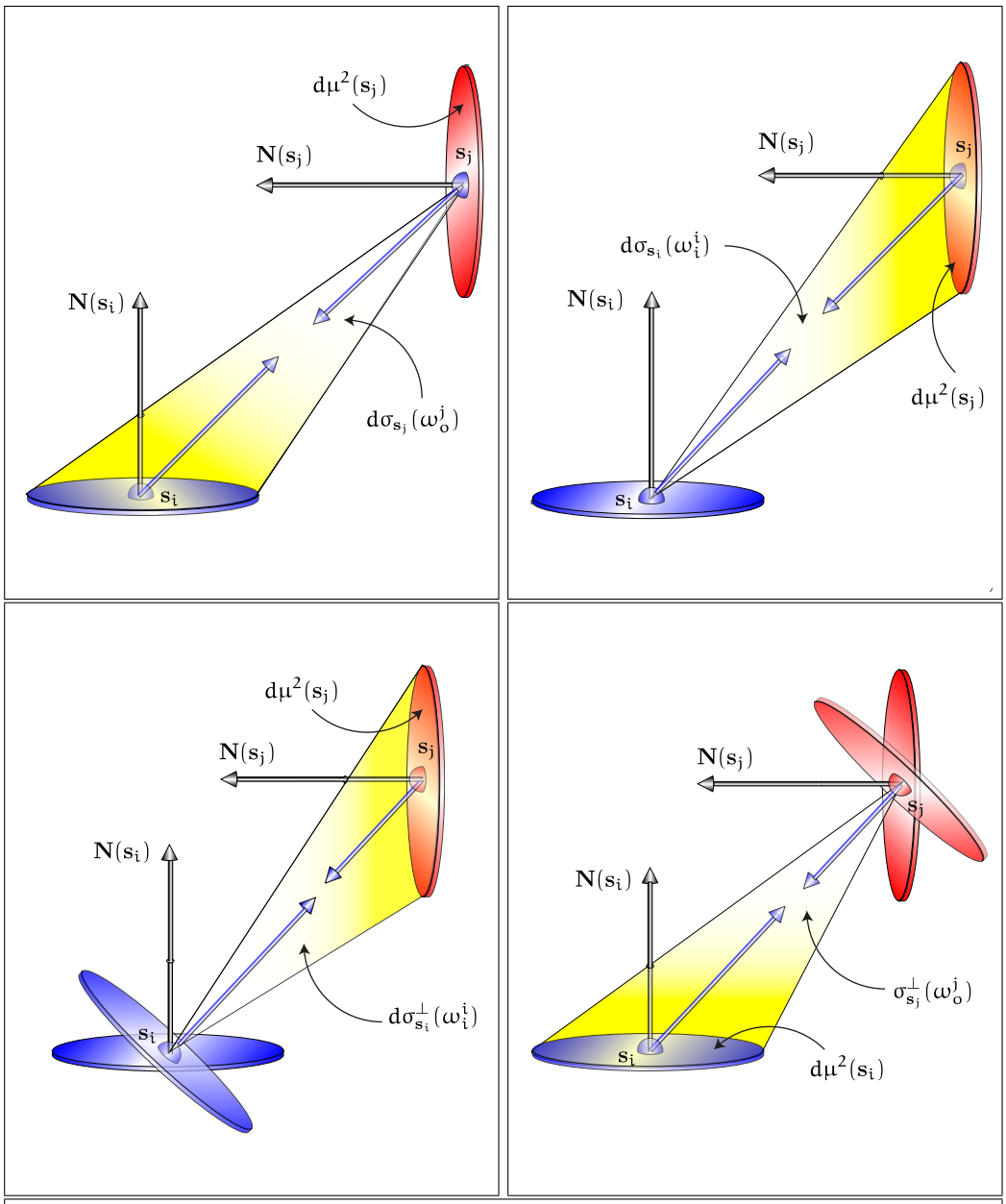


FIGURE 3.9: RADIANCE INVARIANCE IN A VACUUM. Radiance does not attenuate within a vacuum with distance. Since most sensors, such as the human eye or photographic cameras, are sensitive to radiance, the principle of radiance invariance explains why a given surface produces the same visual impression due to its emitted radiance at all viewing distances.

i) directly proportional to the radiance of the emitting surface,

ii) directly proportional to the surface areas of the emitter and the receiver,

iii) inversely proportional to the square of the distance between the emitter and the receiver, and

iv) dependent on the orientation of the surface normals with respect to the line connecting the two involved surface patches.

The radiance invariance principle plays the central role in rendering algorithms approximating the global illumination problem in a vacuum, since it guarantees the inversion of the optical path. Ray Tracing (664)

REMARK 3.5 *Usually, all radiometric quantities using radiance are spectral distributions, but for simplicity in CG, radiance is assumed to be a vector* $L = (R, G, B)$*, where* R, G*, and* B *are the intensities for the selected wavelengths of the red, green, and blue wavelength band. Since we use radiance as a scalar quantity in all of our equations, this means, that an equation using radiance can only be interpreted as it is valid for a single of the three bands of wavelength.*

3.4 IRRADIANCE

Based on the concept of radiant power we now define a further radiometric quantity: *Irradiance*. Each ray tracing based rendering algorithm must compute irradiance since it is the quantity that is measured within a pixel of the image plane. Radiant Power (249)

DEFINITION 3.3 (Irradiance, E) *Let* $\Phi_i(\mathbf{x})$ *be the radiant power from all directions over the hemisphere incident at point* $\mathbf{x}$ *on a real or hypothetical surface* A*, see Figure 3.10. Then,* irradiance, E*, is defined as:*

$$E(\mathbf{x}) \stackrel{def}{=} \frac{d\Phi_i(\mathbf{x})}{d\mu^2(\mathbf{x})}. \tag{3.45}$$

Defined as radiant power area density*, irradiance is a function of the spatial variable* $\mathbf{x}$ *and is measured in units of* $\left[\frac{W}{m^2}\right]$*. In the literature it is also known as* incident radiant power area density*, or* incident radiant flux area density*.*

Irradiance can easily be computed from incident radiance L_i by integration at point $\mathbf{x}$ about the hemisphere $\mathcal{H}^2_i(\mathbf{x})$, thus via: Radiance (250)

$$E(\mathbf{x}) = \int_{\mathcal{H}^2_i(\mathbf{x})} L_i(\mathbf{x}, \omega_i)\, d\sigma^{\perp}_{\mathbf{x}}(\omega_i). \tag{3.46}$$

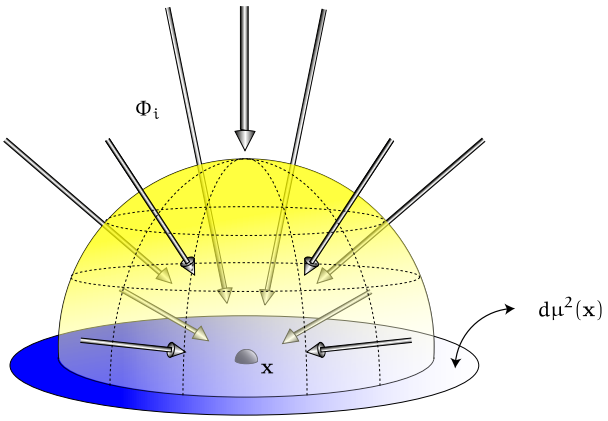


FIGURE 3.10: IRRADIANCE. Because the number of photons received at a single point is zero, we cannot talk about the amount of light received at a single point on a surface. Instead, we has to talk about the average number of photons in small infinitesimal intervals of space or time. This means, that radiometric quantities are expressed in terms of photon densities. Irradiance, defined as the spatial derivative of flux, is the amount of incident light at a surface point $\mathbf{x}$ per unit area.

This relation then implies that irradiance, expressed in terms of differential quantities, can also be treated as a directional quantity, namely:

$$dE(\mathbf{x}, \omega_i) \overset{(3.15)}{=} L_i(\mathbf{x}, \omega_i)\, d\sigma^{\perp}_{\mathbf{x}}(\omega_i), \tag{3.47}$$

where irradiance is interpreted as the radiance that would fall on a small bit of surface oriented facing the direction ω_i.

As already mentioned, irradiance is a one-sided, surface-oriented property, that is, a function of the position on a specific surface. Only in the case where irradiance is constant at all points, we can neglect from its spatial dependence. We can also talk about irradiance at a point $\mathbf{x}$ in space by specifying a surface normal $\mathbf{N}$ and defining the irradiance $E(\mathbf{x}, \mathbf{N}(\mathbf{x}))$ as the irradiance that falls on a small surface oriented facing the direction $\mathbf{N}$. As it is the radiometric quantity for describing radiation incident on a surface, the calculation of irradiance plays an important role in rendering procedures for shading a pixel.

LEMMA 3.1 (Lambert's Cosine Law) *Let us assume that a light beam arrives at point* $\mathbf{s}_i$ *at a flat surface* A_i *from direction* ω_i^i. Lambert's cosine law *states that the irradiance on the surface varies as the cosine between the incident light direction* ω_i^i *and the surface normal* $\mathbf{N}(\mathbf{s}_i)$ *at any point* $\mathbf{s}_i \in A_i$, *see Figure 3.11. Mathematically, this can*

be expressed as follows:

$$\begin{aligned} E(\mathbf{s}_i) &= \left|\langle \mathbf{N}(\mathbf{s}_i), \omega_i^i \rangle\right| E(\mathbf{s}_i^\perp) && (3.48)\\ &= \left|\cos\theta_i^i\right| E(\mathbf{s}_i^\perp), && (3.49)\end{aligned}$$

where $E(\mathbf{s}_i^\perp)$ *is the irradiance measured at the projection of point* $\mathbf{s}_i$ *on the projected surface patch* $A_i^\perp$ *of* A_i *in direction* ω_i^i*, thus the irradiance of the cross section of the light beam.*

PROOF 3.1 *Let us consider the flux* Φ *through the surface patch* $A_i^\perp$ *caused by the light beam, see Figure 3.11. Obviously, this flux is given by* Radiant Flux (249)

$$\Phi = E(\mathbf{s}_i^\perp)\, \mu^2(A_i^\perp). \tag{3.50}$$

Due to the principle of radiance invariance in free space, the same flux Φ *falls as* Φ_i *on the larger surface area* A_i*, producing an irradiance of size:* Radiance Invariance (253)

$$\begin{aligned} E(\mathbf{s}_i) &= \frac{\Phi}{\mu^2(A_i)} && (3.51)\\ &\overset{(3.50)}{=} \frac{E(\mathbf{s}_i^\perp)\, \mu^2(A_i^\perp)}{\mu^2(A_i)} && (3.52)\\ &= E(\mathbf{s}_i^\perp) \frac{\mu^2(A_i)\left|\cos\theta_i^i\right|}{\mu^2(A_i)} && (3.53)\\ &= \left|\cos\theta_i^i\right| E(\mathbf{s}_i^\perp). && (3.54)\end{aligned}$$

The cosine law states that the irradiance falling on any surface varies with the cosine of the incident angle. The perceived measurement is orthogonal to the incident flux and is reduced at oblique angles, which causes light to spread out over a wider area than it would be if the incident directions would be perpendicular to the measurement. Obviously, irradiance is proportional to the density of incident rays and inversely proportional to the distance between these rays.

EXAMPLE 3.2 (Irradiance on the Image Plane) *From the introductory chapter it is known that every ray tracing algorithm computes the projection of a scene onto the image plane of the involved camera system. Now, ray tracing, as a ray-based method, uses the radiometric concept of radiance while the quantity measured within a pixel of the image plane is irradiance. Then, the question arises: How is it possible to express irradiance to be measured via radiance carried by a ray?* Radiance (250)

For that purpose let us consider Figure 3.12 where a lens of diameter d *is located at a distance* f *from the image plane and at distance* z *from a scene object. Due to the principle of radiance invariance and the projection property of the lens, the flux incident at the pixel* $\square_i$ *with directions from the solid angle subtended by the lens* $\oplus$ Radiance Invariance (253) Radiant Power (249)

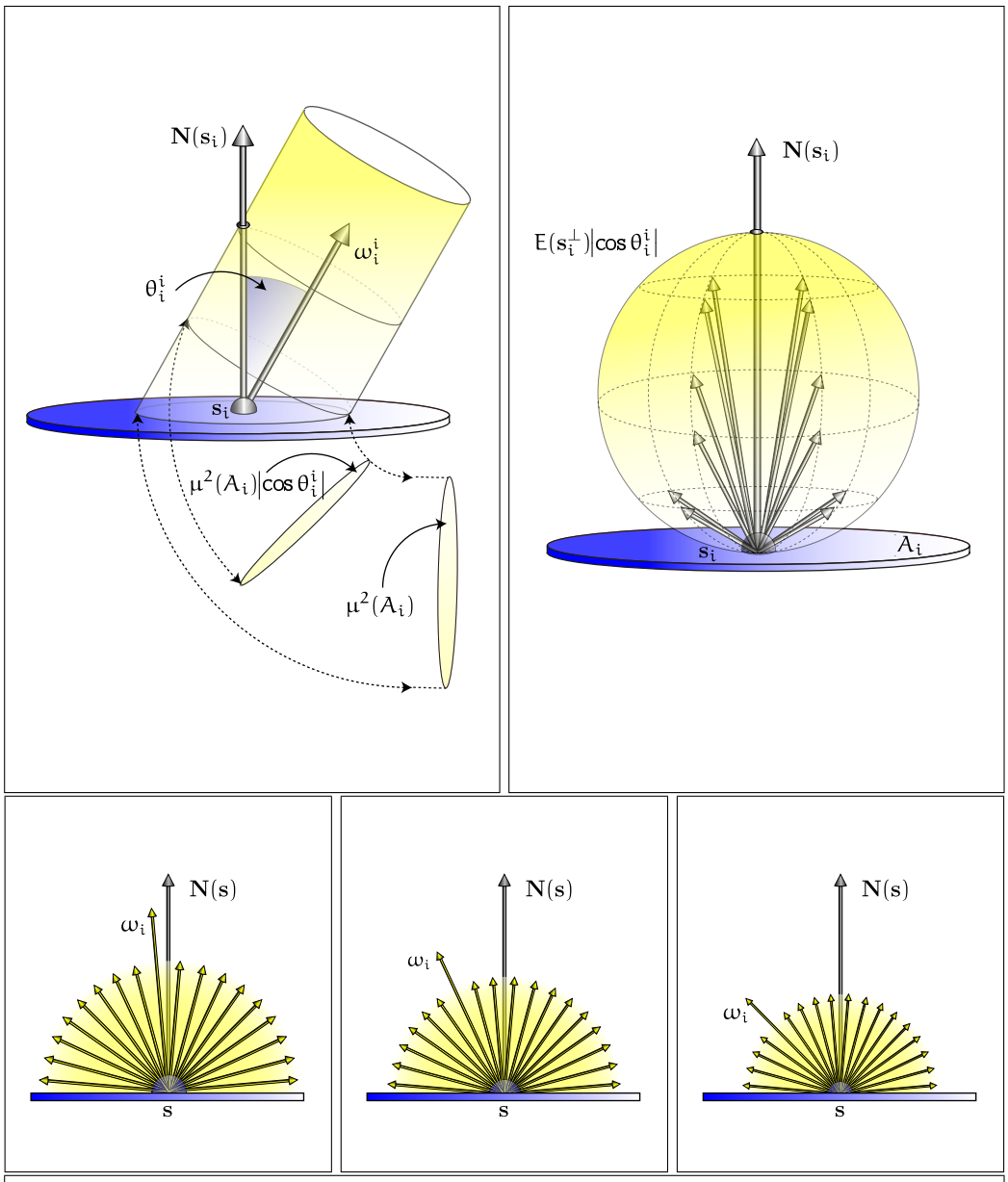


FIGURE 3.11: GEOMETRY FOR DERIVING LAMBERT'S COSINE LAW. Lambert's Cosine Law states that the irradiance on the surface varies as the cosine between the incident light direction ω_i^i and the surface normal $\mathbf{N}(s_i)$ at any point $s_i \in A_i$.
The three lower images show that the amount of reflected radiance at surface point s changes with the angle between the surface normal $\mathbf{N}(s)$ and the incident direction ω_i due to Lambert's Cosine Law.

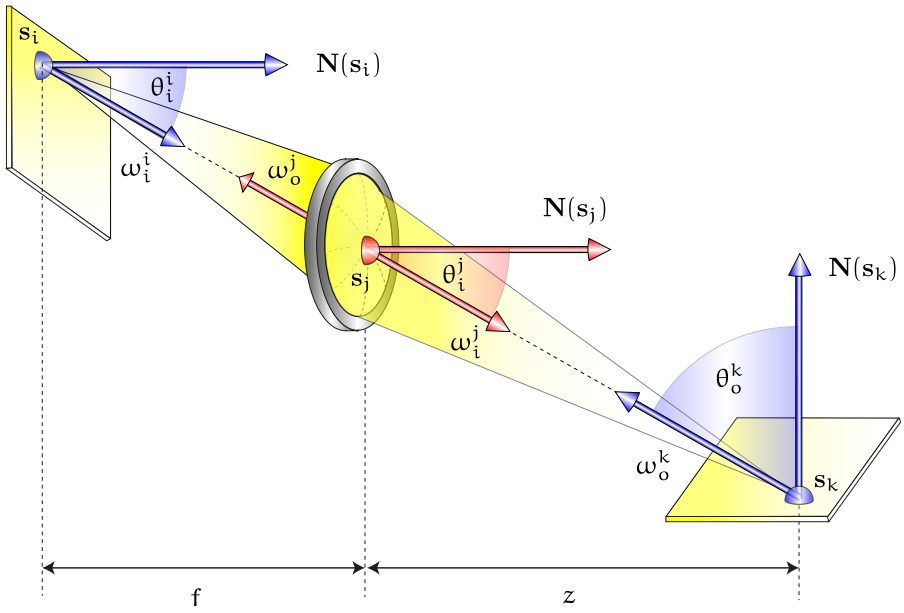


FIGURE 3.12: IRRADIANCE ON THE IMAGE PLANE. The irradiance measured at pixel $\square_i$ can be evaluated as the quotient of the incident flux at $\square_i$ and the area $\mu^2(\square_i)$.

as seen from the pixel is given by

$$\Phi_i = \int_{\square_i} \left(\int_{\oplus} L_i(s_i, \omega_i^i)\, d\sigma^{\perp}_{s_i}(\omega_i^i) \right) d\mu^2(s_i), \tag{3.55}$$

where $s_i \in \square_i$, $d\sigma^{\perp}_{s_i}$ *is the projected solid angle subtended by the lens,* $\oplus$, *as seen from* s_i. Projected Solid Angle (88)

Now, the optical properties of the lens—radiant power flows lossless through the lens—ensures, that the flux Φ_i *corresponds to the flux exitant from the projection* P_k *of the pixel onto a surface object and the solid angle subtended by the lens as seen from* $s_k \in P_k$, *that is,*

$$\Phi_i = \Phi_o \tag{3.56}$$

$$= \int_{P_k} \left(\int_{\oplus} L_o(s_k, \omega_o^k)\, d\sigma^{\perp}_{s_k}(\omega_o^k) \right) d\mu^2(s_k). \tag{3.57}$$

Under the assumption that the radiance distribution is constant in the scene,

the exitant radiance L_o *can be moved outside the integral and we get:*

$$\Phi_o = \int_{P_k} L_o \left(\int_{\oplus} d\sigma^{\perp}_{\mathbf{s}_k}(\omega_o^k) \right) d\mu^2(\mathbf{s}_k) \tag{3.58}$$

$$= \mu^2(P_k)\, L_o \int_{\oplus} d\sigma^{\perp}_{\mathbf{s}_k}(\omega_o^k). \tag{3.59}$$

Transforming the projected solid angle measure into the Lebesgue area measure, for details see Equation 2.199, leads to:

$$\Phi_o \overset{(2.186)}{=} \mu^2(P_k)\, L_o \int_{\oplus} \frac{\left|\cos\theta_i^j \cos\theta_o^k\right|}{\|\mathbf{s}_k - \mathbf{s}_j\|_2^2}\, d\mu^2(\mathbf{s}_j), \tag{3.60}$$

where the lens is denoted as $\oplus$.

Replacing the distance between the points $\mathbf{s}_j$ *and* $\mathbf{s}_k$ *via the definition of the cosine of angle* $\cos\theta_i^j$*, namely,*

$$\frac{z}{\cos\theta_i^j}, \tag{3.61}$$

then the integrand will not be dependent of the integration variable $\mathbf{s}_j$*, that is, it can be moved in front of the integral. This implies, that the exitant flux,* Φ_o*, can be written as:*

$$\Phi_o = \mu^2(P_k)\, L_o \frac{\left|\cos\theta_i^j \cos\theta_o^k\right|}{\left(\frac{z}{\cos\theta_i^j}\right)^2} \int_{P_j} d\mu^2(\mathbf{s}_j) \tag{3.62}$$

$$= \mu^2(P_k)\, L_o \frac{\left|\cos^3\theta_i^j \cos\theta_o^k\right|}{z^2} \int_{P_j} d\mu^2(\mathbf{s}_j) \tag{3.63}$$

$$= \frac{\pi}{4}\left(\frac{d}{z}\right)^2 A_k\, L_o \left|\cos^3\theta_i^j \cos\theta_o^k\right|, \tag{3.64}$$

where we have used $\mu^2(\oplus) = \frac{\pi d^2}{4}$ *and* $\mu^2(P_k) = \int_{P_k} d\mu^2(\mathbf{s}_k) = A_k$.

Due to Definition (3.3) and the condition that radiant power flows lossless through the lens, the irradiance to be measured at pixel $\square_i$ *can be written as:*

$$E(\mathbf{s}_i) = \frac{\Phi_i}{\mu^2(\square_i)} \approx \frac{\pi}{4} L_o \frac{A_k}{A_i}\left(\frac{d}{z}\right)^2 \left|\cos^3\theta_i^j \cos\theta_o^k\right|, \tag{3.65}$$

where we have assumed that it holds: $\mu^2(\square_i) = A_i$.

Let us now consider the solid angles subtended by the surface patch P_k *as well as the pixel* $\square_i$ *a little bit closer. As* P_k *corresponds to the projection of the pixel* $\square_i$

Solid Angle (83)

via the lens $\oplus$ *onto a region within the scene, the corresponding solid angles* $d\sigma_{\mathbf{s}_j}(\omega_i^j)$ *and* $d\sigma_{\mathbf{s}_j}(\omega_o^j)$ *are given by*

$$d\sigma_{\mathbf{s}_j}(\omega_i^j) = \frac{A_k \left|\cos\theta_o^k\right|}{\|\mathbf{s}_k - \mathbf{s}_j\|_2^2} \tag{3.66}$$

$$= \frac{A_k \left|\cos\theta_o^k\right|}{\left(\frac{z}{\cos\theta_i^j}\right)^2} \tag{3.67}$$

and

$$d\sigma_{\mathbf{s}_j}(\omega_o^j) = \frac{A_i \left|\cos\theta_i^j\right|}{\|\mathbf{s}_i - \mathbf{s}_j\|_2^2} \tag{3.68}$$

$$= \frac{A_i \left|\cos\theta_i^j\right|}{\left(\frac{f}{\cos\theta_i^j}\right)^2} \tag{3.69}$$

have the same size. That is, the term $A_k \left|\cos\theta_o^k\right|$ *can then be expressed by*

$$A_k \left|\cos\theta_o^k\right| = A_i \left|\cos\theta_i^j\right| \frac{\left(\frac{z}{\cos\theta_i^j}\right)^2}{\left(\frac{f}{\cos\theta_i^j}\right)^2} \tag{3.70}$$

$$= A_i \left|\cos\theta_i^j\right| \left(\frac{z}{f}\right)^2. \tag{3.71}$$

Using this relation in Equation (3.65), then the irradiance $E(\mathbf{s}_i)$ *corresponds to*

$$E(\mathbf{s}_i) = \frac{\pi}{4} L_o \left(\frac{d}{f}\right)^2 \cos^4\theta_i^j. \tag{3.72}$$

From this formula we conclude that the irradiance $E(\mathbf{s}_i)$ can be measured via the exitant radiance from P_k influenced by the geometry of the camera lens system, thus, the focal length f, the lens diameter d, and the off-axis angle $\cos\theta_i^j$. Radiance (250)

REMARK 3.6 (Sensors are Sensitive to Radiance) *The above example shows, that, the irradiance at a pixel is directly proportional to the radiance coming from object surfaces, even we say: The image irradiance is proportional to the scene radiance, where the factor of proportionality is given by the geometry of the optical device. Obviously, when passing through the lens radiance is transformed into irradiance. This proportionality, together with the property of radiance invariance, implies that radiance is the quantity that global illumination algorithms must compute and display to the observer.*

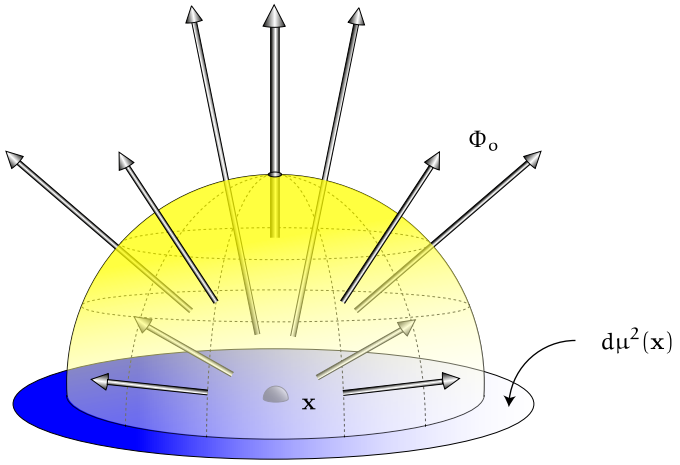


FIGURE 3.13: RADIOSITY. The total power leaving point $\mathbf{x}$ on a surface per unit area on the surface is called radiosity.

3.5 RADIOSITY

A radiometric concept similar to irradiance is *radiosity*. While irradiance is defined as the radiant power area density incident at a surface, radiosity is defined as the radiant power leaving a surface patch in all directions.

Irradiance (257)

Radiant Power (249)

DEFINITION 3.4 (Radiosity, B) *Let $\Phi_o(\mathbf{x})$ be the radiant power exitant at point $\mathbf{x}$ on a real or hypothetical surface A flowing in all directions over the hemisphere about $\mathbf{x}$, see Figure 3.13. Then,* radiosity, B, *is defined as:*

$$\mathrm{B}(\mathbf{x}) \stackrel{\mathrm{def}}{=} \frac{\mathrm{d}\Phi_o(\mathbf{x})}{\mathrm{d}\mu^2(\mathbf{x})}, \tag{3.73}$$

that is, radiosity is the exitant radiant power per unit area at a surface.

Radiosity is a function of the spatial variable $\mathbf{x}$ and is measured in units of $\left[\frac{W}{m^2}\right]$. In the literature it is also known as radiant power area density, *or* radiant exitance, *denoted by* M, *and it refers to the radiant flux leaving a surface patch from a point $\mathbf{x}$.*

Since radiosity is defined as the exitant radiant flux area density, radiosity can also be computed from radiance, namely via

Radiance (250)

$$B(\mathbf{x}) = \int_{\mathcal{H}^2_o(\mathbf{x})} L_o(\mathbf{x}, \omega_o)\, d\sigma^\perp_{\mathbf{x}}(\omega_o). \tag{3.74}$$

This relation then implies that radiosity expressed in terms of differential quantities can also be treated as a directional quantity:

$$dB(\mathbf{x}, \omega) \overset{(3.15)}{=} L_o(\mathbf{x}, \omega_o)\, d\sigma^\perp_{\mathbf{x}}(\omega_o). \tag{3.75}$$

EXAMPLE 3.3 *Let us consider a circular surface patch P_j, illuminated by a point light source located at point $\mathbf{s}_i$ that reflects the constant incoming radiance $L_i(\mathbf{s}_j\omega^j_i) = C$ depending on the cosine between the surface normal and the outgoing direction ω^j_o, see Figure 3.14. Then, the associated exitant radiance function is given by* $\|\cdot\|_2$ (861)

$$L_o(\mathbf{s}_j, \omega^j_o) = C\left|\cos\theta^j_o\right| \quad \forall \mathbf{s}_j \in P_j, \tag{3.76}$$

where $\cos\theta^j_o = \langle \mathbf{N}(\mathbf{s}_j), \omega^j_o\rangle$ is the angle between the normal at point $\mathbf{s}_j$ on the surface patch and the outgoing direction ω^j_o with $C = const.$ In this case, it holds for the radiosity at point $\mathbf{s}_j$ of this surface patch:

$$\begin{aligned}
B(\mathbf{s}_j) &= \int_{\mathcal{H}^2_o(\mathbf{s}_j)} L_o(\mathbf{s}_j, \omega^j_o)\, d\sigma^\perp_{\mathbf{s}_j}(\omega^j_o) && (3.77)\\
&= C\int_{\mathcal{H}^2_o(\mathbf{s}_j)} \left|\cos\theta^j_o\right|\, d\sigma^\perp_{\mathbf{s}_j}(\omega^j_o) && (3.78)\\
&\overset{(2.192)}{=} C\int_{[0,2\pi)}\int_{[0,\frac{\pi}{2}]} \cos^2\theta^j_o\, \sin\theta^j_o\, d\mu(\theta^j_o) d\mu(\phi^j_o) && (3.79)\\
&= C\,2\pi\frac{-\cos^3\theta^j_o}{3}\Big|_0^{\frac{\pi}{2}} = \frac{C\,2\pi}{3}. && (3.80)
\end{aligned}$$

With $C = 100\frac{W}{sr\cdot m^2}$ and $r = 0.1\,m$ then we obtain for the radiosity at point $\mathbf{s}_j$ of this surface:

$$B(\mathbf{s}_j) = \frac{200\,\pi}{3} \qquad \left[\frac{W}{m^2}\right]. \tag{3.81}$$

The power of the patch can now be computed by integrating the radiosity B over the entire patch, thus, Radiant Power (249)

$$\begin{aligned}
\Phi &= \int_{P_j} B(\mathbf{s}_j)\, d\mu^2(\mathbf{s}_j) && (3.82)\\
&= \frac{200\,\pi}{3}\frac{1}{10^2}\pi && (3.83)\\
&= \frac{2}{3}\pi^2 \qquad [W]. && (3.84)
\end{aligned}$$

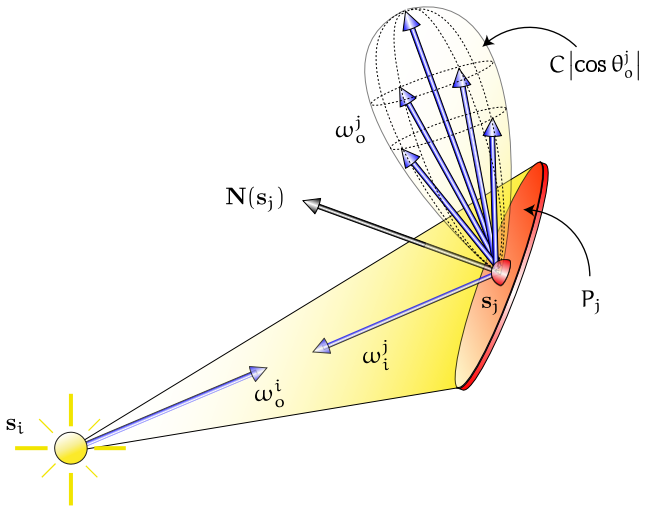


FIGURE 3.14: RADIOSITY. A point light source illuminates a surface patch P_j that reflects the incoming radiance depending on the cosine between the surface normal at point s_j and the outgoing direction ω_o^j.

EXAMPLE 3.4 (A First Approach for Deriving the Classical Radiosity Equation) *From our introductory chapter it is known that the stationary light transport equation in free space is a complicated integral equation. One idea for solving this equation is to simplify the underlying problem. By assuming that all object surfaces in a scene are Lambertian reflectors, the reflected radiance at all surfaces is constant in all directions. Due to Equation (3.32) from Example 3.1 then it holds:* SLTEV (399)

$$L_o(s_i, \omega_o^i) = \frac{\Phi}{A\pi} \overset{(3.73)}{=} \frac{B(s_i)}{\pi} \tag{3.85}$$

Lambertian Reflector (349) *for all points s_i on an object surface and all directions over the hemisphere about point s_i. Multiplying both sides of the SLTEV from Equation (1.4) by the constant π yields:*

$$\pi L_o(s_i, \omega_o^i) = \pi L_e(s_i, \omega_o^i) + \pi \int_{\mathcal{H}_i^2(s_i)} f_s(s_i, \omega_i^i \to \omega_o^i)\, L_i(s_i, \omega_i^i)\, d\sigma^{\perp}_{s_i}(\omega_i^i). \tag{3.86}$$

Radiance Invariance (253) *Since we consider the light transport in free space, the principle of radiance invariance holds and we can express the incident radiance $L_i(s_i, \omega_i^i)$ under the integral by the exitant radiance $L_o(s_j, \omega_o^j)$, where it holds $s_j = \gamma(s_i, \omega_i^i)$ and $\omega_o^j = -\omega_i^i$, so we*

get:

$$\pi L_o(s_i, \omega_o^i) = \pi L_e(s_i, \omega_o^i) + \pi \int_{\mathcal{H}_i^2(s)} f_s(s_i, \omega_i^i \to \omega_o^i)\, L_o(s_j, \omega_o^j)\, d\sigma_{s_i}^{\perp}(\omega_i^i). \quad (3.87)$$

The relationship between radiance and radiosity from Equation (3.85) then leads to:

$$B(s_i) = B_e(s_i) + \int_{\mathcal{H}_i^2(s)} f_s(s_i, \omega_i^i \to \omega_o^i)\, B(s_j, \omega_o^j)\, d\sigma_{s_i}^{\perp}(\omega_i^i). \quad (3.88)$$

This equation does not correspond to the classical radiosity integral equation *as it is known from computer graphics but it shows the idea behind the development of the radiosity integral equation. We show the exact derivation of the radiosity equation in Chapter 10.*

Radiosity Integral Equation (782)

REMARK 3.7 (Radiant Exitance) *Due to [190, Sillion & Puech 1994], the standard quantity used to characterize light sources is exitance. Similar to radiosity, see Equation (3.74),* radiant exitance *is also expressed as an integral over the hemisphere, namely the integral of the emitted radiance,*

$$M(x) = \int_{\mathcal{H}_o^2(x)} L_e(x, \omega_o)\, d\sigma_x^{\perp}(\omega_o), \quad (3.89)$$

thus, the portion of radiance due to internal emission.

3.6 RADIANT INTENSITY

Almost all radiometric quantities introduced in the last sections represent area densities, so, they can not be used to describe the radiant behavior of point light sources. To characterize the energy distribution of a point light source in a scene, we are forced to introduce a further radiometric quantity: *Radiant intensity.*

Point Light Source (50)

DEFINITION 3.5 (Radiant Intensity, I) *Let $\Phi(\omega)$ be the radiant power incident on, passing through, or emerging from a point on a real or hypothetical surface A in space into direction ω, see Figure 3.15. Then,* radiant intensity, *denoted by* I, *is defined as:*

Radiant Power (249)

$$I(\omega) \stackrel{def}{=} \frac{d\Phi(\omega)}{d\sigma_x(\omega)}, \quad (3.90)$$

that is, radiant intensity is radiant power per unit solid angle.

Radiant intensity is a function of the directional variable ω from or toward the point x for which it is defined and it is measured in units of $\left[\frac{W}{sr}\right]$. In the literature it is also known as radiant flux solid angle density.

Solid Angle (83)

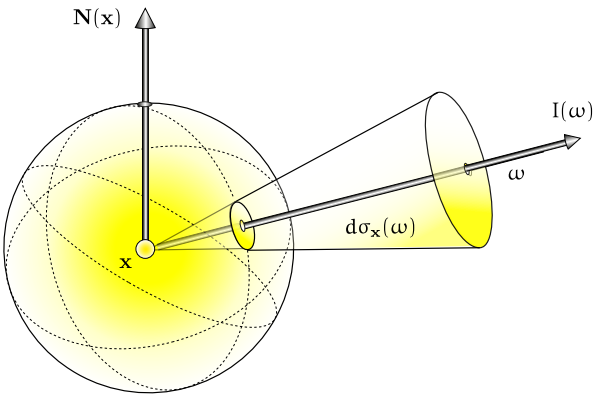


FIGURE 3.15: RADIANT INTENSITY. The radiant energy carried by photons leaving from, passing through or arriving at point $\mathbf{x}$ per unit solid angle $d\sigma_{\mathbf{x}}$ around a fixed direction ω per unit time is called radiant intensity.

Since it is defined as the radiant flux solid angle density, radiant intensity can also be computed from radiance, namely via:

Radiance (250)

$$I(\omega) = \int_A L(\mathbf{x}, \omega)\, d\mu^2(\mathbf{x}^\perp). \tag{3.91}$$

We have defined radiant intensity in terms of a point in space and a direction arriving at or leaving this point, that is, it is also important to say what the point is and at which direction we are interested in. Due to its definition as a solid angle density of radiant flux, radiant intensity is a useful concept for describing the radiant behavior of point light sources, or sources, that are very far away from the observer such as stars at the firmament.

Radiant Flux (249)

In the following two examples, we will apply the concept of radiant intensity to derive the *Inverse Square Law* and the formulation of radiance in terms of flux emanating from a point light source.

Isotropy (335)

EXAMPLE 3.5 (The Inverse Square Law) *Let us assume an* isotropic *point light source, that is, an emitter that radiates photons uniformly from a point* $\mathbf{x} \in \mathbb{R}^3$ *in all directions* ω_o*, thus* $I(\omega_o) = C$*. The flux emitted by this source is then given by*

$$\Phi_e = \int_{S^2(\mathbf{x})} I(\omega_o)\, d\sigma(\omega_o) \tag{3.92}$$

$$\overset{I(\omega_o)=C}{=} 4\pi C, \tag{3.93}$$

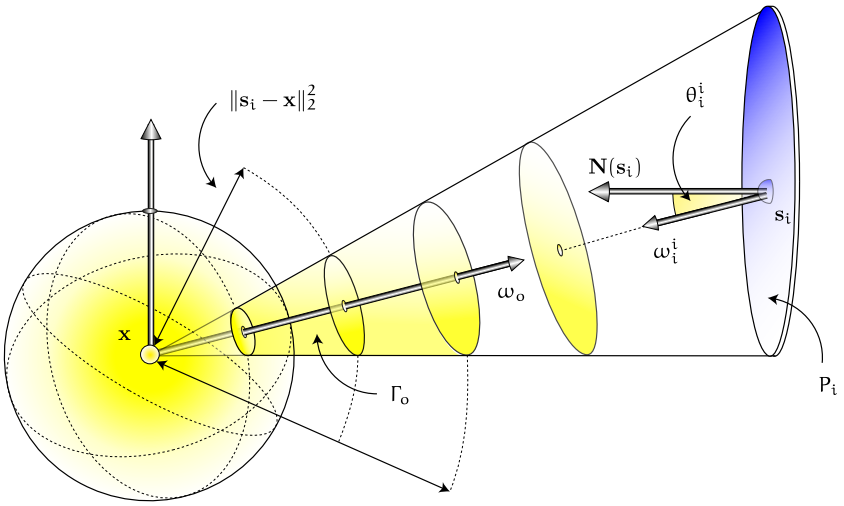


FIGURE 3.16: THE INVERSE SQUARE LAW. The irradiance on a surface patch decreases with the square of its distance to a point light source.

that is, our point light source has the radiant intensity

$$C = \frac{\Phi_e}{4\pi}. \tag{3.94}$$

As it is easily seen from Figure 3.16, the irradiance at point s_i on the surface patch P_i with Lebesgue measure A_i can be evaluated via the flux emitted from point x into solid angle Γ_o subtended by the patch P_i as seen from s_i. As the emitted flux Φ_e falls as incident flux Φ_i at the surface patch P_i, and the patch is small compared to its distance from x, we get:

Irradiance (257)

Radiant Flux (249)

$$E(s_i) \overset{(3.45)}{=} \frac{d\Phi_i(s_i, \omega_i)}{d\mu^2(P_i)} \tag{3.95}$$

$$\overset{(3.93)}{=} \frac{\int_{\Gamma_o} I(\omega_o)\, d\sigma(\omega_o)}{\mu^2(P_i)} \tag{3.96}$$

$$\overset{(3.94)}{=} \frac{\Phi_e}{4\pi} \frac{\int_{\Gamma_o} d\sigma(\omega_o)}{\mu^2(P_i)} \tag{3.97}$$

$$\overset{(2.196)}{=} \frac{\Phi_e}{4\pi} \frac{\int_{P_i} \frac{|\cos\theta_i^i|}{\|s_i - x\|_2^2}\, d\mu^2(s_i)}{\mu^2(P_i)} \tag{3.98}$$

$$= \frac{\Phi_e}{4\pi} \frac{|\cos\theta_i^i|}{\|s_i - x\|_2^2}, \tag{3.99}$$

where $\|\mathbf{s}_i - \mathbf{x}\|_2^2$ *is the distance of point* $\mathbf{x}$ *to point* $\mathbf{s}_i$ *on the surface patch* P_i *and* θ_i^i *is the angle between the surface normal at* $\mathbf{s}_i$ *and the direction* ω_i^i *towards the light source. Relation (3.99) is called the* inverse square law*,*

$$E(\mathbf{s}_i) = I(\omega)\frac{\left|\cos\theta_i^i\right|}{\|\mathbf{s}_i - \mathbf{x}\|_2^2}, \tag{3.100}$$

Irradiance (257) *since the irradiance on the surface patch decreases with the square of its distance to the point light source.*

The statement of the inverse square law should be intuitive since a surface that is close to a point light source will receive more photons per area than a surface that Solid Angle (83) *subtends the same solid angle but is further away from the light source, see Figure 3.16.*

EXAMPLE 3.6 (Radiance from a Point Light Source) *In classic ray tracing algorithms, the light sources illuminating a scene are assumed to be point light sources. In* Shadow Ray (14) *implementations of such algorithms shadow rays are fired in direction to the light* Radiance (250) *sources for estimating the direct illumination. Now, since rays carry radiance, but the intensity of point light sources is often given as the power of the source, it is* Radiant Power (249) *required to transform the quantity power to radiance.*

Irradiance (257) *From the example above it is known that irradiance at point* $\mathbf{s}_i$ *onto a small surface patch* P_i *is given by*

$$E(\mathbf{s}_i) = \frac{\Phi_e}{4\pi}\frac{\left|\cos\theta_i^i\right|}{\|\mathbf{s}_i - \mathbf{x}\|_2^2}, \tag{3.101}$$

where we assume that the point light source is located at $\mathbf{x} \in \mathbb{R}^3$ *and* θ_i^i *is the angle between the surface normal of* P_i *at* $\mathbf{s}_i$ *and the direction* ω_i^i *towards the light source.*

Radiant Flux (249) *To express the radiance in terms of flux, we can now utilize the concept of the* Dirac δ-distribution (118) *Dirac δ-distribution. Based on the Dirac δ-distribution, the radiance incident at point* $\mathbf{s}_i$ *from direction* $\omega_i^i = -\omega_o$ *can be expressed by*

$$\begin{aligned}
E(\mathbf{s}_i) &\overset{(3.99)}{=} \frac{\Phi_e}{4\pi}\frac{\left|\cos\theta_i^i\right|}{\|\mathbf{s}_i - \mathbf{x}\|_2^2} && (3.102)\\
&\overset{(2.302)}{=} \int_{\mathcal{H}_i^2(\mathbf{s}_i)} \frac{\Phi_e}{4\pi}\frac{|\cos\theta_i|}{\|\mathbf{s}_i - \mathbf{x}\|_2^2}\delta(\theta_i - \theta_i^i)\delta(\phi_i - \phi_i^i)\,d\mu(\theta_i)d\mu(\phi_i) && (3.103)\\
&\overset{(2.309)}{=} \int_{\mathcal{H}_i^2(\mathbf{s}_i)} \frac{\Phi_e}{4\pi\|\mathbf{s}_i - \mathbf{x}\|_2^2} && (3.104)\\
&\qquad \delta(\cos\theta_i - \cos\theta_i^i)\delta(\phi_i - \phi_i^i)\,|\cos\theta_i|\sin\theta_i\,d\mu(\theta_i)d\mu(\phi_i). && (3.105)
\end{aligned}$$

Radiometric Quantity	Symbol	Definition	Unit
Energy	Q	$-$	J
Radiant Flux	Φ	$\frac{dQ}{dt}$	W
Radiance	L	$\frac{d\Phi}{d\mu^2 d\sigma^\perp}$	$\frac{W}{m^2\,sr}$
Irradiance	E	$\frac{d\Phi}{d\mu^2}$	$\frac{W}{m^2}$
Radiosity	B	$\frac{d\Phi}{d\mu^2}$	$\frac{W}{m^2}$
Radiant Intensity	I	$\frac{d\Phi}{d\sigma}$	$\frac{W}{sr}$

TABLE 3.1: RADIOMETRIC DEFINITIONS.

Expressing the last equation in terms of incident directions leads to

$$E(s_i) = \int_{\mathcal{H}^2_i(s_i)} \underbrace{\frac{\Phi_e}{4\pi\|s_i - x\|_2^2}\,\delta_\sigma(\omega_i - \omega_i^i)}_{L_i(s_i,\omega_i^i)}\, d\sigma^\perp_{s_i}(\omega_i) \tag{3.106}$$

$$= \int_{\mathcal{H}^2_i(s_i)} L_i(s_i,\omega_i^i)\, d\sigma^\perp_{s_i}(\omega_i). \tag{3.107}$$

Due to Equation (3.46) we conclude, that the integrand in (3.106) can be interpreted as the incident radiance at point s_i. This means that the incident radiance L_i at point s_i coming from direction ω_i^i can be rephrased with the help of the Dirac δ-distribution as:

$$L_i(s_i,\omega_i^i) = \frac{\Phi_e}{4\pi\|x - s_i\|_2^2}\,\delta_\sigma(\omega_i - \omega_i^i) \tag{3.108}$$

$$\overset{(3.101)}{=} E(s_i)\;\delta_\sigma(\omega_i - \omega_i^i). \tag{3.109}$$

This relation is fundamental, as it expresses radiance in terms of irradiance of a light ray and thus builds a bridge between radiometry and geometric optics.

The radiometric definitions presented in this section are summarized in Table 3.1.

REMARK 3.8 *A great deal of confusion concerns the use and misuse of the term intensity. Some folks use it for $\frac{W}{sr}$, some use it for $\frac{W}{m^2}$ and others use it for $\frac{W}{m^2\cdot sr}$. It is quite clearly defined in the SI system, in the definition of the base unit of luminous intensity, the candela. Some attempt to justify alternate uses by adding adjectives like optical, used for $\frac{W}{m^2}$, or specific, used for $\frac{W}{m^2\cdot sr}$, but this practice only adds to*

the confusion. The underlying concept is quantity per unit solid angle [148, Palmer 1999].

3.7 A LITTLE BIT OF PHOTOMETRY

Photometry is the science of measuring visible light in units weighted according to the sensitivity of our visual system. Because the sensitivity of our eye varies with the wavelength, the perceived brightness of a monochromatic radiation at $550\,\mathrm{nm}$ is different from that of
Radiance (250) $700\,\mathrm{nm}$, even if the radiances are the same. Thus, a light source emitting a radiance of one watt per square meter per steradiant of green light, for example, appears much brighter than the same source emitting a radiance of one watt per square meter per steradiant of red or blue light.

Radiant Energy (248) Now, in photometry we do not measure watts of radiant energy, but rather it is attempted to determine the subjective impression caused by stimulating the human eye-
Radiant Power (249) brain visual system with radiant power. As the eye is a non linear detector of light, this task is very complicated. Light varies not only with wavelength, but also with the amount of radiant flux, whether the light is constant or flickering, the spatial complexity of the scene being perceived, the adaptation of the iris and retina, the psychological and physiological state of the observer, and many other variables, [13, Ashdown 1994]. That is, when we talk about brightness—which can be seen as a psycho-physical sensation—it is not sufficient only to consider the radiometric quantities radiance, radiosity, or power of a source.

The foundations of photometry were laid in 1729 by Pierre Bouguer, who discussed photometric principles in terms of the light source of his time: a wax candle. Thus, the wax candle also became the basis of the central concept of point light source in photometric theory. Today, the international standard is a theoretical point light source emitting a luminous intensity of one candela, that is, it emits monochromatic radiation with a frequency of $540\,\mathrm{THz}$ having a radiant intensity of $\frac{1}{683}\left[\frac{W}{\mathrm{sr}}\right]$, [48, Dutré 2003].

To characterize the average human visual response to light, a standardized spectral response function, the so-called *luminous efficiency function*, V, is used, see Figure 3.7. It was introduced in 1924 by the Commission Internationale de l'Eclairage, or CIE, after testing with over one hundred observers to visually match the brightness of monochromatic light sources with different wavelengths under certain conditions. As a statistical model of the human visual response to light, the luminous efficiency function shows the photopic luminous efficiency of the human visual system depending on the wavelength of light. It provides a weighting function that, together with the candela, can be used to convert radiometric quantities into their photometric analogues. That is, the only difference between radiometry and photometry lies in the units of measurement.

By defining a so-called *linear, radiometric-photometric transition operator* [28,
Linear Operator (53) Boss & al. 2011], with the conversion factor $\mathrm{K_m} = 683$, $\mathbf{Y}_{\mathrm{RP}}$, via

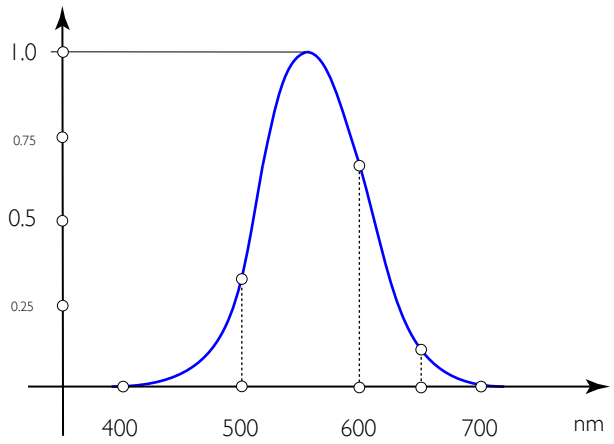


FIGURE 3.17: 1988 C.I.E. PHOTOPIC LUMINOUS EFFICIENCY FUNCTION V. The luminous efficiency function quantifies the sensitivity of the human visual system to all wavelengths of light, thus for example, a light source will appear brighter if it emits light of wavelength $600\,\mathrm{nm}$ than the same light source that emits light of $500\,\mathrm{nm}$ or $650\,\mathrm{nm}$ wavelength.

$$\left(\mathbf{Y}_{\mathrm{RP}} f\right)(\mathbf{x},\omega) \stackrel{\mathrm{def}}{=} K_m \int_{[0,\infty)} V(\lambda) f(\mathbf{x},\omega,\lambda)\, d\mu(\lambda), \qquad \left[\frac{\mathrm{lm}}{\mathrm{m}^2\cdot\mathrm{sr}}\right], \tag{3.110}$$

each radiometric quantity can then be transformed to its corresponding photometric quantity, that is, $\mathbf{Y}_{\mathrm{RP}}$ builds a bridge between radiometry and photometry.

EXAMPLE 3.7 (Luminance) Luminance *is the photometric equivalent of radiance. Using the radiometric-photometric transition operator* $\mathbf{Y}_{\mathrm{RP}}$*, luminance* $L_V(\mathbf{x},\omega)$ *can then be computed via:* Radiance (250)

$$\begin{aligned} L_V(\mathbf{x},\omega) &= \left(\mathbf{Y}_{\mathrm{RP}} L\right)(\mathbf{x},\omega) && (3.111)\\ &\stackrel{\mathrm{def}}{=} K_m \int_{[0,\infty)} V(\lambda)\, L(\mathbf{x},\omega,\lambda)\, d\mu(\lambda). && (3.112) \end{aligned}$$

Based on this formula, a light source emitting a constant spectral radiance $L(\mathbf{x},\omega,\lambda) = C \neq 0$ *only over a band of wavelength in the range between* $300\,\mathrm{nm}$ *and* $800\,\mathrm{nm}$ *has luminance of*

$$\begin{aligned} L_V(\mathbf{x},\omega) &\stackrel{\mathrm{def}}{=} C K_m \int_{[300\,\mathrm{nm},800\,\mathrm{nm})} V(\lambda)\, d\mu(\lambda) && (3.113)\\ &= C\cdot 683\cdot 107 = C\cdot 7.30\cdot 10^4 \frac{\mathrm{lm}}{\mathrm{m}^2\cdot\mathrm{sr}}, && (3.114) \end{aligned}$$

Photometric Quantity	Symbol	Definition	Unit
Luminous Energy	Q_V	–	talbot
Luminous Flux	Φ_V	$\frac{dQ_V}{dt}$	lm
Luminance	L_V	$\frac{d\Phi_V}{d\mu^2 d\sigma^\perp}$	$\frac{lm}{m^2\,sr}$
Illuminance	E_V	$\frac{d\Phi_V}{d\mu^2}$	$\frac{lm}{m^2}$
Luminous Exitance	B_V	$\frac{d\Phi_V}{d\mu^2}$	$\frac{lm}{m^2}$
Luminous Intensity	I_V	$\frac{d\Phi_V}{d\sigma}$	$\frac{lm}{sr}$

TABLE 3.2: PHOTOMETRIC DEFINITIONS.

where the maximum luminous efficacy, $K_m = 683\frac{lm}{W}$, *and the value of the integral is* 107 nm, *for more details see [28, Boss & al. 2011].*

REMARK 3.9 *As all rendering algorithms simulate the light transport in a scene, they also compute radiances of particular wavelengths at visible points. That is, for computing an image that is to be displayed on a device, it is not only necessary to map the radiance value into RGB-values. Here, we also have to account for the capability of a device to reproduce the correct perceptual response as well as the brightness present in the real scene.*

Radiance (250)

SPD (40)

In this book, we use radiometry exclusively. The visual response by an observer can then be added as a post-process, also called tone-mapping.

The photometric definitions presented in this section are summarized in Table 3.2.

3.8 REFERENCE LITERATURE AND FURTHER READING

Our construct of the particle space, introduced in Section 3.1, is based on the concept of phase space from [10, Arvo 1995]. He uses an approach based on a set of axioms that corresponds to the observable behavior of photons, and uses concepts from measure theory to derive some radiometric quantities. In [220, Veach 1997] this approach is extended to a more general class of radiometric quantities.

Pat Hanrahan's section *Transport Theory* in [36, Cohen & Wallace 1993] was helpful for us when counting particles and photons in Section 3.1. A similar, more intuitive

description of the principle of radiance invariance as we did it in Section 3.3 can be found in [190, Sillion & Puech 1994] and [50, Dutré & al. 2003].

Excellent and easily understandable introductions to radiometry can be found in [167, Preisendorfer 1958] as well as in the books by [127, McCluney 1994], [147, Palmer & Grant 2009], and [166, Preisendorfer 1976]. A detailed discussion on light and physically based lighting and shading models with respect to the concerns of CG is contained in [37, Comninos 2006]. Many other books about global illumination, such as [36, Cohen & Wallace 1993], [190, Sillion & Puech 1994], [13, Ashdown 1994], [68, Glassner 1995], [95, Jensen 2001], [50, Dutré & al. 2003], [187, Shirley & Morley 2003], [51, Dutré & al. 2006], and [158, Pharr & Humphreys 2004], [159, Pharr & Humphreys 2010] address the most important radiometric quantities briefly in a useful and easily understandable way. Apart from [191, Slusallek 1995] also [48, Dutré & al. 2003] and [1, Akinene-Möller & al. 2008] are other good resources on the topics of radiometry and in particular on photometry. Nice, brief and easily understandable overviews of radiometric quantities and units that are used in this section can also be found in [148, Palmer 1999] and in [48, Dutré & al. 2003]. A number of graphics paper available online provide also good coverage of some of the material presented here, for example Steve Marschner's lecture notes on radiometry [125, Marschner 2009], [61, Fleet & Hertzmann 2005], and [168, Preisendorfer & Tyler 1958].

In [165, Preisendorfer 1965] an attempt is made for a strict axiomatic formulation of radiometric quantities based on the wave nature of light and in [136, Nicodemus 1963] the concept of *basic radiance* is introduced, useful when deriving symmetric BSDFs.

CHAPTER FOUR

MATHEMATICAL FORMULATIONS OF STATIONARY LIGHT TRANSPORT

The goal of this chapter is to derive generally valid equations that describe the transport of light particles through participating media or a vacuum. Now, a mathematical description of the transport of light through a scene, where all properties of light is accounted for, is, without any restrictions to the underlying particle model, not really possible. So, we will firstly consider the transport of abstract particles through a scene with participating media. This requires the characterization of the most important properties of particles that can affect their motion through a medium or a vacuum, that is, we are forced to consider all those effects which imply changes in the distribution of particles in a scene. For that, we assume that all these particles subject a series of limitations and can be modeled via elements of the so-called *particle space*. Our particles are moving collision-free at a constant velocity and can be described at each moment by their current position as well as their direction of motion. We also assume that they do not possess any internal states, such as polarization, frequency, charge, or spin. Nevertheless, the resulting *particle transport equation* is an extremely complicated *integro-differential equation* that is not really usable for the field of realistic rendering. That is, it is not sufficient, only to restrict our discussion with respect to the properties of particles, but we also have to restrict our consideration to the behavior of particles at participating media or when interacting with object surfaces in a vacuum. Hence, we will approximate the scattering behavior at matter particles within participating media and the reflection or refraction behavior at object surfaces by so-called *bidirectional scattering* and *reflectance distribution functions*. Assigning the abstract particles photon character, the resulting particle transport equation is transformed—using the radiometric quantities derived in the previous chapter—into a mathematical formulation that ultimately explains the photon transport.

OVERVIEW OF THIS CHAPTER. Based on the measure theoretic concept of the particle space we will describe the transport of abstract particles within participating media in form of balance equations, resulting in the *scalar version of the particle transport equation*. By expressing the *reflection, transmission*, and *scattering behavior* of light in terms of so- Section 4.1 Section 4.2

called *bidirectional scattering functions*, then we obtain a geometric-physically correct mathematical formulation of the interaction of light particles with object surfaces in participating media and in free space. These functions also work as a basis for a mathematical description of light processes playing out within participating media and at the boundaries of a given scene including also *subsurface scattering*. Afterwards, we talk about the most important properties of different light sources. We will derive the *stationary light transport equation in participating media*, and in a *vacuum*, both in scalar form, i.e. neglecting the polarization properties of light. These equations can be considered as the mathematical formulations of the global illumination problem in computer graphics based on geometrical optics. Afterwards, we devote to the dual problem of light transport. That is, we present the *stationary importance transport equation in a vacuum*, as the adjoint equation of the stationary light transport in a vacuum and introduce the concept of importance as the adjoint quantity to radiance. Finally, we present with the definition of the *measurement equation*, an elegant mathematical formulation, which simulates the measurement process of light.

Section 4.3
Section 4.4
Section 4.5
Section 4.6

For all these discussions, we have decided to use a notation based on measure theory for describing phenomena occurring in particle transport, introduced in Chapter 2. Thus, we will always incorporate the measures we are working with in the given relations in a mathematically correct manner. Though physicists may be unaccustomed to this, it is mathematically consistent and appears to us to be advantageous, especially when it comes to Monte Carlo integration to be introduced in one of the following chapters.

4.1 PARTICLE AND LIGHT TRANSPORT IN PARTICIPATING MEDIA AND IN A VACUUM

In participating media particles can be generated at any moment within the media by chemical or thermal processes, that is, new particles can be injected into the system due to so-called *emission* processes. Now, the number of particles in a participating media can not only be increased by emission, it can also be decreased by *absorption* processes. During its travel through a participating medium, a particle can also collide with matter particles of the medium. As result of such a collision, a particle can change its direction of motion, so for example, it can be *out-scattered* from its original flow of direction, or the particle can be *in-scattered* from any other direction into a particular direction.

All these processes are not possible in a vacuum. In a vacuum, the number of particles always remains the same, that is, particles cannot be destroyed or newborn. If we further assume that particles are moving collision-free, a particle can also change its original direction of motion only due to reflection or refraction, if it interacts with existing surfaces in a scene.

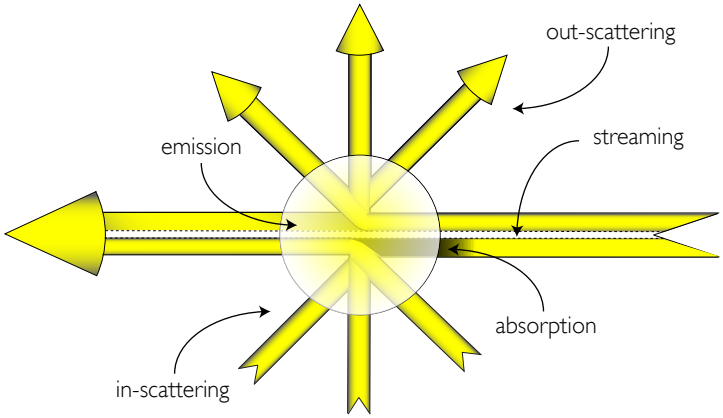


FIGURE 4.1: POSSIBLE INTERACTION OF LIGHT WITH PARTICIPATING MEDIA. We distinguish three categories of particle behavior when light interacts with participating media: Emission, streaming, and collisions. Collisions are further divided into absorption and scattering, where particles can be in- and out-scattered.

In the following sections, we will derive equations that describe the transport of abstract particles in participating media and in a vacuum. This requires the characterization of the particles we are interested in and of all above mentioned interactions of particles with matter that can affect the travel of a particle. That is, we are forced to consider all those effects which implies changes in the distribution of particles in a scene. So, let us assume that our particles are small and numerous and that they are moving collision-free with constant velocity through a medium or a vacuum, where at each moment a particle can be described by its current position $\mathbf{x} \in \mathbb{R}^3$ and its direction of motion ω. The rather complex nature of the resulting transport equation, along with the fact that in mathematics one often finds solution methods for specific cases of integro-differential equations, then suggests to transform this type of equation into a class of equations considerably more suitable for finding solution procedures, namely *linear integral equations*. Formulated in scalar representation the particle transport equation can then be interpreted as a mathematical description of the global illumination problem in a vacuum or participating media, restricted to geometrical optics. Assigning the abstract particles photon character using Einstein's formula $E = h\nu$, we then transform with the help of radiometric quantities, the particle transport equation into a mathematical formulation that explains the *photon transport*: the so-called *stationary light transport equation in scalar form*. Later, it serves as a basis for deriving the *light transport equation in participating media* and the *light transport equation in free space*, which for the field of computer graphics are

Section 4.1.1
Section 4.1.2
Section 4.1.3

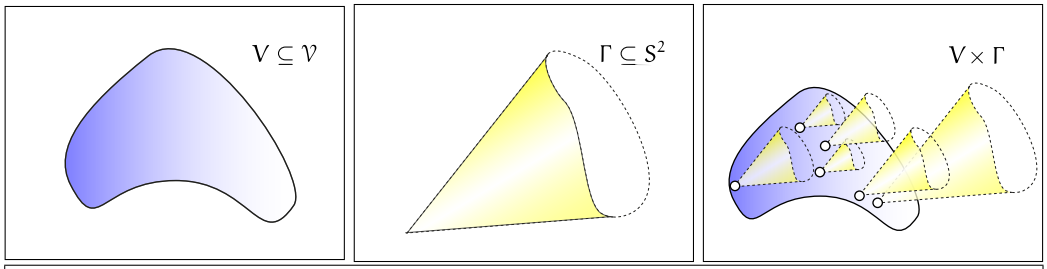


FIGURE 4.2: A SUBSPACE OF THE PARTICLE SPACE, Ξ. A finite volume $V \subseteq \mathcal{V}$ and a finite solid angle $\Gamma \subseteq S^2$. The Cartesian product of $V \subseteq \mathcal{V}$ and $\Gamma \subseteq S^2$ results in the 5-dimensional subspace $V \times \Gamma$ of particle space $\mathbb{R}^3 \times S^2$.

of fundamental importance.

4.1.1 THE STATIONARY PARTICLE TRANSPORT EQUATION IN INTEGRO-DIFFERENTIAL FORM

$\mathcal{V}$ (41) Let $V \times \Gamma$ with $V \subseteq \mathcal{V}$ and $\Gamma \subseteq S^2$ be a fixed volume of the particle space Ξ, see Figure 4.2.
Particle Space (244) Since we are only interested in *stationary* distributions we can assume that each volume
Particle Flux (246) has a fixed number of particles. This assumption then implies that the particle flux is independent on time, that is: The flow of particles into and out of $V \times \Gamma$ must exactly balance.

To derive a balance equation for particle transport we group the processes that are responsible for changes in the number of particles in $V \times \Gamma$ into gains and losses and equate these two quantities.

Let us assume that all particle behaviors fall into one of the following categories: *emission*, *in-scattering*, *out-scattering*, *streaming*, and *absorption*, where each of these processes can change the number of particles in $V \times \Gamma$. While emission and in-scattering increase the number of particles in $V \times \Gamma$, out-scattering and absorption decrease the number of particles in $V \times \Gamma$, see Figure 4.3.

For the following discussion, let us denote the number of particles, emitted or absorbed per unit of time in volume $V \times \Gamma$, by $\mathcal{E}$ and $\mathcal{A}$ respectively and the number of particles flowing into and out of $V \times \Gamma$ by $\mathcal{S}_{\text{in}}$ as well as $\mathcal{S}_{\text{out}}$. Furthermore, we denote with $\mathcal{C}_{\text{in}}$ and $\mathcal{C}_{\text{out}}$ the particles that scatter into $V \times \Gamma$ and those that scatter out of $V \times \Gamma$, see Figure 4.3. Then, the associated balance equation must satisfy the relation

$$\underbrace{\mathcal{E} + \mathcal{C}_{\text{in}} + \mathcal{S}_{\text{in}}}_{\text{gains}} = \underbrace{\mathcal{A} + \mathcal{C}_{\text{out}} + \mathcal{S}_{\text{out}}}_{\text{losses}} \tag{4.1}$$

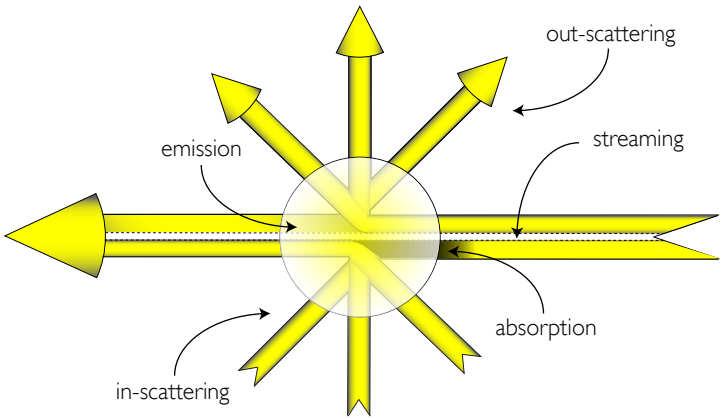


FIGURE 4.3: PHYSICAL PROCESSES RESPONSIBLE FOR CHANGES IN THE NUMBER OF PARTICLES IN PARTICIPATING MEDIA. We distinguish five types of processes that can change the number of particles within participating media: Emission, absorption, in- and out-scattering, as well as streaming.

or equivalently

$$\underbrace{\mathcal{E} + \mathcal{C}_{in} + \mathcal{S}_{in}}_{gains} - \left(\underbrace{\mathcal{A} + \mathcal{C}_{out} + \mathcal{S}_{out}}_{losses} \right) = 0. \tag{4.2}$$

Based on the concept of the particle density, n, the number $\mathcal{N}(t)$ of particles which are contained at time t in V and which move in directions of Γ is given by:

Particle Density (245)
$d\sigma$ (87)
$d\mu^3$ (82)

$$\mathcal{N}(t) \stackrel{def}{=} \int_V \int_\Gamma n(x, \omega, t)\, d\sigma_x(\omega)\, d\mu^3(x). \tag{4.3}$$

As already mentioned, we are only interested in stationary distributions of particles, that is, we can assume that the properties of the medium and the sources are time independent. For the change in number of particles in time interval $d\mu(t)$ then it must hold:

$$\frac{d\mathcal{N}(t)}{d\mu(t)} = 0 \qquad \left[\frac{1}{s}\right], \tag{4.4}$$

or, using Equation (4.2), equivalently:

$$\frac{d\mathcal{N}(t)}{d\mu(t)} \quad = \quad \mathcal{E} + \mathcal{C}_{in} - \left(\underbrace{(\mathcal{S}_{out} - \mathcal{S}_{in})}_{\mathcal{S}} + \mathcal{A} + \mathcal{C}_{out} \right) = 0. \tag{4.5}$$

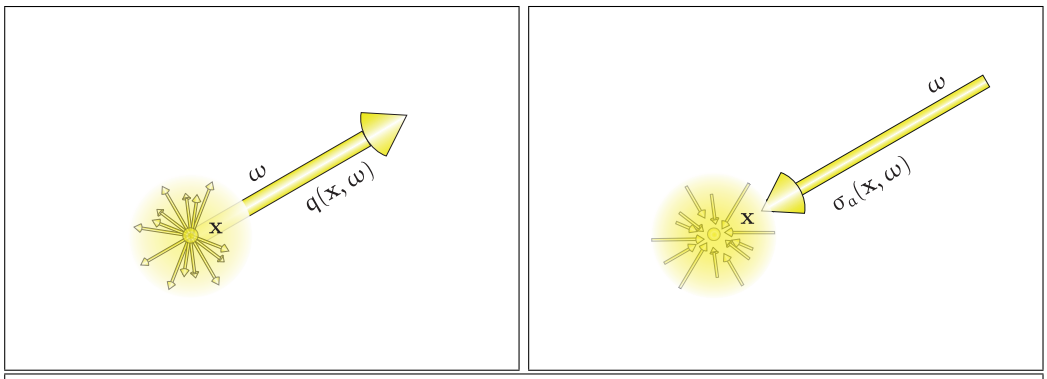


FIGURE 4.4: EMISSION AND ABSORPTION OF ABSTRACT PARTICLES. The particle space source function q emits abstract particles at point $\mathbf{x}$ in direction ω. In contrast to q, the function σ_a absorbs particles at point $\mathbf{x}$ coming from direction ω.

To formulate a closed mathematical expression that—based on the above balance equation—characterizes the particle transport, we now need a formal description of the individual components of Equation (4.5) occurring in a volume of the particle space taking into account participating media. Therefore, we must mathematically formulate all above-mentioned processes involved in particle transport and use them in Equation (4.5).

EMISSION. The first process we want to analyze is the *emission* of particles in participating media. It is responsible for the creation of new particles by one or more physical processes. Denoted by $\mathcal{E}$, the *emission term* can simply be described by using a so-called *particle space source function*, q,

$$q : \Xi \times T \to \mathbb{R}^{\geq 0} \tag{4.6}$$

also denoted as an *emission function*. It returns the number of particles created per unit volume, per unit solid angle and per unit time. The emission function is measured in units of $\frac{1}{m^3 \cdot sr \cdot s}$. Based on the emission function q the emission term, $\mathcal{E}$, is then defined as:

$$\mathcal{E} \stackrel{\text{def}}{=} \int_V \int_\Gamma q(\mathbf{x}, \omega) d\sigma_{\mathbf{x}}(\omega)\, d\mu^3(\mathbf{x}), \qquad \left[\frac{1}{s}\right] \tag{4.7}$$

it gives the number of particles emitted by all sources within the volume $V \times \Gamma$, see Figure 4.4.

ABSORPTION. Let us now analyze the absorption term, $\mathcal{A}$, describing the *absorption* behavior of particles. Similarly to emission, we now assume an *absorption coefficient*, σ_a,

$$\sigma_a : \Xi \to [0, \infty] \tag{4.8}$$

with

$$(\mathbf{x}, \omega) \mapsto \sigma_a(\mathbf{x}, \omega). \tag{4.9}$$

Under the condition, that the involved participating medium is isotropic, the absorption coefficient is independent of ω, thus, it has units of $\frac{1}{m}$ and gives the probability density function with which the particles at point $\mathbf{x}$, traveling in direction ω, will be absorbed. The absorption term, $\mathcal{A}$, is then defined via

Probability Density Function (176)

$$\mathcal{A} \stackrel{\text{def}}{=} \int_V \int_\Gamma \sigma_a(\mathbf{x}, \omega)\, \nu\, n(\mathbf{x}, \omega)\, d\sigma_{\mathbf{x}}(\omega)\, d\mu^3(\mathbf{x}) \qquad \left[\frac{1}{s}\right]. \tag{4.10}$$

REMARK 4.1 *Usually, the absorption coefficient is defined in transport theory as a density function and interpreted as the probability with which a photon traveling through a participating medium will be absorbed per unit length and per unit solid angle. Often, the direction from which the photon enters the center of absorption plays no role, i.e. the medium is assumed to be isotropic. Unfortunately, this is frequently not the case, especially for natural participating media [44, Ditchburn 1991]. So, for example, the blue coloring of the sky results from the fact that the absorption behavior of a particle demonstrates both direction dependence and wavelength dependence.*

Probability Density Function (176)

STREAMING. Next, we are interested in the number of particles with directions in Γ that either escape from or enter into the volume V simply by *streaming*. The change in $\mathcal{N}(t)$ due to streaming can now be written as the net flow of particles with directions in Γ that pass through the boundary ∂V of $V \times \Gamma$. Since the net flow through a surface depends only on the component of the particle flow which is normal to the patch, we define a so-called *streaming term*, $\mathcal{S}$, by:

$\partial\mathcal{V}$ (41)

$d\sigma_s$ (87)

$$\mathcal{S} \stackrel{\text{def}}{=} \int_{\partial V} \int_\Gamma \nu\, n(\mathbf{s}, \omega)\, \langle \omega, \mathbf{N}(\mathbf{s}) \rangle\, d\sigma_{\mathbf{s}}(\omega)\, d\mu^2(\mathbf{s}), \qquad \left[\frac{1}{s}\right] \tag{4.11}$$

where $\mathbf{N}(\mathbf{s})$ is the normal at point $\mathbf{s} \in \partial V$ and the inner product $\langle \mathbf{N}(\mathbf{s}), \omega \rangle$ indicates the direction of the net flow. A positive value signalizes a net flow into the volume V, thus an increase of $\mathcal{N}(t)$, while a negative value signalizes a net flow out of V, thus, a decrease of $\mathcal{N}(t)$.

$\langle \cdot, \cdot \rangle$ (845)

Now, the streaming term, as defined above, is expressed as an integral over ∂V. Further below, it will be clear that it is advantageous to convert this surface integral into a volume integral. Using the *Gauss Divergence Theorem*,[1] the *streaming term* $\mathcal{S}$ can also be expressed as an integral over the volume $V \times \Gamma$, namely as:

$\partial\mathcal{V}$ (41)

∇ (53)

[1] Assume V is a compact subset of $\mathbb{R}^s$ with piecewise smooth boundary ∂V. Let furthermore $\mathbf{F}$ be a continuous, differentiable, vector-valued function defined on a neighborhood of S, then we have

$$\int_V \langle \nabla, \mathbf{F}(\mathbf{x}) \rangle d\mu^3(\mathbf{x}) = \int_{\partial V} \langle \mathbf{F}(\mathbf{s}), \mathbf{n}(\mathbf{s}) \rangle d\mu^2(\mathbf{s}). \tag{4.12}$$

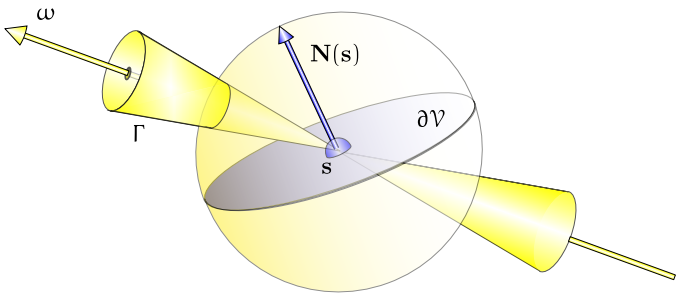


FIGURE 4.5: STREAMING BEHAVIOR OF ABSTRACT PARTICLES. The calculation of the number of particles streaming through the subset $\partial V \times \Gamma$ of the phase space Ξ by means of integration over the boundary $\partial V \in \mathcal{M}$.

$$\mathcal{S} \overset{\text{Gauss}}{=} \int_V \int_\Gamma \langle \omega, \nabla \rangle \, \nu \, n(\mathbf{x}, \omega) \, d\sigma_{\mathbf{x}}(\omega) \, d\mu^3(\mathbf{x}). \tag{4.13}$$

SCATTERING. With the definition of the entities $\mathcal{E}, \mathcal{A}$, and $\mathcal{S}$, the only thing that remains to complete our characterization of particle transport is the analysis of *scattering*, which, compared to emission and absorption, turns out to be slightly more complex.

Integral Kernel (127)

Mathematically, *scattering* is described by introducing a linear integral kernel κ,

$$\kappa : \mathcal{V}^o \times S^2 \times S^2 \to [0, 1] \tag{4.14}$$

the so-called *volume scattering kernel*, κ, with

$$(\mathbf{x}, \omega, \omega') \mapsto \kappa(\mathbf{x}, \omega \to \omega'). \tag{4.15}$$

Probability Density Function (176)

Its units are $\frac{1}{m \cdot sr}$ and it gives the probability that a particle at $\mathbf{x}$ moving in direction ω will be deflected into the new direction ω'.

Considering the scattering of a particle more exactly, then we must distinguish between two scattering processes: *in-scattering* and *out-scattering*. By in-scattering we mean the process in which particles, regardless of their starting point, are transmitted into the direction of the currently observed stream, whereby out-scattering describes the change in the particles original path. So, for those particles $\mathcal{C}_{in}$ arriving at point $\mathbf{x} \in V$ from all directions ω' over the unit sphere and scattered into direction $\omega \in \Gamma$ the following expression holds:

$$\mathcal{C}_{in} \overset{\text{def}}{=} \int_V \int_\Gamma \int_{S^2(\mathbf{x})} \kappa(\mathbf{x}, \omega' \to \omega) \, \nu \, n(\mathbf{x}, \omega') \, d\sigma_{\mathbf{x}}(\omega') d\sigma_{\mathbf{x}}(\omega) d\mu^3(\mathbf{x}) \qquad \left[\frac{1}{s}\right]. \tag{4.16}$$

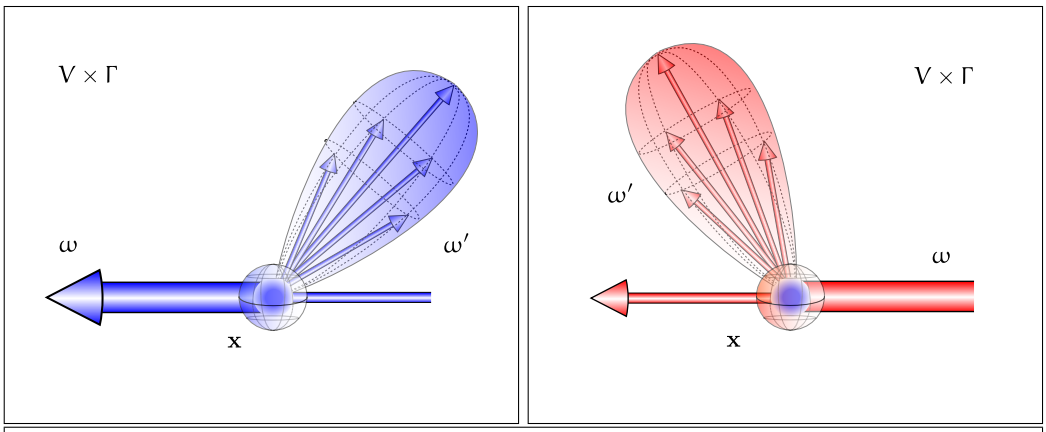


FIGURE 4.6: IN AND OUT-SCATTERING OF ABSTRACT PARTICLES. In-scattering is the process by which particles at point $\mathbf{x}$ flowing from directions ω' are added to $V \times \Gamma$. Out-scattering describes the change in the particles original path at a point $\mathbf{x} \in V$ in directions $\omega' \notin \Gamma$.

Obviously, out-scattered particles leave their original direction and are no longer being relevant for our considerations. We get the number of particles, $\mathcal{C}_{\text{out}}$, that at point $\mathbf{x}$ are forced to change their original direction $\omega \in \Gamma$ into any other $\omega' \in S^2$ by:

$$\mathcal{C}_{\text{out}} \stackrel{\text{def}}{=} \int_V \int_\Gamma \nu n(\mathbf{x}, \omega) \int_{S^2(\mathbf{x})} \kappa(\mathbf{x}, \omega \to \omega')\, d\sigma_{\mathbf{x}}(\omega') d\sigma_{\mathbf{x}}(\omega) d\mu^3(\mathbf{x}) \qquad \left[\frac{1}{s}\right]. \quad (4.17)$$

REMARK 4.2 *Note that the inner integration in $\mathcal{C}_{\text{in}}$ and $\mathcal{C}_{\text{out}}$ is over the entire sphere of directions, which means that we account for particles whose directions lie within Γ both before and after scattering. Since we are only interested in the net change due to scattering, that is, the difference between $\mathcal{C}_{\text{in}}$ and $\mathcal{C}_{\text{out}}$, the unwanted fraction will be canceled out, since it appears on both sides of our balance equation.*

THE STATIONARY PARTICLE TRANSPORT EQUATION IN INTEGRO-DIFFERENTIAL FORM. Let us consider once more the inner integral in Equation (4.17), thus the integral over the kernel κ, then it is possible to replace this integral by a function, σ_s, that is,

$$\sigma_s(\mathbf{x}, \omega) \stackrel{\text{def}}{=} \int_{S^2(\mathbf{x})} \kappa(\mathbf{x}, \omega \to \omega')\, d\sigma_{\mathbf{x}}(\omega'), \qquad \left[\frac{1}{m}\right] \quad (4.18)$$

where σ_s is called the *out-scattering coefficient*. In analogy to the absorption coefficient σ_a, the out-scattering coefficient gives the probability that a particle will suffer a scattering collision per unit distance traveled.

Using the out-scattering coefficient σ_s in Equation (4.17), the out-scattering term

$\mathcal{C}_{\text{out}}$ can then be simplified described similarly to the absorption term $\mathcal{A}$, as:

$$\mathcal{C}_{\text{out}} \stackrel{\text{def}}{=} \int_V \int_\Gamma \sigma_s(\mathbf{x},\omega)\, v\, n(\mathbf{x},\omega)\, d\sigma_{\mathbf{x}}(\omega) d\mu^3(\mathbf{x}). \qquad \left[\frac{1}{s}\right] \tag{4.19}$$

The Formulas (4.10) and (4.19) of the absorption term and the out-scattering term now imply the introduction of a so-called *extinction coefficient* defined by,

$$\sigma_t(\mathbf{x},\omega) \stackrel{\text{def}}{=} \sigma_a(\mathbf{x},\omega) + \sigma_s(\mathbf{x},\omega), \qquad \left[\frac{1}{m}\right] \tag{4.20}$$

which is the probability that a particle will be subjected to either kind of collision per unit distance traveled, i.e. the particle is absorbed or scattered.

Based on the extinction coefficient σ_t the balance equation can now be written as:

$$\mathcal{S} + \underbrace{(\mathcal{A} + \mathcal{C}_{\text{out}})}_{\mathcal{C}_{\text{ext}}} = \mathcal{E} + \mathcal{C}_{\text{in}} \tag{4.21}$$

thus

$$\mathcal{S} + \mathcal{C}_{\text{ext}} = \mathcal{E} + \mathcal{C}_{\text{in}}. \tag{4.22}$$

Because V and Γ have been chosen arbitrarily and the terms of Equation (4.22) all have the same integration domain, it follows that equality must also hold for the integrands. Removing the two outer integrals then leads to the *stationary particle transport equation in intego-differential form*:

$$\begin{aligned} \langle \omega, \nabla \rangle v\, n(\mathbf{x},\omega) \quad &+ \quad \sigma_t(\mathbf{x},\omega)\, v\, n(\mathbf{x},\omega) \\ &= \quad q(\mathbf{x},\omega) + \int_{S^2(\mathbf{x})} \kappa(\mathbf{x},\omega' \to \omega)\, v\, n(\mathbf{x},\omega')\, d\sigma_{\mathbf{x}}(\omega'). \end{aligned} \tag{4.23}$$

Replacing the particle density n according to Equation (3.18) by the flux $\frac{\Phi_P(\mathbf{x},\omega)}{v}$, where v is the velocity of a particle, we get the *stationary particle transport equation in intego-differential form*, expressed in terms of particle flux:

$$\begin{aligned} \langle \omega, \nabla \rangle \Phi_P(\mathbf{x},\omega) \quad &+ \quad \sigma_t(\mathbf{x},\omega)\, \Phi_P(\mathbf{x},\omega) \\ &= \quad q(\mathbf{x},\omega) + \int_{S^2(\mathbf{x})} \kappa(\mathbf{x},\omega' \to \omega)\, \Phi_P(\mathbf{x},\omega')\, d\sigma_{\mathbf{x}}(\omega'). \end{aligned} \tag{4.24}$$

An equation of this type is denoted as an *ordinary integro-differential equation*. It is valid only in interior points of the underlying volume. Thus, finding a solution to the particle transport equation also requires setting up the boundary conditions, which correspond to a formal description of particle distribution at the surfaces delimiting our system.

STATIONARY PARTICLE TRANSPORT EQUATIONS WITH BOUNDARY CONDITION. Out of a variety of possible formulations of boundary conditions we restrict our further considerations only to the two simplest types: *explicit* and *implicit boundary* conditions.

We call a boundary condition *explicit* if the number of particles leaving a surface point is independent of the number of incident particles. The case where particles leaving a point on a surface depends on incident particles is specified as an *implicit boundary condition*.

By means of the concept of explicit boundary conditions we can now simply describe the emission behavior of particle sources at the boundaries of $V \times \Gamma$ by selecting a particle space source function

$$\Phi_P(s, \omega) = q_b(s, \omega) \tag{4.25}$$

defined on $\partial V \times S^2$. It returns the number of particles entering in the system that are created by emission processes on the involved boundaries. An implicit boundary condition could be defined by an integral transformation of the form

$$\Phi_P(s, \omega) = \int_{S^2} \kappa_b(s, \omega' \to \omega)\, \Phi_P(s, \omega')\, d\sigma_s(\omega'), \tag{4.26}$$

where κ_b is the *surface scattering kernel* measured in units $\frac{1}{sr}$. This construct then expresses the number of particles reflected at surface point s in any direction ω as a weighted sum of the incoming particles at s, where the weighting can depend on the incoming and outgoing directions.

Combining an explicit and an implicit boundary condition—where we additionally integrate over all locations on the boundary ∂V where emission processes can occur—we finally arrive at the *stationary particle transport equation with implict and explicit boundary conditions*, also briefly called *SPTE in integro-differential form*, $\partial\mathcal{V}$ (41)

$$\begin{aligned} \langle \omega, \nabla \rangle \Phi_P(x, \omega) \quad &+ \quad \sigma_t(x, \omega)\, \Phi_P(x, \omega) \\ &= \quad q(x, \omega) + \int_{S^2(x)} \kappa(x, \omega' \to \omega)\, \Phi_P(x, \omega')\, d\sigma_x(\omega'), \end{aligned} \tag{4.27}$$

where it holds $x \in \mathcal{V}^o$ in all interior points x of the volume $V \times \Gamma$, and $\mathcal{V}^o$ (41)

$$\Phi_P(s, \omega) = q_b(s, \omega) + \int_{S^2(s)} \kappa_b(s, \omega' \to \omega)\, \Phi_P(s, \omega')\, d\sigma_s(\omega'), \tag{4.28}$$

are the associated boundary conditions. This equation then describes, together with their boundary condition, the transport of abstract particles in a closed system of the particle space.

4.1.2 THE STATIONARY PARTICLE TRANSPORT EQUATION IN INTEGRAL FORM

As already mentioned, the stationary particle transport equation is an integro-differential equation because it contains the derivative and the integral of the unknown function Φ_P.
Integro-differential Equation (127) Compared to the theory of solving differential equations and integral equations, the theory
Φ_P (246) for solving such a type of equation is not as well developed. Therefore, for exploring a
Section 2.3.3 variety of solution methods, we will make use of the advantage to transform integro-differential equations into differential equations or integral equations. Since, it has proven advantageous to convert our particle transport equation into an integral equation, we will now present this approach. It allows to involve the boundary conditions into the integral equation rather than maintaining them separately as a set of constraints.

In the first step of the following discussion we convert, by means of appropriate transformations, our particle transport equation into a differential equation. Integrating this differential equation then incorporates the boundary conditions and results in our desired integral equation.

From our discussion in Section A.5 it is known that the operator $\langle \omega, \nabla \rangle$ from Equation
Directional Derivative (871) (4.27) is a directional derivative in direction ω. Such a derivative can be written as

$$\langle \omega, \nabla \rangle \Phi_P(\mathbf{x}, \omega) \overset{(A.30)}{=} \frac{\partial}{\partial \mu(\alpha)} \Phi_P(\mathbf{x} + \alpha\omega, \omega) \Big|_{\alpha=0} \tag{4.29}$$

$$= -\frac{\partial}{\partial \mu(\alpha)} \Phi_P(\mathbf{x} - \alpha\omega, \omega) \Big|_{\alpha=0} \tag{4.30}$$

and is interpreted as the change in flux along the ray $\mathbf{r} = \mathbf{x} + \alpha\omega, \alpha \in \mathbb{R}$, see Figure 4.7.

With respect to the particle transport equation from Formula (4.27) we are now interested in what happens along the ray $\mathbf{x} - \alpha\omega$, that is, in backward direction towards surface point $\mathbf{s}$. For that purpose, let us replace the operator $\langle \omega, \nabla \rangle$ in Equation (4.27) by the directional derivative from Equation (4.30). Then the stationary particle transport equation in all inner points $\mathbf{x}$ of the volume $V \times \Gamma$ takes on the form

$$\frac{\partial \Phi_P(\mathbf{x} + \alpha\omega, \omega)}{\partial \mu(\alpha)} + \sigma_t(\mathbf{x} - \alpha\omega, \omega)\, \Phi_P(\mathbf{x} - \alpha\omega, \omega) \tag{4.31}$$

$$= q(\mathbf{x} - \alpha\omega, \omega) + \int_{S^2(\mathbf{x} - \alpha\omega)} \kappa(\mathbf{x} - \alpha\omega, \omega' \to \omega)\, \Phi_P(\mathbf{x} - \alpha\omega, \omega')\, d\sigma_{\mathbf{x} - \alpha\omega}(\omega').$$

Particle Space Flux (246) This expression can now be interpreted as follows: The number of particles at an inner
Scattering (284) point $\mathbf{x}$ within a participating medium results from scattering, absorption, or emission
Absorption (282) processes that occur on the ray $\mathbf{x} - \alpha\omega$ in backward direction to a surface point $\mathbf{s}$, while
Emission (282) the number of particles at a surface point is predetermined via the boundary conditions,
Boundary Conditions (287) see Figure 4.8.

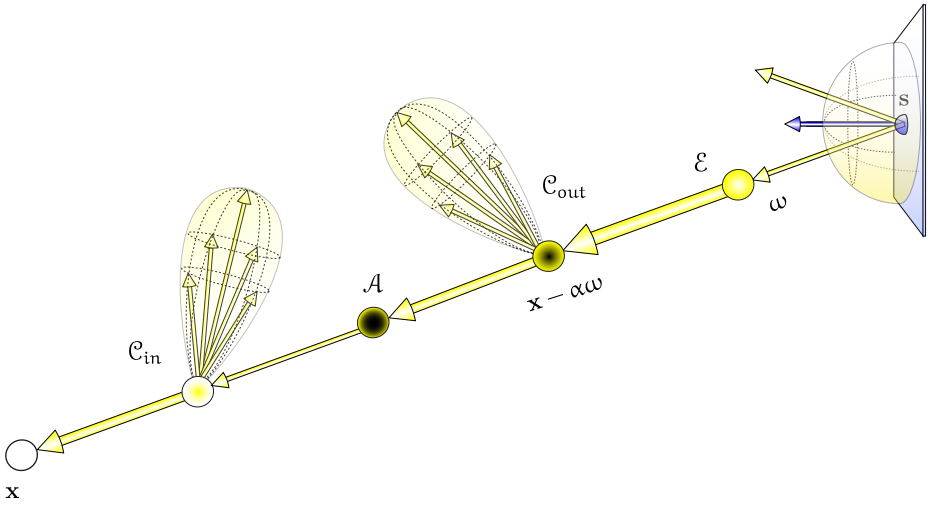


FIGURE 4.7: THE CHANGE IN FLUX ALONG THE RAY $\mathbf{r} = \mathbf{x} - \alpha\omega, \alpha \in \mathbb{R}$. Between the points $\mathbf{x} \in \mathbb{R}^3$, $\mathbf{x} - \alpha\omega, \alpha > 0$ and $\mathbf{s}$, the amount of flux can change due to emission, absorption, and scattering processes.

To further simplify our notation, we define a so-called *gain function* $Q(\mathbf{x}, \omega)$ that combines the emission and the in-scattering term

$$Q(\mathbf{x} - \alpha\omega, \omega) \tag{4.32}$$
$$\stackrel{def}{=} q(\mathbf{x} - \alpha\omega, \omega) + \int_{S^2(\mathbf{x}-\alpha\omega)} \kappa(\mathbf{x} - \alpha\omega, \omega' \to \omega)\, \Phi_P(\mathbf{x} - \alpha\omega, \omega')\, d\sigma_{\mathbf{x}-\alpha\omega}(\omega').$$

With the new functions

$$\widehat{\Phi_P}(\alpha) \stackrel{def}{=} \Phi_P(\mathbf{x} - \alpha\omega, \omega) \tag{4.33}$$
$$\widehat{Q}(\alpha) \stackrel{def}{=} Q(\mathbf{x} - \alpha\omega, \omega) \tag{4.34}$$
$$\widehat{\sigma_t}(\alpha) \stackrel{def}{=} \sigma_t(\mathbf{x} - \alpha\omega, \omega) \tag{4.35}$$

Equation (4.31) can be represented as a linear, first-order, ordinary differential equation of the form

$$-\frac{d}{d\mu(\alpha)}\widehat{\Phi_P}(\alpha) + \widehat{\Phi_P}(\alpha)\,\widehat{\sigma_t}(\alpha) = \widehat{Q}(\alpha) \tag{4.36}$$
$$\frac{d}{d\mu(\alpha)}\widehat{\Phi_P}(\alpha) - \widehat{\Phi_P}(\alpha)\,\widehat{\sigma_t}(\alpha) = -\widehat{Q}(\alpha), \tag{4.37}$$

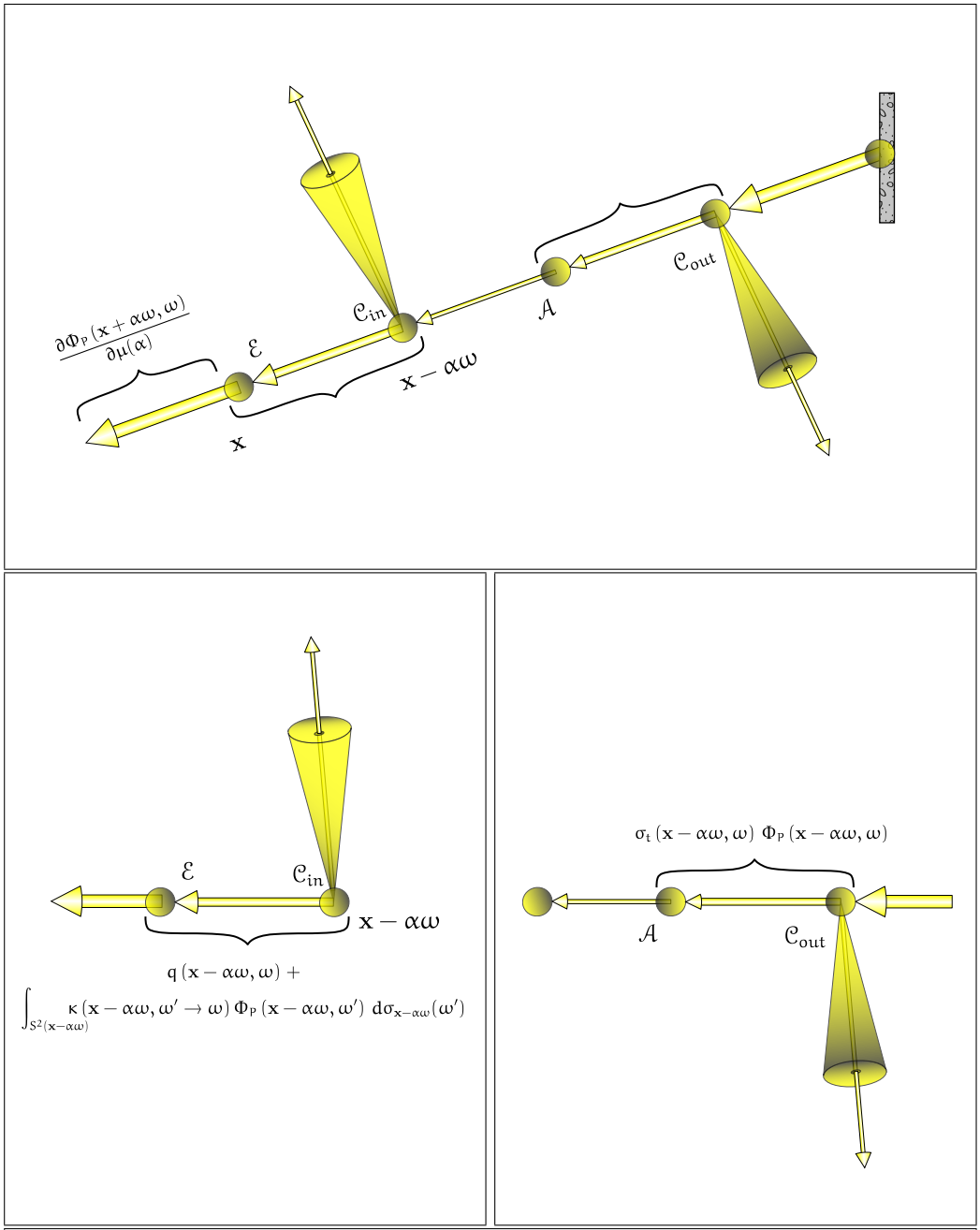


FIGURE 4.8: PARTICLE FLUX OVER THE RAY $\mathbf{r} = \mathbf{x} + \alpha\omega$. The change in the number of particles at point $\mathbf{x}$ in the Euclidean space $\mathbb{R}^3$ is a composite of emission, $\mathcal{E}$, absorption, $\mathcal{A}$, and scattering, $\mathcal{S}_{in}, \mathcal{S}_{out}$, processes along the ray $\mathbf{r} = \mathbf{x} - \alpha\omega$, $0 < \alpha \leq d_\partial(\mathbf{x}, \omega)$.

that can be solved by means of an integrating factor. For that purpose, first we search a function $h(\alpha)$, which, multiplied with the left-hand side of Equation(4.37), corresponds to the derivative of $\widehat{\Phi_P}(\alpha)\, h(\alpha)$, i.e., it should hold:

$$\left(\frac{d}{d\mu(\alpha)}\widehat{\Phi_P}(\alpha) - \widehat{\Phi_P}(\alpha)\,\widehat{\sigma}_t(\alpha)\right) h(\alpha) \quad = \quad \left(\frac{d}{d\mu(\alpha)}\widehat{\Phi_P}(\alpha)\right) h(\alpha) + \widehat{\Phi_P}(\alpha)\left(\frac{d}{d\mu(\alpha)}h(\alpha)\right). \tag{4.38}$$

Multiplication on the left shows that the first term on each side is equal. This leads to

$$-\widehat{\Phi_P}(\alpha)\,\widehat{\sigma}_t(\alpha)h(\alpha) = \widehat{\Phi_P}(\alpha)\frac{d}{d\mu(\alpha)}h(\alpha). \tag{4.39}$$

or equivalently

$$\widehat{\sigma}_t(\alpha) = -\frac{\frac{d}{d\mu(\alpha)}h(\alpha)}{h(\alpha)}. \tag{4.40}$$

Finally, we integrate this relationship along the ray $\mathbf{r} = \mathbf{x} - \alpha\omega$ from 0 to α resulting in

$$\ln\left(h(\alpha)\right) = \int_{[0,\alpha]} -\widehat{\sigma}_t(\xi)\; d\mu(\xi), \tag{4.41}$$

i.e., the function we are seeking is

$$h(\alpha) = \exp\left(\int_{[0,\alpha]} -\widehat{\sigma}_t(\xi)\; d\mu(\xi)\right). \tag{4.42}$$

Now, multiplying Equation (4.37) by $h(\alpha)$ yields

$$\left(\frac{d}{d\mu(\alpha)}\widehat{\Phi_P}(\alpha)\right) h(\alpha) - \widehat{\Phi_P}(\alpha)\,\widehat{\sigma}_t(\alpha)h(\alpha) = -\widehat{Q}(\alpha)h(\alpha) \tag{4.43}$$

which—due to the fact that the left-hand side corresponds to the derivative of the product $\widehat{\Phi_P}(\alpha)\, h(\alpha)$—can also be written as

$$\frac{d}{d\mu(\alpha)}\left(\widehat{\Phi_P}(\alpha)\cdot h(\alpha)\right) = -\widehat{Q}(\alpha)h(\alpha). \tag{4.44}$$

Integrating this expression from 0 to α yields

$$\int_{[0,\alpha]} \frac{d}{d\mu(\xi)}\left(h(\xi)\widehat{\Phi_P}(\xi)\right)\; d\mu(\xi) = -\int_{[0,\alpha]} h(\xi)\widehat{Q}(\xi)\, d\mu(\xi) \tag{4.45}$$

thus

$$h(\xi)\widehat{\Phi_P}(\xi)\Big|_{\xi=0}^{\xi=\alpha} = -\int_{[0,\alpha]} h(\xi)\widehat{Q}(\xi)\,d\mu(\xi). \tag{4.46}$$

Rephrasing Equation (4.46) with $h(0) = 1$ then leads to

$$\widehat{\Phi_P}(0) = h(\alpha)\widehat{\Phi_P}(\alpha) + \int_{[0,\alpha]} h(\xi)\widehat{Q}(\xi)d\mu(\xi). \tag{4.47}$$

Equation (4.47) describes how we can find the number of particles $\widehat{\Phi_P}(0) = \Phi_P(\mathbf{x},\omega)$ in terms of the number of particles arriving along direction ω towards a point $\mathbf{s} = \mathbf{x} - \alpha\omega$
γ (47) at a surface. To find this point we use the ray-casting function, γ, i.e. we compute

$$\mathbf{s} = \gamma(\mathbf{x},\omega), \tag{4.48}$$

which returns the intersection point $\mathbf{s}$ of the ray starting point $\mathbf{x}$ in direction ω with any of the surfaces in the scene. To ensure that the ray-casting function is always well-defined,
∂V (41) we assume that the set of object surfaces $\partial\mathcal{V}$ encloses the scene to be rendered, such as an infinitely large, black, and non-reflecting sphere.

To involve the particle number at boundaries into Equation (4.47), we define an exitant boundary function $\Phi_{P,b}$ for any $(\mathbf{s},\omega) \in \partial\mathcal{V} \times S^2$ by

$$\Phi_{P,b}(\mathbf{s},\omega) \stackrel{def}{=} q_b(\mathbf{s},\omega) + \int_{S^2(\mathbf{s})} \kappa_b(\mathbf{s},\omega' \to \omega)\,\Phi_P(\mathbf{s},\omega')\,d\sigma_{\mathbf{s}}(\omega'), \tag{4.49}$$

where q_b is a *surface emission function* and $\Phi_P(\mathbf{s},\omega')$ is the incident number of particles at the surface point $\mathbf{s}$ coming from directions ω'.

Since it is required to generalize Equation (4.47) back to a form that holds for any ray, we will introduce several new functions: The *optical distance function*, τ, is defined by

$$\tau(\mathbf{x},\mathbf{y}) \stackrel{def}{=} \int_{[0,\|\mathbf{y}-\mathbf{x}\|]} \sigma_t\,(\mathbf{x} - \alpha\omega, \omega)\;d\mu(\alpha) \tag{4.50}$$

and the *path absorption function*, β, is given by

$$\beta(\mathbf{y},\mathbf{x}) \stackrel{def}{=} \exp(-\tau(\mathbf{x},\mathbf{y})). \tag{4.51}$$

Obviously, the latter is a more general form of the integrating factor h from above since it holds

$$\beta(\mathbf{y},\mathbf{x}) = h(\alpha) \tag{4.52}$$

for $\mathbf{y} = \mathbf{x} - \alpha\omega$.

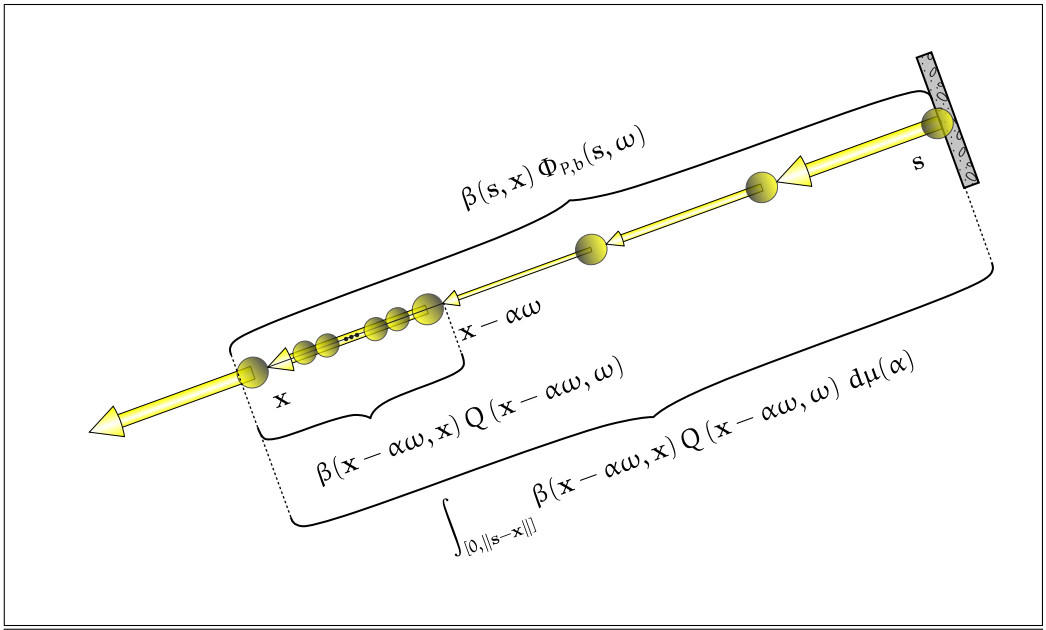


FIGURE 4.9: THE INTEGRAL FORM OF THE PARTICLE TRANSPORT EQUATION. The number of particles at the inner point $\mathbf{x} \in \mathcal{V}^o$ is composed of two components. First, the particles emanating from point $\mathbf{s}$ on a surface that is nearest to $\mathbf{x}$ in direction $-\omega$ reduced by possible out-scattering and absorption processes on the way to $\mathbf{x}$. The other component is found by looking backwards along the ray between $\mathbf{x}$ and $\mathbf{s}$. At each point $\mathbf{x}-\alpha\omega, 0 \le \alpha < d_\partial(\mathbf{x},\omega)$, we add to the volumetric emission the in-scattered particles at that point and we adjust this amount by volumetric effects of out-scattering and absorption as it travels from that point back to $\mathbf{x}$.

REMARK 4.3 *Physically, the optical distance function τ often also called the* optical thickness *can be interpreted as a dimensionless cumulative measure of absorption and out-scattering over distance in the medium, that relates the physical distance to the optical distance. As a measure of penetration depth it is used to characterize optically thick layer, $\tau \gg 1$, and optically thin layers, $\tau \ll 1$.*

Based on the above discussion, the integro-differential form of the SPTE from Formula (4.27) can now be expressed via Equation (4.47) more easily in integral form as

$$\begin{aligned}\Phi_P(\mathbf{x},\omega) &= \beta(\mathbf{s},\mathbf{x})\,\Phi_{P,b}(\mathbf{s},\omega) + \\ &\int_{[0,\|\mathbf{s}-\mathbf{x}\|]}\beta(\mathbf{x}-\alpha\omega,\mathbf{x})\,Q\,(\mathbf{x}-\alpha\omega,\omega)\;d\mu(\alpha),\end{aligned} \tag{4.53}$$

see Figure 4.9.

The complete form of the SPTE in integral form is then given by:

DEFINITION 4.1 (The Stationary Particle Transport Equation in Integral Form, SPTE) *Let* β *be the path absorption function from Relation (4.51) and* q_b, q *as well as* κ_b, κ *denote particle sources respectively describe the scattering behavior of particles at surfaces or within participating media, then, the one-speed* stationary particle transport equation in integral form, *also briefly denoted as* SPTE, *is defined as:*

$$\begin{aligned}
&\Phi_P(\mathbf{x},\omega)\\
&\overset{(4.49,4.54)}{=} \beta(\mathbf{s},\mathbf{x})\left(q_b(\mathbf{s},\omega)+\int_{S^2(\mathbf{s})}\kappa_b(\mathbf{s},\omega'\to\omega)\,\Phi_P(\mathbf{s},\omega')\,d\sigma_{\mathbf{s}}(\omega')\right)+\\
&\qquad\int_{[0,\|\mathbf{s}-\mathbf{x}\|]}\beta(\mathbf{x}',\mathbf{x})\left(q(\mathbf{x}',\omega)+\right.\\
&\qquad\qquad\left.\int_{S^2(\mathbf{x}')}\kappa(\mathbf{x}',\omega'\to\omega)\,\Phi_P(\mathbf{x}',\omega')\,d\sigma_{\mathbf{x}'}(\omega')\right)d\mu(\alpha),
\end{aligned}$$

where $\mathbf{x}' = \mathbf{x} - \alpha\omega$.

Stationary PTE (294) Particle Space Flux (246) Out-Scattering (284) Absorption (282) Emission (282)

Physically, the stationary particle transport equation can be interpreted as the number of particles at point $(\mathbf{x},\omega)$ composed of two components. One of the components describes particles emanating from point $\mathbf{s}$ on a surface that is nearest to $\mathbf{x}$ in direction $-\omega$ reduced by possible out-scattering and absorption processes on the way to $\mathbf{x}$. The other component is found by looking backwards along the ray between $\mathbf{x}$ and $\mathbf{s}$. At each point we add to the volumetric emission the in-scattered particles at that point and adjust this amount by volumetric effects of out-scattering and absorption as it travels from that point to $\mathbf{x}$.

REMARK 4.4 *The integral form of the stationary particle transport equation is, as our discussion makes clear, equivalent to the integro-differential equation from (4.27). Furthermore, it offers the additional advantage of explicitly including the necessary* Boundary Conditions (287) *boundary conditions rather than requiring a separate equation. In its scalar version* Section 4.4 *it serves as the basis for deriving all of the integral equations describing the global* Section 4.5 *illumination problem based on the principles of geometric optics.*

REMARK 4.5 *In the field of computer graphics we will encounter primarily the stationary particle transport equation as a fundamental mathematical formulation describing particle transport albeit in a somewhat altered and considerably simplified form. As the central formulation describing the global illumination problem we will* Section 4.4 *encounter the* light transport equation in participating media *and the* light transport equation in free space *as well as the adjoint of the light transport equation in a vac-* Section 4.5 *uum, the* importance equation in a vacuum. *In Chapter 6, we then develop for these equations solutions techniques that are based on probability theoretical approaches.*

4.1.3 THE STATIONARY LIGHT TRANSPORT EQUATION IN INTEGRAL FORM

The transport equation derived in the previous section is not based on any plausible physical concept but rather, it is rooted at the model of abstract particles moving through a medium. Now, however, when we consider particle transport under the assumptions governing photon transport then the nature of a photon requires to incorporate photon-specific properties into our transport equation, i.e. the speed and energy of light, as well as the behavior of light quanta at the boundaries of the given scene.

Photon (246)

From Equation (3.16) and (3.18) it follows that the relationship between the incident flux of particles and incident radiance is given by:

Particle Flux (246)
Radiance (250)

$$L_i(\mathbf{x}, \omega_i) = \hbar\nu \, \Phi_P(\mathbf{x}, \omega_i) \qquad \left[\frac{W}{m^2 \cdot sr}\right]. \tag{4.54}$$

Linking not only Φ_P, but also the particle source functions q to the physical quantity of radiant power, then we must replace the surface and volume emission terms q_b and q from Equation (4.54) by two radiance emitting functions, the *volume radiance emission function*, ϵ, defined by:

q (282)

$$\epsilon(\mathbf{x}, \omega_o) \overset{def}{=} \hbar\nu \, q(\mathbf{x}, \omega_o) \tag{4.55}$$

$$\overset{\nu=\frac{c}{\lambda}}{=} \frac{\hbar c}{\lambda} q(\mathbf{x}, \omega_o) \qquad \left[\frac{W}{m^3 \cdot sr}\right] \tag{4.56}$$

and the *surface radiance emission function*, $\epsilon_b(\mathbf{s}, \omega_o)$, given by:

$$\epsilon_b(\mathbf{s}, \omega_o) \overset{def}{=} \hbar\nu \, q_b(\mathbf{s}, \omega_o) \tag{4.57}$$

$$\overset{\nu=\frac{c}{\lambda}}{=} \frac{\hbar c}{\lambda} q_b(\mathbf{s}, \omega_o) \qquad \left[\frac{W}{m^2 \cdot sr}\right]. \tag{4.58}$$

Using the concept of radiance, then the stationary particle transport equation in integral form can be expressed as:

Radiance (250)
PTE in Integral Form (294)

DEFINITION 4.2 (The Stationary Light Transport Equation, SLTE) *Assume* $\mathbf{s}$ *is a point on a surface* $M \in \partial\mathcal{V}$, $\mathbf{x}, \mathbf{x}' \in \mathcal{V}^o$ *are inner points of a participating medium,* $L_b(\mathbf{s}, \omega_o)$ *and* $Q_o(\mathbf{x}', \omega_o)$ *are the outgoing radiance from surface point* $\mathbf{s}$*, respectively from point* $\mathbf{x}'$ *within a participating medium in direction* ω_o *attenuated by the path absorption function* β*, and* $L_i(\mathbf{x}, \omega_i)$ *is the incident radiance at point* $\mathbf{x}$ *from direction* ω_i*, then:*

$\partial\mathcal{V}$ (41)
$\mathcal{V}^o$ (41)
β (292)

$$L_i(\mathbf{x}, \omega_i) = \beta(\mathbf{s} \to \mathbf{x}) \, L_b(\mathbf{s}, \omega_o) + \int_{[0, \|\mathbf{s}-\mathbf{x}\|]} \beta(\mathbf{x}' \to \mathbf{x}) \, Q_o(\mathbf{x}', \omega_o) \, d\mu(\alpha). \tag{4.59}$$

is called the stationary light transport equation*, also briefly denoted as the* stationary LTE*, or* SLTE*, for an illustration see Figure 4.10.*

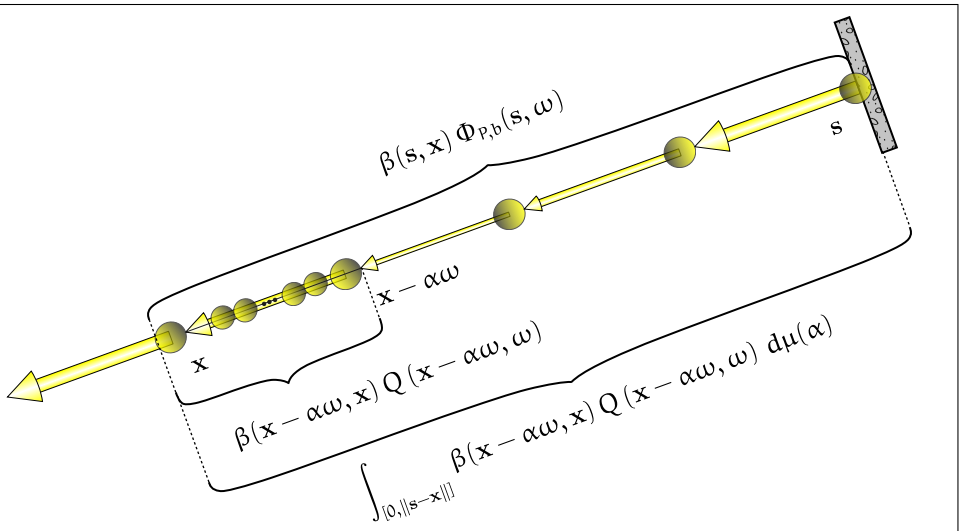


FIGURE 4.10: THE INTEGRAL FORM OF THE STATIONARY LIGHT TRANSPORT EQUATION. The number of particles at the inner point $\mathbf{x} \in \mathcal{V}^o$ is composed of two components. First, the particles emanating from point $\mathbf{s}$ on a surface that is nearest to $\mathbf{x}$ in direction $-\omega$ reduced by possible out-scattering and absorption processes on the way to $\mathbf{x}$, thus $\beta(\mathbf{s},\mathbf{x})\epsilon_b(\mathbf{s},-\omega)$. The other component is found by looking backwards along the ray between $\mathbf{x}$ and $\mathbf{s}$. At each point $\mathbf{x}-\alpha\omega, 0 \leq \alpha \leq d_\partial(\mathbf{x},\omega)$, we add to the volumetric emission the in-scattered particles at that point and we adjust this amount by volumetric effects of out-scattering and absorption as it travels from that point back to $\mathbf{x}$.

The first term on the right-hand side of the equal sign describes the radiance coming from surface point $\mathbf{s}$, which is nearest to an inner volume point $\mathbf{x}$ in direction ω_o, attenuated by possible out-scattering and absorption processes on the way to $\mathbf{x}$. The second term describes the radiance emitted or scattered in direction ω_o at all points $\mathbf{x}'$ between $\mathbf{x}$ and $\mathbf{s}$ and attenuated by out-scattering and absorption processes on the way to $\mathbf{s}$.

Writing out the SLTE in full, where we use the boundary radiance function L_b defined in accordance with Equation (4.49) and the gain function Q_o from Equation (4.32), then we can get the *stationary light transport equation expressed in terms of incident radiance*:

DEFINITION 4.3 (The Stationary Light Transport Equation Expressed in Incident Radiance, SLTE) *Let* $\mathbf{s}$ *be a point on a surface* $M \in \partial\mathcal{V}$, $\mathbf{x},\mathbf{x}' \in \mathcal{V}^o$ *are inner points of a participating medium,* $L_i(\mathbf{s},\omega_i')$ *and* $L_i(\mathbf{x}',\omega_i')$ *are the incident radiance at surface point* $\mathbf{s}$, *respectively at point* $\mathbf{x}' = \mathbf{x} - \alpha\omega_o$ *within a participating medium coming from direction* ω_i', *attenuated by the path absorption function* β, *then the* stationary

∂V (41) $\mathcal{V}^o$ (41) β (292)

light transport equation expressed in terms of incident radiance *has the form:*

$$\begin{aligned} & L_i(\mathbf{x},\omega_i) \\ = \;& \beta(\mathbf{s}\to\mathbf{x})\left(\epsilon_b(\mathbf{s},\omega_o)+\int_{S^2(\mathbf{s})}\kappa_b(\mathbf{s},\omega_i'\to\omega_o)\,L_i(\mathbf{s},\omega_i')\,d\sigma_{\mathbf{s}}(\omega_i')\right)+ \\ & \int_{[0,\|\mathbf{s}-\mathbf{x}\|]}\beta(\mathbf{x}'\to\mathbf{x})\left(\epsilon(\mathbf{x}',\omega_o)+\int_{S^2(\mathbf{x}')}\kappa(\mathbf{x}',\omega_i'\to\omega_o)\,L_i(\mathbf{x}',\omega_i')\,d\sigma_{\mathbf{x}'}(\omega_i')\right)d\mu(\alpha) \end{aligned} \tag{4.60}$$

Using the formulas (4.50) and (4.51), then the stationary light transport equation can also be written as:

$$\begin{aligned} & L_i(\mathbf{x},\omega_i) \\ = \;& e^{-\left(\int_{[0,\|\mathbf{s}-\mathbf{x}\|]}\sigma_t(\mathbf{x}',\omega_o)\,d\mu(\alpha)\right)}\,\epsilon_b(\mathbf{s},\omega_o)+ \\ & e^{-\left(\int_{[0,\|\mathbf{s}-\mathbf{x}\|]}\sigma_t(\mathbf{x}',\omega_o)\,d\mu(\alpha)\right)}\int_{S^2(\mathbf{s})}\kappa_b(\mathbf{s},\omega_i'\to\omega_o)\,L_i(\mathbf{s},\omega_i')\,d\sigma_{\mathbf{s}}(\omega_i')+ \\ & \int_{[0,\|\mathbf{s}-\mathbf{x}\|]}e^{-\left(\int_{[0,\alpha]}\sigma_t(\mathbf{x}',\omega_o)\,d\mu(\alpha)\right)}\,\epsilon(\mathbf{x}',\omega_o)d\mu(\alpha)+ \\ & \int_{[0,\|\mathbf{s}-\mathbf{x}\|]}e^{-\left(\int_{[0,\alpha]}\sigma_t(\mathbf{x}',\omega_o)\,d\mu(\alpha)\right)}\left(\int_{S^2(\mathbf{x}')}\kappa(\mathbf{x}',\omega_i'\to\omega_o)\,L_i(\mathbf{x}',\omega_i')\,d\sigma_{\mathbf{x}'}(\omega_i')\right)d\mu(\alpha). \end{aligned} \tag{4.61}$$

4.2 BIDIRECTIONAL DISTRIBUTION FUNCTIONS

Using the particle model we initially discussed the emission, absorption, and scattering behavior of abstract particles at points in the Euclidean space incorporating potential interaction with a participating medium. So, all these phenomena together characterize the transport of abstract particles in participating media. Mathematically, this was expressed in form of transport equations, so-called Fredholm integral equations of the 2^{nd} kind. Now, we want to turn our attention more closely to the following two questions: How can we mathematically formulate, based on the principles of physics, the process of interaction of light—thus, reflection, refraction, and transmission of light—at surfaces in a vacuum or in a participating medium? And: How can we involve these results into the abstract formulas of the light transport equations?

Section 4.1.1
Section 4.1.2
Fredholm IE 2^{nd} kind (127)
Section 4.1.3

For that purpose, we first discuss the *principles of geometric optics* a little bit in more detail, that is, we talk about *reflection* and *refraction* of light at object surfaces and discuss its interaction with various materials. In principle, the behavior of light can mathematically be described by so-called *absorption, emission*, and *scattering functions*, characterizing the involved medium or the object surfaces in the scene. Somehow combined and used as parameters, it is then possible to simulate the interaction of light

Section 4.2.1

with object surfaces or within participating media in a mathematical way. The only thing we have to do is, to replace the theoretical scattering kernel in our light transport equations by these new constructs. The idea behind it is the construction of a high dimensional function, a so-called *bidirectional scattering-surface reflectance-distribution function*. It should offer information about the quantity of light arriving at a point of the observed surface and given off again at another point due to the scattering properties of the material. If we then restrict this model, which not only describes scattering but also subsurface scattering, further, such that the reflected light ray also leaves the surface at the point, where it arrives, then we ultimately get a simple mathematical formulation of scattering properties of materials as well: the concept of the BSDF. Finally, we then extend the concept of the BSDF to the case where we also consider light transport in participating media, which results in the construct of the *phase function*.

Section 4.2.2
Section 4.2.3
Section 4.2.4

4.2.1 PRINCIPLES OF GEOMETRIC OPTICS AS BASIS FOR BIDIRECTIONAL DISTRIBUTION FUNCTIONS

Normally you would have to use Maxwell's equations [44, Ditchburn 1991], [27, Born & Wolf 1999] for plane or spherical waves to explain the propagation of light through a scene, but there are other much more simpler models. Thus, *geometric optics*, often also referred to as *ray optics*, describes light propagation in terms of so-called *light rays*. Such a light ray is a theoretical construct that can be seen as the normal of a wavefront in direction of the propagation of light, see Figure 4.11. Physically interpreted as an infinitely thin light bundle starting at surface point $\mathbf{s}$ and going in direction $\omega \in S^2(\mathbf{s})$, a ray $\mathbf{r}$,

Ray (11)

$$\mathbf{r} = \mathbf{s} + \alpha\omega, \qquad \alpha \in \mathbb{R}^{>0} \tag{4.62}$$

can be used as an abstraction to approximately model how light will propagate.

Geometric optics provides rules for propagating a ray through a system where the path taken by the light ray indicates how the associated wave will propagate. Although it does not work for optical effects such as diffraction, interference, or polarization, as well as scattering of light at particles within a participating medium, the principle of the light ray is an enormous simplification for describing optical phenomena. Geometric optics is a good approximation when the wavelength of light is very small compared with the size of structures with which the light interacts.

For the following discussion of interaction of light at object surfaces, we assume that light propagates only on straight lines. We will not consider any wave specific effects such as diffraction, where light bends around objects. We will also not consider media with varying indices of refraction. Furthermore, we assume that light travels at infinite speed

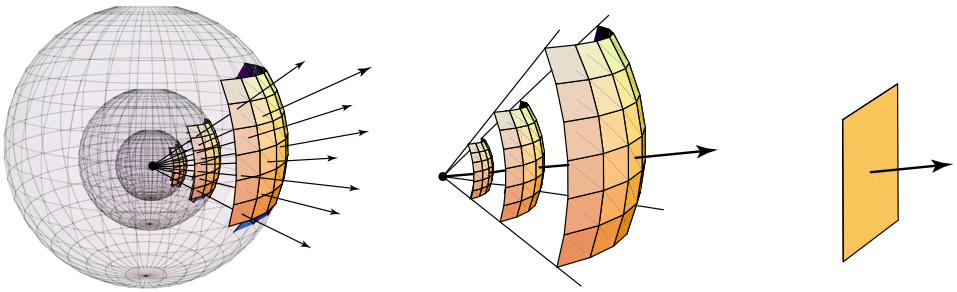


FIGURE 4.11: THE CONCEPT OF LIGHT RAY. A light ray defined as the normal in direction of a spherical wave. Spherical waves, a sector of a spherical wave, and a plane wave with a normal in direction of its propagation.

through a medium and that it is not influenced by other physical factors such as gravity or magnetic fields.

We begin this section with a short overview of the interaction of light with various materials. Afterwards we devote our interest to the most relevant light phenomenon in computer graphics, the reflection of light at object surfaces, and we discuss the physical effect of refraction of light at interfaces between to different media. As the interaction of light with planar interfaces between two substances is strongly related to the Fresnel equations, we also introduce the concepts of reflectance and transmittance of light which allows to determine the amount of incident light that is reflected or refracted at an object surface.

Section 4.2.1.1
Section 4.2.1.2
Section 4.2.1.3

4.2.1.1 INTERACTION OF LIGHT WITH VARIOUS MATERIALS

From the variety of different lighting phenomena that can be observed in nature, we are mainly interested in the interaction of light at surfaces, as well as with particles in participating media, namely: The interaction of light at boundaries between different types of media and the scattering and absorption of light within media or materials of real world.

Materials and media found in nature can roughly be characterized as being homogenous or non-homogenous. While homogenous substances have a constant composition with the same optical properties, non-homogenous are composed of two or more different types of homogenous materials or media. So, non-homogenous substances have different optical properties, which mainly depend on their composition. We coarsely distinguish between two classes of homogenous substances: *opaque* and *transparent* materials or media.

Typical opaque materials are conductors, such as iron, copper, or alluminium, but also some types of plastic or wood, which are bad conductors. A substance is defined to be opaque if it prevents light from passing through the material or the medium. Here, light is *reflected* in any direction, depending on the incident direction and the material of the surface, see Figure 4.12. Contrary to opaque substances, light can pass through transparent materials or media. The phenomenon of light traveling trough the material is called *transmission*, and the process of the abrupt change of direction when the beam enters into the other medium is called *refraction*. This effect can be observed at glass, water, or different types of liquids. Transparent materials and media are also called *dielectrics*, since they do not conduct electrical currents, see Figure 4.12.

Reflection (300)
Refraction (305)

Apart from opaque and transparent substances, there is also a class of materials and media that allow light to travel through a substance, where the light is diffusely *scattered* in all directions when it collides with particles or atoms in the substance. This type of material is called *translucent*. A similar effect like scattering in translucent materials occurs when a light beam travels through a non-homogenous medium. Here, a light beam is also scattered in all directions if it collides with particles of the medium, resulting in the diffusion of the light beam. Scattering processes do not only depend on the wavelength of the individual photons, but they are also dependent on the form and the size of the suspended particles, for a detailed discussion, see Section 4.2.4. Examples of translucent substances are paper, wax, snow, and smoke, see Figure 4.12.

Scattering (375)

Another phenomenon of light that can be observed when light travels through a material is absorption. Absorption corresponds to the collision of a photon with an atom of the material where the photon is distracted and its energy is stored as heat in the material, see Figure 4.12.

4.2.1.2 REFLECTION OF LIGHT

When light strikes the surface of an object it encounters a great network of close-fitting atoms, where a great amount of the original light ray will be scattered backwards. This effect is called *reflection*. Generally speaking, reflection is the change in direction of a light ray in particular at an interface between two different media so that the ray returns in its original direction. In computer graphics we distinguish coarsely between five types of reflection: *specular*, *diffuse*, *mixed*, *retro-reflective*, and *gloss reflection*.

DEFINITION 4.4 (Ideal Specular Reflection) Ideal specular reflection *is the mirror-like reflection of light at a perfectly smooth surface, where a light ray, incoming from a single direction* ω_i*, is reflected into a single outgoing direction* $M_{\mathbf{N}}(\omega_i)$*. As shown in Figure 4.13, for the reflected ray* $\omega_r \stackrel{\text{def}}{=} M_{\mathbf{N}}(\omega_i)$ *it holds:*

$$M_{\mathbf{N}}(\omega_i) = 2\langle \mathbf{N}(\mathbf{s}), \omega_i \rangle \, \mathbf{N}(\mathbf{s}) - \omega_i, \tag{4.63}$$

where $\mathbf{N}(\mathbf{s})$ *is the surface normal at point* $\mathbf{s}$*.*

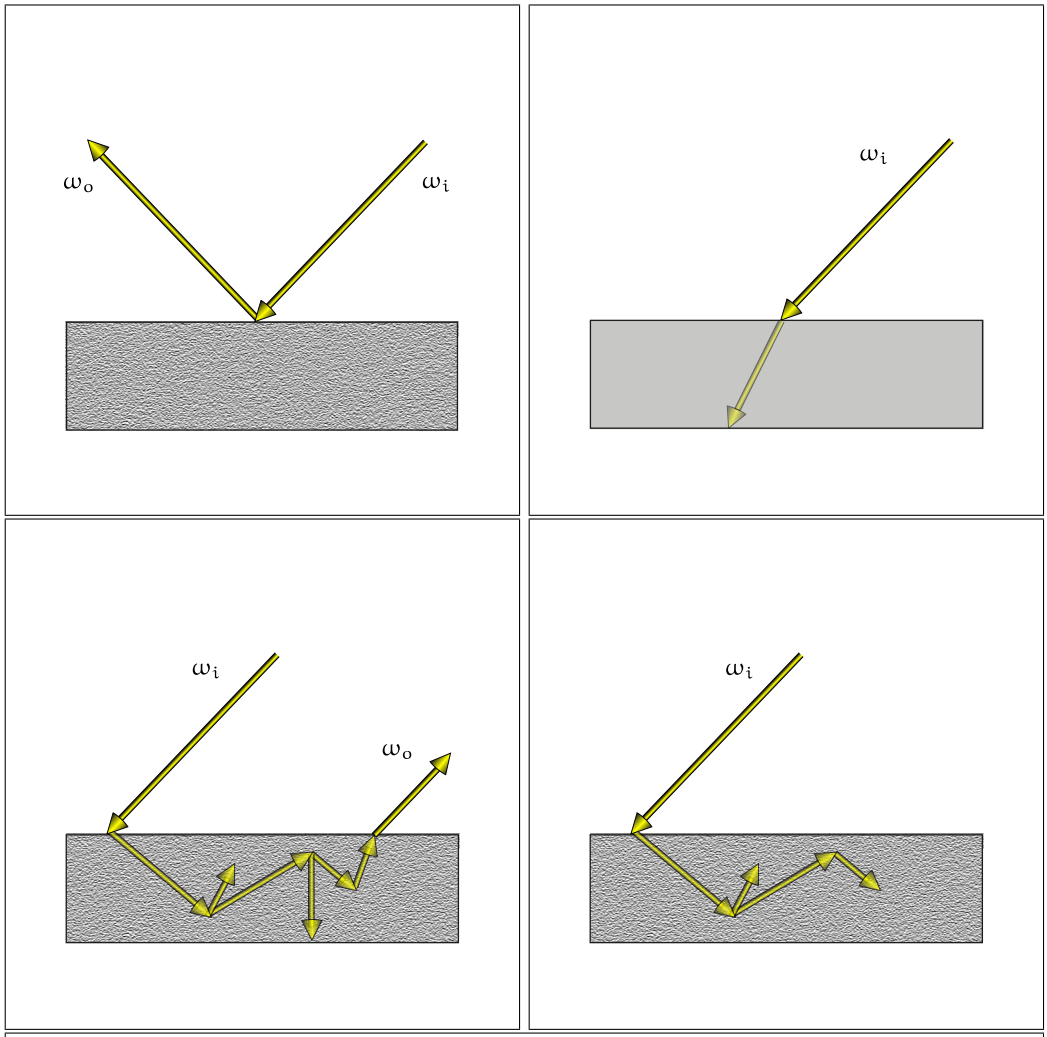


FIGURE 4.12: INTERACTION OF LIGHT WITH VARIOUS MATERIALS. Depending on the underlying material or the involved media, in computer graphics we coarsely distinguish between four different types of light interaction at materials or within participating media: *Reflection*, *refraction* or *transmission*, *scattering*, and *absorption*.
Opaque materials prevent light from passing through the material, so, an incident light ray is reflected at the surface. While light can pass collision free through transparent materials—the light ray is refracted when it arrives at an object surface—in translucent materials, a beam of photons is scattered in all directions when they collide with particles or atoms in the substance. But a light ray can also be absorbed within a medium, if the involved photons are distracted from their original direction of motion where their energy is stored as heat in the material.

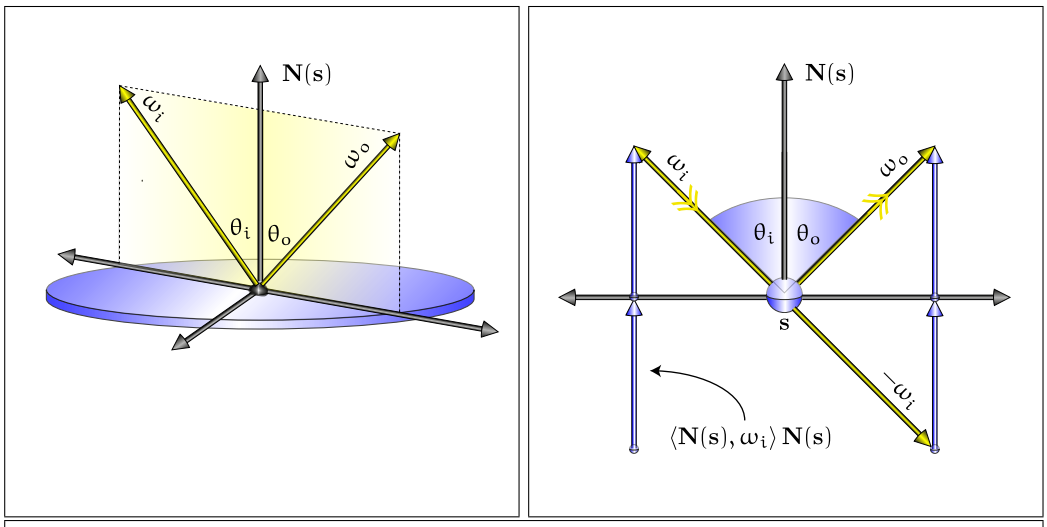


FIGURE 4.13: THE GEOMETRY FOR IDEAL SPECULAR REFLECTION AND THE CALCULATION OF THE REFLECTED RAY. The incident ray, the exitant ray, and the normal vector at surface point $\mathbf{s}$ lie in the same plane. The right image shows like the reflected ray $\omega_r = \omega_o$ can efficiently be computed via the projection of ω_i onto the normal vector at point $\mathbf{s}$ and direction ω_i.

This property is described by the law of reflection, *which states that the direction of the incident ray* ω_i *and the direction of the reflected ray* $\omega_r \equiv \omega_o \stackrel{\text{def}}{=} M_{\mathbf{N}}(\omega_i)$ *makes the same angle with respect to the surface normal, that is,*

$$\theta_o = \theta_i. \tag{4.64}$$

An additionally defining characteristic of specular reflection is that ω_i, ω_o *and the surface normal* $\mathbf{N}(\mathbf{s})$ *lie in the same plane, i.e. the three vectors are linear dependent, see Figure 4.13.*

Linear Dependent (857)

Ideal specular reflection occurs at surfaces, that are perfectly smooth, such as mirrors or very highly polished metals. Such surfaces can be seen as composed of many tiny microfacets that are perfectly aligned with surface normals that point in the same direction and that reflect light in a simple, predictable way. At these surfaces, ideal specular reflection generates images that are, due to the reflection law, upright and have the same distance behind the surface as the objects are in front of the surface. Slight reflection also occurs at interfaces between two different media when light travels from a medium such as air into a medium such as water or glass. Here, as we will see further below, a tiny fraction of light is reflected from the interface while the remainder is refracted. Ideal specular reflection is, as the name already expresses, not a real reflection behavior of surfaces that occurs in nature.

Fresnel Equations (306)

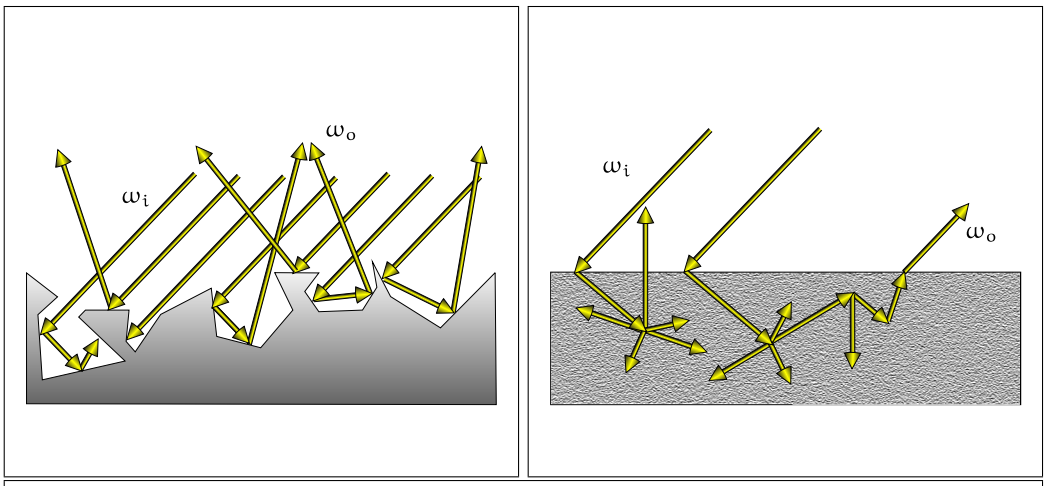


FIGURE 4.14: SURFACE GEOMETRIES FOR DIFFUSE REFLECTION. A light beam falling on a diffuse surface can bounces off in all directions due to the microscopic irregularities at a surface. Ideal diffuse reflection can also result of a change of propagation direction of light rays due to collisions with particles and atoms within the material.

Another type of an ideal reflection effect is *ideal diffuse reflection*.

DEFINITION 4.5 (Ideal Diffuse Reflection) *A material surface is called* ideal diffuse reflective *or simply* ideal diffuse *if it reflects light uniformly in all directions with the same energy.*

Diffuse reflection occurs if light strikes an ideal rough or an ideal granular surface, that can be seen to be composed of many randomly distributed microfacets with surface normals that are uniformly distributed over the entire hemisphere, see Figure 4.14. When a light beam hits such a surface, it is splitted in infinitely many light rays, which are reflected in a random fashion in all directions due to the microscopic irregularities at the surface, or they penetrate into the material, where they collide with particles and atoms and their original direction of motion is changed. This means, that an image can not be formed. Diffuse reflection can be seen as the complement to ideal specular reflection, i.e. if a surface is completely non-specular, the reflected light will be evenly spread over the entire hemisphere surrounding the surface, see Figure 4.15.

Diffuse reflection is the reason why objects illuminate other objects in the surrounding area. This also ensures that light reflected from objects that are not shiny or specular, such as paper, walls, or ground, reaches areas not directly in view of a light source. An often used model in computer graphics for simulating diffuse reflection is the *Lambertian illumination model*, in which light is equally reflected in all directions.

Lambert's Reflection Model (350)

In real world, reflection of light is neither ideal specular nor ideal diffuse. Due to [68,

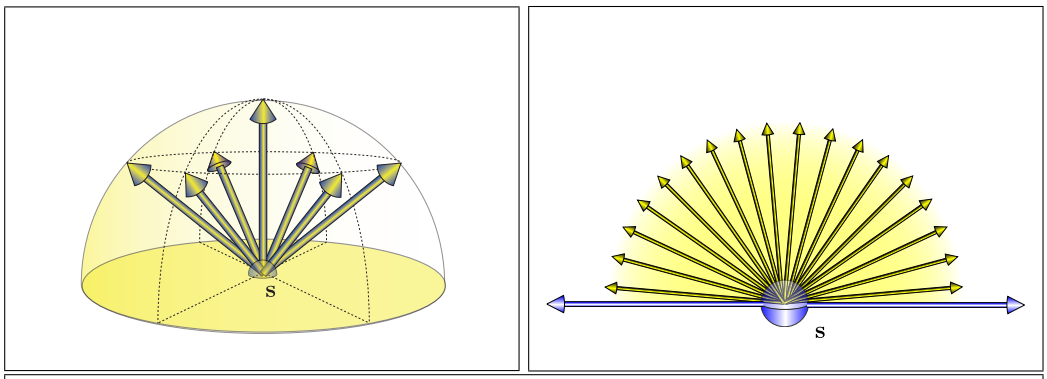


FIGURE 4.15: IDEAL DIFFUSE REFLECTION AT A SURFACE. A light beam incident at point s on a diffuse surface is scattered diffusely in all directions over the hemisphere $\mathcal{H}^2_{\mathrm{i}}(\mathbf{s})$. The reflected light will be evenly spread over the entire hemisphere surrounding a surface point.

Glassner 1995], it can rather be described as a combination of these different reflection types. Therefore, in the following we will define three such reflection effects, which are useful in rendering algorithms.

DEFINITION 4.6 (Mixed Reflection) Mixed reflection *is the combination of ideal specular reflection and ideal diffuse reflection.*

A material that simulates mixed reflection behavior can be modeled by a weighted combination of ideal diffuse and ideal specular components.

DEFINITION 4.7 (Retro Reflection) *The reflection behavior of a material is called* retro-reflective *if the energy of the incident light ray is reflected in directions close to itself over a wide range of incident directions.*

Almost all materials are a little bit retro-reflective, but we call only those that retro-reflect most of their incident energy as *retro-reflectors* [68, Glassner 1995]. Last but not least, we present *gloss reflection*, see Figure 4.16.

DEFINITION 4.8 (Gloss Reflection) *A material surface is referred to as* gloss *if its reflection behavior can be seen as a combination of mixed reflection and a mirror-like appearance of a rough surface. Here, light from an infinitesimal thin light beam is scattered and spread into some finite solid angle typically around the perfect mirrored reflection of the incoming ray.*

In the above definition, a rough surface can be interpreted as a composition of not uniformly distributed many micro-facets with surface normals that are distributed around

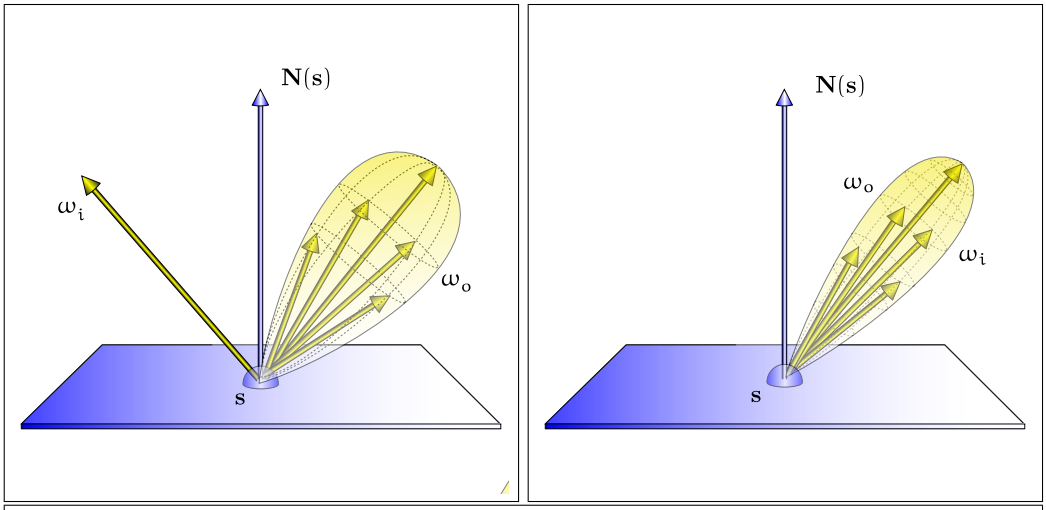


FIGURE 4.16: GLOSS AND RETRO-REFLECTIVE REFLECTION. At a gloss surface, an incident light beam is scattered and spread into some finite solid angle typically around the perfect reflection direction, where the degree of glossiness determines the lobe of set of possible reflected directions. At a retro-reflected surface, the reflected ray is close to the incident ray.

the average surface normal. Gloss reflection is generated by higher weighting of the specular reflection compared to the diffuse reflection part.

4.2.1.3 REFRACTION OF LIGHT

When light strikes the interface between two different media it encounters a great net of close-fitting atoms. These atoms scatter both a little fraction of light backwards and the bigger part of light in forward direction where the forward scattered direction is snapped off with respect to the original direction of the light ray. This effect occurs because the incident light wave changes its velocity and its wavelength when entering into the new medium. This deflection is referred to as *refraction*. It is dependent on both the two media involved and the direction of light transfer.

DEFINITION 4.9 (Ideal Specular Refraction) Ideal specular refraction *or simply* specular refraction *occurs at interfaces between two media when light travels from a medium with index of refraction* η_i *into another medium with index of refraction* η_t, *where*

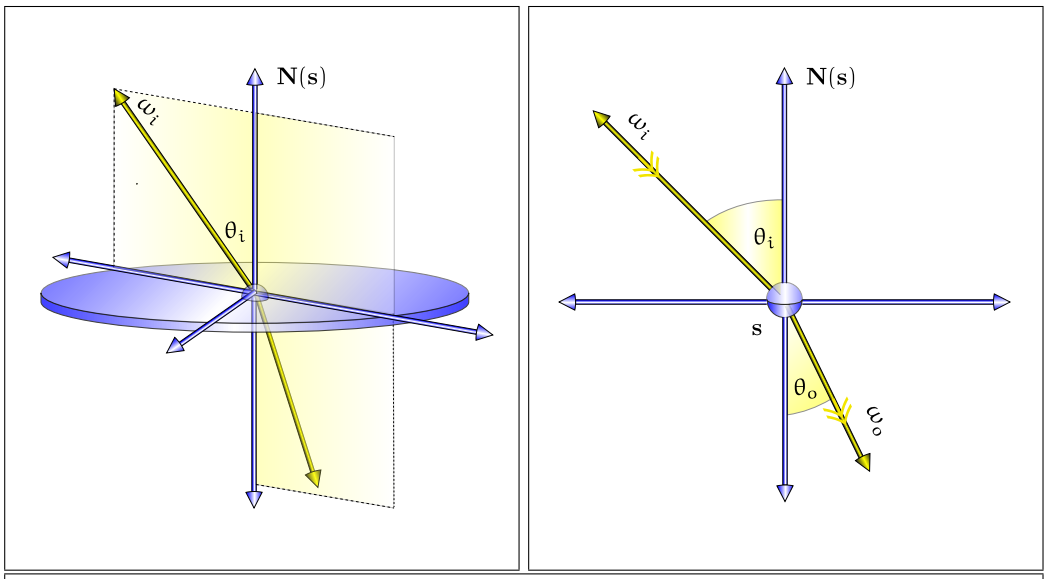


FIGURE 4.17: THE GEOMETRY OF SPECULAR REFRACTION. The transmitted ray ω_t is refracted in direction to the surface normal if the refractive index $\eta_t > \eta_i$ else, it is refracted away from the surface normal. Additionally, the transmitted ray ω_t and the incoming ray ω_i lie in the same plane as the surface normal $\mathbf{N}(\mathbf{s})$.

it holds: $\eta_i \neq \eta_t$. *The refracted ray* $R_{\mathbf{N}}(\omega_i)$ *can be written as:*

$$R_{\mathbf{N}}(\omega_i) = -\frac{\eta_i}{\eta_t}\omega_i + \mathbf{N}\left(\frac{\eta_i}{\eta_t}\langle \mathbf{N}(\mathbf{s}), \omega_i\rangle - \sqrt{1 - \left(\frac{\eta_i}{\eta_t}\right)^2 (1 - \langle \mathbf{N}(\mathbf{s}), \omega_i\rangle^2)}\right), \quad (4.65)$$

where $\mathbf{N}(\mathbf{s})$ *is the surface normal at point* $\mathbf{s}$ *in direction to the incident medium, see Figure 4.17. This property is described by* Snell's law, *which states that the angle of the incident ray* ω_i *and the angle of the transmitted ray* $\omega_t \stackrel{\text{def}}{=} R_{\mathbf{N}}(\omega_i)$ *with respect to* $\mathbf{N}(\mathbf{s})$ *satisfies the following equation*

$$\eta_t \sin(\theta_t) = \eta_i \sin(\theta_i). \quad (4.66)$$

An additionally defining characteristic of specular refraction is that ω_i, ω_o *and the surface normal* $\mathbf{N}(\mathbf{s})$ *lie in the same plane, i.e. the three vectors are linear dependent.*

Linear Dependent (857)

THE FRESNELS EQUATIONS. Even when a smooth surface exhibits only specular reflection or specular refraction not all of the light is necessarily reflected or refracted. Solving Maxwell's equations for a light wave traveling from a medium of a given refractive index η_i

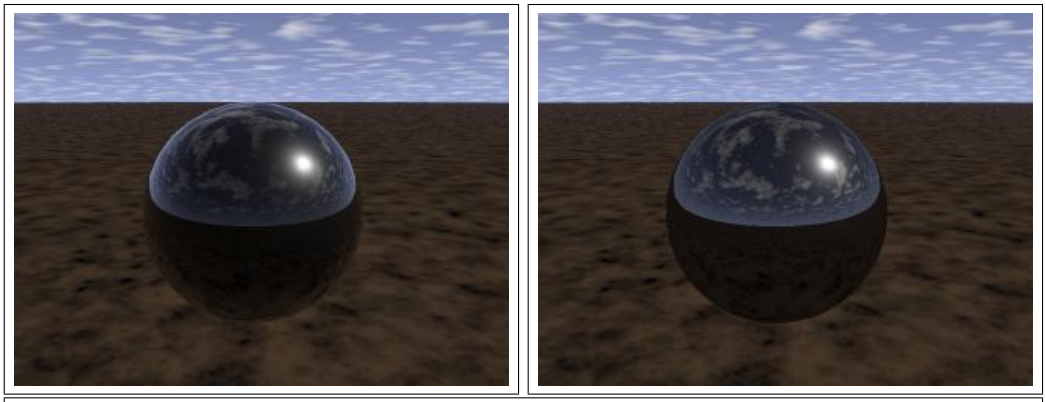


FIGURE 4.18: FRESNEL EFFECT FOR DIELECTRICA. Two black and opaque spheres made of a dielectrica. The Fresnel effect reduces the amount of reflected light for normally incident light. The other sphere uses a constant reflectance with angle. Image courtesy of Stephen H. Westin, Cornell University.

into another medium with refractive index η_t while striking a polished boundary between the two media results in the so-called *Fresnel equations*. These equations can then be used to predict how much light is reflected, and how much is refracted in the situations described above. The fresnel effect, as the visible result of the solution to the Fresnel equations, is the observation that things get more reflective at grazing angles, see Figure 4.18. It can be interpreted as direct consequence of the electromagnetic nature of light considered as a wave consisting of an an electric and a magnetic field component perpendicular to each other and the fact that the electric field cannot be discontinuous even when the wave meets a discontinuity in the refractive index at a polished surface.

In physics, the Fresnel equations are derived directly from the continuity conditions of the electric and the magnetic field components of an incident electromagnetic wave as well as the normal component at the surface boundary between the media, see [27, Born & Wolf 1999] or [80, Hecht 2001]. Due to the fact that the Fresnel equations are solutions of the Maxwell equations at smooth boundaries, they incorporate the *polarization of light*, that is, the orientation of the *electric field vector* with respect to the incident plane, spanned by the surface normal as well as the incident, and reflected wave. It is the Fresnel effect that is responsible for the polarization of the outgoing wave after a specular interaction with a smooth surface, even if the incoming wave is unpolarized.

Now, the orientation of the electric field vector of the incident light wave can be parallel or perpendicular to the plane of incidence. That is, we have to distinguish between two forms of Fresnel equations: the Fresnel equations, where the polarization of the incident light is parallel to the surface, and the Fresnel equations, where the polarization of the incident light is perpendicular to the surface. Additionally, as conductors, unlike dielectrics, do not transmit light—some of the incident light is absorbed by the

material and transformed into heat—there are also two types of Fresnel equations: one for *dielectrics* and one for *conductors*. So, we have four equations resulting in four Fresnel factors used to define two Fresnel reflectance types: the *Fresnel reflectance* and the *Fresnel transmittance*.

DEFINITION 4.10 (Fresnel Equations for Dielectrics) *Let us assume an incident light wave coming from direction* ω_i *is reflected in direction* $\omega_r \stackrel{\text{def}}{=} M_{\mathbf{N}}(\omega_i)$ *at a smooth non-conducting surface within a medium with refractive index* η_i*, where some of the amount of light is refracted in direction* $\omega_t \stackrel{\text{def}}{=} R_{\mathbf{N}}(\omega_i)$ *in a medium with refractive index* η_t*. Then, the* Fresnel equations for dielectrics *are given by:*

$$r_{\parallel} = \frac{\eta_t \cos\theta_i - \eta_i \cos\theta_t}{\eta_t \cos\theta_i + \eta_i \cos\theta_t} \tag{4.67}$$

as well as

$$r_{\perp} = \frac{\eta_i \cos\theta_i - \eta_t \cos\theta_t}{\eta_i \cos\theta_i + \eta_t \cos\theta_t}, \tag{4.68}$$

where $r_{\parallel}$ *is also called the* Fresnel reflectance for parallel polarized light *and* $r_{\perp}$ *is the* Fresnel reflectance for perpendicular polarized light*.*

DEFINITION 4.11 (Fresnel Equations for Conductors) *Let us assume an incident light wave coming from direction* ω_i *is reflected in direction* ω_o *at a smooth conducting surface with refractive index* η*. Let furthermore* κ *be the* absorption coefficient *of the conductor, thus the amount of incident light that is absorbed, that is, which is not reflected in direction* ω_o*. Then, the* Fresnel equations for conductors *are given by:*

$$r_{\parallel} = \frac{(\eta^2 + \kappa^2)\cos^2\theta_i - 2\eta\cos\theta_i + 1}{(\eta^2 + \kappa^2)\cos^2\theta_i + 2\eta\cos\theta_i + 1} \tag{4.69}$$

as well as

$$r_{\perp} = \frac{(\eta^2 + \kappa^2) - 2\eta\cos\theta_i + \cos^2\theta_i}{(\eta^2 + \kappa^2) + 2\eta\cos\theta_i + \cos^2\theta_i} \tag{4.70}$$

where $r_{\parallel}$ *is also called the* Fresnel reflectance for parallel polarized light *and* $r_{\perp}$ *is also called the* Fresnel reflectance for perpendicular polarized light*.*

In simple ray tracers, the Fresnel reflectance is often controlled by a so-called reflectivity parameter, commonly a constant value valid over the entire surface. But as we can see from the Formulas (4.67) - (4.70), the Fresnel equations are at least directionally dependent on the incident direction. Under the assumption that light is unpolarized, we can define a quantity, called *Fresnel reflectance*, F_r, by the arithmetic average of the parallel and perpendicular Fresnel reflectance, namely by:

TABLE 4.1: INDICES OF REFRACTION FOR A VARIETY OF MEDIA. The refractive index, η, is defined as $\eta \stackrel{def}{=} \frac{c}{c_m}$, where c is the speed of light in a vacuum and c_m is the speed of light in the medium. Table data are copied from [158, Pharr & Humphreys 2004].

Medium	Index of Refraction
Vacuum	1.0
Air at sea level	1.00029
Ice	1.31
Water	1.333
Fused quartz	1.46
Glass	1.5 - 1.6
Sapphire	1.77
Diamond	2.42

TABLE 4.2: REPRESENTATIVE MEASURED VALUES OF η AND κ FOR A FEW CONDUCTORS. Table data are copied from [158, Pharr & Humphreys 2004].

Object	η	κ
Gold	0.370	2.820
Silver	0.177	3.638
Copper	0.167	2.63
Steel	2.485	3.433

DEFINITION 4.12 (The Fresnel Reflectance, F_r) *Let $r_\parallel$ and $r_\perp$ denote the parallel, respectively, the perpendicular Fresnel reflectance introduced in the Definitions 4.10 and 4.11, then the* Fresnel reflectance, F_r, *is defined as:*

$$F_r : S^2 \longrightarrow \mathbb{R}^{\geq 0} \tag{4.71}$$

$$\omega_i \longmapsto F_r(\omega_i) \stackrel{def}{=} \frac{r_\perp^2 + r_\parallel^2}{2}. \tag{4.72}$$

Now, since dielectrica also transmit light and the Fresnel effect does not include absorption, based on the principle of conservation of energy, we can also introduce the so-called *Fresnel transmittance* for dielectrica by:

Conservation of Energy (332)

DEFINITION 4.13 (The Fresnel Transmittance, F_t) *Since dielectrica also transmit light, due to conservation of energy, we can define the so-called Fresnel transmittance, F_t, for dielectric by:*

$$F_t : S^2 \longrightarrow \mathbb{R}^{\geq 0} \tag{4.73}$$

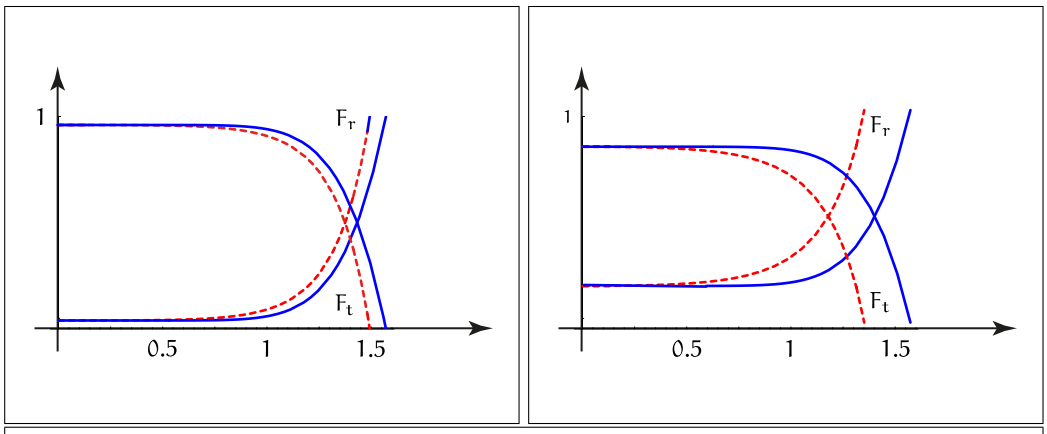


FIGURE 4.19: SCHLICK'S APPROXIMATION OF FRESNEL FUNCTIONS F_r AND F_t. The Fresnel functions F_r and F_t with corresponding approximations according to Schlick. The graphs of the functions show the transition from a vacuum ($\eta_1 = 1$) to glass ($\eta_2 = 1.5$) and respectively diamond ($\eta_2 = 2.42$).

$$\omega_i \longmapsto F_t(\omega_i) \stackrel{def}{=} 1 - F_r(\omega_i). \tag{4.74}$$

REMARK 4.6 (Schlick's Approximation for the Fresnel Equations of Dielectrica) *Used in a rendering algorithm, the computation of* F_r *and* F_t *requires, due to the evaluations of the cosines, enormous costs in computation time. Therefore, in [178, Schlick 1993] an approximation for Relation (4.72) is suggested as follows:*

$$F_r(\theta_i) \stackrel{def}{=} F_r(0) + (1 - F_r(0))(1 - \cos\theta_i)^5. \tag{4.75}$$

Since the approximation error in Schlick's formula is less than the restriction to consider unpolarized instead of polarized light, see [141, Olano & al. 2002], the usage of Formula (4.75) in practice is nearly as good as the real Fresnel equations.

REMARK 4.7 *As we know from the phenomenon of ideal reflection, also ideal specular refraction occurs at interfaces, that are perfectly smooth, such as water surfaces, see Figure 4.20. Even at ideal refractive surface, not all of the incident light is refracted, but some of the light will be reflected in the mirrored directions. So, ideal specular refraction is, as the name already expresses, not a real refraction behavior of surfaces that does occurs in nature.*

TOTAL INTERNAL REFLECTION. Let us consider once more Formula (4.65) of the refracted ray $R_N(\omega_i)$ from Definiton 4.9. If the value under the square root is negative, which can

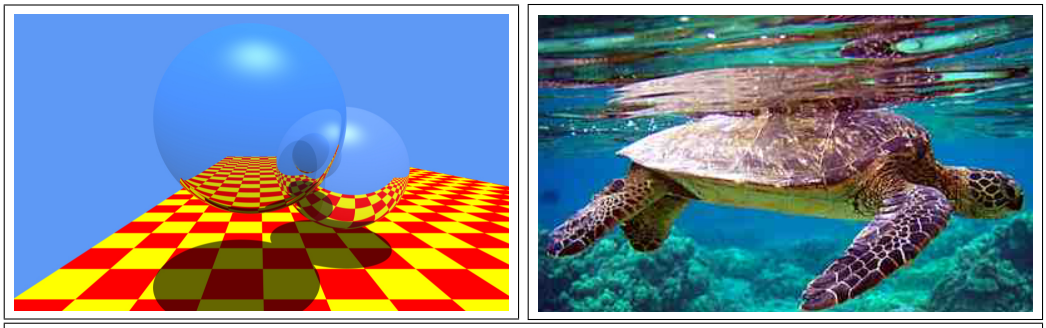


FIGURE 4.20: SPECULAR REFRACTION AND TOTAL INTERNAL REFLECTION. The original scene from Turner Whitted for visualizing refraction using ray tracing. Image courtesy of Paul Solt. The Green turtle, Chelonia mydas and his total internal reflection on the right hand side is a copy from Wikimedia Commons.

only occurs if $\eta_i > \eta_t$, then the incoming ray is not refracted, it is totally reflected into the incident medium and there can be no transited energy. This effect is called *total internal reflection*.

Total internal reflection occurs, when light, coming from a medium with a larger index of refraction to a material with lower index, arrives at the interface between the two media at an angle greater than the so-called *critical angle*, θ_c, given by:

$$\theta_c \stackrel{def}{=} \arcsin\left(\frac{\eta_t}{\eta_i}\right). \tag{4.76}$$

Under these conditions, the ray is totally reflected at the boundary back into the medium from where it comes. No light can pass through the boundary. A typical situation where this phenomenon can be observed is, when light passes from glass or water to air, see Figure 4.20, or when a beam of light passes through a prism.

4.2.2 THE MATHEMATICAL MODEL OF THE BIDIRECTIONAL REFLECTANCE-DISTRIBUTION FUNCTION

In Section 4.2.1.1 we have seen that, depending on the physical properties of a material or a medium, light interacts at object surfaces or at interfaces between participating media in different ways. This complicated light-matter dynamic depends on the physical properties of light as well as the physical composition and characteristic of the matter.

Considering objects of translucent materials such as skin, marble, snow, or wax then a light beam striking the surface of such an object enters the material and scatters around before leaving the surface at another position, see the lower left image in Figure 4.21. This behavior of a light beam is different from that striking a polished or a rough surface

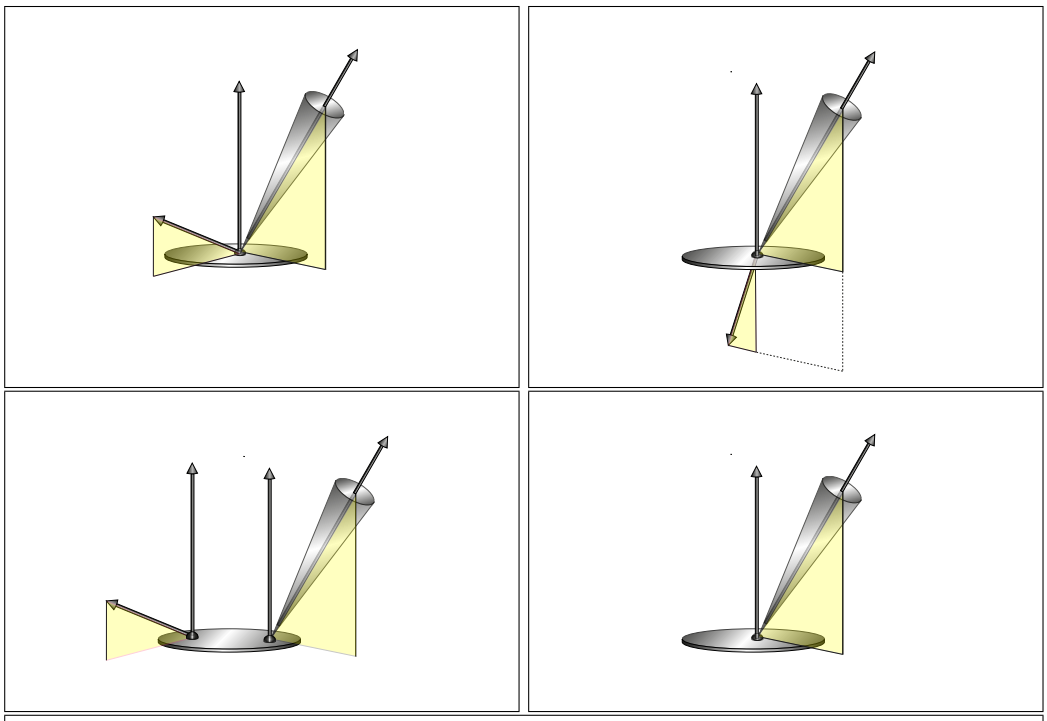


FIGURE 4.21: THE GEOMETRY OF INTERACTION OF LIGHT WITH MATERIALS. Light can be reflected at a surface point and it can be refracted at the same point where it enters into a material. Furthermore, it can also enter at a point into a material and leave the material at another point. Last but not least, light can enter into a surface but due to absorption within the material no light leaves the surface.

of metal, as shown in Section 4.2.1.1. Here, the incident light wave does not penetrate appreciably into the material but the largest part of incident electromagnetic energy is reflected specularly or diffusely at the same point. Yet another behavior can be observed, if a light beam interacts at the boundary between two media of different refraction indices, as for example between air and glass, or glass and water. The incident light wave will then be refracted at the boundary—depending on the refraction indices of the two involved media—and penetrates into the other medium. Apart from these phenomena, it is also possible, that a light beam on the whole is neither reflected nor refracted at an object surface. This can be observed if a light wave penetrates deep in a material and the largest part of its energy is absorbed, see Figure 4.21 and Figure 4.22.

Section 4.2.1.2
Section 4.2.1.3

Now, to involve all these scattering phenomena of light at the boundaries $\partial\mathcal{V}$ of a participating medium into our stationary light transport equation, we will try to described these phenomena mathematically by so-called *bidirectional reflectance-distribution functions*.

$\partial\mathcal{V}$ (41)
SLTEV (4.38)

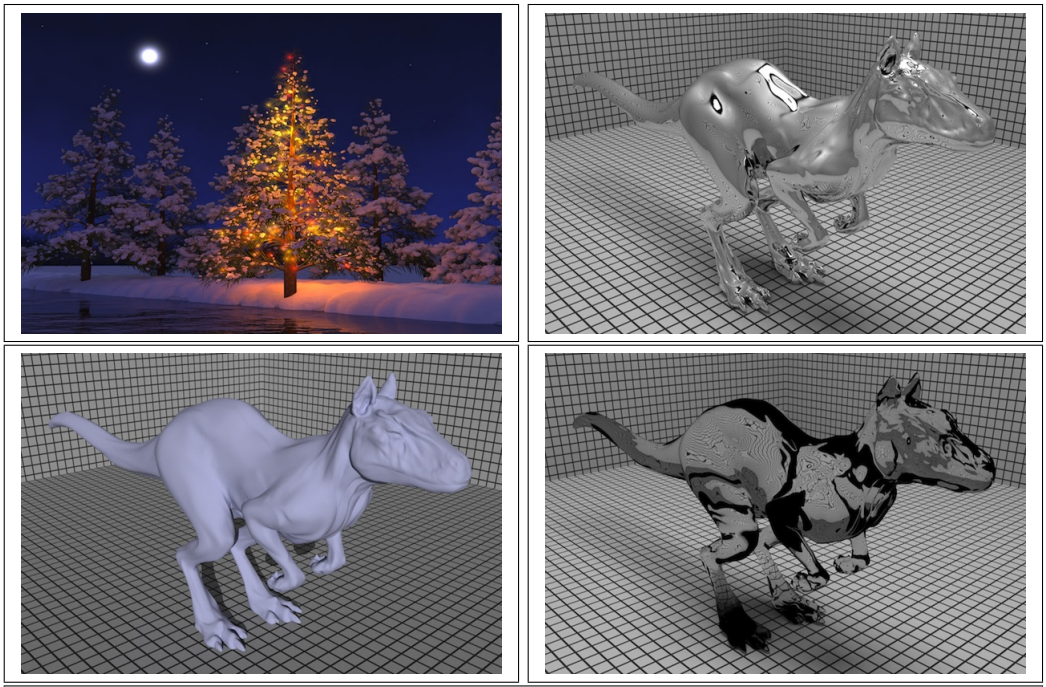


FIGURE 4.22: INTERACTION OF LIGHT WITH MATERIALS. Simulation of subsurface scattering in snow. The Killeroo model rendered via pbrt with ideal specular reflection, ideal diffuse reflection, and ideal specular transmission. Image Courtesy of Matt Pharr and Greg Humphreys.

As we will see, a bidirectional reflectance-distribution function can mathematically be interpreted as the kernel of a linear integral operator between two function spaces. In some cases, it is not a function in our usual sense, but rather a distribution, also called a *generalized function*, that makes sense only inside integrals. That is, bidirectional distribution functions will take the job of the theoretically introduced scattering kernels κ_b in the light transport equations, such as for example the SLTE from Definition 4.3.

Linear Integral Operator (130)
Distribution (117)

Based on Nicodemus' definition of the *bidirectional reflectance-distribution function*, we will derive in this section the *bidirectional scattering-surface reflectance-distribution function*. Also referred to as the *BSSRDF*, it is the fundamental quantity that usually characterizes the scattering properties of a material. In its generalized form, it is an element of a 12-dimensional function space, and can be considered as the most general description of light scattering at object surfaces. Due to its complexity—one has to be evaluated for 12 variables—the generalized BSSRDF is not really applicable in rendering algorithms. Hence, we present the BSSRDF in a form in which it is also used in many

Section 4.2.2.1
Function Space (28)

rendering algorithms for simulating subsurface scattering, that is, our BSSRDF does not account for light phenomena such as phosphorescence, fluorescence, and the wavelength of incoming light.

Now, in scenes without subsurface scattering also the simplified 8-dimensional BSSRDF is too costly to evaluate, hence, we additionally abstract from further variables occurring in a BSSRDF. This leads us to the class of the so-called BRDFs and BTDFs, thus the class of *bidirectional-reflectance* and *bidirectional-transmission distribution functions*. A BRDF and a BTDF can then be taken as the fundamental quantities for the optical characterization of an object. They are the descriptions of the interaction of light at object surfaces as well as at boundaries between media that will be used in any rendering algorithm. We then discuss the physical properties of BRDF and BTDF and introduce the concepts of *reflectance* and *transmittance*. Finally, we present the most important BRDF models used in rendering algorithms and show how BRDFs can be measured and represented.

Section 4.2.2.2
Section 4.2.2.3
Section 4.2.2.4
Section 4.2.2.5

Now, due to [135, Nicodemus & al. 1977], a bidirectional reflectance-distribution function *is a derivative, a distribution function, relating the irradiance to its contribution to the reflected radiance in another direction.*

Irradiance (257)
Radiance (250)

Such a function can mathematically be described as follows:

$$f_r(s, \omega_i \to \omega_o) \stackrel{def}{=} \frac{dL_o(s, \omega_o)}{dE_i(s, \omega_i)} = \frac{dL_o(s, \omega_o)}{L_i(s, \omega_i)\, d\sigma^{\perp}_s(\omega_i)}, \tag{4.77}$$

where s denotes a point on an object surface from $\partial\mathcal{V}$ and ω_i and ω_o correspond to incident and exitant directions. Note, the illumination comes from the differential solid angle $d\sigma^{\perp}_s(\omega_i)$, while we are measuring the reflected radiance only along a single direction ω_o.

$\partial\mathcal{V}$ (41)
$d\sigma^{\perp}$ (89)

The questions that now arise are: Why is a BRDF defined in this way? Would it not be better to define a BRDF as the ratio of the reflected radiance to incident radiance, or perhaps as the ratio of the reflected irradiance to incident irradiance, as illustrated in Figure 4.23. To answer these questions it is important to know how the physical quantity of light can be measured.

Irradiance (257)

4.2.2.1 SUBSURFACE SCATTERING AND THE BSSRDF

When light strikes a surface of any non-metallic material some of it penetrates into the material, and is there absorbed, or scattered hundreds of times around before it leaves the surface at a different location. Unlike absorption, which changes the light's amount but not its direction of propagation, scattering changes—due to any discontinuities within the material, such as air bubbles, other particles, or density variations—the direction but not the amount of light. This effect of light is called *subsurface scattering*. Typical

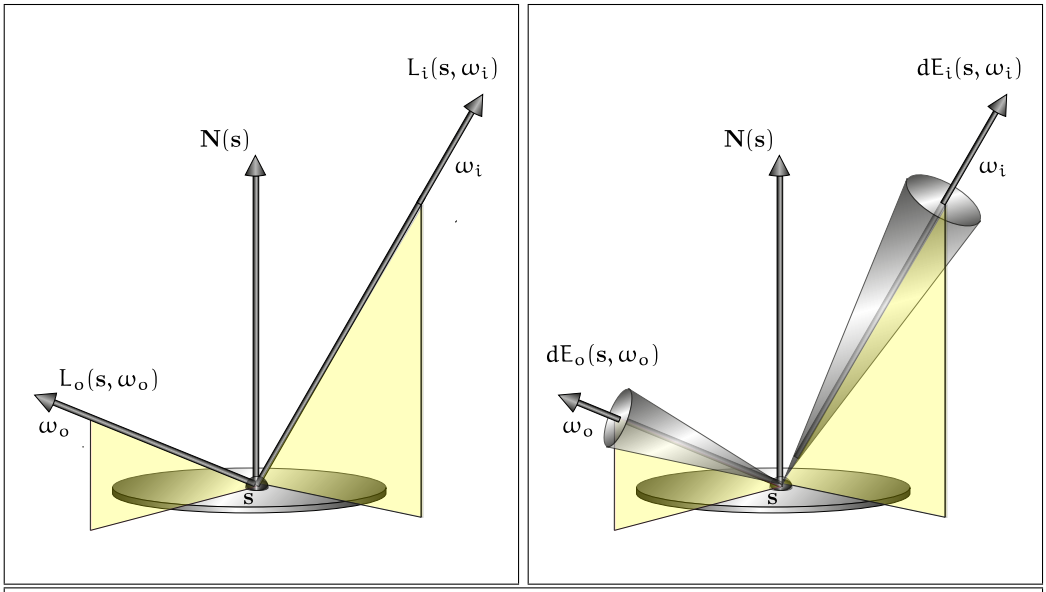


FIGURE 4.23: TWO DIFFERENT MODELS TO DEFINE THE BIDIRECTIONAL REFLECTANCE-DISTRIBUTION FUNCTION. Due to [135, Nicodemus & al. 1977], a bidirectional reflectance-distribution function is a derivative, a distribution function, relating the irradiance from direction ω_i at surface point s to its contribution to the reflected radiance in direction ω_o. Why is a BRDF defined in this way? Would it not be better to define a BRDF as the ratio of the reflected radiance to incident radiance, or perhaps as the ratio of the reflected irradiance to incident irradiance?

materials, where subsurface scattering occurs are translucent materials. Here some incident light reflects off the first surface as gloss, while some enters the material and undergoes multiple scattering within the material resulting in a diffuse pattern of reflectance. In a second interaction, light is scattered and transmitted through the object, emerging on a side in a diffuse pattern. As a result, color can be seen in both diffuse reflectance and transmittance, depending on how the object is viewed. Thus, translucent materials often have soft appearance where light bleeds through thin slabs of the material, see Figure 4.24.

Let us now consider Figure 4.25, it illustrates the geometry underlying the process of subsurface scattering. From our definition of flux it is evident that the incident flux from direction ω_i at a small infinitesimal surface patch $d\mu^2(s_i)$ is given by:

Flux (249)

$$d^2\Phi_i(s_i, \omega_i) \overset{(3.2)}{=} L_i(s_i, \omega_i)\, d\sigma^{\perp}_{s_i}(\omega_i)\, d\mu^2(s_i), \tag{4.78}$$

where $d\sigma^{\perp}_{s_o}$ corresponds to the projected solid angle as seen from point s_i.

Projected Solid Angle (89)

Depending on the surface properties, light particles then enter the surface, collide with the atoms in the material, and are subject to a variety of absorption and scattering processes in the material. Since absorbed photons leave the system, we are mostly

FIGURE 4.24: SIMULATION OF SUBSURFACE SCATTERING. These images show how a face model, milk, and a translucent marble bust can be rendered using the BSSRDF model. The box scene with a translucent white box, and the Utah teapot consisting of translucent material. Image courtesy of Henrik Wann Jensen, UCSD.

interested in the scattering processes, that is, we must observe the differential scattered radiance dL_o in direction (s_o, ω_o). The exitant radiance $L_o(s_o, \omega_o)$ leaving point s_o in direction ω_o can now be viewed as a composition of contributions $dL_o(s_i \to s_o, \omega_i \to \omega_o)$ from the incident flux from directions ω_i within the solid angle $d\sigma^{\perp}_{s_i}(\omega_i)$. A sensor that registers the surface from direction ω_o would measure the flux

Radiance (250)

$d\sigma^{\perp}$ (89)

$$d^3\Phi_o(s_o, \omega_o) \overset{(3.2)}{=} dL_o(s_i \to s_o, \omega_i \to \omega_o)\, d\sigma^{\perp}_{s_o}(\omega_o)\, d\mu^2(s_o) \tag{4.79}$$

from direction ω_i, where $d\sigma_{s_o}$ corresponds to the aperture of the measurement device.

Now, we are interested in defining a function that describes the reflecting properties of a surface. Constructing, in accordance with the definition of the reflection degree from optics, a function via the ratio of the outgoing flux to the incident flux, i.e.:

$$\frac{d^3\Phi_o(s_0, \omega_o)}{d^2\Phi_i(s_i, \omega_i)} = \frac{dL_o(s_i \to s_o, \omega_i \to \omega_o)\, d\sigma^{\perp}_{s_o}(\omega_o)\, d\mu^2(s_o)}{L_i(s_i, \omega_i)\, d\sigma^{\perp}_{s_i}(\omega_i)\, d\mu^2(s_i)}, \tag{4.80}$$

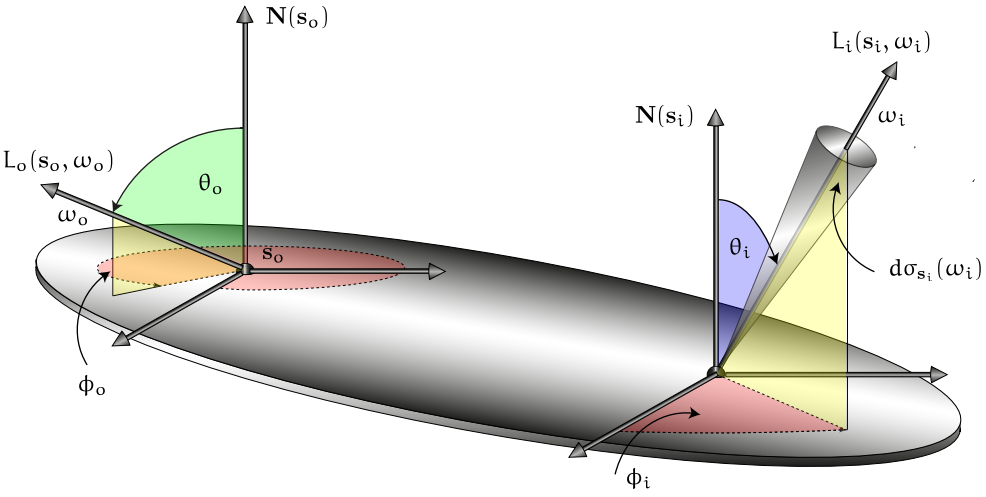


FIGURE 4.25: THE BSSRDF. The bidirectional scattering-surface reflectance-distribution function serves as a model for subsurface scattering. It is a distribution function, that relates the irradiance from direction ω_i at a surface point s_i to its contribution to the reflected radiance in direction ω_o at another surface point s_o.

results in the problem that this term contains three differential quantities in the numerator, but only two in the denominator.

Let us consider Equation (4.80) a little bit more closely. Since the incident radiance L_i and the incident projected solid angle $d\sigma^{\perp}_{s_i}$ are directly proportional to dL_o—doubling the incident radiance respectively the incident solid angle leads to doubling the reflected radiance—the term (4.80) is invariant with respect to these two quantities. It should also be clear that Relation(4.80) is not invariant with respect to the outgoing solid angle $d\sigma_{s_o}$—obviously, it not proportional to $d\sigma_{s_o}$—since the exitant solid angle only depends on the measuring device but does not depends on the incident radiance falling on the surface patch.

$d\sigma^{\perp}$ (89)

From this discussion, we now conclude, that it is not a good idea to define a bidirectional reflectance-distribution function as proposed in Relation (4.80). As the exitant radiance is proportional to the incident irradiance we define, in accordance with [135, Nicodemus & al. 1977], a BSSRDF as follows:

DEFINITION 4.14 (Bidirectional Scattering-Surface reflectance-distribution Function, BSSRDF) *Let $\partial\mathcal{V}$ be the set of all 2-dimensional surfaces of scene objects in $\mathbb{R}^3$, s_i, s_o be points on the same surface $A \in \partial\mathcal{V}$, and $\mathcal{H}^2_i$ and $\mathcal{H}^2_o$ denotes the incident and exi-*

$\partial\mathcal{V}$ (41)

Measurable Function (98)

tant hemispheres, which refer to the same set of directions. We call the measurable function S defined by:

$$S : \partial\mathcal{V} \times \partial\mathcal{V} \times \mathcal{H}_i^2 \times \mathcal{H}_o^2 \to [0, \infty] \tag{4.81}$$

with

$$S(\mathbf{s}_i \to \mathbf{s}_o, \omega_i \to \omega_o) \stackrel{\text{def}}{=} \frac{d^2 L_o(\mathbf{s}_i \to \mathbf{s}_o, \omega_i \to \omega_o)}{d^2\Phi_i(\mathbf{s}_i, \omega_i)} \tag{4.82}$$

$$= \frac{d^2 L_o(\mathbf{s}_i \to \mathbf{s}_o, \omega_i \to \omega_o)}{dE_i(\mathbf{s}_i, \omega_i)\, d\mu^2(\mathbf{s}_i)} \tag{4.83}$$

$$= \frac{d^2 L_o(\mathbf{s}_i \to \mathbf{s}_o, \omega_i \to \omega_o)}{L_i(\mathbf{s}_i, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}_i}(\omega_i)\, d\mu^2(\mathbf{s}_i)} \tag{4.84}$$

the bidirectional scattering-surface reflectance-distribution function*, also briefly denoted as the* BSSRDF*.*

The BSSRDF is the quantity that can be used to characterize all non-metallic materials. It has units of $\left[\frac{1}{sr \cdot m^2}\right]$ and describes the ratio of exitant radiance L_o at point $\mathbf{s}_o$ in direction ω_o with respect to the entire incident flux Φ_i coming from direction ω_i at point $\mathbf{s}_i$. A BSSRDF can also be interpreted as a probability distribution function that for a given incoming direction at an entering point returns the probability of light outgoing at some other point in any outgoing direction. As each of the directions ω_i and ω_o can be parametrized by the azimuth angle ϕ and the zenith angle θ, and $\mathbf{s}_i$ as well as $\mathbf{s}_o$ are points on any surface of $\partial\mathcal{V}$, the BSSRDF is an element of an 8-dimensional function space. Usually, a BSSRDF is also dependent on the wavelength of the incoming light, which is in computer graphics approximated often by defining and evaluating the BSSRDF separately per color channel. Thus, taking into account also the light phenomena of fluorescence and phosphorescence, then a BSSRDF has 12 degrees of freedom, [69, Goesele 2004], [124, Lensch 2005]. Such a complex function is costly to evaluate even if it does not take the wavelength of the light and the time into consideration.

CDF (171)

$\omega = (\theta, \phi)$ (834)

Linear Function Space (28)

REMARK 4.8 *It should be clear that measuring a 12-dimensional BSSRDF is not really possible, hence in practice rendering algorithms make only use of BSSRDFs of the type defined above, i.e. of no more than 8 degrees of freedom. Depending on the scattering event, which we observe, we can make further simplifications with respect to the representation of the BSSRDF, for this, see Section 4.2.2.2.*

Given the description of the incident illumination, via the concept of the BSSRDF, it is now possible to derive an equation that predict the appearance of a surface. Reformulating Relation (4.82) as follows:

$$d^2 L_o(\mathbf{s}_i \to \mathbf{s}_o, \omega_i \to \omega_o) = S(\mathbf{s}_i \to \mathbf{s}_o, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}_i}(\omega_i)\, d\mu^2(\mathbf{s}_i) \tag{4.85}$$

and subsequently integrating patch $A(\mathbf{s}_i)$ over the upper hemispheres about $\mathbf{s}_i$, we arrive at the *subsurface scattering equation.*

DEFINITION 4.15 (Subsurface-Scattering Equation) *The* subsurface-scattering equation *indicates the fraction of incident radiance at all point $\mathbf{s}_i \in \partial\mathcal{V}$ of a surface patch from all directions that is scattered only in direction ω_o at point $\mathbf{s}_o$, it is defined as*

$$L_{sseq}(\mathbf{s}_o, \omega_o) \stackrel{def}{=} \int_{A(\mathbf{s}_i)} \int_{\mathcal{H}_i^2(\mathbf{s}_i)} S(\mathbf{s}_i \to \mathbf{s}_o, \omega_i \to \omega_o)\, L_i(\mathbf{s}_i, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}_i}(\omega_i)\, d\mu^2(\mathbf{s}_i). \quad (4.86)$$

4.2.2.2 SCATTERING AT OBJECT SURFACES, THE BRDF AND THE BTDF

For most applications in computer graphics the usage of a BSSDRF makes— due to the computational effort for evaluating a BSSRDF—no sense for describing the reflection behavior of a surface. So, we are interested in finding an other function that describes the interaction of light at surfaces similar to the BSSRDF, but that is easier to evaluate.

Let us assume, we have to render a material such as polished hard metal, glass, or clear water. Compared with materials such as skin, wood, stone, or snow, as considered in the foregoing section, these materials posses only inhomogeneities much smaller than the smallest visible wavelength of light. That is, they can be considered to be optically homogeneous and hence, the scale of their scattering is extremely small, or they will not scatter light traveling through it [1, Akinene-Möller & al.].

Thus, the scattered light is re-emitted from the surface very close to its original entry point, which means, that the effect of subsurface scattering can be neglected. Depending on the material properties of the surface, the interaction of light can be reduced to a scattering process at the surface. Under the assumption of a homogeneous material the BSSRDF is then no more dependent on the spatial parameters $\mathbf{s}_i$ and $\mathbf{s}_o$ but only on the distance $\|\mathbf{s}_o - \mathbf{s}_i\|_2$ between these two points. This results in a function parameterized in terms of two directions and a displacement on the surface, thus 6 degrees of freedom. Such a kind of a bidirectional distribution function is called a *bidirectional subsurface-scattering distribution function*, also denoted as a BSSDF. With respect to the definition of a bidirectional reflectance-distribution function, the discussion from above now suggest to abstract in Definition 4.14 from the integration over the surface patch. This then leads to the definition of the *bidirectional reflectance-distribution function.*

DEFINITION 4.16 (Bidirectional reflectance-distribution Function, BRDF) *Let us assume $\partial\mathcal{V}$ be a set of 2-dimensional surfaces in $\mathbb{R}^3$, $\mathbf{s}$ be a point on any surface $A \in \partial\mathcal{V}$, $\mathcal{H}_i^2$ and $\mathcal{H}_o^2$ be the incident and exitant hemisphere, which refer to the same set of directions, see Figure 4.26. We call the measurable function, f_r, defined by:*

$\partial\mathcal{V}$ (41)

Measurable Function (98)

$$f_r : \partial\mathcal{V} \times \mathcal{H}_i^2 \times \mathcal{H}_o^2 \to [0, \infty] \quad (4.87)$$

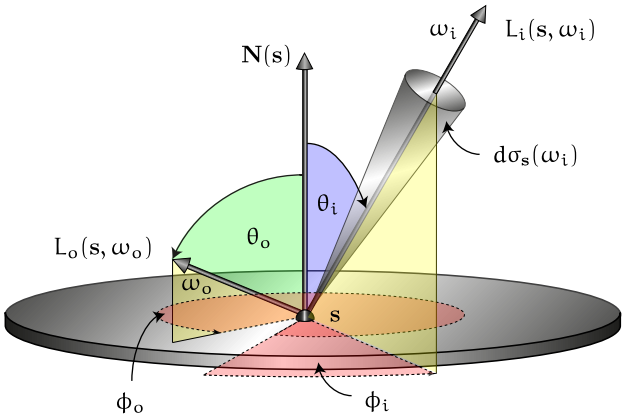


FIGURE 4.26: THE BIDIRECTIONAL REFLECTANCE-DISTRIBUTION FUNCTION. The BRDF is a four-dimensional function, defined on $\partial\mathcal{V} \times \mathcal{H}^2_i \times \mathcal{H}^2_o$ with values in $[0, \infty]$. It describes how much incident light, coming from direction ω_i, is reflected at a surface point in the outgoing direction ω_o.

with

$$f_r(s, \omega_i \to \omega_o) \stackrel{def}{=} \frac{dL_o(s, \omega_i \to \omega_o)}{dE(s, \omega_i)} = \frac{dL_o(s, \omega_i \to \omega_o)}{L_i(s, \omega_i)\, d\sigma^{\perp}_s(\omega_i)}, \tag{4.88}$$

the bidirectional reflectance-distribution function*, also briefly denoted as the* BRDF.

Let us assume that light striking a point on a surface is reflected at the same location then every BRDF is an approximation of a BSSRDF. Obviously, the BRDF is defined as the ratio of the reflected differential radiance L_o at point s in direction ω_o with respect to the differential irradiance E_i coming from direction ω_i at s. The BRDF can thus be taken as the fundamental quantity for the optical characterization of surfaces. Due to its definition it has units of $\left[\frac{1}{sr}\right]$. Like the BSSRDF, a BRDF can also be interpreted as a probability density function that for a given incoming direction at a point returns the probability of light emanating in any outgoing direction. Dependent on a spatial variable and two directions, the BRDF is an element of a 6-dimensional function space. In reality, the BRDF is wavelength dependent, which is in computer graphics approximated usually by defining and evaluating the BRDF separately per color channel. Discussing the reflectance behavior at homogeneous surfaces, then also the spatial variable can be omitted since the reflectance properties of the material do not vary with spatial position. The associated BRDF is then only 4-dimensional. A further simplification of the BRDF

Radiance (250)
Irradiance (257)
PDF (176)
Linear Function Space (28)

can be achieved by assuming that the underlying material is isotropic, such as steel and alluminium. This leads to BRDFs that are only dependent on the zenith angles θ_i and θ_o and the difference of the azimuth angle $\phi_o - \phi_i$, thus 3-dimensional, see Figure 4.27.

Isotropic BRDF (335)

Similar to the construction of the subsurface-scattering equation from Relation (4.86), we can now construct a reflectance equation, based on the definition of the BRDF that yields information on the reflectance appearance of a surface. Rephrasing Relation (4.88) with respect to the outgoing radiance and integrating over the positive hemisphere around observation point $\mathbf{s}$ then results in the so-called *reflectance equation*:

DEFINITION 4.17 (Reflectance Equation) *The* reflectance equation*, often also called* reflection equation*, indicates the quantity of incident radiance at point* $\mathbf{s}$ *from all directions that is reflected only in direction* ω_o*, it is defined as:*

$$L_o(\mathbf{s}, \omega_o) \stackrel{def}{=} \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i). \tag{4.89}$$

REMARK 4.9 *The reflectance equation plays a central role in our rendering algorithms for solving the stationary light transport equation in free space. It is this quantity that any Monte Carlo rendering algorithms has to evaluate at each hit point* $\mathbf{s}$ *on purely reflective surfaces, thus surfaces, that are not light sources.*

Chapter 9

From our discussions in Section 4.2.1.2 we know that optical discontinuities in a material are responsible for scattering of light at the interface between two different substances. Let us now consider the interaction of light at a flat, perfectly smooth, and polished surface, whose optical irregularities are much smaller than the smallest wavelength of light. At such a surface, which is of course not possible in reality, these irregularities have no effect on the light flow. Due to the surface properties, the photons contained in a light beam from direction ω_i incident at point $\mathbf{s}$ do not interact with the material, that is, they can not enter the material, but bounce at the surface according to the law of reflection in just one direction ω_o. Since ideal specular reflection occurs without loss of energy, then the reflected radiance is given by the incident radiance, i.e. it must hold:

Law of Reflection (300)

$$L_o(\mathbf{s}, \omega_o) = L_i(\mathbf{s}, \omega_i), \tag{4.90}$$

with $\theta_o = \theta_i$ and $\phi_o = \phi_i \pm \pi$. Under these conditions, now the following theorem is valid:

THEOREM 4.1 (The Ideal Specular BRDF) *Given be a surface* $A \in \partial\mathcal{V}$ *which satisfies the condition from Equation (4.90) for all* $\mathbf{s} \in A$*. Then, the reflection behavior of* A *can*

$\partial\mathcal{V}$ (41)

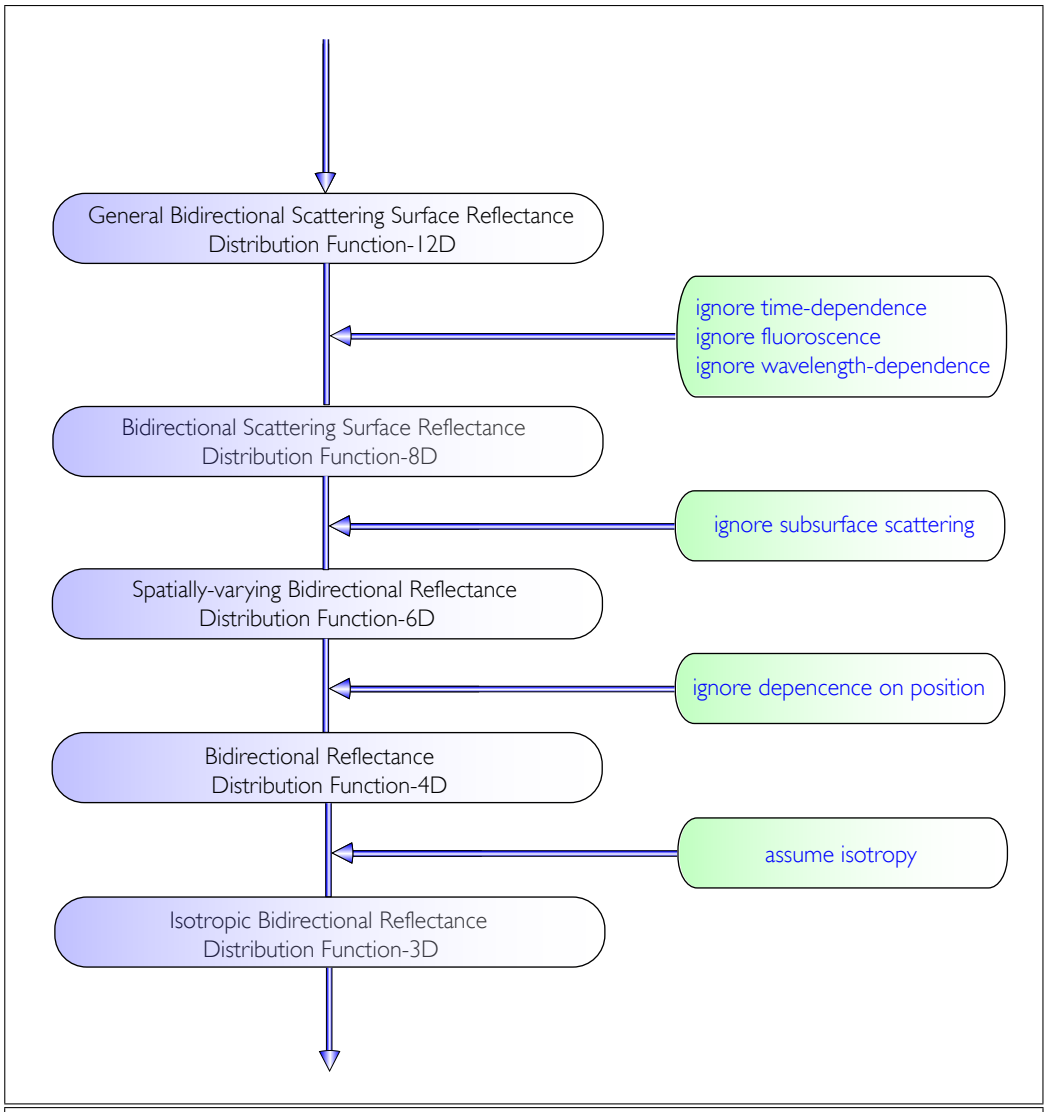


FIGURE 4.27: TAXONOMY OF APPEARANCE REPRESENTATIONS. Since measuring a 12-dimensional function is not really possible in a rendering algorithm, typically assumptions about the form of the bidirectional scattering surface reflectance-distribution function are required. Ignoring fluorescence and time, as well as wavelength-dependence leads to a 8D function, but also this function is not really suitable for rendering. Neglecting subsurface scattering, that is, using $\mathbf{s}_i = \mathbf{s}_o$, results in a BRDF of 6 dimensions. If the reflectance behavior of a surface is additionally independent on the position, then the associated BRDF is a four-dimensional function. Last but not least, the scattering at isotropic surfaces can be simulated via a 3D bidirectional scattering surface reflectance-distribution function.

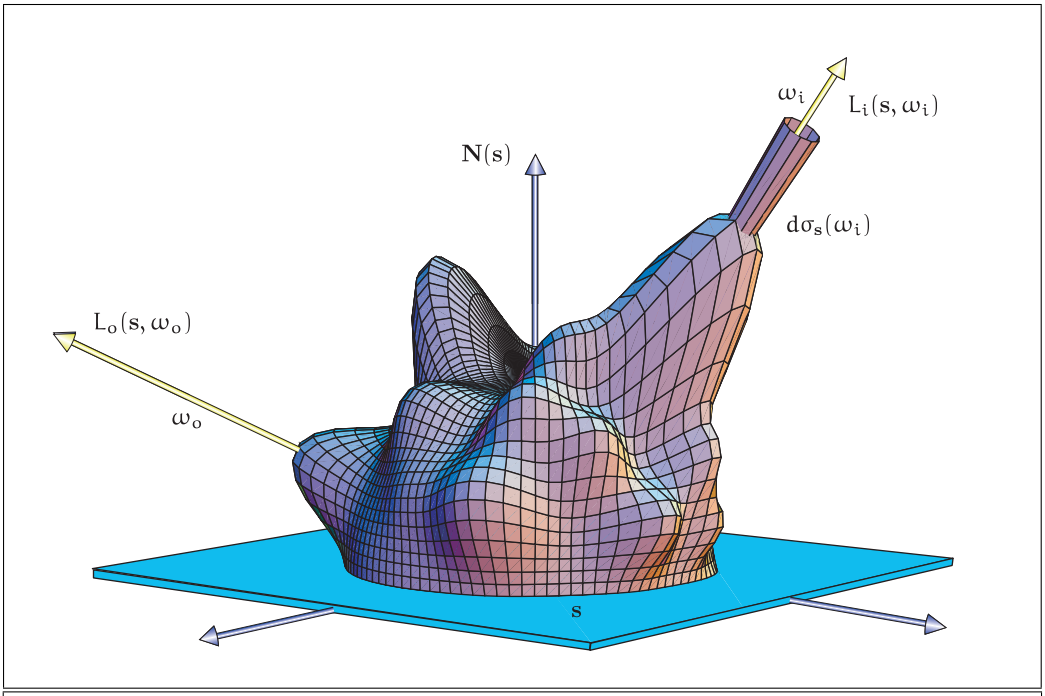


FIGURE 4.28: VISUALIZATION OF AN ARTIFICIALLY GENERATED BRDF.

be described by a so-called ideal specular BRDF, $f_r^{\vee}$, *which is given by:*

$$f_r^{\vee}(s, \omega_i \rightarrow \omega_o) \tag{4.91}$$

$$= \delta_{\sigma^\perp}(\omega_i - \omega_o) \tag{4.92}$$

$$= \frac{\delta_\sigma(\omega_i - \omega_o)}{|\cos\theta_i|} \tag{4.93}$$

$$= \frac{\delta(\cos\theta_i - \cos\theta_o)}{|\cos\theta_i|}\,\delta(\phi_i - (\phi_o \pm \pi)), \tag{4.94}$$

where $\omega_o \stackrel{def}{=} M_N(\omega_i)$ *corresponds to the mirrored direction of* ω_i.

PROOF 4.1 *Due tue Definition 4.17, the ideal specular BRDF, which we are seeking, must satisfy the reflectance equation from Relation (4.89), i.e.:*

$$L_o(s, \omega_o) = \int_{\mathcal{H}_i^2(s)} f_r^{\vee}(s, \omega_i \rightarrow \omega_o)\, L_i(s, \omega_i)\, d\sigma_s^{\perp}(\omega_i). \tag{4.95}$$

Since we are only interested in ideal specular reflection, the integral has to be evaluated only in direction to the reflected ray $r(s, \omega_o)$, *thus* $\omega_o = (\theta_o, \phi_o)$ *with*

$\theta_o = \theta_i$ *and* $\phi_o = \phi_i \pm \pi$. *This means that the integral has a singularity at* ω_o. *From measure theory it is known that the measure of a finite or countable infinite set is zero, that is, it is not possible to find a function which satisfies the conditions (4.89) and (4.90).*

Dirac δ-Distribution (118)

Fortunately, both conditions can be satisfied by the Dirac δ-distribution. From Example 2.49 we know that the exitant radiance L_o *can be written in terms of the incident radiance* L_i *with the help of the Dirac δ-distribution, so we get:*

$$L_o(s,\omega_o) = \int_{\mathcal{H}^2_i(s)} \delta_{\sigma^\perp}(\omega_i - \omega_o)\, L_i(s,\omega_i)\, d\sigma^\perp_s(\omega_i) \tag{4.96}$$

$$= \int_{\mathcal{H}^2_i(s)} \frac{\delta_\sigma(\omega_i - \omega_o)}{|\cos\theta_i|}\, L_i(s,\omega_i)\, |\cos\theta_i|\, d\sigma_s(\omega_i), \tag{4.97}$$

which with Relation (2.328) can be written as:

$$L_o(s,\omega_o) \tag{4.98}$$

$$= \int_{\mathcal{H}^2_i(s)} \underbrace{\frac{\delta(\cos\theta_i - \cos\theta_o)}{|\cos\theta_i|}\, \delta(\phi_i - (\phi_o \pm \pi))}_{f^\vee_r(s,\omega_i \to \omega_o) \stackrel{def}{=}}\, L_i(s,\omega_i))\, d\sigma^\perp_s(\omega_i)$$

$$\stackrel{(2.192)}{=} \int_{\mathcal{H}^2_i(s)} f^\vee_r(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma^\perp_s(\omega_i). \tag{4.99}$$

REMARK 4.10 (The Image Range of a BRDF) *Defined as the fraction of two differential quantities, where both, the numerator and the denominator are non-negative quantities, the BRDF can only assume non-negative values. Furthermore, the representation of an ideal specular BRDF as a Dirac δ-distribution implies that a BRDF is a mapping,* f_r*:*

Dirac δ-Distribution (118)

$$f_r : \partial\mathcal{V} \times \mathcal{H}^2_i \times \mathcal{H}^2_o \to [0,\infty] \tag{4.100}$$

that is, f_r *can also assume the value infinity.*

In real world there are no ideal specular surfaces that reflect light lossless in only a single direction. Reflection at real surfaces entails always also loss of energy. From Section 4.2.1.2 it is known that the fraction of light, which is reflected in the mirrored direction ω_o, is specified by the Fresnel reflectance for unpolarized light from Relation (4.72), that is:

$$L_o(s,\omega_o) = F_r(\omega_i)\, L_i(s,\omega_i), \tag{4.101}$$

with $\theta_o = \theta_i$ and $\phi_o = \phi_i \pm \pi$. A BRDF, describing the specular reflectance behavior at a flat and smooth surface, is then given by the following lemma:

∂V (41) LEMMA 4.1 (The Specular BRDF) *Given be a surface* $A \in \partial\mathcal{V}$, *that satisfies for all* $s \in A$ *the condition from Equation (4.101). Then, the reflection behavior of* A *can be described by a so-called* specular BRDF, $f_r^\vee$, *which is given by:*

$$\begin{aligned} f_r^\vee(s, \omega_i \to \omega_o) &= F_r(\omega_i)\, \delta_{\sigma^\perp}(\omega_i - \omega_o) && (4.102)\\ &= F_r(\omega_i)\, \frac{\delta_\sigma(\omega_i - \omega_o)}{|\cos\theta_i|} && (4.103)\\ &= F_r(\omega_i)\, \frac{\delta(\cos\theta_i - \cos\theta_o)}{|\cos\theta_i|}\, \delta(\phi_i - (\phi_o \pm \pi)) && (4.104) \end{aligned}$$

with $\omega_o \overset{def}{=} M_N(\omega_i)$.

PROOF 4.1 *With the same arguments as in the proof to the previous theorem, the reflected radiance* $L_o(s, \omega_o)$ *can now be written as:*

$$\begin{aligned} & L_o(s, \omega_o) \\ &= F_r(\omega_i)\, L_i(s, (\theta_i, \phi_o \pm \pi)) && (4.105)\\ &\overset{(2.328)}{=} \int_{\mathcal{H}_i^2(s)} \underbrace{F_r(\omega_i) \frac{\delta(\cos\theta_i - \cos\theta_o)}{|\cos\theta_i|}\, \delta(\phi_i - (\phi_o \pm \pi))}_{f_r^\vee(s, \omega_i \to \omega_o) \overset{def}{=}} L_i(s, \omega_i)\, |\cos\theta_i|\, d\sigma_s(\omega_i) \\ &\overset{(2.192)}{=} \int_{\mathcal{H}_i^2(s)} f_r^\vee(s, \omega_i \to \omega_o)\, L_i(s, \omega_i)\, d\sigma_s^\perp(\omega_i). && (4.106) \end{aligned}$$

Now, the most surfaces in real world are not perfectly smooths since they posses optical irregularities, that are much greater than the smallest wavelength of light. The photons of a light ray falling at such a surface now interact with these irregularities in such a way that they are also reflected according to the reflection law, but now due to the irregularities and the discontinuities in the material in all directions of the hemisphere. Although we can only approximate this effect under very specially prepared experimental conditions, this idealization plays a fundamental role in computer graphics, especially in algorithms based on the finite element approach. This type of reflection is called *ideal diffuse reflection* and the surface underlying them is called a *Lambertian reflector*. Since reflected radiance in any direction is equal, that is,

Chapter 10

Lambertian Reflector (349)

$$L_o(s, \omega_o) = L_o(s, \omega_o') \tag{4.107}$$

for all $\omega_o, \omega_o' \in \mathcal{H}_i^2(s)$, a BRDF describing the ideal diffuse reflectance behavior at such a perfectly rough surface is then given by the following theorem:

THEOREM 4.2 (The Ideal Diffuse BRDF) *Given be a surface* $M \in \partial\mathcal{V}$, *that satisfies for all* $s \in A$ *the condition from Equation (4.107). Then, the reflection behavior of* A *can be described by a so-called* diffuse BRDF, f_r^o, *which is constant, thus:* ∂V (41)

$$f_r^o(s, \omega_i \to \omega_o) = C, \tag{4.108}$$

where C is a positive real number that is smaller than infinity.

Reflection Equation (321) PROOF 4.2 *From the reflection equation we conclude that for a BRDF f_r^o, characterizing diffuse material properties, the following must hold:*

$$\begin{aligned} 0 &= L_o(s,\omega_o) - L_o(s,\omega_o') && (4.109)\\ &= \int_{\mathcal{H}_i^2(s)} \underbrace{(f_r^o(s,\omega_i \to \omega_o) - f_r^o(s,\omega_i \to \omega_o'))}_{=0} L_i(s,\omega_i)\, d\sigma_s^\perp(\omega_i), && (4.110)\end{aligned}$$

as $L_i(s,\omega_i) \neq 0$. This then implies

$$f_r^o(s,\omega_i \to \omega_o) = f_r^o(s,\omega_i \to \omega_o'). \tag{4.111}$$

This means that the fraction of incident radiance,, reflected in directions ω_o and ω_o', is the same for ω_i and all $\omega_o, \omega_o' \in \mathcal{H}_o^2(s)$. Therefore we can conclude that the assigned BRDF is constant and can be written as

$$f_r^o \overset{(4.88)}{=} \frac{dL_o(s,\omega_i \to \omega_o)}{dE_i(s,\omega_i)} = C, \tag{4.112}$$

and that the reflected radiance is proportional to the incident irradiance, since it holds:

$$\begin{aligned} L_o(s,\omega_o) &= f_r^o \int_{\mathcal{H}_i^2(s)} L_i(s,\omega_i)\, d\sigma_s^\perp(\omega_i) && (4.113)\\ &\overset{(3.46)}{=} f_r^o\, E(s). && (4.114)\end{aligned}$$

REMARK 4.11 (Isotropic BRDF) *Due to [135, Nicodemus & al. 1977] a BRDF, such* Isotropy (335) *as the ideal diffuse BRDF, is called* isotropic, *as the reflected radiance is a constant with the same value for all outgoing directions ω_o regardless of how it is irradiated.*

REMARK 4.12 *In the following section, we will pick up the above theorem once more* Conservation Of Energy (332) *and derive an exact formula for the diffuse BRDF based on the conservation of energy.*

After formulating the reflection behavior at surfaces with the help of BRDFs, now it still remains to formulate the refraction at the interface between two media by a similar concept: the *bidirectional transmission-distribution function.*

DEFINITION 4.18 (Bidirectional Transmission-Distribution Function, BTDF) *Let us as-* $\partial\mathcal{V}$ (41) *sume $\partial\mathcal{V}$ be a set of 2-dimensional surfaces in $\mathbb{R}^3$, s be a point on any surface $A \in \partial\mathcal{V}$, $\mathcal{H}_i^2$ and $\mathcal{H}_t^2$ be the incident and transmitted hemisphere, with $\mathcal{H}_t^2 = -\mathcal{H}_i^2$.* Measurable Function (98) *We call the measurable function, f_t, defined by:*

$$f_t : \partial\mathcal{V} \times \mathcal{H}_i^2(s) \times \mathcal{H}_t^2(s) \to [0,\infty] \tag{4.115}$$

with

$$f_t(s, \omega_i \to \omega_t) \stackrel{def}{=} \frac{dL_t(s, \omega_i \to \omega_t)}{dE(s, \omega_i)} = \frac{dL_t(s, \omega_i \to \omega_t)}{L_i(s, \omega_i)\, d\sigma_s^{\perp}(\omega_i)}, \tag{4.116}$$

the bidirectional transmission-distribution function, *also briefly denoted as the* BTDF, *see Figure 4.29.*

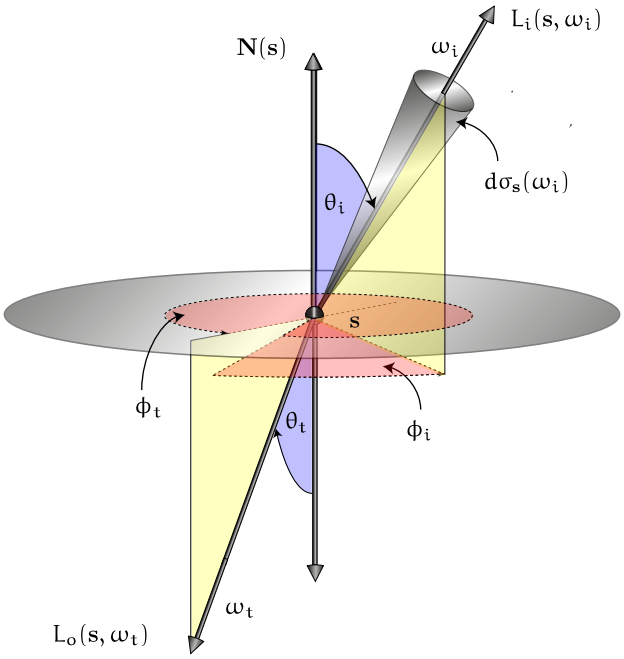


FIGURE 4.29: BIDIRECTIONAL TRANSMISSION-DISTRIBUTION FUNCTION. The BTDF is a four-dimensional function defined on $\partial\mathcal{V} \times \mathcal{H}_i^2 \times \mathcal{H}_t^2$ with values in $[0, \infty]$. It describes how much incident light, coming from direction ω_i, is refracted at a surface point in the refracted direction ω_t.

Let us study the transition of a light ray from a medium with a smaller refractive index into a medium with a higher refractive index, e.g. from air to glass. Due to Snell's law from Definition 4.9 the ray is refracted at the interface between the two media in direction to the normal. This means that transmitted light coming from the whole incident hemisphere $\mathcal{H}_i^2(s)$ about any point s no longer fills the entire hemisphere $\mathcal{H}_t^2(s)$ on the opposite side of the interface, see Figure 4.30.

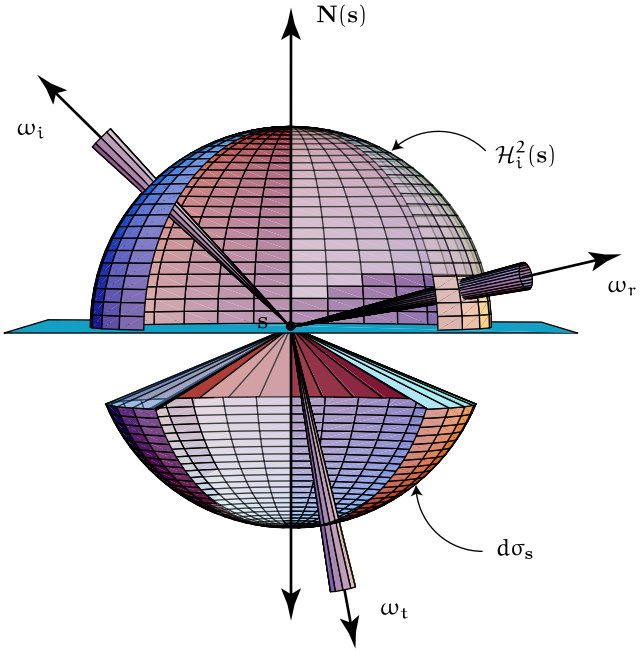


FIGURE 4.30: GEOMETRY FOR DEFINING THE BTDF. The geometry for describing refraction at a surface point. Obviously, the wave incident at point $\mathbf{s}$ is refracted in direction ω_t, where it holds: $\omega_t \in \left[\sin^{-1}\left(\frac{\eta_i}{\eta_t}\right), \pi\right] \times [0, 2\pi]$.

∂V (41) Refraction Index (305)

THEOREM 4.3 (The Ideal Transmitted BTDF) *Let $A \in \partial V$ be the interface between two media with refraction indices η_i and η_t. $L_i(\mathbf{s}, \omega_i)$ denotes as usual the incident radiance at point $\mathbf{s}$ coming from direction ω_i and $L_t(\mathbf{s}, \omega_t)$ is the radiance refracted at $\mathbf{s}$ in the transmitted direction ω_t. The transmission behavior of A can then be described by a so-called* ideal transmitted BTDF, f_t, *which is given by:*

$$f_t(\mathbf{s}, \omega_i \to \omega_t) = \frac{\eta_t^2}{\eta_i^2}\delta_{\sigma^\perp}(\omega_i - \omega_t) \tag{4.117}$$

$$= \frac{\eta_t^2}{\eta_i^2}\frac{\delta_\sigma(\omega_i - \omega_t)}{|\cos\theta_i|} \tag{4.118}$$

$$= \frac{\eta_t^2}{\eta_i^2}\frac{\delta(\cos\theta_i - \cos\theta_t)}{|\cos\theta_i|}\delta(\phi_i - (\phi_t \pm \pi)), \tag{4.119}$$

with $\omega_t \stackrel{\text{def}}{=} R_{\mathbf{N}}(\omega_i)$.

PROOF 4.3 *Let us discuss the case where light arrives from a medium with smaller refractive index η_i and enters a medium with a greater refractive index η_t. Discussing*

the phenomenon of ideal loss less refraction as a consequence of the law of conservation of energy and the law of refraction, the light energy coming from the entire hemisphere around a small surface patch $d\mu(A)$ *is compressed into a solid angle* $d\sigma_s$ *over the patch, that occupied only a subset of the exitant hemisphere, see Figure 4.30. This means that the transmitted radiance increases as light crosses the interface.*

Considering the angular parameterization (θ, ϕ) *of* ω *and using the definition of the projected solid angle from Equation (2.192) then it holds:*

$(\theta, \phi, 1)$ (832)

$$\frac{d^2\Phi_t(s,\omega_t)}{d^2\Phi_i(s,\omega_i)} = \frac{L_t(s,\omega_t)\,|\cos\theta_t|\,\sin\theta_t\,d\theta_t\,d\phi_t}{L_i(s,\omega_i)\,|\cos\theta_i|\,\sin\theta_i\,d\theta_i\,d\phi_i} \tag{4.120}$$

$$\overset{(4.66)}{=} \frac{L_t(s,\omega_t)\,\eta_i}{L_i(s,\omega_i)\,\eta_t}\frac{|\cos\theta_t|\,d\theta_t\,d\phi_t}{|\cos\theta_i|\,d\theta_i\,d\phi_i}. \tag{4.121}$$

Due to the conservation of energy it holds

Conservation of Energy (332)

$$\frac{L_t(s,\omega_t)\,\eta_i}{L_i(s,\omega_i)\,\eta_t}\frac{|\cos\theta_t|\,d\theta_t\,d\phi_t}{|\cos\theta_i|\,d\theta_i\,d\phi_i} = 1 \tag{4.122}$$

which, slightly reformulated, leads to:

$$L_t(s,\omega_t) = L_i(s,\omega_i)\frac{\eta_t}{\eta_i}\frac{|\cos\theta_i|\,d\theta_i\,d\phi_i}{|\cos\theta_t|\,d\theta_t\,d\phi_t}. \tag{4.123}$$

Due to [136, Nicodemus 1963], by differentiating Snell's law with respect to the azimuth angle θ*, we obtain from*

$$\eta_i \sin\theta_i = \eta_t \sin\theta_t \tag{4.124}$$

the relation

$$\eta_i \cos\theta_i d\theta_i = \eta_t \cos\theta_t d\theta_t. \tag{4.125}$$

According to the law of refraction obviously it holds $\phi_t = \phi_i + \pi$*. Differentiating this relation with respect to* ϕ *yields:*

$$d\phi_i = d\phi_t. \tag{4.126}$$

Using these identities in Equation (4.123), then we get:

$$L_t(s,\omega_t) = \frac{\eta_t^2}{\eta_i^2} L_i(s,\omega_i), \tag{4.127}$$

that is, with respect to Equation (4.123) the transmitted radiance increases by a factor of $\frac{\eta_t^2}{\eta_i^2}$*.*

In analogy to our explanations to the Dirac δ*-distribution, the value of* L_t *at*

Dirac δ-Distribution (118)

point $\mathbf{s}$ *in direction* ω_t *can be evaluated as:*

$$
\begin{aligned}
& L_t(\mathbf{s},\omega_t) && (4.128)\\
\overset{(4.127)}{=}\ & \int_{\mathcal{H}^2_i(\mathbf{s})} \delta_{\sigma^\perp}(\omega_i-\omega_t)\,\frac{\eta_t^2}{\eta_i^2}\,L_i(\mathbf{s},\omega_i)\,d\sigma^\perp_{\mathbf{s}}(\omega_i),\\
=\ & \int_{\mathcal{H}^2_i(\mathbf{s})} \frac{\delta_{\sigma}(\omega_i-\omega_t)}{|\cos\theta_i|}\,\frac{\eta_t^2}{\eta_i^2}\,L_i(\mathbf{s},\omega_i)\,|\cos\theta_i|\,d\sigma_{\mathbf{s}}(\omega_i),\\
=\ & \int_{\mathcal{H}^2_i(\mathbf{s})} \frac{\eta_t^2}{\eta_i^2}\,\frac{\delta(\cos\theta_i-\cos\theta_t)}{|\cos\theta_i|}\,\delta(\phi_i-(\phi_o\pm\pi))\,L_i(\mathbf{s},\omega_i)\,|\cos\theta_i|d\sigma_{\mathbf{s}}(\omega_i),\\
=\ & \int_{\mathcal{H}^2_i(\mathbf{s})} \underbrace{\frac{\eta_t^2}{\eta_i^2}\,\frac{\delta(\cos\theta_i-\cos\theta_t)}{|\cos\theta_i|}\,\delta(\phi_i-(\phi_o\pm\pi))}_{f_t(\mathbf{s},\omega_i\to\omega_t)\overset{def}{=}}\,L_i(\mathbf{s},\omega_i)\,d\sigma^\perp_{\mathbf{s}}(\omega_i),
\end{aligned}
$$

which implies the following form of the BTDF:

$$f_t(\mathbf{s},\omega_i\to\omega_t)=\frac{\eta_t^2}{\eta_i^2}\frac{\delta(\cos\theta_i-\cos\theta_t)}{|\cos\theta_i|}\delta(\phi_i-(\phi_o\pm\pi)). \tag{4.129}$$

Similar to the case of ideal specular surfaces, in real world there are no perfectly smooth surfaces at which light can be refracted lossless in only a single direction. Like reflection, also refraction at real surfaces entails always loss of energy. To denote the fraction of incident energy, that is transmitted to the outgoing direction, we can use the Fresnel reflectance for unpolarized light from Relation (4.72), that is:

$$L_t(\mathbf{s},\omega_t)=(1-F_r(\omega_i))\,L_i(\mathbf{s},\omega_i), \tag{4.130}$$

with $\theta_t=\arcsin\left(\frac{\eta_i}{\eta_t}\sin\theta_i\right)$ and $\phi_o=\phi_i\pm\pi$. A BTDF, describing the specular transmittance behavior at a flat and smooth surface is then given by the following lemma.

$\partial\mathcal{V}$ (41) Refraction Index (305)

LEMMA 4.2 (The Transmitted BRDF) *Let* $A\in\partial\mathcal{V}$ *be the interface between two media with refraction indices* η_i *and* η_t. $L_i(\mathbf{s},\omega_i)$ *denotes as usual the incident radiance at point* $\mathbf{s}$ *coming from direction* ω_i *and* $L_t(\mathbf{s},\omega_t)$ *is the radiance refracted at* $\mathbf{s}$ *in the transmitted direction* ω_t. *The transmission behavior of* A *can then be described by a so-called* transmitted BTDF, f_t, *which is given by:*

$$
\begin{aligned}
f_t(\mathbf{s},\omega_i\to\omega_t) &= (1-F_r(\omega_i))\,\frac{\eta_t^2}{\eta_i^2}\,\delta_{\sigma^\perp}(\omega_i-\omega_t) && (4.131)\\
&= (1-F_r(\omega_i))\,\frac{\eta_t^2}{\eta_i^2}\,\frac{\delta_{\sigma}(\omega_i-\omega_t)}{|\cos\theta_i|} && (4.132)\\
&= (1-F_r(\omega_i))\,\frac{\eta_t^2}{\eta_i^2}\,\frac{\delta(\cos\theta_i-\cos\theta_t)}{|\cos\theta_i|}\,\delta(\phi_i-(\phi_t\pm\pi)), && (4.133)
\end{aligned}
$$

with $\omega_t\overset{def}{=}R_{\mathbf{N}}(\omega_i)$.

PROOF 4.2 *The proof is identical to Theorem 4.3, we leave it to the interested reader.*

4.2.2.3 PHYSICAL PROPERTIES OF BRDF AND BTDF, AND THE CONCEPTS OF REFLECTANCE AND TRANSMITTANCE

We mention it once again, BRDFs are approximations to the real interaction of light at surfaces. To be physically valid, so a BRDF has to satisfy some basic physical properties.

Now, from the Maxwell equations, see [27, Born & Wolf 1999] or [80, Hecht 2001], it is known, that the energy of light is proportional to the square of the amplitude of the electric field. That is, light is a non-negative physical quantity, and thus, also a BRDF must be non-negative, see Remark 4.10. Furthermore, a BRDF must also satisfy two other properties: *Helmholtz reciprocity* and *conservation of energy*. Just in the case, that all these three properties are satisfied by a BRDF, we call a BRDF *physically plausible*. Note, as we shall see further below using the example of a Lambertian BRDF, a physically plausible BRDF must not be physically possible or physically correct.

BRDF (320)

Lambertian BRDF (349)

While, the non-negativity of a BRDF is obviously—it is guaranteed by its definition as the fraction of two non-negative radiometric quantities, see Remark 4.10—the two other properties require to be studied a little more closely.

HELMHOLTZ RECIPROCITY. By *Helmholtz reciprocity* we mean that as a result of the symmetry of the Maxwell equations, the value of a BRDF remains the same even when the incoming and outgoing direction of the involved light waves are swapped. That is, swapping the direction of light does not change the amount of light that is reflected. The principle of Helmholtz reciprocity is illustrated in Figure 4.31. Mathematically, this property can be expressed as:

$$f_r(\mathbf{s}, \omega_i \to \omega_o) = f_r(\mathbf{s}, \omega_o \to \omega_i) \tag{4.134}$$

for all $\omega_i \in \mathcal{H}^2_i(\mathbf{s}), \omega_o \in \mathcal{H}^2_o(\mathbf{s})$. This is also the reason why we can use the notation

$$f_r(\mathbf{s}, \omega_i \leftrightarrow \omega_o) \stackrel{def}{=} f_r(\mathbf{s}, \omega_i \to \omega_o) = f_r(\mathbf{s}, \omega_o \to \omega_i) \tag{4.135}$$

in our equations.

As we will see further below, the principle of Helmholtz reciprocity is unique to reflection and is not valid for surfaces that transmit light.

REMARK 4.13 *In Section 9.3 we will shown, that it is the property of a BRDF satisfying the principle of Helmholtz reciprocity why so-called* bidirectional algorithms *can be used to solve the global illumination problem. Since such algorithms compute the radiance distribution in a scene by constructing paths starting at the eye of the observer and from the light sources at the same time, it is urgent necessary that the involved BRDFs, in those cases called* BSDFs*, satisfy the Helmholtz reciprocity.*

Radiance (250)

BSDF (371)

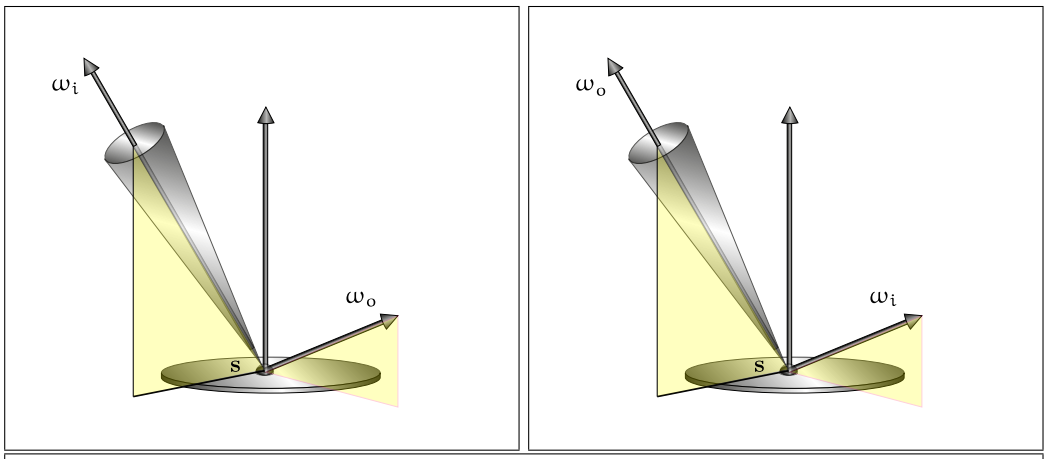


FIGURE 4.31: THE PRINCIPLE OF THE HELMHOLTZ RECIPROCITY. For all pairs of directions ω_i and ω_o it holds: $f_r(s, \omega_i \to \omega_o) = f_r(s, \omega_o \to \omega_i)$, i.e. swapping the direction of light does not change the amount of light that is reflected.

The example of the Phong BRDF will show, that a reciprocity failure is less serious than a failure in the property of conservation of energy. So, ray tracers, based on the classic principle of ray tracing, use non-reciprocal BRDFs and deliver under certain conditions acceptable images.

Phong BRDF (250)

Whitted-style Ray Tracing (664)

THE PRINCIPLE OF CONSERVATION OF ENERGY. The *principle of conservation of energy* has to do with the scattering of light during the light-matter interaction. As a direct consequence of the second law of thermodynamics, it states that the total energy of reflected light from a surface point always has a positive value and may never exceed the energy of incident light at that point, see Figure 4.32.

$d\sigma^\perp$ (89)

So, a BRDF must satisfy the following inequality for all directions $\omega_i \in \mathcal{H}^2_i(s)$:

$$0 \leq \int_{\mathcal{H}^2_o(s)} f_r(s, \omega_i \to \omega_o)\, d\sigma^\perp_s(\omega_o) \leq 1. \tag{4.136}$$

The integral between the inequality signs in Inequality (4.136) is also called the *directional-hemisperical reflectance* of a surface and is denoted by

$$\rho(\omega_i \to \mathcal{H}^2_o) \stackrel{def}{=} \int_{\mathcal{H}^2_o(s)} f_r(s, \omega_i \to \omega_o)\, d\sigma^\perp_s(\omega_o) \tag{4.137}$$

with $\omega_i \in \mathcal{H}^2_i(s)$.

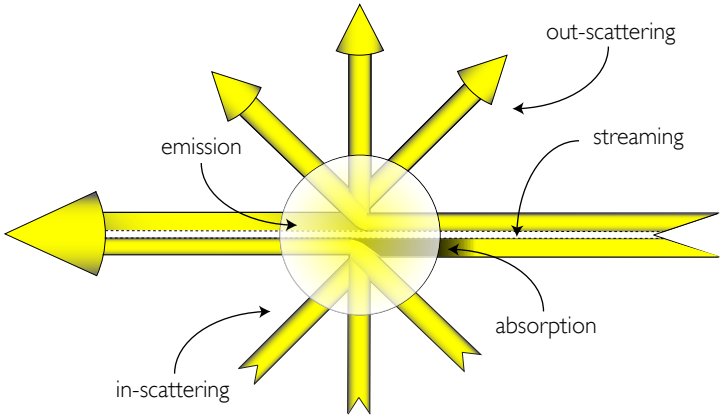


FIGURE 4.32: THE PRINCIPLE OF CONSERVATION OF ENERGY. The total amount of energy in an isolated system remains constant over time. Thus, energy can neither be created nor destroyed, it can only be transformed from one state to another.

THEOREM 4.4 (The Energy Conservation Condition of the BRDF) *Let f_r be a physically valid BRDF, then for all $\omega_i \in \mathcal{H}^2_i(\mathbf{s})$ it must hold:*

$$\int_{\mathcal{H}^2_o(\mathbf{s})} f_r(\mathbf{s}, \omega_i \rightarrow \omega_o)\, d\sigma^\perp_{\mathbf{s}}(\omega_o) \leq 1. \tag{4.138}$$

PROOF 4.4 *Let $\omega_i \in \mathcal{H}^2_i(\mathbf{s})$ be fixed. According to [220, Veach 1997], we assume that the incident power is concentrated in a single direction ω_i. Then, the incident radiance $L_i(\mathbf{s}, \omega)$ can be written as $L_i(\mathbf{s}, \omega) = 1 \cdot \delta_{\sigma^\perp}(\omega - \omega_i)$.* Dirac δ-distribution (118)

Applying the definition of the Dirac δ-distribution and the Property (2.307) then leads to

$$\begin{aligned} E(\mathbf{s}) &= \int_{\mathcal{H}^2_i(\mathbf{s})} L_i(\mathbf{s}, \omega)\, d\sigma^\perp_{\mathbf{s}}(\omega) && (4.139) \\ &= \int_{\mathcal{H}^2_i(\mathbf{s})} \delta_{\sigma^\perp}(\omega - \omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega) \overset{(2.307)}{=} 1 && (4.140) \end{aligned}$$

and

$$L_o(\mathbf{s},\omega_o) = \int_{\mathcal{H}_i^2(\mathbf{s})} f_r(\mathbf{s},\omega\to\omega_o)\,L_i(\mathbf{s},\omega)\,d\sigma_{\mathbf{s}}^{\perp}(\omega) \tag{4.141}$$

$$= \int_{\mathcal{H}_i^2(\mathbf{s})} f_r(\mathbf{s},\omega\to\omega_o)\,\delta_{\sigma^\perp}(\omega-\omega_i)\,d\sigma_{\mathbf{s}}^{\perp}(\omega) \tag{4.142}$$

$$\overset{(2.307)}{=} f_r(\mathbf{s},\omega_i\to\omega_o). \tag{4.143}$$

Radiant Exitance (267)

Using this relation in the formula for computing the radiant exitance then we get:

$$M(\mathbf{s}) = \int_{\mathcal{H}_o^2(\mathbf{s})} L_o(\mathbf{s},\omega_o)\,d\sigma_{\mathbf{s}}^{\perp}(\omega_o) \tag{4.144}$$

$$= \int_{\mathcal{H}_o^2(\mathbf{s})} f_r(\mathbf{s},\omega_i\to\omega_o)\,d\sigma_{\mathbf{s}}^{\perp}(\omega_o). \tag{4.145}$$

Since the law of conservation of energy implies

$$0 \le M \le E, \tag{4.146}$$

so the total amount of energy reflected at a surface point over all directions must be less than or equal to the total amount of incident energy at this point. Applying the Relations (4.140) and (4.145) then proves our theorem

$$0 \le \int_{\mathcal{H}_o^2(\mathbf{s})} f_r(\mathbf{s},\omega_i\to\omega_o)\,d\sigma_{\mathbf{s}}^{\perp}(\omega_o) \le 1. \tag{4.147}$$

REMARK 4.14 (The Principle of Superposition) *As two light rays incident at point* $\mathbf{s}$ *on a surface has no influence on each other, reflection behaves linearly. That is, the total amount of light reflected by the surface in a specific direction* ω_o *is given by the hemispherical integral over all possible incoming directions* ω_i *around* $\mathbf{s}$*. This then results in the reflection equation from (4.89), namely:*

$$L_o(\mathbf{s},\omega_o) \overset{def}{=} \int_{\mathcal{H}_i^2(\mathbf{s})} f_r(\mathbf{s},\omega_i\to\omega_o)\,L_i(\mathbf{s},\omega_i)\,d\sigma_{\mathbf{s}}^{\perp}(\omega_i). \tag{4.148}$$

REMARK 4.15 (Fluorescence and Phosphorescence) *It should also be mentioned that a BRDF is generally also dependent on the wavelength of the incident light as well as the reflected light. That is, a physically correct BRDF should also simulate the light effect of fluorescence. Now, in computer graphics we can get rid of a further dimension by discretizing the wavelength in so-called wavelength bands. Since the human visual system has only three kind of receptors we use the so-called* RGB-band*, thus, the three bands* red, green, *and* blue*. We then use the BRDF together with every such band to describe the wavelength dependance of reflection.*

Additionally, we want also assume that the time between the arrival of the incident light and the emittance of the reflected light at a point $\mathbf{s}$ *is negligible, that is, also phosphorescence can not be simulated by our BRDFs.*

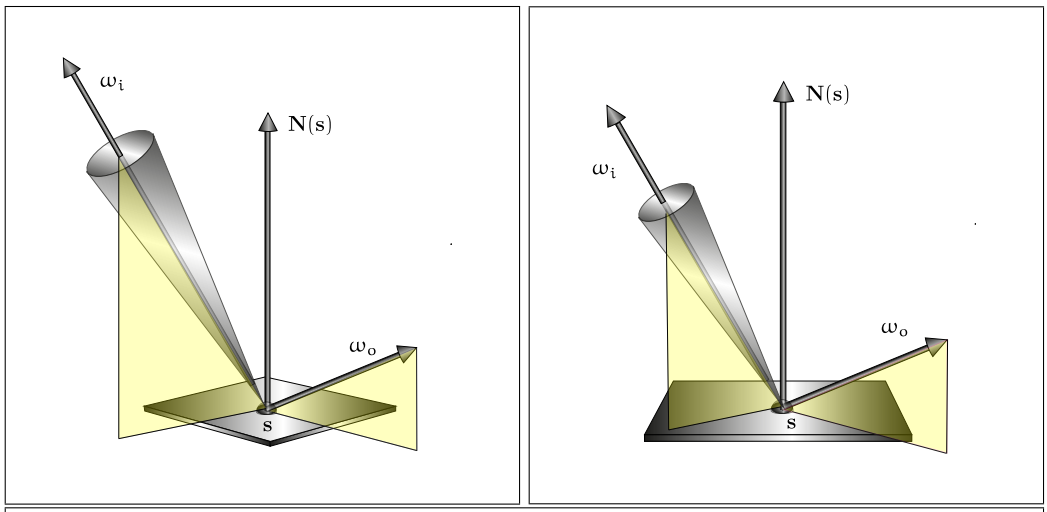


FIGURE 4.33: REFLECTION PROPERTY OF AN ANISOTROPIC BRDF. The amount of reflected light may change with rotations of the surface about the normal $\mathbf{N}(\mathbf{s})$ due to anisotropy.

ISOTROPIC AND ANISOTROPIC BRDFS. BRDFs, as the mathematical formulation of the reflection behavior at material surfaces, can be partitioned into two classes: *isotropic* BRDFs and *anisotropic* BRDFs.

The term isotropic describes the reflectance properties of a surface, which are invariant with respect to rotation of the surface underlying the surface normal. That is, the perceived percentage of light reflected at a point does not change if the surface under the viewer is rotated. Thus, BRDFs simulating light reflection at isotropic materials are independent on the azimuth angle ϕ_i and ϕ_r. They are, apart of θ_i, θ_r only still dependent on the difference $\phi_r - \phi_i$. This implies, that isotropic BRDFs have only three degrees of freedom instead of four, as the BRDF was generally defined. So, an isotropic BRDF can be written as:

$$f_{r,iso}(\mathbf{x}, \theta_i \rightarrow (\theta_r, \phi_r - \phi_i)) \tag{4.149}$$

with $\theta_i, \theta_r \in \left[0, \frac{\pi}{2}\right]$ und $\phi_i, \phi_r \in [0, 2\pi)$. Many materials like flat plastic, steel, and aluminium have isotropic BRDFs.

In general, BRDFs are not isotropic, but rather anisotropic. So, we call a BRDF anisotropic, if it describes the reflection behavior of a material that changes the perceived percentage of light reflected with respect to rotation of the surface around the surface normal, see Figure 4.33. In practice, most real-world material surfaces are anisotropic to some degree. This effect can be observed in particular by the illumination of brushed metal, for an exampl, vinyl, compact discs or several types of textiles or hair, see Figure

4.34.

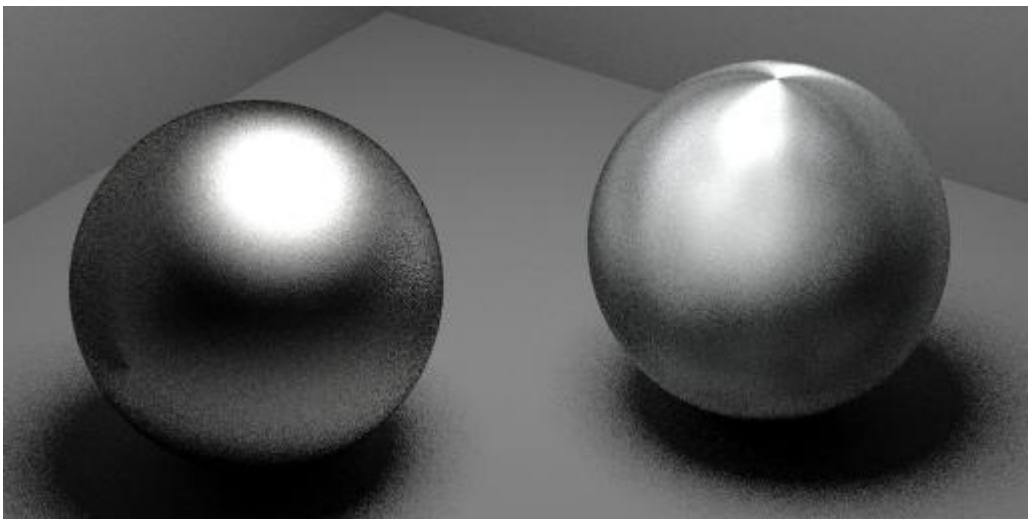


FIGURE 4.34: COMPARISON OF ISOTROPIC VS. ANISOTROPIC ALUMINUM BRDF. The spheres are rendered with two different BRDFs for brushed aluminum. Note the different specular highlight shapes from the anisotropic model in the right image. The surface orientation affects the appearance, revealing the anisotropic behavior of the reflected light. Image courtesy by Stephen H. Westin, Cornell University.

GENERALIZED REFLECTANCE. The BRDF, defined as a derivative, is primarily useful as a theoretical construct for describing the interaction of light at object surfaces since infinitesimal elements such as the solid angle do not include measurable amounts of radiant flux. This can be demonstrated in particular at the example of the ideal specular BRDF which, in form of a Dirac δ-distribution, can take on values from 0 to infinity. This means that the ideal specular BRDF, formulated as a Dirac δ-distribution, is not suitable for practical applications.

σ (87)
Flux (249)
Ideal Specular BRDF (325)
Dirac δ-distribution (118)

So, we need a physically concept that describes the process of interaction of light at surfaces expressed in terms of the BRDF, and which is also useful in practice. This concept is given in the definition of the *generalized reflectance*.

Generally, reflectance is a directional quantity, that is, a function of the reflected and incident amount of light at a surface, which is also wavelength and polarization dependent on the incident and reflected light. However, we discuss the reflectance in a wavelength and polarization independent version.

DEFINITION 4.19 (Generalized Reflectance, ρ.) *The ratio of flux reflected from a differential surface patch* $d\mu^2(\mathbf{s})$ *around a point* $\mathbf{s} \in \partial\mathcal{V}$ *into direction* ω_o*, respectively a*

∂V (41)

finite solid angle Υ_o*, to the incident flux at* $d\mu^2(s)$ *through direction* ω_i*, respectively a finite solid angle* Υ_i*, is called the* generalized reflectance *of* $d\mu^2(s)$*, commonly denoted by* ρ*. It is defined as:*

$$\rho(s, \Upsilon_i \to \Upsilon_o) \stackrel{def}{=} \frac{d\Phi_o(s, \Upsilon_o)}{d\Phi_i(s, \Upsilon_i)} \tag{4.150}$$

with

$$d\Phi_o(s, \Upsilon_o) \stackrel{def}{=} \begin{cases} d\mu^2(s)\, L_o(s, \omega_o)\, d\sigma_s^\perp(\omega_o) & \text{if } \Upsilon_o = \omega_o \\ d\mu^2(s) \int_{\Upsilon_o} L_o(s, \omega_o)\, d\sigma_s^\perp(\omega_o) & \text{otherwise} \end{cases} \tag{4.151}$$

and

$$d\Phi_i(s, \Upsilon_i) \stackrel{def}{=} \begin{cases} d\mu^2(s)\, L_i(s, \omega_i)\, d\sigma_s^\perp(\omega_i) & \text{if } \Upsilon_i = \omega_i \\ d\mu^2(s) \int_{\Upsilon_i} L_i(s, \omega_i)\, d\sigma_s^\perp(\omega_i) & \text{otherwise.} \end{cases} \tag{4.152}$$

Due to the principle of conservation of energy, for physically valid materials it has to hold:

$$0 \le \rho(s, \Upsilon_i \to \Upsilon_o) \le 1. \tag{4.153}$$

We defined the reflectance in dependence on incident and exitant solid angle Υ_i and Υ_o. Since any of these both solid angles can be chosen as infinitesimal, finite, or as the entire hemisphere, there are nine different reflectances. Due to [36, Cohen &. Wallace, 1993], the names of these reflectances can be formed by combining the words: *directional*, *conical*, and *hemispherical*, where directional corresponds to a differential solid angle, conical stands for a finite solid angle, and hemispherical corresponds to a solid angle over the entire hemisphere.

The most interesting types of reflectance in computer graphics are the *directional-hemispherical reflectance*, already known from our discussion about the conservation of energy, as well as the concept of the *hemispherical-hemispherical reflectance*.

LEMMA 4.3 (Hemispherical-hemispherical Reflectance) *Based on the definition of the generalized reflectance, let us assume* $\Upsilon_i = \mathcal{H}_i^2$ *and* $\Upsilon_o = \mathcal{H}_o^2$*. Let us further assume, that the incident radiation is uniform and isotropic within the incident solid angle* Υ_i*, then the* hemispherical-hemispherical reflectance *can be expressed in terms of the BRDF as:*

$$\rho(s, \mathcal{H}_i^2 \to \mathcal{H}_o^2) \quad = \quad \frac{1}{\pi} \int_{\mathcal{H}_i^2(s)} \int_{\mathcal{H}_o^2(s)} f_r(s, \omega_i \to \omega_o)\, d\sigma_s^\perp(\omega_i)\, d\sigma_s^\perp(\omega_o), \tag{4.154}$$

where both, $\mathcal{H}_i^2$ *and* $\mathcal{H}_o^2$*, correspond to the upper and the lower hemisphere. The hemispherical-hemispherical reflectance is a constant that gives the fraction of incident light, reflected by a surface under the condition that the incident light is the same from all directions.*

PROOF 4.3 *Substituting* $\Upsilon_i = \mathcal{H}^2_i(\mathbf{s})$ *and* $\Upsilon_o = \mathcal{H}^2_o(\mathbf{s})$ *with* $\mathbf{s} \in \partial\mathcal{V}$ *in Equation (4.150) then we get:*

$$\rho(\mathbf{s}, \mathcal{H}^2_i \to \mathcal{H}^2_o) \overset{(4.150)}{=} \frac{d\Phi_o(\mathbf{s}, \mathcal{H}^2_o)}{d\Phi_i(\mathbf{s}, \mathcal{H}^2_i)} \tag{4.155}$$

$$\overset{(3.15)}{=} \frac{d\mu^2(\mathbf{s}) \int_{\mathcal{H}^2_o(\mathbf{s})} L_o(\mathbf{s}, \omega_o)\, d\sigma^\perp_{\mathbf{s}}(\omega_o)}{d\mu^2(\mathbf{s}) \int_{\mathcal{H}^2_i(\mathbf{s})} L_i(\mathbf{s}, \omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i)} \tag{4.156}$$

$$= \frac{\int_{\mathcal{H}^2_o(\mathbf{s})} L_o(\mathbf{s}, \omega_o)\, d\sigma^\perp_{\mathbf{s}}(\omega_o)}{\int_{\mathcal{H}^2_i(\mathbf{s})} L_i(\mathbf{s}, \omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i)} \tag{4.157}$$

$$\overset{(4.89)}{=} \frac{\int_{\mathcal{H}^2_o(\mathbf{s})} \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_o)}{\int_{\mathcal{H}^2_i(\mathbf{s})} L_i(\mathbf{s}, \omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i)}. \tag{4.158}$$

Using the condition that the incident radiation is uniform and isotropic within the incident beam, then L_i *is constant and can be moved outside the integrals, in both, the numerator and denominator, so that it cancels out resulting in:*

$$\rho(\mathbf{s}, \mathcal{H}^2_i \to \mathcal{H}^2_o) = \frac{\int_{\mathcal{H}^2_o(\mathbf{s})} \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, d\sigma^\perp_{\mathbf{s}}(\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_o)}{\int_{\mathcal{H}^2_i(\mathbf{s})} d\sigma^\perp_{\mathbf{s}}(\omega_i)} \tag{4.159}$$

$$\overset{(2.300)}{=} \frac{1}{\pi} \int_{\mathcal{H}^2_o(\mathbf{s})} \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, d\sigma^\perp_{\mathbf{s}}(\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_o). \tag{4.160}$$

Due to [135, Nicodemus & al. 1977], the condition that L_i *is constant in all points and in all directions within the incident angle is fairly well approximated in any well-designed reflectometer, i.e. the hemispherical-hemispherical reflectance is very useful for describing the observed reflectance of a surface.*

REMARK 4.16 (Further Reflectance Concepts) *Based on the definition of the generalized reflectance, we can derive—dependent on the choice of the incident and exitant solid angles* Υ_i *and* Υ_o*—further reflectance concepts. Thus, e.g. the* directional-hemispherical reflectance *and the* hemispherical-directional reflectance *defined by:*

$$\rho(\mathbf{s}, \omega_i \to \mathcal{H}^2_o) = \int_{\mathcal{H}^2_o(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, d\sigma^\perp_{\mathbf{s}}(\omega_o), \tag{4.161}$$

respectively

$$\rho(\mathbf{s}, \mathcal{H}^2_i \to \omega_o) = \frac{d\sigma^\perp(\omega_o)}{\pi} \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, d\sigma^\perp_{\mathbf{s}}(\omega_i) \tag{4.162}$$

as well as the directional-directional reflectance*:*

$$\rho(\mathbf{s}, \omega_i \to \omega_o) = d\sigma^\perp(\omega_o)\, f_r(\mathbf{s}, \omega_i \to \omega_o). \tag{4.163}$$

For one thing they give the amount of light arriving at point $\mathbf{s}$ *from direction* ω_i *reflected in all directions of the hemisphere, see Relation (4.136), and for another the portion of incident light over the whole hemisphere reflected in direction* ω_o*. Together with the hemispherical-hemispherical reflectance, these reflectance types are the most important for our further discussions.*

REMARK 4.17 (Shortened Notation of the Four Common Types of Reflectance) *For abbreviation of our future formulas, we use the following shortened notations of reflectance types:*

$$\rho_{dd}(\mathbf{s}) \equiv \rho(\mathbf{s}, \omega_i \to \omega_o) \tag{4.164}$$
$$\rho_{dh}(\mathbf{s}) \equiv \rho(\mathbf{s}, \omega_i \to \mathcal{H}^2_o) \tag{4.165}$$
$$\rho_{hd}(\mathbf{s}) \equiv \rho(\mathbf{s}, \mathcal{H}^2_i \to \omega_o) \tag{4.166}$$
$$\rho_{hh}(\mathbf{s}) \equiv \rho(\mathbf{s}, \mathcal{H}^2_i \to \mathcal{H}^2_o). \tag{4.167}$$

Based on Theorem 4.2, now, we will derive an exact formula for the ideal diffuse BRDF f^o_r, which also shows a relation between our theoretical concept of the BRDF and the rather practical concept of reflectance.

LEMMA 4.4 (The Ideal Diffuse and the Diffuse BRDF, f^o_r) *The ideal diffuse BRDF* f^o_r*, which satisfies the conservation of energy, is given by the following formula:*

$$f^o_r(\mathbf{s}, \omega_i \to \omega_o) = \frac{1}{\pi}. \tag{4.168}$$

Since there are no ideal diffuse surfaces in real world, the BRDF approximating the reflection behavior of light at a real diffuse surface is given by:

$$f^o_r(\mathbf{s}, \omega_i \to \omega_o) = \frac{\rho_{hh}(\mathbf{s})}{\pi}. \tag{4.169}$$

PROOF 4.4 *We know from Theorem 4.2 that the ideal diffuse BRDF* f^o_r *is constant. Requiring that* f^o_r *also satisfies the conservation of energy then, with the definition of the hemispherical-hemispherical reflectance* ρ_{hh}*, from Lemma (4.3), it must hold:*

$$\rho_{hh}(\mathbf{s}) = \frac{1}{\pi}\int_{\mathcal{H}^2_i(\mathbf{s})}\int_{\mathcal{H}^2_o(\mathbf{s})} f^o_r(\mathbf{s}, \omega_i \to \omega_o)\, d\sigma^\perp_{\mathbf{s}}(\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_o) \tag{4.170}$$
$$= \frac{1}{\pi} f^o_r \int_{\mathcal{H}^2_i(\mathbf{s})}\int_{\mathcal{H}^2_o(\mathbf{s})} d\sigma^\perp_{\mathbf{s}}(\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_o) \tag{4.171}$$
$$= \frac{1}{\pi} f^o_r \int_{\mathcal{H}^2_i(\mathbf{s})} d\sigma^\perp_{\mathbf{s}}(\omega_i) \int_{\mathcal{H}^2_o(\mathbf{s})} d\sigma^\perp_{\mathbf{s}}(\omega_o) \tag{4.172}$$
$$\overset{(2.300)}{=} \frac{1}{\pi} f^o_r\, \pi^2 \tag{4.173}$$
$$= f^o_r \pi. \tag{4.174}$$

With Relation (4.174), we now obtain a formulation for the diffuse BRDF f_r^o *as a quotient of the hemispherical-hemispherical reflectance and the constant* π*, namely:* Lambertian Reflector (349)

$$f_r^o(s, \omega_i \to \omega_o) = \frac{\rho_{hh}(s)}{\pi}. \tag{4.175}$$

Ignoring any possible absorption of light particles in the underlying reflection process then it even holds:

$$f_r^o(s, \omega_i \to \omega_o) \overset{\rho_{hh}(s)=1}{=} \frac{1}{\pi}. \tag{4.176}$$

REMARK 4.18 *The result from Lemma 4.4 can also be achieved via the definition of the directional-hemispherical reflectance instead using the hemispherical-hemispherical reflectance. Under the assumption, that the ideal diffuse BRDF* f_r^o *is constant it holds:*

$$\begin{aligned} \rho_{dh}(s) &= \int_{\mathcal{H}_o^2(s)} f_r(s, \omega_i \to \omega_o)\, d\sigma_s^\perp(\omega_o) && (4.177)\\ &= f_r^o \int_{\mathcal{H}_o^2(s)} d\sigma_s^\perp(\omega_o) && (4.178)\\ &\overset{(2.300)}{=} f_r^o\, \pi && (4.179) \end{aligned}$$

which implies

$$f_r^o(s, \omega_i \to \omega_o) = \frac{\rho_{dh}(s)}{\pi}. \tag{4.180}$$

GENERALIZED TRANSMITTANCE, τ. Similar to the concept of reflectance, we can now also define the *transmittance* of a surface. As both concepts are defined as the fraction of outgoing flux to incident flux, we define the transmittance, τ, in terms of the already known concept of reflectance.

DEFINITION 4.20 (Generalized Transmittance) *Let* $A \in \partial V$ *be the interface between two media with refraction indices* η_i *and* η_t*. The ratio of flux transmitted from a differential surface patch* $d\mu^2(s)$ *around a point* $s \in \partial V$ *into direction* ω_t*, respectively a finite solid angle* Υ_t*, to the incident flux at* $d\mu^2(s)$ *through direction* ω_i*, respectively a finite solid angle* Υ_i*, is called the* generalized transmittance *of* $d\mu^2(s)$*, denoted by* τ*. It is defined as* ∂V (41)

$$\tau(s, \Upsilon_i \to \Upsilon_t) \overset{def}{=} \frac{d\Phi_t(s, \Upsilon_t)}{d\Phi_i(s, \Upsilon_i)} \tag{4.181}$$

with

$$d\Phi_t(s, \Upsilon_t) \overset{def}{=} \begin{cases} d\mu^2(s)\, L_o(s, \omega_t)\, d\sigma_s^\perp(\omega_t) & \text{if } \Upsilon_t = \omega_t \\ d\mu^2(s) \int_{\Upsilon_t} L_o(s, \omega_t)\, d\sigma_s^\perp(\omega_t) & \text{otherwise} \end{cases} \tag{4.182}$$

and

$$d\Phi_i(s,\Upsilon_i) \stackrel{def}{=} \begin{cases} d\mu^2(s)\, L_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i) & \text{if } \Upsilon_i = \omega_i \\ d\mu^2(s) \int_{\Upsilon_i} L_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i) & \text{otherwise.} \end{cases} \tag{4.183}$$

Using Equation (4.127) for expressing the transmitted radiance L_t in terms of incident radiance L_i then the transmittance, τ, can also be expressed in terms of the reflectance ρ, namely by:

$$\tau(s,\Upsilon_i \to \Upsilon_t) = \frac{\eta_t^2}{\eta_i^2}\rho(s,\Upsilon_i \to \Upsilon_t). \tag{4.184}$$

REMARK 4.19 (Reflectance and Transmittance of Non-absorbing and Non-emitting Surfaces) *With our definition of reflectance and transmittance, obviously it holds:*

$$\begin{aligned} \frac{d\Phi_o(s,\Upsilon_t)}{d\Phi_i(s,\Upsilon_i)} &= \frac{d\Phi_r(s,\Upsilon_r) + d\Phi_t(s,\Upsilon_t)}{d\Phi_i(s,\Upsilon_i)} && (4.185)\\ &= \frac{d\Phi_r(s,\Upsilon_r)}{d\Phi_i(s,\Upsilon_i)} + \frac{d\Phi_t(s,\Upsilon_t)}{d\Phi_i(s,\Upsilon_i)} && (4.186)\\ &= \rho(s,\Upsilon_i \to \Upsilon_t) + \tau(s,\Upsilon_i \to \Upsilon_t). && (4.187) \end{aligned}$$

As for non-absorbing and non-emitting surfaces it holds $\frac{d\Phi_o}{d\Phi_i} = 1$, we can then express the reflectance of such a kind of surface in terms of transmittance or vice-versa, thus:

$$\begin{aligned} \rho(s,\Upsilon_i \to \Upsilon_t) &= 1 - \tau(s,\Upsilon_i \to \Upsilon_t) && (4.188)\\ \tau(s,\Upsilon_i \to \Upsilon_t) &= 1 - \rho(s,\Upsilon_i \to \Upsilon_t) && (4.189) \end{aligned}$$

as well as in terms of the involved refraction indices, that is:

$$\begin{aligned} \rho(s,\Upsilon_i \to \Upsilon_t) &\stackrel{(4.184)}{=} \frac{1}{1+\frac{\eta_t^2}{\eta_i^2}} && (4.190)\\ &= \frac{\eta_i^2}{\eta_i^2+\eta_t^2} && (4.191) \end{aligned}$$

$$\begin{aligned} \tau(s,\Upsilon_i \to \Upsilon_t) &\stackrel{(4.184)}{=} \frac{1}{1+\frac{\eta_i^2}{\eta_t^2}} && (4.192)\\ &= \frac{\eta_t^2}{\eta_t^2+\eta_i^2}. && (4.193) \end{aligned}$$

REMARK 4.20 (Shortened Notation of the Four Common Types of Transmittance) *For abbreviation of our future formulas, we use the following shortened notation of transmittance types:*

$$\tau_{dd}(\mathbf{s}) \equiv \tau(\mathbf{s}, \omega_i \to \omega_o) = \frac{\eta_t^2}{\eta_i^2}\rho(\mathbf{s}, \omega_i \to \omega_o) \tag{4.194}$$

$$\tau_{dh}(\mathbf{s}) \equiv \tau(\mathbf{s}, \omega_i \to \mathcal{H}_o^2) = \frac{\eta_t^2}{\eta_i^2}\rho(\mathbf{s}, \omega_i \to \mathcal{H}_o^2) \tag{4.195}$$

$$\tau_{hd}(\mathbf{s}) \equiv \tau(\mathbf{s}, \mathcal{H}_i^2 \to \omega_o) = \frac{\eta_t^2}{\eta_i^2}\rho(\mathbf{s}, \mathcal{H}_i^2 \to \omega_o) \tag{4.196}$$

$$\tau_{hh}(\mathbf{s}) \equiv \tau(\mathbf{s}, \mathcal{H}_i^2 \to \mathcal{H}_o^2) = \frac{\eta_t^2}{\eta_i^2}\rho(\mathbf{s}, \mathcal{H}_i^2 \to \mathcal{H}_o^2). \tag{4.197}$$

COMPONENTS OF A BRDF AND A BTDF. The ideal reflection types described in this and the previous section can not be found in nature. In real world, we encounter frequently a reflection type that can be specified to be between these limiting cases: the so-called *directional diffuse reflection*, also referred to as *glossy reflection*. Here, light from an infinitesimal thin light beam is scattered and spread into some finite solid angle typically around the perfect reflection direction ω_o, see Figure 4.35. Due to [190, Sillion & Puech 1994], a detailed description of directional diffuse reflection is difficult, because it requires accurate statements about the interaction of light with the irregularities of the surface, which are comparable with the wavelength of light. Since physical optics, thus the field of optics, which considers light as a wave, is outside the scope of this book, we are not interested in the wave properties of light. Based on the principles of geometric optics, then an approach to approximate such a type of BRDF is the decomposition of a BRDF into a diffuse part, f_r^o, a specular part, $f_r^\vee$, and a so-called *glossy* or *directional diffuse part*, f_r^{gl}, that is:

Glossy Reflection (304)
f_r^o (339)
$f_r^\vee$ (325)

$$f_r = f_r^o + f_r^\vee + f_r^{gl}, \tag{4.198}$$

see Figure 4.36.

For the most BRDF models—as shown in Section 4.2.2.5—this composition is to restrictive. So, the ideal specular component $f_r^\vee$ is mostly already incorporated in the glossy component—which we then simply call the specular component f_r^{sp}—that is, we can write:

$$f_r^{sp} = f_r^\vee + f_r^{gl}. \tag{4.199}$$

This means that a BRDF is commonly assumed to be the composition of a diffuse and a specular term, namely:

$$f_r = f_r^o + f_r^{sp}. \tag{4.200}$$

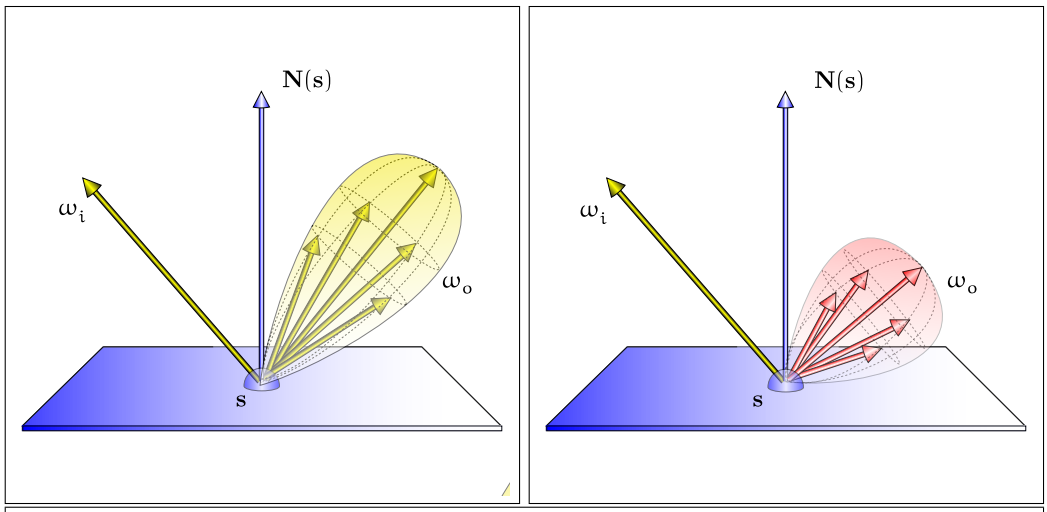


FIGURE 4.35: GLOSS OR DIRECTIONAL DIFFUSE REFLECTION. Visualization of a gloss, or a directional-diffuse reflective surface. Note: The degree of glossiness determines the lobe of set of possible reflected directions. The surface visualized in the left image is more gloss than the surface in the right image. This is also the reason why the lobe of outgoing directions in the left image is smaller than in the image on the right-hand side.

REMARK 4.21 (Composition of a BTDF) *Similar to the representation of a BRDF from Equation (4.198), also a BTDF* f_t *can be composed of three more basic transmission models, namely: a diffuse, a specular, and a glossy refraction model. So, a BTDF can also be written as*

$$f_t = f_t^o + f_t^\vee + f_t^{gl}, \tag{4.201}$$

where f_t^o, $f_t^\vee$, *and* f_t^{gl} *are defined as the counter parts of* $f_r^o, f_r^\vee$, *and* f_r^{gl}.

4.2.2.4 MEASUREMENTS AND REPRESENTATIONS OF BRDFS

In the last sections we introduced the mathematical construct of the bidirectional distribution function and presented theoretical constructed BRDFs, which describe the reflectance behavior at idealized smooth and diffuse surfaces. One question that now often arises is: How is it possible to represent the reflection behavior at real-world surfaces by means of a BRDF? In principle, there are mainly two ways to obtain such a BRDF that can be used as a reflection model in realistic image synthesis algorithms. One way is to construct closed-form mathematical functions derived from analytical models using physical principles, and the other is based on the resampling of BRDF data, acquired by empirical measurements of real-world surfaces.

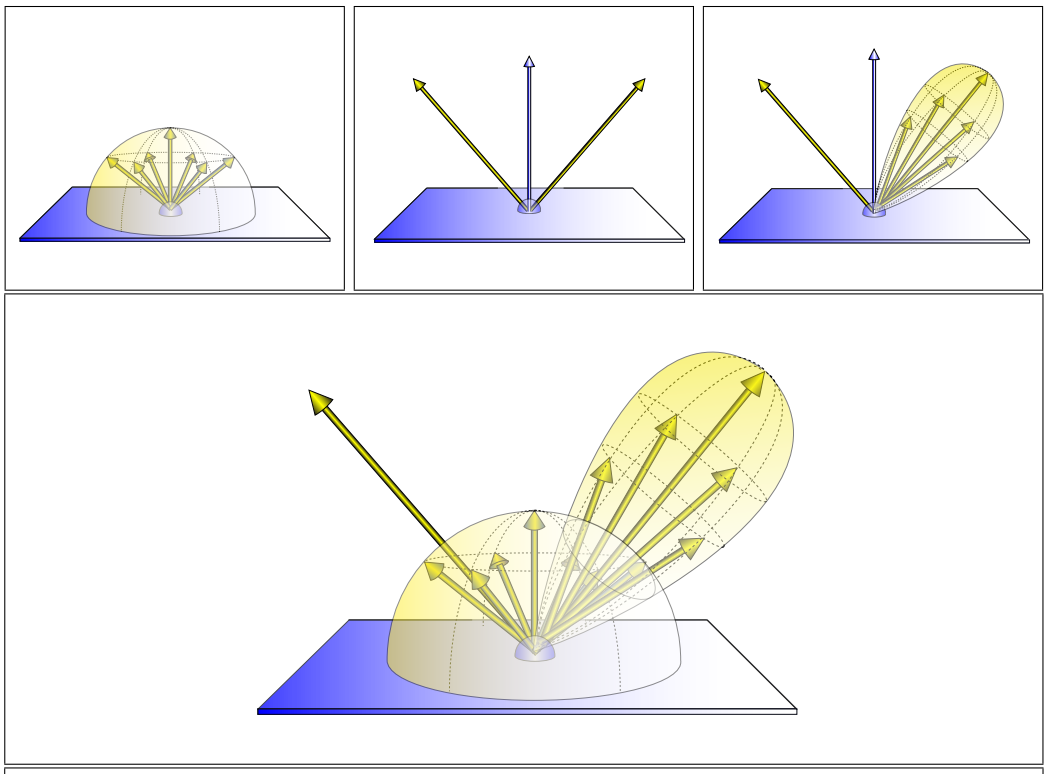


FIGURE 4.36: COMPOSITION OF A BRDF. As real materials are not perfectly diffuse or specular, in practice, a BRDF, as shown in the lower image, is often treated as a sum of three different components: A Lambertian diffuse component, f_r^o, an ideal specular component, $f_r^\vee$, and a glossy or directional diffuse component, f_r^{gl}.

BRDF REPRESENTATION VIA ANALYTICAL MODELS. BRDFs, represented by closed-form functions, are generally derived from some physical theory of how light reflects from a surface. Apart from the incident and exitant directions ω_i and ω_o they depend on other parameters, which are used to control the reflectance properties and the surface appearance of a material. Such a function can be given by a very simple but also a relatively complex expression.

Reflectance (336)

Thus, as we will see in the next section, there are actually many simple theoretical models—such as the *Phong* or the *Blinn-Phong illumination model*—which can be used to describe a very wide range of visually interesting lighting effects. Although they are rarely based on exact physical properties of light and matters, they can passably model some materials, in particular plastics. Other, more complex and more physically plausible BRDF models, such as the *Cook-Torrance BRDF* or the *Ward* illumination model attempt to model physical reality.

Phong Model (353)
Blinn-Phong Model (357)
Cook-Torrence Model (361)
Ward Model (369)

ACQUIRED BRDF DATA AND MEASURING A BRDF. In many situations, analytical approaches of a BRDF cannot accurately model the reflection of light at a real world surface. Considering e.g. materials such as leather or hair, it would be a very large challenge to approximate the reflection properties of such complex materials by closed-form BRDFs. Since in particular organic materials have such a very complex microstructure, which is hard to simulate theoretically, it is better to measure the light reflected by a surface instead of to create a closed-form function to approximate it. A BRDF created in this way is called a *measured BRDF* and data produced in this way are often referred to as *acquired BRDF data*.

A device used for measuring BRDFs is called a *gonioreflectometer*, see Figure 4.37. The sample to be measured is placed at the center of the device, while a light source and detector are moving about the hemisphere above the sample, and measure the reflectance for every few degrees. The output of such a measurement is a list of values, parameterized by ω_i and ω_o, which can then be used as a lookup table during rendering. This is the reason, why we also speak of a *tabulated BRDFs*. Depending on the sampling rate, n_i incident direction samples and n_o exitant direction samples, such a table, of size $n_i \cdot n_o$, can be very large. Since in a rendering algorithm is often difficult to determine which analytical model should be used to simulate certain visual lighting effects, it is often advantageous to use such measured BRDFs.

After discussing, how we obtain a BRDF as a very large set of data, we are now interested in techniques to store and compute them efficiently. In [175, Rusinkiewicz 1997] a variety of such techniques is presented from those one is in particular interesting for us: the method of storing BRDFs by projecting them onto *spherical harmonics*.

BRDF REPRESENTATION BY SPHERICAL HARMONICS. Obviously, a radiance distribution, L, can be interpreted as a function defined on the unit sphere or the hemisphere, that is, it can be expressed as a series expansion using the spherical harmonic basis functions $Y_{l,m}$. Thus, L can be written as

Spherical Harmonics (123)

$Y_{l,m}$ (124)

$$L(\mathbf{x},\omega) = \sum_{l=0}^{\infty}\sum_{m=-l}^{l} C_{l,m}Y_{l,m}(\omega) \tag{4.202}$$

$$\overset{i=l(l+1)+m}{=} \sum_{i}^{n} C_i Y_i(\omega), \tag{4.203}$$

where the coefficients, $C_{l,m}$, of the series are determined in an analogous way as the coefficients of a Fourier series expansion of a function. If the function L is known, or at least known at a number of samples, then the coefficients $C_{l,m}$ can be computed via Formula (2.344), or they can be approximated by solving a series of linear equations using the function values at these samples. Forming a n-dimensional vector $\mathbf{C} = (C_1, \ldots, C_n)^T$

Fourier Coefficients (38)

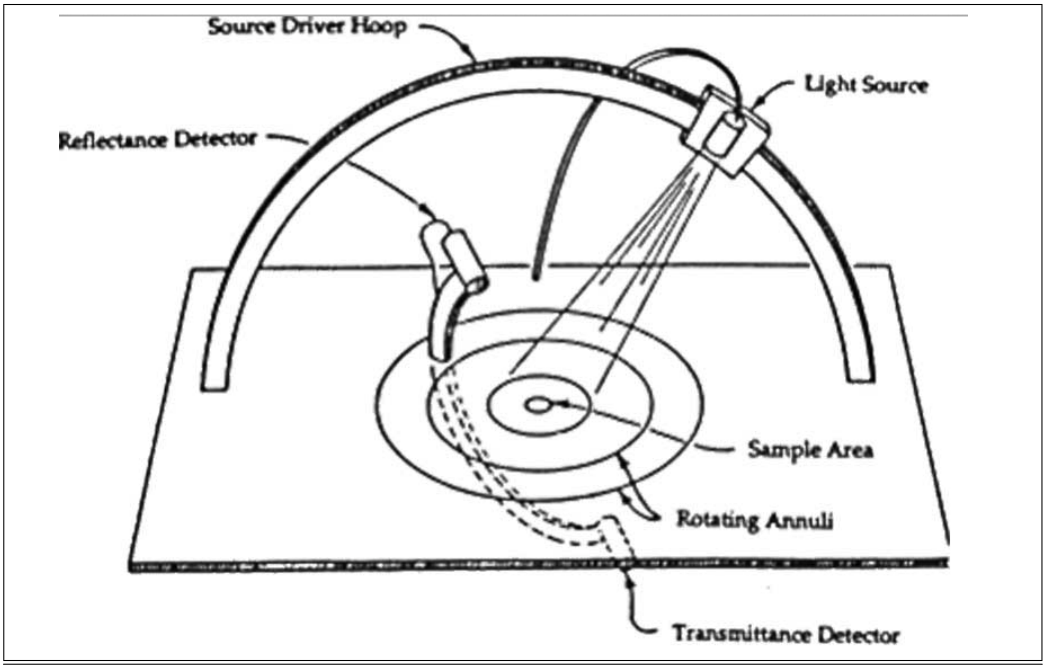


FIGURE 4.37: GONIOREFLECTOMETER FOR MEASURING BRDF DATA. Gonioreflectometer devices are expensive. Since they are also quite slow, a full BRDF measurement at even a low resolution can easily cost thousands of dollars. In addition, due to the number of moving parts in the device, a gonioreflectometer can produce data, which are quite noisy.

using the coefficients $C_{l,m}$, multiplying $\mathbf{C}$ with another vector, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, whose components consist of the basis functions, $Y_{l,m}$, implies that Equation (4.202) becomes

$$L(\mathbf{x}, \omega) = \langle \mathbf{C}, \mathbf{Y}(\omega) \rangle. \tag{4.204}$$

Considering the vector $\mathbf{C}$ as a function of incident direction, then, due to [92, Kainulainen 2003], a BRDF, f_r can be approximated by the dot product

$$f_r(\mathbf{x}, \omega_i \rightarrow \omega_o) \approx \langle \mathbf{C}(\omega_i), \mathbf{Y}(\omega_o) \rangle. \tag{4.205}$$

Fourier Series (39)

$Y_{0,0}$ (124)

Similar to the theory of the Fourier series, the high-order terms represent the high frequency components of the distribution, the low-order terms the low frequency components. That is, while ideal diffuse surfaces can be described via $Y_{0,0}$, the simulation of glossier or more specular-appearing surfaces demand the calculation of further coefficients in the series of expansion of spherical harmonics from Equation (2.343).

EXAMPLE 4.1 (A First Encounter with Precomputed Radiance Transfer, PRT) Precomputed radiance transfer, *also briefly denoted as* PRT, *can be interpreted as a global illumination model, which is well suited for real-time rendering, since the bulk of computation occurs offline. The algorithm separates the light sources in a scene from the transport properties, projects the transport properties to spherical harmonics, and combines them at run time with the illumination functions, which can be made unique. On this way, it is possible to compute the shadowing of an illumination point via evaluation of a dot product, in the case of diffuse reflection, and a vector-matrix multiplication for specular reflection [122, Lauschke 2006].*

Let us consider a scene where all surfaces are assumed to be Lambertian reflectors and the existing light sources are far away, then the reflection equation can be written as

Lambertian Reflector (349)

$$L_o(\mathbf{s},\omega_o) = \int_{\mathcal{H}^2_i(\mathbf{s})} f^o_r(\mathbf{s},\omega_i \to \omega_o) L_i(\mathbf{s},\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i). \tag{4.206}$$

Now, our goal is to express the reflection equation as a product of two functions, which, represented as finite series of spherical harmonic basis functions, are approximations of the illumination function L_i *and the transport property* f^o_r *and* $\cos\theta_i$, *hidden in the differential projected solid angle* $d\sigma^\perp$. *As the diffuse BRDF* $f^o_r = \frac{\rho_{dh}}{\pi}$ *is constant, it holds:*

$$L_o(\mathbf{s},\omega_o) = \frac{\rho_{dh}(\mathbf{s})}{\pi} \int_{\mathcal{H}^2_i(\mathbf{s})} L_i(\mathbf{s},\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i) \tag{4.207}$$

$$= \frac{\rho_{dh}(\mathbf{s})}{\pi} \int_{\mathcal{H}^2_i(\mathbf{s})} L_i(\mathbf{s},\omega_i)\, |\cos\theta_i|\, d\sigma_{\mathbf{s}}(\omega_i). \tag{4.208}$$

Let us assume that it holds $\frac{\rho_{dh}}{\pi} = \frac{1}{\pi}$, *that is, ideal diffuse reflection occurs at surface point* $\mathbf{s}$, *then Relation (4.208) can be written as*

$$L_o(\mathbf{s},\omega_o) \overset{Y_{0,0}(\omega_i)=\frac{1}{\sqrt{4\pi}}}{=} \frac{2}{\sqrt{\pi}} Y_{0,0}(\omega_i) \int_{\mathcal{H}^2_i(\mathbf{s})} L_i(\mathbf{s},\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i) \tag{4.209}$$

$$= \frac{2\,Y_{0,0}(\omega_i)}{\sqrt{\pi}} \int_{\mathcal{H}^2_i(\mathbf{s})} L_i(\mathbf{s},\omega_i)\, |\cos\theta_i|\, d\sigma_{\mathbf{s}}(\omega_i). \tag{4.210}$$

Writing now the function L_i *as a finite series of spherical harmonic basis functions, thus*

$$L_i(\mathbf{s},\omega_i) = \sum_{l=0}^{n} \sum_{m=-l}^{l} C_{l,m} Y_{l,m}(\omega_i) \tag{4.211}$$

with SH coefficients

$$C_{l,m} = \int_{\mathcal{H}^2_i(\mathbf{s})} L_i(\mathbf{s},\omega_i) Y_{l,m}(\omega_i)\, d\sigma_{\mathbf{s}}(\omega_i) \tag{4.212}$$

and substituting the incident radiance L_i *in Equation (4.208) by the Expression (4.211), then we obtain*

$$L_o(s,\omega_o) = \frac{1}{\pi}\int_{\mathcal{H}^2_i(s)}\left(\sum_{l=0}^{n}\sum_{m=-l}^{l} C_{l,m}Y_{l,m}(\omega_i)\right)|\cos\theta_i|\,d\sigma_s(\omega_i) \quad (4.213)$$

$$= \frac{1}{\pi}\sum_{l=0}^{n}\sum_{m=-l}^{l}\int_{\mathcal{H}^2_i(s)} C_{l,m}Y_{l,m}(\omega_i)\,|\cos\theta_i|\,d\sigma_s(\omega_i) \quad (4.214)$$

$$\overset{(2.344)}{=} \frac{1}{\pi}\sum_{l=0}^{n}\sum_{m=-l}^{l} C_{l,m}\underbrace{\int_{\mathcal{H}^2_i(s)} Y_{l,m}(\omega_i)\,|\cos\theta_i|\,d\sigma_s(\omega_i)}_{=C'_{l,m}} \quad (4.215)$$

$$= \frac{1}{\pi}\sum_{l=0}^{n}\sum_{m=-l}^{l} C_{l,m}C'_{l,m}. \quad (4.216)$$

The last equation can be considered as the product of $\frac{1}{\pi}$ *and the dot product of the vectors consisting of the SH coefficients* $C_{l,m}$ *and* $C'_{l,m}$.

4.2.2.5 BRDF MODELS

The reflection behavior of most materials occurring in nature is so complex that it can not simply be described by a specular, a diffuse, or a glossy BRDF. The BRDFs of such materials are very complex, and in most cases, an exact simulation of the reflection behavior is not possible. Today, computer graphics distinguishes between four classes of BRDFs depending on the sources they come from:

Section 4.2.2.5.1
- *idealized* BRDF models,

Section 4.2.2.5.2
- *phenomenological* BRDF models,

Section 4.2.2.5.3
- *physical-based* or *physics-inspired* BRDF models, and

Section 4.2.2.5.4
- BRDF models based on *measured* data.

Let us now present the most interesting examples from any of these classes of BRDF models.

4.2.2.5.1 IDEALIZED BRDF MODELS

In Section 4.2.2.2, we have introduced the concept of idealized BRDF. So, we know the ideal specular BRDF, $f_r^{\vee}$, and the ideal diffuse BRDF, f_r^{o}. They represent formulas that describe the behavior of light at idealized surfaces.

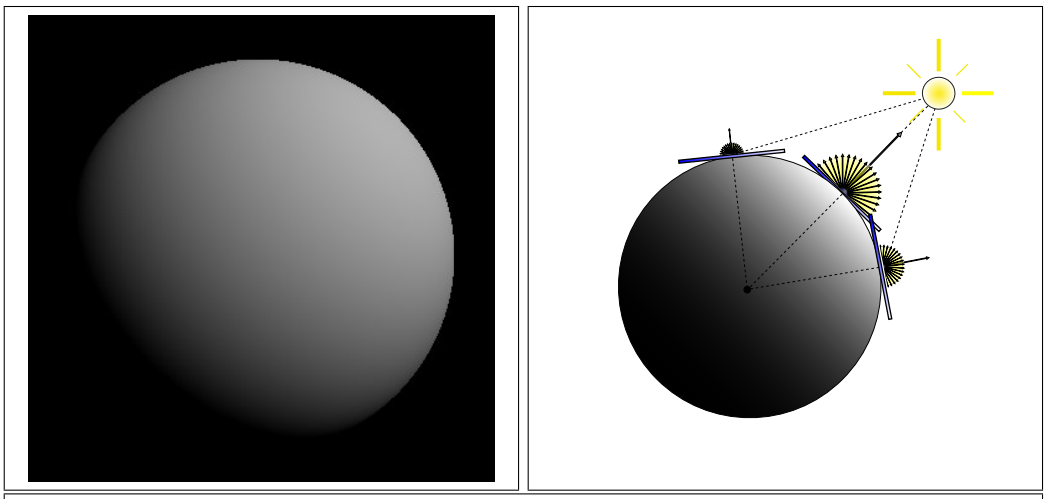


FIGURE 4.38: SPHERES RENDERED WITH THE LAMBERTIAN ILLUMINATION MODEL. The left image shows a sphere rendered with the Lambertian BRDF. The image on the right-hand side illustrates the dependence of irradiance on the angle between the surface normal at an illumination point and the direction towards a light source.

THE LAMBERTIAN ILLUMINATION MODEL. The *Lambertian illumination model* assumes that light incident at a surface is scattered in such a way, that the appearance of the surface is independent of the observer's angle of view, see Figure 4.38. Mathematically expressed, this means, that light falling on any surface point is uniformly scattered in all directions. As described in detail in Section 4.2.1.2, there are two reason that are responsible for this effect: the microgeometry of the underlying rough surface or the phenomen of subsurface scattering of light penetrated into the material. Such a perfectly diffuse surface is also called a *Lambertian reflector*. Although they mathematically conserve energy, Lambertian reflectors are impossible in nature due to thermodynamic reasons. A typical example of a Lambertian reflector is unfinished wood.

DEFINITION 4.21 (The Lambertian BRDF) *Let f_r^o denote the ideal diffuse BRDF furthermore let ρ_{dh} as well as ρ_{hh} denote the directional-hemispherical as well as the hemispherical-hemispherical reflectance. Then, the* Lambertian BRDF f_r^{LA} *is defined as:* ρ_{dh} (338) ρ_{hh} (337)

$$f_r^{LA}(\mathbf{s}) \overset{def}{=} f_r^o(\mathbf{s}) \tag{4.217}$$

$$\overset{(4.175)}{=} \frac{\rho_{hh}(\mathbf{s})}{\pi} \tag{4.218}$$

$$\overset{(4.180)}{=} \frac{\rho_{dh}(\mathbf{s})}{\pi}. \tag{4.219}$$

REMARK 4.22 *Obviously, the Lambertian BRDF is physically plausible, since it satis-*

fies the condition of non-negativity, Helmholtz-reciprocity and energy conservation. We leave this simple proof to the interested reader.

The definition above shows, that the Lambertian BRDF is independent on a directional variable, which is also the reason, why we can express it without any directional variable. Based on the Lambertian BRDF, the *Lambertian illumination model* can now be defined by:

DEFINITION 4.22 (The Lambertian Illumination Model) *Let ω_i be the direction vector of the incident radiance at surface point $\mathbf{s}$ and $\mathbf{N}(\mathbf{s})$ denotes as usual the normal at surface point $\mathbf{s}$. Then, the* Lambertian illumination model *is defined by:*

$$L_o(\mathbf{s}) \stackrel{def}{=} \int_{\mathcal{H}_i^2(\mathbf{s})} f_r^{LA}(\mathbf{s})\, L_i(\mathbf{s},\omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i) \tag{4.220}$$

$$= f_r^{LA}(\mathbf{s}) \int_{\mathcal{H}_i^2(\mathbf{s})} L_i(\mathbf{s},\omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i) \tag{4.221}$$

$$\stackrel{(3.46)}{=} f_r^{LA}(\mathbf{s})\, E(\mathbf{s}) \tag{4.222}$$

$$= f_r^{LA}(\mathbf{s})\, L_i(\mathbf{s})\, |\cos\theta_i|, \tag{4.223}$$

where $\cos\theta_i$ is the angle between the normal at point $\mathbf{x}$ and the direction towards the light source.

Obviously, the reflected radiance is proportional to the incident irradiance—it is dependent on the angle between the surface normal and the incident light ray—so, it is constant and has the same value in all directions. This means, that the surface does not change brightness as you move your eyes.

The Lambertian illumination is already known since more than 200 years. Since it is very simple, it is one of the most widely used models in computer graphics. The Lambertian illumination model is a purely local illumination model that only accounts for direct light coming from light sources and neglect any reflection on other surfaces. Using the absolute value of the cosine in Formula (4.223) circumvents the problem that the reflected radiance can be negative. Although it is not physically plausible, it is a good approximation for many dull, matte surfaces in real world, such as paper and completely flat paint.

Physically Plausible BRDF (331)

REMARK 4.23 (The Diffuse Reflection Coefficient) *Often, the Lambertian illumination model is also be expressed in terms of a so-called* diffuse reflection coefficient k_d*, a material constant measured in unit $\left[\frac{1}{sr}\right]$, which is defined by:*

$$k_d \stackrel{def}{=} \frac{\rho_{dh}(\mathbf{s})}{\pi} = f_r^{LA}(\mathbf{s}), \tag{4.224}$$

thus,

$$L_o(\mathbf{s}) = k_d\, E(\mathbf{s}). \tag{4.225}$$

In the following example, we show the use of a Lambertian illumination model in a small scene illuminated by a point light source.

EXAMPLE 4.2 *Given be a diffuse surface illuminated by an isotropic point source. Due to our discussion from Example 3.5, the contribution of the light source at point* $\mathbf{x}$ *to surface point* $\mathbf{s}$ *is obviously given by:*

$$E(\mathbf{s}) \overset{(3.99)}{=} \frac{\Phi_e}{4\pi} \frac{|\langle \mathbf{N}(\mathbf{s}), \omega_i \rangle|}{\|\mathbf{s}-\mathbf{x}\|_2^2}, \tag{4.226}$$

where θ_i *is the angle between the surface normal at point* $\mathbf{s}$ *and the direction towards the point light source.*

Based on a Lambertian illumination model, the light reflected at scene point $\mathbf{s}$ *in any direction* $\omega_o \in \mathcal{H}_o^2$ *is*

$$\begin{aligned} L_o(\mathbf{s}, \omega_o) &= k_d E(\mathbf{s}) && (4.227)\\ &= k_d \frac{\Phi_e}{4\pi} \frac{|\langle \mathbf{N}(\mathbf{s}), \omega_i \rangle|}{\|\mathbf{s}-\mathbf{x}\|_2^2} && (4.228)\\ &\overset{(4.224)}{=} \frac{\rho_{dh}(\mathbf{s})\Phi_e}{4\pi^2} \frac{|\cos\theta_i|}{\|\mathbf{s}-\mathbf{x}\|_2^2}. && (4.229) \end{aligned}$$

4.2.2.5.2 PHENOMENOLOGICAL BRDF MODELS

Phenomenological BRDF models are based on scientific methods, which attempt to detect the essential and meaningful in the phenomena of light interaction at object surfaces in a scene to be rendered. Contrary to practical, empirical approaches, these methods choose intuitive accesses to the problem to be solved. The most phenomenological BRDFs are simple equations, controlled by only a few parameters, which are used to describe the properties of real-world surfaces. Since they are generally simple to evaluate and often deliver surprisingly good results, phenomenological BRDF models still play a central role in realistic rendering.

THE PHONG ILLUMINATION MODEL. The still mostly used illumination model in computer graphics is the *Phong model*, derived 1975 and named after his founder Bui Tuong Phong [160, Phong 1975]. The Phong model is phenomenological motivated and tries to reflect the effect of gloss observed at surfaces. Since it is represented in a simple mathematical formula, it is also efficient to evaluate. The classic *Phong Shading Equation* is usually defined in terms of a point light source and composed of a diffuse and a specular part. It is not based on any correct defined physical quantity, but on intensity values assigned to light sources and has the form

$$S_p = C_p \left(\cos i \,(1-d) + d\right) + W(i) \cos^n s, \tag{4.230}$$

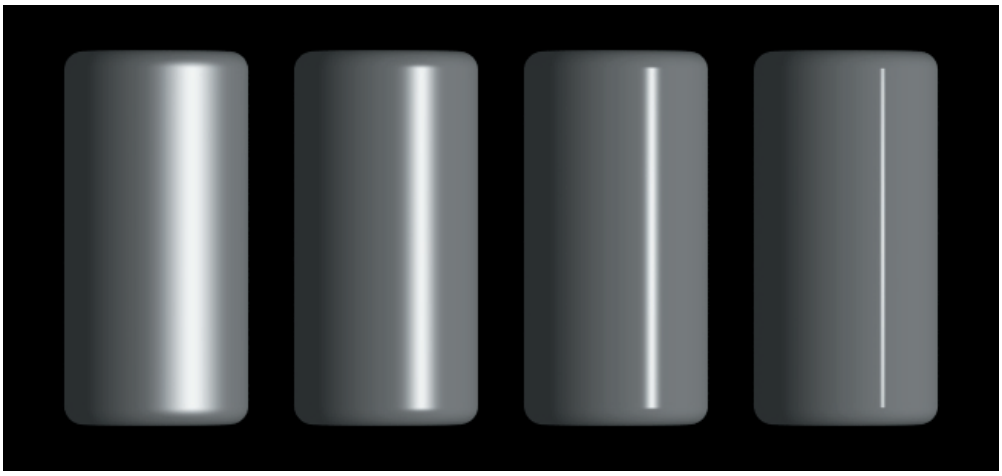


FIGURE 4.39: CYLINDERS RENDERED WITH THE PHONG ILLUMINATION MODEL. Four cylinders rendered with specular highlights. All cylinders are shaded with a directional light. As e increases—from left to right $e = 5, 20, 100, 1000$—the highlights become smaller and the cylinders look more shiny. Image courtesy of Kevin Suffern, University of Technology, Sidney.

where S_p is the shading at point p, d is the environmental diffuse reflection component, C_p is the reflection coefficient of the object at point p for a certain wavelength, $W(i)$ is a function giving the ratio of specularly reflected light to incident light as a function of the incident angle i, s is the angle between the specular direction and the observer, and n is a power for modelling the specular reflected light. The values of $W(i)$ and n are adjusted as free parameters for the picture, without any physical basis for the adjustments. Typical values are $0.1 < W(i) < 0.8$ and $1 < n < 10$. When used strictly as a BRDF representation, the diffuse reflection component is zero ($d = 0$) [160, Phong 1975]. Images rendered with the Phong model are shown in Figure 4.39.

Expressed in terms of radiance and with a constant ambient term replacing global illumination, the Phong shading equation can be written as:

$$L_o(\mathbf{s}, \omega_o) \stackrel{\text{def}}{=} k_a L_a + \left(k_d \langle \mathbf{N}(\mathbf{s}), \omega_i \rangle + k_s \langle \omega_r, \omega_o \rangle^{k_e}\right) L_i(\mathbf{s}, \omega_i), \tag{4.231}$$

where ω_i denotes the incoming direction of light at point $\mathbf{s}$, ω_o is the direction to the viewer, and ω_r is the direction of the reflected light. The coefficient k_e is called the Phong exponent, detailed explained further below, and k_d as well as k_s are material constants, which, also under the conditions $k_d \leq 1$ and $k_s \leq 1$, not ensure conservation of energy.

As we are interested in BRDFs, we will transform the classic Phong illumination model into a reflection model based on the concept of a bidirectional reflectance distribution function. For this, we define the *Phong BRDF* by:

DEFINITION 4.23 (The Phong BRDF) *Let ω_r be the perfect specular reflected direction vector of the incident radiance at surface point $\mathbf{s}$ coming from direction ω_i and ω_o be a vector in direction to the viewer. Let furthermore ρ_{dh} be the directional-hemispherical reflectance, and ρ_{dd} the so-called directional-directional reflectance, two material constants commonly chosen from interval $[0,1]$. Then, the* Phong BRDF, *f_r^{PH}, is defined as:*

ρ_{dh} (338)
ρ_{dd} (338)

$$f_r^{PH}(\mathbf{s}, \omega_i \to \omega_o) \stackrel{def}{=} \frac{\rho_{dh}(\mathbf{s})}{\pi} + \rho_{dd}(\mathbf{s}) \frac{\langle \omega_r, \omega_o \rangle^{k_e}}{\langle \mathbf{N}(\mathbf{s}), \omega_i \rangle} \tag{4.232}$$

$$= \frac{\rho_{dh}(\mathbf{s})}{\pi} + \rho_{dd}(\mathbf{s}) \frac{\cos^{k_e} \theta_{ro}}{|\cos \theta_i|} \tag{4.233}$$

with Phong exponent $k_e \in \mathbb{N}$ *and surface normal $\mathbf{N}(\mathbf{s})$, see Figure 4.40.*

With the diffuse reflection coefficient $k_d \stackrel{def}{=} \frac{\rho_{dh}}{\pi}$ and the specular reflection coefficient k_s given by $k_s \stackrel{def}{=} \rho_{dd}$ the Phong BRDF can then be written in the more commonly used form:

$$f_r^{PH}(\mathbf{s}, \omega_i \to \omega_o) = k_d + k_s \frac{\cos^{k_e} \theta_{ro}}{|\cos \theta_i|}. \tag{4.234}$$

The specular term in f_r^{PH} controls the color and the Phong exponent the expansion of the gloss. Thus, very glossy surfaces can be modeled by a large Phong exponent, $k_e > 100$, and more matte surfaces by smaller values of k_e, such as 1. Ideal specular reflection can be modeled if k_e goes to infinity. This can be interpreted in such a way that the specular reflection is approximated by a cone centered around the direction ω_r with an exponentially decreasing radiance, see Figure 4.41 and 4.42. Apart from the specular component, the Phong BRDF contains also a diffuse component, which is expressed by the Lambertian BRDF, that is, the diffuse reflection coefficient k_d from the foregoing section.

k_d (350)

REMARK 4.24 *Notice the left image from Figure 4.41, where the cosine-lobe of the Phong BRDF penetrates the surface for grazing incident directions. As we will seen, when discussing direct illumination this situation makes no problems, but we have a problem when discussing indirect illumination, unless we do something about it.*

Direct Illumination (410)
Indirect Illumination (410)

Furthermore we note, that for angles between ω_o and ω_r greater than $\frac{\pi}{2}$, the dot product $\langle \omega_r, \omega_o \rangle$ is negative, in these cases we have to clamp $\langle \omega_r, \omega_o \rangle$ to positive values in all of our formulas.

Due to the cosine-term in the denominator of the specular part, the Phong BRDF does not satisfy the Helmholtz reciprocity, this can easily be shown by:

$$f_r^{PH}(\mathbf{s}, \omega_i \to \omega_o) = k_d + k_s \frac{\cos^{k_e} \theta_{ro}}{|\cos \theta_i|} \tag{4.235}$$

$$\neq k_d + k_s \frac{\cos^{k_e} \theta_{ro}}{|\cos \theta_o|} \tag{4.236}$$

$$= f_r^{PH}(\mathbf{s}, \omega_o \to \omega_i). \tag{4.237}$$

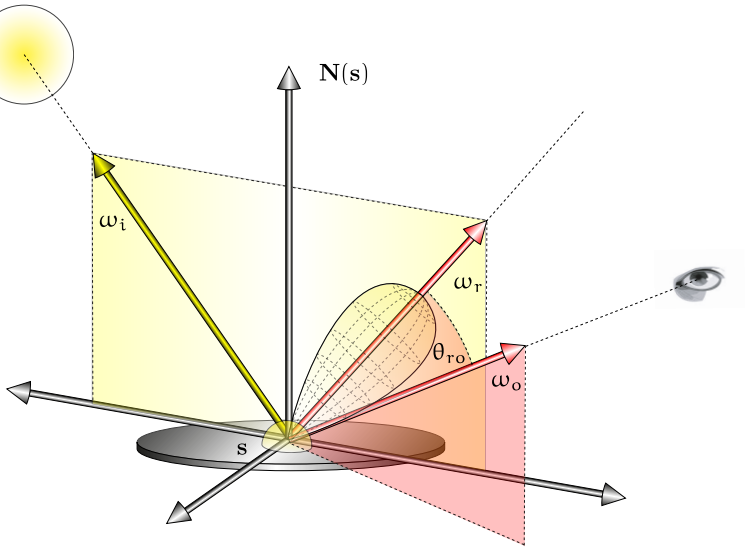


FIGURE 4.40: THE GEOMETRY UNDERLYING THE PHONG BRDF. Light incident at a surface point s from a direction ω_i is specularly reflected in direction ω_r. The Phong BRDF approximates the reflected light as a cone around the mirrored direction ω_r with exponentially decreasing intensity. This means, that the amount of light reflected in direction ω_o to the viewer is depending on the cone angle $\cos\theta_{ro}$ and the roughness k_e of the surface.

Reciprocity (331)
Conservation of Energy (332)

In addition, the Phong BRDF also violates the law of conservation of energy, as—due to the cosine in the denominator—the amount of reflected light for incident directions near to the surface will be unbounded. That is, the Phong BRDF is not physically plausible.

REMARK 4.25 (The Reciprocal Phong BRDF) *There is an easy way to make the Phong BRDF symmetric, namely, by canceling the cosine-term from the denominator of the specular component, $f_r^{\vee}$, of the BRDF f_r^{PH}. This then leads to the so-called* reciprocal Phong BRDF, *f_r^{rPH}:*

$$f_r^{rPH} \stackrel{def}{=} k_d + k_s \cos^{k_e}\theta_{ro}. \tag{4.238}$$

Based on the Phong BRDF, the Phong illumination model for n light sources is now defined by:

DEFINITION 4.24 (The Phong Illumination Model) *Let us assume there are n extended far away area light sources $\sun_1,\ldots,\sun_n$ in a scene to be rendered, let further more*

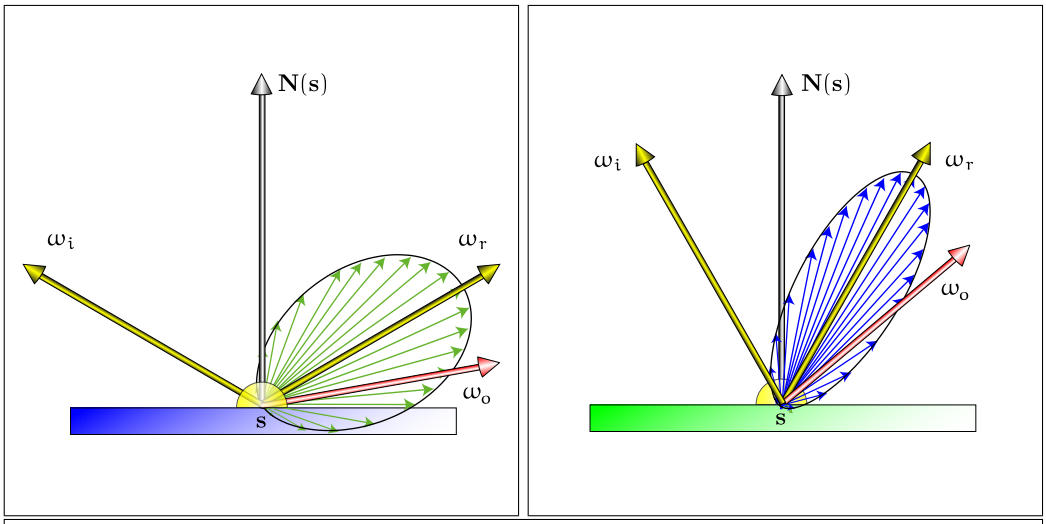


FIGURE 4.41: COSINE POWER LOBE. A wide and a narrow Phong lobe produced by a large respectively a small value of k_e. Notice also how the wide Phong lobe penetrates the surface. For grazing angle, almost half the lobe will be below the surface.

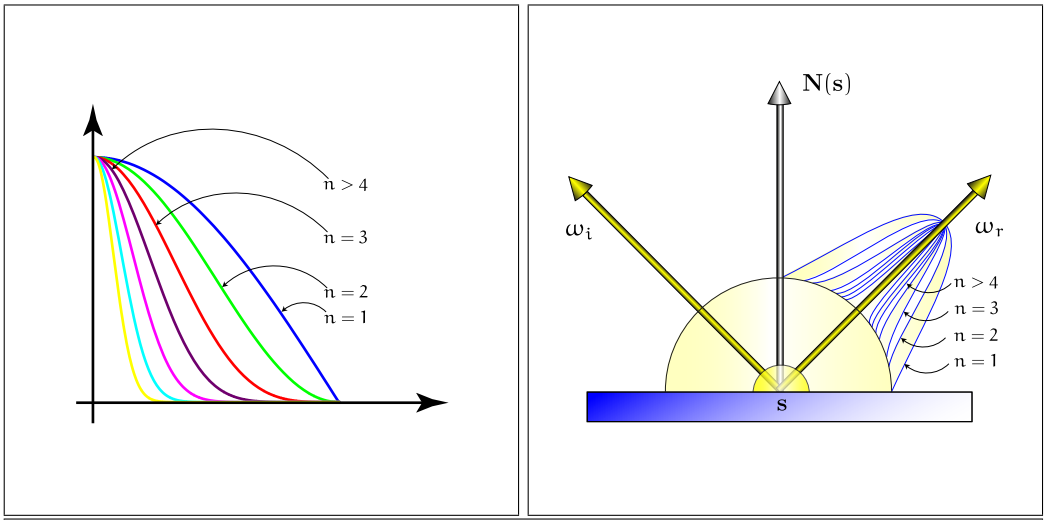


FIGURE 4.42: THE INFLUENCE OF EXPONENT k_e ON THE PHONG BRDF. Left, plots of the function $\cos\theta^n$ for $n = 1, 2, \ldots, 7$ in the interval $[0, \frac{\pi}{2}]$. In the right image, the variation of the reflected radiance at surface point s is shown—as a function of the direction of the viewing vector ω_o, the mirrored direction ω_r, and the Phong exponent $k_e = 1, 2, \ldots, 7$—as it rotates in all possible directions about the point of interest.

f_r^{PH} *be the Phong BRDF. Then, the* Phong illumination model *is defined as:*

$$L_o(s,\omega_o) \overset{def}{=} k_a L_a + \sum_{j=1}^{n} \int_{\mathcal{H}_i^2(s)} f_r^{PH}(s,\omega_i^j \to \omega_o)\, L_i(s,\omega_i^j)\, d\sigma_s^\perp(\omega_i^j), \quad (4.239)$$

$$= k_a L_a + \sum_{j=1}^{n} \int_{\mathcal{H}_i^2(s)} f_r^{PH}(s,\omega_i^j \to \omega_o)\, L_e(l_j,-\omega_i^j)\, d\sigma_s^\perp(\omega_i^j), \quad (4.240)$$

where the constant $k_a L_a$ *is* the ambient illumination *representing indirect illumination, and* l_j *are points on the light source* $\sun_j$ *reachable in direction* ω_i^j.

REMARK 4.26 (The Phong Illumination Model for Point Light Sources) *In the case where we only consider* n *point light sources* $*_1,\dots,*_n$ *the Phong illumination model can be expressed by using Dirac-*δ *distributions as:*

Dirac-δ Distribution (118)

$$L_o(s,\omega_o) = k_a L_a + \sum_{j=1}^{n} \left(k_d \langle \mathbf{N}(s), \omega_i^j \rangle + k_s \langle \omega_r, \omega_o \rangle^{k_e} \right) L_i(s,\omega_i^j) \quad (4.241)$$

$$= k_a L_a + \sum_{j=1}^{n} \left(k_d \left|\cos\theta_i^j\right| + k_s \left|\cos^{k_e}\theta_{or}^j\right| \right) L_e(*_j,-\omega_i^j), \quad (4.242)$$

with $*_j = \gamma(s,\omega_i^j)$.

Note, the additional factor $\left|\cos\theta_i^j\right|$ *at the diffuse coefficient* k_d *respectively the absence of this factor in the denominator of the specular term. Both changes are required due to the fact that the associated BRDF must be integrated with respect to projected solid angle measure* $\sigma^\perp$.

$\sigma^\perp$ (88)

Like the Lambertian reflection model, also the Phong illumination model is a purely local illumination model, which only accounts for direct light coming from light sources and neglect any reflection on other surfaces. Due to its definition, the exitant radiance in direction to the viewer, thus $L_o(s,\omega_o)$, consist of three components: First an ambient term $k_a L_a$, which simulates the indirect light in the scene. Second, the diffuse part, thus the radiance at s which comes from all directions ω_i^j with respect to the light sources. And finally, the specular fraction of light coming from all light sources. The parameter n, determining the size of the highlight, can be used to control the roughness, $n < 2$—or shininess $n > 30$—of the surface.

REMARK 4.27 (Wavelength Dependance in The Phong Illumination Model) *As the radiometric quantity radiance is used in the Phong BRDF, and radiance must be considered as a three-dimensional vector from the RGB-color system, the color of an object can be controlled by evaluating the Phong illumination model at three wavelengths. For that, we only have to replace the reflection coefficients* $k_x, x \in \{a,d,s\}$ *in the Phong model by three-dimensional vectors* $\mathbf{k}_x = (k_{xr}^T, k_{xg}^T, k_{xb}^T)$ *which are used*

to control the color of the corresponding component. So, for rendering a purely blue object, we choose: $\mathbf{k}_d = (0,0,1)^T$.

THE BLINN-PHONG MODEL. A widely-used variant of the Phong reflection model was introduced in [25, Blinn 1977]. Its importance does not lie in the fact, that it is physically more accurate than the Phong illumination model, but because it avoids to compute the reflection vector which makes it faster to compute.

Let us assume, the light source and the viewer are infinitely far away from the observation point, then the incident direction ω_i of a light ray and the exitant direction ω_o towards the viewer are constant over the whole scene. Since the reflected direction ω_r is expensive to calculate, in [25, Blinn 1977] it is suggested to use a vector $\mathbf{H}$ which is the direction to a hypothetical surface that is oriented in direction halfway between the light vector and the direction towards the viewer:

DEFINITION 4.25 (The Blinn-Phong BRDF, f_r^{BP}) *Let ω_r be the perfect specular reflected direction vector of the incident radiance at surface point $\mathbf{s}$ coming from direction ω_i and ω_o be a vector in direction to the viewer. Let furthermore ρ_{dh} be the directional-hemsipherical reflectance and ρ_{dd} the directional-directional hemispherical reflectance, both material constants commonly chosen from the interval $[0,1]$. Then, the* Blinn-Phong BRDF f_r^{BP}, *see Figure 4.43, is defined as:*

ρ_{dh} (338) ρ_{dd} (338)

$$f_r^{BP}(\mathbf{s}, \omega_i \to \omega_o) \stackrel{def}{=} \frac{\rho_{dh}(\mathbf{s})}{\pi} + \rho_{dd}(\mathbf{s}) \frac{\langle \mathbf{N}(\mathbf{s}), \mathbf{H} \rangle^{k_e}}{\langle \omega_i, \mathbf{N}(\mathbf{s}) \rangle} \tag{4.243}$$

$$= \frac{\rho_{dh}(\mathbf{s})}{\pi} + \rho_{dd}(\mathbf{s}) \frac{\cos^{k_e} \theta_{\mathbf{H}}}{|\cos \theta_i|} \tag{4.244}$$

with Phong exponent $k_e \in \mathbb{N}$, *surface normal* $\mathbf{N}(\mathbf{s})$ *and half vector* $\mathbf{H} = \frac{\omega_o + \omega_i}{2}$.

With the diffuse reflection coefficient $k_d = \frac{\rho_{dh}}{\pi}$ *and the specular reflection coefficient* $k_s = \rho_{dd}$ *the Blinn-Phong BRDF can then be written in the more commonly used form*

$$f_r^{PH}(\mathbf{s}, \omega_i \to \omega_o) = k_d + k_s \frac{\cos^{k_e} \theta_{\mathbf{H}}}{|\cos \theta_i|}. \tag{4.245}$$

Since the half vector $\mathbf{H}$ corresponds to the normal of the surface that reflects the incident light ray in the ideal mirrored direction ω_o, the closer $\mathbf{N}$ and $\mathbf{H}$ are, the brighter the specular highlight will be.

Based on the Blinn-Phong BRDF, the Blinn-Phong illumination model for n light sources is now defined by:

DEFINITION 4.26 (The Blinn-Phong Illumination Model) *Let us assume there are n extended far away area light sources in a scene to be rendered. The* Blinn-Phong Illu-

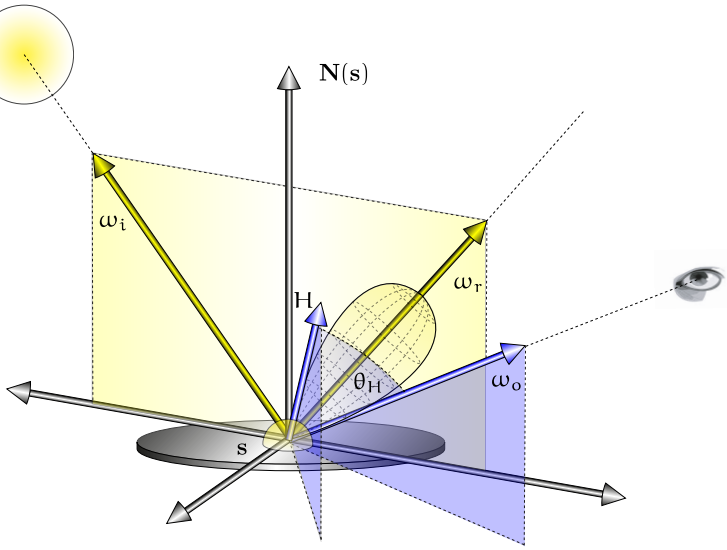


FIGURE 4.43: THE GEOMETRY UNDERLYING THE BLINN -PHONG BRDF. Light incident at surface point s from a direction ω_i is specularly reflected in direction ω_r. The Phong BRDF approximates the reflected light as a cone around the mirrored direction ω_r with exponentially decreasing intensity. This means, that the amount of light reflected in direction ω_o to the viewer is depending on the cone angle $\cos\theta_H$ and the roughness k_e of the surface.

mination model *is defined as:*

$$L_o(s,\omega_o) \stackrel{def}{=} k_a L_a + \sum_{j=1}^{n} \int_{\mathcal{H}_i^2} f_r^{BP}(s,\omega_i^j \to \omega_o)\, L_i(s,\omega_i^j)\, d\sigma_s^{\perp}(\omega_i^j), \qquad (4.246)$$

where the constant $k_a L_a$ *is* the ambient illumination*, representing indirect illumination, and* ω_i^j *are directions with respect to light source* j.

REMARK 4.28 (The Blinn-Phong Illumination Model for Point Light Sources) *In the case where we only consider point light sources, the Blinn-Phong illumination model can be expressed by*

$$L_o(s,\omega_o) \stackrel{def}{=} k_a L_a + \sum_{j=1}^{n} \left(k_d \langle \mathbf{N}(s), \omega_i^j \rangle + k_s \langle \mathbf{N}(s), \mathbf{H}_j \rangle^{k_e} \right) L_i(s,\omega_i^j). \qquad (4.247)$$

Note, the additional factor $\langle \mathbf{N}(s), \omega_i^j \rangle$ *at the diffuse coefficient* k_d *respectively the absence of this factor in the denominator of the specular term. Both changes are*

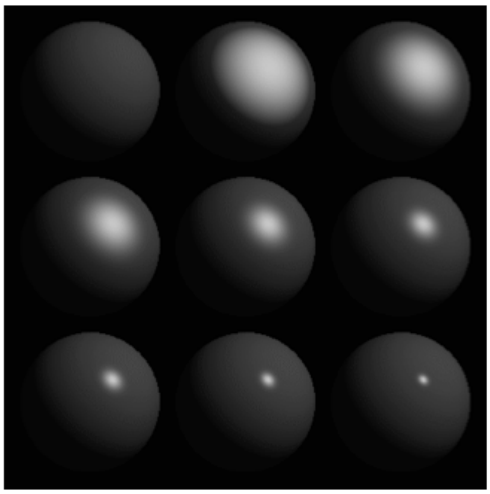


FIGURE 4.44: IMAGES RENDERED WITH THE BLINN-PHONG ILLUMINATION MODEL. The above spheres illustrate specular reflections. Here, the Phong exponent, k_e representing the shininess of the surface, is varied from 0, left-upper sphere, to a value of $k_e > 100$, right-lower sphere. The spheres on the left hand side are rendered using the Phong BRDF, the images on the right are rendered via the Blinn-Phong illumination model.

required due to the fact that the approriate BRDF must be integrated with respect to projected solid angle measure $\sigma^{\perp}$. $\sigma^{\perp}$ (88)

REMARK 4.29 *Altough it is neither reciprocal nor energy-conserving, the Blinn-Phong model plays an important role in a wide range of graphics accelerators as it is the lightning model used in standard OpenGL. Images rendered via the Blinn-Phong model are shown in Figure 4.44.* Helmholtz Reciprocity (331) Conservation of Energy (332)

Also the Blinn-Phong model can be made symmetric by canceling the cosine-term from the denominator of the specular component.

THE MODIFIED PHONG MODEL. Since the Phong BRDF is neither reciprocal nor energy conservative, the Phong BRDF is, as already mentioned above, not *physically plausible*. In [117, Lafortune & Willems 1977] these two mayor problems are handled by normalizing the Phong BRDF. Helmholtz Reciprocity (331) Conservation of Energy (332)

LEMMA 4.5 (The Modified Phong BRDF) *Let* f_r^{PH} *be the Phong BRDF from Equation*

(4.232). Defining

$$f_r^{PH,MOD}(\mathbf{s},\omega_i\to\omega_o) \overset{def}{=} \frac{\rho_{dh}(\mathbf{s})}{\pi}+\frac{\rho_{dd}(\mathbf{s})\,(k_e+2)}{2\pi}\langle\omega_r,\omega_o\rangle^{k_e} \tag{4.248}$$

$$= \frac{\rho_{dh}(\mathbf{s})}{\pi}+\frac{\rho_{dd}(\mathbf{s})\,(k_e+2)}{2\pi}\cos^{k_e}\theta_{ro} \tag{4.249}$$

with $\rho_{dh}+\rho_{dd}\leq 1$*, which, using the diffuse and specular coefficients* k_d *and* k_s*, can also be formulated as:*

$$f_r^{PH,MOD}(\mathbf{s},\omega_i\to\omega_o) = k_d + k_s\frac{k_e+2}{2}\cos^{k_e}\theta_{ro}, \tag{4.250}$$

then $f_r^{PH,MOD}$ *is a physically plausible BRDF satisfying the Helmholtz reciprocity and the conservation of energy.*

PROOF 4.5 *To show that* f_r^{PH} *satisfies the Helmholtz reciprocity is straightforward, and the conservation of energy condition of* f_r^{PH} *follows from the following discussion: The maximum of reflected energy in the Phong BRDF at point* $\mathbf{s}$ *occurs when* ω_i *and* $\mathbf{N}(\mathbf{s})$ *are parallel to each other. Obviously, this also implies* $\omega_r=\mathbf{N}(\mathbf{s})$*, that is:* $\langle\omega_r,\omega_o\rangle=\langle\mathbf{N}(\mathbf{s}),\omega_o\rangle=\cos\theta_o$*.*

Computing the directional-hemispherical reflectance ρ_{dh} *then yields*

$$0 \overset{(4.153)}{\leq} \rho_{dh}(\mathbf{s},\omega_i\to\mathcal{H}^2_o) \tag{4.251}$$

$$\overset{(4.161)}{=} \int_{\mathcal{H}^2_o(\mathbf{s})} f_r^{PH,MOD}(\mathbf{s},\omega_i\to\omega_o)\,d\sigma^\perp_{\mathbf{s}}(\omega_o) \tag{4.252}$$

$$= \int_{\mathcal{H}^2_o(\mathbf{s})} k_d + k_s\frac{k_e+2}{2}\,\langle\mathbf{N}(\mathbf{s}),\omega_o\rangle^{k_e}\,d\sigma^\perp_{\mathbf{s}}(\omega_o) \tag{4.253}$$

$$= \int_{\mathcal{H}^2_o(\mathbf{s})} k_d\,d\sigma^\perp_{\mathbf{s}}(\omega_o) + \int_{\mathcal{H}^2_o(\mathbf{s})} k_s\frac{k_e+2}{2}\cos^{k_e}\theta_o\,d\sigma^\perp_{\mathbf{s}}(\omega_o) \tag{4.254}$$

$$= k_d\pi + k_s\frac{k_e+2}{2}\int_{\mathcal{H}^2_o(\mathbf{s})}\cos^{k_e}\theta_o\,d\sigma^\perp_{\mathbf{s}}(\omega_o) \tag{4.255}$$

$$= k_d\pi + k_s\frac{k_e+2}{2}\int_{[0,\frac{\pi}{2}]}\int_{[0,2\pi)}\cos^{k_e+1}\theta_o\sin\theta_o\,d\mu(\phi_o)d\mu(\theta_o) \tag{4.256}$$

$$= k_d\pi + k_s 2\pi\frac{k_e+2}{2}\int_{[0,\frac{\pi}{2}]}\cos^{k_e+1}\theta_o\sin\theta_o\,d\mu(\theta_o) \tag{4.257}$$

$$= k_d\pi + k_s\pi(k_e+2)\frac{(-1)}{k_e+2}\cos\theta_o\Big|_0^{\frac{\pi}{2}} \tag{4.258}$$

$$= k_d\pi + k_s\pi = \rho_{dh}+\rho_{dd}. \tag{4.259}$$

The condition $\rho_{dh}+\rho_{dd}\leq 1$ *then proofs the lemma.*

4.2.2.5.3 PHYSICAL-BASED OR PHYSICS-INSPIRED BRDF MODELS

With the Phong illumination model we can simulate plastic-like surfaces well, but it shows significant deficiencies when modeling high-gloss, metallic surfaces, or the reflection at rough materials. Here we are forced to go back to reflectance models that in some sense are based, at least in part, on the underlying physics of reflection. So, physical-based or physics-inspired BRDF models can be characterized as to attempt to mimic physical reality. That is, they attempt to use theoretical constructs from geometric and physical optics to build a reflection model that is closer to reality than phenomenological illumination models.

Since already the simple phenomenon of reflection is such a complex process, it also makes no sense to combine all possible effects of light into a single lighting model from which we expect that it is efficient and reasonable usable in practice. So, we have to decide what is important to the application of the model in practice. This means, that physical-based illumination models in computer graphics can not model the light interaction as it works in real world.

In the following, we will present a physically based illumination model whose variation in reflectivity is based on microscopically rough surfaces induced by randomly oriented specular microfacets: *The Cook-Torrance Illumination Model*, [39, Cook & Torrance 1982]. Although it is physically based, it corresponds to a completely local process, i.e. the incident and exitant light ray arrive and leave the surface at the same point. Furthermore, effects such as multiple scattering and polarization, which would require the study of the classical electromagnetic wave theory [91, Ishimaru 1997], [27, Born & Wolf 1999], [217, Tsang & al. 2000], and [216, Tsang & Kung 2001] are neglected in this model.

THE COOK-TORRENCE ILLUMINATION MODEL. The perhaps most important physical-based illumination model in computer graphics is the *Cook-Torrance Illumination Model*. It is wavelength dependent and in practice very well to describe metals such as copper and gold, but it can also be used to render materials with different degree of roughness.

Cook and Torrence picked up the idea of [215, Torrence & Sparrow 1967], that any rough surface can be composed by many randomly placed tiny v-shaped grooves lined with flat perfectly specular mirrors, called *microfacets*, see the images in the top row of Figure 4.45. As with the microfacets also their directions are randomly distributed over a surface, the surface is statistically described by a distribution function—commonly, with a strong peak in direction to the macroscopic surface normal—that gives the probability that a microfacet has a particular orientation. Via this distribution function then we can model different types of surface roughness: the greater the variation of the microfacet normals, the rougher the surface is, while smooth surfaces have relatively small variation in the microfacet normals, see Figure 4.45. This model is only correct if the wavelength of light is smaller than the roughness of the surface.

The Cook-Torrence model works with a BRDF that is composed of two components: a specular component, $f_r^\vee$, for handling the specularly reflected light and an ideal diffuse $f_r^\vee$ (325)

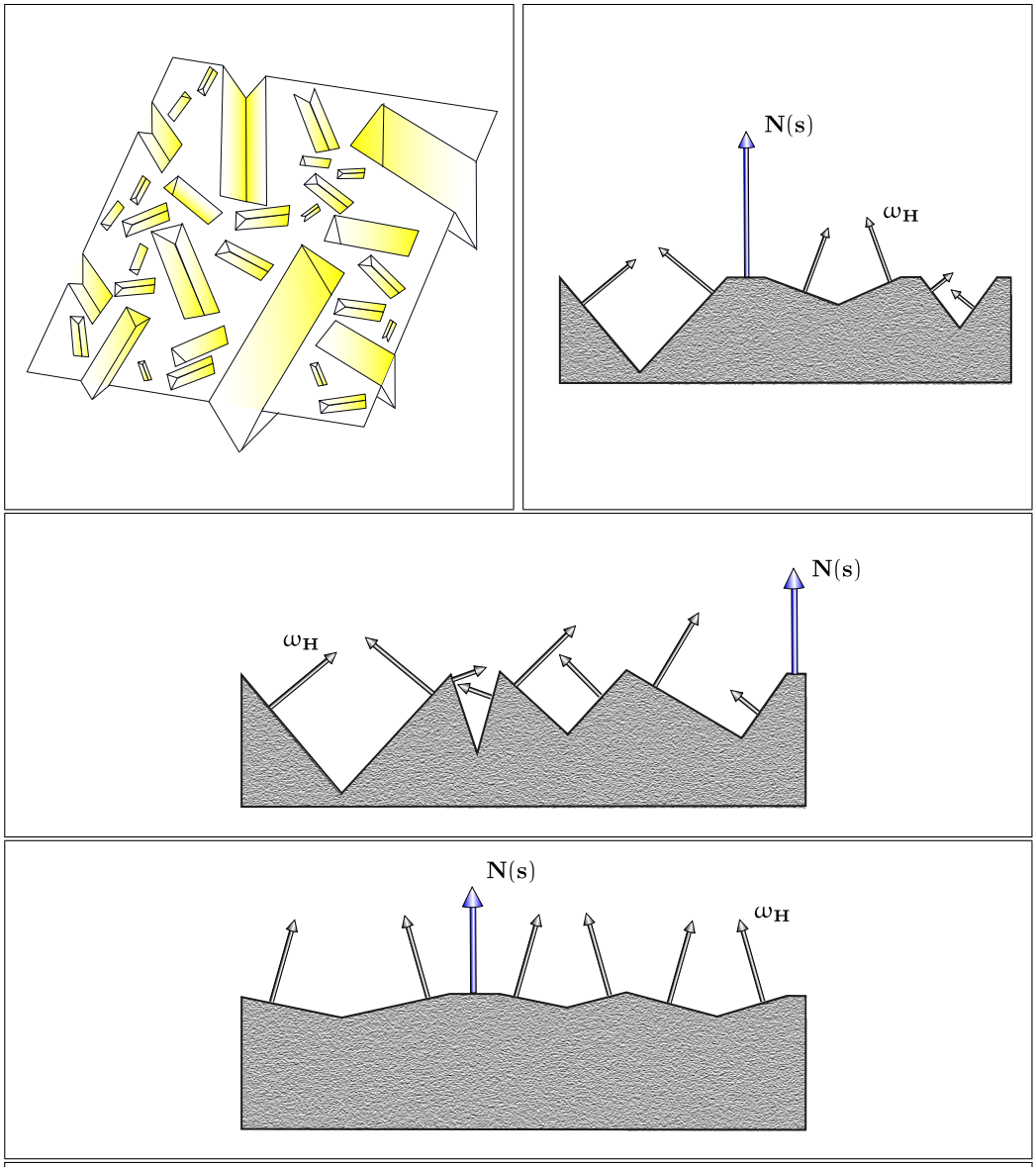


FIGURE 4.45: MICROFACETS MODEL. A surface is composed of many randomly placed small v-shaped grooves lined with flat mirrors, so-called microfacets. So, the roughness of a surface can statistically be described by a distribution function that gives the probability that a microfacet has a normal in a particular direction, see the right image in the top row, it represents the cross section of a rough surface. Obviously it holds: The greater the variation of the microfacet normals, $\omega_{\mathbf{H}}$, the rougher the surface is, this fact is visualized in the image in the center. The lower image demonstrates a smooth surface, where we have only a relatively small variation in the microfacet normals.

component, f_r^o, for handling the diffusely reflected light, that is, the Cook-Torrance BRDF is of the form: f_r^o (339)

$$f_r^{CT} = k_d f_r^o + k_s f_r^{\vee}, \tag{4.260}$$

where we assume—to ensure that the Cook-Torrance BRDF is energy conserving—that the reflection coefficients k_d and k_s follow the constraint: $k_d + k_s \leq 1$. k_d (350) k_s (353)

The diffuse component, f_r^o, is as usual chosen as the standard constant diffuse term:

$$f_r^o \stackrel{def}{=} \frac{\rho_{dh}}{\pi}. \tag{4.261}$$

Compared with f_r^o, the specular component, $f_r^{\vee}$, of the Cook-Torrance BRDF is much more complex. For an derivation of $f_r^{\vee}$, let us consider a light beam incident at a rough surface, modeled via a large set of microfacets. Since the microfacets are assumed to be ideal mirrors, only those microfacets, whose surface normal is given by $\omega_{\mathbf{H}} = \frac{\omega_i + \omega_o}{\|\omega_i + \omega_o\|_2}$, reflect the incident light from direction ω_i in the mirrored direction $M_{\omega_{\mathbf{H}}}(\omega_i) = \omega_o$, see Figure 4.46. Obviously, the fraction of microfacets that participate in the reflection of light from ω_i to ω_o is given via the value of the distribution function D at the half vector $\omega_{\mathbf{H}}$, that is, $D(\omega_{\mathbf{H}})$.

Then, the differential flux $d\Phi_i$ incident on microfacets with surface normal $\omega_{\mathbf{H}}$ is given by: Φ (249)

$$d^2\Phi_i(\mathbf{s}, \omega_i) = L_i(\mathbf{s}, \omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i)\, d\mu^2(m_{\omega_{\mathbf{H}}}), \tag{4.262}$$

where $m_{\omega_{\mathbf{H}}}$ are all microfacets with surface normal $\omega_{\mathbf{H}}$, μ^2 is the Lebesgue area measure, and $\mathbf{s}$ are points on these microfacets. Lebesgue Area Measure (82)

Using the relation $d\sigma_{\mathbf{s}}^{\perp}(\omega_i) = d\sigma_{\mathbf{s}}(\omega_i)\cos\theta_{\mathbf{H}}$, see Figure 4.46, then we can also write: $d\sigma_{\mathbf{s}}^{\perp}$ (88)

$$d^2\Phi_i(\mathbf{s}, \omega_i) = L_i(\mathbf{s}, \omega_i)\, d\sigma_{\mathbf{s}}(\omega_i)\cos\theta_{\mathbf{H}}\, d\mu^2(m_{\omega_{\mathbf{H}}}). \tag{4.263}$$

The Lebesgue area measure of all active microfacets $m_{\omega_{\mathbf{H}}}$ can now easily be estimated by the Distribution function D, namely as:

$$d\mu^2(m_{\omega_{\mathbf{H}}}) = D(\omega_{\mathbf{H}})\, d\sigma_{\mathbf{s}}(\omega_{\mathbf{H}})\, d\mu^2(A), \tag{4.264}$$

where A denotes the surface in which we are interested in and $\omega_{\mathbf{H}}$ are microfacet normals lying within the differential solid angle $d\sigma_{\mathbf{s}}(\omega_{\mathbf{H}})$. Inserting this relation into Equation (4.263) yields: $d\sigma_s$ (87)

$$d^2\Phi_i(\mathbf{s}, \omega_i) = L_i(\mathbf{s}, \omega_i)\, d\sigma_{\mathbf{s}}(\omega_i)\, D(\omega_{\mathbf{H}})\, d\sigma_{\mathbf{s}}^{\perp}(\omega_{\mathbf{H}})\, d\mu^2(A). \tag{4.265}$$

As all microfacets are ideal specular reflectors, the reflected differential flux $d\Phi_o$ can be expressed in terms of the incident flux and the Fresnel reflectance F_r by: F_r (309)

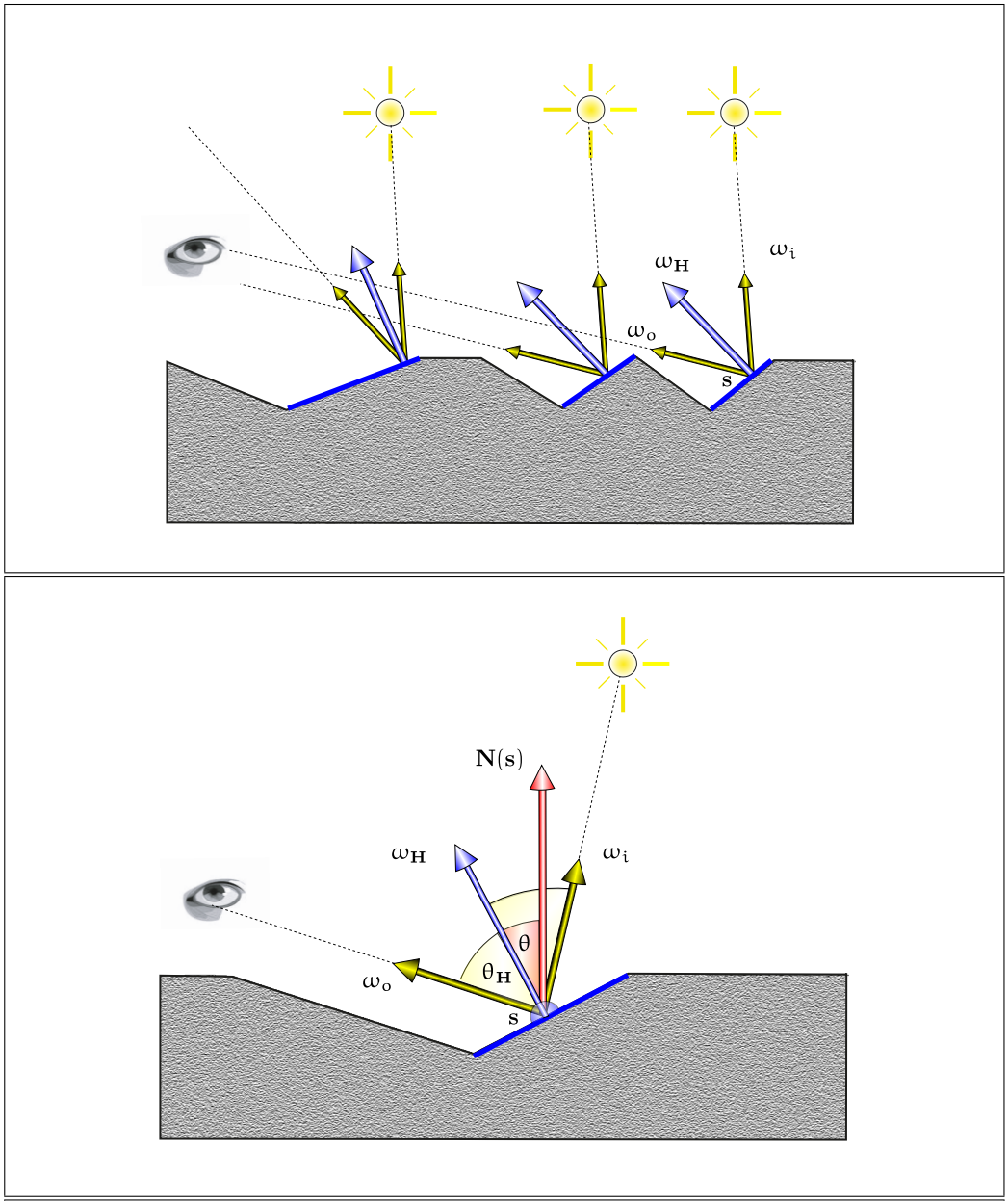


FIGURE 4.46: MICROFACETS GEOMETRY. For perfectly specular microfacets, only those with normals $\omega_{\mathbf{H}}$ reflect light incident from direction ω_i into direction $\omega_o = M_{\omega_{\mathbf{H}}}(\omega_i)$. Note: The angle between the macroscopic surface normal $\mathbf{N}$ and $\omega_{\mathbf{H}}$, the surface normal of a microfacet, is denoted by θ, while the incoming and outgoing angle of the light ray is denoted by $\theta_{\mathbf{H}}$.

$$d^2\Phi_o(s,\omega_o) = F_r(\omega_o)d^2\Phi_i(s,\omega_i) \quad (4.266)$$

$$\overset{(4.265)}{=} F_r(\omega_o)\,L_i(s,\omega_i)\,d\sigma_s(\omega_i)\,D(\omega_H)\,d\sigma_s^{\perp}(\omega_H)\,d\mu^2(A). \quad (4.267)$$

Now, let us consider Figure 4.47. Obviously, all light incident at the microfacet in the lower right image is also reflected in direction to the viewer, which does not hold for the other two cases. In the upper image, a light source illuminates a microfacet, but a fraction of the light reflected in direction to the viewer is blocked by a part of the opposite microfacet. This means, that only a fraction of the illumination reaches the viewer. We call this effect *masking*.

Using the sine law for triangles then due to some trigonometric formulae, we get:

$$G = 1 - \frac{m}{l} \quad (4.268)$$

$$= 1 - \frac{\sin m}{\sin l} \quad (4.269)$$

$$= 1 - \frac{\sin\left(\theta_H + 2\theta - \frac{\pi}{2}\right)}{\sin\left(\frac{\pi}{2} - \theta_H\right)} \quad (4.270)$$

$$= 2 \cdot \frac{\cos\theta \cdot \cos\left(\theta + \theta_H\right)}{\cos\theta_H}. \quad (4.271)$$

Setting the angles $\cos\theta = \langle N, \omega_H\rangle$, $\cos\left(\theta + \theta_H\right) = \langle N, \omega_o\rangle$, and $\cos\theta_H = \langle \omega_o, \omega_H\rangle$, then we get for the probability of masking:

$$G = \min\left(1, \frac{2\langle N, \omega_H\rangle\langle N, \omega_o\rangle}{\langle \omega_o, \omega_H\rangle}\right). \quad (4.272)$$

A similar case is illustrated in the lower left image of Figure 4.47. Here, a part of the microfacet visible from the viewer lies in the shadow area of another illuminated microfacet. Also this means that only a part of the illumination reaches the viewer. The term *shadowing* is used to describe this interference effect due to incident light. Substituting the direction vectors ω_i by ω_o in the above formula, leads directly to the corresponding formula for shadowing, namely:

$$G = \min\left(1, \frac{2\langle N, \omega_H\rangle\langle N, \omega_i\rangle}{\langle \omega_i, \omega_H\rangle}\right). \quad (4.273)$$

The probability of neither shadowing nor masking taking place can be approximated by the minimum of these two probabilities, namely:

$$G(\omega_i \to \omega_o) \overset{def}{=} \min\left\{1, \frac{2\langle N, \omega_H\rangle\langle N, \omega_o\rangle}{\langle \omega_o, \omega_H\rangle}, \frac{2\langle N, \omega_H\rangle\langle N, \omega_i\rangle}{\langle \omega_i, \omega_H\rangle}\right\}, \quad (4.274)$$

where N is the macroscopic surface normal through a point illuminated from direction ω_i, and considered from direction ω_o and ω_H corresponds to the microfacet normal, for a detailed discussion of the geometry term, see [25, Blinn 1977].

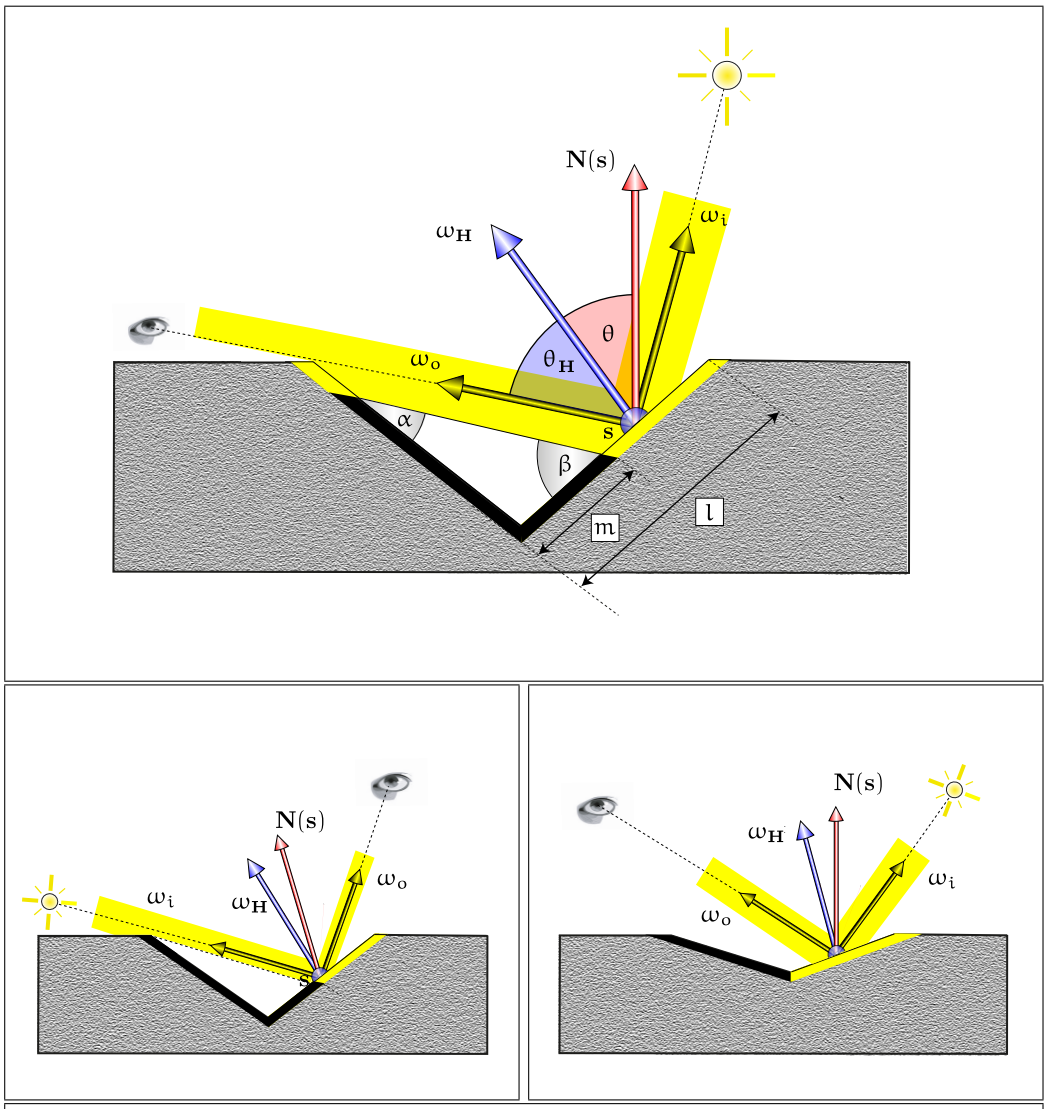


FIGURE 4.47: GEOMETRICAL ATTENUATION IN THE MICROFACETS MODEL. All three images show the interaction of light with a microfacet. When the incident direction of a light beam begins to approach the mean surface, then interference effects come into play. This can be seen in the upper image where some reflected light is trapped. This effect is called *masking*. The degree of masking is dependent on the ratio $\frac{m}{l}$ which describes the proportionate amount of a microfacet contributing to reflected light, namely: $G = 1 - \frac{m}{l}$. For *shadowing* the situation is geometrically identical with the role of the directions ω_i and ω_o interchanged, see the lower left image. In the lower right image, all incident light is reflected in direction to the viewer, this means, there is no attenuation in the reflected light due to the underlying geometry, so it holds: $G = 1$.

Inserting the geometry term into Equation (4.267) yields:

$$\begin{aligned} d^2\Phi_o(s,\omega_o) &= F_r(\omega_o)\, G(\omega_i \to \omega_o)\, D(\omega_{\mathbf{H}}) \cdot \\ &\quad L_i(s,\omega_i)\, d\sigma_s(\omega_i)\, d\sigma_s^{\perp}(\omega_{\mathbf{H}})\, d\mu^2(A). \end{aligned} \tag{4.275}$$

Obviously, the radiance exitant at point s on any active microfacet in direction ω_o is given by:

$$L_o(s,\omega_o) \overset{(3.15)}{=} \frac{d^2\Phi_o(s,\omega_o)}{d\sigma^{\perp}(\omega_o)\, d\mu^2(A)} \tag{4.276}$$

$$= F_r(\omega_o)\, G(\omega_i \to \omega_o)\, D(\omega_{\mathbf{H}}) \cdot \frac{L_i(s,\omega_i)\, d\sigma_s(\omega_i)\, d\sigma_s^{\perp}(\omega_{\mathbf{H}})\, d\mu^2(A)}{d\sigma^{\perp}(\omega_o)\, d\mu^2(A)} \tag{4.277}$$

$$= F_r(\omega_o)\, G(\omega_i \to \omega_o)\, D(\omega_{\mathbf{H}}) \cdot \frac{L_i(s,\omega_i)\, d\sigma_s(\omega_i)\, d\sigma_s^{\perp}(\omega_{\mathbf{H}})}{d\sigma^{\perp}(\omega_o)}. \tag{4.278}$$

Now, due to Definition 4.88, an associated BRDF is defined as:

$$f_r(s,\omega_i \to \omega_o) \overset{def}{=} \frac{dL_o(s,\omega_o)}{L_i(s,\omega_i) d\sigma_s^{\perp}(\omega_i)} \tag{4.279}$$

$$\overset{(4.278)}{=} F_r(\omega_o)\, G(\omega_i \to \omega_o)\, P(\omega_{\mathbf{H}}) \cdot \tag{4.280}$$

$$\frac{L_i(s,\omega_i)\, d\sigma_s(\omega_i)\, d\sigma_s^{\perp}(\omega_{\mathbf{H}})}{d\sigma^{\perp}(\omega_o)\, L_i(s,\omega_i) d\sigma_s^{\perp}(\omega_i)} \tag{4.281}$$

$$= \frac{F_r(\omega_o)\, G(\omega_i \to \omega_o)\, D(\omega_{\mathbf{H}})}{\cos\theta_i \cos\theta_o} \frac{d\sigma_s^{\perp}(\omega_{\mathbf{H}})}{d\sigma(\omega_o)}. \tag{4.282}$$

According to the Law of reflection—for an illustration, see Figure 4.46—the polar angle θ_o between ω_i and ω_o is equal to $2\theta_{\mathbf{H}}$, while the azimuthal $\phi_o = \phi_{\mathbf{H}}$. Using this two relation, then the second term in Equation (4.282) can be rephrased as:

Law of Reflection (300)

$$\frac{d\sigma_s^{\perp}(\omega_{\mathbf{H}}) \cos\theta_{\mathbf{H}}}{d\sigma(\omega_o)} \overset{(2.186),(2.192)}{=} \frac{\sin\theta_{\mathbf{H}} \cos\theta_{\mathbf{H}}\, d\mu(\theta_{\mathbf{H}}) d\mu(\phi_{\mathbf{H}})}{\sin\theta_o\, d\mu(\theta_o) d\mu(\phi_o)} \tag{4.283}$$

$$= \frac{\sin\theta_{\mathbf{H}} \cos\theta_{\mathbf{H}}\, d\mu(\theta_{\mathbf{H}}) d\mu(\phi_{\mathbf{H}})}{\sin(2\theta_{\mathbf{H}}) d\mu(2\theta_{\mathbf{H}}) d\mu(\phi_{\mathbf{H}})} \tag{4.284}$$

$$= \frac{\sin\theta_{\mathbf{H}} \cos\theta_{\mathbf{H}}\, d\mu(\theta_{\mathbf{H}})}{\sin(2\theta_{\mathbf{H}}) 2\, d\mu(\theta_{\mathbf{H}})} \tag{4.285}$$

$$= \frac{\sin\theta_{\mathbf{H}} \cos\theta_{\mathbf{H}}}{2\cos\theta_{\mathbf{H}} \sin\theta_{\mathbf{H}} \cdot 2} = \frac{1}{4}. \tag{4.286}$$

Using this relation in Equation (4.282) implies the following formulation of the specular component of the Cook-Torrence BRDF:

$$f_r(s,\omega_i \to \omega_o) = \frac{F_r(\omega_o)\, G(\omega_i \to \omega_o)\, D(\omega_{\mathbf{H}})}{4 \cos\theta_i \cos\theta_o}. \tag{4.287}$$

DEFINITION 4.27 (The Cook-Torrance BRDF, f_r^{CT}) *Let* F_r *be the Fresnel reflectance from Equation (4.72),* G*, represents the geometrical attenuation factor from above, and* D *denotes a probability distribution function of microfacet orientations...*

k_d (350) k_s (353)

Let furthermore $\frac{\rho_{dh}}{\pi}$ *be the diffuse reflection coefficient and* $\frac{\rho_{dd}}{\pi}$ *the coefficient of specular reflection, both, material constants commonly chosen from the interval* $[0,1]$*. Then, the* Cook-Torrance BRDF, f_r^{CT}*, is defined as:*

$$f_r^{CT}(\mathbf{s}, \omega_i \to \omega_o) \stackrel{def}{=} k_d f_r^o + k_s f_r^{\vee} \tag{4.288}$$

$$= \frac{\rho_{dh}(\mathbf{s})}{\pi} + \frac{\rho_{dd}(\mathbf{s})}{\pi} \frac{F_r(\omega_o)\, G(\omega_i \to \omega_o)\, D(\omega_{\mathbf{H}})}{4 \cos\theta_i \cos\theta_o}. \tag{4.289}$$

REMARK 4.30 (The Choice of the Microfacet Distribution Function) *There a various possibilities to model the roughness of the underlying surface. So, in [25, Blinn 1977] a Gaussian distribution function for* D *was used:*

$$D(\cos\theta_{\mathbf{H}}) \stackrel{def}{=} c \exp\left(-\frac{\theta_{\mathbf{H}}^2}{m^2}\right), \tag{4.290}$$

where c *is an arbitrary constant, and* m *denotes the root mean square slope of the facets. This distribution function is simple, fast to compute, and is good at matching reality as well as being pretty fast. But is has the ominous constant* c *which should be chosen to normalize the BRDF* f_r^{CT}*.*

Another, more accurate, but computationally more complex distribution function is the Beckmann distribution function. It is one of the most common microfacet distribution, does not need the ominous constant as in the Blinn model, and it is defined by:

$$D(\cos\theta_{\mathbf{H}}) \stackrel{def}{=} \frac{1}{m^2 \cos^4\theta_{\mathbf{H}}} \exp\left(-\frac{\tan^2\theta_{\mathbf{H}}}{m^2}\right). \tag{4.291}$$

When m *is small, such as* $m = 0.1$*, the microfacet slopes vary only slightly from macroscopy surface normal, this means, that the reflection is highly directional. Large* m*, i.e.* m *near one, imply that the associated v-grooves are deep, resulting in a rough surface that spreads out the light it reflects.*

REMARK 4.31 *To model surfaces that have multiple scales of roughness, [39, Cook & Torrance 1982] suggest to use a weighted sum of distribution functions:*

$$D(\cos\theta_{\mathbf{H}}) = \sum_{i=j}^{n} w_j D_j(\cos\theta_{\mathbf{H}}), \tag{4.292}$$

where the sum of the weights w_j *yields one.*

REMARK 4.32 *There are a series of other physically based BRDF models, such as the* He-model, *the* Oren-Nayar BRDF *or the* Ashikhmin-Shirley reflection model.

The He-model, *[85, He & al. 1991], is one of the most comprehensive BRDF models. In literature, it is often also known as the He-Torrance model. It accounts for example many effects such as polarization of light and subsurface scattering as well as the interaction of light at anisotropic surfaces. But the He-model has the disadvantage, that it is computationally very intensive. The* Oren-Nayar BRDF, *[142, Oren & Nayar 1994], is particularly interesting as it applies microfacet theory to Lambertian microfacets.*

The Ashikhmin-Shirley BRDF *incorporates some ideas from the Schlick and the Lafortune BRDF, introduced in [178, Schlick 1993] and [115, Lafortune & al. 1997]. It is a modern version of the Phong BRDF, or rather of the Blinn-Phong BRDF, that uses the same exponentiated cosine-lobe. The Ashikhmin-Shirley BRDF is physically plausible and models a number of physical effects, [14, Ashikhmin and Shirley 2000].*

4.2.2.5.4 BRDF MODELS BASED ON MEASURED DATA

Finally, we present with the *Ward BRDF* a further reflection model based on empirical data introduced to fit measured data.

WARD MODEL. In [231, Ward 1992] the approach was made to generate a BRDF which is not only easily implemented but is also useful for precisely matching the observed behavior from measurements at most occurring materials. Using only a few simple parameters, the Ward model is not only easy to control, but can be sampled efficiently for Monte Carlo integration. The Ward BRDF is derived for describing the reflection behavior of light at isotropic surfaces, but can be extended in a straightforward way to describe reflection at anisotropic surfaces.

Similar to the Torrance-Sparrow model, also the Ward model uses a Gauss distribution to describe the irregularities of a surface in a stochastically way. Since the geometry as well as the Fresnel term—typically involved in physically-based BRDFs—is difficult to integrate, in Wards reflectance model they are approximated by a term normalizing the BRDF.

DEFINITION 4.28 (The Isotropic Ward BRDF) *The* Ward BRDF *for isotropic surfaces is defined as:* Isotropy (335)

$$f^{W}_{r,iso}(\mathbf{s}, \omega_i \to \omega_o) = \frac{\rho_{dh}(\mathbf{s})}{\pi} + \rho_{dd}(\mathbf{s}) \frac{\exp^{-\tan^2 \frac{\langle \mathbf{H}(\mathbf{s}), \mathbf{N}(\mathbf{s}) \rangle}{\alpha^2}}}{4\pi\alpha^2 \sqrt{\langle \omega_i, \mathbf{N}(\mathbf{s}) \rangle \langle \omega_o, \mathbf{N}(\mathbf{s}) \rangle}}, \tag{4.293}$$

where ρ_{dd}, and ρ_{dh} are the directional-directional as well as the directional-hemispherical reflectances, α is the standard deviation of the microfacet slope, $\mathbf{H}$ is the half vector between ω_i and ω_o, and $\mathbf{N}$ is the surface normal at point $\mathbf{s}$. ρ_{dd} (338) ρ_{dh} (338) Standard Deviation (213)

Obviously, the isotropic Ward BRDF is the sum of two components. The diffuse term, given by $\frac{\rho_{dh}}{\pi}$, and a Gaussian gloss lobe defined by two parameters, the directional-directional reflectance ρ_{dd} and the microfacet slope α, simulating the roughness of the underlying surface. While the specular reflectance ρ_{dd} controls the magnitude of the lobe, α governs the width of the lobe.

REMARK 4.33 *Letting the roughness parameter α go to infinity, the surface gets perfectly diffuse, that is, the Ward BRDF is independent of the direction and can be written as:*

$$f_{r,iso}^{W,o}(\mathbf{s}, \omega_i \to \omega_o) \overset{(4.169)}{=} \frac{\rho_{dh}(\mathbf{s})}{\pi}. \tag{4.294}$$

Ignoring the diffuse term and letting the roughness parameter vanish, the Ward BRDF simulates ideal reflection at a smooth surface, and the BRDF is given by

$$f_{r,iso}^{W,\vee}(\mathbf{s}, \omega_i \to \omega_o) \overset{(4.104)}{=} \rho_{dd}(\mathbf{s}) \frac{\delta(\omega_i - \omega_o)}{|\cos\theta_i|}, \tag{4.295}$$

with $\omega_o = M_{\mathbf{N}}(\omega_i)$.

Extending the Gaussian distribution in Wards reflection model to surfaces with two perpendicular slope distributions α_x and α_y, then leads to *Wards anisotropic BRDF*.

Anisotropy (335) DEFINITION 4.29 (The Anisotropic Ward BRDF) *The Ward BRDF for anisotropic surfaces is defined as:*

$$f_r^W(\mathbf{s}, \omega_i \to \omega_o) = \frac{\rho_{dh}(\mathbf{s})}{\pi} + \rho_{dd}(\mathbf{s}) \frac{\exp^{-\tan^2\delta\left(\frac{\cos^2\phi}{\alpha_x^2} + \frac{\sin^2\phi}{\alpha_y^2}\right)}}{4\pi\alpha_x\alpha_y\sqrt{\langle\omega_i, \mathbf{N}(\mathbf{s})\rangle\langle\omega_o, \mathbf{N}(\mathbf{s})\rangle}}, \tag{4.296}$$

Standard Deviation (213) *where α_x and α_y are the standard deviations of the microfacet slope in x- and y-direction, $\mathbf{H}$ is the half vector between ω_i and ω_o, $\mathbf{N}$ is the surface normal at point $\mathbf{s}$, ϕ is the azimuth angle of the half vector projected into the surface plane.*

As in the isotropic case, the anisotropic Ward BRDF is also a sum of two components. The diffuse term, given by ρ_{dh}, a Gaussian anisotropic gloss lobe, now, defined by three parameters, ρ_{dd} and the surface slopes in the two orthogonal directions x and y-direction. While the specular reflectance controls the magnitude of the lobe, α_x and α_y govern the width of the lobe in the principal directions of anisotropy [229, Walter 2005].

Helmholtz Reciprocity (331) Contrary to the Phong model, the Ward BRDF is physically correct, i.e. it satisfies
Conservation of Energy (332) the Helmholtz reciprocity and the conservation of energy. Under the assumption, that the reflectances ρ_{dh} and ρ_{dd} together are smaller than one and the values α respectively α_x and α_y are not to large, we get a physically correct BRDF. This can be proofed via comparing of measured data from real materials with the BRDF. Such comparisons then

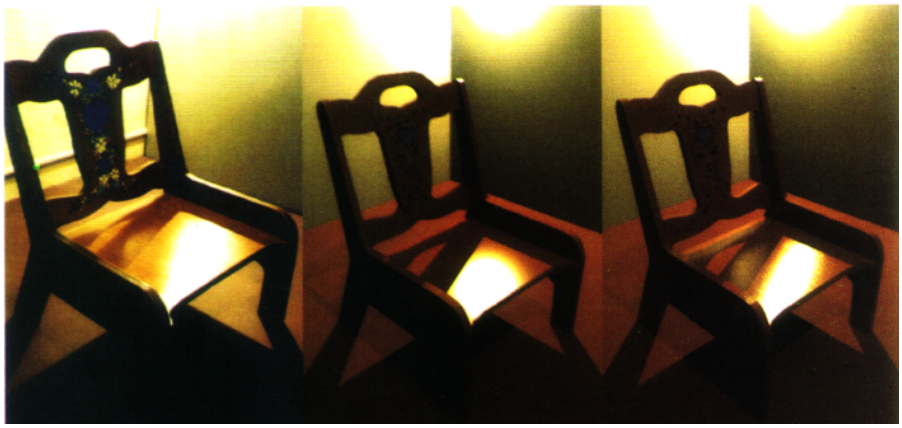


FIGURE 4.48: EXAMPLES OF THE ANISOTROPIC WARD BRDFS. Varnished wood comparison. On the left shows a photograph of a childs chair. The center shows a simulation of the chair using the isotropic Gaussian model with a strictly deterministic calculation. On the right shows a hybrid deterministic and stochastic simulation of the chair using the elliptical Gaussian model. Courtesy of Greg Ward Larson.

show that the Ward model can be used for modeling many materials, but that there are also materials that can not be approximated via this model. Images rendered via the anisotropic Ward model are shown in Figure 4.48.

Another interesting BRDF, similar to the Ward model, was published in [178, Schlick & al. 1993]. The Schlick BRDF is a combination of an empiric and a theoretical model. Since the Schlick BRDF uses only a few parameters, it is simple and efficient to evaluate, but offers possibilities to describe reflections in a preferably physical-plausible manner. This enable the Schlick BRDF to be used in hardware implementations.

4.2.3 BIDIRECTIONAL SCATTERING DISTRIBUTION FUNCTION

In computer graphics the two light phenomena of reflection and refraction are more and more treated together under the synonym of scattering. The mathematical concept behind scattering is the *bidirectional scattering distribution function* also briefly denoted as *BSDF*. Unlike the BRDF and BTDF, the BSDF is not a key concept in radiometry, but it plays a major role in computer graphics, and in our future theoretical and practical considerations, as it frees us to distinguish between reflection and transmission at surfaces. This, then makes our equations easier to handle. Chapter (3)

DEFINITION 4.30 (Bidirectional Scattering Distribution Function) *Let $\partial\mathcal{V}$ be the set of all* $\partial\mathcal{V}$ (41)

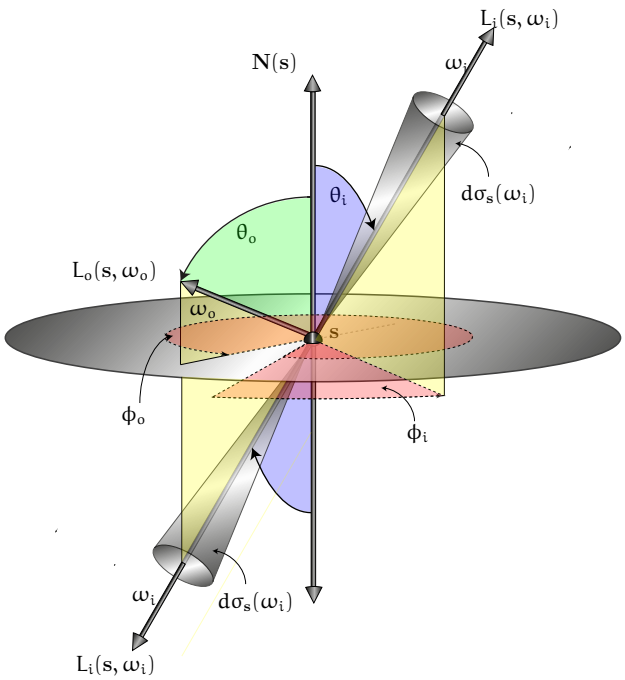


FIGURE 4.49: THE GEOMETRY OF THE BIDIRECTIONAL SCATTERING DISTRIBUTION FUNCTION. The BSDF is a four-dimensional function, defined on $\partial V \times S_i^2 \times S_o^2$ with values in $[0,\infty]$. It describes how much incident light, coming from direction ω_i, is reflected at a surface point in the outgoing direction ω_o.

2-dimensional surfaces in $\mathbb{R}^3$, s *be a point on any surface* $A \in \partial\mathcal{V}$, *furthermore* S_i^2 *be the set of all incident directions and* S_o^2 *be the set of all exitant directions around* s. *We call the measurable function* f_s *defined on*

Measurable Function (98)

$$f_s : \partial\mathcal{V} \times S_i^2 \times S_o^2 \to [0,\infty] \tag{4.297}$$

with

$$f_s(s,\omega_i \to \omega_o) \stackrel{def}{=} \frac{dL_o(s,\omega_i \to \omega_o)}{dE(s,\omega_i)} = \frac{dL_o(s,\omega_i \to \omega_o)}{L_i(s,\omega_i)d\sigma_s^\perp(\omega_i)}, \qquad \left[\frac{1}{sr}\right] \tag{4.298}$$

the bidirectional scattering distribution function, *also briefly the* BSDF, *see Figure 4.49.*

BRDF (320)

Due to its definition, the BSDF can be interpreted as the union of an upper and a

BTDF (330) lower BRDF as well as an upper and a lower BTDF, i.e. instead four, there is only still one single function with which we have to work. The advantage of this construct is that we do not more need to handle the light behavior at each side of the involved surface separately by a BRDF and an BTDF, which in the theory leads to more simpler equations.

To be physically valid, a BSDF must satisfy the same properties as a BRDF, that is, a BSDF should be a non-negative function, satisfy the principle of Helmholtz reciprocity,

$$f_s(\mathbf{s}, \omega_i \to \omega_o) = f_s(\mathbf{s}, \omega_o \to \omega_i) \tag{4.299}$$

and that of conservation of energy, also

$$\int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_o)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_o) \leq 1. \tag{4.300}$$

For deriving a BSDF that holds for arbitrary physically valid material, let us follow [221, Veach 1998]:

THEOREM 4.5 *Let f_s be the BSDF for a physically valid surface, which is either the boundary of an opaque object or the interface between two non-absorbing media. Provided that there are no external magnetic fields, then we have:*

$$\frac{f_s(\mathbf{s}, \omega_i \to \omega_o)}{\eta_o^2} = \frac{f_s(\mathbf{s}, \omega_o \to \omega_i)}{\eta_i^2}, \tag{4.301}$$

where η_o and η_i are functions of the exitant respectively the incident directions ω_o and ω_i, [221, Veach 1998].

PROOF 4.5 *Let us assume a scene is given in a vacuum or in a participating medium. Let furthermore $A \in \partial V$ be an opaque surface or an interface between two non-absorbing media. Considering the light, falling on a small area $d\mu^2(\mathbf{s})$ around point $\mathbf{s}$ from a small differential solid angle $d\sigma(\omega_i)$ around the direction ω_i, which is scattered—i.e. reflected or refracted, specularly or non-specularly—toward another cone $d\sigma(\omega_o)$ around ω_o. Then, the differential incident flux from ω_i, scattered in directions within $d\sigma(\omega_o)$ is equal to*

$$\begin{aligned} d\Phi_{io} &= L_o(\mathbf{s}, \omega_o)\, d\mu^2(\mathbf{s})\, d\sigma^{\perp}_{\mathbf{s}}(\omega_o) && (4.302)\\ &= f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i)\, d\mu^2(\mathbf{s})\, d\sigma^{\perp}_{\mathbf{s}}(\omega_o), && (4.303) \end{aligned}$$

while the flux incident from directions ω_o at $\mathbf{s}$ scattered in directions from $d\sigma(\omega_i)$ is given by:

$$\begin{aligned} d\Phi_{oi} &= L_i(\mathbf{s}, \omega_i)\, d\mu^2(\mathbf{s})\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) && (4.304)\\ &= f_s(\mathbf{s}, \omega_o \to \omega_i)\, L_i(\mathbf{s}, \omega_o)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_o)\, d\mu^2(\mathbf{s})\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i). && (4.305) \end{aligned}$$

Now, in free space as well as in a non-absorbing medium it holds $d\Phi_{io} = d\Phi_{oi}$, *that is,*

$$\begin{aligned} d\Phi_{io} &= d\Phi_{oi} && (4.306)\\ f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i) &= f_s(\mathbf{s}, \omega_o \to \omega_i)\, L_i(\mathbf{s}, \omega_o). && (4.307) \end{aligned}$$

Substituting Kirchhoff's equilibrium radiance law, see [221, Veach 1998], that is,

$$\frac{L_i(\mathbf{s}, \omega_i)}{\eta_i^2} = \frac{L_i(\mathbf{s}, \omega_o)}{\eta_o^2} \tag{4.308}$$

in the foregoing equation yields:

$$\frac{f_s(\mathbf{s}, \omega_i \to \omega_o)}{\eta_o^2} = \frac{f_s(\mathbf{s}, \omega_o \to \omega_i)}{\eta_i^2}, \tag{4.309}$$

where η_o *and* η_i *are functions of the exitant respectively the incident directions* ω_o *and* ω_i, *the so-called* refraction indices *of the incident and the exitant media.*

REMARK 4.34 *The result of the above theorem has important consequences for deriving bidirectional algorithms for solving the global illumination problem.*

Bidirectional Path Tracing (717)

Bidirectional algorithms, as we will present them in Section 9.3, are based on the idea to connect two independently generated subpaths, one starting from a light source and the other starting from the eye. As we have seen above, non-symmetric scattering occurs whenever light is refracted. This means that a bidirectional algorithm that uses a non-symmetric BSDF has to use different scattering rules depending on weather paths are started from the eye or from a light source, which finally results in two different transport equations.

Monte Carlo Light Tracing (710)
Monte Carlo Path Tracing (692)

Let us now reformulate Equation (4.298) as follows, i.e. writing

$$dL_o(\mathbf{s}, \omega_i \to \omega_o) = f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i) \tag{4.310}$$

SSEQ (319)

and integrating similar to our procedures for deriving the subsurface scattering equation over all incident directions of the unit sphere, then we get the so-called *scattering equation*:

DEFINITION 4.31 (Scattering Equation) *The* scattering equation *indicates the fraction of incident radiance at point* $\mathbf{s}$ *from all directions that is scattered only in direction* ω_o, *it is given by:*

$$L_o(\mathbf{s}, \omega_o) \overset{def}{=} \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i). \tag{4.311}$$

The scattering equation can be used to predict the appearance of the surface, given a description of the incident illumination.

Reflectance (336)

REMARK 4.35 (Reflectances Defined on the Basis of the BSDF) *For the sake of completeness, we also mention that the concept of reflectance, introduced in Section 4.2.2.3 can also be transferred on the BSDF by simply replacing the BRDF in all existing equations by the BSDF, thus:*

$$\rho_{dd} \equiv \rho(\omega_i \to \omega_o) \tag{4.312}$$
$$\rho_{ds} \equiv \rho(\omega_i \to S_o^2) \tag{4.313}$$
$$\rho_{sd} \equiv \rho(S_i^2 \to \omega_o) \tag{4.314}$$
$$\rho_{ss} \equiv \rho(S_i^2 \to S_o^2). \tag{4.315}$$

Via the notion of the reflectance, also the theoretical concept of the BSDF can be expressed in terms of the rather more practical used concept of reflectance.

REMARK 4.36 (Components of a BSDF) *Based on the definition of the BRDF and the BTDF, a BSDF can also be seen as composed of a diffuse part, f_s^o, a specular part, $f_s^\vee$, and and a glossy or directional diffuse part, f_s^{gl}, that is, we will often use the BSDF in the following form*

$$f_s = f_s^o + f_s^\vee + f_s^{gl}. \tag{4.316}$$

4.2.4 PHASE FUNCTIONS

When a light quantum collides with a particle that has an index of refraction different from its environment, the photon is refracted from its original direction of motion. This phenomenon of refraction is not only dependent on the size and shape of the particle, but also on the wavelength and the angle of the incident photon with respect to the viewer. Thus, the scattering properties of a particle can be described by a so-called *scattering profile*, [164, Prahl 1988], or in other words by a *phase function*, which differs in general from particle to particle.

In this section, we introduce the concept of the *phase function*. As a tool for describing the directional dependent scattering behavior of photons striking particles in a participating medium. It plays an important role in volume scattering. We also talk about properties of the phase function and present the most interesting analytical models of phase functions for the field of computer graphics. Let us start with the definition of a phase function.

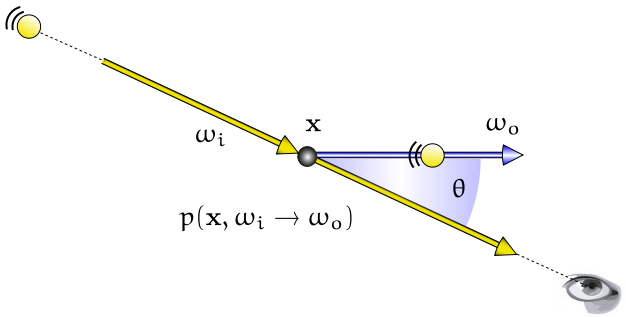


FIGURE 4.50: THE GEOMETRY UNDERLYING THE DEFINITION OF THE PHASE FUNCTION. For illustrating the scattering behavior of photons at particles in participating media we go back to the convention commonly used in the scattering literature. The incident direction ω_i always points toward the point where scattering happens and the outgoing direction ω_o points away from it. Obviously, this notation is different from that used for BRDFs. To conform this notation with that used for BRDFs, we have to replace in our formulas for the light transport the incident direction ω_i in the associated phase function by $-\omega_i$.

DEFINITION 4.32 (Phase Function, p) *Let us consider a photon coming from direction ω_i which collides with a particle in a participating medium at point $x \in \mathcal{V}^o$. Depending on the shape and size of the particle as well as the incident direction ω_i, then the photon will be refracted in direction ω_o, see Figure 4.50. The fraction of light at point x scattered from direction ω_i into direction ω_o can now be described by a function, the so-called* phase function p*, that is, a mapping* $\mathcal{V}^o$ (41)

$$p : \mathcal{V}^o \times S_i^2 \times S_o^2 \to [0, \infty] \tag{4.317}$$

with

$$(x, \omega_i \to \omega_o) \mapsto p(x, \omega_i \to \omega_o), \tag{4.318}$$

and the normalization property

$$\int_{S^2(x)} p(x, \omega_i \to \omega_o)\, d\sigma_x(\omega_o) = 1 \qquad \left[\frac{1}{sr}\right]. \tag{4.319}$$

REMARK 4.37 *From Figure 4.50 it can be seen that we use a different convention for the direction vectors at a scattering event from the one usually used in computer graphics when considering scattering at a surface. Instead of facing both directions away from the scattering effect, we have used the widely spread convention for scattering at a particle, where the incoming direction vector points toward the scattering point.*

Most phase functions are symmetrical around the incident direction ω_i depending only on the angle between the incident and exitant direction, thus $\langle \omega_i, \omega_o \rangle = \cos\theta$, hence they are often written as functions of the variable $\cos\theta$, namely by:

$$p(\cos\theta). \tag{4.320}$$

EXAMPLE 4.3 (1-dimensional Phase Functions) *The normalization condition for a* 1*-dimensional phase function has the form:*

$$\begin{aligned}
\int_{S^2(\mathbf{x})} p\left(\langle \omega_i, \omega_o \rangle\right) d\sigma_{\mathbf{x}}(\omega_o) &\overset{(2.186)}{=} \int_{[0,2\pi)} \int_{[0,\pi]} p(\cos\theta) \sin\theta \, d\mu(\theta) d\mu(\phi) && (4.321)\\
&= 2\pi \int_{[0,\pi]} p(\cos\theta) \sin\theta \, d\mu(\theta) && (4.322)\\
&= 2\pi \int_{[\pi,0]} p(\cos\theta) d\mu(\cos\theta) && (4.323)\\
&= 2\pi \int_{[\cos\pi, \cos 0]} p(x) d\mu(x) && (4.324)\\
&= \underbrace{2\pi \int_{[-1,1]} p(x) d\mu(x)}_{1}, && (4.325)
\end{aligned}$$

that is, with the result from Example 2.44, the normalization factor is $\frac{1}{4\pi}$.

REMARK 4.38 *Commonly, we will only consider* 1D *phase functions depending on the variable* $\cos\theta$. *Only in rare cases, such as when we have to consider media with a crystalline structure, it is required to discuss phase functions depending on more, in particular, two directions.*

REMARK 4.39 *Due to its definition, the phase function can be interpreted as a probability density function defined on the probability space* $(S^2, \mathfrak{B}(S^2), \mathbb{P})$ *in this case it gives the probability that a photon incident at point* $\mathbf{x}$ *from direction* ω_i *will be scattered in a differential solid angle around direction* ω_o.

PDF (176)
Probability Space (163)

PHYSICAL PROPERTIES OF PHASE FUNCTIONS. Considered as a PDF, a phase function must be a non-negative measurable function, that is, it must hold:

PDF (176)
Measurable Function (98)

$$p(\mathbf{x}, \omega_i \to \omega_o) \geq 0 \tag{4.326}$$

for all $\omega_i, \omega_o \in S^2$.

It should also be clear that the normalization condition (4.319) entails energy conservation. This is, as we will see further below, an important property which makes it possible to replace the hypothetical scattering kernel in the stationary particle transport equation, thus the SPTE, by the phase function.

Scattering Kernel (284)
SPLTE (294)

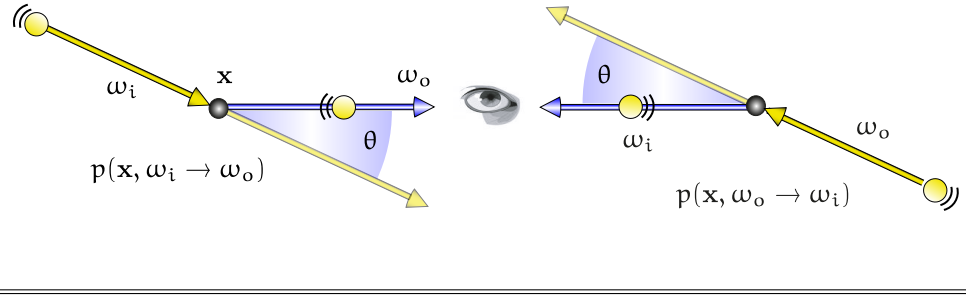


FIGURE 4.51: HELMHOLTZ RECIPROCITY OF THE PHASE FUNCTION. The value of the function remains unchanged if the direction of light is reversed, that is, the location of the viewer and the light source can be swapped. This can mathematically by expressed via $p(\mathbf{x}, \omega_i \to \omega_o) = p(\mathbf{x}, \omega_o \to \omega_i)$.

As known from our concept of the BRDF, also phase functions satisfy the principle of *Helmholtz reciprocity*, that is, the value of a phase function remains equal even if the incoming and outgoing direction of the involved photons are interchanged. So, reversing the direction of light does not change the fraction of light that is scattered. The reciprocity property of the phase function is illustrated in Figure 4.51, mathematically it can be expressed as:

BRDF (320)

Helmholtz Reciprocity (331)

$$p(\mathbf{x}, \omega_i \to \omega_o) = p(\mathbf{x}, \omega_o \to \omega_i) \tag{4.327}$$

for all $\omega_i, \omega_o \in S^2(\mathbf{x})$. This is also the reason why the notation

$$p(\mathbf{x}, \omega_i \leftrightarrow \omega_o) \stackrel{\text{def}}{=} p(\mathbf{x}, \omega_i \to \omega_o) = p(\mathbf{x}, \omega_o \to \omega_i) \tag{4.328}$$

is justified.

EXAMPLE 4.4 (1-dimensional Phase Functions, Continued) *The Helmholtz reciprocity of 1-dimensional phase functions can easily be shown by*

$$\cos\theta = \langle \omega_i, \omega_o \rangle = \langle \omega_o, \omega_i \rangle = \cos\theta \tag{4.329}$$

thus

$$p(\langle \omega_i, \omega_o \rangle) = p(\langle \omega_o, \omega_i \rangle). \tag{4.330}$$

To specify the preferred scattering direction of a particle, a parameter called the *average cosine*, also denoted as the *asymmetry parameter* is used. In the literature this parameter is often denoted by g and defined as follows:

DEFINITION 4.33 (The Asymmetry Parameter, g) *The* asymmetry parameter, g, *also known as the* average cosine *is defined as the integral over all directions of the phase function multiplied by the cosine of the angle between* ω_i *and* ω_o*, thus:*

$$g \overset{\text{def}}{=} \int_{S^2(\mathbf{x})} p(\mathbf{x}, \omega_i \to \omega_o)\, \langle \omega_i, \omega_o \rangle \, d\sigma_{\mathbf{x}}(\omega_o) \tag{4.331}$$

$$\overset{(2.191)}{=} \int_{S^2(\mathbf{x})} p(\mathbf{x}, \omega_i \to \omega_o) \cos\theta \, d\sigma_{\mathbf{x}}(\omega_o) \tag{4.332}$$

with $\cos\theta = \langle \omega_i, \omega_o \rangle$.

Using the monotonicity of the Lebesgue integral from Lemma 2.2 in the definition of the asymmetry factor, then we can conclude that the average cosine g is a real number from $[-1, 1]$ since it holds:

$$-1 \overset{(4.319)}{=} -\int_{S^2(\mathbf{x})} p(\mathbf{x}, \omega_i \to \omega_o)\, d\sigma_{\mathbf{x}}(\omega_o) \tag{4.333}$$

$$\overset{-1 \le \cos\theta}{\le} \int_{S^2(\mathbf{x})} p(\mathbf{x}, \omega_i \to \omega_o)\, \cos\theta \, d\sigma_{\mathbf{x}}(\omega_o) \tag{4.334}$$

$$\overset{(4.331)}{=} g \tag{4.335}$$

$$g \overset{(4.331)}{=} \int_{S^2(\mathbf{x})} p(\mathbf{x}, \omega_i \to \omega_o)\, \cos\theta \, d\sigma_{\mathbf{x}}(\omega_o) \tag{4.336}$$

$$\overset{\cos\theta \le 1}{\le} \int_{S^2(\mathbf{x})} p(\mathbf{x}, \omega_i \to \omega_o)\, d\sigma_{\mathbf{x}}(\omega_o) \tag{4.337}$$

$$\overset{(4.319)}{=} 1. \tag{4.338}$$

This allows us to control the degree of anisotropy of the medium via the asymmetry factor g in such a way that a negative value of g indicates that particles are scattered preferably backwards, and $g > 0$ indicates that particles are scattered rather in forward direction. The value $g = 0$ simulates isotropic scattering. Isotropic scattering means, that energy is distributed equally in forward and backward direction. The greater the value of g, the more scattering occurs close to the incident direction ω_i, in the case of forward-scattering, or $-\omega_i$ for backward-scattering.

Isotropy (335)

EXAMPLE 4.5 (1-dimensional Phase Functions, Continued) *For the asymmetry parameter*

g *we get:*

$$g \overset{(4.331)}{=} \int_{S^2(\mathbf{x})} p\left(\langle \omega_i, \omega_o \rangle\right) \langle \omega_i, \omega_o \rangle \, d\sigma_{\mathbf{x}}(\omega_o) \tag{4.339}$$

$$\overset{(2.186)}{=} \int_{[0,2\pi]} \int_{[0,\pi)} p(\cos\theta) \cos\theta \sin\theta \, d\mu(\theta) d\mu(\phi) \tag{4.340}$$

$$= 2\pi \int_{[0,\pi)} p(\cos\theta) \cos\theta \sin\theta \, d\mu(\theta) \tag{4.341}$$

$$= 2\pi \int_{[\pi,0)} p(\cos\theta) \cos\theta d\mu(\cos\theta) \tag{4.342}$$

$$= 2\pi \int_{[-1,1)} p(x) \, x \, d\mu(x). \tag{4.343}$$

BRDF (320) As known from our discussions about BRDFs, so, there are also a variety of different phase function models, which range from parametrized models—usually defined by a small number of parameters—over measured data, to analytical models, incorporating the shape and material of the involved particles. We will now present the most relevant analytical SPLTE (294) phase functions, which can be used in the SPTE as scattering kernels.

ISOTROPIC PHASE FUNCTION. The simplest example of an analytical phase function is the *isotropic phase function*. It is a constant and defined by:

$$p_{iso}(\mathbf{x}, \omega_i \to \omega_o) = \frac{1}{4\pi}, \tag{4.344}$$

where the factor $\frac{1}{4\pi}$ results from the normalization condition (4.319). The isotropic phase function has units $\left[\frac{1}{sr}\right]$. It can be seen as the equivalent of diffuse reflection and can be interpreted in such a way that a photon, when it collides with a particle, will be scattered with the same probability in all directions over the unit sphere, see Figure 4.52.

HENYEY-GREENSTEIN PHASE FUNCTION. A commonly used non-isotropic phase function is the *Henyey-Greenstein phase function* [83, Henyey & Greenstein 1941]. It is based on an empirical model for simulating scattering of radiation in the galaxy and can be used to model a large variety of different scattering types, such as scattering in oceans, clouds, skin, stone, and much more. The Henyey-Greenstein phase function, see Figure 4.53, is intuitively controlled by the asymmetry parameter g. In its normalized form, the Henyey-Greenstein phase function is defined as:

$$p_{HG}(\mathbf{x}, \omega_i \to \omega_o) \overset{def}{=} \frac{1}{4\pi} \frac{1 - g^2}{(1 + g^2 - 2g\langle \omega_i, \omega_o \rangle)^{\frac{3}{2}}} \tag{4.345}$$

$$= \frac{1}{4\pi} \frac{1 - g^2}{(1 + g^2 - 2g\cos\theta)^{\frac{3}{2}}}. \tag{4.346}$$

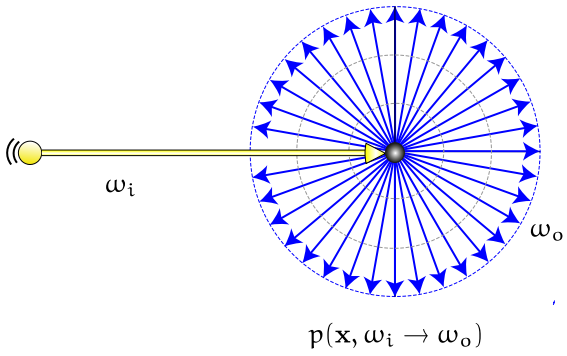


FIGURE 4.52: ISOTROPIC PHASE FUNCTION. The amount of light scattered by particles does not depend on the direction of the incident light ray and the viewing direction.

An other common non-isotropic phase functions is the *modified Henyey-Greenstein phase function*, [164, Prahl 1988], it has the form:

$$p(x, \omega_i \rightarrow \omega_o) \stackrel{def}{=} \frac{1}{4\pi}\left(\beta + (1-\beta)\frac{1-g^2}{(1+g^2-2g\langle\omega_i,\omega_o\rangle)^{\frac{3}{2}}}\right) \quad (4.347)$$

$$= \frac{1}{4\pi}\left(\beta + (1-\beta)\frac{1-g^2}{(1+g^2-2g\cos\theta)^{\frac{3}{2}}}\right), \quad (4.348)$$

where the first term represents the amount of light scattered isotropically and the second term contains the Henyey-Greenstein function. For $\beta = 0$, this phase functions reduces to the Henyey-Greenstein phase function.

For modeling more complex scattering properties of particles in [95, Jensen 2001] combinations of Henyey-Greenstein phase functions are suggested, such as:

$$p(x, \omega_i \rightarrow \omega_o) \stackrel{def}{=} \sum_{i=1}^{n} w_i \left(\frac{1-g_i^2}{4\pi\left(1+g_i^2-2g_i\langle\omega_i,\omega_o\rangle\right)^{\frac{3}{2}}}\right) \quad (4.349)$$

$$= \sum_{i=1}^{n} w_i \left(\frac{1-g_i^2}{4\pi\left(1+g_i^2-2g_i\cos\theta\right)^{\frac{3}{2}}}\right), \quad (4.350)$$

where $\sum_{i=1}^{n} w_i = 1$ is a sum of weights and g_i controls the shape of each lobe. Such combinations of Henyey-Greenstein phase functions can give very realistic results when forward scattering, $g > 0$, and backward scattering lobes, $g < 0$, are used.

SCHLICK PHASE FUNCTION. Since the shape of the Henyey-Greenstein phase function is similar to an ellipsoid, [24, Schlick & al. 1993] recommend to approximate it by an

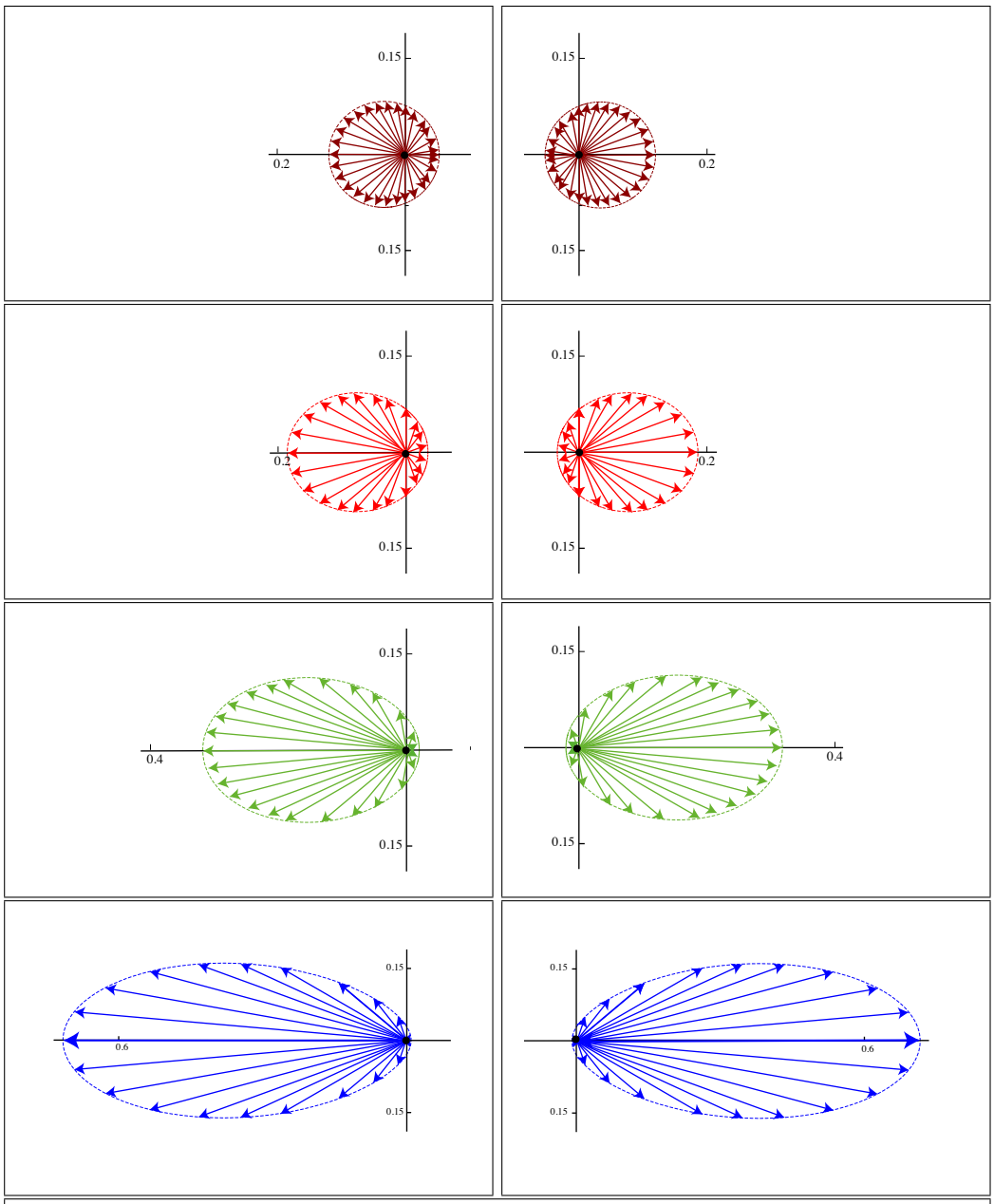


FIGURE 4.53: HENYEY-GREENSTEIN PHASE FUNCTIONS. The polar plots of the Henyey-Greenstein phase function for anisotropy parameter $g = \pm 0.2, \pm 0.4, \pm 0.6, \pm 0.8$. The plots show the intensity of scattering as a function of ω_i and ω_o in forward and backward direction. Higher values of g increase the probability of scattering in forward direction, while smaller values scatters light more in backward direction.

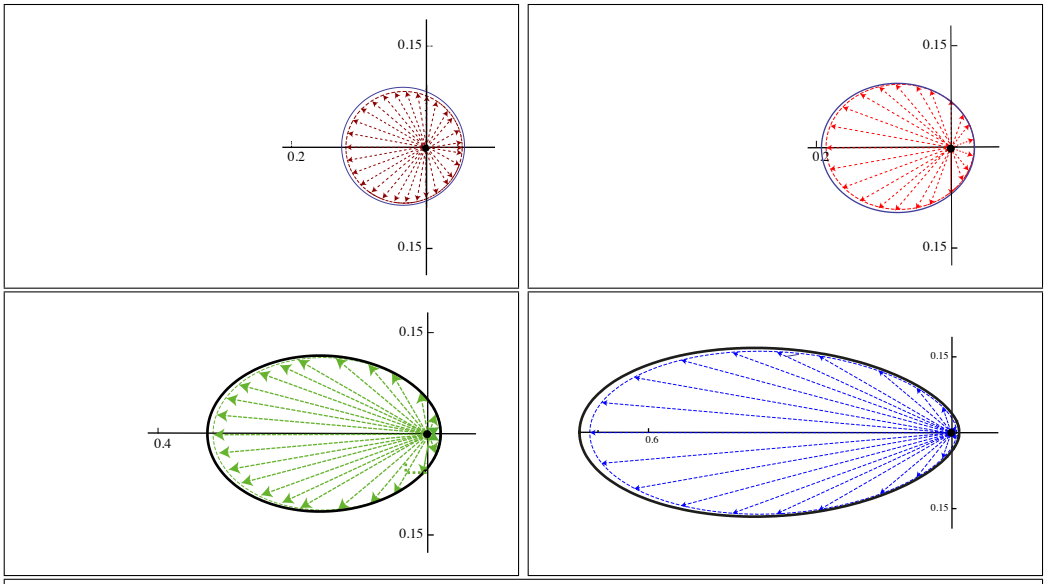


FIGURE 4.54: SCHLICK PHASE FUNCTIONS. The polar plots of the Schlick phase function for anisotropy parameter value $g = 0.2, 0.4, 0.6, 0.8$. The plots show the intensity of scattering as a function of ω_I and ω_o in forward direction. Higher values of g increase the probability of scattering in forward direction.

ellipsoid which would eliminate the relatively costly computation of the $\frac{3}{2}$ exponent in the denominator of the Henyey-Greenstein phase function, see Figure 4.54. *Schlick's phase function* is defined as:

$$p(\mathbf{x}, \omega_i \rightarrow \omega_o) \stackrel{def}{=} \frac{1-k^2}{4\pi(1+k\langle\omega_i,\omega_o\rangle)^2} \tag{4.351}$$

$$= \frac{1-k^2}{4\pi(1+k\cos\theta)^2}. \tag{4.352}$$

Here, $k \in]-1,1[$ acts similarly to the parameter g in the Henyey-Greenstein phase function, that is, it controls the preferred direction of the scattering. This means: $k = 0$ corresponds to isotropic scattering, $k > 0$ is forward scattering, and $k < 0$ results in backward scattering. Due to [159, Pharr and Humphreys 2010] an accurate approximation to the Henyey-Greenstein phase function is given by the polynomial

$$k \approx 1.55g - 0.55g^3 \tag{4.353}$$

for intermediate values of k.

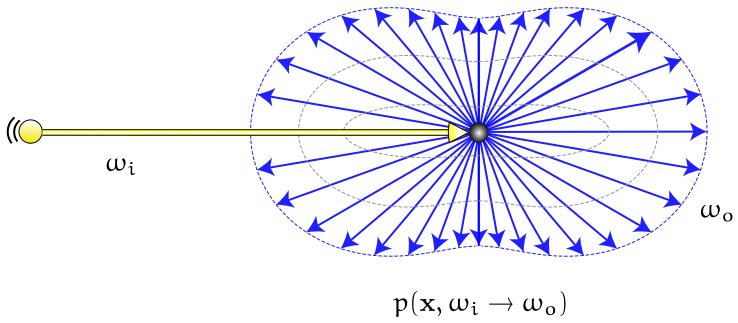


FIGURE 4.55: RAYLEIGH PHASE FUNCTION. When particles are somewhat smaller than the wavelength of light, the scattering at these particles can be described by the Rayleigh phase function. The Rayleigh phase function is isotropic around the incident direction.

RAYLEIGH PHASE FUNCTION. Another type of an often used non-isotropic phase function is the *Rayleigh phase function*, named after Lord Rayleigh who explained the blue color of the sky as a result of light scattering predominantly in the blue end of the spectrum with the help of this function [33, Chandrasekhar 960] and [19, Beckmann & al. 1987].

Rayleigh scattering models the scattering for extremely small spherical particles such as molecules of the air. It can be extended to scattering from particles up to about a tenth of the wavelength of the light. The Rayleigh phase function is defined as

$$p(x, \omega_i \to \omega_o) \stackrel{\text{def}}{=} \frac{1}{4\pi}\frac{3}{4}\left(1 + \langle \omega_i, \omega_o \rangle^2\right) \tag{4.354}$$

$$= \frac{1}{4\pi}\frac{3}{4}(1 + \cos^2\theta), \tag{4.355}$$

see Figure 4.55.

REMARK 4.40 (Lorenz-Mie Scattering) *In connection with Rayleigh scattering let us also mention Lorenz-Mie scattering, which is based on a more complex theory. Lorenz-Mie scattering is derived from Maxwell's equations and can be used to describe scattering by spherical particles, whose size is comparable to the wavelength of light, such as water droplets of fog. Lorenz-Mie theory can be used to derive phase functions for a homogeneous collection of spherical particles where any ratio of diameter to wavelength is allowed. In [140, Nishita & al. 1987] two empirically derived approximations to the complicated Lorenz-Mie scattering functions for foggy atmospheres*

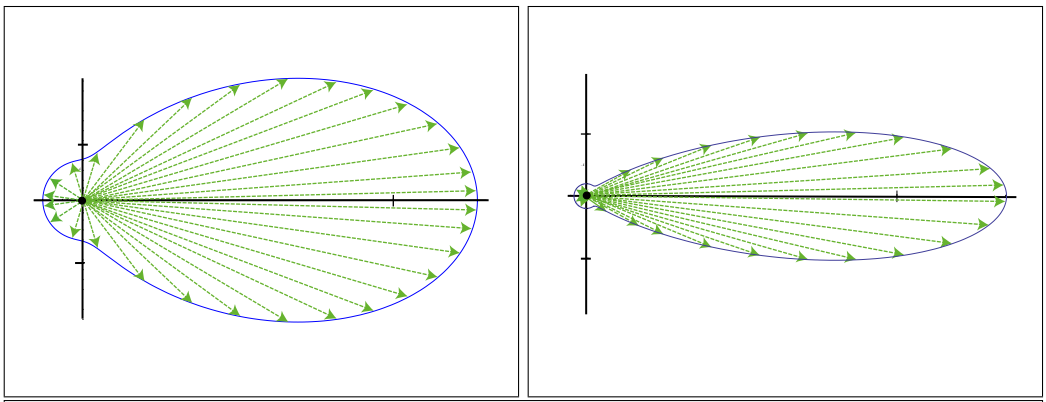


FIGURE 4.56: LORENZ-MIE PHASE FUNCTIONS. The polar plots of the hazy, left image, and the murky phase functions on the right side. They are approximations to the behavior of scattering in foggy atmospheres based on Lorenz-Mie theory.

are presented, one for hazy *atmospheres:*

$$P_{MH}(\mathbf{x}, \omega_i \rightarrow \omega_o) \stackrel{\text{def}}{=} \frac{1}{4\pi}\left(\frac{1}{2} + \frac{9}{2}\left(\frac{1+\cos\theta}{2}\right)^2\right), \tag{4.356}$$

and one for murky *atmospheres:*

$$P_{MH}(\mathbf{x}, \omega_i \rightarrow \omega_o) \stackrel{\text{def}}{=} \frac{1}{4\pi}\left(\frac{1}{2} + \frac{33}{2}\left(\frac{1+\cos\theta}{2}\right)^{32}\right), \tag{4.357}$$

see Figure 4.56.

4.3 LIGHT SOURCES

For rendering a virtual scene, we need light sources that are responsible for illuminating the existing objects. Now, in Section 2.1.3 we have already meet two different types of light sources: ideal point light sources and area light sources. In Chapter 3 we then discussed some properties of these types of light sources. In the current section, we will now extend our interest in light sources, so we will discuss more properties of point and area light sources and present some further types of light sources useful for rendering different light effects.

POINT LIGHT SOURCES. In Definition 2.12 we introduced an ideal point light source as the center of a spherical field of light, where light is uniformly radiated in all directions.

Point lights are idealized light sources, that do not exist in the real world. They can rather be considered as abstractions of light sources that are far away compared to the size of the light sources. Physically, a point light is a source of radiant power, that adds, in a given period of time, a certain amount of energy to the environment. Now, due to the point singularity at the source, it is difficult to describe the energy distribution of a point light source via the radiometric concept of radiance. That is, the remaining radiometric quantities to describe the illumination behavior of a point light source can only be: irradiance, or radiant intensity. As the Inverse Square Law from Example 3.5 shows, the irradiance contribution of a point light source varies inversely to the square of the distance between the point light source and the illuminated surface. In deed, this is mathematically correct, but often results in unnatural lighting of a scene. Therefore, a useful measure for modeling the illumination behavior of a point light source is the radiometric quantity radiant intensity, introduced in Section 3.6. In contrast to irradiance, radiant intensity, as Equation (3.94) shows, does not change with distance from the light source.

Radiant Power (249) Radiance (250) Irradiance (257) Radiant Intensity (267) Radiant Intensity (267)

EXAMPLE 4.6 (Direct Illumination due to Point Light Sources) *Let us consider a scene consisting of opaque surfaces illuminated by a point light source, $*$, located at point $\mathbf{x} \in \mathbb{R}^3$. Obviously, the direct illumination at surface point $\mathbf{s}$ can then be described by the reflectance equation, that is:*

Reflectance Equation (321)

$$L_o(\mathbf{s}, \omega_o) = \int_{\mathcal{H}_i^2(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i). \tag{4.358}$$

Since the incident radiance at point $\mathbf{s}$ only comes from a single direction ω_i^l towards the point light $*$, instead to integrate over the whole upper hemisphere, we only has to account for direction ω_i^l. Using a Dirac δ-distribution, the reflected radiance can be expressed as:

Dirac-δ Distribution (117)

$$\begin{aligned} L_o(\mathbf{s}, \omega_o) &= \int_{\mathcal{H}_i^2(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, \delta_{\sigma^\perp}(\omega_i - \omega_i^l)\, L_i(\mathbf{s}, \omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i) && (4.359)\\ &\overset{(2.302)}{=} f_r(\mathbf{s}, \omega_i^l \to \omega_o)\, L_i(\mathbf{s}, \omega_i^l) && (4.360)\\ &= f_r(\mathbf{s}, \omega_i^l \to \omega_o)\, L_e(\mathbf{x}, -\omega_i^l) && (4.361) \end{aligned}$$

with $\mathbf{x} = \gamma(\mathbf{s}, \omega_i^l)$. Using Relation (3.99), then the reflected radiance at point $\mathbf{s}$ in direction ω_o can also be expressed in terms of emitted radiant power of the point light $*$, namely by:

$$L_o(\mathbf{s}, \omega_o) = f_r(\mathbf{s}, \omega_i^l \to \omega_o)\, \frac{\Phi_e(\mathbf{x})}{4\pi \|\mathbf{s} - \mathbf{x}\|_2^2} \left|\cos \theta_i^l\right|. \tag{4.362}$$

Equation (4.362) says that in a conventional implementation of a ray tracer, where only point light sources are used, the exitant radiance at point $\mathbf{s}$ in direction

Radiance (250)

Radiant Power (249) *ω can be computed via the radiant power emitted from the light source.*

Assuming, that the scene is illuminate by a finite set of point lights $*_1, \ldots, *_n$, then the reflected radiance at point $\mathbf{s}$ in direction ω_o is given by:

$$\begin{aligned} L_o(\mathbf{s}, \omega_o) &= \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) && (4.363)\\ &= \sum_{j=1}^{n} f_r(\mathbf{s}, \omega_i^{l_j} \to \omega_o)\, L_i(\mathbf{s}, \omega_i^{l_j}) \left|\cos\theta_i^{l_j}\right| && (4.364)\\ &= \sum_{j=1}^{n} f_r(\mathbf{s}, \omega_i^{l_j} \to \omega_o)\, L_e(\mathbf{x}_j, -\omega_i^{l_j}) \left|\cos\theta_i^{l_j}\right| && (4.365)\\ &= \sum_{j=1}^{n} f_r(\mathbf{s}, \omega_i^{l_j} \to \omega_o)\, \frac{\Phi_e(\mathbf{x}_j)}{4\pi \left\|\mathbf{x}_j - \mathbf{s}\right\|_2^2} \left|\cos\theta_i^{l_j}\right|, && (4.366) \end{aligned}$$

with $\mathbf{x}_j = \gamma(\mathbf{s}, \omega_i^{l_j})$ and $\left|\cos\theta_i^{l_j}\right|$ corresponds to the cosine between the surface normal at point $\mathbf{s}$ and direction $\omega_i^{l_j}$ towards $*_j$. We leave the details of this derivation to the interested reader as an exercise.

Although the concept of the point light source is an idealization, point lights play a central role in rendering algorithms as they serve as the basis of a series of other types of light sources.

AREA LIGHT SOURCES. Based on the concept of the point light source, we introduced in Definition 2.13 the type of the area light source as a 2-dimensional surface, whose points act as ideal point light sources. Since all light sources in real world have some amount of surface area, area light sources are the real light emitters that should be simulated in rendering algorithms. In contrast to point light sources that illuminate a surface point $\mathbf{s}$ from only a single direction, area light sources illuminate such a point from a range of directions, namely, the solid angle subtended by the light source as seen from $\mathbf{s}$. This implies soft shadows and smooth light effects, see Figure 4.57.

EXAMPLE 4.7 (Direct Illumination due to Area Light Sources) *A well-known problem in almost all rendering algorithms is the computation of direct illumination at a surface point $\mathbf{s}$ due to area lights, $☼_1, \ldots, ☼_n$, thus, the evaluation of the integral*

$$\begin{aligned} L_o(\mathbf{s}, \omega_o) &= \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) && (4.367)\\ &= \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, L_e(\gamma(\mathbf{s}, \omega_i), -\omega_i))\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i), && (4.368) \end{aligned}$$

where $\gamma(\mathbf{s}, \omega_i)$ are points at one of the area light sources $☼_1, \ldots, ☼_n$.

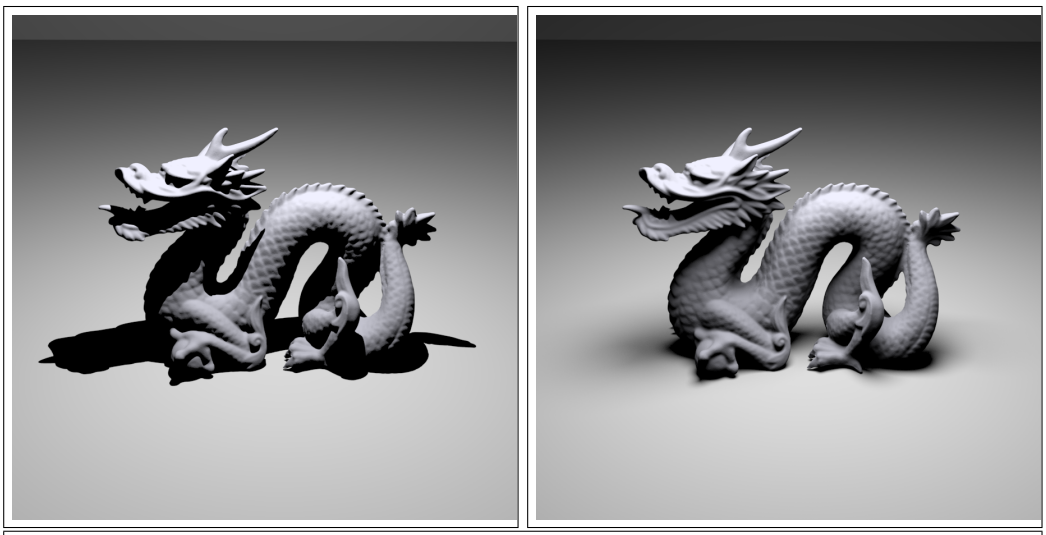


FIGURE 4.57: A SURFACE ILLUMINTATED BY A POINT LIGHT SOURCE AND AN AREA LIGHT SOURCE. On the left, a point light source illuminates a surface from only a single direction. In contrast, the illumination of the same surface by an area light source. The surface point $\mathbf{s}$ receives light from directions within the solid angle subtended by the source and seen from $\mathbf{s}$. Image courtesy of Pharr and Humphreys.

As we will see in Section 4.4.2.2 the integration domain for direct illumination can be changed from $\mathcal{H}^2_i(\mathbf{s})$ to the union of solid angles of the light sources, that is, instead of to evaluate the integral over the whole hemisphere, we integrate the reflectance equation over $\bigcup_{j=1}^{n} \sphericalangle_j$ or the set of surfaces within the scene. This approach will lead to more efficient sampling techniques for computing the direct illumination at a surface point. Sampling of area light sources is a central point when discussing Monte Carlo path tracing in Section 9.1.

Section 6.7.4

REMARK 4.41 (Approximating an Area Light Source by an Array of Point Light Sources) *The effect of an area light can simply be approximated by a set $*_1, \ldots, *_n$ of point light sources that are uniformly or randomly distributed on a flat or curved shape. As the light comes from point light sources, the direct light integral from above can*

be captured using Dirac δ-distributions, that is:

$$\begin{aligned} L_o(s,\omega_o) &= \int_{\mathcal{H}^2_i(s)} f_r(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma^\perp_s(\omega_i) && (4.369)\\ &= \int_{\bigcup_{j=i}^{n} {}^*j} f_r(s,\omega_i \to \omega_o)\, \delta_{\sigma^\perp}(\omega_i \to \omega_i^{l_j})\, L_i(s,\omega_i))\, d\sigma^\perp_s(\omega_i) && (4.370)\\ &= \sum_{j=1}^{n} f_r(s,\omega_i^{l_j} \to \omega_o)\, L_e(\gamma(s,\omega_i^{l_j}),-\omega_i^{l_j}), && (4.371) \end{aligned}$$

where $\gamma(s,\omega_i^{l_j})$ *are points on one of the area light sources* $\sun_1,\ldots,\sun_n$.

DIRECTIONAL LIGHT SOURCES. *Directional lights*, often also known as *distant light sources*, are a special kind of point light sources, whose light travels only in a single direction through the whole scene to be rendered. Thus, a directional light can be considered as an approximation of a point or an area light source, far away compared to the size of the scene illuminated by the source.

DEFINITION 4.34 (Ideal Directional Light Source) *An* ideal directional light source, *often also called a* distant light source, *or a* light source at infinity, *corresponds to a point light source,* L_e, *or an area light source,* $L_e^{\partial\mathcal{V}}$, *that emits its light in a single direction* ω_o *over the unit sphere or the upper hemisphere.*

Due to the property that an ideal directional light source emits light particles only in a single direction ω_o, the emission of a directional light source can be quantified by measuring the power through a unit area surface perpendicular to direction ω_o. That is, irradiance is the radiometric quantity useful for measuring the emission of an ideal distant light source.

Irradiance (257)

EXAMPLE 4.8 (The Sun, the Prototype of an Ideal Directional Light Source) *A typical example of a distant light is the sun as considered from the earth. Since the solid angle subtended by the earth as seen from the sun corresponds to a tiny patch within the spherical field of the sun, radial rays emitted from the sun to an object become closer to parallel at the object gets farther away. That is, in the case of the earth that is so far away from the sun, the illumination effectively arrives in parallel beams.*

EXAMPLE 4.9 (Direct Illumination due to Directional Light Sources) *Let us now consider a scene consisting of opaque surfaces, illuminated by a directional light source—with or without surface area—then, the direct illumination at surface point* s *is described by the reflectance equation, that is:*

Reflectance Equation (321)

$$L_o(s,\omega_o) = \int_{\mathcal{H}^2_i(s)} f_r(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma^\perp_s(\omega_i). \qquad (4.372)$$

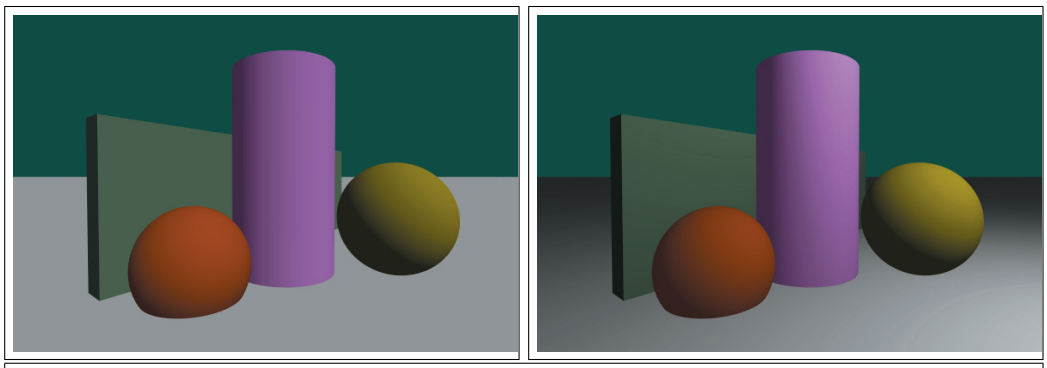


FIGURE 4.58: A SCENE RENDERED WITH DIRECTIONAL AND POINT LIGHT. On the left, a scene, consisting of two spheres, a cylinder, an axis-aligned box, and a plane, all rendered with directional illumination. As you can see, the color of the ground plane is constant, see Equation (4.374). The reason for that is, that the outgoing radiance is independent of ω_o and ω_i^l is the same at all surface points s. On the right, the same scene rendered with a point light source. Image Courtesy of Kevin Suffern, University of Technology, Sydney.

Since the incident radiance L_i *at any point* s *comes from a single direction* ω_l *towards the directional light, the reflected radiance can be expressed in terms of a Dirac* δ*-distribution, namely, by:*

Dirac δ-Distribution (117)

$$L_o(s,\omega_o) = \int_{\mathcal{H}_i^2(s)} f_r(s,\omega_i \to \omega_o)\,\delta_{\sigma^\perp}(\omega_i-\omega_i^l)\,L_i(s,\omega_i)\,d\sigma_s^\perp(\omega_i) \quad (4.373)$$

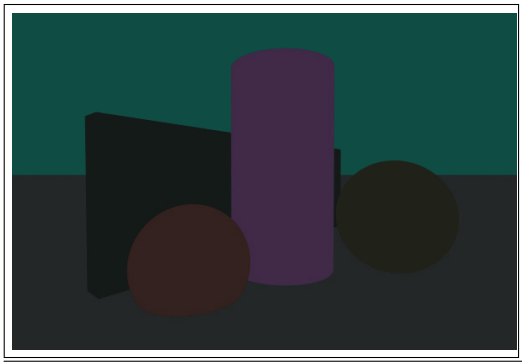
$$\overset{(2.302)}{=} f_r(s,\omega_i^l \to \omega_o)\,L_i(s,\omega_i^l). \quad (4.374)$$

SPOT LIGHT SOURCES. Another type of a light source also based on the concept of the point light, are *spot light sources*. Instead to emit their light in all directions, such as a point light source, or, in a single direction, as in the case of a directional light source, spot light sources emit light in a cone of directions starting at their locations within a scene.

DEFINITION 4.35 (Spot Light Source) *A* spot light source, *as commonly defined in computer graphics, corresponds to a point light source* L_e *at point* $x \in \partial\mathcal{V}$ *that emits its light in directions* ω_o *of a finite solid angle* Γ *on the unit sphere or the upper hemisphere about* x.

Solid Angle (83)

Since we defined spot light sources via the concept of the point light source, they also generate hard shadows. This disadvantage can be removed, by using an outer solid angle, Γ_o, who contains the solid angle Γ, that is, it must hold: $\Gamma \subset \Gamma_o$. Such a spot light then fully illuminates all objects inside the inner cone of angles, while the region of directions between Γ and Γ_o can be considered as a transition zone where the illumination weakens from full illumination to no illumination, such that points outside Γ_o are not illuminated



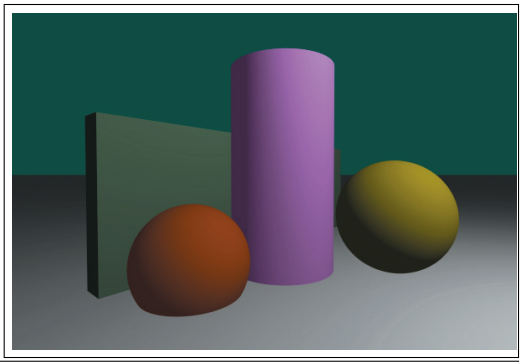


FIGURE 4.59: A SCENE RENDERED WITH AMBIENT AND POINT LIGHT. On the left, a scene, consisting of two spheres, a cylinder, an axis-aligned box, and a plane, all rendered with ambient illumination. As you can see, hopefully, all objects are rendered with constant colors, where each material can reflect a different fraction of the ambient illumination. On the right, the same scene rendered with a point light source. Image Courtesy of Kevin Suffern, University of Technology, Sydney.

at all. The same effect can also be achieved by using an area light source $L_e^{\partial V}$ instead of a point light as source of a spot light.

AMBIENT LIGHT SOURCES. Since objects in shadow or facing away from light sources within a scene are completely black—although just about all surfaces receive a little bit of light from somewhere—rendered images of such scenes appear highly unrealistic. Even if it is not approximately not correct for real scenes, the visual quality of such images can significantly improved by using a very simple model of indirect light, a so-called *ambient light source*, see Figure 4.59.

DEFINITION 4.36 (Ambient Light Source) *An* ambient light source*, is a hypothetical light source, that emits a constant radiance value* $L_a(\mathbf{s}, \omega_o) = C, C > 0$ *at all points and directions within a scene to be rendered.* Radiance (250)

There is no physical analog to an ambient light source in real world. Ambient light is typically used in computer graphics to approximate global illumination in scenes where no indirect light exists. Evidently, ambient light increases the level of background illumination, which implies that effects of other light sources are softened.

EXAMPLE 4.10 (Direct Illumination due to Ambient Light Sources) *For opaque Lambertian surfaces the reflected radiance due to ambient light at a surface point* $\mathbf{s}$ *in any* Lambertian Reflectors (349)

direction ω_o is given by:

$$L_o(s,\omega_o) \overset{(4.89)}{=} \int_{\mathcal{H}_i^2(s)} f_r^{LA}\,(L_a(\gamma(s,\omega_i),-\omega_i)) + L_i(s,\omega_i))\; d\sigma_s^{\perp}(\omega_i) \tag{4.375}$$

$$= \int_{\mathcal{H}_i^2(s)} f_r^{LA}\,(C + L_i(s,\omega_i))\, d\sigma_s^{\perp}(\omega_i) \tag{4.376}$$

$$= \frac{\rho_{dh}(s)}{\pi}\int_{\mathcal{H}_i^2(s)} C\, d\sigma_s^{\perp}(\omega_i) + \frac{\rho_{dh}(s)}{\pi}\int_{\mathcal{H}_i^2(s)} L_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i) \tag{4.377}$$

$$= \frac{\rho_{dh}(s)}{\pi}\left(C\,\pi + \int_{\mathcal{H}_i^2(s)} L_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i)\right), \tag{4.378}$$

f_r^{LA} (349) *where f_r^{LA} is the Lambertian BRDF and $L_a(\gamma(s,\omega_i)-\omega_i) = C$ is the ambient light contribution at point s. A similar formula can be derived for an arbitrary BRDF, namely:*

$$L_o(s,\omega_o) = C\int_{\mathcal{H}_i^2(s)} f_r(s,\omega_i \to \omega_o)\, d\sigma_s^{\perp}(\omega_i) + \tag{4.379}$$

$$\int_{\mathcal{H}_i^2(s)} f_r(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i). \tag{4.380}$$

We leave the derivation of this formula to the interested reader as a simple exercise.

REMARK 4.42 *Apart from the types of light sources introduced above, it is possible to define much more types of light sources in rendering, such as projection, sky, and textured lights, as well as goniophotometric diagram lights etc., for a detailed discussion see [158, Pharr & Humphreys 2004] or [1, Akenine-Möller & al. 2008].*

4.4 THE STATIONARY LIGHT TRANSPORT IN PARTICIPATING MEDIA AND IN A VACUUM

In Section 4.1.3 we have derived the stationary light transport equation in participating media in integral form. It is the governing equation that describes the behavior of light within a medium that absorbs, emits, and scatters light.

As we have seen, in contrast to the integro-differential form, the integral form of the SLTE also incorporates the interaction of light at the boundaries of the medium. That is, under certain conditions the SLTE can also be used to describe the light transport in a vacuum, where there is no absorption, emission, or scattering at all except on surfaces. To utilize this property of the SLTE for our further concerns with respect to considering the light transport in a vacuum or participating media, the only thing we have to do

is: To choose the parameters within the SLTE that describe the absorption, emission, or scattering behavior of light in an appropriate way.

So, in the next two section, we will first consider some of the special cases that the SLTE subsumes. Afterwards, we will devote to the stationary light transport equation valid in a vacuum, the *SLTEV*, in computer graphics denoted as the *rendering equation*, [98, Kajiya 1986]. It describes the equilibrium distribution of radiance in a scene under vacuum conditions. We also show that the SLTEV can be expressed exclusively in terms of exitant or incident quantities instead of the above introduced mixed-form of incident and exitant radiance. Via a change in the integration measure then we transform the spherical integral within the SLTEV into a surface integral. This will be of great advantage if we discuss direct illumination at a surface point and opens up several different approaches for solving the global illumination problem.

Section 4.4.1

Section 4.4.2

4.4.1 THE STATIONARY LIGHT TRANSPORT EQUATION IN PARTICIPATING MEDIA

Recall, the stationary light transport equation has the form

SLTE (297)

$$\begin{aligned} & L_i(\mathbf{x}, \omega_i) \\ = \quad & \beta(\mathbf{s} \to \mathbf{x}) \left(\epsilon_b(\mathbf{s}, \omega_o) + \int_{S^2(\mathbf{s})} \kappa_b(\mathbf{s}, \omega_i' \to \omega_o)\, L_i(\mathbf{s}, \omega_i')\, d\sigma_{\mathbf{s}}(\omega_i') \right) + \\ & \int_{[0, \|\mathbf{s}-\mathbf{x}\|]} \beta(\mathbf{x}' \to \mathbf{x}) \left(\epsilon(\mathbf{x}', \omega_o) + \int_{S^2(\mathbf{x}')} \kappa(\mathbf{x}', \omega_i' \to \omega_o)\, L_i(\mathbf{x}', \omega_i')\, d\sigma_{\mathbf{x}'}(\omega_i') \right) d\mu(\alpha). \end{aligned} \tag{4.381}$$

Replacing the theoretically constructed scattering kernels in this equation, thus the surface scattering kernel κ_b at point $\mathbf{s}$ by the BSDF f_s, i.e.:

$$k_b(\mathbf{x}, \omega_i' \to \omega_o) \quad = \quad f_s(\mathbf{x}, \omega_i' \to \omega_o) \tag{4.382}$$

and the volume scattering kernel κ at $\mathbf{x}'$ by the phase function p multiplied by the scattering coefficient σ_s thus,

$$k(\mathbf{x}, \omega_i' \to \omega_o) \quad = \quad \sigma_s(\mathbf{x}, \omega_o)\, p(\mathbf{x}, -\omega_i' \to \omega_o), \tag{4.383}$$

then the stationary light transport equation can be represented in a form which is much more useful for the purposes of computer graphics, namely the *stationary light transport equation in participating media*.

DEFINITION 4.37 (The Stationary Light Transport Equation in Participating Media, SLTE)
Let $\mathbf{s}$ *be a point on a surface* $M \in \partial\mathcal{V}$, $\mathbf{x}, \mathbf{x}' \in \mathcal{V}^o$ *are inner points of a participating* ∂V (41)
medium, $L_i(\mathbf{s}, \omega_i')$ *and* $L_i(\mathbf{x}', \omega_i')$ *describe the incident radiance at surface point* $\mathbf{s}$, V° (41)
respectively at point $\mathbf{x}' = \mathbf{x} + \alpha\omega_i$ *within a participating medium coming from direction*
β (292) ω_i', *attenuated by the path absorption function* β. *Then, the fundamental equation of light transport that governs the behavior of light in a medium that absorbs, emits, and scatters is called the* stationary light transport equation in participating media, *also briefly denoted as* SLTE. *It is given by*

$$\begin{aligned} & L_i(\mathbf{x}, \omega_i) \\ &= \beta(\mathbf{s} \to \mathbf{x}) \left(L_e(\mathbf{s}, \omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i' \to \omega_o)\, L_i(\mathbf{s}, \omega_i')\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i') \right) + \\ & \quad \int_{[0, \|\mathbf{s}-\mathbf{x}\|]} \beta(\mathbf{x}' \to \mathbf{x}) \left(L_e(\mathbf{x}', \omega_o) + \right. \\ & \qquad \left. \sigma_s(\mathbf{x}', \omega_o) \int_{S^2(\mathbf{x}')} p(\mathbf{x}', -\omega_i' \to \omega_o)\, L_i(\mathbf{x}', \omega_i')\, d\sigma_{\mathbf{x}'}(\omega_i') \right) d\mu(\alpha), \end{aligned} \tag{4.384}$$

see Figure 4.60

REMARK 4.43 *Note, the different integration measures used in the SLTE. While the integration measure within the first integral is with respect to the projected solid angle-*
BSDF (371) *—the BSDF operates on irradiance instead of radiance—we integrate the second term with respect to the solid angle measure. Note also the reverse incident direction due*
Phase Function (376) *to the definition of the direction vectors occurring in the phase function.*

Scattering Equation (374) REMARK 4.44 *Using the definition of the scattering equation in the first term of Equation (4.384) of Definition 4.37, then we get the commonly used form of the light transport in participating media, see [152, Pauly 1999], namely:*

$$\begin{aligned} & L_i(\mathbf{x}, \omega_i) \\ &= \beta(\mathbf{s} \to \mathbf{x})\, L_o(\mathbf{s}, \omega_o) \\ & \quad \int_{[0, \|\mathbf{s}-\mathbf{x}\|]} \beta(\mathbf{x}' \to \mathbf{x}) \left(L_e(\mathbf{x}', \omega_o) + \right. \\ & \qquad \left. \sigma_s(\mathbf{x}', \omega_o) \int_{S^2(\mathbf{x}')} p(\mathbf{x}', -\omega_i' \to \omega_o)\, L_i(\mathbf{x}', \omega_i')\, d\sigma_{\mathbf{x}'}(\omega_i') \right) d\mu(\alpha). \end{aligned} \tag{4.385}$$

Now, finding a solution to the SLTE—in which every point of the observed volume may communicate with every point on a boundary of an object as well as any point within a participating medium—is a very costly and time consuming task. But if we limit the effects considered by simplifying the corresponding equation and lowering its computational costs,

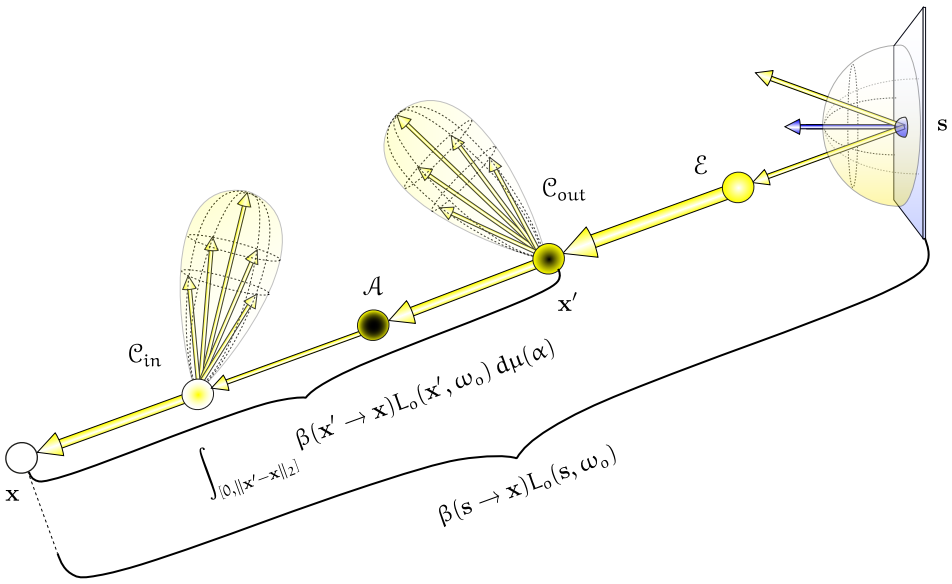


FIGURE 4.60: THE STATIONARY LIGHT TRANSPORT EQUATION IN PARTICIPATING MEDIA. The radiance incident at $\mathbf{x}$ from direction ω_i is the sum of the reduced radiance from the nearest visible surface point $\mathbf{s}$ and the reduced, accumulated, and scattered radiance L_o along the line connecting $\mathbf{x}$ and $\mathbf{s}$.

we can derive procedures for simulating different light effects that yield respectable results with a reasonable amount of effort. For that purpose, let us observe the following cases:

EXAMPLE 4.11 (Non-scattering Media) *The abstraction of scattering in the integral form of the stationary light transport equation is a preferred and promising method for representing participating media such as so-called* particle clouds.

Setting the scattering coefficient $\sigma_s \equiv 0$ *implies that the gain function* Q_o *reduces to the volume emission at inner points of the medium, that is, the incident radiance* $L_i(\mathbf{x}, \omega_i)$ *can be written as:*

Gain Function (289)

$$L_i(\mathbf{x}, \omega_i) = \beta(\mathbf{s} \to \mathbf{x}) \left(L_e(\mathbf{s}, \omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i' \to \omega_o)\, L_i(\mathbf{s}, \omega_i')\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i') \right) + \int_{[0, \|\mathbf{s}-\mathbf{x}\|]} \beta(\mathbf{x}' \to \mathbf{x})\, L_e(\mathbf{x}', \omega_o) d\mu(\alpha) \quad (4.386)$$

with $\omega_o = -\omega_i$. *This implies that the incident radiance at point* $\mathbf{x}$ *from direction* ω_i *can easily be calculated by integration over the emissions along all points lying on*

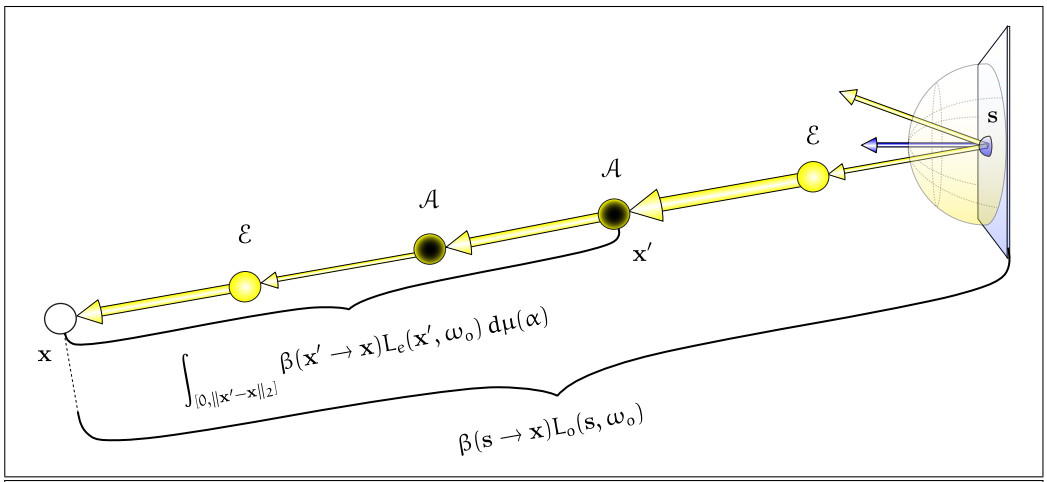


FIGURE 4.61: THE SLTE IN NON-SCATTERING MEDIA. On its way from surface point s to the volumetric point $x \in \mathcal{V}^o$, a beam of photons is only subject to absorption or emission events. There are no scattering events between the surface points s and x.

the ray from x to s, reduced by any potential occurrence of thermal absorption plus the attenuated radiance coming from the surface, see Figure 4.61.

EXAMPLE 4.12 (Non-absorbing and Non-emitting Media) *Let us now consider the light transport in non-absorbing and non-emitting media. Setting the emitted radiance $L_e \equiv 0$ implies, that the gain function Q_o reduces to the radiance in-scattered at inner points of the medium, that is, with $\beta \equiv 1$ the incident radiance $L_i(x, \omega_i)$ can be written as:*

$$L_i(x, \omega_i) = L_e(s, \omega_o) + \int_{S^2(s)} f_s(s, \omega_i' \to \omega_o)\, L_i(s, \omega_i')\, d\sigma_s^{\perp}(\omega_i') + \int_{[0, \|s-x\|]} \sigma_s(x', \omega_o) \int_{S^2(x')} p(x', -\omega_i' \to \omega_o)\, L_i(x', \omega_i')\, d\sigma_{x'}(\omega_i') d\mu(\alpha) \tag{4.387}$$

with $\omega_o = -\omega_i$, see Figure 4.62.

EXAMPLE 4.13 (Non-absorbing and Non-scattering Media) *Let us assume we have a medium, which does not scatter and does not absorb. Setting the scattering coefficient $\sigma_s \equiv 0$ implies, that the gain function Q_o reduces to the volume emission at inner points of the medium, that is, with $\beta \equiv 1$, due to the condition of a non-*

absorbing medium, we get:

$$L_i(\mathbf{x},\omega_i) = L_e(\mathbf{s},\omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s},\omega_i' \to \omega_o)\, L_i(\mathbf{s},\omega_i')\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i') + \int_{[0,\|\mathbf{s}-\mathbf{x}\|]} L_e(\mathbf{x}',\omega_o)\, d\mu(\alpha) \tag{4.388}$$

with $\omega_o = -\omega_i$, *see Figure 4.63.*

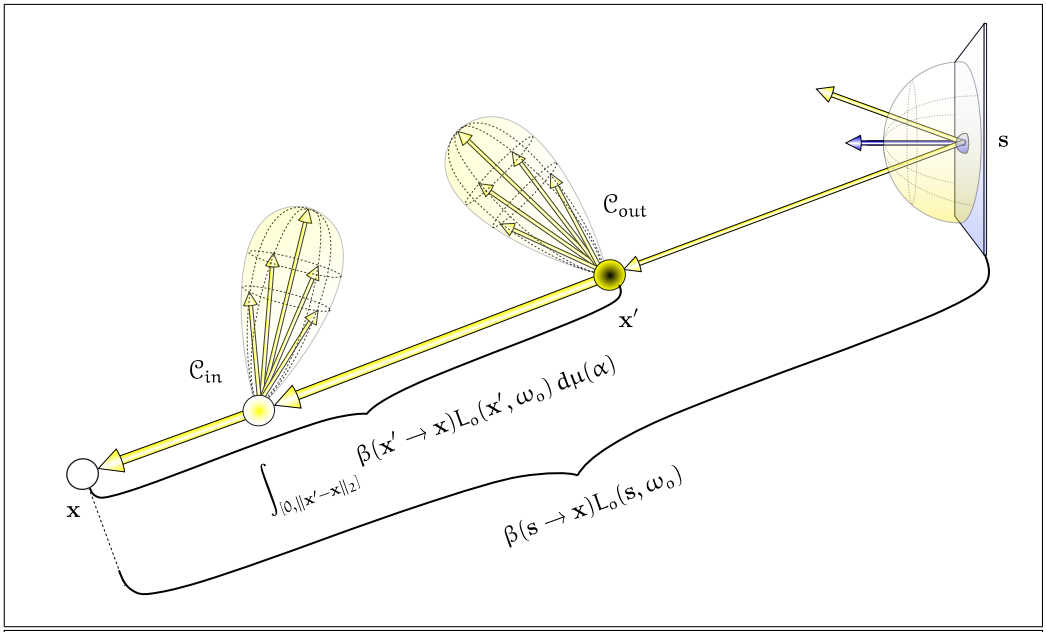


FIGURE 4.62: THE SLTE IN NON-ABSORBING AND NON-EMITTING MEDIA. On its way from surface point $\mathbf{s}$ to the volumetric point $\mathbf{x} \in \mathcal{V}^o$, a beam of photons is only subject to scattering events. There are neither absorption nor emission events between the surface points $\mathbf{s}$ and $\mathbf{x}$.

EXAMPLE 4.14 (Non-emitting and Non-scattering Media) *Finally, let us discuss the light transport in non-emitting and non-scattering media, thus neglecting the scattering coefficient,* σ_s, *as well as the volumetric emission function,* L_e. *The choice of* $\sigma_s \equiv L_e \equiv 0$ *leads to the gain function* $Q_o \equiv 0$. *Under these conditions, the SLTE reduces to the interaction of light at the boundaries of the media, that is, the SLTE can be written as:*

$$L_i(\mathbf{x},\omega_i) = \beta(\mathbf{s} \to \mathbf{x}) \left(L_e(\mathbf{s},\omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s},\omega_i' \to \omega_o)\, L_i(\mathbf{s},\omega_i')\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i') \right) \tag{4.389}$$

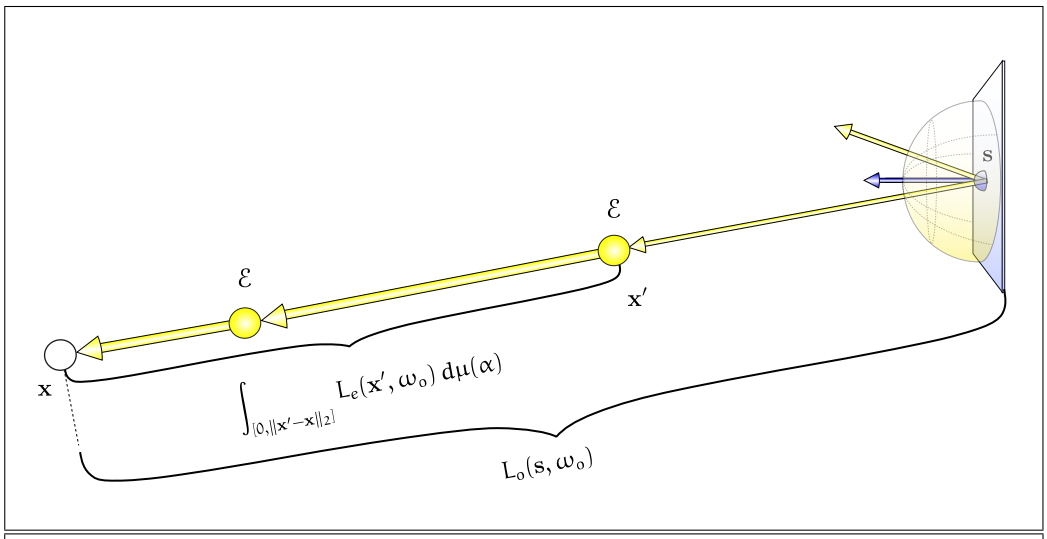


FIGURE 4.63: THE SLTE IN NON-SCATTERING AND NON-ABSORBING MEDIA. On its way from surface point s to the volumetric point $x \in \mathcal{V}^o$, a beam of photons is only subject to emission processes. There are no scattering or absorption events between the surface points s and x.

with $\omega_o = -\omega_i$, *see Figure 4.64.*

4.4.2 THE STATIONARY LIGHT TRANSPORT EQUATION IN A VACUUM

Based on the SLTE from (4.384), we will now derive two different but equivalent formulations of the light transport in a vacuum: the *stationary light transport equation* under vacuum conditions in its *spherical* as well as its 3-*point form*.

THE SPHERICAL FORM OF THE STATIONARY LIGHT TRANSPORT EQUATION IN A VACUUM. Let us consider the light transport under vacuum conditions in a closed scene composed of a finite set $\partial\mathcal{V}$ of 2-dimensional surfaces. Then, we can ignore the effects of absorption and
$\partial\mathcal{V}$ (41) scattering as there is no medium involved. Thus, the general equation characterizing light transport is reduced to the calculation of radiance at the boundaries of object surfaces, that is to say, to the formulation of the boundary conditions.

DEFINITION 4.38 (The Spherical Form of the SLTE in a Vacuum, SLTEV) *Let* $L_o(s, \omega_o)$ *be the exitant radiance at surface point* s *in direction* ω_o, $L_e(s, \omega_o)$ *be the radiance*

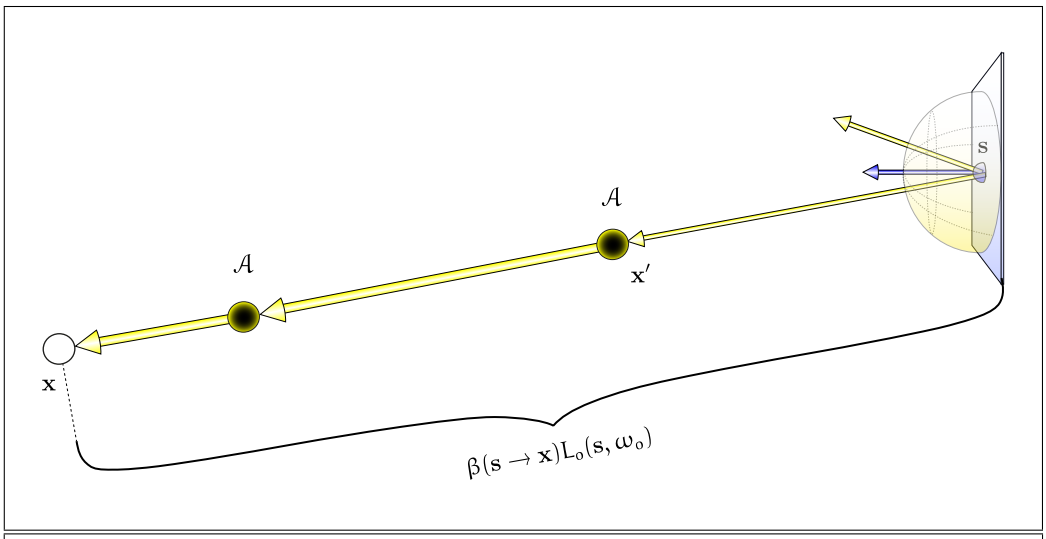


FIGURE 4.64: THE SLTE IN NON-SCATTERING AND NON-EMITTING MEDIA. On its way from surface point $\mathbf{s}$ to the volumetric point $\mathbf{x} \in \mathcal{V}^o$, a beam of photons is only subject to absorbing processes. There are no scattering or emission events between the surface points $\mathbf{s}$ and $\mathbf{x}$.

emitted at $\mathbf{s}$ *in direction* ω_o*, and* $L_i(\mathbf{s}, \omega_i)$ *denotes the incident radiance at point* $\mathbf{s}$ *coming from directions* $\omega_i \in S^2(\mathbf{s})$*. Then, the equation*

$$L_o(\mathbf{s}, \omega_o) = L_e(\mathbf{s}, \omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i), \tag{4.390}$$

that describes the scattering behavior of light at an object surface in a vacuum via the BSDF f_s *is called the* stationary light transport equation in a vacuum*, also briefly denoted as* SLTE in a vacuum*, or* SLTEV*, see Figure 4.65.*

REMARK 4.45 (The Hemispherical Form of the SLTE in a Vacuum) *Under the condition that all object surfaces in Equation (4.390) are opaque, it suffices to integrate in Equation (4.390) over the upper hemisphere instead of the entire unit sphere. Then, the SLTE in a vacuum can also be written as:*

$$L_o(\mathbf{s}, \omega_o) = L_e(\mathbf{s}, \omega_o) + \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i), \tag{4.391}$$

see Figure 4.66.

REMARK 4.46 *The SLTEV—in its spherical as well as in its hemispherical form—can also be seen as a result of the conservation of energy from physics. Thus, energy*

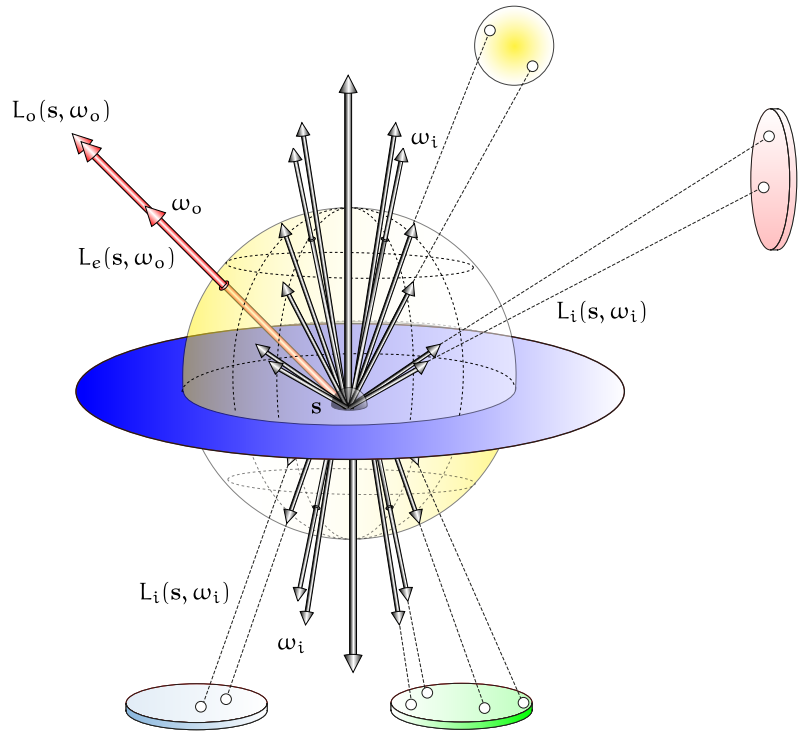


FIGURE 4.65: THE SPHERICAL FORM OF THE SLTE IN A VACUUM. The exitant radiance at point s in direction ω_o equals the self-emitted exitant radiance $L_e(s, \omega_o)$ plus any incident radiance from the illuminating sphere that is reflected at s in direction ω_o.

balance on a surface means that exitant radiance must be equal to emitted radiance plus the fraction of incident radiance which is scattered. That is, the SLTEV must be composed of two expressions: a self-emitted term L_e, describing the energy that comes from light sources, and a term which looks like the scattering equation from Relation (4.311). That is, the SLTEV subsumes the quantity of radiance emitted from point s on a surface A in direction ω_o as a sum of an emission term L_e related to point s in direction ω_o together with the BSDF f_s, which gives information about radiance incident from all directions and emitted in direction ω_o.

REMARK 4.47 (The Rendering Equation, REQ) *In the literature, the Equations (4.390) and (4.391) are also often denoted as the* rendering equation *in spherical as well as in hemispherical form. Due to [98, Kajiya 1986], the rendering equation is written*

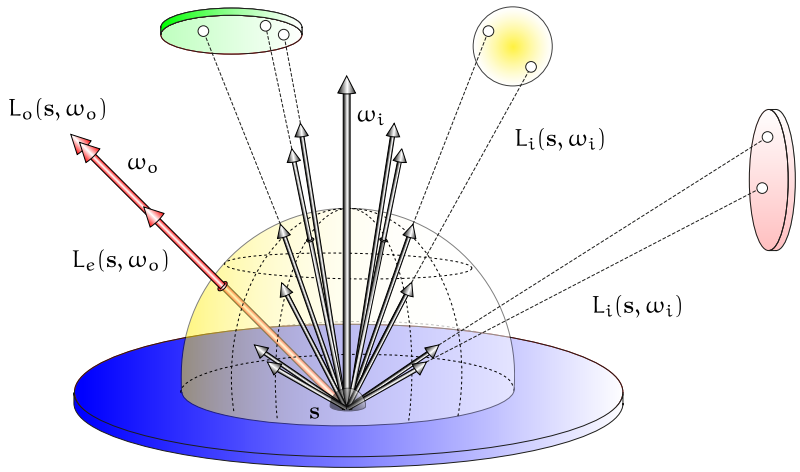


FIGURE 4.66: THE HEMISPHERICAL FORM OF THE SLTE IN A VACUUM. The exitant radiance at point s in direction ω_o equals the self-emitted exitant radiance $L_e(s, \omega_o)$ plus any incident radiance from the illuminating hemisphere that is reflected at s in direction ω_o.

In its original form as:

$$I(x, x') = g(x, x') \left(\epsilon(x, x') + \int_S \rho(x, x', x'') \, I(x, x'') \, dx'' \right) \tag{4.392}$$

where

$I(x, x')$	*is related to the intensity of light passing from point* x *to point* x'
$g(x, x')$	*is a geometry term*
$\epsilon(x, x')$	*is related to the intensity of light emitted from point* x' *to point* x
$\rho(x, x', x'')$	*is related to the intensity of light scattered from* x'' *to* x *by a patch of surface* x'.

THE 3-POINT FORM OF THE STATIONARY LIGHT TRANSPORT EQUATION IN A VACUUM. Instead of projecting the entire scene to the unit sphere, and then integrating with respect to projected solid angle, it is also possible to use the boundaries of scene objects as integration domains. That is, we integrate over the surfaces in the scene. If we go this way,

then, as we have seen in Section 2.3, we obtain the scattering equation as well as the reflection equation in a form, that satisfies the requirements to a Fredholm integral equation of the 2^{nd} kind. Based on this new integration strategy, we get another mathematically equivalent formulation for the light transport in a vacuum, the so-called 3-*point forms* of the SLTEV. It opens up several different approaches for solving the global illumination problem.

Let us recall Example 2.51, where we performed the process of transforming the $\sigma^{\perp}$ (89) projected solid angle measure $\sigma^{\perp}$ to the Lebesgue area measure μ^2. Adapted to this μ^2 (82) derivation, we can define the 3-*point form of the SLTE in a vacuum* as follows:

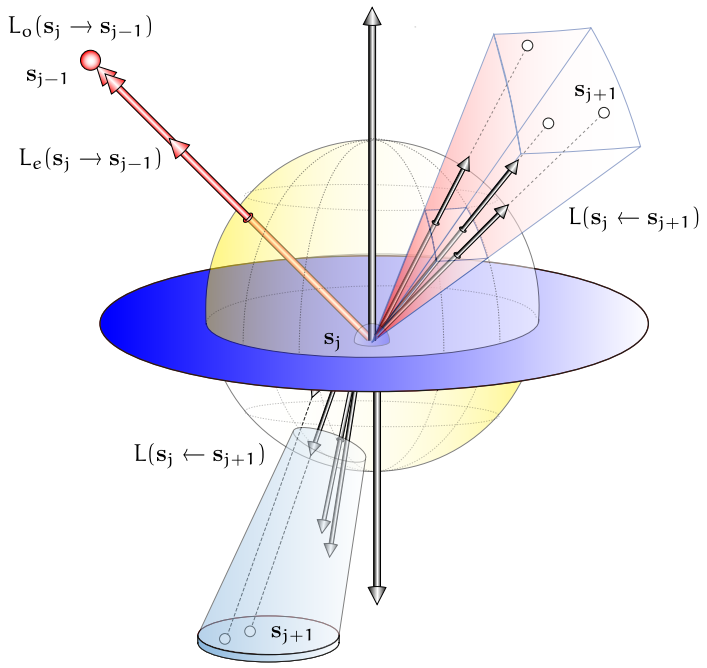


FIGURE 4.67: THE 3-POINT FORM OF THE SLTEV. The exitant radiance at point s_j in direction towards point s_{j-1} equals the self-emitted exitant radiance $L_e(s_j \to s_{j+1})$ from s_j plus any incident radiance at point s_j that comes from all points s_{j+1} at surfaces of the scene that are visible from s_j.

Exitant Function (48) DEFINITION 4.39 (The 3-point Form of the STLE in a Vacuum) *Let us express the exitant*

radiance from point $\mathbf{s}_j$ *outgoing in direction* ω_o^j *towards point* $\mathbf{s}_{j-1}$ *by*

$$L(\mathbf{s}_j \rightarrow \mathbf{s}_{j-1}) \equiv L_o(\mathbf{s}_j, \omega_o^j), \tag{4.393}$$

where $\mathbf{s}_{j-1} = \gamma(\mathbf{s}_j, \omega_o^j)$*, see Figure 4.67. Let us furthermore formulate the BSDF at point* $\mathbf{s}_j$ *by*

γ (47)
BSDF (371)

$$f_s(\mathbf{s}_{j+1} \rightarrow \mathbf{s}_j \rightarrow \mathbf{s}_{j-1}) \equiv f_s(\mathbf{s}_j, \omega_i^j \rightarrow \omega_o^j) \tag{4.394}$$

with $\mathbf{s}_{j+1} = \gamma(\mathbf{s}_j, \omega_i^j)$*, and the* geometry term, $\mathcal{G}$*, by*

$\mathcal{G}$ (129)

$$\mathcal{G}(\mathbf{s}_{j+1} \leftrightarrow \mathbf{s}_j) \stackrel{def}{=} \mathcal{V}(\mathbf{s}_{j+1} \leftrightarrow \mathbf{s}_j) \frac{\left|\cos\theta_o^{j+1} \cos\theta_i^j\right|}{\|\mathbf{s}_{j+1} - \mathbf{s}_j\|^2} \, d\mu^2(\mathbf{s}_{j+1}), \tag{4.395}$$

with the visibility function $\mathcal{V}$*. Now, the SLTEV can be written as an integral over all surfaces* $\partial\mathcal{V}$*, thus:*

Visibility Function (45)
$\partial\mathcal{V}$ (41)

$$\begin{aligned} L(\mathbf{s}_j \rightarrow \mathbf{s}_{j-1}) \quad &= \quad L_e(\mathbf{s}_j \rightarrow \mathbf{s}_{j-1}) + \\ &\int_{\partial\mathcal{V}} f_s(\mathbf{s}_{j+1} \rightarrow \mathbf{s}_j \rightarrow \mathbf{s}_{j-1}) \, L(\mathbf{s}_j \leftarrow \mathbf{s}_{j+1}) \, \mathcal{G}(\mathbf{s}_{j+1} \leftrightarrow \mathbf{s}_j) \, d\mu^2(\mathbf{s}_{j+1}). \end{aligned} \tag{4.396}$$

This Equation is denoted as the 3-point form of the SLTEV*, see Figure 4.67.*

4.4.2.1 FORMULATIONS OF THE SLTEV BASED ON EXITANT AND INCIDENT RADIANCE

A closer look at the different forms of the SLTEV from above shows that the outgoing radiance as well as the incident radiance appear within the corresponding equations. Thus, in the spherical form of the SLTEV, the exitant radiance from point $\mathbf{s}$ in direction ω_o is composed of light emitted from $\mathbf{s}$, and of light that arrives from all direction ω_i around $\mathbf{s}$. We have a similar situation for the 3-point form of the SLTEV. Here the exitant radiance from point $\mathbf{s}_j$ towards point $\mathbf{s}_{j-1}$ is composed of the quantity of light emitted from $\mathbf{s}_j$ in direction to $\mathbf{s}_{j-1}$, and light that comes from all points $\mathbf{s}_{j+1}$ on surfaces which are visible from $\mathbf{s}_j$.

At this point, we can in turn utilize the principle of radiance invariance along a light ray in a vacuum. Due to this fact, the incident radiance $L_i(\mathbf{s}, \omega_i)$ at surface point $\mathbf{s}$ can be replaced by the exitant radiance $L_o(\mathbf{s}', -\omega_i)$, where $\mathbf{s}' = \gamma(\mathbf{s}, \omega_i)$. Applied to the spherical form of the SLTE in a vacuum, this results in the *spherical form of the SLTEV on the basis of exitant radiance.*

Radiance Invariance (253)

DEFINITION 4.40 (The Spherical Form of the SLTEV Based on Exitant Radiance) *Let* $\mathbf{s}$ *and* $\mathbf{s}' = \gamma(\mathbf{s}, \omega_i)$ *be two mutually visible points on different surfaces from* ∂V*, see Figure*

4.68. Then, the spherical form of the SLTEV expressed in terms of exitant radiance *is given by*

$$\mathrm{L_o}(\mathbf{s},\omega_o) = \mathrm{L_e}(\mathbf{s},\omega_o) + \int_{S^2(\mathbf{s})} \mathrm{f_s}(\mathbf{s},\omega_i \to \omega_o)\,\mathrm{L_o}(\gamma(\mathbf{s},\omega_i),-\omega_i)\,\mathrm{d}\sigma^{\perp}_{\mathbf{s}}(\omega_i) \quad (4.397)$$

$$= \mathrm{L_e}(\mathbf{s},\omega_o) + \int_{S^2(\mathbf{s})} \mathrm{f_s}(\mathbf{s},\omega_i \to \omega_o)\,\mathrm{L_o}(\mathbf{s}',-\omega_i)\,\mathrm{d}\sigma^{\perp}_{\mathbf{s}}(\omega_i). \quad (4.398)$$

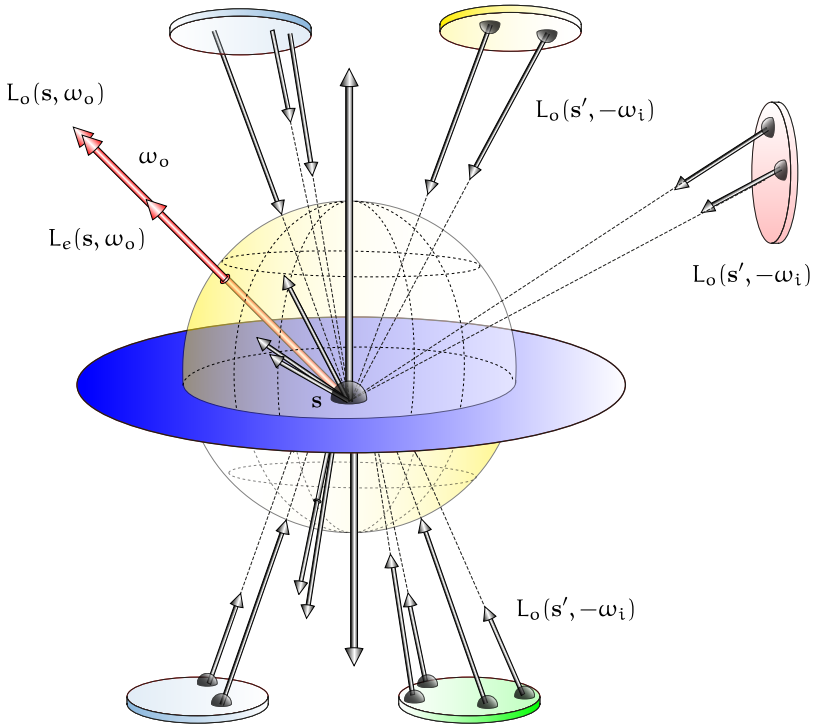


FIGURE 4.68: THE SPHERICAL FORM OF THE SLTEV BASED ON EXITANT RADIANCE. The exitant radiance at point $\mathbf{s}$ in direction ω_o equals the self-emitted exitant radiance $\mathrm{L_e}(\mathbf{s},\omega_o)$ plus any exitant radiance from the illuminating sphere that is reflected at $\mathbf{s}$ in direction ω_o.

This can be interpreted in such a way that the radiance outgoing from point $\mathbf{s}$ in direction ω_o consist of a self-emitted contribution from $\mathbf{s}$ in direction ω_o, as well as the amount of exitant radiance coming from points $\gamma(\mathbf{s},\omega_i)$ reachable about directions $\omega_i \in S^2(\mathbf{s})$.

Radiance Invariance (253)

Applying the principle of radiance invariance also on the 3-point form of the SLTEV

from (4.396), we then obtain the corresponding *3-point form of the SLTEV based on exitant radiance*, see Figure 4.69.

DEFINITION 4.41 (The 3-point form of the SLTEV Based on Exitant Radiance) *Suppose* $\mathbf{s}_{j-1}$ *is a point lying on a surface* $M \in \partial\mathcal{V}$ *visible from point* $\mathbf{s}_j$ *on an other surface. Let us further assume that* $\mathbf{s}_{j+1}$ *is a point of a surface also visible from* $\mathbf{s}_j$*, see Figure 4.69. Then, the* 3-point form of the SLTEV based on exitant radiance *is defined by*

$$L(\mathbf{s}_j \to \mathbf{s}_{j-1}) = L_e(\mathbf{s}_j \to \mathbf{s}_{j-1}) + \int_{\partial\mathcal{V}} f_s(\mathbf{s}_{j+1} \to \mathbf{s}_j \to \mathbf{s}_{j-1})\, L(\mathbf{s}_{j+1} \to \mathbf{s}_j)\, \mathcal{G}(\mathbf{s}_{j+1} \leftrightarrow \mathbf{s}_j)\, d\mu^2(\mathbf{s}_{j+1}). \tag{4.399}$$

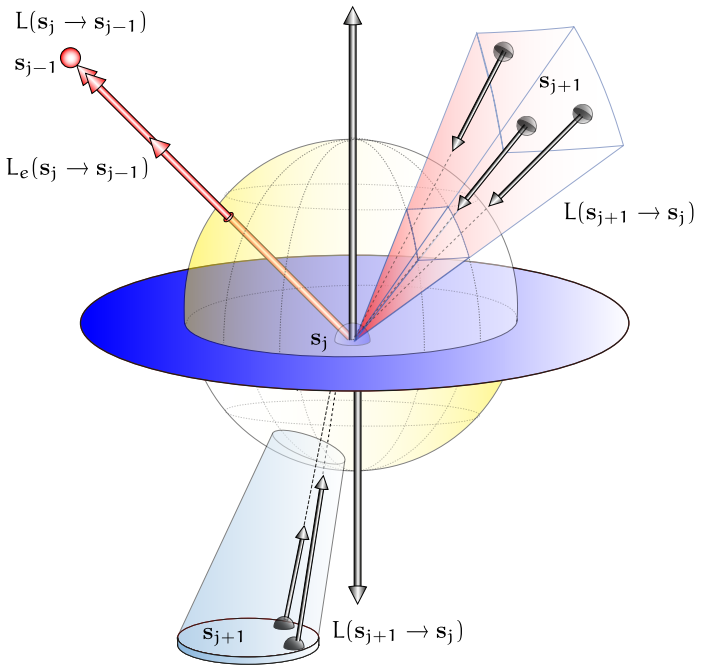


FIGURE 4.69: THE 3-POINT FORM OF THE SLTEV BASED ON EXITANT RADIANCE. The exitant radiance at point $\mathbf{s}_j$ equals the self-emitted incident radiance $L_e(\mathbf{s}_j \to \mathbf{s}_{j-1})$ plus any exitant radiance coming from visible points $\mathbf{s}_{j+1}$ that is reflected at point $\mathbf{s}_j$ in direction to $\mathbf{s}_{j-1}$.

DEFINITION 4.42 (The Spherical Form of the SLTEV Based on Incident Radiance) *Let* $\mathbf{s}$

and $\mathbf{s}' = \gamma(\mathbf{s}, \omega_i)$ *be two mutually visible points on different surfaces from* ∂V*, see Figure 4.70. Then, the* spherical form of the SLTEV expressed in terms of incident radiance *is given by*

$$\begin{aligned} L_i(\mathbf{s}, \omega_i) &= L_e(\mathbf{s}, \omega_i) + \int_{S^2(\gamma(\mathbf{s},\omega_i))} f_s(\gamma(\mathbf{s}, \omega_i), \omega_i' \to -\omega_i)\, L_i(\gamma(\mathbf{s}, \omega_i), \omega_i')\, d\sigma^{\perp}_{\gamma(\mathbf{s},\omega_i)}(\omega_i') \quad (4.400) \\ &= L_e(\mathbf{s}, \omega_i) + \int_{S^2(\mathbf{s}')} f_s(\mathbf{s}', \omega_i' \to -\omega_i)\, L_i(\mathbf{s}', \omega_i')\, d\sigma^{\perp}_{\mathbf{s}'}(\omega_i') \quad (4.401) \end{aligned}$$

This can be interpreted in such a way that the radiance incident at point $\mathbf{s}$ from direction ω_i is composed of an emission term in direction ω_i, as well as the amount of incident radiance at points $\mathbf{s}'$ visible from $\mathbf{s}$ in directions $\omega_i \in S^2$, which is reflected at surfaces in direction to $\mathbf{s}$.

When we keep this little game going the other way around now, then we obtain the *SLTEV based on incident radiance in 3-point form*, thus:

DEFINITION 4.43 (The 3-point form of the SLTEV Based on Incident Radiance) *Suppose* $\mathbf{s}_{j+1}$ *and* $\mathbf{s}_{j-1}$ *are points on two different surfaces of* ∂V *visible from a given point* $\mathbf{s}_j$*, see Figure 4.71. The* 3-point form of the SLTEV based on incident radiance *is defined by*

$$\begin{aligned} L(\mathbf{s}_{j-1} \leftarrow \mathbf{s}_j) &= L_e(\mathbf{s}_{j-1} \leftarrow \mathbf{s}_j) + \int_{\partial V} f_s(\mathbf{s}_{j-1} \leftarrow \mathbf{s}_j \leftarrow \mathbf{s}_{j+1})\, L_i(\mathbf{s}_j \leftarrow \mathbf{s}_{j+1})\, G(\mathbf{s}_j \leftrightarrow \mathbf{s}_{j+1})\, d\mu^2(\mathbf{s}_{j+1}). \quad (4.402) \end{aligned}$$

REMARK 4.48 (The Hemispherical Forms of the Light Transport Equation in a Vacuum) *It should be clear that when rendering scenes consisting only of opaque surfaces, it makes sense to replace the scattering equation in the above spherical forms of the SLTEV by the reflectance equation. In this cases the integration goes over the upper hemisphere* $\mathcal{H}^2_i$ *instead of* S^2*.*

REMARK 4.49 *The above introduced two approaches to consider the scattering equation in the different forms of the SLTEV, once as an integral over the areas of the scene objects and once as an integral over the unit sphere, also lead to two different strategies for solving the SLTEV in a rendering procedure. Thus, we will see that the scattering equation in the spherical forms of the SLTEV are solved by Monte Carlo methods via sampling a number of directions from a distribution of directions on the unit sphere and casting rays to evaluate the integrand. In contrast to this procedure, the scattering equation in the 3-point forms of the SLTEV is solved by Monte Carlo methods, which choose a number of points on surfaces according to distributions over the surface areas, and making use of the visibility function to compute the coupling of*

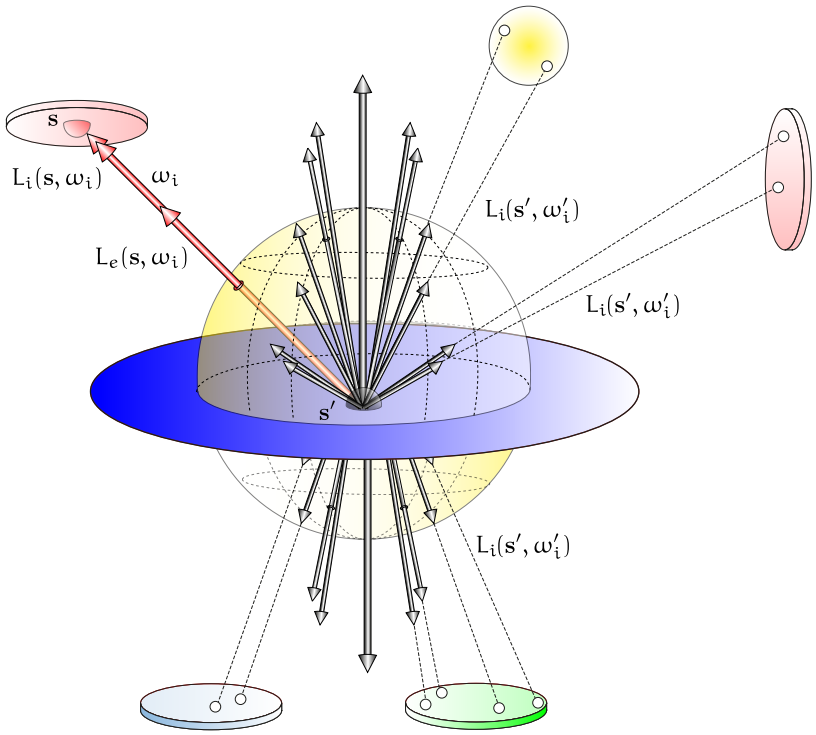


FIGURE 4.70: THE SPHERICAL FORM OF THE SLTEV BASED ON INCIDENT RADIANCE.
The incident radiance at point s in direction ω_o equals the self-emitted exitant radiance $L_e(s, \omega_o)$ plus any incident radiance from the illuminating sphere that is reflected at s in direction ω_o.

the points. As we will see, in many Monte Carlo rendering algorithms the spherical forms of the SLTEV are used for generating paths in a scene to be rendered. While this is a convenient method of direction sampling, the 3-point forms of the SLTEV underlies primarily those methods based on finite element techniques. Here, in particular radiosity procedures and path-integral formulations, as well as the sampling of surface light sources are prime usage examples.

4.4.2.2 DIRECT AND INDIRECT ILLUMINATION FORMULATION OF THE SLTEV

Let us consider once more the stationary light transport equation in a vacuum in spherical form. With respect to surface point s it can be written as sum of the emitted radiance at

SLTEV Spherical Form (399)

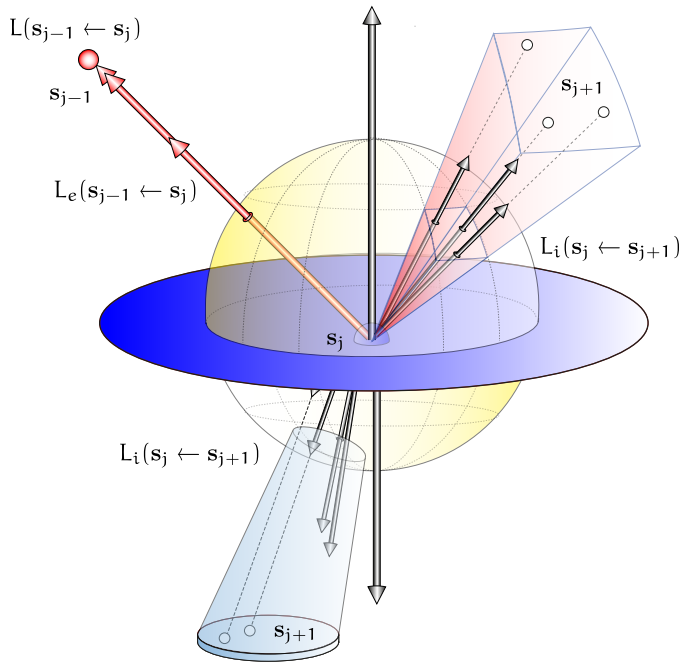


FIGURE 4.71: THE 3-POINT FORM OF THE SLTEV BASED ON INCIDENT RADIANCE. The incident radiance at point s_{j-1} equals the self-emitted incident radiance $L_e(s_j \to s_{j-1})$ from s_j plus any incident radiance at s_j coming from all points s_{j+1} at surfaces within the the scene that are visible from s_j.

point s in direction ω_o and the scattering equation from (4.311), thus,

$$L_o(s, \omega_o) = L_e(s, \omega_o) + \int_{S^2(s)} f_s(s, \omega_i \to \omega_o)\, L_i(s, \omega_i)\, d\sigma_s^{\perp}(\omega_i). \tag{4.403}$$

Obviously, the integration domain of the SLTEV can now be split into two disjoint sets: The projection of all regions of light sources, visible from the center of the unit sphere, onto the unit sphere, thus the set $\text{☼}^{\perp}$, and the complement of this set, that is, $\overline{\text{☼}^{\perp}} = S^2 \setminus \text{☼}^{\perp}$, see Figure 4.72. Using this stratification of the integration domain of the SLTEV, then we get due to the linearity property of the Lebesgue integral with respect to the integration domain the following decomposition of the exitant radiance at point s

Lebesgue Integral (105)

in direction ω_o:

$$L_o(s,\omega_o) = L_e(s,\omega_o) + \int_{☼^\perp \cup \overline{☼^\perp}} f_s(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^\perp(\omega_i) \tag{4.404}$$

$$= L_e(s,\omega_o) + \int_{☼^\perp} f_s(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^\perp(\omega_i) + \int_{\overline{☼^\perp}} f_s(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^\perp(\omega_i). \tag{4.405}$$

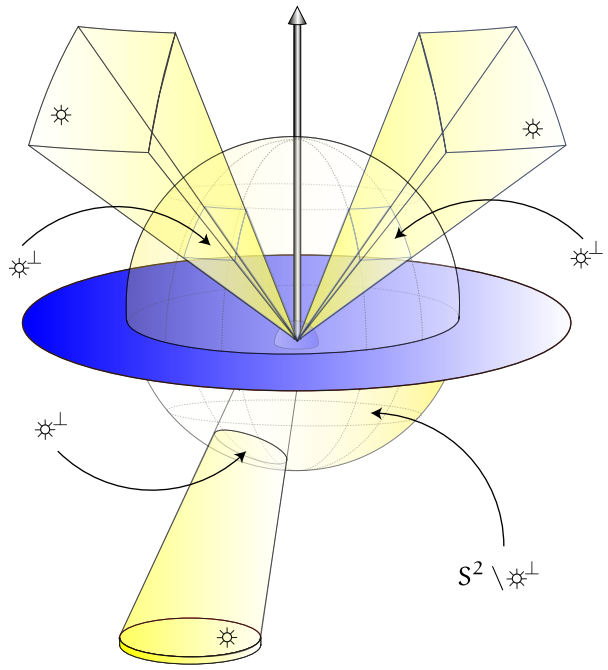


FIGURE 4.72: THE PROJECTION OF LIGHT SOURCES ONTO THE UNIT SPHERE S^2. Stratification of the unit sphere about point s in the strata $☼^\perp$ and $S^2(s) \setminus ☼^\perp$ for computing the direct and indirect illumination on an opaque surface. As you can easily see in the figure, the projection of a light source onto the unit sphere can be a more or less complicated subset of points, in particular if the projections are not disjoint.

Since the incident radiance $L_i(s,\omega_i)$ in the first integral comes from points $s' = \gamma(s,\omega_i)$ at light sources, we can also express $L_i(s,\omega_i)$, due to the principle of radiance

(47) Radiance Invariance (253)

invariance in a vacuum, by the emitted radiance at these points, that is, the SLTEV takes on the form

$$\begin{aligned} L_o(s,\omega_o) &= L_e(s,\omega_o) + \\ &\underbrace{\int_{\Omega^\perp} f_s(s,\omega_i \to \omega_o)\, L_e(\gamma(s,\omega_i),-\omega_i)\, d\sigma_s^\perp(\omega_i)}_{L^{\leftarrow}(s,\omega_o)} + \\ &\underbrace{\int_{\Omega^\perp} f_s(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^\perp(\omega_i)}_{L^{\Leftarrow}(s,\omega_o)}. \end{aligned} \tag{4.406}$$

This means, that the light reflected at point s in direction ω_o is a composition of an emission term, $L_e(s,\omega_o)$, a *direct illumination* component, $L^{\leftarrow}(s,\omega_o)$, and an *indirect illumination* component, $L^{\Leftarrow}(s,\omega_o)$. Mathematically, this can be expressed as follows:

$$L_o(s,\omega_o) = L_e(s,\omega_o) + L^{\leftarrow}(s,\omega_o) + L^{\Leftarrow}(s,\omega_o). \tag{4.407}$$

Here, $L^{\leftarrow}(s,\omega_o)$ can be interpreted as the radiance that arrives at s directly from light sources which is reflected in direction ω_o, and $L^{\Leftarrow}(s,\omega_o)$ is the radiance emitted by light sources which arrives at s over surfaces and is reflected into direction ω_o, see Figure 4.73.

Depending on the number, position, direction, and the shape of the light sources as well as the object surfaces within a scene, the solid angles subtended by the integration domains of the integrals from Equation (4.406) as seen from point s can be very complicated. Evaluating these integrals, represented in spherical-form, means, that just these projections must be computed. Due to [10, Arvo 1995] this is not a trivial but a very tricky task. Only in the case where all of our light sources are point light sources, it is trivial. Let us consider this case in the following example.

Solid Angle (83)

EXAMPLE 4.15 *Let us assume we have a scene illuminated by n point light sources $\{*_1,\ldots,*_n\}$. Obviuously, the projection $*^\perp$ of the point light sources onto the unit sphere is a null set, that is, due to the properties of the Lebesgue integral, the direct illumination $L^{\leftarrow}(s,\omega_o)$ must be zero.*

Now, due to Definition 2.12 point light sources are locations of infinitely high power. As they can be defined via the Dirac δ-distribution, we can replace the radiance incident at point s from direction ω_i by the radiance, emitted from point light source $*_j$ in direction ω_o^j, thus,

Dirac δ-distribution (118)

$$L_e(*_j,\omega_o^j) = \delta(\omega_i \to -\omega_o^j)\, L_e(\gamma(s,\omega_i),-\omega_i), \tag{4.408}$$

where we assume: $\gamma(s,\omega_i) = *_j$ and $\omega_o^j = -\omega_i$, see Figure 4.73.

Splitting the incident radiance L_i within the SLTEV

$$L_o(s,\omega_o) = L_e(s,\omega_o) + \int_{S^2(s)} f_s(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^\perp(\omega_i) \tag{4.409}$$

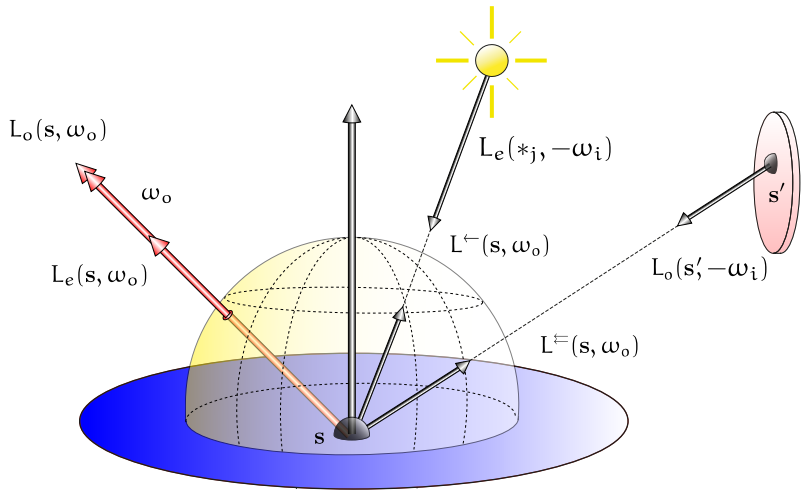


FIGURE 4.73: DIRECT AND INDIRECT ILLUMINATION FORMULATION OF THE SLTEV.
The light reflected at point s in direction ω_o is a composition of an emission term, $L_e(s,\omega_o)$, a direct illumination component, $L^{\leftarrow}(s,\omega_o) = L_e(*_j,-\omega_i)$, and an indirect illumination component, $L^{\leftleftarrows}(s,\omega_o) = L_o(s,-\omega_i)$.

into the emitted radiance L_e *from point light sources and the incident radiance* L_i *from non emitting surfaces and using this in the above formula, then we get:*

$$L_o(s,\omega_o) = L_e(s,\omega_o) + \int_{S^2(s)} f_s(s,\omega_i \to \omega_o)\, L_e\left(\bigcup_{j=1}^{n} *_j,-\omega_i\right) d\sigma_s^{\perp}(\omega_i) + \int_{S^2(s)} f_s(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i). \quad (4.410)$$

Using Relation (4.408) and the fact that $*^{\perp}$ *is a null set, then we can write the SLTEV in the following form:*

$$L_o(s,\omega_o) = L_e(s,\omega_o) + \int_{S^2(s)} f_s(s,\omega_i \to \omega_o)\, \delta(\omega_i \to -\omega_o^j)\, L_e(\gamma(s,\omega_i),-\omega_i)\, d\sigma_s^{\perp}(\omega_i) + \quad (4.411)$$

$$\int_{S^2(s)} f_s(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i), \quad (4.412)$$

where $\gamma(\mathbf{s}, \omega_i) \in \{*_1, \ldots, *_n\}$.

The δ-distribution in the first integral then allows to write the first integral as a sum over the set of point light sources, that is,

$$\begin{aligned} L_o(\mathbf{s}, \omega_o) &= L_e(\mathbf{s}, \omega_o) + \sum_{j=1}^{n} f_s(\mathbf{s}, -\omega_o^j \to \omega_o)\, L_e(*_j, \omega_o^j) \left|\cos\theta_i\right| + \\ &\int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i). \end{aligned} \tag{4.413}$$

Obviously, the radiance exitant at point $\mathbf{s}$ *in direction* ω_o *is the sum of a self-emitted fraction, the radiance emitted from all point light sources* $*_j$ *in the directions* ω_o^j *towards* $\mathbf{s}$ *and reflected to* ω_o*, and the indirect illumination incident at* $\mathbf{s}$.

REMARK 4.50 *In Example 6.45 we will show the equivalence of the stratification of the integration domain of the SLTEV and* next event estimation*, a variance reduction technique for solving linear integral equations. This technique is used in many rendering algorithms based on stochastic principles as a procedure for getting better images more efficiently and with only little more effort.*

Section 6.7.4

Section 6.6

Section 6.7.4

In Example 6.46 we will illustrate, that the direct illumination can easily be solved by transforming the spherical integral into a surface integral. This does not require the computation of the projection of the light sources onto the unit sphere. The integration domain of a surface integral can then easily be sampled using variance reduction techniques from Monte Carlo integration.

3-point Form of SLTEV (402)

Section 6.7.4

4.5 THE IMPORTANCE TRANSPORT EQUATION IN A VACUUM

In Chapter 10 we will discuss the radiosity method, a finite element based technique for solving the light transport problem. Radiosity methods divide the scene to be rendered into a finite set of surface patches and use the energy that is reflected from these patches to determine the radiosity of a surface point to be shaded.

Radiosity (264)

Now let us assume, that a radiosity algorithm has partitioned a scene into millions of small patches, where we have to evaluate the stationary light transport equation for each of those patches. But if we are only interested in a small sector of the scene, then only a fraction of all surfaces have a significant impact on the final image. That is, if can develop a method that determines just these important surfaces, then we do not need so much effort spent in computing the radiosities of unimportant surfaces. The idea behind such an algorithm leads to the fundamental concept of *importance*, firstly inspired by works in connection with Monte Carlo simulations for the neutron transport.

SLTEV (398)

In Section 2.3.2 we have shown, that for a given integral equation, there exists infinitely many adjoint equations, each with a different source term. The solution to an adjoint equation with a source term at the most important part of the function domain is called importance since it indicates how much the different parts of the domain contribute to the solution at the most important part, [35, Christensen 2003].

Adjoint Equation (65)
Source Term (127)

Let us assume, we are interested in measuring the radiance

$$L_o(s,\omega_o) = L_e(s,\omega_o) + \int_{S^2(s)} f_s(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i) \qquad (4.414)$$

with respect to any arbitrary function W_e, that is, we want to compute the quantity

$$\langle W_e, L_o \rangle. \qquad (4.415)$$

We have shown in Section 2.3.2, that to any integral operator of a Fredholm type integral equation of the 2^{nd} kind, there exists the adjoint operator. Since the integral kernel within the stationary light transport equation in a vacuum is given by the BSDF, the adjoint operator can easily be obtained by switching the involved variables in the BSDF. With respect to the SLTEV this means, that we can define the adjoint integral operator, f_s^*, via:

Integral Operator (130)
Fredholm Integral Equation (127)
SLTEV (398)
BSDF, (371)

$$f_s^*(s,\omega_i \to \omega_o) \stackrel{def}{=} f_s(s,\omega_o \to \omega_i), \qquad (4.416)$$

that is, we simply reverse the incident and the exitant direction within the BSDF.

Based on any arbitrary function W_e and the adjoint operator f_s^* we can now define the adjoint of the stationary light transport equation within a vacuum:

SLTEV (398)

DEFINITION 4.44 (The Stationary Importance Transport Equation in a Vacuum) *Let W_e be any given function from $\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \zeta^{\perp})$ and f_s^* be a linear integral operator given by*

$\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \zeta^{\perp})$ (111)

$$f_s^*(s,\omega_i \to \omega_o) \stackrel{def}{=} f_s(s,\omega_o \to \omega_i), \qquad (4.417)$$

then there exists a function $W_i \in \mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \zeta^{\perp})$ that satisfies the adjoint equation

$$W_o(s,\omega_o) = W_e(s,\omega_o) + \int_{S^2(s)} f_s^*(s,\omega_i \to \omega_o)\, W_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i) \qquad (4.418)$$

$$= W_e(s,\omega_o) + \int_{S^2(s)} f_s(s,\omega_o \to \omega_i)\, W_i(s,\omega_i)\, d\sigma_s^{\perp}(\omega_i), \qquad (4.419)$$

the so-called stationary importance transport equation in a vacuum, *also briefly denoted as,* SITEV, *see Figure 4.74.*

As you can see, the structure of the stationary light transport and the stationary importance transport equation in a vacuum is the same. Obviously, the quantities W_e

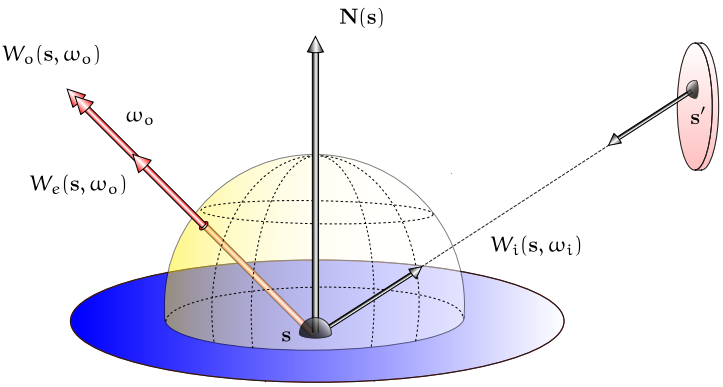


FIGURE 4.74: THE STATIONARY IMPORTANCE TRANSPORT EQUATION IN A VACUUM. The SITEV can be interpreted as a transport equation for the influence of the detector, which describes the transport of importance through a scene.

and W_i play the same roles within the importance equation as the emitted radiance L_e and the incident radiance L_i play in the SLTEV. This is also the reason, why we call W_e the *emitted importance* and W_i the *incident importance*.

IncidentFunction (48)
Radiance (250)

Importance is the adjoint of light. It is transported like light, but in the reverse direction. While the light sources, described by the source term L_e, are the most interested construct for the SLTEV, for the SITEV, we can consider the eye, a virtual camera, or the directly visible parts of a scene, as the most relevant regions. That is, the importance equation can be interpreted as a transport equation for the influence of the detector. The transport rules known from the SLTEV can be applied equally well to the sensors, by treating the responsivity as an emitted quantity, namely the above mentioned emitted importance.

Source Term (127)

REMARK 4.51 *In a global illumination algorithms, the concept of importance can be used as a tool for reducing the computational effort needed, since it can be used to control which parts of scene are relevant for computing an image, thus for example, W_e can be defined as an exitant function that only takes non-zero values for points of a pixel and directions from the solid angle subtended by the camera lens as seen from the pixel.*

REMARK 4.52 (Alternative Formulations of the Stationary Importance Transport Equation

in a Vacuum) *Mostly, importance is represented as an exitant directional quantity,* $W_o(\mathbf{s}, \omega_o)$*, from a point* $\mathbf{s}$ *in direction* ω_o*. As light and importance are adjoint, it can easily be shown that the principle of radiance invariance can also be transferred to importance, that is, even importance is* invariant *along a ray within a vacuum. This then implies that importance can also be expressed as an incident function* W_i*, where it holds:*

Radiance Invariance (253)

$\gamma(\cdot,\cdot)$ (47)

$$W_i(\mathbf{s}, \omega_i) = W_o(\gamma(\mathbf{s}, \omega_i), -\omega_i). \tag{4.420}$$

Using Relation (4.420) in the SITEV then we get the spherical form of the SITEV expressed in terms of exitant importance, namely,

$$W_o(\mathbf{s}, \omega_o) \quad = \quad W_e(\mathbf{s}, \omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_o \to \omega_i)\, W_o(\gamma(\mathbf{s}, \omega_i), -\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i). \tag{4.421}$$

Similar to our procedure in Section 4.4.2.1 we can now also derive the remaining incident and exitant formulations of the stationary importance transport equation in a vacuum in spherical, hemispherical, or 3-point form. We leave the details of these derivations to the interested reader as simple exercises.

In [47, Dutré 1996] the approach is made, to define importance in a similar manner as we have introduced radiance in Section 3.3, namely, as the fraction of light that indicates its contribution to the region of most interest.

REMARK 4.53 (Importance) *Let* $\mathbf{s}$ *be a point upon a surface* $\partial\mathcal{V}$ *and* ω_o *any direction over the unit sphere around* $\mathbf{s}$*. Then, the* importance *or the* potential *of point* $\mathbf{s}$ *in direction* ω*, denoted by* $W(\mathbf{s}, \omega)$*, is defined as the flux through* $\partial H \times \Gamma \subset \partial\mathcal{V} \times S^2$ *as the result of the differential flux from point* $\mathbf{s}$ *in direction* ω_o*, thus,*

$\partial\mathcal{V}$ (41)

Flux (249)

$$W(\mathbf{s}, \omega_o) \quad \stackrel{def}{=} \quad \frac{d^2\Phi(\partial H \times \Gamma)}{d^2\Phi(\mathbf{s}, \omega_o)} \tag{4.422}$$

$$= \quad \frac{d^2\Phi(\partial H \times \Gamma)}{L(\mathbf{s}, \omega_o) d\sigma^{\perp}_{\mathbf{s}}(\omega_o) d\mu^2(\mathbf{x})}. \tag{4.423}$$

Obviously, importance is a five-dimensional, dimensionless quantity that varies with position and direction. Often, it is also denoted as potential. [150, Pattanaik & Mudur 1993].

4.6 THE MEASUREMENT EQUATION

Until now, we were only interested in computing and specifying the radiance distribution within a scene. Now, let's look at using sensors to measure the incident radiance and to

generate images.

Global Illumination Problem

The main task in computer graphics is to find a solution of the global illumination problem, in other words, the construction of an image that visualize a scene specified by the global illumination problem. A solution to this can be computed via a two step process: First, we have to determine the region of the scene, which is relevant for the final image, and then the illumination at all points within this region has to be computed by solving the corresponding stationary light transport equation. Since an image is stored as an array of pixels $\square_1, \ldots, \square_{s_x \cdot s_y}$, where each pixel covers a finite subregion of the viewing frustum, we have to perform a set of real-valued measurements $\mathcal{M}_1, \ldots, \mathcal{M}_{s_x \cdot s_y}$. Each measurement then corresponds to the output of a hypothetical sensor that responds to the radiance incident upon it. This is characterized by the *emitted importance function*, W_e^j, which varies according to the position and direction at which light strikes the sensor. It specifies the importance of light arriving along each ray to the corresponding measurement $\mathcal{M}_j, 1 \leq j \leq s_x \cdot s_y$.

Radiance (250)

W_e^j (416)

The total response of a sensor can then be determined by integrating the product of the incident radiance and the emitted importance function, that is, by evaluating the so-called *measurement equation*, which is defined as follows:

DEFINITION 4.45 (The Measurement Equation) *Let $\mathbf{s}$ be a point on an object surface from $\partial\mathcal{V}$, $L_i(\mathbf{s}, \omega)$ be the incident radiance at point $\mathbf{s}$ from direction ω and $W_e(\mathbf{s}, \omega)$ be the exitant importance of $\mathbf{s}$ in direction ω. Then, the sensor response $\mathcal{M}$ can be computed by a measurement of the form*

Radiance (250)

Importance (416)

$$\mathcal{M} \overset{\text{def}}{=} \int_{\partial\mathcal{V}} \int_{S^2(\mathbf{s})} W_e(\mathbf{s}, \omega)\, L_i(\mathbf{s}, \omega)\, d\sigma_{\mathbf{s}}^{\perp}(\omega)\, d\mu^2(\mathbf{s}) \qquad [W] \tag{4.424}$$

$$= \int_{\partial\mathcal{V}} \int_{S^2(\mathbf{s})} W_e(\mathbf{s}, \omega)\, L_i(\mathbf{s}, \omega)\, \underbrace{\langle \mathbf{N}(\mathbf{s}), \omega \rangle}_{|\cos\theta_{\mathbf{s}}|}\, d\sigma_{\mathbf{s}}\, d\mu^2(\mathbf{s}) \tag{4.425}$$

$$= \int_{\partial\mathcal{V}} \int_{S^2(\mathbf{s})} W_e(\mathbf{s}, \omega)\, L_i(\mathbf{s}, \omega)\, |\cos\theta_{\mathbf{s}}|\, d\sigma_{\mathbf{s}}\, d\mu^2(\mathbf{s}), \tag{4.426}$$

where we assumed, that the sensors are part of the scene so that we can integrate over their surfaces. Due to [137, Nicodemus 1978], we call $\mathcal{M}$ the measurement equation.

Throughput Measure (94)

REMARK 4.54 *Using Definition 2.26 of the throughput measure, the measurement equation can also be written in the following form*

$$\mathcal{M} \overset{(2.209)}{=} \int_{\partial\mathcal{V} \times S^2(\mathbf{s})} W_e(\mathbf{r})\, L_i(\mathbf{r})\, d\zeta^{\perp}(\mathbf{r}) \tag{4.427}$$

$$\overset{(2.265)}{=} \langle W_e, L_i \rangle, \tag{4.428}$$

Linear Functional (55)

with $\mathbf{r} = (\mathbf{s}, \omega) \in \mathcal{R}^{\partial\mathcal{V}}$, where the linear functional $\langle \cdot, \cdot \rangle$ is defined by the inner product

Ray Space (44) *defined on the $\mathcal{L}^2$-function space over $\mathcal{R}^{\partial\mathcal{V}}$ with respect to the projected throughput measure.*

In our discussions, the measurement equation is always be interpreted as the response of a sensor or a sensor element, such as a pixel $\square_j$ on the image plane of a camera that measures the incident radiance upon it. These sensors are usually virtual and do not interfere with the light transport in the scene, where the responsivity W_e^j of the sensor $\square_j$ is zero almost everywhere, except for points and directions that lie within the solid angle subtended by the pixel and the camera lens used. We use the measurement equation then in the form

Radiance (250)

$$\mathcal{M}_j \stackrel{def}{=} \int_{\square_j}\int_{S^2(\mathbf{s})} W_e^j(\mathbf{s},\omega)\, L_i(\mathbf{s},\omega)\, d\sigma_{\mathbf{s}}^{\perp}(\omega)\, d\mu^2(\mathbf{s}), \qquad [W] \tag{4.429}$$

where $\bigcup_{j=1}^{s_x\cdot s_y}\square_j \subset \partial\mathcal{V}$ corresponds to the image plane, with s_x, s_y as the resolution in x and y-direction. Each measurement $\mathcal{M}_j$ works with a different responsivity function W_e^j, which can be used to model arbitrary lens systems as well as linear filters used for antialiasing.

EXAMPLE 4.16 (The Importance Function of a Pinhole Camera Model) *The simplest camera model which we can use in a rendering algorithm is the pinhole camera, see Figure 4.75. This camera type is specified by the eye, $\mathbf{e} \in \mathbb{R}^3$, of an observer or a virtual camera and a rectangular image plane in front of the eye. An image is constructed by performing the central projection of the scene objects to the image plane, which corresponds to a pixel array of dimension $s_x \cdot s_y$.*

When rendering an image, then we have to determine the radiance passing through each pixel of the image plane, that is, we have to compute the inner product $\langle W_e^j(\mathbf{s},\omega), L_i(\mathbf{s},\omega)\rangle$ for all points $\mathbf{s} \in \square_j$ and all directions ω within the solid angle subtended by the pixel $\square_j$ as seen from the eye.

Since the aperture of such a camera is a single point, the importance function underlying a pinhole camera must incorporate a Dirac δ-distribution, that is, it can be replaced by a product of the form

Dirac δ-distribution (117)

$$W_e^j(\mathbf{s},\omega) = \delta_\sigma(\omega - \omega_e)\, f_j(\mathbf{s}) \tag{4.430}$$

where $\omega_e = \frac{\mathbf{e}\to\mathbf{s}}{\|\mathbf{e}\to\mathbf{s}\|_2}$ is the direction outgoing from eye point $\mathbf{e}$ through point $\mathbf{s}$ within pixel $\square_j$ of the image plane and f_j is a normalized reconstruction filter function for pixel $\square_j$.

Replacing the importance function W_e^j in the measurement equation by Relation (4.430), then we get:

$$\begin{aligned}\mathcal{M}_j &= \int_{\square_j}\int_{S^2(\mathbf{s})} \delta_\sigma(\omega-\omega_e)\, f_j(\mathbf{s})\, L_i(\mathbf{s},\omega)\, d\sigma_{\mathbf{s}}^{\perp}(\omega)\, d\mu^2(\mathbf{s}) && (4.431)\\ &= \int_{\square_j} f_j(\mathbf{s})\, L_i(\mathbf{s},\omega_e)\, |\cos\theta_e|\, d\mu^2(\mathbf{s}). && (4.432)\end{aligned}$$

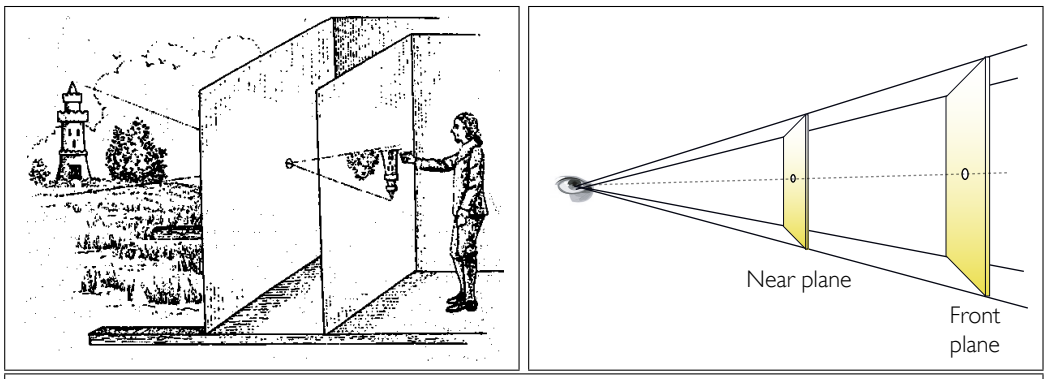


FIGURE 4.75: REAL AND VIRTUAL PINHOLE CAMERA MODEL. The image on the left shows a real pinhole camera, also called *camera obscura*. The pinhole of a real camera obscura is located between the scene objects and the film plane. The image is constructed by a central projection of the scene objects onto the film plane, that is, the image is inverted about a line through the pinhole and perpendicular to the image plane. On the right, we have a virtual pinhole camera, as used in CG. Here, the film plane is located in front of the hole at the near plane, and the hole corresponds to the eye point. Image courtesy of NN.

A more efficient choice of the importance than that from Equation (4.430) could be the Dirac distribution $\delta_{\sigma^\perp}$:

$$W_e^j(\mathbf{s},\omega) = \frac{\delta_\sigma(\omega-\omega_e)}{|\cos\theta_\mathbf{s}|}\, f_j(\mathbf{s}) \tag{4.433}$$

$$= \delta_{\sigma^\perp}(\omega-\omega_e)\, f_j(\mathbf{s}) \tag{4.434}$$

then the cosine in the measurement equation can be removed and the measurement equation can be expressed in a more simplified form, namely by

$$\mathcal{M}_j = \int_{\square_j} f_j(\mathbf{s})\, L_i(\mathbf{s},\omega_e)\, d\mu^2(\mathbf{s}). \tag{4.435}$$

We will encouter Equation (4.435) when discussing rendering algorithms in connection with antialiasing.

Section 8.4.3

Flux (249) REMARK 4.55 *Since the measurement equation measures the flux through a pixel, but the human eye is sensitive to radiance rather than to flux, the flux through a pixel is usually converted to an average radiance value L^j_{avg} defined as:*

$$L^j_{avg} \stackrel{def}{=} \frac{\mathcal{M}_j}{\int_{\partial\mathcal{V}}\int_{S^2} W_e^j(\mathbf{s},\omega)\, d\sigma^\perp_\mathbf{s}(\omega)\, d\mu^2(\mathbf{s})}. \tag{4.436}$$

REMARK 4.56 (The Pixel Equation) *Using more realistic camera models we can enhance the realism of rendered images. Then, effects such as depth of field or motion blur can*

Section 8.4.3.2 *be simulated by using a lens camera model and adapting the measurement equation to include an integration over a finite exposure time. The measurement equation,* Section 8.4.3.3 *then also often called the* pixel equation*, see [107, Kolb & al, 1995], has the form*

$$\mathcal{M}_j \stackrel{def}{=} \int_{\square_j} \int_{\Gamma} \int_{\Lambda} \int_{T} L(T(s,\omega,\lambda),\lambda,t)\, S(s,\omega,t)\, f(s,\lambda)\, d\mu(t) d\mu(\lambda) d\sigma_s^{\perp}(\omega) d\mu^2(s), \quad (4.437)$$

where f *is the sensor response on the pixel* $\square_j$*,* T *corresponds to the lens optics,* S *models the behavior of the shutter,* T *is the exposure time,* Γ *is the solid angle subtended by the lens as seen from pixel* $\square_j$ *and* Λ *is the band of wavelength of visible light.*

The pixel equation can also be simplified by putting the above dependencies in the importance function W_e^j*, that is, we can also write:*

$$\mathcal{M}_j \stackrel{def}{=} \int_{\square_j} \int_{\Gamma} \int_{T} W_e^j(s,\omega,t)\, L_i(s,\omega,t)\, d\mu(t) d\sigma_s^{\perp}(\omega) d\mu^2(s), \quad (4.438)$$

where W_e^j *simulates the behavior of the shutter* S *and the lens, as well as the pixel filter function* f *from above. This is the form of the measurement we use in Section 8.4.3 when extending distribution ray tracing in order to render more realistic images.* Section 8.4

4.7 REFERENCE LITERATURE AND FURTHER READING

There is a vast literature on a detailed discussion of optics in general, but we give only a few representative titles, such as [44, Ditchburn 1991], [79, Hecht 1975], [80, Hecht 2001], [59, Feynman 1985], [155, Perez 1996], and [27, Born & Wolf 1999]. [44, Ditchburn 1991] presents a single theory of light, integrating two fields. So, it is shown that quantum theory is a natural development of wave theory, and that together they constitute a single valid theory of light. The book is aimed at students with an intermediate-level knowledge of physics. [59, Feynman 1985] is an adaptation for the general reader of four lectures on quantum electrodynamics. [80, Hecht 2001] starts with the historical development of optics, and provides readers with the most up-to-date coverage of optics. This textbook covers the whole field of optics in a comprehensible manner. As typical for a book from Schaums Outline series, [79, Hecht 1975] is a good reference—with many exercises—for the rather practical oriented reader. The basic concepts of optics in particular with respect to the field of global illumination are also discussed in detail in [66, Glassner 1989], [67, Glassner 1995], [68, Glassner 1995-2], [233, Watt 1992], [186, Shirley 2002], [187, Shirley & Morley 2003], [158, Pharr & Humphreys 2004], and [159, Pharr & Humphreys 2010].

Our approach for deriving scalar versions of the light transport equation, that is the whole Section 4.1, is mainly based on [9, Arvo 1993], [68, Glassner 1995], and [191, Slusallek 1995]. There are many other possible approaches for deriving the light transport equations. So, in [50, Dutré & al. 2003], [51, Dutré & al. 2006], [95, Jensen 2001], [158, Pharr & Humphreys 2004], and [159, Pharr & Humphreys 2010] a rather intuitively approach is chosen, which begins by considering derivatives of flux such as in [33, Chandrasekhar 1960] and [91, Ishimaru 1997]. [163, Pomraning 1973] describes a Lagrangian approach, and in [165, Preisendorfer 1965] a rigorous axiomatic derivation of the general transport equation is presented, which is based on measure theory.

The concept of the bidirectional reflectance distribution functions is from [135, Nicodemus & al. 1977]. According to [135, Nicodemus & al. 1977], we start with the derivation of the BSSRDF. The concept of the BSSRDF is also described in [195, Snyder & Wan 1998] and [196, Snyder 1998]. Also [159, Pharr & Humphreys 2010] deals with this topic. Although the BSSRDF is a time-consuming function in realistic rendering, until today, a series of papers have dealt with it. One of the first papers that deals with subsurface scattering was [75, Hanrahan & Krueger 1993], this paper presents a model for subsurface scattering in layered surfaces in terms of one-dimensional linear transport theory. In [94, Jensen & al. 2001] a simple model for subsurface light transport in translucent materials is introduced. [93, Jensen & Buhler 2002] present an efficient two-pass rendering technique for translucent materials, and in [128, Mertens & al. 2003] a novel approach is presented to efficiently render local subsurface scattering effects.

We derive the BRDF in accordance to [135, Nicodemus & al. 1977], that is, starting with the BSSRDF and the assumption, that it is not dependent on the spatial parameters $\mathbf{s}_\mathrm{i}$ and $\mathbf{s}_\mathrm{o}$. Of great help for our presentation of the BRDF were [175, Rusinkiewicz 1997], [76, Hanrahan & al. 2000], [229, Walter 2005], and [238, Wynn 2006]. Another useful reference was [141, Olano & al. 2002], where the topic was discussed very beautiful in detail. The concept of the BSDF was firstly mentioned in [82, Heckbert 1991]. Unlike the BRDF and BTDF, the BSDF is not a key concept in radiometry, but it plays a major role in computer graphics, and in our theoretical and practical considerations, as it frees us to distinguish between reflection and transmission at surfaces.

Easily readable introductory treatments on spherical harmonics are given in [43, Dempski & al. 2005] and [190, Sillion 1994].

An exhaustive treatment of the concept of the phase function is given by [218, van de Hulst 1981], [156, Petty 2006], [26, Bohren & al. 2004], and in particular in [33, Chandrasekhar 1960]. [154, Pegoraro 2010] is a good reference for the use of the phase function in global illumination.

There is a series of graphics textbooks that provided us with a good overview, in particular, about the different types of light sources and their properties used in computer graphics. These are [158, Pharr & Humphreys 2004], [205, Suffern 2007], and [159, Pharr & Humphreys 2010]. While the books by Pharr and Humphreys discuss the concept of the light source with the focus on implementing light emitters for a physically based rendering

system, in [205, Suffern 2007], also reference is made to the maths describing the direct illumination at a surface point due to a specific light source. We go a similar way, but we derive our formulas in terms of radiance instead of a radiance scaling factor and a color as in [205, Suffern 2007]. Another excellent source with respect to the description of different types of the various light source was [1, Akenine-Möller & al. 2008]. We also remark, that [1, Akinene-Möller & al. 2008] and [158, Pharr & Humphreys 2004], [159, Pharr & Humphreys 2010] apart from the light sources introduced in Section 4.3, present further types of light sources, such as environmental lightning, projection, sky and textured lights, as well as goniophotometric diagram lights. Last but not least, a nice and intuitively description of the most important types of light sources for games programming can be found in [43, Dempski & Viale 2005]. We recommend this text to the reader not so familiar with calculus and higher mathematics.

Our approach for introducing the importance equation is similar to the approach chosen in [221, Veach 1998]. We also uses the concept of the adjoint operator, but on a very low level. So, we have already shown in detail in Chapter 2 that based on an inner product the solution of an operator equation multiplied by the source function of the associated adjoint equation can be reduced to the inner product of the solution of the adjoint equation and the source function of the associated direct problem. In [47, Dutré 1996] and [50, Dutré & al. 2003] the importance equation is introduced based on the concept of importance, namely, as the fraction of light that indicates its contribution to the region of most interest. An overview of the use of importance in speeding up rendering is given in [35, Christensen 2003]. Here, an attempt is made to clarify the various uses of adjoints and importance in rendering by unifying them into a single framework.

The concept of the measurement equation is from [137, Nicodemus 1978] and more realistic camera systems are discussed in [107, Kolb & al. 1995].

CHAPTER FIVE

MATHEMATICAL MODELS OF LIGHT AND IMPORTANCE TRANSPORT

In the beginning God created the heavens and the earth. Now the earth was formless and empty, darkness was over the surface of the deep, and the Spirit of God was hovering over the waters. And God said:

$$\begin{aligned} \nabla \cdot \mathbf{E} &= \frac{\sigma}{\epsilon} \\ \nabla \cdot \mathbf{B} &= 0 \\ \nabla \times \mathbf{E} &= -\frac{\partial}{\partial t}\mathbf{B} \\ \nabla \times \mathbf{B} &= \mu\sigma\mathbf{E} + \mu\epsilon\frac{\partial}{\partial t}\mathbf{E} \end{aligned}$$

and there was light. God saw that the light was good and he separated the light from the darkness. God called the light day *and the darkness* night*. And there was evening, and there was morrning—the first day.*

GENESIS 1:3-5

Due to Definition 1.3, the global illumination problem is given by the quadruple

$$\left(\mathcal{V}, L_e, \bigcup_{i=1}^{n} f_{s_i}, W_e\right), \tag{5.1}$$

and consists in evaluating the measurement equation, i.e. the linear functional: Measurement Equation (416)

$$\mathcal{M} = \langle W_e, L_i \rangle \tag{5.2}$$

for all pixels of the image plane, that is, it is just this equation which must be solved by any rendering algorithm. Linear Functional (55)

In this chapter, we present a light transport framework—for the case of light transport in a vacuum mainly developed in [221, Veach 1998]—that addresses precisely this goal. Based on mathematical concepts such as measures, function spaces, inner products, and linear operators, it also serves as the basis to an operator model of light transport in a vacuum and in participating media. Due to [221, Veach 1998], many aspects of this framework are new, since it does not make assumptions about the symmetry of the used BSDFs and thus, leads to a richer structure than previous approaches. So, it does adequately describe the relationships between light and importance transport, recursive evaluation, or between incident and exitant transport quantities. As a result, we get different but equivalent and symmetrical formulations of the measurement equation. This has the advantage that solutions for the global illumination problem can be derived in various ways: by distributing radiance from light sources into the scene, and collecting the incident or exitant radiance at pixel sets that must be visible from the final image, or by distributing importance from sources and collecting them at locations that are illuminated by light sources of the scene. Based on this framework, it is also possible to distribute radiance and importance into the scene simultaneously, and to compute their interaction at common points of intersection.

Section 5.1 OVERVIEW OF THIS CHAPTER. In this chapter, we talk about operator models for light transport in a vacuum and in participating media. Thus, we first present the operator model of light transport in a vacuum developed in [219, Veach 1996] and [221, Veach 1998]. We then use this model as the fundament for a new operator model of the light transport in participating media. For that purpose we extend the light propagation and light scattering operators introduced in [219, Veach 1996] by additional multiplication and integral operators resulting in a light transport model in participating media. Afterwards, we turn to the dual problem of light transport. We present an operator model of the
Section 5.2 importance transport in a vacuum using the *stationary importance transport equation in a vacuum*, the adjoint equation of the SLETV. Based on these two models then we derive four basic transport operator equations which, applied to the measurement equa-
Section 5.3 tion, lead to four different but equivalent algorithms for solving the global illumination problem. Next, we devote to the construct of the *path integral* and discuss the path inte-
Section 5.4 gral formulation of light transport based on the concept of the *continuous path measure*. Endowed with this new measure, we construct a specific measurable space, the so-called *path space*, that allows to describe the light transport as well as the importance transport, instead of an integral equation, as a simple integral over all possible paths within a scene. Here, we restrict our discussion not only to the case of light transport in a vacuum, but we also consider the light transport in participating media. We conclude this chapter with a short overview of a further mathematical model of light and importance transport, the
Section 5.5 *global reflectance distribution function*. The idea behind the GRDF is the concept of the BRDF. But compared with a BRDF, the GRDF is able to compute the behavior of light in an environment, independent of the initial lighting or viewpoint conditions. Thus, the concept of the GRDF can be interpreted as a combination of radiance and importance

transport and allows us to describe the global illumination problem in a very short and elegant way.

5.1 OPERATOR MODELS FOR LIGHT TRANSPORT

For generating a photorealistic image, a great class of rendering algorithm computes the radiance distribution at all visible points within a scene, which is obviously a computational costly task. This can be done by determining the flux incident at all pixels of the image plane. For that purpose, we have to solve the SLTE at points visible through the pixels of the image plane, and have to evaluate the linear functional,

Radiance (250)
Flux (249)
SLTE (394)
Linear Functional (55)

$$\mathcal{M} \stackrel{\text{def}}{=} \langle W_e, \mathrm{L_i} \rangle, \tag{5.3}$$

for these pixel sets.

Measurement Equation (416)

In the above inner product, W_e and $\mathrm{L_i}$ are functions from the function space $\mathcal{L}^2(\mathcal{R}, \zeta)$, where W_e corresponds to the emitted importance at points and directions from the ray space $\mathcal{R}$ and $\mathrm{L_i}$ is the unknown incident radiance—thus, solutions of the stationary light transport equation in participating media—at all these points and directions, see Figure 5.1.

Inner Product (859)
$\mathcal{L}^2(\mathcal{R}, \zeta)$ (111)
Emitted Importance (416)
$\mathcal{R}$ (44)

From our discussions in Chapter 4, as well as from Figure 5.2 follows that the light transport in participating media is composed of three processes, an emission, an absorption, and a scattering process. All these processes can be observed at surfaces of objects within a scene and at small particles within eventually involved participating media. That is, points at surfaces as well as points within media can serve as emitters, absorbers, and apart from scattering at a point on a surface, light can also be scattered at small particles within participating media.

Emission (282)
Absorption (282)
Scattering (284)

Based on this observation, the SLTE can then be split into four terms: two emission terms, one for emission at object surfaces and one for emission at particles within participating media, and two scattering terms, even for surfaces and particles within media, where each emission and scattering term is combined with an absorption term. Slightly rephrased, and expressed in terms of exitant and incident radiance, the SLTE can then be

SLTE (394)
Exitant & Incident Functions (48)

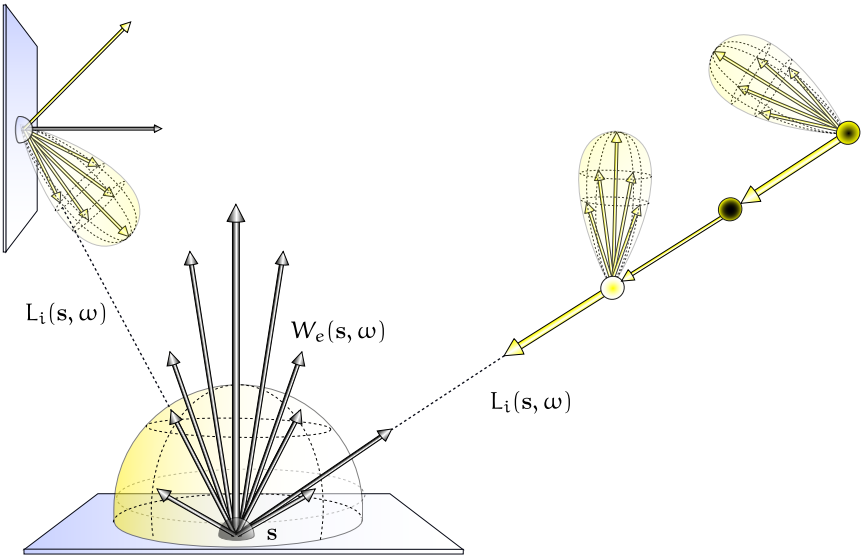


FIGURE 5.1: EVALUATING THE MEASUREMENT EQUATION. For all points s within a pixel, the product of the exitant importance in all directions, $W_e(s,\omega)$, over the unit sphere about s is multiplied with the incoming radiance $L_i(s,\omega)$.

written as:

$$
\begin{aligned}
&L_i(x,\omega_i) \\
&= \underbrace{\beta(s\to x)}_{\text{volume attenuation}}\ \underbrace{L_{e,o}(s,\omega_o)}_{\text{emission at surface}} + \int_{[0,\|x'-x\|]} \underbrace{\beta(x'\to x)}_{\text{volume attenuation}}\ \underbrace{L_{e,o}(x',\omega_o)\,d\mu(\alpha)}_{\text{emission within volume}} + \\
&\quad \underbrace{\beta(s\to x)}_{\text{volume attenuation}} \underbrace{\int_{S^2(s)} f_s(s,\omega_i'\to\omega_o)\,L_i(s,\omega_i')\,d\sigma_s^{\perp}(\omega_i')}_{\text{scattering at surface}} + \\
&\quad \int_{[0,\|x'-x\|]} \underbrace{\beta(x'\to x)}_{\text{volume attenuation}}\ \underbrace{\sigma_s(x')\int_{S^2(x')} p(x',-\omega_i'\to\omega_o)\,L_i(x',\omega_i')\,d\sigma_{x'}(\omega_i')d\mu(\alpha)}_{\text{scattering within volume}},
\end{aligned}
\tag{5.4}
$$

$d_{\partial V}$ (47) where $s=\gamma(x,\omega_i), x'=x+\alpha\omega_i, 0<\alpha<d_{\partial V}(x,\omega_i)$. The two first terms on the right-hand side describe the emission processes and the two last terms describe the scattering processes, all endowed with an attenuation factor, β, due to subsequent absorption processes within participating media.

Now, integral equations of such a complexity are commonly not analytically solvable,

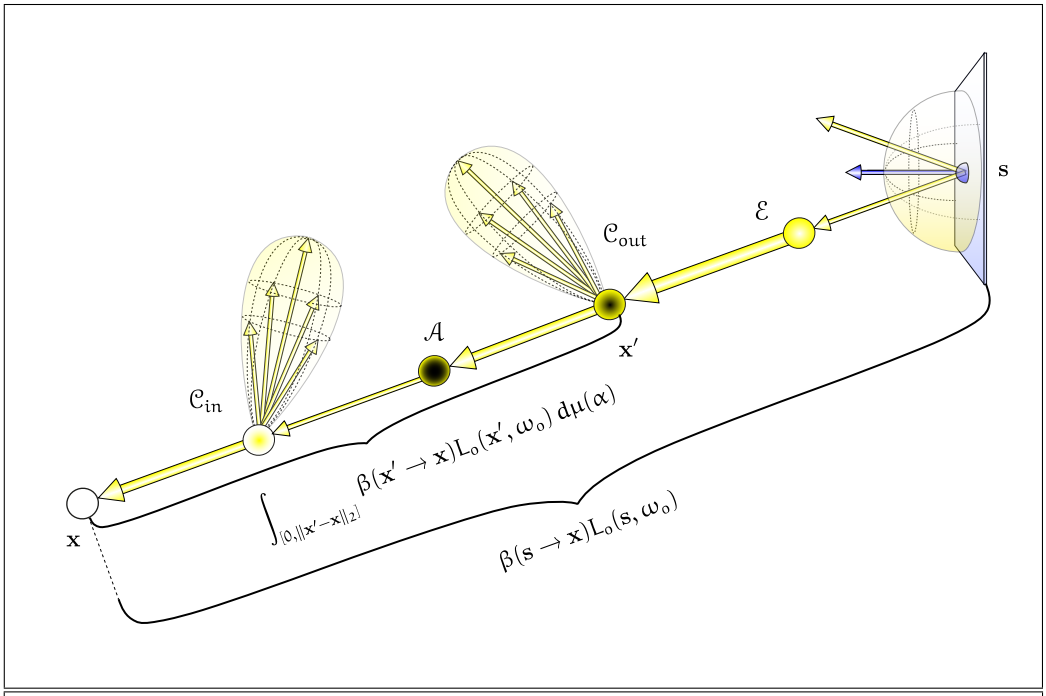


FIGURE 5.2: GEOMETRY OF THE SLTE. The light transport in participating media is composed of three processes, an emission, an absorption, and a scattering process, $\mathcal{E}, \mathcal{A}$ and $\mathcal{C}$. All these processes can be observed at surfaces of objects within a scene and at small particles within media. That is, points at surfaces as well as points within media can serve as emitters, absorbers, and scatterers. The radiance incident at volumetric point $\mathbf{x}$ from direction ω is the sum of the attenuated and reflected radiance at surface point $\mathbf{s}$ as well as the exitant, attenuated radiance that comes from all volumetric points between $\mathbf{x}$ and $\mathbf{s}$

except for very trivial cases. Unfortunately, this also holds for the SLTE. As we know, the SLTE is a Fredholm type integral equation, and we also know that integral equations of this type are solvable via the Neumann series approach. That is: If we can formulate the SLTE from Equation (5.4) as a linear operator equation of the form

SLTE (394)
Fredholm Integral Equation (127)
Neumann Series Approach (135)
Linear Operator Equation (61)

$$\mathrm{L_i} = \mathrm{L}_{e,\mathrm{i}} + \mathbf{T}\,\mathrm{L_i}, \tag{5.5}$$

we have a chance to compute an approximate solution of Equation (5.4) via the Neumann series approach. But this requires the identification of the exitant radiance function $\mathrm{L}_{e,\mathrm{i}}$ as the source or driving function of an integral equation as well as the incident radiance function $\mathrm{L_i}$, and the construction of a linear operator $\mathbf{T}$ on the Lebesgue space $\mathcal{L}^2(\mathcal{R}, \zeta)$.

$\mathcal{L}^2(\mathcal{R}, \zeta)$ (111)
Neumann Series Approach (135)
Driving Function (127)
Linear Operator (53)

Our main goal in the section is precisely to develop a mathematical framework that serves as the basis to operator models of light transport, in both their incident and exitant

forms. Before we will derive such a mathematical framework for the general case of light transport in participating media let us first of all simplify our discussion by considering light transport under vacuum conditions.

5.1.1 AN OPERATOR MODEL FOR LIGHT TRANSPORT IN A VACUUM

$\partial\mathcal{V}$ (41) Let us consider the light transport in a closed scene composed of a finite set $\partial\mathcal{V}$ of 2-dimensional surfaces in the Euclidean space $\mathbb{R}^3$ under vacuum conditions. Then, we can ignore the effects of emission, absorption, and scattering at small particles, as there is no medium involved. That is, the general equation characterizing the light transport is reduced to the calculation of radiance at the boundaries of object surfaces, so to say, to the formulation of the boundary conditions.

With respect to Equation (5.4), light transport under vacuum conditions then means that we have to neglect volumetric scattering, volumetric emission, as well as attenuation induced by absorption processes within participating media. That is, at all inner points
$\mathcal{V}^o$ (41) $\mathbf{x}'$ of a volume $\mathcal{V}^o$ we can assume that apart from

$$L_{e,o}(\mathbf{x}', \omega_o) = 0 \tag{5.6}$$

it also hold

$$\sigma_s(\mathbf{x}') = 0 \tag{5.7}$$

and their is no attenuation of radiance between two points $\mathbf{x}$ and $\mathbf{x}'$, that is,

$$\beta(\mathbf{x} \to \mathbf{x}') = 1. \tag{5.8}$$

Based on these assumptions and the relation $\omega_o' = -\omega_i$ in Equation (5.4), then the stationary light transport equation in a vacuum can be formulated in terms of exitant and incident radiance as follows

$$L_i(\mathbf{s}, \omega_i) = L_{e,o}(\mathbf{s}', \omega_o') + \int_{S^2(\mathbf{s}')} f_s(\mathbf{s}', \omega_i' \to \omega_o')\, L_i(\mathbf{s}', \omega_i')\, d\sigma^{\perp}_{\mathbf{s}'}(\omega_i'), \tag{5.9}$$

γ (47) where it holds: $\mathbf{s}' = \gamma(\mathbf{s}, \omega_i)$ for $\mathbf{s}, \mathbf{s}' \in \partial\mathcal{V}$, see Figure 5.3.

It is just this equation that can be considered as our starting point for the derivation of an operator model of light transport in a vacuum.

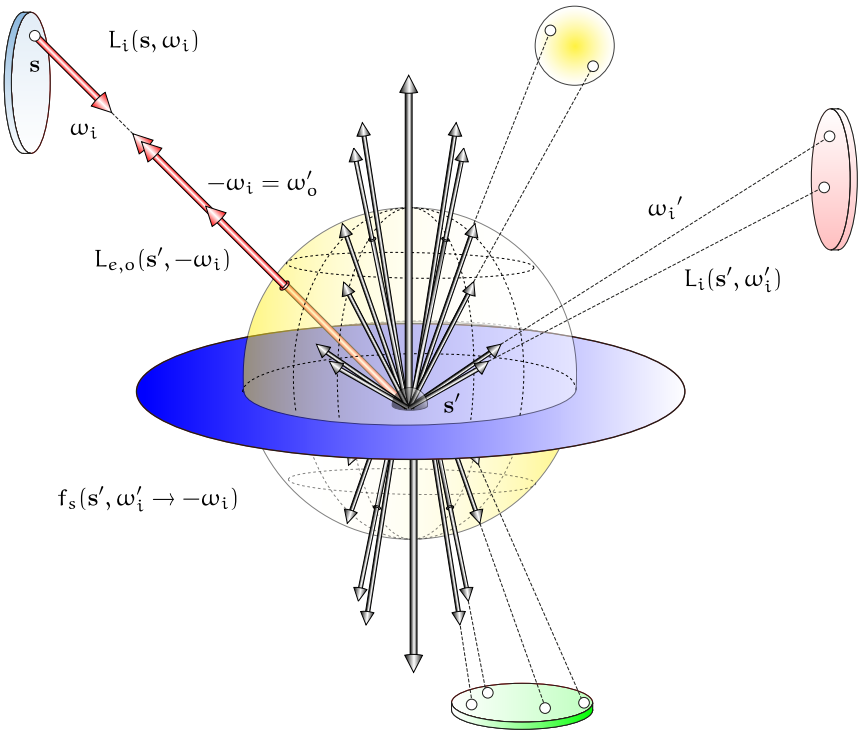


FIGURE 5.3: GEOMETRY OF THE SLTEV. The light transport in a vacuum is composed of an emission and a scattering process. Both processes can be observed only at surfaces of objects within a scene. That is, only points at surfaces can serve as emitters, and light can also only be scattered at object surfaces.

5.1.1.1 THE LIGHT PROPAGATION AND THE LIGHT SCATTERING OPERATOR IN A VACUUM

Let us consider the source function $L_{e,o}(s', \omega'_o)$ from the SLTEV a little bit more closely. Due to the principle of radiance invariance this exitant quantity can also be expressed in terms of incident radiance. Namely, applying the ray casting function γ to the surface point s obviously leads to

Radiance Invariance (253)
Incident Functions (48)
γ (47)

$$L_{e,o}(s', \omega'_o) \overset{s'=\gamma(s,\omega_i)}{=} L_{e,o}(\gamma(s, \omega_i), \omega'_o) \tag{5.10}$$

$$\overset{(3.34)}{=} L_{e,i}(s, \omega_i), \tag{5.11}$$

see Figure 5.3.

SLTEV (398) Using this relation, then the SLTEV can be expressed only in terms of incident radiance as

$$L_i(\mathbf{s},\omega_i) = \underbrace{L_{e,i}(\mathbf{s},\omega_i)}_{\text{emission from surface}} + \underbrace{\int_{S^2(\mathbf{s}')} f_s(\mathbf{s}',\omega_i' \to \omega_o')\, L_i(\mathbf{s}',\omega_i')\, d\sigma^{\perp}_{\mathbf{s}'}(\omega_i')}_{\text{scattering at surface}} \quad (5.12)$$

where $\mathbf{s}' = \gamma(\mathbf{s},\omega_i)$ and $\mathbf{s},\mathbf{s}'$ are points on surfaces from $\partial\mathcal{V}$. That is, the light transport in free space can be considered as a composition of an emission and a scattering event, $\partial\mathcal{V}$ (41) both occurring at object surfaces.

Radiance Invariance (253) Evidently, this partitioning is a consequence of the principle of radiance invariance and the behavior of light within a vacuum, since light incident at a surface point is only influenced by scattered and emitted light propagated from surfaces. The light transport in free space can also be considered as a series of alternating *light scattering* and *light propagation processes*. The *propagation process* describes the travel of photons along lines between surfaces within the vacuum, while the *scattering process* provides information about the interactions of photons at the surfaces.

Linear Operator (53) $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}},\zeta^{\perp})$ (111) We will now mathematically capture any of these processes by means of linear operators acting on the Lebesgue space $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}},\zeta^{\perp})$. With the help of these operators, we are then in a position to formulate the physical processes of propagation and scattering of light in a precise mathematical way.

Exitant & Incident Functions (48) THE LIGHT PROPAGATION OPERATOR IN A VACUUM. It is known that an exitant function $L_o(\mathbf{s},\omega_o)$ measures the radiance leaving a surface point $\mathbf{s}$ in direction ω_o, while an incident radiance function $L_i(\mathbf{s},\omega_i)$ measures radiance arriving at this point from direction ω_i. Light Ray (11) Based on the model of the light ray, geometric optics assumes that light, emitted at surfaces, propagates along straight lines through a vacuum. Compared with the light transport in participating media, the energy of a light ray is, due to the principle of Radiance Invariance (253) radiance invariance, not attenuated in a vacuum. Based on these considerations, a physical Linear Operator (53) process, such as the propagation or the scattering of light, can be interpreted as a linear operator, see Example 2.22, where this operator maps an incident radiance function L_i onto an exitant radiance function L_o.

$\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}},\zeta^{\perp})$ (111) DEFINITION 5.1 (The Light Propagation Operator in a Vacuum, $\mathbf{G}^{\partial\mathcal{V}}$) *Let* $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}},\zeta^{\perp})$ *be the space of square Lebesgue-integrable functions defined on* $\mathcal{R}^{\partial\mathcal{V}}$. *Then, the* light propagation operator in a vacuum

$\mathcal{R}^{\partial\mathcal{V}}$ (44)

$$\begin{aligned} \mathbf{G}^{\partial\mathcal{V}} : \mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}},\zeta^{\perp}) &\longrightarrow \mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}},\zeta^{\perp}) \\ h_o(\mathbf{s}',\omega_o) &\mapsto h_i(\mathbf{s},\omega_i) = (\mathbf{G}^{\partial\mathcal{V}} h_o)(\mathbf{s},\omega_i) \end{aligned}$$

is defined by

$$(\mathbf{G}^{\partial\mathcal{V}} h_o)(\mathbf{s},\omega_i) \stackrel{\text{def}}{=} \begin{cases} h_o(\gamma(\mathbf{s},\omega_i),-\omega_i) & \text{if } d_{\partial\mathcal{V}}(\mathbf{s},\omega_i) < \infty \\ 0 & \text{else,} \end{cases} \quad (5.13)$$

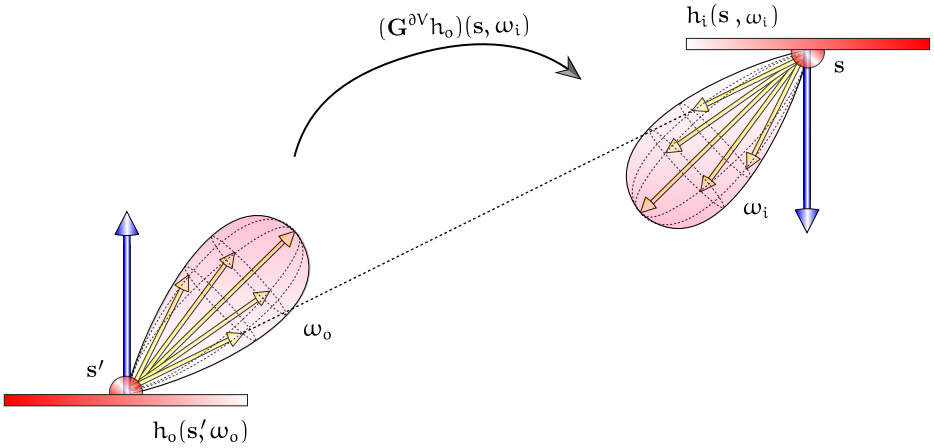


FIGURE 5.4: THE LIGHT PROPAGATION OPERATOR IN A VACUUM. The propagation operator $\mathbf{G}^{\partial\mathcal{V}}$ maps an exitant function $h_o(s',\omega_o)$ onto the incident function $h_i(s,\omega_i) = (\mathbf{G}^{\partial\mathcal{V}}h_o)(s,\omega_i)$ with $\gamma(s,\omega_i) = s'$.

where $s, s' \in \partial\mathcal{V}$ *with* $s' = \gamma(s,\omega_i)$, $\omega_o = -\omega_i$, *and* $d_{\partial\mathcal{V}}$ *is the* boundary distance function *of the ray casting function* γ, *see Figure 5.4.* $d_{\partial\mathcal{V}}$ (47) γ (47)

The light propagation operator can now be interpreted in such a way that it maps an exitant function h_o, *defined on* $\partial\mathcal{V}$, *onto an incident function* $h_i = (\mathbf{G}^{\partial\mathcal{V}}h_o)$, *which is also defined on points from* $\partial\mathcal{V}$. $\partial\mathcal{V}$ (41)

REMARK 5.1 *Applied to an exitant radiance function* L_o, *the operator* $\mathbf{G}^{\partial\mathcal{V}}$ *returns the incident radiance function* $L_i = \mathbf{G}^{\partial\mathcal{V}}L_o$ *as result of the propagation of light from object surfaces. If the function* L_o *measures photons exitant from points of* $\partial\mathcal{V}$, *the function* L_i *obviously measures these photons after propagation incident on surfaces.*

Let us now suppose that the emitted radiance in all points of the $\partial\mathcal{V} \times S^2$ is given by $L_{e,o}(s,\omega_o)$. With the help of the light propagation operator $\mathbf{G}^{\partial\mathcal{V}}$ from above, we can then define an incident emitted radiance function $L_{e,i}$ in terms of $L_{e,o}$ by

$$L_{e,i}(s,\omega_i) \overset{(3.34)}{=} L_{e,o}(\gamma(s,\omega_i),-\omega_i) \tag{5.14}$$

$$\overset{(5.13)}{=} (\mathbf{G}^{\partial\mathcal{V}}L_{e,o})(s,\omega_i). \tag{5.15}$$

With respect to Equation (5.12) and the associated operator equation this means that $\mathbf{G}^{\partial\mathcal{V}}L_{e,o}$ corresponds to the source function of a Fredholm integral equation of the Source Function (127)

2^{nd} kind, i.e. it plays the role of the constant in the operator equation from Relation (5.5).

THE LOCAL LIGHT SCATTERING OPERATOR. In addition to the surface emission term, the SLTEV from Equation (5.12) still contains the scattering equation. Similar to the definition of the light propagation operator in a vacuum, we now define a *local light scattering operator* by:

DEFINITION 5.2 (The Local Light Scattering Operator, $\mathbf{K}^{\partial\mathcal{V}}$) *Let $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp)$ be the space of square Lebesgue-integrable functions defined on ray space $\mathcal{R}^{\partial\mathcal{V}}$. Then, the* local light scattering operator *in a vacuum*

$$\begin{aligned} \mathbf{K}^{\partial\mathcal{V}} : \mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp) &\longrightarrow \mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp) \\ h_i(\mathbf{s}, \omega_i) &\mapsto h_o(\mathbf{s}, \omega_o) = (\mathbf{K}^{\partial\mathcal{V}} h_i)(\mathbf{s}, \omega_o) \end{aligned}$$

is defined as

$$(\mathbf{K}^{\partial\mathcal{V}} h_i)(\mathbf{s}, \omega_o) \stackrel{def}{=} \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, h_i(\mathbf{s}, \omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i) \tag{5.16}$$

where $\mathbf{s} \in \partial\mathcal{V}$, see Figure 5.5.

The local light scattering operator can be interpreted in such a way that it maps an incident function h_i, defined on $\partial\mathcal{V}$, onto an exitant function $h_o = (\mathbf{G}^{\partial\mathcal{V}} h_i)$, which is also defined on points from $\partial\mathcal{V}$.

REMARK 5.2 *Applied to an incident radiance function L_i the operator $\mathbf{K}^{\partial\mathcal{V}}$ returns the exitant radiance function L_o as result from a single scattering operation at an object surface, thus $L_o = \mathbf{K}^{\partial\mathcal{V}} L_i$. If the function L_i measures photons just before their arrival at a surface point, then L_o measures photons after scattering.*

5.1.1.2 THE LIGHT TRANSPORT OPERATOR EQUATION IN A VACUUM

In the previous section we derived the light propagation and the local light scattering operator, and obtained, with $\mathbf{G}^{\partial\mathcal{V}} L_{e,o}$ as the source function, the first component of the linear operator equation associated with Equation (5.5). Now, the question that arises is: How should we construct the operator $\mathbf{T}$?

THE LIGHT TRANSPORT OPERATOR IN A VACUUM. Let us consider the scattering equation of the SLTEV. Obviously, it can be expressed in form of a linear operator equation by applying the local light scattering operator $\mathbf{K}^{\partial\mathcal{V}}$ on the incident radiance function L_i,

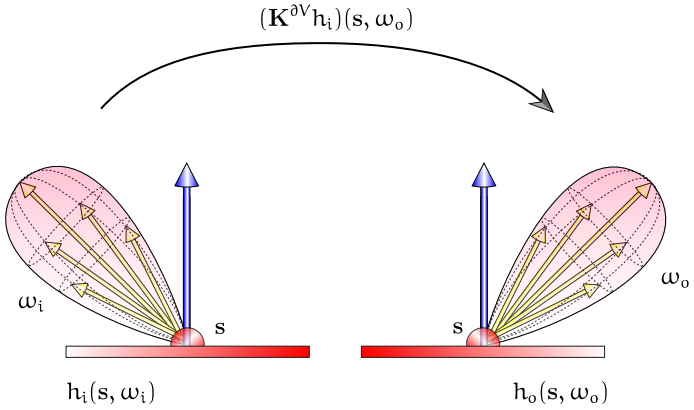


FIGURE 5.5: THE LOCAL LIGHT SCATTERING OPERATOR IN A VACUUM. The local light scattering operator maps an incident function $h_i(s,\omega_i)$ onto the exitant function $h_o(s,\omega_o) = (\mathbf{K}^{\partial\mathcal{V}}h_i)(s,\omega_o)$.

thus:

$$\int_{S^2(s')} f_s(s',\omega_i' \to \omega_o')\, L_i(s',\omega_i')\, d\sigma^{\perp}_{s'}(\omega_i') \overset{(5.16)}{=} \left(\mathbf{K}^{\partial\mathcal{V}} L_i\right)(s',\omega_o') \tag{5.17}$$

$$= \underbrace{\left(\mathbf{K}^{\partial\mathcal{V}} L_i\right)}_{L_o}(\gamma(s,\omega_i),-\omega_i) \tag{5.18}$$

$$= L_o(\gamma(s,\omega_i),-\omega_i) \tag{5.19}$$

with $\omega_o' = -\omega_i$ and $s' = \gamma(s,\omega_i)$.

Now, Equation (5.19) is of the same form as Equation (5.15), that is: the radiance after a scattering event can also be expressed in terms of incident radiance by

Exitant & Incident Functions (48)

$$L_o(\gamma(s,\omega_i),-\omega_i) \overset{(5.13)}{=} (\mathbf{G}^{\partial\mathcal{V}} L_o)(s,\omega_i) \tag{5.20}$$

$$\overset{(5.19)}{=} (\mathbf{G}^{\partial\mathcal{V}}\mathbf{K}^{\partial\mathcal{V}} L_i)(s,\omega_i). \tag{5.21}$$

This means: Applying the propagation operator $\mathbf{G}^{\partial\mathcal{V}}$ to the scattering operator $\mathbf{K}^{\partial\mathcal{V}}$ delivers the fraction of light propagated to some point s after scattering at surface point s' in direction $-\omega_i$. Inspired by [219, Veach 1996] we denote the composition $\mathbf{G}^{\partial\mathcal{V}}\mathbf{K}^{\partial\mathcal{V}}$ of the light propagation and the local light scattering operator as the *light transport operator in a vacuum*:

$\mathbf{G}^{\partial\mathcal{V}}$ (430)
$\mathbf{K}^{\partial\mathcal{V}}$ (432)

DEFINITION 5.3 (The Light Transport Operator in a Vacuum, $\mathbf{T}_{\mathrm{L_i}}^{\partial\mathcal{V}}$) *Let* $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp)$ *be the space of square Lebesgue-integrable functions defined on the ray space* $\mathcal{R}^{\partial\mathcal{V}}$. *Then, the* light transport operator

$\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp)$ (111) $\mathcal{L}(\mathcal{R})$ (44)

$$\begin{aligned} \mathbf{T}_{\mathrm{L_i}}^{\partial\mathcal{V}} : \mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp) &\longrightarrow \mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^\perp) \\ \mathrm{h_i}(\mathbf{s}, \omega_\mathrm{i}) &\mapsto \mathrm{h}'_\mathrm{i}(\mathbf{s}', \omega'_\mathrm{i}) = (\mathbf{T}_{\mathrm{L_i}}^{\partial\mathcal{V}} \mathrm{h_i})(\mathbf{s}', \omega'_\mathrm{i}) \end{aligned}$$

is defined by

$$\mathbf{T}_{\mathrm{L_i}}^{\partial\mathcal{V}} \stackrel{\mathrm{def}}{=} \mathbf{G}^{\partial\mathcal{V}} \mathbf{K}^{\partial\mathcal{V}}, \tag{5.22}$$

$\mathbf{G}^{\partial\mathcal{V}}$ (430) $\mathbf{K}^{\partial\mathcal{V}}$ (432)

where $\mathbf{G}^{\partial\mathcal{V}}$ *is the light propagation operator in a vacuum and* $\mathbf{K}^{\partial\mathcal{V}}$ *is the local light scattering operator, see Figure 5.6*

REMARK 5.3 *Applied to an incident radiance function* $\mathrm{L_i}$, $\mathbf{T}_{\mathrm{L_i}}^{\partial\mathcal{V}}$ *returns the incident function* $\mathbf{T}_{\mathrm{L_i}}^{\partial\mathcal{V}} \mathrm{L_i}$, *that is: The light transport operator simulates a single scattering step followed by a single light propagation step.*

THE LIGHT TRANSPORT OPERATOR EQUATION IN A VACUUM. Combining the Expressions (5.15) and (5.21) results in the following operator formulation for the SLTEV expressed in incident radiance

$$\begin{aligned} \mathrm{L_i} &= \mathbf{G}^{\partial\mathcal{V}} \mathrm{L}_{e,\mathrm{o}} + \mathbf{G}^{\partial\mathcal{V}} \mathbf{K}^{\partial\mathcal{V}} \mathrm{L_i} && (5.23) \\ &\stackrel{(5.15)}{=} \mathrm{L}_{e,\mathrm{i}} + \mathbf{G}^{\partial\mathcal{V}} \mathbf{K}^{\partial\mathcal{V}} \mathrm{L_i}. && (5.24) \end{aligned}$$

With the source function $\mathrm{L}_{e,\mathrm{i}}$ and the definition of the light transport operator, we are now ready to represent the stationary light transport equation in a vacuum from Equation (5.12) in a more simpler form, namely as a linear operator equation:

Linear Operator Equation (61)

DEFINITION 5.4 (The Incident Light Transport Operator Equation in a Vacuum) *Based on the light propagation operator* $\mathbf{G}^{\partial\mathcal{V}}$ *and the local light scattering operator* $\mathbf{K}^{\partial\mathcal{V}}$, *the* incident light transport operator equation in a vacuum *associated with the SLTEV is given by*

$\mathbf{G}^{\partial\mathcal{V}}$ (430) $\mathbf{K}^{\partial\mathcal{V}}$ (432) SLTEV (398)

$$\mathrm{L_i} = \mathrm{L}_{e,\mathrm{i}} + \mathbf{T}_{\mathrm{L_i}}^{\partial\mathcal{V}} \mathrm{L_i}. \tag{5.25}$$

Recall, solving the light transport equation in a vacuum is equivalent to finding a solution of the light transport operator equation from Definition 5.4. Under the condition that the light transport operator is contracting, i.e. $\|\mathbf{T}_{\mathrm{L_i}}^{\partial\mathcal{V}}\| < 1$, then it holds—due to our discussion about the solution of linear operator equations:

Linear Operator Equation (61)

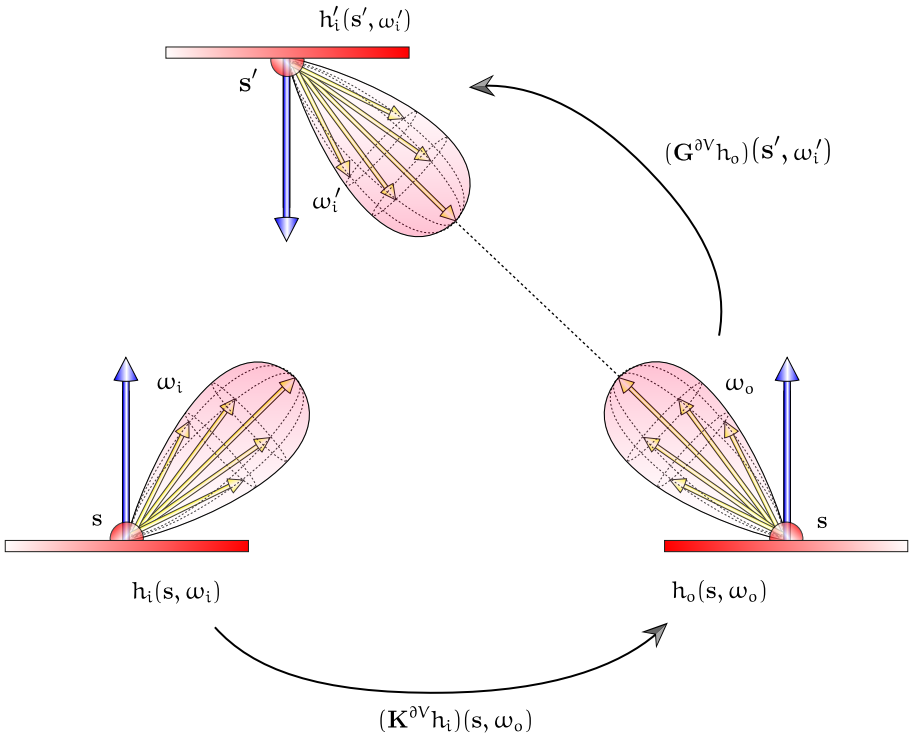


FIGURE 5.6: THE LIGHT TRANSPORT OPERATOR IN A VACUUM. The local light scattering operator $\mathbf{T}_{L_i}^{\partial V}$ is a composition of the local light scattering operator $\mathbf{K}^{\partial V}$ and the light propagation operator $\mathbf{G}^{\partial V}$, that is, an incident function $h_i(s, \omega_i)$ is mapped via $\mathbf{K}^{\partial V}$ to an exitant function $h_o(s, \omega_o)$. The light propagation operator then maps this function again to an incident function $h'_i(s', \omega'_i)$.

$$
\begin{aligned}
L_i &= L_{e,i} + \mathbf{T}_{L_i}^{\partial V} L_i && (5.26)\\
L_i - \mathbf{T}_{L_i}^{\partial V} L_i &= L_{e,i} && (5.27)\\
(\mathbf{I} - \mathbf{T}_{L_i}^{\partial V}) L_i &= L_{e,i} && (5.28)\\
L_i &= \underbrace{(\mathbf{I} - \mathbf{T}_{L_i}^{\partial V})^{-1}}_{\stackrel{def}{=} \mathbf{S}_{L_i}^{\partial V}} L_{e,i}, && (5.29)
\end{aligned}
$$

that is,

$$
\mathbf{S}_{L_i}^{\partial V} L_{e,i} = (\mathbf{I} - \mathbf{T}_{L_i}^{\partial V})^{-1} L_{e,i} \tag{5.30}
$$

is the exact solution, which we are seeking. According to [221, Veach 1998], we call $\mathbf{S}_{L_i}^{\partial V}$

the *solution operator* of the stationary light transport equation in a vacuum.

Measurement Equation (416) Replacing the incident radiance in the measurement equation from Relation (5.3) by
Neumann Series (135) the product of the solution operator $\mathbf{S}_{L_i}^{\partial\mathcal{V}}$ of the Neumann series and the emitted radiance $L_{e,i}$, then the measurement equation can be written as:

$$\mathcal{M} \stackrel{def}{=} \langle W_e, L_i \rangle \tag{5.31}$$

$$\stackrel{(5.29)}{=} \langle W_e, \mathbf{S}_{L_i}^{\partial\mathcal{V}} L_{e,i} \rangle. \tag{5.32}$$

REMARK 5.4 *From Equation (2.387) we know that the solution operator* $\mathbf{S}_{L_i}^{\partial\mathcal{V}}$ *can be*
Neumann Series (135) *expressed in form of a Neumann series, namely:*

$$L_i = \sum_{i=0}^{\infty} \mathbf{T}_{L_i}^{\partial\mathcal{V}^i} L_{e,i} \tag{5.33}$$

$$= L_{e,i} + \mathbf{T}_{L_i}^{\partial\mathcal{V}} L_{e,i} + \mathbf{T}_{L_i}^{\partial\mathcal{V}^2} L_{e,i} + \mathbf{T}_{L_i}^{\partial\mathcal{V}^3} L_{e,i} + \dots \tag{5.34}$$

$$= L_{e,i} + \mathbf{G}^{\partial\mathcal{V}}\mathbf{K}^{\partial\mathcal{V}} L_{e,i} + (\mathbf{G}^{\partial\mathcal{V}}\mathbf{K}^{\partial\mathcal{V}})^2 L_{e,i} + (\mathbf{G}^{\partial\mathcal{V}}\mathbf{K}^{\partial\mathcal{V}})^3 L_{e,i} + \dots. \tag{5.35}$$

Obviously, the light transport in a vacuum can be interpreted as composed of two physical processes: Propagation of light between surfaces and scattering of light at surfaces, that is, light incident at a scene point comes directly from surface emitters, $L_{e,i}$*, or indirectly via multiple scattering at surface points,* $\mathbf{T}_{L_i}^{i}$ *for* $i \geq 1$*. Hence, for*
Ray Space (44) *describing the distribution of light in ray space it suffices to determine the amount of light emitted from existing sources and to formulate the light transport operator* $\mathbf{T}_{L_i}^{\partial\mathcal{V}}$*.*

Let us finish this section with an example that illustrates the derivation of an operator equation for the light transport in an idealized scene, a self-emitting sphere composed of a diffuse material.

EXAMPLE 5.1 *Given be a self-emitting sphere composed of diffuse material. Then, the*
BSDF (371) *reflection behavior of the sphere can be described by a constant BSDF* $\frac{\rho_{dh}}{\pi}$*. The incident radiance at surface point* s_j *on the sphere can be computed by solving the*
SLTEV (398) *SLTEV, that is,*

$$L_i(s_j \to s_{j-1}) = L_e(s_j \to s_{j-1}) + \int_{\partial\mathcal{V}\times\partial\mathcal{V}} f_s(s_{j+1} \to s_j \to s_{j-1})\, L_i(s_{j+1} \to s_j)\, \mathcal{G}(s_{j+1} \leftrightarrow s_j)\, d\mu^2(s_{j+1}). \tag{5.36}$$

see Figure 5.7.

Due to the fact that in a sphere all points are visible to each other it holds

$$L_i(s_j \to s_{j-1}) = L_e(s_j \to s_{j-1}) + \frac{\rho_{dh}}{\pi} \int_{\partial\mathcal{V}\times\partial\mathcal{V}} L_i(s_{j+1} \to s_j) \frac{|\cos\theta_o^{j+1} \cos\theta_i^j|}{\|s_{j+1} - s_j\|_2^2}\, d\mu^2(s_{j+1}). \tag{5.37}$$

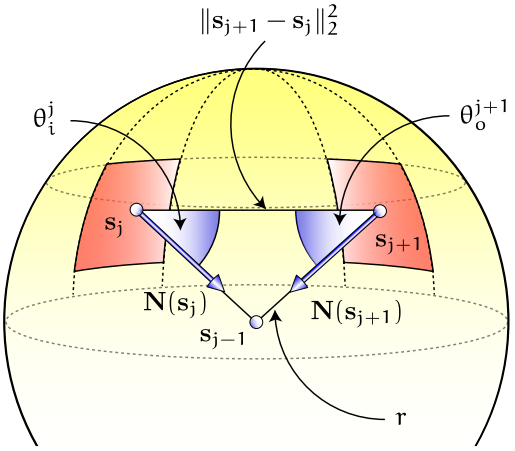


FIGURE 5.7: THE GEOMETRY OF A DIFFUSE SELF-EMITTED SPHERE FOR COMPUTING RADIANCE.

Obviously, we have: $\cos\theta_o^{j+1} = \cos\theta_i^j = \frac{\|s_{j+1}-s_j\|_2^2}{2R}$, *that is,*

$$L_i(s_j \to s_{j-1}) = L_e(s_j \to s_{j-1}) + \frac{\rho_{dh}}{4\pi R^2}\int_{\partial\mathcal{V}\times\partial\mathcal{V}} L_i(s_{j+1} \to s_j)\, d\mu^2(s_{j+1}). \quad (5.38)$$

With, $\mathbf{G}^{\partial\mathcal{V}} \equiv 1$ *and* $\mathbf{K}^{\partial\mathcal{V}} = \frac{\rho_{dh}}{4\pi R^2}$, *Equation (5.38) can be written as an operator equation of the form*

$$L_i(s_j \to s_{j-1}) = L_e(s_j \to s_{j-1}) + \sum_{j=1}^{\infty}\left(\frac{\rho_{dh}}{4\pi R^2}\right)^j L_e(s_{j+1} \to s_j) \quad (5.39)$$

$$\overset{|\frac{\rho_{dh}}{4\pi R^2}|<1}{=} L_e(s_j \to s_{j-1}) + \frac{\rho_{dh}}{4\pi R^2} L_e(s_{j+1} \to s_j) + \quad (5.40)$$

$$\left(\frac{\rho_{dh}}{4\pi R^2}\right)^2 L_e(s_{j+2} \to s_{j+1}) + \ldots \quad (5.41)$$

$$\overset{L_e = C}{=} \frac{1}{1 - \frac{\rho_{dh}}{4\pi R^2}} C.$$

5.1.2 AN OPERATOR MODEL FOR LIGHT TRANSPORT IN PARTICIPATING MEDIA

Based on the discussion of the derivation of a linear operator equation for the light trans- Linear Operator Equation (61)

port in a vacuum in Section 5.1.1.2, now we devote our interest to the development of a linear operator equation for the case of light transport in participating media.

For that purpose, we will split the stationary light transport equation valid in participating media into four terms: two emission terms, one for emission at object surfaces and one for emission at particles within participating media, and two scattering terms, even for surfaces and particles within media, where each emission and scattering term is combined with an absorption term. With the light propagation operator $\mathbf{G}^{\partial\mathcal{V}}$, the local light scattering operator $\mathbf{K}^{\partial\mathcal{V}}$ from our discussion about an operator model for light transport within a vacuum, and the construction of two new operators $\overline{\mathbf{G}}^{\mathcal{V}^o}$ and $\overline{\mathbf{K}}^{\mathcal{V}^o}$, that describe the light transport only in participating media, then we can formulate the SLTE as a linear operator equation. Expressed as a Neumann series, this operator equation can be solved as it is done in the previous section for the operator equation underlying the stationary light transport equation within a vacuum.

5.1.2.1 THE LIGHT PROPAGATION AND THE LIGHT SCATTERING OPERATOR IN PARTICIPATING MEDIA

$\mathbf{G}^{\partial\mathcal{V}}$ (430) $\mathbf{K}^{\partial\mathcal{V}}$ (432) SLTE (296)

Let us recall the SLTE from Equation (5.4). Using the light propagation operator in vacuum, $\mathbf{G}^{\partial\mathcal{V}}$, and the local light scattering operator, $\mathbf{K}^{\partial\mathcal{V}}$, then the SLTE can be written, slightly rephrased in the following mixed operator-integral equation formulation:

$$\begin{aligned} &L_i(\mathbf{x},\omega_i) \\ &= \underbrace{\beta(\mathbf{s}\to\mathbf{x})}_{\text{volume attenuation}} \underbrace{\left(\mathbf{G}^{\partial\mathcal{V}}L_{e,o}+\mathbf{T}^{\partial\mathcal{V}}_{L_i}L_i\right)}_{\text{SLTEV operator equation}} + \\ &\int_{[0,\|\mathbf{x}'-\mathbf{x}\|]} \underbrace{\beta(\mathbf{x}'\to\mathbf{x})}_{\text{volume attenuation}} \underbrace{L_e(\mathbf{x}',\omega_o)d\mu(\alpha)}_{\text{emission within volume}} + \\ &\int_{[0,\|\mathbf{x}'-\mathbf{x}\|]} \underbrace{\beta(\mathbf{x}'\to\mathbf{x})}_{\text{volume attenuation}} \underbrace{\sigma_s(\mathbf{x}')\int_{S^2(\mathbf{x}')} p(\mathbf{x}',-\omega_i'\to\omega_o)\,L_i(\mathbf{x}',\omega_i')d\sigma_{\mathbf{x}'}(\omega_i')d\mu(\alpha)}_{\text{scattering within volume}}, \end{aligned} \tag{5.42}$$

where $\mathbf{G}^{\partial\mathcal{V}}L_{e,o}$ and $\mathbf{T}^{\partial\mathcal{V}}_{L_i}L_i$ are square-integrable functions from $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}},\zeta^\perp)$.

Obviously, the product of the terms $\beta(\mathbf{s}\to\mathbf{x})$ and $\left(\mathbf{G}^{\partial\mathcal{V}}L_{e,o}+\mathbf{T}^{\partial\mathcal{V}}_{L_i}L_i\right)$ on the right hand side can be interpreted as the light transport operator equation in a vacuum attenuated by the path absorption function, β, while the two integrals describe the light emission and scattering processes in volumetric points of a medium combined with absorption processes.

Our goal in this section is to express the *emission-within-volume-term* and the *scattering-within-volume-term* of the SLTE from Equation (5.4) by two corresponding linear operators, $\overline{\mathbf{G}}^{\mathcal{V}^o}$ and $\overline{\mathbf{K}}^{\mathcal{V}^o}$, valid in participating media. Combining these two operators with the light propagation operator $\mathbf{G}^{\partial\mathcal{V}}$ and the local light scattering operator $\mathbf{K}^{\partial\mathcal{V}}$ from

our discussion of an operator model for light transport within a vacuum then leads to an operator equation for the SLTE of type

$$L_i = \beta(\mathbf{s} \to \mathbf{x}) \left(\mathbf{G}^{\partial\mathcal{V}} L_{e,o} + \mathbf{T}^{\partial\mathcal{V}}_{L_i} L_i \right) + \overline{\mathbf{G}}^{\mathcal{V}^o} L_{e,o} + \overline{\mathbf{T}}^{\mathcal{V}^o}_{L_i} L_i, \tag{5.43}$$

with the transport operator $\overline{\mathbf{T}}^{\mathcal{V}^o}_{L_i} = \overline{\mathbf{G}}^{\mathcal{V}^o} \overline{\mathbf{K}}^{\mathcal{V}^o}$, valid within participating media.

Now, since $\mathbf{G}^{\partial\mathcal{V}}$ and $\mathbf{K}^{\partial\mathcal{V}}$ are only valid in a vacuum, we have to extend these two operators by the path absorption function β, which describes absorption within participating media. As a result, we get a so-called *light propagation operator*, $\overline{\mathbf{G}}^{\mathcal{V}^o}$, and a *local light scattering operator in participating media*, $\overline{\mathbf{K}}^{\mathcal{V}^o}$. Together with their analogues from light transport in a vacuum, these constructs then allow to formulate the SLTE from Relation (5.42) as a linear operator equation of the same form as it is known from the last section, namely,

Path Absorption Function (292)

$$\begin{aligned}
L_i &= \beta(\mathbf{s} \to \mathbf{x}) \left(\mathbf{G}^{\partial\mathcal{V}} L_{e,o} + \mathbf{T}^{\partial\mathcal{V}}_{L_i} L_i \right) + \overline{\mathbf{G}}^{\mathcal{V}^o} L_{e,o} + \overline{\mathbf{T}}^{\mathcal{V}^o}_{L_i} L_i && (5.44)\\
&= \beta(\mathbf{s} \to \mathbf{x}) \mathbf{G}^{\partial\mathcal{V}} L_{e,o} + \overline{\mathbf{G}}^{\mathcal{V}^o} L_{e,o} + \beta(\mathbf{s} \to \mathbf{x}) \mathbf{T}^{\partial\mathcal{V}}_{L_i} L_i + \overline{\mathbf{T}}^{\mathcal{V}^o}_{L_i} L_i && (5.45)\\
&= \underbrace{\left(\beta(\mathbf{s} \to \mathbf{x}) \mathbf{G}^{\partial\mathcal{V}} + \overline{\mathbf{G}}^{\mathcal{V}^o} \right) L_{e,o}}_{L_{e,i} = \overline{\mathbf{G}} L_{e,o}} + \underbrace{\left(\beta(\mathbf{s} \to \mathbf{x}) \mathbf{T}^{\partial\mathcal{V}}_{L_i} + \overline{\mathbf{T}}^{\mathcal{V}^o}_{L_i} \right) L_i}_{\overline{\mathbf{T}}_{L_i} L_i} && (5.46)\\
&= L_{e,i} + \overline{\mathbf{T}}_{L_i} L_i. && (5.47)
\end{aligned}$$

The goal of this and the following section is the exact derivation of the operator equation 5.47 and all needed mathematical constructs. We begin with the derivation of the light propagation operators and the local light scattering operators in participating media.

THE LIGHT PROPAGATION OPERATORS IN PARTICIPATING MEDIA. For modifying the light propagation operator $\mathbf{G}^{\partial\mathcal{V}}$ and the local light scattering operator $\mathbf{K}^{\partial\mathcal{V}}$ such that they are also valid in a participating medium, in a first step, we have to adapt the underlying function spaces, and then we have to extend these operators by a so-called *multiplication operator* given by the path absorption function β.

Absorption Function (292)

Recall, the light propagation operator $\mathbf{G}^{\partial\mathcal{V}}$ and the local light scattering operator $\mathbf{K}^{\partial\mathcal{V}}$ are linear mappings between the space $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta)$, that is, the space of square Lebesgue-integrable functions defined on the ray space $\mathcal{R}^{\partial\mathcal{V}}$. Now, apart from the interaction of light at surfaces from $\partial\mathcal{V}$, the light transport in scenes with participating media also takes into account the interaction of light at small particles from $\mathcal{V}^o$. This means, that the emitted as well as the incident radiance functions in the SLTE are all together functions defined on the extended ray space $\mathcal{R}$, that is, functions from the Lebesgue space $\mathcal{L}^2(\mathcal{R}, \overline{\zeta})$, thus,

$\mathcal{L}^2(\mathcal{R}, \overline{\zeta})$ (112)

$$\mathcal{L}^2(\mathcal{R}, \overline{\zeta}) \stackrel{\text{def}}{=} \left\{ f \in \mathcal{L}(\mathcal{R}) \,\middle|\, \|f\|_{\mathcal{L}^2} < \infty,\ f = f^{\partial\mathcal{V}} + f^{\mathcal{V}^o} \right\}, \tag{5.48}$$

where $f^{\partial\mathcal{V}}(\mathbf{x},\omega) \equiv 0$ for $\mathbf{x} \in \mathcal{R}\setminus\partial\mathcal{V}$, $f^{\mathcal{V}^o} \equiv 0$ for $\mathbf{x} \in \mathcal{R}\setminus\mathcal{V}^o$ and the volume $\mathcal{V} = \mathcal{V}^o \cup \partial\mathcal{V}$.

We can now introduce a so-called surface light propagation operator in participating media which describes the flow of light from surfaces passing participating media.

DEFINITION 5.5 (The Surface Light Propagation Operator in Participating Media) *Given* $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ (112) *be the space of square Lebesgue-integrable functions,* $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$*, defined on the ray* $\mathcal{R}$ (44) *space* $\mathcal{R}$*. Then, the* surface light propagation operator in participating media

$$\begin{aligned} \overline{\mathbf{G}}^{\partial\mathcal{V}} : \mathcal{L}^2(\mathcal{R},\overline{\zeta}) &\longrightarrow \mathcal{L}^2(\mathcal{R},\overline{\zeta}) \\ h_o(\mathbf{x},\omega_o) &\mapsto h_i(\mathbf{x},\omega_i) = (\overline{\mathbf{G}}^{\partial\mathcal{V}} h_o)(\mathbf{x},\omega_i) \end{aligned}$$

is defined by

$$(\overline{\mathbf{G}}^{\partial\mathcal{V}} h_o)(\mathbf{x},\omega_i) \stackrel{def}{=} \begin{cases} \beta(\mathbf{s}\to\mathbf{x})\, h_o(\gamma(\mathbf{x},\omega_i),-\omega_i) & \text{if } d_{\partial\mathcal{V}}(\mathbf{x},\omega_i) < \infty \\ 0 & \text{else,} \end{cases} \tag{5.49}$$

γ (47) *where* $\mathbf{x} \in \mathcal{V}$*,* $\mathbf{s} = \gamma(\mathbf{x},\omega_i) \in \partial\mathcal{V}$*, and* $\omega_o = -\omega_i$*.*

REMARK 5.5 *The surface light propagation operator* $\overline{\mathbf{G}}^{\partial\mathcal{V}}$ *can be interpreted as it maps an exitant function* h_o *to an incident function* $h_i = \overline{\mathbf{G}}^{\partial\mathcal{V}} h_o$ *attenuated by the value of the path absorption function* β*. Considered a little bit closer, you can say that* $\overline{\mathbf{G}}^{\partial\mathcal{V}}$ *operates only on the* surface-component $h_o^{\partial\mathcal{V}}$ *of the function* h_o*. Due to its definition, it holds*

$$\overline{\mathbf{G}}^{\partial\mathcal{V}} \underbrace{h_o}_{\in\mathcal{L}^2(\mathcal{R},\overline{\zeta})} = \overline{\mathbf{G}}^{\partial\mathcal{V}} \underbrace{\left(h_o^{\partial\mathcal{V}} + h_o^{\mathcal{V}^o}\right)}_{\in\mathcal{L}^2(\mathcal{R},\overline{\zeta})} \tag{5.50}$$

$$\stackrel{(5.49)}{=} \overline{\mathbf{G}}^{\partial\mathcal{V}} h_o^{\partial\mathcal{V}} + \underbrace{\overline{\mathbf{G}}^{\partial\mathcal{V}} h_o^{\mathcal{V}^o}}_{\equiv 0} \tag{5.51}$$

$$= \underbrace{\overline{\mathbf{G}}^{\partial\mathcal{V}} h_o^{\partial\mathcal{V}}}_{\in\mathcal{L}^2(\mathcal{R},\overline{\zeta})}. \tag{5.52}$$

Obvioulsy, $\overline{\mathbf{G}}^{\partial\mathcal{V}}$ *maps the* volume-component $h_o^{\mathcal{V}^o}$ *of* h_o *to zero. Although it only operates on the component* $h_o^{\partial\mathcal{V}}$ *of* h_o*, nevertheless it returns a function* $h_i^{\partial\mathcal{V}} + h_i^{\mathcal{V}^o} \in \mathcal{L}^2(\mathcal{R},\overline{\zeta})$*.*

$\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ (112) EXAMPLE 5.2 *Let us consider an exitant radiance function* $L_e \in \mathcal{L}^2(\mathcal{R},\overline{\zeta})$*. If it hold* $L_e(\mathbf{x},\omega_o) \neq 0$ *for a finite set of points from* $\mathcal{V}$ *and directions from* S^2*, then the function* L_e *can be used to describe point light sources on surfaces and within participating*

media. Applying the operator $\overline{\mathbf{G}}^{\partial\mathcal{V}}$ *to* L_e *leads to*

$$\underbrace{\overline{\mathbf{G}}^{\partial\mathcal{V}}\mathrm{L}_e}_{\in\mathcal{L}^2(\mathcal{R},\overline{\zeta})} = \overline{\mathbf{G}}^{\partial\mathcal{V}}\left(\mathrm{L}_e^{\partial\mathcal{V}} + \mathrm{L}_e^{\mathcal{V}^o}\right) \tag{5.53}$$

$$\overset{(5.49)}{=} \overline{\mathbf{G}}^{\partial\mathcal{V}}\mathrm{L}_e^{\partial\mathcal{V}} + \underbrace{\overline{\mathbf{G}}^{\partial\mathcal{V}}\mathrm{L}_e^{\mathcal{V}^o}}_{\equiv 0} \tag{5.54}$$

$$= \underbrace{\overline{\mathbf{G}}^{\partial\mathcal{V}}\mathrm{L}_e^{\partial\mathcal{V}}}_{\in\mathcal{L}^2(\mathcal{R},\overline{\zeta})}, \tag{5.55}$$

that is, the operator $\overline{\mathbf{G}}^{\partial\mathcal{V}}$ *propagates the light coming from point light sources on surfaces to points in* $\mathcal{V}$ *that can be reached on straight lines from the surface point light sources. It does not take into account the light emitted from point light sources in participating media.*

Apart from the process of light propagation at surfaces, the light transport within participating media is also characterized by light that comes from scattering or emission events at volumetric points. Also this process can mathematically be captured by a linear operator, the so-called *volume light propagation operator in participating media*, $\overline{\mathbf{G}}^{\mathcal{V}^o}$, given on the Lebesgue space $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$.

DEFINITION 5.6 (The Volume Light Propagation Operator in Participating Media) *Given be the space of square Lebesgue-integrable functions,* $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$*, defined on ray space* $\mathcal{R}$*. Then, the* volume light propagation operator in participating media

$\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ (112)
$\mathcal{R}$ (44)

$$\begin{aligned}\overline{\mathbf{G}}^{\mathcal{V}^o} : \mathcal{L}^2(\mathcal{R},\overline{\zeta}) &\longrightarrow \mathcal{L}^2(\mathcal{R},\overline{\zeta})\\ h_o(\mathbf{x},\omega_o) &\mapsto h_i(\mathbf{x},\omega_i) = (\overline{\mathbf{G}}^{\mathcal{V}^o}h_o)(\mathbf{x},\omega_i)\end{aligned}$$

is defined by

$$(\overline{\mathbf{G}}^{\mathcal{V}^o}h_o)(\mathbf{x},\omega_i) \overset{\mathrm{def}}{=} \begin{cases}\int_{[0,\|\mathbf{x}'-\mathbf{x}\|]}\beta(\mathbf{x}'\to\mathbf{x})\,h_o(\mathbf{x}',-\omega_i)\,\mathrm{d}\mu(\alpha) & \text{if } \mathbf{x}'=\mathbf{x}+\alpha\omega_i\in\mathcal{V}^o\\ 0 & \text{else,}\end{cases} \tag{5.56}$$

where $\mathbf{x}'$ *is a point within a medium lying on the line from* $\mathbf{x}\in\mathcal{V}$ *to the closest point* $\mathbf{s}=\gamma(\mathbf{x},\omega_i)$ *on a surface.*

γ (47)

REMARK 5.6 *The volume light propagation operator* $\overline{\mathbf{G}}^{\mathcal{V}^o}$ *can be interpreted as it maps an exitant function* h_o *to an incident function* $h_i=\overline{\mathbf{G}}^{\partial\mathcal{V}}h_o$ *attenuated by the value of the path absorption function* β*. Contrary to the operator* $\overline{\mathbf{G}}^{\partial\mathcal{V}}$*, which only operates on the surface-component* $h_o^{\partial\mathcal{V}}$ *of* h_o*, the volume light propagation operator* $\overline{\mathbf{G}}^{\mathcal{V}^o}$ *only*

operates on the volume-component $h_o^{\mathcal{V}^o}$ *of* h_o*, that is, applied to the exitant function* h_o*, we get:*

$$\overline{\mathbf{G}}^{\mathcal{V}^o} \underbrace{h_o}_{\in\mathcal{L}^2(\mathcal{R},\overline{\zeta})} = \overline{\mathbf{G}}^{\mathcal{V}^o} \underbrace{\left(h_o^{\partial\mathcal{V}} + h_o^{\mathcal{V}^o}\right)}_{\in\mathcal{L}^2(\mathcal{R},\overline{\zeta})} \tag{5.57}$$

$$\overset{(5.56)}{=} \underbrace{\overline{\mathbf{G}}^{\mathcal{V}^o} h_o^{\partial\mathcal{V}}}_{\equiv 0} + \overline{\mathbf{G}}^{\mathcal{V}^o} h_o^{\mathcal{V}^o} \tag{5.58}$$

$$= \underbrace{\overline{\mathbf{G}}^{\mathcal{V}^o} h_o^{\mathcal{V}^o}}_{\in\mathcal{L}^2(\mathcal{R},\overline{\zeta})}. \tag{5.59}$$

Thus, $\overline{\mathbf{G}}^{\mathcal{V}^o}$ *maps the* surface-component $h_o^{\partial\mathcal{V}}$ *of* h_o *to zero. Although it only operates on the component* $h_o^{\mathcal{V}^o}$ *of* h_o*, nevertheless it returns a function* $h_i^{\partial\mathcal{V}} + h_i^{\mathcal{V}^o} \in \mathcal{L}^2(\mathcal{R},\overline{\zeta})$.

EXAMPLE 5.3 *Let us consider an exitant radiance function* $L_e \in \mathcal{L}^2(\mathcal{R},\overline{\zeta})$ *once more, where* $L_e(\mathbf{x},\omega_o) \neq 0$ *for a set of points from* $\mathcal{V}$ *and directions from* S^2*. Applying the operator* $\overline{\mathbf{G}}^{\mathcal{V}^o}$ *to* L_e *leads to*

$\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ (112)

$$\underbrace{\overline{\mathbf{G}}^{\mathcal{V}^o} L_e}_{\in\mathcal{L}^2(\mathcal{R},\overline{\zeta})} = \overline{\mathbf{G}}^{\mathcal{V}^o}\left(L_e^{\partial\mathcal{V}} + L_e^{\mathcal{V}^o}\right) \tag{5.60}$$

$$\overset{(5.5)}{=} \underbrace{\overline{\mathbf{G}}^{\mathcal{V}^o} L_e^{\mathcal{V}^o}}_{\in\mathcal{L}^2(\mathcal{R},\overline{\zeta})}, \tag{5.61}$$

that is, the operator $\overline{\mathbf{G}}^{\mathcal{V}^o}$ *propagates the light coming from point light sources within participating media to points in* $\mathcal{V}$ *that can be reached on straight lines. It does not take into account the light emitted from point light sources lying on object surfaces.*

Based on these new concepts, we are now able to define the *light propagation operator in participating media* as sum of the surface light propagation operator and the volume light propagation operator.

DEFINITION 5.7 (The Light Propagation Operator in Participating Media) *Given be the space of square Lebesgue-integrable functions,* $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$*, defined on the ray space* $\mathcal{R}$*. Then, the* light propagation operator in participating media

$\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ (112), $\mathcal{R}$ (44)

$$\begin{aligned} \overline{\mathbf{G}} : \mathcal{L}^2(\mathcal{R},\overline{\zeta}) &\longrightarrow \mathcal{L}^2(\mathcal{R},\overline{\zeta}) \\ h_o(\mathbf{x},\omega_o) &\mapsto h_i(\mathbf{x},\omega_i) = \left(\overline{\mathbf{G}} h_o\right)(\mathbf{x},\omega_i) \end{aligned}$$

is defined as the sum of the surface light propagation operator $\overline{\mathbf{G}}^{\partial\mathcal{V}}$ *and the* volume

$\overline{\mathbf{G}}^{\partial\mathcal{V}}$ (440)

$\overline{\mathbf{G}}^{\mathcal{V}^o}$ (441) light propagation operator $\overline{\mathbf{G}}^{\mathcal{V}^o}$, *that is, for any* $\mathbf{x} \in \mathcal{V}$ *it is given by*

$$(\overline{\mathbf{G}} h_o)(\mathbf{x}, \omega_i) \stackrel{\text{def}}{=} (\overline{\mathbf{G}}^{\partial\mathcal{V}} h_o)(\mathbf{x}, \omega_i) + (\overline{\mathbf{G}}^{\mathcal{V}^o} h_o)(\mathbf{x}, \omega_i). \tag{5.62}$$

REMARK 5.7 *Applied to an exitant radiance function* L_o*, the operator* $\overline{\mathbf{G}}$ *returns the incident radiance function* L_i *resulting from the propagation of light from points* $\mathbf{s}$ *on object surfaces and from all points* $\mathbf{x}'$ *on the line to* $\mathbf{s}$ *within participating media, thus* $L_i = \overline{\mathbf{G}} L_o$*. If the function* L_o *measures photons exitant from points of* $\mathcal{V} = \partial\mathcal{V} \cup \mathcal{V}^o$*, the function* L_i *obviously returns the fraction of these photons that propagates through the medium.*

Let us show how the concept of the light propagation operator in participating media works when applied to an exitant radiance function

EXAMPLE 5.4 (Emission in Participating Media Expressed in Incident Radiance) *Let us suppose that the emitted radiance in all scene points* $\mathbf{x} \in \mathcal{V}$ *is given by* $L_{e,o}(\mathbf{x}, \omega_o)$*. Applying the light propagation operator in participating media to* $L_{e,o}(\mathbf{x}, \omega_o)$ *leads to*

$$\begin{aligned}
(\overline{\mathbf{G}} L_{e,o})(\mathbf{x}, \omega_i) &\stackrel{\text{def}}{=} ((\overline{\mathbf{G}}^{\partial\mathcal{V}} + \overline{\mathbf{G}}^{\mathcal{V}^o}) L_{e,o})(\mathbf{x}, \omega_i) && (5.63)\\
&= (\overline{\mathbf{G}}^{\partial\mathcal{V}} L_{e,o})(\mathbf{x}, \omega_i) + (\overline{\mathbf{G}}^{\mathcal{V}^o} L_{e,o})(\mathbf{x}, \omega_i) && (5.64)\\
&= \underbrace{\beta(\mathbf{s} \to \mathbf{x}) L_{e,o}(\mathbf{s}, \omega_o)}_{(\overline{\mathbf{G}}^{\partial\mathcal{V}} L_{e,o})(\mathbf{x}, \omega_i)} + \\
&\quad \underbrace{\int_{[0, \|\mathbf{x}' - \mathbf{x}\|]} \beta(\mathbf{x}' \to \mathbf{x})\, L_{e,o}(\mathbf{x}', \omega_o) \mathrm{d}\mu(\alpha)}_{(\overline{\mathbf{G}}^{\mathcal{V}^o} L_{e,o})(\mathbf{x}, \omega_i)}, && (5.65)
\end{aligned}$$

where $\mathbf{s}$ *is the closest visible surface point reachable from* $\mathbf{x}$ *in direction* ω_i*, and* $\mathbf{x}'$ *are all points lying on the line starting in* $\mathbf{x}$ *and passing through the medium.*

Obviously, the above equation describes the emission processes of the SLTE from (5.4), at surface and volumetric points, expressed in incident radiance. Since $(\overline{\mathbf{G}} L_{e,o})$ *is an incident quantity, we use in the following discussion the identity*

SLTE (296)

$$L_{e,i}(\mathbf{x}, \omega_i) \stackrel{\text{def}}{=} (\overline{\mathbf{G}} L_{e,o})(\mathbf{x}, \omega_i). \tag{5.66}$$

If the emitted radiance is known in all scene points, then also the function $\overline{\mathbf{G}} L_{e,o}$ is known, since it can be evaluated via the path absorption function β. This means: $\overline{\mathbf{G}} L_{e,o}$ can be interpreted as the driving function of a Fredholm integral equation of the 2^{nd} kind, and thus can be used as the constant in our operator equation from (5.5).

Path Absorption Function (292)

Driving Function (127)

THE LOCAL SCATTERING OPERATOR IN PARTICIPATING MEDIA. Obviously, light passing through a medium can come from scattering events at surfaces or from scattering events

within a medium, thus, similar to the definition of the light propagation operator $\overline{\mathbf{G}}$ also the *light scattering operator* $\overline{\mathbf{K}}$ must be composed of a *surface light scattering operator* and a *volume light scattering operator*.

$\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ (112) $\mathcal{R}$ (44)

DEFINITION 5.8 (The Local Surface Light Scattering Operator in Participating Media) *Let $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ be the space of square Lebesgue-integrable functions defined on the ray space $\mathcal{R}$. Then, the* local surface light scattering operator in participating media

$$\begin{aligned}\overline{\mathbf{K}}^{\partial\mathcal{V}} : \mathcal{L}^2(\mathcal{R},\overline{\zeta}) &\longrightarrow \mathcal{L}^2(\mathcal{R},\overline{\zeta})\\ h_i(\mathbf{x},\omega_i) &\mapsto h_o(\mathbf{x},\omega_o) = (\overline{\mathbf{K}}^{\partial\mathcal{V}} h_i)(\mathbf{x},\omega_o)\end{aligned}$$

is defined by

$$(\overline{\mathbf{K}}^{\partial\mathcal{V}} h_i)(\mathbf{x},\omega_o) \stackrel{\text{def}}{=} \begin{cases} \int_{S^2(\mathbf{s})} f_s(\mathbf{s},\omega_i \to \omega_o)\, h_i(\mathbf{s},\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) & \text{if } \mathbf{s}=\mathbf{x}\in\partial\mathcal{V}\\ 0 & \text{else.}\end{cases} \tag{5.67}$$

REMARK 5.8 *The local surface light scattering operator $\overline{\mathbf{K}}^{\partial\mathcal{V}}$ can be interpreted as it maps an incident function h_i to an exitant function $h_o = \overline{\mathbf{G}}^{\mathcal{V}^o} h_o$. Like $\overline{\mathbf{G}}^{\partial\mathcal{V}}$, so $\overline{\mathbf{K}}^{\partial\mathcal{V}}$ only operates on the surface component $h_i^{\partial\mathcal{V}}$ of the incident function h_i, the volume component $h_i^{\mathcal{V}^o}$ is mapped to zero. Nevertheless, $\overline{\mathbf{G}}^{\partial\mathcal{V}}$ returns an exitant function from $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$.*

In a similar way, we now define the local volume light scattering operator in participating media, as the operator, that describes only the scattering behavior at points within a participating medium. This operator only operates on the volume-component of an incident function h_i, and maps the surface-component of h_i to zero.

$\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ (112) $\mathcal{R}$ (44)

DEFINITION 5.9 (The Local Volume Light Scattering Operator in Participating Media) *Let $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ be the space of square Lebesgue-integrable functions defined on the ray space $\mathcal{R}$. Then, the* local surface light scattering operator in participating media

$$\begin{aligned}\overline{\mathbf{K}}^{\mathcal{V}^o} : \mathcal{L}^2(\mathcal{R},\zeta) &\longrightarrow \mathcal{L}^2(\mathcal{R},\zeta)\\ h_i(\mathbf{x},\omega_i) &\mapsto h_o(\mathbf{x},\omega_o) = (\overline{\mathbf{K}}^{\mathcal{V}^o} h_i)(\mathbf{x},\omega_o)\end{aligned}$$

is defined by

$$(\overline{\mathbf{K}}^{\mathcal{V}^o} h_i)(\mathbf{x},\omega_o) \stackrel{\text{def}}{=} \begin{cases} \sigma_s(\mathbf{x}) \int_{S^2(\mathbf{x})} p(\mathbf{x},-\omega_i \to \omega_o)\, h_i(\mathbf{x},\omega_i)\, d\sigma_{\mathbf{x}}(\omega_i) & \text{if } \mathbf{x}\in\mathcal{V}\\ 0 & \text{else.}\end{cases} \tag{5.68}$$

REMARK 5.9 *The volume light scattering operator $\overline{\mathbf{K}}^{\mathcal{V}^o}$ can be interpreted as it maps an incident function h_i, to an exitant function $h_o = \overline{\mathbf{K}}^{\mathcal{V}^o} h_i$. Although this operator*

reduces the surface-component of h_i*, it returns an exitant function from the Lebesgue space* $\mathcal{L}^2(\mathcal{R}, \overline{\zeta})$.

Ultimately, we can now define the local light scattering operator within participating media as sum of the linear operators $\overline{\mathbf{K}}^{\partial\mathcal{V}}$ and $\overline{\mathbf{K}}^{\mathcal{V}^o}$.

DEFINITION 5.10 (The Local Light Scattering Operator in Participating Media) *Given be the space of square Lebesgue-integrable functions,* $\mathcal{L}^2(\mathcal{R}, \overline{\zeta})$*, defined on the ray space* $\mathcal{R}$*. Then, the* light scattering operator in participating media

$\mathcal{L}^2(\mathcal{R}, \overline{\zeta})$ (112) $\mathcal{R}$ (44)

$$\begin{aligned} \overline{\mathbf{K}} : \mathcal{L}^2(\mathcal{R}, \overline{\zeta}) &\longrightarrow \mathcal{L}^2(\mathcal{R}, \overline{\zeta}) \\ h_i(\mathbf{x}, \omega_i) &\mapsto h_o(\mathbf{x}, \omega_o) = \left(\overline{\mathbf{K}} h_i\right)(\mathbf{x}, \omega_o) \end{aligned}$$

is defined as the sum of the local surface light scattering operator $\overline{\mathbf{K}}^{\partial\mathcal{V}}$ *and the* local volume light scattering operator $\overline{\mathbf{K}}^{\mathcal{V}^o}$ *by*

$$(\overline{\mathbf{K}} h)(\mathbf{x}, \omega_o) \stackrel{def}{=} (\overline{\mathbf{K}}^{\partial\mathcal{V}} h_i)(\mathbf{x}, \omega_o) + (\overline{\mathbf{K}}^{\mathcal{V}^o} h_i)(\mathbf{x}, \omega_o). \tag{5.69}$$

REMARK 5.10 *Applied to an incident radiance function* L_i*, the operator* $\overline{\mathbf{K}}$ *returns the exitant radiance function* L_o *as result from a single scattering operation at object surfaces and within a medium, thus* $L_o = \overline{\mathbf{K}} L_i$*. If the function* L_i *measures photons just before their arrival at surfaces or volumetric points, then* L_o *measures photons after the scattering.*

Let us show how the concept of the local light scattering operator in participating media works when applied to an incident radiance function

EXAMPLE 5.5 (Scattering in Participating Media Expressed in Incident Radiance) *Let us assume, that* $L_i(\mathbf{x}, \omega_i)$ *denotes the incident radiance at scene points* $\mathbf{x}$ *from direction* ω_i*. Applying the local light scattering operator in participating media to* $L_i(\mathbf{x}, \omega_i)$*, then we get:*

$$\begin{aligned} (\overline{\mathbf{K}} L_i)(\mathbf{x}, \omega_i) &\stackrel{(5.69)}{=} ((\overline{\mathbf{K}}^{\partial\mathcal{V}} + \overline{\mathbf{K}}^{\mathcal{V}^o}) L_i)(\mathbf{x}, \omega_o) && (5.70) \\ &= (\overline{\mathbf{K}}^{\partial\mathcal{V}} L_i)(\mathbf{x}, \omega_o) + (\overline{\mathbf{K}}^{\mathcal{V}^o} L_i)(\mathbf{x}, \omega_o) && (5.71) \\ &\stackrel{(5.67),(5.68)}{=} \int_{S^2(\mathbf{s})} f_s(\mathbf{x}, \omega_i \to \omega_o)\, L_i(\mathbf{x}, \omega_i)\, d\sigma^{\perp}_{\mathbf{x}}(\omega_i) + && (5.72) \\ &\qquad \sigma_s(\mathbf{x}) \int_{S^2(\mathbf{x})} p(\mathbf{x}, -\omega_i \to \omega_o)\, L_i(\mathbf{x}, \omega_i)\, d\sigma_{\mathbf{x}}(\omega_i). \end{aligned}$$

Obvioulsy the above equation describes the scattering processes of the SLTE from Equation (5.4) expressed in terms of incident radiance.

SLTE (296)

5.1.2.2 THE LIGHT TRANSPORT OPERATOR EQUATION IN PARTICIPATING MEDIA

In Section 5.1.1.2 we have seen, that the stationary light transport within a vacuum can be described by an operator equation of type

$$L_i = L_{e,i} + \mathbf{T}_{L_i}^{\partial\mathcal{V}} L_i, \tag{5.73}$$

where $L_{e,i}$ corresponds to the emitted radiance within the scene, expressed as an incident quantity, L_i is the unknown incident radiance, and $\mathbf{T}_{L_i}^{\partial\mathcal{V}}$ is the associated light transport operator valid in a vacuum, defined on the square Lebesgue-integrable function space $\mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^{\perp})$.

Now, the structure of the SLTE from Relation (5.4) implies, that also an operator equation for the stationary light transport in participating media can be written in form of an operator equation of the type

$$L_i = L_{e,i} + \overline{\mathbf{T}}_{L_i} L_i, \tag{5.74}$$

thus, as sum of the emitted radiance within the scene and a linear transport operator valid in participating media, where $L_{e,i}$ and L_i are square Lebesgue-integrable functions from the function space $\mathcal{L}^2(\mathcal{R}, \overline{\zeta})$.

From Example 5.4 it is known that the two emission terms of the SLTE from Relation (5.4) can be expressed via the linear light propagation operator in participating media, namely by

$$L_{e,i} \overset{(5.66)}{=} \overline{\mathbf{G}} L_{e,o}. \tag{5.75}$$

Thus, to achieve our goal, the only thing that remains is to express the scattering-at-surface-term and the scattering-within-volume-term from Equation (5.4) via the linear operators $\overline{\mathbf{G}}$ and $\overline{\mathbf{K}}$. But this can be done easily: Obviously, the composition of the surface light propagation operator $\overline{\mathbf{G}}^{\partial\mathcal{V}}$ and the local surface light scattering operator $\overline{\mathbf{K}}^{\partial\mathcal{V}}$ and the application of this new operator to an incident function L_i, leads to the scattering-at-surface-term of the SLTE, thus,

$\overline{\mathbf{G}}^{\partial\mathcal{V}}$ (440)
$\overline{\mathbf{K}}^{\partial\mathcal{V}}$ (444)

$$\left(\overline{\mathbf{G}}^{\partial\mathcal{V}} \overline{\mathbf{K}}^{\partial\mathcal{V}} L_i\right)(\mathbf{s}, \omega_i) = \beta(\mathbf{s} \to \mathbf{x}) \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i' \to \omega_o)\, L_i(\mathbf{s}, \omega_i')\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i') \tag{5.76}$$

where $\omega_o = -\omega_i$. On the other side, combining $\overline{\mathbf{G}}^{\mathcal{V}^o}$ and $\overline{\mathbf{K}}^{\mathcal{V}^o}$, applied to an incident function L_i, leads to the scattering-within-volume-term from Equation (5.4), thus,

$\overline{\mathbf{G}}^{\mathcal{V}^o}$ (441)
$\overline{\mathbf{K}}^{\mathcal{V}^o}$ (444)

$$\begin{aligned}(\overline{\mathbf{G}}^{\mathcal{V}^o} \overline{\mathbf{K}}^{\mathcal{V}^o} L_i)(\mathbf{x}\omega_i) = \\ \int_{[0, \|\mathbf{x}'-\mathbf{x}\|]} \beta(\mathbf{x}' \to \mathbf{x})\, \sigma_s(\mathbf{x}') \int_{S^2(\mathbf{x}')} p(\mathbf{x}', -\omega_i' \to \omega_o)\, L_i(\mathbf{x}', \omega_i')\, d\sigma_{\mathbf{x}'}(\omega_i') d\mu(\alpha).\end{aligned} \tag{5.77}$$

This then suggest the idea, to define the light transport operator in participating media as a combination of the light propagation operator $\overline{\mathbf{G}}$ and the local light scattering operator in participating media $\overline{\mathbf{K}}$

$\overline{\mathbf{G}}$ (442)
$\overline{\mathbf{K}}$ (445)

Obviously, applying the operator $\overline{\mathbf{G}}$ to the operator $\overline{\mathbf{K}}$ delivers the fraction of light incident at point $\mathbf{x}$ from directions ω_i after scattering at a surface point $\mathbf{s}$ and at points within a medium that lie on a line from $\mathbf{x}$ to $\mathbf{s}$ in directions ω_i. Mathematically, this can be expressed as follows:

$\overline{\mathbf{G}}$ (442)
$\overline{\mathbf{K}}$ (445)

$$
\begin{aligned}
(\overline{\mathbf{GK}}\mathrm{L}_i)(\mathbf{x},\omega_i) &= \left(\left(\overline{\mathbf{G}}^{\partial\mathcal{V}}+\overline{\mathbf{G}}^{\mathcal{V}^o}\right)\left(\overline{\mathbf{K}}^{\partial\mathcal{V}}+\overline{\mathbf{K}}^{\mathcal{V}^o}\right)\mathrm{L}_i\right)(\mathbf{x},\omega_i) && (5.78)\\
&= \left(\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}}\mathrm{L}_i\right)(\mathbf{x},\omega_i)+\underbrace{\left(\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\mathcal{V}^o}\mathrm{L}_i\right)(\mathbf{x},\omega_i)}_{\mathrm{L}_i\equiv 0}+ && (5.79)\\
&\quad \underbrace{\left(\overline{\mathbf{G}}^{\mathcal{V}^o}\overline{\mathbf{K}}^{\partial\mathcal{V}}\mathrm{L}_i\right)(\mathbf{x},\omega_i)}_{\mathrm{L}_i\equiv 0}+\left(\overline{\mathbf{G}}^{\mathcal{V}^o}\overline{\mathbf{K}}^{\mathcal{V}^o}\mathrm{L}_i\right)(\mathbf{x},\omega_i). && (5.80)
\end{aligned}
$$

Due to its definitions, the operators $\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\mathcal{V}^o}$ and $\overline{\mathbf{G}}^{\mathcal{V}^o}\overline{\mathbf{K}}^{\partial\mathcal{V}}$ each return the function $\mathrm{L}_i \equiv 0$. Intuitively, this should be clear, since $\overline{\mathbf{K}}^{\mathcal{V}^o}$ maps the surface-component of L_i to zero, and the following application of the operator $\overline{\mathbf{G}}^{\partial\mathcal{V}}$ works only on this surface-component. The same also holds for $\overline{\mathbf{G}}^{\mathcal{V}^o}\overline{\mathbf{K}}^{\partial\mathcal{V}}$ vice versa, here, $\overline{\mathbf{K}}^{\partial\mathcal{V}}$ maps the volume-component of L_i to zero, and $\overline{\mathbf{G}}^{\mathcal{V}^o}$ works only on this component. That is, $(\overline{\mathbf{GK}}\mathrm{L}_i)(\mathbf{x},\omega_i)$ can be written as

$$
\left(\overline{\mathbf{GK}}\mathrm{L}_i\right)(\mathbf{x},\omega_i)=\left(\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}}\mathrm{L}_i\right)(\mathbf{x},\omega_i)+\left(\overline{\mathbf{G}}^{\mathcal{V}^o}\overline{\mathbf{K}}^{\mathcal{V}^o}\mathrm{L}_i\right)(\mathbf{x},\omega_i). \tag{5.81}
$$

Compared with the light transport operator $\mathbf{T}^{\partial\mathcal{V}}_{\mathrm{L}_i}$ in a vacuum, the linear operator $\left(\overline{\mathbf{GK}}\mathrm{L}_i\right)(\mathbf{x},\omega_i)$ from Equation (5.81) plays the same role in a medium, thus, it can be interpreted as the light transport operator in participating media.

Together with the emission operator from Equation (5.75) we then get the following operator formulation for the SLTE

$$
\mathrm{L}_i=\overline{\mathbf{G}}\mathrm{L}_{e,o}+\overline{\mathbf{GK}}\mathrm{L}_i. \tag{5.82}
$$

As in the previous section, we denote the composition of the light propagation and the local light scattering operator in participating media as the *light transport operator in participating media*.

DEFINITION 5.11 (The Light Transport Operator in Participating Media) *Let* $\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ *be the space of square Lebesgue-integrable functions defined on the ray space* $\mathcal{R}$. *Then, the* light transport operator

$\mathcal{L}^2(\mathcal{R},\overline{\zeta})$ (112)
$(\mathcal{R},\zeta)$ (44)

$$
\begin{aligned}
\overline{\mathbf{T}}_{\mathrm{L}_i}:\mathcal{L}^2(\mathcal{R},\overline{\zeta}) &\longrightarrow \mathcal{L}^2(\mathcal{R},\overline{\zeta})\\
h_i(\mathbf{x},\omega_i) &\mapsto \left(\overline{\mathbf{T}}_{\mathrm{L}_i}h_i\right)(\mathbf{x},\omega_i)
\end{aligned}
$$

is defined by

$$\overline{\mathbf{T}}_{\mathrm{L_i}} \stackrel{\mathrm{def}}{=} \overline{\mathbf{G}}\overline{\mathbf{K}} \tag{5.83}$$

$$= \overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}} + \overline{\mathbf{G}}^{\mathcal{V}^o}\overline{\mathbf{K}}^{\mathcal{V}^o} \tag{5.84}$$

$$= \overline{\mathbf{T}}^{\partial\mathcal{V}}_{\mathrm{L_i}} + \overline{\mathbf{T}}^{\mathcal{V}^o}_{\mathrm{L_i}} \tag{5.85}$$

$\overline{\mathbf{G}}$ (443) $\overline{\mathbf{K}}$ (445) *where $\overline{\mathbf{G}}$ is the light propagation operator and $\overline{\mathbf{K}}$ is the local light scattering operator, both valid in participating media, see Figure 5.8. The light transport operator $\overline{\mathbf{T}}_{\mathrm{L_i}}$ maps an incident radiance function $\mathrm{L_i}$ to the incident function $\overline{\mathbf{T}}_{\mathrm{L_i}}\mathrm{L_i}$, that is: The light transport operator simulates a single scattering step of the light transport in participating media.*

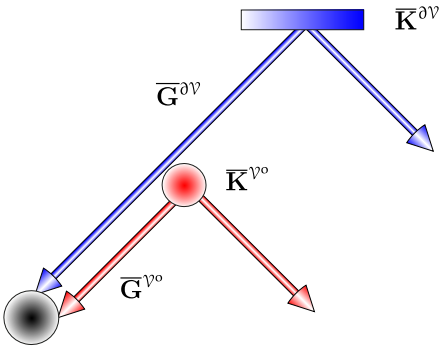


FIGURE 5.8: THE LIGHT TRANSPORT OPERATOR IN PARTICIPATING MEDIA, $\overline{\mathbf{T}}$. Light incident at a surface or a volumetric point is scattered via $\overline{\mathbf{K}}^{\partial\mathcal{V}}$ respectively $\overline{\mathbf{K}}^{\mathcal{V}^o}$ and propagated via $\overline{\mathbf{G}}^{\partial\mathcal{V}}$ respectively $\overline{\mathbf{G}}^{\mathcal{V}^o}$ to the point to be shaded.

With the source function $\overline{\mathbf{G}}\mathrm{L}_{e,o}$ and the definition of the light transport operator $\overline{\mathbf{T}}_{\mathrm{L_i}}$, we are now ready to represent the *light transport equation in participating media* from Equation (5.4) in a much simpler form, namely as a linear operator equation.

Linear Operator Equation (61)

DEFINITION 5.12 (The Light Transport Operator Equation in Participating Media) *Based on the light propagation operator $\overline{\mathbf{G}}$ and the local light scattering operator $\overline{\mathbf{K}}$, the* light transport operator equation *associated with the SLTE is given by*

$\overline{\mathbf{G}}$ (443) $\overline{\mathbf{K}}$ (445)

$$\mathrm{L_i} = \overline{\mathbf{G}}\mathrm{L}_{e,o} + \overline{\mathbf{T}}_{\mathrm{L_i}}\mathrm{L_i} \tag{5.86}$$

$$= \mathrm{L}_{e,i} + \overline{\mathbf{T}}_{\mathrm{L_i}}\mathrm{L_i}, \tag{5.87}$$

where we use the identity $L_{e,i} = \overline{\mathbf{G}} L_{e,o}$ *for the emitted radiance from Equation (5.66).*

With the same argumentation as in the case of the light transport in a vacuum—that is, under the condition that the light transport operator is contracting, i.e. $\|\overline{\mathbf{T}}_{L_i}\| < 1$—the solution operator $\overline{\mathbf{S}}_{L_i}$ of the above linear operator equation, given by

$$\overline{\mathbf{S}}_{L_i} = \left(\mathbf{I} - \overline{\mathbf{T}}_{L_i}\right)^{-1}, \tag{5.88}$$

can also be written in form of a Neumann series as:

$$\begin{aligned} L_i &= \sum_{i=0}^{\infty} \overline{\mathbf{T}}_{L_i}^i L_{e,i} && (5.89)\\ &= L_{e,i} + \overline{\mathbf{T}}_{L_i} L_{e,i} + \overline{\mathbf{T}}_{L_i}^2 L_{e,i} + \overline{\mathbf{T}}_{L_i}^3 L_{e,i} + \dots . && (5.90) \end{aligned}$$

The light transport in participating media can then be interpreted as composed of two physical processes: Propagation and scattering of light at surfaces as well as at small particles within media. For describing the distribution of light in ray space it suffices to determine the amount of light emitted from surfaces or from volumetric sources and to formulate the light transport operator $\overline{\mathbf{T}}_{L_i}$.

For a detailed analysis of light transport in participating media, let us write our light transport operator equation in terms of the surface and volume propagation as well as the scattering operators, thus:

$$\begin{aligned} L_i &= \sum_{i=0}^{\infty} (\overline{\mathbf{G}}\overline{\mathbf{K}})^i L_{e,i} && (5.91)\\ &= L_{e,i} + \overline{\mathbf{G}}\overline{\mathbf{K}} L_{e,i} + (\overline{\mathbf{G}}\overline{\mathbf{K}})^2 L_{e,i} + (\overline{\mathbf{G}}\overline{\mathbf{K}})^3 L_{e,i} + \cdots && (5.92)\\ &\overset{(5.81)}{=} L_{e,i} + \left(\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}} + \overline{\mathbf{G}}^{\mathcal{V}^\circ}\overline{\mathbf{K}}^{\mathcal{V}^\circ}\right) L_{e,i} + && (5.93)\\ &\quad \left(\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}} + \overline{\mathbf{G}}^{\mathcal{V}^\circ}\overline{\mathbf{K}}^{\mathcal{V}^\circ}\right)^2 L_{e,i} + \left(\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}} + \overline{\mathbf{G}}^{\mathcal{V}^\circ}\overline{\mathbf{K}}^{\mathcal{V}^\circ}\right)^3 + \cdots \\ &= L_{e,i} + \overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}} L_{e,i} + \overline{\mathbf{G}}^{\mathcal{V}^\circ}\overline{\mathbf{K}}^{\mathcal{V}^\circ} L_{e,i} + \overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}}\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}} L_{e,i} + \cdots && (5.94)\\ &= L_{e,i} + \\ &\quad \overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}} L_{e,i} + \overline{\mathbf{G}}^{\mathcal{V}^\circ}\overline{\mathbf{K}}^{\mathcal{V}^\circ} L_{e,i} + && (5.95)\\ &\quad \overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}}\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}} L_{e,i} + \overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}}\overline{\mathbf{G}}^{\mathcal{V}^\circ}\overline{\mathbf{K}}^{\mathcal{V}^\circ} L_{e,i} + \\ &\quad \overline{\mathbf{G}}^{\mathcal{V}^\circ}\overline{\mathbf{K}}^{\mathcal{V}^\circ}\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}} L_{e,i} + \overline{\mathbf{G}}^{\mathcal{V}^\circ}\overline{\mathbf{K}}^{\mathcal{V}^\circ}\overline{\mathbf{G}}^{\mathcal{V}^\circ}\overline{\mathbf{K}}^{\mathcal{V}^\circ} L_{e,i} + \cdots , \end{aligned}$$

see Figure 5.9. Obviously, the light, incident at a scene point comes directly from surface and volume emitters, or indirectly over multiple scattering at surfaces and/or volumetric points.

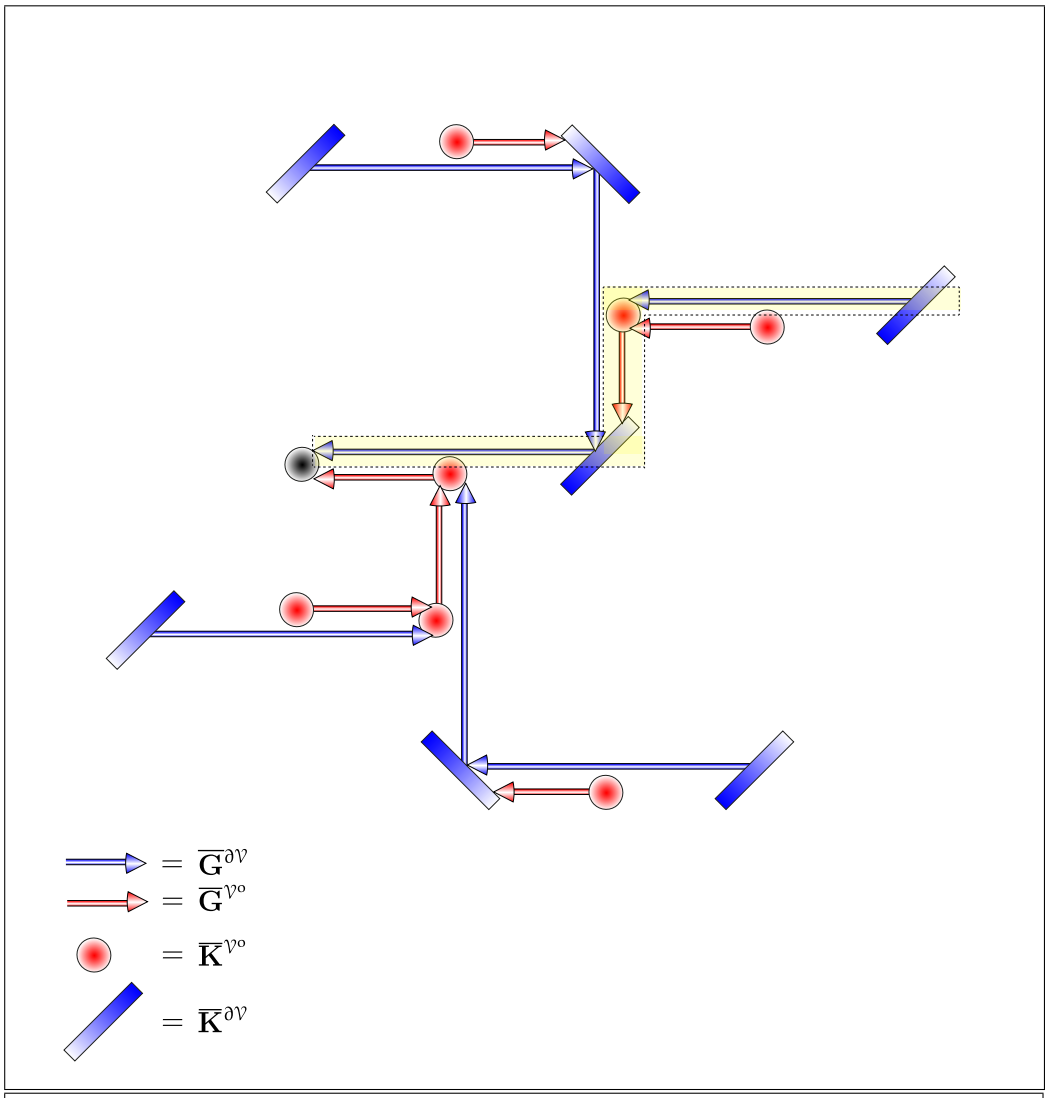


FIGURE 5.9: THE LIGHT TRANSPORT OPERATOR IN PARTICIPATING MEDIA. Let us assume that the black point has to be shaded. The figure illustrates the recursive application of the light transport operator $\overline{\mathbf{T}} = \overline{\mathbf{G}}\overline{\mathbf{K}}$, with $\overline{\mathbf{G}} = \overline{\mathbf{G}}^{\partial\mathcal{V}} + \overline{\mathbf{G}}^{\partial\mathcal{V}}$ and $\overline{\mathbf{K}} = \overline{\mathbf{K}}^{\partial\mathcal{V}} + \overline{\mathbf{K}}^{\partial\mathcal{V}}$ for evaluating the first three terms of the associated Neumann series. The blue and the red arrow corresponds to light propagation from surface respectively volumetric points and the red circle as well as the blue brick corresponds to volumetric respectively scattering at surface points. The yellow labeled path represents the transport path $\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}}\overline{\mathbf{G}}^{\mathcal{V}^o}\overline{\mathbf{K}}^{\mathcal{V}^o}\overline{\mathbf{G}}^{\partial\mathcal{V}}\overline{\mathbf{K}}^{\partial\mathcal{V}}$.

Replacing the incident radiance in the measurement equation from (5.3) by the product of the solution operator $\overline{\mathbf{S}}$ of the Neumann series and the emitted radiance, $L_{e,i}$, then a measurement can also be written as the linear functional:

Measurement Equation (416)
Neumann Series (135)
Linear Functional (55)

$$\mathcal{M} \overset{def}{=} \langle W_e, L_i \rangle \tag{5.96}$$

$$\overset{L_i = \overline{\mathbf{S}}_{L_i} L_{e,i}}{=} \langle W_e, \overline{\mathbf{S}}_{L_i} L_{e,i} \rangle. \tag{5.97}$$

5.2 AN OPERATOR MODEL FOR IMPORTANCE TRANSPORT IN A VACCUM

The fundamental result from Section 2.1.6, where we discussed adjoint operator equations, was the identity

$$\langle f(x), i(x) \rangle = \langle g(x), h(x) \rangle \tag{5.98}$$

for a pair of adjoint equations

$$f(x) = g(x) + (\mathbf{T}f)(x) \tag{5.99}$$

$$h(x) = i(x) + (\mathbf{T}^* h)(x). \tag{5.100}$$

In Section 2.3.2 then we have transferred this result to the case of Fredholm integral equations of the 2^{nd} kind. Applied to the incident stationary light transport equation and the incident stationary importance transport equation within a vacuum then identity from Equation (5.98) can be written as:

SLTEV (398)
SITEV (413)

$$\langle L_{e,o}, W_i \rangle = \langle W_e, L_i \rangle. \tag{5.101}$$

This means, that the measurement equation can not only be solved via computing the radiance at important points and directions of the scene, but also via computing the importance at illuminated points and directions.

Measurement Equation (416)

In the following two sections, we will now derive shortly the tools that are required also to formulate an operator model for importance transport in a vacuum. As we will see in Section 9.2, this will lead to new insights and possibilities to write new rendering algorithms.

Importance (415)

5.2.1 THE IMPORTANCE PROPAGATION AND THE IMPORTANCE SCATTERING OPERATOR IN A VACUUM

To derive an operator model for importance transport, similar to the model of light transport, we have to construct linear operators on a corresponding function space. Since importance flows in the opposite direction to radiance, this function space has to be defined

Radiance (250)

on the reversible ray space $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$ introduced in Box 2.3.

Let $\mathcal{L}(\widetilde{\mathcal{R}}^{\partial\mathcal{V}})$ be the space of real-valued functions defined on the reversible ray space $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$. We now equipped this function space with a measure based on the throughput measure $\zeta^{\perp}$. For this, we have to construct a bijective mapping $\mathfrak{m}$ between $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$ and the ray space $\mathcal{R}^{\partial\mathcal{V}}$, by

$\zeta^{\perp}$ (94)
$\mathcal{R}^{\partial\mathcal{V}}$ (44)

$$\begin{aligned} \mathfrak{m} : \widetilde{\mathcal{R}}^{\partial\mathcal{V}} &\rightarrow \mathcal{R}^{\partial\mathcal{V}} && (5.102)\\ \widetilde{\mathbf{r}} = (\mathbf{s},\omega) &\mapsto \mathfrak{m}(\mathbf{s},\omega) = (\gamma(\mathbf{s},\omega),-\omega), && (5.103)\end{aligned}$$

which ensures that any ray of $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$ has an image in $\mathcal{R}^{\partial\mathcal{V}}$, that is, the reversible ray $\widetilde{\mathbf{r}} = (\mathbf{s},\omega)$ is mapped to its associated ray in $\mathcal{R}^{\partial\mathcal{V}}$.

Borel σ-algebra (865)

Based on the function $\mathfrak{m}$, we can now construct a measure $\widetilde{\zeta}^{\perp}$ on the Borel σ-algebra $\mathfrak{B}(\widetilde{\mathcal{R}}^{\partial\mathcal{V}})$ defined by

$$\begin{aligned} \widetilde{\zeta}^{\perp}(B) &\stackrel{\text{def}}{=} (\zeta^{\perp}\circ\mathfrak{m})(B) && (5.104)\\ &= \left((\mu^2\times\sigma^{\perp})\circ\mathfrak{m}\right)(B) && (5.105)\end{aligned}$$

for all $B \in \mathfrak{B}(\widetilde{\mathcal{R}}^{\partial\mathcal{V}})$.

REMARK 5.11 *Obviously, the measure $\widetilde{\zeta}^{\perp}$ assigns a set B of reversible rays the throughput measure of its image $\mathfrak{m}(B)$.*

Endowed with the measure $\widetilde{\zeta}^{\perp}$, we then define the Lebesgue space $\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}},\widetilde{\zeta}^{\perp})$ by

$$\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}},\widetilde{\zeta}^{\perp}) \stackrel{\text{def}}{=} \left\{ f\in\mathcal{L}(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}) \,\middle|\, \|f\|_{\mathcal{L}^2} < \infty \right\}, \tag{5.106}$$

where the $\mathcal{L}^2$-norm is given by the inner product

$$\begin{aligned} \|f\|_{\mathcal{L}^2} \stackrel{\text{def}}{=} \langle f,f\rangle &= \int_{\partial\mathcal{V}\times S^2(\mathbf{s})} |f(\widetilde{\mathbf{r}})|^2 \,\mathrm{d}\widetilde{\zeta}^{\perp}(\widetilde{\mathbf{r}}) && (5.107)\\ &\stackrel{(5.105)}{=} \int_{\partial\mathcal{V}}\int_{S^2(\mathbf{s})} |f(\mathfrak{m}(\mathbf{s},\omega))|^2 \,\mathrm{d}\mu^2(\mathbf{s})\,\mathrm{d}\sigma^{\perp}_{\mathbf{s}}(\omega). && (5.108)\end{aligned}$$

Thus, $\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}},\widetilde{\zeta}^{\perp})$ contains all functions which, square-integrated over the unit sphere at all surface points $\partial\mathcal{V}$, deliver a finite value.

THE IMPORTANCE PROPAGATION OPERATOR IN A VACUUM. Based on the construct of the Lebesgue space $\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}},\widetilde{\zeta}^{\perp})$, we now define—in accordance to the derivation of the light propagation operator—the *importance propagation operator in a vacuum* by:

$\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}},\widetilde{\zeta}^{\perp})$ (452)
$\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$ (44)

DEFINITION 5.13 (The Importance Propagation Operator in a Vacuum) *Let $\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}},\widetilde{\zeta}^{\perp})$ be the space of square Lebesgue-integrable functions defined on $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$. Then, the* im-

portance propagation operator in a vacuum

$$\begin{aligned} \mathbf{G}^{\partial\mathcal{V}^*} : \mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \widetilde{\zeta}^{\perp}) &\longrightarrow \mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \widetilde{\zeta}^{\perp}) \\ h_o(s', \omega_o') &\mapsto h_i(s, \omega_i) = (\mathbf{G}^{\partial\mathcal{V}^*} h_o)(s, \omega_i) \end{aligned}$$

is defined by

$$(\mathbf{G}^{\partial\mathcal{V}^*} h_o)(s, \omega_i) \stackrel{\text{def}}{=} \begin{cases} h_o(\gamma(s, \omega_i), -\omega_i) & \text{for } (s, \omega_i) \in \widetilde{\mathcal{R}}^{\partial\mathcal{V}} \\ 0 & \text{otherwise,} \end{cases} \tag{5.109}$$

where $s \in \partial\mathcal{V}$, $\omega_o' = -\omega_i$, *and* γ *is the ray casting function with* $\gamma(s, \omega_i) = s'$. γ (47)

So, importance propagation operator can be interpreted as a mapping that maps an exitant function h_o, *defined on surfaces from* $\partial\mathcal{V}$, *onto an incident function* $h_i = (\mathbf{G}^{\partial\mathcal{V}^*} h_o)$, *which is also defined on points from* $\partial\mathcal{V}$. $\partial\mathcal{V}$ (41)

REMARK 5.12 *Applied to an exitant radiance function* W_o, *the operator* $\mathbf{G}^{\partial\mathcal{V}^*}$ *returns the incident importance function* $W_i = \mathbf{G}^{\partial\mathcal{V}^*} W_o$ *as result of the propagation of importance from object surfaces. If the function* W_o *measures importons exitant from any point, the function* W_i *obviously measures these importons after propagation incident on surfaces.*

REMARK 5.13 *Due to the definition of the light propagation operator in a vacuum* $\mathbf{G}^{\partial\mathcal{V}}$ *and its adjoint, the importance propagation operator in a vacuum* $\mathbf{G}^{\partial\mathcal{V}^*}$ *it is relatively straightforward to show that it holds:*

$$\mathbf{G}^{\partial\mathcal{V}} = \mathbf{G}^{\partial\mathcal{V}^*}, \tag{5.110}$$

that is, $\mathbf{G}^{\partial\mathcal{V}}$ *is self-adjoint. For a detailed proof see [221, Veach 1998].*

THE LOCAL IMPORTANCE SCATTERING OPERATOR IN A VACUUM. In addition to the surface emission term, the SITEV from Definition (4.44) still contains a scattering term similar to the scattering equation of the SLTEV. We now define a *local importance scattering operator* by: Scattering Equation (374)

DEFINITION 5.14 (The Local Importance Scattering Operator) *Let* $\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \widetilde{\zeta}^{\perp})$ *be the space of square Lebesgue-integrable functions defined on* $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$. *Then, the* local importance scattering operator $\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \widetilde{\zeta}^{\perp})$ (452) $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$ (44)

$$\begin{aligned} \mathbf{K}^{\partial\mathcal{V}^*} : \mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^{\perp}) &\longrightarrow \mathcal{L}^2(\mathcal{R}^{\partial\mathcal{V}}, \zeta^{\perp}) \\ h_i(s, \omega_i) &\mapsto h_o(s, \omega_o) = (\mathbf{K}^{\partial\mathcal{V}^*} h_i)(s, \omega_o) \end{aligned}$$

is defined as

$$(\mathbf{K}^{\partial\mathcal{V}^*} h_i)(s, \omega_o) \stackrel{\text{def}}{=} \int_{S^2(s)} f_s^*(s, \omega_i \to \omega_o)\, h_i(s, \omega_i)\, d\sigma_s^{\perp}(\omega_i), \tag{5.111}$$

where $\mathbf{s} \in \partial\mathcal{V}$. ∂V (41)

The local light scattering operator can be interpreted as it maps an incident function h_i*, defined on* $\partial\mathcal{V}$*, onto an exitant function* $h_o = (\mathbf{G}^{\partial\mathcal{V}^*} h_i)$*, which is also defined on points from* $\partial\mathcal{V}$*.*

Exitant & Incident Functions (48)

REMARK 5.14 *Applied to an incident importance function* W_i *the operator* $\mathbf{K}^{\partial\mathcal{V}^*}$ *returns the exitant importance function* W_o *as result from a single scattering operation at an object surface, thus* $W_o = \mathbf{K}^{\partial\mathcal{V}^*} W_i$*. If the function* W_i *measures importons just before their arrival at a surface point, then* W_o *measures importons after scattering.*

REMARK 5.15 *Obviously, the definition of the local importance scattering operator,* $\mathbf{K}^{\partial\mathcal{V}^*}$*, is well defined, since it holds:*

$$
\begin{aligned}
&\langle h, \mathbf{K}^{\partial\mathcal{V}} g \rangle \\
&= \int_{\partial\mathcal{V}} \int_{S^2(\mathbf{s})} h(\mathbf{s}, \omega_o) \underbrace{\left(\int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o) g(\mathbf{s}, \omega_i) \, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) \right)}_{\mathbf{K}^{\partial\mathcal{V}} g} d\sigma^{\perp}_{\mathbf{s}}(\omega_o) d\mu^2(\mathbf{s}) \\
&= \int_{\partial\mathcal{V}} \int_{S^2(\mathbf{s})} \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o) \, h(\mathbf{s}, \omega_o) d\sigma^{\perp}_{\mathbf{s}}(\omega_o) \, g(\mathbf{s}, \omega_i) \, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) d\mu^2(\mathbf{s}) \\
&= \int_{\partial\mathcal{V}} \int_{S^2(\mathbf{s})} \underbrace{\left(\int_{S^2(\mathbf{s})} f^*_s(\mathbf{s}, \omega_o \to \omega_i) \, h(\mathbf{s}, \omega_o) \, d\sigma^{\perp}_{\mathbf{s}}(\omega_o) \right)}_{\mathbf{K}^{\partial\mathcal{V}^*} h} g(\mathbf{s}, \omega_i) \, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) d\mu^2(\mathbf{s}) \\
&= \langle \mathbf{K}^{\partial\mathcal{V}^*} h, g \rangle .
\end{aligned}
$$

Here we have used the adjoint BSDF f^*_s*, defined by,*

$$
f^*_s(\mathbf{s}, \omega_i \to \omega_o) \stackrel{\text{def}}{=} f_s(\mathbf{s}, \omega_o \to \omega_i), \tag{5.112}
$$

Helmholtz Reciprocity (331)
Theorem of Fubini-Tonelli (115)

the Helmholtz reciprocity of the adjoint BSDF, and the Theorem of Fubini-Tonelli for iterated integrals to change the order of integration.

5.2.2 THE IMPORTANCE TRANSPORT OPERATOR EQUATION IN A VACUUM

In analogy to the definition of the light transport operator, we can now define the *importance transport operator in a vacuum*, $\mathbf{T}^{\partial\mathcal{V}}_{W_i}$, by the composition of the local importance scattering operator and the importance propagation operator. For that purpose, let us assume that the BSDF f_s is symmetric at all points of $\partial\mathcal{V}$, so that it holds: $\mathbf{K}^{\partial\mathcal{V}} = \mathbf{K}^{\partial\mathcal{V}^*}$.

$\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \widetilde{\zeta}^{\perp})$ (452) $\widetilde{\mathcal{R}}^{\partial\mathcal{V}}$ (48)

DEFINITION 5.15 (The Importance Transport Operator in a Vacuum, $\mathbf{T}^{\partial\mathcal{V}^*}_{W_i}$) *Let $\mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \widetilde{\zeta}^{\perp})$ be the space of square Lebesgue-integrable functions defined on the reversible ray space $\mathcal{L}(\widetilde{\mathcal{R}})$. Then, the* importance transport operator in a vaccum

$$\begin{aligned} \mathbf{T}^{\partial\mathcal{V}}_{W_i} : \mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \widetilde{\zeta}^{\perp}) &\longrightarrow \mathcal{L}^2(\widetilde{\mathcal{R}}^{\partial\mathcal{V}}, \widetilde{\zeta}^{\perp}) \\ h_i(\mathbf{s}, \omega_i) &\mapsto h'_i(\mathbf{s}, \omega_i) = (\mathbf{T}^{\partial\mathcal{V}}_{W_i} h_i)(\mathbf{s}, \omega_i) \end{aligned}$$

is defined by

$$\mathbf{T}^{\partial\mathcal{V}}_{W_i} \stackrel{\text{def}}{=} \mathbf{G}^{\partial\mathcal{V}^*}\mathbf{K}^{\partial\mathcal{V}^*} \tag{5.113}$$

where $\mathbf{G}^{\partial\mathcal{V}^*}$ is the importance propagation operator in a vacuum and $\mathbf{K}^{\partial\mathcal{V}^*}$ is the local importance scattering operator in a vacuum. $\mathbf{G}^{\partial\mathcal{V}^*}$ (453) $\mathbf{K}^{\partial\mathcal{V}^*}$ (453)

REMARK 5.16 *Applied to an incident importance function W_i, the transport operator $\mathbf{T}^{\partial\mathcal{V}}_{W_i}$ returns the incident function $\mathbf{T}^{\partial\mathcal{V}}_{W_i} W_i$, that is: The importance transport operator simulates a single scattering step of importance transport at a surface in a vacuum.*

DEFINITION 5.16 (The Importance Transport Operator Equation in a Vaccum) *Based on the importance propagation operator $\mathbf{G}^{\partial\mathcal{V}^*}$ and the local importance scattering operator $\mathbf{K}^{\partial\mathcal{V}^*}$, the* incident importance transport operator equation in a vacuum *associated with the SITEV is given by* SITEV (413)

$$W_i = W_{e,i} + \mathbf{T}^{\partial\mathcal{V}}_{W_i} W_i. \tag{5.114}$$

Analogous to the development of a solution to the light transport operator equation in a vacuum, the importance transport operator equation can—under the condition that the importance transport operator $\mathbf{T}^{\partial\mathcal{V}}_{W_i}$ is contracting—also be written in form of a Neumann series. For that, $\mathbf{S}^{\partial\mathcal{V}}_{W_i}$ denotes the importance transport solution operator given by Neumann Series (135)

$$\mathbf{S}^{\partial\mathcal{V}}_{W_i} \stackrel{\text{def}}{=} \left(\mathbf{I} - \mathbf{T}^{\partial\mathcal{V}}_{W_i}\right)^{-1} \tag{5.115}$$

then it holds:

$$\begin{aligned} W_i &= \mathbf{S}^{\partial\mathcal{V}}_{W_i} W_{e,i} && (5.116) \\ &\stackrel{(5.115)}{=} (\mathbf{I} - \mathbf{T}^{\partial\mathcal{V}}_{W_i})^{-1} W_{e,i} && (5.117) \\ &\stackrel{(2.385)}{=} \sum_{i=0}^{\infty} \mathbf{T}^{\partial\mathcal{V}\,i}_{W_i} W_{e,i} && (5.118) \\ &= W_{e,i} + \mathbf{T}^{\partial\mathcal{V}}_{W_i} W_{e,i} + \mathbf{T}^{\partial\mathcal{V}\,2}_{W_i} W_{e,i} + \cdots. && (5.119) \end{aligned}$$

Obviously, the importance transport in a vacuum can be interpreted as composed of two processes: Propagation of importance between surfaces and scattering of importance

at surfaces, that is, the importance incident at a scene point comes directly from surface emitters or indirectly via multiple scattering at surface points. Hence, for describing the distribution of importance in a scene it suffices to determine the amount of importance emitted from existing sources and to formulate the importance transport operator $\mathbf{T}_{W_i}^{\partial\mathcal{V}}$.

BSDF (371) REMARK 5.17 *The requirement of the symmetry of the BSDF is very restrictive, since already such simple processes like the refraction of light at interfaces can not be modeled by a symmetric BSDF. As one can see from Lemma 4.2, the term* $(1 -$
BTDF (330) $F_r(\omega_i))\frac{\eta_t^2}{\eta_i^2}$ *in the BTDF does not satisfy the condition of symmetry, that is, the scattering operator* $\mathbf{K}^{\partial\mathcal{V}}$ *can not be self-adjoint.*

Using the same rules for scattering in both transport operator equations, [221, Veach 1998] suggest, as a consequence of Kirchhoffs law of thermo dynamics, to choose the reflection operator $\mathbf{K}^{\partial\mathcal{V}}$ *as:*

$$(\mathbf{K}^{\partial\mathcal{V}}L_i)(s,\omega_o) \stackrel{def}{=} \int_{S^2} f_s(s,\omega_i \to \omega_o)L_i(s,\omega_i)\,d\sigma_s^{\perp}(\omega_i) \tag{5.120}$$

with the symmetry condition

$$\frac{f_s(s,\omega_i \to \omega_o)}{\eta_o^2} = \frac{f_s(s,\omega_o \to \omega_i)}{\eta_i^2}. \tag{5.121}$$

This property ensures that $\frac{f_s(s,\omega_i\to\omega_o)}{\eta_o^2}$ *is a symmetric function, which leads to the self-adjointness of the reflection operator* $\mathbf{K}^{\partial\mathcal{V}}$.

5.3 FOUR BASIC TRANSPORT OPERATOR MODELS OF LIGHT TRANSPORT IN A VACUUM

$\mathbf{G}^{\partial\mathcal{V}}$ (430) Using the light and importance propagation operators $\mathbf{G}^{\partial\mathcal{V}}$ and $\mathbf{G}^{\partial\mathcal{V}^*}$ as well as of the
$\mathbf{G}^{\partial\mathcal{V}^*}$ (453) local scattering operators $\mathbf{K}^{\partial\mathcal{V}}$ and $\mathbf{K}^{\partial\mathcal{V}^*}$ within the incident light, respectively, the im-
$\mathbf{K}^{\partial\mathcal{V}}$ (432) portance transport operator equations leads to four mathematical representations for the
$\mathbf{K}^{\partial\mathcal{V}^*}$ (453) basic quantities, radiance and importance, namely:

$$L_i = \mathbf{G}^{\partial\mathcal{V}}L_o \quad \text{and} \quad W_i = \mathbf{G}^{\partial\mathcal{V}^*}W_o \tag{5.122}$$

$$L_o = \mathbf{K}^{\partial\mathcal{V}}L_i \quad \text{and} \quad W_o = \mathbf{K}^{\partial\mathcal{V}^*}W_i. \tag{5.123}$$

That is, the light transport operator equation can be expressed in terms of incident radiance, namely as,

$$L_i \stackrel{(5.25)}{=} L_{e,i} + \mathbf{T}_{L_i}^{\partial\mathcal{V}}L_i \tag{5.124}$$

and its adjoint equation, the importance transport operator equation, can be expressed in terms of exitant importance,

$$\begin{aligned} W_o &\overset{(5.111)}{=} \mathbf{K}^{\partial\mathcal{V}^*} W_i && (5.125)\\ &\overset{(5.114)}{=} \mathbf{K}^{\partial\mathcal{V}^*}\left(W_{e,i} + \mathbf{T}^{\partial\mathcal{V}}_{W_i} W_i\right) && (5.126)\\ &= \mathbf{K}^{\partial\mathcal{V}^*} W_{e,i} + \mathbf{K}^{\partial\mathcal{V}^*}\mathbf{G}^{\partial\mathcal{V}^*}\mathbf{K}^{\partial\mathcal{V}^*} W_i && (5.127)\\ &\overset{(5.111)}{=} W_{e,o} + \underbrace{\mathbf{K}^{\partial\mathcal{V}^*}\mathbf{G}^{\partial\mathcal{V}^*}}_{\mathbf{T}^{\partial\mathcal{V}}_{W_o}} W_o && (5.128)\\ &= W_{e,o} + \mathbf{T}^{\partial\mathcal{V}}_{W_o} W_o. && (5.129) \end{aligned}$$

Furthermore, we get the incident importance transport operator equation expressed in incident importance

$$W_i \overset{(5.113)}{=} W_{e,i} + \mathbf{T}^{\partial\mathcal{V}}_{W_i} W_i, \tag{5.130}$$

and for its adjoint, the exitant light transport operator equation, it holds:

$$\begin{aligned} L_o &\overset{(5.16)}{=} \mathbf{K}^{\partial\mathcal{V}} L_i && (5.131)\\ &\overset{(5.24)}{=} \mathbf{K}^{\partial\mathcal{V}}\left(L_{e,i} + \mathbf{T}^{\partial\mathcal{V}}_{L_i} L_i\right) && (5.132)\\ &= \mathbf{K}^{\partial\mathcal{V}} L_{e,i} + \mathbf{K}^{\partial\mathcal{V}}\mathbf{G}^{\partial\mathcal{V}}\mathbf{K}^{\partial\mathcal{V}} L_i && (5.133)\\ &\overset{(5.16)}{=} L_{e,o} + \underbrace{\mathbf{K}^{\partial\mathcal{V}}\mathbf{G}^{\partial\mathcal{V}}}_{\mathbf{T}^{\partial\mathcal{V}}_{L_o}} L_o && (5.134)\\ &= L_{e,o} + \mathbf{T}^{\partial\mathcal{V}}_{L_o} L_o. && (5.135) \end{aligned}$$

The solution operators to the above equation can now easily be derived. So, it holds for the solution operator of the incident light transport operator equation within a vacuum:

$$\mathbf{S}^{\partial\mathcal{V}}_{L_i} \overset{(5.32)}{=} \left(\mathbf{I} - \mathbf{T}^{\partial\mathcal{V}}_{L_i}\right)^{-1} = \left(\mathbf{I} - \mathbf{G}^{\partial\mathcal{V}}\mathbf{K}^{\partial\mathcal{V}}\right)^{-1} \tag{5.136}$$

and for the incident importance transport operator equation within a vacuum we get:

$$\mathbf{S}^{\partial\mathcal{V}}_{W_i} \overset{(5.115)}{=} \left(\mathbf{I} - \mathbf{T}^{\partial\mathcal{V}}_{W_i}\right)^{-1} = \left(\mathbf{I} - \mathbf{G}^{\partial\mathcal{V}^*}\mathbf{K}^{\partial\mathcal{V}^*}\right)^{-1}, \tag{5.137}$$

as well as

$$\mathbf{S}^{\partial\mathcal{V}}_{L_o} \overset{(5.32)}{=} \left(\mathbf{I} - \mathbf{T}^{\partial\mathcal{V}}_{L_o}\right)^{-1} = \left(\mathbf{I} - \mathbf{K}^{\partial\mathcal{V}}\mathbf{G}^{\partial\mathcal{V}}\right)^{-1} \overset{(5.137)}{=} \mathbf{S}^{\partial\mathcal{V}\,*}_{W_i} \tag{5.138}$$

and

$$\mathbf{S}^{\partial\mathcal{V}}_{W_o} \overset{(5.115)}{=} \left(\mathbf{I} - \mathbf{T}^{\partial\mathcal{V}}_{W_o}\right)^{-1} = \left(\mathbf{I} - \mathbf{K}^{\partial\mathcal{V}^*}\mathbf{G}^{\partial\mathcal{V}^*}\right)^{-1} \overset{(5.136)}{=} \mathbf{S}^{\partial\mathcal{V}^*}_{L_i} \tag{5.139}$$

for the exitant formulations of the light and importance transport operator equation within a vacuum.

Based on these results, we can now derive four different but equivalent formulations of the measurement equation: The first two of these measurements suggest to solve the global illumination problem via solving the incident formulations of the stationary light transport equation respectively the stationary importance transport equation, thus,

Measurement Equation (416

$$\mathcal{M} \overset{(4.428)}{=} \langle W_{e,o}, L_i \rangle \tag{5.140}$$

and

$$\begin{aligned}
\mathcal{M} &\overset{(4.428)}{=} \langle W_{e,o}, L_i \rangle && (5.141)\\
&\overset{(5.32)}{=} \langle W_{e,o}, \mathbf{S}^{\partial\mathcal{V}}_{L_i} L_{e,i} \rangle && (5.142)\\
&= \langle W_{e,o}, \mathbf{S}^{\partial\mathcal{V}}_{L_i} \mathbf{G}^{\partial\mathcal{V}} L_{e,o} \rangle && (5.143)\\
&\overset{(2.95)}{=} \langle \mathbf{G}^{\partial\mathcal{V}^*} \mathbf{S}^{\partial\mathcal{V}^*}_{L_i} W_{e,o}, L_{e,o} \rangle && (5.144)\\
&\overset{(5.139)}{=} \langle \mathbf{G}^{\partial\mathcal{V}^*} \underbrace{\mathbf{S}^{\partial\mathcal{V}}_{W_o} W_{e,o}}_{W_o}, L_{e,o} \rangle && (5.145)\\
&= \langle W_i, L_{e,o} \rangle. && (5.146)
\end{aligned}$$

That is, since the incident function W_i can be expressed via the adjoint of the solution operator of the light operator equation from Equation (5.30) applied to the emitted importance function $W_{e,o}$, namely by,

$$W_i = \mathbf{G}^{\partial\mathcal{V}^*} \mathbf{S}^{\partial\mathcal{V}^*}_{L_i} W_{e,o}, \tag{5.147}$$

there are two possibilities to solve the global illumination problem within a vacuum via incident formulation of the transport equations.

Apart from these both possibilities, there are still two other ways that can be used to solve the measurement equation. These methods are based on the exitant formulations of the transport equations, thus, it holds:

$$\begin{aligned}
\mathcal{M} &\overset{(5.146)}{=} \langle W_i, L_{e,o} \rangle && (5.148)\\
&= \langle \mathbf{G}^{\partial\mathcal{V}^*} W_o, L_{e,o} \rangle && (5.149)\\
&\overset{(2.95)}{=} \langle W_o, \mathbf{G}^{\partial\mathcal{V}} L_{e,o} \rangle && (5.150)\\
&= \langle W_o, L_{e,i} \rangle && (5.151)
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{M} &\overset{(4.428)}{=} \langle W_{e,o}, L_i \rangle && (5.152)\\
&= \langle W_{e,o}, \mathbf{G}^{\partial\mathcal{V}} L_o \rangle && (5.153)\\
&\overset{(2.95)}{=} \langle \mathbf{G}^{\partial\mathcal{V}^*} W_{e,o}, L_o \rangle && (5.154)\\
&= \langle W_{e,i}, L_o \rangle. && (5.155)
\end{aligned}$$

REMARK 5.18 *Note: The four measurements*

$$\mathcal{M} = \langle W_{e,o}, L_i \rangle, \tag{5.156}$$
$$\mathcal{M} = \langle W_o, L_{e,i} \rangle, \tag{5.157}$$
$$\mathcal{M} = \langle W_{e,i}, L_o \rangle \tag{5.158}$$
$$\mathcal{M} = \langle W_i, L_{e,o} \rangle, \tag{5.159}$$

correspond exactly to the quantities which we discuss in Section 2.3.2. Each of these inner products is composed of the solution of the direct equation and the source function of the associated adjoint equation or vice versa. Obviously, the concept of the adjoint operator allows us to evaluate measurements in a variety of ways leading to new insights and approaches for solving the light transport equation within a vacuum.

Thus, a possible method for measuring the flux through a sensor could be to compute the radiance distribution at all points within the scene combined with the importance emitted by the sensor. Vice versa, we can also compute the importance distribution at all points within the scene, and combine it with the radiance emitted by the light sources. Together, these equations specify many ways in which measurements can be made.

Photon Flux (249)
Radiance (250)
Importance (415)

5.4 THE PATH INTEGRAL MODEL OF LIGHT TRANSPORT

Usually, the global illumination problem is solved by evaluating the measurement equation at different points and directions. For that purpose, an infinite-dimensional linear integral equation, that describes the light transport in a vacuum, has to be solved. As we will see in Section 6.7.1, all methods for solving such integral equations comes with the disadvantage, that paths, starting at the eye or from a light source, have to be generated and evaluated recursively. This means that we can only find the next evaluation point by locally evaluating the current point.

Global Illumination Problem (6)
Integral Equation (127)
SLTEV (398)
Measurement Equation (416)

Instead of solving the global illumination problem by means of the well-known solution methods for integral equations a new approach was proposed in [221, Veach 1998]. The idea was developed by [198, Spanier & Gelbard] and is based on the concept of the so-called *path integral*. Here, the light transport problem is transformed into a simple integration problem, such, that each measurement can be written in the form

Global Illumination Problem (6)

$$\mathcal{M}_j = \int_{\mathbf{P}^\infty} f_j(\bar{\mathbf{x}}) d\mu_\infty(\bar{\mathbf{x}}), \tag{5.160}$$

where $\mathbf{P}^\infty$ is the *set of transport paths* of finite length within a scene, μ_∞ is the so-called *continuous path measure* defined on this space of paths, and f_j is a *measurement*

Measure (79)

contribution function.

Since it describes a measurement via an ordinary integral, the path integral formulation of the light transport problem has also a much simpler structure than the commonly used form of the measurement equation, defined via a dot product on a special function space. Additionally, it takes a more global view, since paths can now be considered as samples in the integration domain of all possible paths. This allows to use general-purpose integration methods for solving this integral such as multiple importance sampling.

Section 6.6.9

The path integral formulation of the light transport does not need to know the mathematical concepts of the direct and adjoint equation or the differences between the radiometric quantities of light and importance. Nevertheless, it allows to construct paths in arbitrary ways, that is, by starting at any node of a path, and extending the path outwards in both directions. This then leads to sampling strategies such as bidirectional path tracing and the Metropolis light transport, which are often better described using the path integral formulation.

Adjoint Equation (131)
Importance (415)
Section 9.3
Section 6.5.3.2

Section 5.4.1

In this section, we present the path integral model of light transport in a vacuum, that is, we construct a measure space consisting of the space of all paths of finite length between object surfaces $\mathbf{P}^\infty$, the associated continuous path measure, μ_∞, and the concept of the measurement contribution function. All these components are then used to build the path integral model of the light transport in a vacuum. Afterwards, we extend the path integral model of the light transport in a vacuum to the path integral model of the light transport in participating media. This requires to drill the path space $\mathbf{P}^\infty$, to include also all path of finite lengths that start, end, or pass through participating media, which involves also to construct a new extended continuous path measure.

Section 5.4.2

5.4.1 THE PATH INTEGRAL MODEL OF LIGHT TRANSPORT IN A VACUUM

Measurement Equation (417)

Our goal in this section is to express each measurement

$$\mathcal{M}_j \quad = \quad \langle W_e^j, L_i \rangle \tag{5.161}$$

as an integral of type

$$\mathcal{M}_j = \int_{\mathbf{P}^\infty} f_j(\overline{\mathbf{x}}) \, d\mu_\infty(\overline{\mathbf{x}}), \tag{5.162}$$

where $\mathbf{P}^\infty$ is the integration domain, f_j corresponds to the integrand, and μ_∞ is the integration measure. To do this, let us now derive these constructs with respect to the stationary light transport equation in a vacuum.

Measure (79)

THE PATH SPACE AND THE CONTINUOUS PATH MEASURE. For the following discussion let $\partial\mathcal{V} = \cup_{i=0}^{n}\partial\mathcal{V}_i$ be a finite set of 2-dimensional surfaces in $\mathbb{R}^3$, such as triangles and rectangles, that can be used to model the objects in a scene to be rendered.

$\partial\mathcal{V}$ (41)

Let us consider the Cartesian product of $k+1$ object surfaces $\partial\mathcal{V}_{i_j} \in \partial\mathcal{V}$, thus,

Cartesian Product (829)

$$\mathbf{P}_{\overline{i_0\dots i_k}} \stackrel{def}{=} \partial\mathcal{V}_{i_0} \times \partial\mathcal{V}_{i_1} \times \cdots \times \partial\mathcal{V}_{i_k}, \tag{5.163}$$

where $0 \leq i_j \leq k$ and $\partial\mathcal{V}_{i_j} \neq \partial\mathcal{V}_{i_{j+1}}$ for $0 \leq j \leq k-1$. Obviously, $\mathbf{P}_{\overline{i_0\dots i_k}}$ can be interpreted as the set of paths $\overline{\mathbf{x}} = \mathbf{x}_{i_0}\mathbf{x}_{i_1}\dots\mathbf{x}_{i_k}$ of length k, starting at $\partial\mathcal{V}_{i_0}$ and ending at $\partial\mathcal{V}_{i_k}$ with $\mathbf{x}_{i_j}$ from $\partial\mathcal{V}_{i_j}$ for $0 \leq j \leq k$. We can now extend this set to the set $\mathbf{P}_k$, i.e. the set of all paths of length k starting and ending at any object surface of $\partial\mathcal{V}$ by defining

$$\mathbf{P}_k \stackrel{def}{=} \bigcup_{\overline{i_0\dots i_k}\in\{0,1,\dots,n\}} \mathbf{P}_{\overline{i_0\dots i_k}}, \tag{5.164}$$

where two neighbored components of the tuple $\overline{i_0\dots i_k}$ must always be different, see Figure 5.10, where a scene is shown consisting of three surfaces.

Based on this construct, we can now define the *path space*, $\mathbf{P}^\infty$, that is, the space of all paths of finite length over object surfaces from $\partial\mathcal{V}$.

DEFINITION 5.17 (The Path Space of All Paths of Finite Length, $\mathbf{P}^\infty$) *The* path space, $\mathbf{P}^\infty$, *that is, the set of all paths of finite lengths, is defined by*

$$\mathbf{P}^\infty \stackrel{def}{=} \bigcup_{k=1}^{\infty} \mathbf{P}_k. \tag{5.165}$$

Let us now consider the set $\mathfrak{B}(\partial\mathcal{V}_{i_j}), 0 \leq j \leq k$, that is, the set of all subsets of $\partial\mathcal{V}_{i_j}$ generated by open rectangles of $\partial\mathcal{V}_{i_j}$. Since each of the sets $\mathfrak{B}(\partial\mathcal{V}_{i_j})$ is a σ-algebra, namely the Borel σ-algebra over the surface $\partial\mathcal{V}_{i_j}$, it follows from measure theory that the Cartesian product

σ-algebra (828)

$\mathfrak{B}(\cdot)$ (865)

$$\mathfrak{B}(\partial\mathcal{V}_{i_0} \times \partial\mathcal{V}_{i_1} \times \cdots \times \partial\mathcal{V}_{i_k}) \stackrel{def}{=} \mathfrak{B}(\partial\mathcal{V}_{i_0}) \times \cdots \times \mathfrak{B}(\partial\mathcal{V}_{i_k}), \tag{5.166}$$

$0 \leq k \leq n$, is also a σ-algebra. With the help of the Lebesgue area measure μ^2, we can then construct the *continuous path measure* μ_k on $\mathfrak{B}(\partial\mathcal{V}_{i_0} \times \partial\mathcal{V}_{i_1} \times \cdots \times \partial\mathcal{V}_{i_k})$ by

μ^2 (82)

Measure (79)

$$\begin{aligned} \mu_k(\mathbf{B}) &\stackrel{def}{=} \mu_k(B_0 \times \cdots \times B_k) && (5.167)\\ &= \mu^2(B_0) \cdot \ldots \cdot \mu^2(B_k), && (5.168) \end{aligned}$$

where $\mathbf{B} = B_0 \times \cdots \times B_k$ with $B_k \in \partial\mathcal{V}_{i_k}$.

Via the construction of the measure μ_k from Equation (5.167) we now define the *continuous path measure* μ_∞, which allows to extend the path space $\mathbf{P}^\infty$ to the measure space $(\mathbf{P}^\infty, \mathfrak{B}(\mathbf{P}^\infty), \mu_\infty)$.

Measure Space (80)

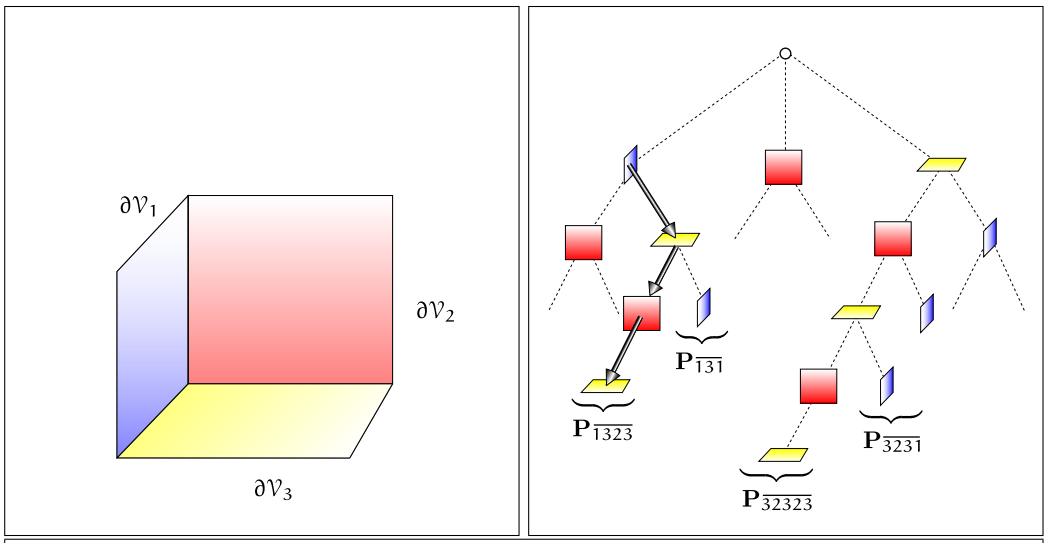


FIGURE 5.10: TYPICAL TRANSPORT PATHS IN A VACUUM. Left, a scene consisting of three object surfaces, it holds: $\partial\mathcal{V} = \{\partial\mathcal{V}_1, \partial\mathcal{V}_2, \partial\mathcal{V}_3\}$. Right, four different sets of paths: $\mathbf{P}_{\overline{1323}}, \mathbf{P}_{\overline{131}}, \mathbf{P}_{\overline{32323}}$ and $\mathbf{P}_{\overline{3231}}$. Thus for example, the set $\mathbf{P}_{\overline{1323}}$ represents all paths of length 3 starting at the left wall, running back, via the floor and the back wall to the left wall. Via the description of a path starting at the root and ending in a leaf, you can recognize that neighbored nodes correspond to different object surfaces. Note: The sets $\mathbf{P}_{\overline{1323}}$ and $\mathbf{P}_{\overline{3231}}$ are sets of different paths, but all paths within these sets have the same length. Obviously, the tree of height $k+1$ corresponds to $\mathbf{P}_k$, that is, the set of all path of length k.

DEFINITION 5.18 (The Measure Space $(\mathbf{P}^\infty, \mathfrak{B}(\mathbf{P}^\infty), \mu_\infty)$) *Let $\mathbf{P}^\infty$ be the space of all paths of all finite lengths and let μ_k be the measure from Equation (5.167) defined on the σ-algebra $\mathfrak{B}(\partial\mathcal{V}_{i_0} \times \partial\mathcal{V}_{i_1} \times \cdots \times \partial\mathcal{V}_{i_k})$. Defining*

Measure (79)

$\mathfrak{B}(\cdot)$ (865)

$$\mu_\infty(\mathbf{B}) \stackrel{\text{def}}{=} \mu_\infty\left(\mathbf{B} \cap \bigcup_{k=1}^{\infty} \mathbf{P}_k\right) \tag{5.169}$$

$$= \mu_\infty\left(\bigcup_{k=1}^{\infty} (\mathbf{B} \cap \mathbf{P}_k)\right) \tag{5.170}$$

$$= \sum_{k=1}^{\infty} \mu_k(\mathbf{B} \cap \mathbf{P}_k), \tag{5.171}$$

where $\mathbf{B} = B_{i_0} \times \cdots \times B_{i_k} \subseteq \partial\mathcal{V}_{i_0} \times \partial\mathcal{V}_{i_1} \times \cdots \times \partial\mathcal{V}_{i_k}$ is a subset of the set of all paths of length k, namely $\mathbf{P}_k$, then μ_∞ is a measure, the so-called continuous path measure. *Applied to the path space $\mathbf{P}^\infty$, the triple $(\mathbf{P}^\infty, \mathfrak{B}(\mathbf{P}^\infty), \mu_\infty)$ is obviously a measure space.*

Measure (79)

Measure Space (80)

EXAMPLE 5.6 *Let us consider once more the scene from Figure 5.10. We are interested in the value of* $\mu_\infty(\mathbf{P}_{\overline{131}})$*, that is, the path measure of paths of length two that start at the left wall and go back via the floor to the left wall, thus,*

$$\begin{aligned}
\mu_\infty(\mathbf{P}_{\overline{131}}) &= \sum_{k=1}^{\infty} \mu_k(\mathbf{P}_{\overline{131}} \cap \mathbf{P}_k) && (5.172)\\
&= \mu_2(\mathbf{P}_{\overline{131}}) && (5.173)\\
&\overset{(5.168)}{=} \mu^2(\partial\mathcal{V}_1)\,\mu^2(\partial\mathcal{V}_3)\,\mu^2(\partial\mathcal{V}_1). && (5.174)
\end{aligned}$$

Similar, we can compute the path measure of the set $\mathbf{P}_{\overline{3231}}$*, thus paths of length three, starting at the floor, running over the back wall to the floor, and ending at the left wall. For paths of this characteristic it holds:*

$$\begin{aligned}
\mu_\infty(\mathbf{P}_{\overline{3231}}) &= \sum_{k=1}^{\infty} \mu_k(\mathbf{P}_{\overline{3231}} \cap \mathbf{P}_k) && (5.175)\\
&= \mu_3(\mathbf{P}_{\overline{3231}}) && (5.176)\\
&\overset{(5.168)}{=} \mu^2(\partial\mathcal{V}_3)\,\mu^2(\partial\mathcal{V}_2)\,\mu^2(\partial\mathcal{V}_3)\,\mu^2(\partial\mathcal{V}_1). && (5.177)
\end{aligned}$$

THE MEASUREMENT CONTRIBUTION FUNCTION. With the path space $\mathbf{P}^\infty$ as integration domain and the continuous path measure μ_∞ as the integration measure, we have two of three constructs that are required for formulating the path integral model of light transport in a vacuum. Now, to represent the measurement equation in form of a path integral, thus, Measurement Equation (416)

$$\mathcal{M}_j = \int_{\mathbf{P}^\infty} f_j(\overline{\mathbf{x}})\, d\mu_\infty(\overline{\mathbf{x}}), \tag{5.178}$$

we only have to choose the integrand f_j.

We know from our discussions in the previous sections that the measurement equation can be written as the inner product of two measurable functions W_e^j and L_i, thus, Measurable Function (98)

$$\begin{aligned}
\mathcal{M}_j &= \langle W_e^j, L_i \rangle && (5.179)\\
&\overset{(5.33),(5.31)}{=} \left\langle W_e^j, \sum_{i=0}^{\infty} \mathbf{T}_{L_i}^{\partial\mathcal{V}^i} L_{e,i} \right\rangle && (5.180)\\
&\overset{(5.22)}{=} \left\langle W_e^j, \sum_{i=0}^{\infty} (\mathbf{G}^{\partial\mathcal{V}}\mathbf{K}^{\partial\mathcal{V}})^i L_{e,i} \right\rangle && (5.181)\\
&= \langle W_e^j, L_{e,i} \rangle + \left\langle W_e^j, \sum_{i=1}^{\infty} (\mathbf{G}^{\partial\mathcal{V}}\mathbf{K}^{\partial\mathcal{V}})^i L_{e,i} \right\rangle. && (5.182)
\end{aligned}$$

Except of the first inner product, any other term is the inner product of one or Linear Operator (53)

more linear operators and the functions W_e^j and $L_{e,i}$. Obviously, the definition of the inner product $\langle\cdot,\cdot\rangle$ identifies a measurement $\mathcal{M}_j$ as an infinite series of integrals. While the integration domain of these integrals corresponds to the Cartesian product of $(\partial\mathcal{V}\times S^2)^i, i \geq 2$, the integrands are given by the products of the emitted importance W_e^j and the single or repeated application of linear operators $\mathbf{G}^{\partial\mathcal{V}}$ and $\mathbf{K}^{\partial\mathcal{V}}$ on the emitted radiance $L_{e,i}$. That is, except of the integrand in the first inner product all other integrands are multi-dimensional integrals, where the dimension of these integrals depends on its position within the series.

$\langle\cdot,\cdot\rangle$ (416) $\mathbf{G}^{\partial\mathcal{V}}$ (430) $\mathbf{K}^{\partial\mathcal{V}}$ (432)

Using the integral representation of the local light scattering operator $\mathbf{K}^{\partial\mathcal{V}}$ and utilizing, with respect to the light propagation operator $\mathbf{G}^{\partial\mathcal{V}}$, the principle of radiance invariance in a vacuum, then the measurement equation can be written in the 3-point form as follows:

Radiance Invariance (253)

$$\begin{aligned}\mathcal{M}_j \overset{(4.428)}{=} &\int_{\partial\mathcal{V}^2} L_e(\mathbf{x}_0 \to \mathbf{x}_1)\, \mathcal{G}(\mathbf{x}_0 \leftrightarrow \mathbf{x}_1)\, W_e^j(\mathbf{x}_1 \to \mathbf{x}_0)\, d\mu^2(\mathbf{x}_0) d\mu^2(\mathbf{x}_1) + \\ &\int_{\partial\mathcal{V}^3} L_e(\mathbf{x}_0 \to \mathbf{x}_1)\, \mathcal{G}(\mathbf{x}_0 \leftrightarrow \mathbf{x}_1)\, f_s(\mathbf{x}_0 \to \mathbf{x}_1 \to \mathbf{x}_2) \mathcal{G}(\mathbf{x}_1 \leftrightarrow \mathbf{x}_2) \cdot \\ &\quad W_e^j(\mathbf{x}_2 \to \mathbf{x}_1)\, d\mu^2(\mathbf{x}_0) d\mu^2(\mathbf{x}_1) d\mu^2(\mathbf{x}_2) + \cdots,\end{aligned} \tag{5.183}$$

3-point Form (403)

where we have used L_e instead of $L_{e,i}$ for expressing the emitted radiance, see Figure 5.11.

A closed formula for the *measurement contribution function*, f_j, separately defined for each path length k, then looks like this:

$$\begin{aligned}f_j(\overline{\mathbf{x}}) \overset{def}{=}\ & L_e(\mathbf{x}_0 \to \mathbf{x}_1) \cdot \\ & \prod_{i=1}^{k-1} \mathcal{G}(\mathbf{x}_{i-1} \leftrightarrow \mathbf{x}_i)\, f(\mathbf{x}_{i-1} \to \mathbf{x}_i \to \mathbf{x}_{i+1}) \cdot \\ & \mathcal{G}(\mathbf{x}_{k-1} \leftrightarrow \mathbf{x}_k)\, W_e^j(\mathbf{x}_k \to \mathbf{x}_{k-1}),\end{aligned} \tag{5.184}$$

whereas $\mathcal{G}$ is the so-called *geometry term* defined by

$$\mathcal{G}(\mathbf{x}_{k-1} \leftrightarrow \mathbf{x}_k) \overset{def}{=} \mathcal{V}(\mathbf{x}_{k-1} \leftrightarrow \mathbf{x}_k) \cdot \frac{\left|\cos\theta_o^{k-1} \cos\theta_i^k\right|}{\|\mathbf{x}_{k-1} - \mathbf{x}_k\|_2^2} \tag{5.185}$$

and $\mathcal{V}$ is the visibility function from Definition 2.1.

EXAMPLE 5.7 *Let us consider once more the scene from Figure 5.10 but now extended by a surface $\partial\mathcal{V}_0$ representing a light source. Obviously, the set $\mathbf{P}_{\overline{0131}}$ from Figure 5.12 corresponds to paths starting at a light source, passing over the left wall and the floor and ending at the left wall of the scene. The integral associated with these*

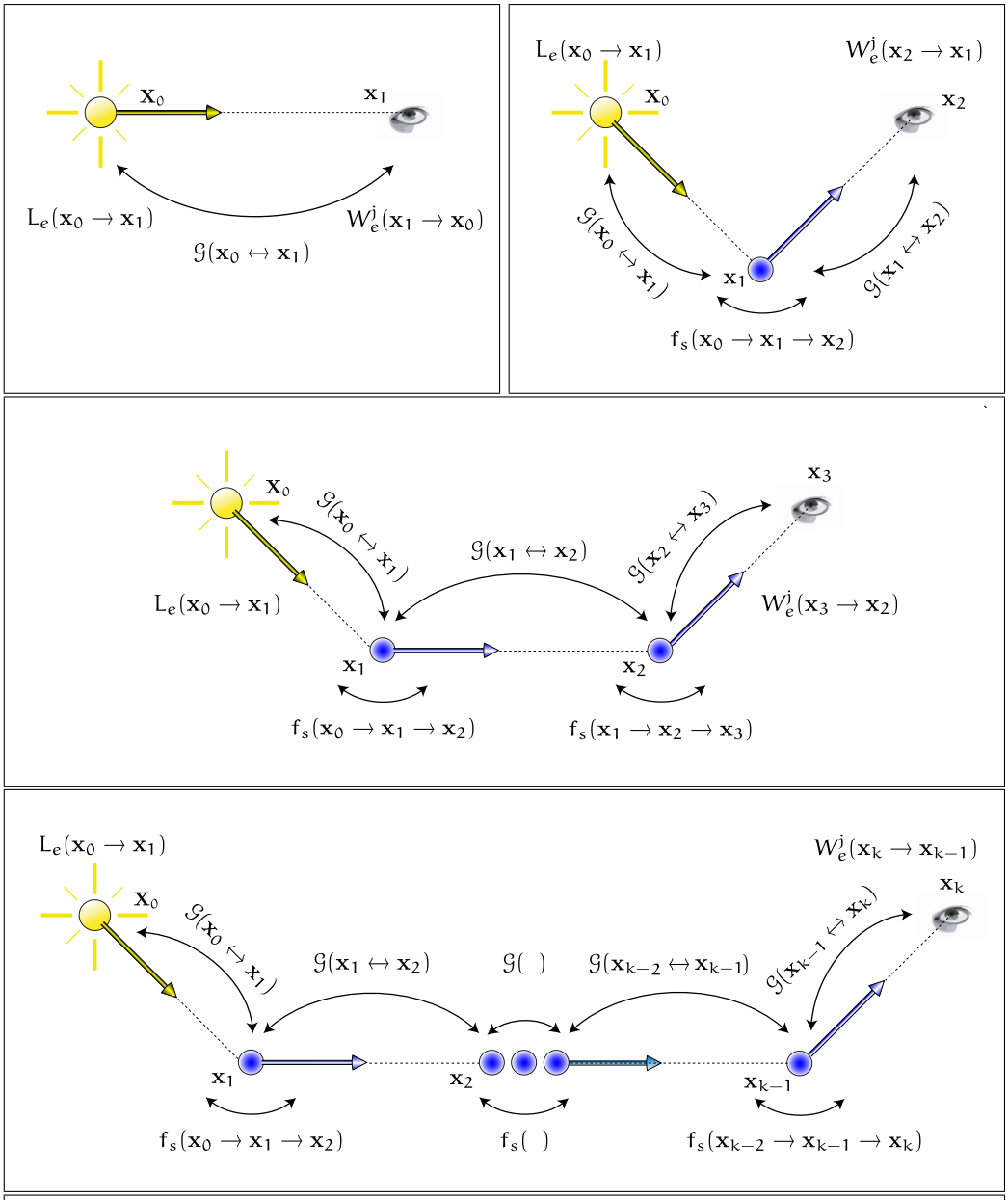


FIGURE 5.11: MEASUREMENT CONTRIBUTION FUNCTION IN A VACUUM. Shown, are paths of lengths $1, 2, 3$ and k. The measurement contribution function f_j corresponds to the product of the emitted radiance at the starting point x_0 and the emitted importance at the end points of the paths as well as the geometry term between two consecutive nodes of each path and the BSDFs at the path nodes between the starting node and the end node.

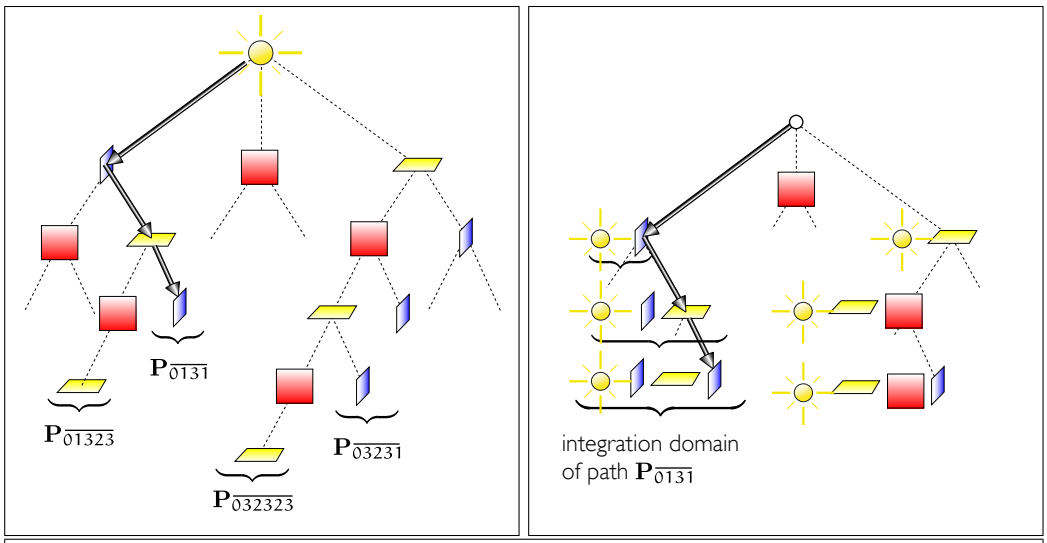


FIGURE 5.12: INTEGRATION OF THE MEASUREMENT CONTRIBUTION FUNCTION IN A VACUUM. Left, a scene consisting of four object surfaces, it holds: $\partial\mathcal{V} = \{\partial\mathcal{V}_0, \partial\mathcal{V}_1, \partial\mathcal{V}_2, \partial\mathcal{V}_3\}$. Right, a small part of the associated measurement contribution function. The symbols at a tree node can be interpreted as the integration domains. Thus, the path $\mathbf{P}_{\overline{0131}}$ of length 3 requires first to integrate via the light source, the left wall, the floor, and the left wall once again.

paths then has the form

$$\int_{\mathbf{P}_{\overline{0131}}} f_j(\overline{\mathbf{x}})\, d\mu_3(\overline{\mathbf{x}}) \tag{5.186}$$

$$\overset{(5.184)}{=} \int_{\partial\mathcal{V}_0\times\partial\mathcal{V}_1\times\partial\mathcal{V}_3\times\partial\mathcal{V}_1} L_e(\mathbf{x}_0 \to \mathbf{x}_1)\, \mathcal{G}(\mathbf{x}_0 \leftrightarrow \mathbf{x}_1) \cdot f_s(\mathbf{x}_0 \to \mathbf{x}_1 \to \mathbf{x}_3)\, \mathcal{G}(\mathbf{x}_1 \leftrightarrow \mathbf{x}_3)\, f_s(\mathbf{x}_1 \to \mathbf{x}_3 \to \mathbf{x}_1') \cdot \mathcal{G}(\mathbf{x}_3 \leftrightarrow \mathbf{x}_1')\, W_e^j(\mathbf{x}_3 \to \mathbf{x}_1')\, d\mu^2(\mathbf{x}_1')\, d\mu^2(\mathbf{x}_3)\, d\mu^2(\mathbf{x}_1)\, d\mu^2(\mathbf{x}_0),$$

where $\mathbf{x}_0 \in \partial\mathcal{V}_0, \mathbf{x}_1 \in \partial\mathcal{V}_1, \mathbf{x}_3 \in \partial\mathcal{V}_3$, *and* $\mathbf{x}_1' \in \partial\mathcal{V}_1$.

Note: This integral contributes only a small fraction of light to the shading of a pixel—namely, the fraction of light that reaches the eye via paths from type $\mathbf{P}_{\overline{0131}}$.

Putting all these things together, then the path integral formulation of the stationary light transport in a vacuum can be defined as follows:

DEFINITION 5.19 (The Path Integral Formulation of Stationary Light Transport in a Vacuum) *Let* $\mathbf{P}^\infty$ *be the space of all paths of finite lengths and* μ_∞ *the continuous path*

measure, then the path integral formulation of stationary light transport in a vacuum *is given by*

$$\begin{aligned} \mathcal{M}_j &= \int_{\mathbf{P}^\infty} f_j(\overline{\mathbf{x}}) d\mu_\infty(\overline{\mathbf{x}}) \qquad (5.187) \\ &= \sum_{k=1}^{\infty} \int_{\partial\mathcal{V}} \dots \int_{\partial\mathcal{V}} L_e(\mathbf{x}_0 \to \mathbf{x}_1) \cdot \prod_{i=1}^{k-1} G(\mathbf{x}_0 \leftrightarrow \mathbf{x}_1)\, f(\mathbf{x}_{i-1} \to \mathbf{x}_i \to \mathbf{x}_{i+1}) \cdot \\ &\qquad G(\mathbf{x}_{k-1} \leftrightarrow \mathbf{x}_k) W_e^j(\mathbf{x}_k \to \mathbf{x}_{k-1})\, d\mu^2(\mathbf{x}_0) \dots d\mu^2(\mathbf{x}_k). \end{aligned}$$

5.4.2 THE PATH INTEGRAL MODEL OF LIGHT TRANSPORT IN PARTICIPATING MEDIA

We will now extend our path integral formulation, introduced in the previous section, to handle participating media. For that purpose, we must also take into account the interaction of light with particles at spatial regions of $\mathbb{R}^3$ filled with participating media. This requires to drill the path space $\mathbf{P}^\infty$, to include also all path of finite lengths that start, end, or pass through participating media, which involves also to construct a new extended continuous path measure.

THE EXTENDED PATH SPACE AND THE EXTENDED CONTINUOUS PATH MEASURE. For the following discussion, let $\partial\mathcal{V} = \cup_{i=0}^{n} \partial\mathcal{V}_i$ be a finite set of surfaces in $\mathbb{R}^3$ that can be used to model the objects in the scene to be rendered, and $\mathcal{V}^o = \cup_{i=0}^{m} \mathcal{V}_i^o$ be a finite set of spatial regions of $\mathbb{R}^3$, where volumetric interactions of light can occur. $\partial\mathcal{V}$ (41) $\mathcal{V}^o$ (41)

Let us now consider the Cartesian product of $k+1$ object surfaces or spatial regions $\overline{\mathcal{V}}_{i_j} \in \mathcal{V} = \partial\mathcal{V} \cup \mathcal{V}^o$, thus, Cartesian Product (829)

$$\overline{\mathbf{P}}_{\overline{i_0 \dots i_k}} \stackrel{def}{=} \overline{\mathcal{V}}_{i_0} \times \overline{\mathcal{V}}_{i_1} \times \dots \times \overline{\mathcal{V}}_{i_k}, \qquad (5.188)$$

where $0 \le i_j \le k$ and $\overline{\mathcal{V}}_{i_j} \neq \overline{\mathcal{V}}_{i_{j+1}}$ for $0 \le j \le k-1$. Obviously $\overline{\mathbf{P}}_{\overline{i_0 \dots i_k}}$ can be considered as the set of paths $\overline{\mathbf{x}} = \mathbf{x}_{i_0} \mathbf{x}_{i_1} \dots \mathbf{x}_{i_k}$ of length k, starting at $\overline{\mathcal{V}}_{i_0}$ and ending at $\overline{\mathcal{V}}_{i_k}$ with $\mathbf{x}_{i_j}$ from $\overline{\mathcal{V}}_{i_j}$ for $0 \le j \le k$. We can now extend this set to the set $\overline{\mathbf{P}}_k$, i.e. the set of all paths of length k starting and ending at any object surface or spatial region of $\mathcal{V}$ by defining

$$\overline{\mathbf{P}}_k \stackrel{def}{=} \bigcup_{\overline{i_0 \dots i_k} \in \{0,1,\dots,n\}} \overline{\mathbf{P}}_{\overline{i_0 \dots i_k}}, \qquad (5.189)$$

where two neighbored components of the tupel $\overline{i_0 \dots i_k}$ must always be different.

Based on this construct, we can now define the *extended path space*, $\overline{\mathbf{P}}^\infty$, that is, the space of all paths of finite lengths over object surfaces or spatial regions within participating media from $\mathcal{V}$.

DEFINITION 5.20 (The Extended Path Space of All Paths of Finite Length, $\overline{\mathbf{P}}^\infty$) *The* extended path space, $\overline{\mathbf{P}}^\infty$, *that is, the set of paths of all finite lengths, is defined by*

$$\overline{\mathbf{P}}^\infty \stackrel{\text{def}}{=} \bigcup_{k=1}^{\infty} \overline{\mathbf{P}}_k. \tag{5.190}$$

Let us now consider the set $\mathfrak{B}(\overline{\mathcal{V}}_{i_j}), 0 \leq j \leq k$, which is either the set generated over open rectangles of $\partial\mathcal{V}_{i_j}$ or the set generated over open cuboids of $\mathcal{V}^o_{i_j}$. Since each of the sets $\mathfrak{B}(\overline{\mathcal{V}}_{i_j})$ is a σ-algebra, namely the Borel σ-algebra over $\overline{\mathcal{V}}_{i_j}$, it follows from measure theory that also the Cartesian product

σ-algebra (828)
$\mathfrak{B}(\cdot)$ (865)
Cartesian Product (829)

$$\mathfrak{B}(\overline{\mathcal{V}}_{i_0} \times \overline{\mathcal{V}}_{i_1} \times \cdots \times \overline{\mathcal{V}}_{i_k}) \stackrel{\text{def}}{=} \mathfrak{B}(\overline{\mathcal{V}}_{i_0}) \times \cdots \times \mathfrak{B}(\overline{\mathcal{V}}_{i_k}), \tag{5.191}$$

$0 \leq k \leq n$, is also a σ-algebra. With the help of the Lebesgue measures μ^2 and μ^3, we can then construct a measure $\overline{\mu}_k$ on $\mathfrak{B}(\overline{\mathcal{V}}_{i_0} \times \overline{\mathcal{V}}_{i_1} \times \cdots \times \overline{\mathcal{V}}_{i_k})$ by

μ^2 (82)

$$\begin{aligned} \overline{\mu}_k(\mathbf{B}) &\stackrel{\text{def}}{=} \overline{\mu}_k(B_0 \times \cdots \times B_k) && (5.192)\\ &= \overline{\mu}_1(B_0) \cdot \ldots \cdot \overline{\mu}_1(B_k), && (5.193) \end{aligned}$$

where $\mathbf{B} = B_0 \times \cdots \times B_k$ and $\overline{\mu}_1(B_j)$ is defined by

Measure (79)

$$\overline{\mu}_1(B_i) \stackrel{\text{def}}{=} \begin{cases} \mu^2(B_j) & \text{if } B_j \in \partial\mathcal{V}_{i_j} \\ \mu^3(B_j) & \text{if } B_j \in \mathcal{V}^o_{i_j}. \end{cases} \tag{5.194}$$

Via the construction of the measure $\overline{\mu}_k$ from Equation (5.192) we now define the *extended continuous path measure* $\overline{\mu}_\infty$, which allows to extend the path space $\overline{\mathbf{P}}^\infty$ to the measure space $(\overline{\mathbf{P}}^\infty, \mathfrak{B}(\overline{\mathbf{P}}^\infty), \overline{\mu}_\infty)$.

Measure (79)
Measure Space (80)

DEFINITION 5.21 (The Extended Measure Space $(\overline{\mathbf{P}}^\infty, \mathfrak{B}(\overline{\mathbf{P}}^\infty), \overline{\mu}_\infty)$) *Let* $\overline{\mathbf{P}}^\infty$ *be the extended path space of all paths of all finite lengths and let* $\overline{\mu}_k$ *be the path measure from Equation (5.192) defined on the* σ*-algebra* $\mathfrak{B}(\overline{\mathcal{V}}_{i_0} \times \overline{\mathcal{V}}_{i_1} \times \cdots \times \overline{\mathcal{V}}_{i_k})$. *Defining*

Measure (79)
$\mathfrak{B}(\cdot)$ (865)

$$\begin{aligned} \overline{\mu}_\infty(\mathbf{B}) &\stackrel{\text{def}}{=} \overline{\mu}_\infty\left(\mathbf{B} \cap \bigcup_{k=1}^{\infty} \overline{\mathbf{P}}_k\right) && (5.195)\\ &= \overline{\mu}_\infty\left(\bigcup_{k=1}^{\infty} (\mathbf{B} \cap \overline{\mathbf{P}}_k)\right) && (5.196)\\ &= \sum_{k=1}^{\infty} \overline{\mu}_k\left(\mathbf{B} \cap \overline{\mathbf{P}}_k\right), && (5.197) \end{aligned}$$

where $\mathbf{B} = B_{i_0} \times \cdots \times B_{i_k} \subseteq \overline{\mathcal{V}}_{i_0} \times \overline{\mathcal{V}}_{i_1} \times \cdots \times \overline{\mathcal{V}}_{i_k}$ *is a subset of the set of all paths of length* k, $\overline{\mathbf{P}}_k$, *then* $\overline{\mu}_\infty$ *is a measure, the so-called* extended continuous path measure. *Applied to the path space* $\overline{\mathbf{P}}^\infty$, *the triple* $(\overline{\mathbf{P}}^\infty, \mathfrak{B}(\overline{\mathbf{P}}^\infty), \overline{\mu}_\infty)$ *is obviously a measure space.*

Measure Space (80)

THE MEASUREMENT CONTRIBUTION FUNCTION. With the extend path space $\overline{\mathbf{P}}^{\infty}$ as integration domain and the extended continuous path measure $\overline{\mu}_{\infty}$ as the integration measure, we have two of three constructs that are required to formulate the path integral model of light transport in participating media. The only thing that we still need to represent the measurement equation in form of a path integral, that is, Measurement Equation (416)

$$\mathcal{M}_j = \int_{\overline{\mathbf{P}}^{\infty}} \bar{f}_j(\bar{\mathbf{x}}) d\overline{\mu}_{\infty}(\bar{\mathbf{x}}), \tag{5.198}$$

is the choice of the integrand $\bar{f}_j$.

Now, from our discussion in the previous sections, we know, that the measurement equation can be written as the inner product of two measurable functions W_e and $L_{e,i}$, thus, Measurement Equation (416) Measurable Function (98)

$$\begin{aligned}
\mathcal{M}_j &= \langle W_e^j, \overline{\mathbf{S}} L_{e,i} \rangle && (5.199)\\
&\overset{(5.88)}{=} \left\langle W_e^j, \sum_{i=0}^{\infty} \overline{\mathbf{T}}_{L_i}^{i} L_{e,i} \right\rangle && (5.200)\\
&\overset{(5.84)}{=} \left\langle W_e^j, \sum_{i=0}^{\infty} \left(\overline{\mathbf{T}}_{L_i}^{\partial\mathcal{V}} + \overline{\mathbf{T}}_{L_i}^{\mathcal{V}^o} \right)^{i} L_{e,i} \right\rangle. && (5.201)
\end{aligned}$$

This equation can then be rephrased in the more readable form, namely as,

$$\begin{aligned}
\mathcal{M}_j = \; & \langle W_e^j, L_{e,i} \rangle + \\
& \langle W_e^j, \overline{\mathbf{T}}_{L_i}^{\partial\mathcal{V}} L_{e,i} \rangle + \langle W_e^j, \overline{\mathbf{T}}_{L_i}^{\mathcal{V}^o} L_{e,i} \rangle + \\
& \langle W_e^j, \overline{\mathbf{T}}_{L_i}^{\partial\mathcal{V}} \overline{\mathbf{T}}_{L_i}^{\partial\mathcal{V}} L_{e,i} \rangle + \langle W_e^j, \overline{\mathbf{T}}_{L_i}^{\partial\mathcal{V}} \overline{\mathbf{T}}_{L_i}^{\mathcal{V}^o} L_{e,i} \rangle + \langle W_e^j, \overline{\mathbf{T}}_{L_i}^{\mathcal{V}^o} \overline{\mathbf{T}}_{L_i}^{\partial\mathcal{V}} L_{e,i} \rangle + \cdots .
\end{aligned}$$

In this equation, the first inner product describes the light transport between a surface or a volumetric emitter and the camera lens or another sensor, while the next both inner products represent the light transport between a light source and the sensor, where a single scattering event at a surface or within a participating medium occurs. The third line then describes the contribution of a source whose emitted light arrives at the eye after two scattering events at surfaces, or within a participating medium, or a scattering event at a surface followed by a scattering event at a point within a medium, and vice versa.

Similar to our discussion in the previous section, the inner product $\langle \cdot, \cdot \rangle$ identifies the measurement $\mathcal{M}_j$ as an infinite series of integrals, where the integration domain is now given by Cartesian products build over $\partial\mathcal{V} \times \mathcal{V}$. The integrands are given by the products of the emitted importance W_e^j and the single or repeated application of the linear transport operators $\mathbf{T}_{L_i}^{\partial\mathcal{V}}$ and $\mathbf{T}_{L_i}^{\mathcal{V}^o}$ on the emitted radiance $L_{e,i}$. That is, except of the integrand in the first inner product all other integrands are integrals over $(\partial\mathcal{V} \times \mathcal{V})^i, i > 2$, where the dimension of the domains depends on its position within the series.

Using the integral representation of the local light scattering operator $\mathbf{K}^{\partial\mathcal{V}}$ and utilizing, with respect to the light propagation operator $\mathbf{G}^{\partial\mathcal{V}}$, the principle of radiance invariance in a vacuum, then the measurement equation can be written in the 3-point form as follows

Radiance Invariance (253)

3-point Form (403)

$$\begin{aligned}\mathcal{M}_j \overset{(4.428)}{=} & \int_{(\partial\mathcal{V}\cup\mathcal{V})^2} L_e(\mathbf{x}_0 \to \mathbf{x}_1)\, \mathcal{G}(\mathbf{x}_0 \leftrightarrow \mathbf{x}_1)\, W_e^j(\mathbf{x}_1 \to \mathbf{x}_0)\, d\overline{\mu}_1(\mathbf{x}_0) d\overline{\mu}_1(\mathbf{x}_1) + \\ & \int_{(\partial\mathcal{V}\cup\mathcal{V})^3} L_e(\mathbf{x}_0 \to \mathbf{x}_1)\, \mathcal{G}(\mathbf{x}_0 \leftrightarrow \mathbf{x}_1)\, f_s(\mathbf{x}_0 \to \mathbf{x}_1 \to \mathbf{x}_2) \mathcal{G}(\mathbf{x}_1 \leftrightarrow \mathbf{x}_2) \\ & \cdot W_e^j(\mathbf{x}_2 \to \mathbf{x}_1)\, d\overline{\mu}_2(\mathbf{x}_0) d\overline{\mu}_2(\mathbf{x}_1) d\overline{\mu}_2(\mathbf{x}_2) + \ldots\end{aligned} \tag{5.202}$$

where we have used L_e instead of $L_{e,i}$ for expressing the emitted radiance, see Figure 5.13.

Using the integral representations of the involved linear integral operators, then we get—analog to our discussion for deriving the measurement contribution function of the stationary light transport in a vacuum—the following relations:

$$f_s(\mathbf{x}_{i-1} \to \mathbf{x}_i \to \mathbf{x}_{i+1}) \overset{def}{=} \begin{cases} f_s(\mathbf{x}_{i-1} \to \mathbf{x}_i \to \mathbf{x}_{i+1}) & \text{if } \mathbf{x}_i \in \partial\mathcal{V} \\ \sigma_s\, p(\mathbf{x}_{i-1} \to \mathbf{x}_i \to \mathbf{x}_{i+1}) & \text{if } \mathbf{x}_i \in \mathcal{V}^o \end{cases} \tag{5.203}$$

for the scattering and the phase functions,

$$L_e(\mathbf{x}_0 \to \mathbf{x}_1) \overset{def}{=} \begin{cases} L_e^{\partial\mathcal{V}}(\mathbf{x}_0 \to \mathbf{x}_1) & \text{if } \mathbf{x}_0 \in \partial\mathcal{V} \\ L_e^{\mathcal{V}^o}(\mathbf{x}_0 \to \mathbf{x}_1) & \text{if } \mathbf{x}_0 \in \mathcal{V}^o \end{cases} \tag{5.204}$$

for the emitted radiance from points on object sources or point emitters in participating media, and

$$\mathcal{G}(\mathbf{x}_{i-1} \leftrightarrow \mathbf{x}_i) \overset{def}{=} \mathcal{V}(\mathbf{x}_{i-1} \leftrightarrow \mathbf{x}_i) \cdot \mathcal{G}'(\mathbf{x}_{i-1} \leftrightarrow \mathbf{x}_i) \cdot \beta(\mathbf{x}_{i-1} \leftrightarrow \mathbf{x}_i) \tag{5.205}$$

with

$$\mathcal{G}'(\mathbf{x}_{i-1} \leftrightarrow \mathbf{x}_i) \overset{def}{=} \begin{cases} \frac{\left|\cos\theta_o^{i-1} \cos\theta_i^i\right|}{\|\mathbf{x}_{i-1} - \mathbf{x}_i\|^2} & \text{if } \mathbf{x}_{i-1}, \mathbf{x}_i \in \partial\mathcal{V} \\ \frac{\left|\cos\theta_o^{i-1}\right|}{\|\mathbf{x}_{i-1} - \mathbf{x}_i\|^2} & \text{if } \mathbf{x}_{i-1} \in \partial\mathcal{V}, \mathbf{x}_i \in \mathcal{V}^o \\ \frac{\left|\cos\theta_i^i\right|}{\|\mathbf{x}_{i-1} - \mathbf{x}_i\|^2} & \text{if } \mathbf{x}_{i-1} \in \mathcal{V}^o, \mathbf{x}_i \in \partial\mathcal{V} \\ \frac{1}{\|\mathbf{x}_{i-1} - \mathbf{x}_i\|^2} & \text{if } \mathbf{x}_{i-1}, \mathbf{x}_i \in \mathcal{V}^o \end{cases} \tag{5.206}$$

for the geometry term.

Putting all these things together, then the path integral formulation of stationary light transport in participating can be defined as

DEFINITION 5.22 (Path Integral Formulation of Stationary Light Transport in Participating Media) *Let $\overline{\mathbf{P}}^\infty$ be the extended space of all paths of all finite lengths and $\overline{\mu}_\infty$ the*

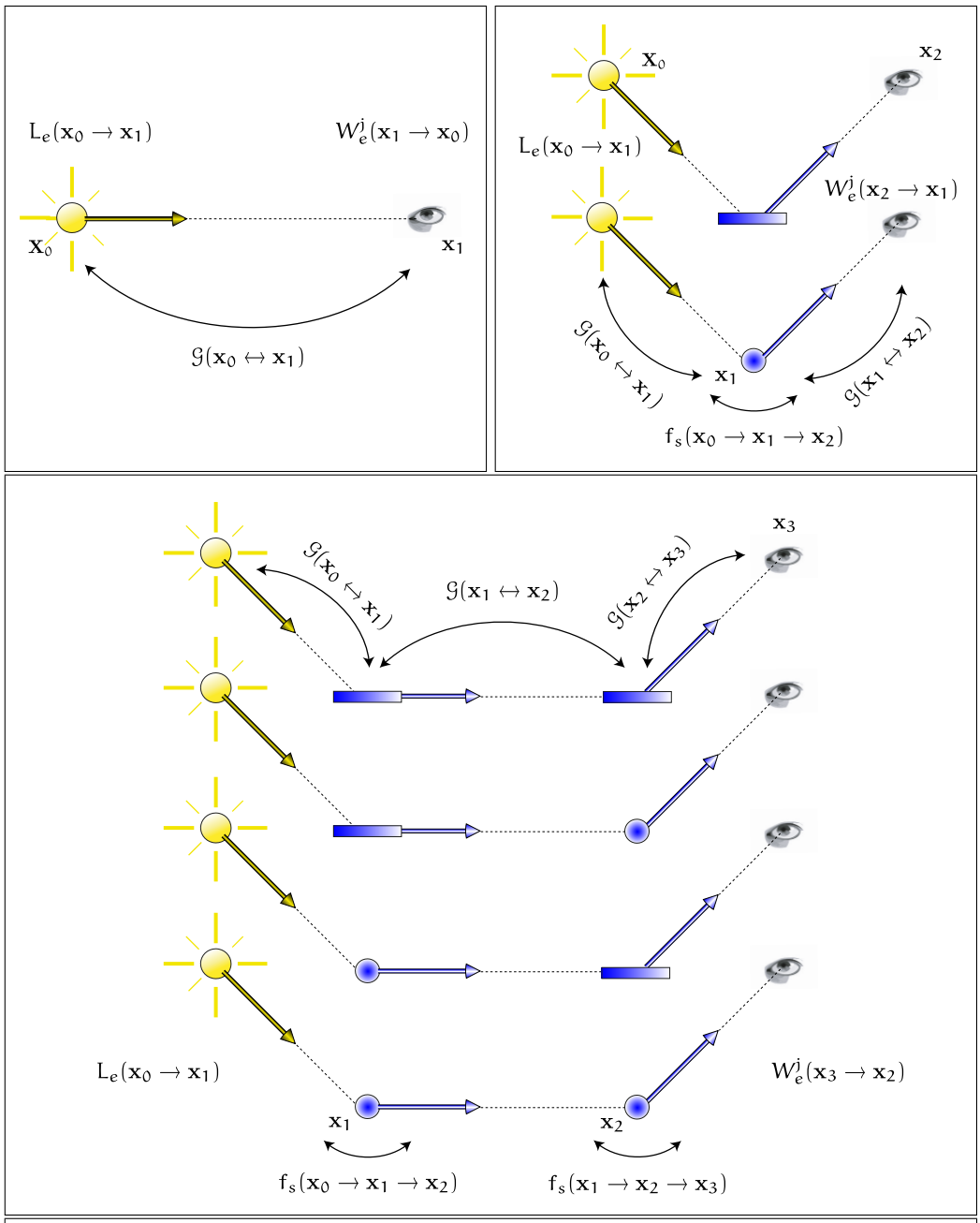


FIGURE 5.13: MEASUREMENT CONTRIBUTION FUNCTION IN PARTICIPATING MEDIA. Shown, are paths of lengths 1, 2 and 3. The measurement contribution function f_j corresponds to the product of the emitted radiance at the starting point x_0 and the emitted importance at the end points of the paths as well as the geometric term between two consecutive nodes of each path and the BSDFs respectively the phase functions at the path nodes between the starting node and the end node.

extended continuous path measure, then the path integral formulation of stationary light transport in participating media *is given by*

$$\mathcal{M}_j = \int_{\overline{\mathbf{P}}^\infty} \bar{f}_j(\overline{\mathbf{x}}) d\overline{\mu}_\infty(\overline{\mathbf{x}}), \tag{5.207}$$

where the extended measurement contribution function $\bar{f}_j$ *is given by*

$$\begin{aligned} \bar{f}_j \stackrel{def}{=} & L_e(\mathbf{x}_0 \to \mathbf{x}_1) \cdot \\ & \prod_{i=1}^{k-1} G(\mathbf{x}_{i-1} \leftrightarrow \mathbf{x}_i) f_s(\mathbf{x}_{i-1} \to \mathbf{x}_i \to \mathbf{x}_{i+1}) \cdot \\ & G(\mathbf{x}_{k-1} \leftrightarrow \mathbf{x}_k) W_e^j(\mathbf{x}_k \to \mathbf{x}_{k-1}). \end{aligned} \tag{5.208}$$

5.5 THE GLOBAL REFLECTANCE DISTRIBUTION FUNCTION

In the last sections we have presented a series of equivalent mathematical formulations of the global illumination problem. So, we have seen, that the global illumination problem can be solved by combining the incident or exitant radiance distribution within the scene with the importance emitted by a sensor. As light and importance transport are dual recursive formulations, the measurement equation, representing the global illumination problem, can also be solved by combining the importance distribution within the scene by the radiance emitted from light sources. With the path integral formulation, we then introduced an elegant mathematical formulation of light transport as a simple integral over the space of all paths within a scene. We will now shortly present another interesting approach for evaluating the measurement equation. Similar as the path integral formulation, it is not of recursive nature, but it is based on a direct function, the so-called *global reflectance distribution function, GRDF*, [116, Lafortune 1996] and [50, Dutré & al. 2003].

SLTEV (398)
SITEV (413)
Measurement Equation (416)

BRDF (320)

THE GLOBAL REFLECTANCE DISTRIBUTION FUNCTION. Let us recall, the BRDF was defined via the ratio of the exitant radiance and the irradiance at a given point. The GRDF is introduced in a similar way, but in contrast to the BRDF, that specifies the local behavior of light reflected at a single surface of the environment, the GRDF describes the interaction of light at all surfaces of the scene. Due to [50, Dutré & al. 2003], the global reflectance distribution function can be defined as follows:

∂V (41)

DEFINITION 5.23 (The Global Reflectance Distribution Function, G_r) *Let* $\partial\mathcal{V}$ *be a set of 2-dimensional surfaces in* $\mathbb{R}^3$, $\mathbf{s}_i, \mathbf{s}_o$ *be points on two different surfaces* $A, A' \in \partial\mathcal{V}$, *and* $\mathcal{H}^2_i$ *as well as* $\mathcal{H}^2_o$ *be the incident and exitant hemispheres at the points* $\mathbf{s}_i, \mathbf{s}_o$, *which refer to the same set of directions. Then, we call the measurable function,* G_r,

Measurable Function (98)

defined by

$$G_r : \partial\mathcal{V} \times \mathcal{H}_i^2 \times \partial\mathcal{V} \times \mathcal{H}_o^2 \to [0, \infty] \tag{5.209}$$

with

$$G_r((s_i, \omega_i) \to (s_o, \omega_o)) \stackrel{def}{=} \frac{d^2 L_o(s_o, \omega_o)}{dE(s_i, \omega_i)\, d\mu^2(s_i)} = \frac{d^2 L_o(s_o, \omega_o)}{L_i(s_i, \omega_i)\, d\sigma^\perp_{s_i}(\omega_i)\, d\mu^2(s_i)}, \tag{5.210}$$

the global reflectance distribution function, *also briefly denoted as the* GRDF.

Obviously, the GRDF gives a measure for the fraction of light emitted at each position and direction within a scene that is eventually radiated through any other position and direction. It has units of $\frac{1}{m^2\, sr}$ and only depends on the geometry of the scene and the material properties of the surfaces within a scene. Due to the principle of radiance invariance in a vacuum, the Helmholtz reciprocity is also valid for the GRDF. Since it is a function of two points and two directions, the GRDF is, similar as the BSSRDF, a function of 8 variables, that requires enormous cost of run time if it has to be sampled within a rendering procedure, or a huge amount of memory is needed if a discretized version is used as an approximation.

Radiance Invariance (253)
Helmholtz Reciprocity (331)
BSSRDF (318)

REMARK 5.19 *It should be clear that the BRDF can be interpreted as a special case of the GRDF. Considering the incident radiance at a single point* s *and identifying this surface point with* s_o*, then Definition 5.23 is identical to the definition of the BRDF.*

BRDF (320)

Furthermore, the definition of the GRDF is not restricted to the radiometric quantities radiance and irradiance, but it can also be defined via the via incident and exitant importance, namely by:

$$G_r((s_i, \omega_i) \to (s_o, \omega_o)) \stackrel{def}{=} \frac{d^2 W_o(s_o, \omega_o)}{W_i(s_i, \omega_i)\, d\sigma^\perp_{s_i}(\omega_i)\, d\mu^2(s_i)}. \tag{5.211}$$

THE MEASUREMENT EQUATION EXPRESSED IN TERMS OF THE GRDF. Let us assume that all surfaces within a scene are opaque, then the measurement equation from Equation (4.45) can be written in the hemispherical form as:

$$\mathcal{M} = \int_{\partial\mathcal{V}} \int_{\mathcal{H}_i^2(s)} W_e(s, \omega)\, L_i(s, \omega)\, d\sigma^\perp_s(\omega)\, d\mu^2(s). \tag{5.212}$$

As the GRDF describes the entire light transport in the scene between two pairs (s, ω) and (s', ω'), the measurement equation can now also be evaluated via computing

$$\begin{aligned} \mathcal{M} &= \int_{\partial\mathcal{V}} \int_{\mathcal{H}_o^2,(s)} \int_{\partial\mathcal{V}} \int_{\mathcal{H}_i^2(s)} L_e(s, \omega)\, G_r((s, \omega) \to (s', \omega'))\, W_e(s', \omega') \\ &\qquad\qquad d\sigma^\perp_{s'}(\omega')\, d\mu^2(s')\, d\sigma^\perp_s(\omega)\, d\mu^2(s) \qquad (5.213) \\ &= \langle L_e, \langle G, W_e \rangle \rangle, \qquad (5.214) \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the inner product defined on the Lebesgue space over $\partial\mathcal{V} \times \mathcal{H}^2_i$ respectively $\partial\mathcal{V} \times \mathcal{H}^2_o$. $\mathcal{L}^2(\cdot,\cdot)$ (110)

REMARK 5.20 *Obviously, the global illumination problem can be captured via the GRDF in a very short and elegant form. This formulation is independent of any initial distributions for self-emitted radiance or importance and, as already mentioned, only dependent on the geometry and the material properties of the surfaces within a scene. This then also leads to a simple representation of the measurement equation, namely as an inner product. For a detailed discussion of the concept of the GRDF and its properties, see [116, Lafortune 1996] and [50, Dutre & al. 2003], [51, Dutre & al. 2006].* $\langle \cdot, \cdot \rangle$ (859)

In Section 9.3, we will encounter bidirectional path tracing *a Monte Carlo algorithm for solving the light transport equation, that can be implemented via use of the global reflectance distribution function, [116, Lafortune 1996].*

5.6 REFERENCE LITERATURE AND FURTHER READING

Our operator model of light and importance transport in a vacuum is based on the corresponding operator models firstly introduced in [220, Veach 1997]. A similar model, based on BRDFs is presented in [10, Arvo 1995]. Both approaches are build on linear function spaces equipped with the necessary formalism from functional analysis as well as measure and integration theory. Compared to [220, Veach 1997] and [10, Arvo 1995], who consider the light and importance transport only in a vacuum, our operator model is also valid for light transport in participating media. A similar, but mathematically less stringent, not on functional analytically concepts based approach can be found in [152, Pauly 1999].

The light and importance transport is also described with the help of a more intuitive operator model in [47, Dutré 1996], [50, Dutré 2003]. We recommend these sources for the more practical oriented reader since no deeper knowledges from functional analysis and measure as well as integration theory are required for understanding. In [47, Dutré 1996], [50, Dutré 2003], you can also find a derivation of the GRDF as an extension of the BRDF to describe multiple scattering effects from all surfaces. As another reference to the GRDF, we recommend [116, Lafortune 1996], where the concept of the global reflectance distribution function was firstly introduced. It served as a template for Section 5.5.

The operator model of importance as the adjoint of light is discussed in detail in [34, Christensen 1995]. In [35, Christensen 2003] then an attempt is made to clarify the various uses of adjoints and importance in rendering by unifying them into a single framework.

The idea to interpret the global illumination problem as an ordinary integral, based on the concept of the path integral, was developed by [198, Spanier & Gelbard 1969].

In [220, Veach 1997] this construct is extended to measures that have natural physical interpretations. We use Eric Veach's construct of the path integral in Section 5.4. Since the path integral as introduced in [220, Veach 1997] is restricted to the light transport within a vacuum, we had to extend all necessary constructs as to hold also in participating media. We have also slightly modified Veach's path measure so that path segments are not accounted for, if they start and end at the same domain.

CHAPTER SIX

MONTE CARLO INTEGRATION

In trying to improve the quality of the synthetic images, we do not expect to be able to display the object exactly as it would appear in reality, with texture, overcast shadows, etc. We hope only to display an image that approximates the real object closely enough to provide a certain degree of realism.
BUI TUONG PHONG, 1975

The term Monte Carlo was first mentioned in connection with a mathematical method in the 1940s during the development of atomic weapons in the laboratories of Los Alamos, USA. The aim of researchers, in particular John von Neumann, Stanislaw Ulam, and Nick Metropolis, was to utilize the advantages of the recently completed ENIAC computer for simulating neutron trajectories using random sampling. In 1949, Metropolis and Ulam published a paper entitled *The Monte Carlo Method*, due to the famous Casino in Monte Carlo, in which the method, which was not entirely new, was presented.

Though the theoretical foundations of the Monte Carlo method had been known for some time, in the absence of electronic data-processors, large-scale application had not been possible, as the modeling of random variables by hand demanded too much time and effort. Thus, the development of the Monte Carlo method into a universally applicable numerical method only became possible as the necessary technical means had been made available.

In the years to follow an increasing number of papers were published on the new method, describing its applicability to a seemingly unending variety of problems in statistical mechanics, particle transport as well as to the solution of economical models. As a result of this popularity, however, the reputation of the Monte Carlo method as a problem solution methodology gradually declined, as the advantages of the method were lost by its universal application to virtually every kind of problem. Things changed in the 1960s, however, when researchers started to examine the kinds of problems to which the Monte

Carlo method could be applied most efficiently. As the ideal application area of the classical Monte Carlo method has shown the solution of a special kind of integrals and integral equations.

The classical Newton-Cotes formulas and Gauss quadratures for evaluating high-dimensional integrals are derived from one-dimensional formulas involving a considerable amount of calculations of the function values at different locations of the integral domain. By contrast, Monte Carlo methods are independent on the dimension of the integral domain and, in particular, of the properties of the involved functions, as well as, in the case of integral equations of the allocated kernels be them continuous or discontinuous. These features, in particular the simplicity of implementation, make Monte Carlo procedures stronger and more effective than ordinary quadrature formulas for solving integrals of higher dimension than six. However, as this method is based on the selection of random variables it has the disadvantage that all the obtained solutions and error boundaries are of statistical nature.

The success of a Monte Carlo method depends on the calculation of appropriate random numbers, where the term random used here in the sense of *randomized*. In addition, as deterministic automata can be used only to generate so-called *pseudorandom numbers* but not to generate truly random numbers, appropriate random number generators and sampling techniques are required.

Compared with other numerical integration methods, Monte Carlo algorithms not only have advantages. One of it's main disadvantages is that apart from the existence of merely probabilistic error boundaries, they are highly dependent upon both, convergence results and error estimations of the involved random numbers. Another disadvantage is that even sufficiently smooth functions lead to the probabilistically slow convergence behavior of order $O\left(\frac{1}{\sqrt{n}}\right)$, which is typical for Monte Carlo integration.

OVERVIEW OF THIS CHAPTER. We begin this excursion into the theory of Monte Carlo integration with a brief review of early works in the development of mathematical algo-
Section 6.1 rithms, in which *deterministic* and *probabilistic approaches* were used for the numerical solution of integrals. Following this, we formulate an integral given over the integration domain $\mathbf{Q}^s \subset \mathbb{R}^s$ as the *expected value of a random variable* defined on the probability
Section 6.2 space generated over the Borel σ-algebra of subsets of $\mathbf{Q}^s$. Via the concept of the *Monte*
Section 6.3 *Carlo estimator* we then get—using the famous *Chebychev inequality* and as a result of the *Central Limit Theorem*—an approximate representation of the integral to be calculated, which can be interpreted as the arithmetic mean of values of the integrand at randomly chosen points of the integration domain. Following this, a summary of the most
Section 6.4 important statements on convergence behavior and run time behavior of the general Monte
Section 6.5 Carlo method will be given. After that, we turn to sampling strategies most frequently applied in practice for the choice of points from $\mathbf{Q}^s$, which are needed for the estimation of integrals. Thus, we present in cases of invertible integrands the *transformation method* as a good sampling procedure used to generate random samples from given probability

distributions and illustrate its efficiency by means of a few important examples. We also discuss *acceptance-rejection sampling* and talk about *Markov chain Monte Carlo*, a method from computational physics, which is often used for difficult sampling problems in high-dimensional spaces and which forms the basis of the *Metropolis light transport algorithm* from global illumination theory. After that, we turn to the field of *variance reduction techniques*, upon which we dwell in some detail. The methods to be presented here are *use of expected values*, the highly efficient procedure of *importance sampling*, based on the choice of a density similar to the integrand, *control variates*, and *stratified sampling*, a method of stratifying the integration domain underlying a given integral into a disjoint union of subdomains. Afterwards, we introduce *Latin hypercube sampling* as a sampling procedure enjoying high estimation in many ray tracing procedures, followed by the sampling strategy of *jittering*, *orthogonal array sampling*, of strong importance to Quasi-Monte Carlo procedures, to be discussed in the next chapter, as well as *antithetic variates*. We finish the section about variance reduction techniques with *multiple importance sampling strategies*, that are based on the idea of using more than one sampling technique to evaluate a given integral, and combining the sample values in a provably good way. At the end of the chapter, we discuss solution methods based on principles of Monte Carlo integration, which can be used for solving integral equations. So, we introduce with the *successive integral substitution* and the *Neumann series* two analytical approaches, which can be involved in probabilistic procedures for solving integral equations. Last but not least, we still present a *random walk-solution* for Fredholm integral equations based on a *discrete-time Markov process*.

Section 6.6

Section 6.7

6.1 MOTIVATING INTEGRATION VIA MONTE CARLO METHODS

Let f be a s-dimensional, real-valued, and Lebesgue-integrable function. Then, the best way to evaluate an integral of type

$$\int_{\mathbf{B}} f(\mathbf{x})\, d\mu^s(\mathbf{x}), \tag{6.1}$$

with $\mathbf{B} \subset \mathbb{R}^s$, would be to solve it analytically, i.e. to express it as some algebraic term involving functions that can easily be evaluated.

In case of $s = 1$, the Fundamental Theorem of Calculus delivers for a great class of functions such a simple evaluable expression by

$$\int_B f(x)\, d\mu(x) = F(b) - F(a), \tag{6.2}$$

where F is any of the infinitely many antiderivatives of f and the integration domain B represents an interval enclosed by points $a, b \in \mathbb{R}$ with $a < b$. Similar, but more complex formulas can be derived for the s-dimensional case via iterating the standard one-dimensional integral or the integral theorems of Gauss, Stokes, and Green, for details, see [174, Rudin 1998].

Now there are many functions that cannot be integrated analytically or that have antiderivatives that are given as infinite series or products of functions. How should we evaluate the integral of such a function?

Let us for example consider the function $f(x) = e^{-x^2}$. Obviously, we have no chance to find an antiderivative that can be written in elementary form. Such integrals cannot be evaluated exactly. For practical use, they must be approximated by means of methods from *numerical analysis*.

For integrating a function, in numerical analysis one looks for algorithms that give an estimate of the integral together with an estimate of the error from the exact value of the integral. Additionally, one demands of these algorithms that they should lead to a good approximate of the result in an reasonable amount of time. So, there is no point in developing algorithms that try to find an exact solution at the expense of run-time. Since the efficiency of an algorithm for evaluating the integral from (6.1) depends on the specific problem, there is also no *perfect* algorithm suitable for all integration problems. In the following sections we will see, that the more one knows about the specific problem, the higher the chances will be to find algorithms which solve the problem efficiently.

In numerical analysis one distinguishs between two classes of algorithms for integrating functions:

Section 6.1.1
- *deterministic procedures*, and

Section 6.2
- *Monte Carlo methods*.

While deterministic procedures evaluates integrals via asymptotic approximations or multiple quadrature techniques, Monte Carlo methods may be described as numerical methods based on random sampling with strong statistical and probabilistic flavor.

6.1.1 APPROXIMATING INTEGRALS VIA DETERMINISTIC METHODS

We know from our discussion in Section 2.3.3.2 that many mathematical problems from everyday life cannot be solved exactly, in particular those based on integral or differential equations. In those situations there are usually two approaches for solving the resulting integration problem:

- analytical approximations to solutions using *asymptotic expansions*, or
- more or less complicated, deterministic algorithms based on *quadratures*.

6.1.1.1 ASYMPTOTIC APPROXIMATIONS

A first type of deterministic methods for solving integrals are so-called *asymptotic methods*. Asymptotic methods for special classes of integrals have long been used to approximate the functions of mathematical physics that have integral representations. Two of the most common techniques are the Laplace method and the method of steepest descent or saddle-point method. This methodology can often be implemented without the use of a computer, and historically an extensive amount of development has gone into these methods, [57, Evans & Swartz 2000].

One of the oldest and most famous asymptotic approximations is *Sterling's formula* for approximating the factorial function $n!$ for large values of $n \in \mathbb{N}_0$:

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}, \tag{6.3}$$

where $\sim$ is used to denote that two functions are asymptotically, or approximately, equivalent.

The aim of asymptotic methods is to find functions that are asymptotically equivalent to the solution of the given integral. One of the most useful tools for finding asymptotic approximations is Taylor's theorem. Thus for example, we can, via the Taylor series of the exponential function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \tag{6.4}$$

easily approximate the integral

$$\int_a^b e^{-x^2}\, d\mu(x) = \int_a^b \sum_{n=0}^{\infty} \frac{(-x^2)^n}{n!}\, d\mu(x) \tag{6.5}$$

by integrating m terms of the Taylor series expansion of e^{-x^2}, namely:

$$\int_a^b e^{-x^2}\, d\mu(x) \approx \sum_{n=0}^{m} \int_a^b (-1)^n \frac{x^{2n}}{n!}\, d\mu(x) \tag{6.6}$$

$$= \sum_{n=0}^{m} \left[(-1)^n \frac{x^{2n+1}}{(2n+1)n!}\right]_a^b. \tag{6.7}$$

Other ways for finding the asymptotic expansions for integrals are the *method of integration by parts*, as well as *Laplace* and *saddle-point approximations*.

While all these methods can be handled very well in the one-dimensional case, their complexity is greatly increased in the multi dimensional case. The additional complexity arises partly from the multi-indexing associated with the multivariable Taylor expansions. In fact some of the expressions become so complicated that calculating these is already a computational problem itself, [57, Evans & Swartz 2000].

6.1.1.2 MULTIPLE QUADRATURE RULES

With the advances in computing power, methods of approximate integration based on *quadrature* became more and more interesting not only for the evaluation of one but also for multi-dimensional integrals.

The historically oldest techniques for solving integrals of type

$$\int_{\mathbf{B}} f(\mathbf{x})\, d\mu^s(\mathbf{x}), \tag{6.8}$$

where $\mathbf{B}$ is any bounded region in $\mathbb{R}^s$, are based on *multiple quadrature rules*. These techniques have their origin in the ancient Greek problem of *squaring the circle*, i.e. the process of inscribing or circumscribing a circle with rectangles or convex polygons of known area for estimating the transcendent number π.

A *multiple quadrature rule of order* n for approximating the integral from (6.8) is a sum of the form

$$\mathcal{I}_n = \sum_{i=1}^{n} w_i f(\mathbf{x}_i), \tag{6.9}$$

whose terms are products composed of so-called *weights*, w_i, and the values of the function f at chosen points $\mathbf{x}_i \in \mathbf{B}$. If such a quadrature rule has been developed with respect to the region $\mathbf{B}$, we can apply the same rule to an integral of the form

$$\int_{\mathbf{B}'} f(T^{-1}(\mathbf{y})) \left| \det \frac{dT^{-1}(\mathbf{y})}{d\mu^s(\mathbf{y})} \right| d\mu^s(\mathbf{y}), \tag{6.10}$$

Bijective Mapping (840) where $T : \mathbf{B} \to \mathbf{B}'$ is a bijective mapping.

Now, since quadrature rules are approximate methods for evaluating the exact value of an integral, they are only useful if the error between the exact value $\mathcal{I}$ and the approximate value of the integral $\mathcal{I}_n$ is sufficiently small for $n \to \infty$, that is, a quadrature rule must satisfies the requirement:

$$|\mathcal{I} - \mathcal{I}_n| \to 0. \tag{6.11}$$

In numerical analysis it is shown that one-dimensional interpolatory rules—as we will present them in the following—have error bounds of order:

$$|\mathcal{I} - \mathcal{I}_n| \leq \sup_{x \in [a,b]} |f^{(n+1)}(x)| \, \frac{\int_a^b \prod_{j=0}^{n} (x - x_j)}{(n+1)!} \, dx, \tag{6.12}$$

where the integration domain of the function f is given by the closed interval $[a, b]$, see [179, Schmeißer & Schirmeier 1976]. In practice this means that increasing the sample size generally improves the approximation although this is not guaranteed. Obviously, Formula

(6.12) implies that in the one-dimensional case quadrature rules have convergence rates of $O(n^{-r})$, provided that the integrand has $r+1$ many continuous derivatives.

Continuous Function (868)

Let us now consider the basic quadrature rules for approximating definite integrals, the so-called *interpolatory rules*, and here particularly, the *Newton-Cotes type* formulae as well as the *Gauss quadrature rules*. For a more detailed discussion on quadrature rules, such as *monomial rules, Bayesian quadrature, lattice rules,* and the class of *adaptive quadrature rules*, we refer to [57, Evans & Swartz 2000].

INTERPOLATORY RULES, THE CASE $s = 1$. A quadrature formula is called an *interpolatory rule* if it is based on integration an interpolatory formula, such as resulting from the Lagrange interpolatory problem with polynomials, see [179, Schmeißer & Schirmeier 1976].

Now, a solution approach for approximating a one-dimensional integral, based on an interpolatory rule, consists in interpolating the function followed by integration the resulting polynomial. As the *interpolation polynomial*, $p \in \mathcal{P}_n$, *in Lagrange form* is given by:

$\mathcal{P}_n$ (58)

$$p(x) \stackrel{def}{=} \sum_{i=0}^{n} q_i(x) f(x_i), \tag{6.13}$$

where apart from

$$q_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \qquad i = 0, 1, \ldots, n \tag{6.14}$$

it also holds

$$p(x_i) = f(x_i), \qquad i = 0, 1, \ldots, n \tag{6.15}$$

then an associated one-dimensional interpolatory rule, the so-called *Newton-Cotes Quadrature Formula* can be defined as follows:

DEFINITION 6.1 (Newton-Cotes Quadrature Formula) *Let $p \in \mathcal{P}_n$ be the Lagrange polynomial for interpolating a one-dimensional, real-valued function f. Then, the quadrature formula*

$$Q_n^{(r)}(f) \stackrel{def}{=} \int_a^b f(x)\, d\mu(x) = \int_a^b p(x)\, d\mu(x) = \sum_{i=0}^{n} f(x_i) \underbrace{\int_a^b q_i(x)\, d\mu(x)}_{w_i} \tag{6.16}$$

is called the Newton-Cotes quadrature formula, *also briefly denoted as the* Newton-Cotes formula, *if the function f is evaluated at the equidistant points set $\{x_0, \ldots, x_n\}$ given by:*

$$x_i = a + i\frac{b-a}{n} \qquad i = 0, 1, \ldots, n. \tag{6.17}$$

Let us now present the two simplest types of Newton-Cotes formula, the *Trapezoid Rule* and Simpson's Rule.

EXAMPLE 6.1 (Trapezoid and Simpson's Rule) *Let us consider the integral*

$$\int_I f(x)\,d\mu(x), \tag{6.18}$$

which we want—due to the linearity property of the Lebesgue integral—decompose into a sum of integrals over equal-sized integration domains $I = [a,b] = \bigcup_{i=1}^{N} [x_{i-1}, x_i]$, *resulting in:*

$$\sum_{i=1}^{N} \int_{[x_{i-1},x_i]} f(x)\,d\mu(x). \tag{6.19}$$

We will now approximate each of these integrals via the Newton-Cotes formula from Definition 6.1. The choice of the parameter $r = 1$ *then leads to the so-called* trapezoid rule, *that is, the integral over* $I_i = [x_{i-1}, x_i]$ *is estimated via the area of a trapezoid build from points* $(x_{i-1}, 0), (x_i, 0), (x_i, f(x_i)$ *and* $(x_{i-1}, f(x_{i-1}))$. *With* $\mu(I_i) = (x_i - x_{i-1}) = \frac{\mu(I)}{n}$, *the trapezoid rule then approximates the integral (6.18) by:*

$$\begin{aligned} Q_n^{(1)}(f) &= \sum_{i=1}^{n} \mu(I_i) \frac{f(x_{i-1}) + f(x_i))}{2} && (6.20)\\ &= \frac{\mu(I)}{2n} \sum_{i=1}^{n} f(x_{i-1}) + f(x_i)), && (6.21) \end{aligned}$$

see the left image in Figure 6.1.

The choice $r = 2$ *results in* Simpson's rule, *thus a weighted average of the function at the endpoints as well as the midpoint of the interval* I_i. *Then, an approximation of the integral from (6.18) via Simpson's rule is given by:*

$$\begin{aligned} Q_n^{(2)}(f) &= \sum_{i=1}^{n} \frac{\mu(I_i)}{6} \left(f(x_{i-1}) + 4f\left(\frac{x_{i-1} + x_i}{2}\right) + f(x_i) \right) && (6.22)\\ &= \frac{\mu(I)}{6n} \sum_{i=1}^{n} \left(f(x_{i-1}) + 4f\left(\frac{x_{i-1} + x_i}{2}\right) + f(x_i) \right), && (6.23) \end{aligned}$$

see the image on the right-hand side of Figure 6.1.

Finally, let us shortly talk about the error of the trapezoid and Simpson's rule.

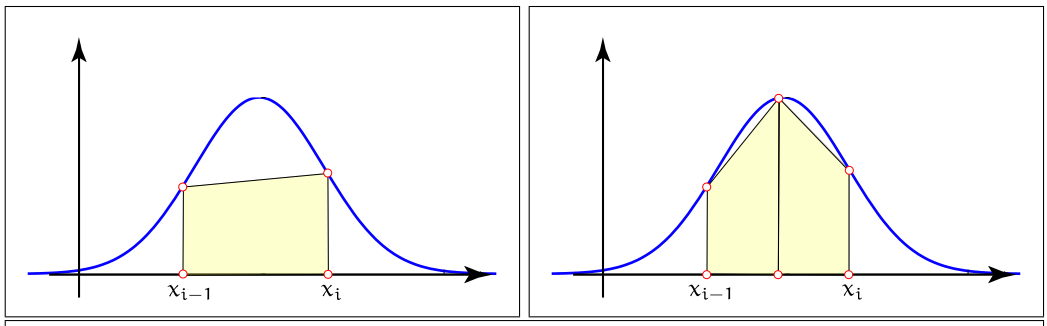


FIGURE 6.1: ONE-DIMENSIONAL NEWTON-COTES FORMULAS. The image on the left visualizes the trapezoid rule where the area under the graph of the function bounded by the endpoints of intervals given by $[x_{i-1}, x_i], 1 \leq i \leq N$ is approximated via the area of a trapezoid. On the right-hand side, Simpson's rule, which is similar to the trapezoid rule, except that we compute the area under a quadratic polynomial approximation instead of a linear approximation as it is done by the trapezoid rule.

The error estimate from Relation (6.12) implies:

$$|Q_n^{(1)} f - \mathcal{I}| \overset{(6.12)}{=} \sum_{i=1}^{n} \frac{\int_{I_i} (x - x_{i-1})(x - x_i)\, dx}{2!} \sup_{x \in I_i} f''(x) \tag{6.24}$$

$$= \sum_{i=1}^{n} \frac{x_i^3 + 3x_i^2 x_{i-1} - 3x_i x_{i-1}^2 + x_{i-1}^3}{2! \cdot 6} \sup_{x \in I_i} f''(x) \tag{6.25}$$

$$= \sum_{i=1}^{n} \frac{\mu(I_i)^3}{12} \sup_{x \in I_i} f''(x) \tag{6.26}$$

$$\overset{\mu(I_i) = \frac{\mu(I)}{n}}{\leq} \frac{\mu(I)^3}{12\, n^2} \sup_{x \in I} f''(x) \tag{6.27}$$

where $\mathcal{I}$ denotes the exact value of the integral from (6.18), and $\mu(I_i) = \mu(I_j), 1 \leq j \leq n$ was assumed. Thus, the error for the trapezoid rule can be estimated as of order $O\left(n^{-2}\right)$, provided that f has at least two continuous derivatives on the integration domain.

A similar error estimation can be derived for Simpson's rule. We leave the proof to the interested reader as an exercise. The error for Simpson's rule is of order $O\left(n^{-4}\right)$, since it can be bounded by the fourth derivative

$$|Q^{(2)} f - \mathcal{I}| = \frac{\mu(I)^5}{2880\, n^4} \sup_{x \in I} f^{iv}(x), \tag{6.28}$$

for more details see [179, Schmeißer & Schirmeier 1976]. Note: Since it is an interpolatory rule based on three points, Simpson's rule is exact when integrating polynomials of degree three or less.

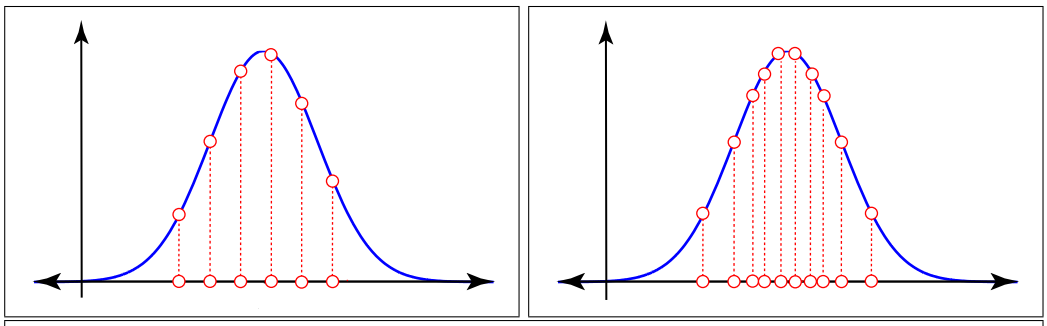


FIGURE 6.2: NEWTON-COTES FORMULAS VS GAUSS RULES. Visualization of a Newton-Cotes and a Gauss quadrature formula for approximating an integral of a real-valued function. As detailed discussed below, Newton-Cotes formulas are based on equidistant partitions of the integration domain, while Gauss rules do not subject to restrictions with respect of the choice of the samples.

As we have seen, the Newton-Cotes type rules approximate an integral by summing up its function values at a set of equidistant chosen points multiplied by appropriately chosen weights. Another popular family of interpolatory integration rules are the *Gauss Qudratures*. Compared with the Newton-Cotes formulae, Gauss quadratures allow the free choice of the supporting points $\{x_0, \ldots, x_n\}$ at which the integrand has to be evaluated, see Figure 6.2. The Gauss quadrature rules depend heavily on the properties of orthogonal polynomials. Since deterministic numerical integration is not the main purpose of our book, we refer the interested reader for a more detailed discussion on this topic to [57, Evans & Swartz 2000].

INTERPOLATORY RULES, THE CASE $s > 1$. Once interpolatory rules had been constructed for the treatment of one-dimensional problems, the next natural step was also to solve multidimensional integrals such as

$$\int_{\mathbf{B}} f(\mathbf{x})\, d\mu^s(\mathbf{x}) = \int_B \ldots \int_B f(x_1, \ldots, x_s) d\mu(x_1) \ldots d\mu(x_s) \tag{6.29}$$

over the integration domain $\mathbf{B} = \underbrace{B \times \cdots \times B}_{s \times times} \subset \mathbb{R}^s$.

Fubini-Tonelli Theorem (115)

For developing multi-dimensional integration rules, the integral from (6.29) is—based on the well-known Theorem of Fubini-Tonelli—considered as an iteration of one-dimensional integrals, where a one-dimensional integration rule is applied in each iteration. Thus, Relation (6.29) can be approximated by:

$$\sum_{i_s=1}^{n} \ldots \sum_{i_1=1}^{n} w_{i_1} \ldots w_{i_s} f(x_{i_1}, \ldots, x_{i_s}), \tag{6.30}$$

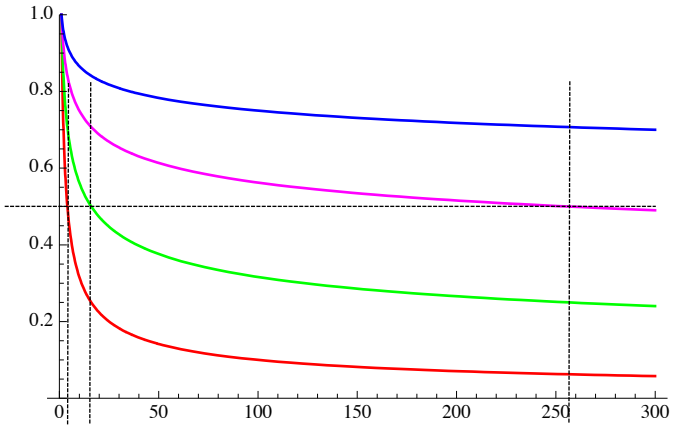


FIGURE 6.3: CONVERGENCE RATES FOR INTERPOLATORY INTEGRATION RULES. Visualization of different convergence rates of an interpolatory integration rule with $r = 2$, and $s = 4, 8, 16$ and $s = 32$. To halve the error, the sample size has to be multiplied by $4, 16, 32$ in the $8, 16$ and 32-dimensional case.

with the weights $w_{i_j}, 1 \leq j \leq s, 1 \leq i_j \leq n$ and the point set $\{x_{i_1}, \ldots, x_{i_s}\}$ constructed over the Cartesian product of the one-dimensional point sets.

The convergence speed of multi-dimensional interpolatory rules is thus predetermined via that of the one-dimensional rules, while the number of samples increases with the dimension s. Therefore, in the s-dimensional case with $N = n^s$ involved samples, the convergence of this procedure is only of order $O\left(N^{\frac{-r}{s}}\right)$. This entails, for large $s \geq 5$, highly complex and time-consuming procedures which are inappropriate for evaluating definite integrals, see Figure 6.3.

Let us finished this little section with an important result which limits the convergence rate of any deterministic quadrature rule: *Bakhvalov's Theorem* [57, Evans & Swartz 2000]. Intuitively Bakhvalov's Theorem says that for any s-dimensional quadrature rule, there exists a function f with r continuous and bounded derivatives for which the error is proportional to $N^{\frac{-r}{s}}$. Due to [220, Veach 1997], Bakhvalov's Theorem implies:

$$|\mathcal{I}_n - \mathcal{I}| > k \cdot N^{-\frac{r}{s}}, \tag{6.31}$$

with a constant $k > 0$ depending on r. Thus even if f has a bounded, continuous first derivative, no quadrature rule has an error bound better than $O(N^{-\frac{1}{s}})$.

6.1.2 THE CURSE OF DIMENSIONALITY

Let us consider the problem of evaluating the volume of the s-dimensional unit sphere S^{s-1}, which is well known to equal

$$\begin{aligned} \nu\mathrm{ol}(S^{s-1}) &= \int_{S^{s-1}} \sin\phi_2 \sin\phi_3^2 \cdots \sin\phi_{s-2}^{s-1} \, d\mu(\phi_{s-1}) \cdots d\mu(\phi_1) && (6.32) \\ &= \frac{\pi^{\frac{s}{2}}}{\Gamma\left(\frac{s}{2}+1\right)}, && (6.33) \end{aligned}$$

where Γ is Euler's gamma function, see [174, Rudin 1998].

As the s-dimensional unit sphere is contained in the s-dimensional cube $I^s = [-1,1]^s$, the integral from (6.33) can be written via a variable transformation as:

χ (839)

$$\nu\mathrm{ol}(S^{s-1}) = \int_{[-1,1]^s} \chi_{S^{s-1}}(\mathbf{x}) d\mu^s(\mathbf{x}). \tag{6.34}$$

A naive method for approximating this integral could be to construct a sequence of points that are uniformly distributed over the s-dimensional cube I^s and averaging the integrand over these points. Due to [57, Evans & Swartz 2000], this can be done by subdividing each axis into n subintervals of equal length and then computing the approximation:

$$\nu\mathrm{ol}(S^{s-1}) \approx \frac{1}{n^s} \sum_{i_1=1}^{n} \cdots \sum_{i_s=1}^{n} I_{S^s}\left(-1+2\frac{i_1-1}{n-1}, \ldots, -1+2\frac{i_k-1}{n-1}\right), \tag{6.35}$$

see Figure 6.4.

Obviously, this approximation leads to an exponential increase in the number of function evaluations, namely n^s. Choosing $n = 10$, then computing the volume of the 10-dimensional unit sphere requires the evaluation of $10,000,000$ function values, which results in enormous costs. Regardless of the accuracy of the approximation, this naive integration method is due to the cost incurred clearly infeasible even for small dimensions of the integration problem.

Additionally, in [57, Evans & Swartz 2000] it is shown, that with increasing dimension the ratio of the volume of the s-dimensional unit sphere to the s-dimensional unit cube is given by

$$\lim_{s\to\infty} \frac{\pi^{\frac{s}{2}}}{\Gamma\left(\frac{s}{2}+1\right) \cdot 2^s} = 0, \tag{6.36}$$

this means that the volume of S^{s-1} is vanishing small compared to the volume of I^s as the dimension rises. As for $s = 2$ this proportion is still $7.85 \cdot 10^{-1}$, it holds for $s = 5$ that the 5-dimensional unit sphere is occupying less than 20 % of the volume of I^5, and for $s = 20$ the ratio between the unit sphere and the unit cube is less than $2.46 \cdot 10^{-7}$.

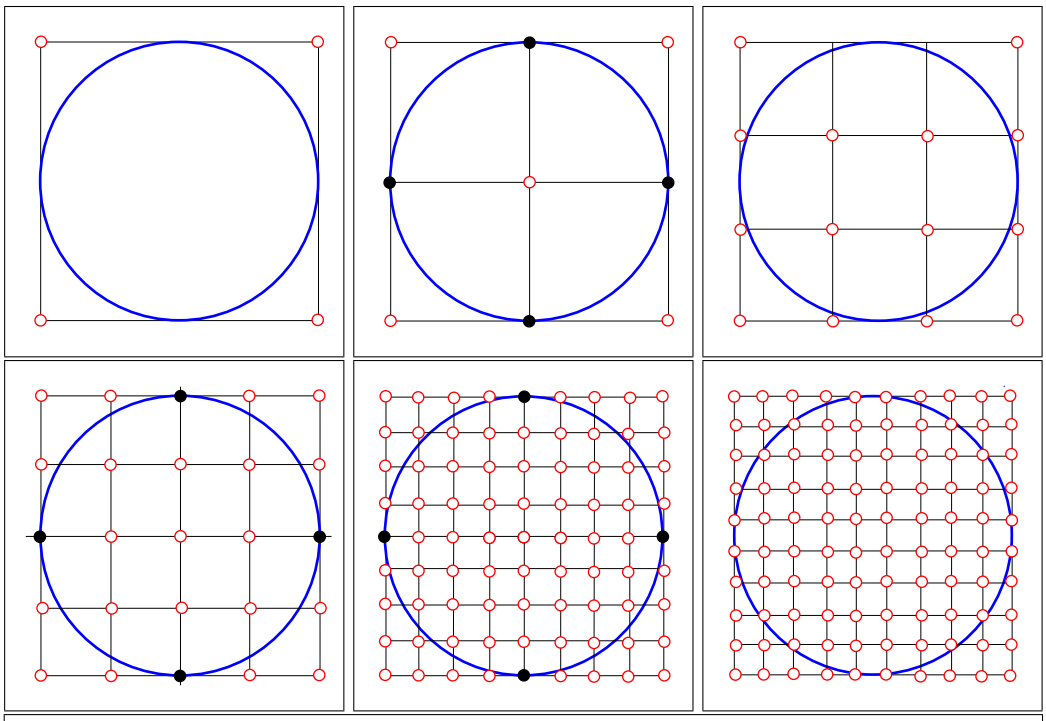


FIGURE 6.4: APPROXIMATING THE VOLUME OF THE 2-DIMENSIONAL UNIT SPHERE.
Six different sampling patterns resulting from subdividing each axis into $n = 1, 2, 3, 4, 8, 9$ subintervals of each length chosen according to Formula for approximating the volume of the 2-dimensional unit sphere.

We conclude from this observation that not only the number of points increase enormously but also the fact that the integrand $I_{S^{s-1}}$ for increasing s does not contribute significantly to the approximation of the integral, since $I_{S^{s-1}} = 0$ for most of these points.

This phenomenon is called *dimensional effect*, or the *curse of dimesionality*. The curse of dimensionality is central to the integration problem, and it is the reason that the efficiency of almost all integration methods decrease as the dimension arises. Note: The curse of dimensionality can not only be interpreted as the exponentially growth in the number of operations, but it must also be viewed in connection with the fact that properties of an integrand which are unproblematic in low dimensions can cause enormous implications in high-dimensional integration problems.

A promising integration technique, which can help to overcome the dimensional effect of integration problems, is Monte Carlo integration, as we will present them in the following section. Monte Carlo integration is a class of integration techniques based on probabilistic approaches which are easily to be constructed and which are independent on the dimension

of the integration problem. So, they have become the most favored methods for solving high-dimensional integrals.

6.2 THE INTEGRAL AS EXPECTED VALUE OF A CONTINUOUS RANDOM VARIABLE

The principle of *Monte Carlo integration* is based on the representation of an integral as expected value of a continuous random variable which must be estimated with the help of stochastic approximation methods.

To explain the theory behind Monte Carlo integration, assume the integral

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \tag{6.37}$$

be given over the integration domain $\mathbf{Q}^s \subset \mathbb{R}^s$ with $0 < \mu^s(\mathbf{Q}^s) < \infty$, where f is a real-valued function from Lebesgue space $\mathcal{L}(\mathbb{R}^s, \mu^s)$, μ^s is the Lebesgue measure on $\mathbb{R}^s$, and $\mathbf{Q}^s$ corresponds to a bounded subset of $\mathbb{R}^s$.

$\mathcal{L}(\mathbb{R}^s, \mu^s)$ (107)
μ^s (82)

THE INTEGRAL AS EXPECTED VALUE OF A CONTINUOUS UNIFORMLY DISTRIBUTED RANDOM VARIABLE. To represent the integral from (6.37) as expected value of a continuous random variable or a random vector, we have to construct a probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ where this random variable can be defined on.

Expected Value of a RV (197)
Probability Space (163)

For that, let us consider the measurable space $(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s))$ with basic set $\mathbf{Q}^s$ and the Borel σ-algebra $\mathfrak{B}(\mathbf{Q}^s)$ of measurable subsets of $\mathbf{Q}^s$. We can easily extend this measurable space to a probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ by defining a probability measure $\mathbb{P}$ via the s-dimensional Lebesgue measure μ^s, namely by:

Measurable Space (80)
$\mathfrak{B}(\mathbf{Q}^s)$ (865)
Probability Measure (80)

$$\mathbb{P}(\mathbf{B}) \stackrel{def}{=} \frac{\mu^s(\mathbf{B})}{\mu^s(\mathbf{Q}^s)} = \frac{\int_{\mathbf{B}} d\mu^s(\mathbf{x})}{\int_{\mathbf{Q}^s} d\mu^s(\mathbf{x})}. \tag{6.38}$$

Because $\mathbb{P}$ is a measure with $\mathbb{P}(\mathbf{Q}^s) = 1$, the measure $\mathbb{P}$ is a probability measure and $\left(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s), \frac{\mu^s}{\mu^s(\mathbf{Q}^s)}\right)$ can be interpreted as our wished probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$.

Let us now consider the uniformly distributed random variable $\mathbf{U}$,

Uniform Distribution (180)

$$\mathbf{U} : (\Omega, \mathfrak{F}(\Omega)) \to (\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s)), \tag{6.39}$$

with

$$\omega \mapsto \mathbf{x} = \mathbf{U}(\omega) = \omega, \tag{6.40}$$

which is mapped via the $\mathbb{R}^s$-$\mathbb{R}$-measurable function $f : \mathbf{x} \mapsto f(\mathbf{x})$ from $\mathbb{R}^s$ to $\mathbb{R}$.

Measurable Function (98)

With $\mathbf{U}$, also the composition $f \circ \mathbf{U}$ is a random variable whose expected value is given by:

$$\begin{aligned} E(f \circ \mathbf{U}) &= \int_{\Omega} f(\mathbf{U}(\omega))\, d\mathbb{P}(\omega) && (6.41)\\ &= \int_{\Omega} f(\omega)\, d\mathbb{P}(\omega) = E(f). && (6.42) \end{aligned}$$

Obviously, reformulating Equation (6.37) leads to

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) = \mu^s(\mathbf{Q}^s) \int_{\mathbf{Q}^s} f(\mathbf{x}) \frac{1}{\mu^s(\mathbf{Q}^s)}\, d\mu^s(\mathbf{x}) \quad (6.43)$$

which, using the probability density function $p_{\mathbf{U}} : \mathbb{R}^s \to \mathbb{R}$, given by

Probability Density Function (176)

$$p_{\mathbf{U}}(\mathbf{x}) \stackrel{def}{=} \begin{cases} \frac{1}{\mu^s(\mathbf{Q}^s)} & \mathbf{x} \in \mathbf{Q}^s \\ 0 & \mathbb{R}^s \setminus \mathbf{Q}^s \end{cases} \quad (6.44)$$

can also be written as:

$$\begin{aligned} \int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) &= \mu^s(\mathbf{Q}^s) \int_{\mathbb{R}^s} f(\mathbf{x})\, p_{\mathbf{U}}(\mathbf{x})\, d\mu^s(\mathbf{x}) && (6.45)\\ &\stackrel{(2.735)}{=} \mu^s(\mathbf{Q}^s) \int_{\mathbb{R}^s} f(\mathbf{x})\, d\mathbb{P}_{\mathbf{U}}(\mathbf{x}) && (6.46)\\ &\stackrel{\mathbf{x}=\mathbf{U}(\omega)}{=} \mu^s(\mathbf{Q}^s) \int_{\Omega} f(\mathbf{U}(\omega))\, d\left(\mathbb{P} \circ \mathbf{U}^{-1}\right)(\omega) && (6.47)\\ &\stackrel{\mathbf{U}(\omega)=\omega}{=} \mu^s(\mathbf{Q}^s) \int_{\Omega} f(\omega)\, d\mathbb{P}(\omega) && (6.48)\\ &\stackrel{(6.42)}{=} \mu^s(\mathbf{Q}^s) E(f), && (6.49) \end{aligned}$$

that is, the integral from (6.37) can be interpreted as expected value of a uniformly distributed random variable given over $\Omega = \mathbf{Q}^s$ multiplied by the s-dimensional volume of the bounded set $\mathbf{Q}^s$.

Expected Value of a RV (197)

EXAMPLE 6.2 (The Integral as Expected Value of a Continuous Uniformly Distributed Random Variable) *As a first practical application for Monte Carlo integration let us represent the following one-dimensional integral*

$$\int_{[0,\frac{\pi}{2}]} \cos(x)\, d\mu(x) \quad (6.50)$$

as expected value of a uniformly distributed random variable defined on a probability space over the base set $\Omega = \left[0, \frac{\pi}{2}\right]$. *As underlying probability space of the uniform random variable* $U : \omega \mapsto U(\omega) = \omega$, *we choose* $\left(\left[0, \frac{\pi}{2}\right], \mathfrak{B}(\left[0, \frac{\pi}{2}\right]), \mathbb{P}\right)$ *with* $\mathbb{P}(B) =$

Uniform Distribution (180)

$\frac{\mu(B)}{\mu([0,\frac{\pi}{2}])}$ *for* $B \in \mathfrak{B}([0,\frac{\pi}{2}])$. *Based on this settings and the* $\mathbb{R}$-$\mathbb{R}$*-measurable function* f *given by* $x \mapsto f(x) = \cos(x)$ *then it holds:*

$$\begin{aligned}
\int_{[0,\frac{\pi}{2}]} \cos(x)\, d\mu(x) &= \frac{\pi}{2}\int_{\mathbb{R}} \cos(x)\frac{2}{\pi}\, d\mu(x) && (6.51)\\
&\overset{p_U(x)=\frac{2}{\pi}}{=} \frac{\pi}{2}\int_{\mathbb{R}} \cos(x)\, d\mathbb{P}_U(x) && (6.52)\\
&\overset{x=U(\omega)}{=} \frac{\pi}{2}\int_{\mathbb{R}} \cos(x)\, d\left(\mathbb{P}\circ U^{-1}\right)(x) && (6.53)\\
&\overset{x=U(\omega)=\omega}{=} \frac{\pi}{2}\int_{\Omega} \cos(\omega)\, d\mathbb{P}(\omega) && (6.54)\\
&\overset{(2.732)}{=} \frac{\pi}{2}E(\cos(U)), && (6.55)
\end{aligned}$$

Expected Value of a RV (197)

that is, the above integral can be interpreted as expected value of the random variable U *multiplied by the length of the closed interval set* $[0,\frac{\pi}{2}]$.

EXAMPLE 6.3 (Antialiasing Interpreted as Expected Value of a Continuous Random Variable) *The image plane of a real or virtual camera can be considered as a pixel array of dimension* $[0,s_x]\times[0,s_y]$ *where* $\mu^2(\square_j)$ *denotes the Lebesgue measure of the area associated with a pixel* $\square_j$. *Now, a well-know problem in computer graphics is antialiased sampling a pixel of the image plane. For this, the value of a pixel must be computed via evaluating the integral*

Lebesgue Measure, μ^2 (82)

$$\int_{\square_j} w(\mathbf{x})L(\mathbf{x})\, d\mu^2(\mathbf{x}), \tag{6.56}$$

where w *corresponds to a weighting function defined on* $\square_j$ *and* $L(\mathbf{x})$ *is the radiance incident at point* $\mathbf{x}\in\square_j$ *over the hemisphere* $\mathcal{H}^2_i(\mathbf{x})$.

Radiance (250)

On the probability space $\left(\square_j, \mathfrak{B}(\square_j), \frac{\mu^2}{\mu^2(\square_j)}\right)$, *where the area of the pixel* $\square_j$ *corresponds to the base set,* $\mathfrak{B}(\square_j)$ *is as usual the Borel* σ*-algebra of measurable sets of* $\square_j$, *and* $\frac{\mu^2}{\mu^2(\square_j)}$ *is a probability measure we then define a uniformly distributed random variable* $\mathbf{U}$,

Borel σ-algebra (865)

Measurable Set (80)

$$\begin{aligned}
\mathbf{U} : (\square_j, \mathfrak{B}(\square_j)) &\to (\mathbb{R}^2, \mathfrak{B}(\mathbb{R}^2)) && (6.57)\\
\omega &\mapsto \mathbf{x} = \mathbf{U}(\omega) = \omega, && (6.58)
\end{aligned}$$

distributed according to the PDF, $p_{\mathbf{U}} : \mathbb{R}^2 \to \mathbb{R}$ *with* $\mathbf{x}\mapsto p_{\mathbf{U}}(\mathbf{x}) = \frac{1}{\mu^2(\square_j)}$.

Using these settings, the integral from Relation (6.56) can then be written as

$$\begin{aligned}\int_{\square_j} w(\mathbf{x})L(\mathbf{x})\,d\mu^2(\mathbf{x}) &= \mu^2(\square_j)\int_{\square_j} w(\mathbf{x})L(\mathbf{x})\frac{1}{\mu^2(\square_j)}\,d\mu^2(\mathbf{x}) && (6.59)\\ &\overset{(2.735)}{=} \mu^2(\square_j)\int_{\square_j} w(\mathbf{x})L(\mathbf{x})\,d\mathbb{P}_{\mathbf{U}}(\mathbf{x}) && (6.60)\\ &\overset{(2.732)}{=} \mu^2(\square_j)\int_{\square_j} w(\mathbf{U}(\omega))L(\mathbf{U}(\omega))\,d\mathbb{P}(\omega) && (6.61)\\ &= \mu^2(\square_j)E\left(w(\mathbf{U})L(\mathbf{U})\right). && (6.62)\end{aligned}$$

A simple technique for reducing aliasing effects is box filtering, *where radiance values, incident at a pixel, are averaged or weighted averaged. Mathematically, box filtering can be described by a weighting function w defined by:*

$$w(\mathbf{x}) \overset{def}{=} \begin{cases} \frac{1}{\mu^2(\square_j)} & \text{if } \mathbf{x} \in \square_j \\ 0 & \text{otherwise} \end{cases}. \tag{6.63}$$

Using box filtering as antialiasing strategy reduces the evaluation of the integral from (6.56) to the computation of the expected value of the random variable $\mathbf{U}$*, namely:*

Expected Value of a RV (197)

$$\int_{\square_j} w(\mathbf{x})L(\mathbf{x})\,d\mu^2(\mathbf{x}) = E(L(\mathbf{U})). \tag{6.64}$$

THE INTEGRAL AS EXPECTED VALUE OF A CONTINUOUS q-DISTRIBUTED RANDOM VARIABLE. In the literature, the process of interpreting a definite integral as expected value of a uniformly random variable and approximating this expected value via a Monte Carlo estimator is also called basic Monte Carlo integration. Now, as we will see in Section 6.6.2 basic Monte Carlo strategies are often due to reasons of efficiency not suitable to estimate the value of an integral. Here, a much better choice would be to choose a random variable whose realizations are distributed according to a well chosen probability distribution. To show that the integral from (6.37) can also be represented as expected value of a q-distributed random variable let us pursue the following way:

Uniform Distribution (180)

Let us consider the probability space $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$, where we choose Ω as $\mathbf{Q}^s$, $\mathfrak{F}(\Omega) = \mathfrak{B}(\mathbf{Q}^s)$ and $\mathbb{P} = \frac{\mu^s}{\mu^s(\mathbf{Q}^s)}$. We now define on the σ-algebra $\mathfrak{B}(\mathbf{Q}^s)$ by means of a measurable, non-negative function q a new measure λ via:

Measurable Function (98)
Borel σ-algebra (865)

$$\lambda(\mathbf{B}) \overset{def}{=} \int_{\mathbf{B}} q(\mathbf{x})\,d\mu^s(\mathbf{x}) \tag{6.65}$$

with $\mathbf{B} \in \mathfrak{B}(\mathbf{Q}^s)$. After that, we construct in accordance to the Radon-Nikodým Theorem a measure $\mathbb{P}_{\mathbf{X}}$ by

Theorem of Radon-Nikodým (176)

$$\mathbb{P}_{\mathbf{X}}(\mathbf{B}) \overset{def}{=} \frac{\lambda(\mathbf{B})}{\lambda(\mathbf{Q}^s)} = \frac{\int_{\mathbf{B}} q(\mathbf{x})\,d\mu^s(\mathbf{x})}{\lambda(\mathbf{Q}^s)}. \tag{6.66}$$

Because $\mathbb{P}_{\mathbf{X}}$ is a measure with $\mathbb{P}_{\mathbf{X}}(\mathbf{Q}^s) = 1$, the measure $\mathbb{P}_{\mathbf{X}}$ is a probability measure and $\left(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s), \mathbb{P}_{\mathbf{X}}\right)$ becomes to a probability space.

Based on the Lebesgue measure μ^s and the above constructed probability measure $\mathbb{P}_{\mathbf{X}}$, the function q can be written as the Radon-Nikodým derivative, thus,

Radon Nykodým Derivative (189)

$$q = \lambda(\mathbf{Q}^s)\frac{d\mathbb{P}_{\mathbf{X}}}{d\mu^s}, \tag{6.67}$$

where the differential measure $d\mu^s$ can be expressed as:

$$d\mu^s = \lambda(\mathbf{Q}^s)\frac{d\mathbb{P}_{\mathbf{X}}}{q}. \tag{6.68}$$

This means that the original integral can be written as:

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \quad = \quad \lambda(\mathbf{Q}^s) \int_{\mathbf{Q}^s} \frac{f(\mathbf{x})}{q(\mathbf{x})}\, d\mathbb{P}_{\mathbf{X}}(\mathbf{x}). \tag{6.69}$$

Measurable Space (80)

In the second step, we now generate a random variable $\mathbf{X}$ on the measurable space $(\mathbb{R}^s, \mathfrak{B}(\mathbb{R}^s))$ distributed according to the PDF $p_{\mathbf{X}} : \mathbb{R}^s \to \mathbb{R}$ with

$$p_{\mathbf{X}}(\mathbf{x}) \stackrel{def}{=} \begin{cases} \frac{q(\mathbf{x})}{\lambda^s(\mathbf{Q}^s)} & \mathbf{x} \in \mathbf{Q}^s \\ 0 & \mathbb{R}^s \setminus \mathbf{Q}^s. \end{cases} \tag{6.70}$$

Measurable Function (98)

As we know, the set of measurable functions is a linear space, that is with $\mathbf{X}, f$ and q, the function $\frac{(f\circ\mathbf{X})(\omega)}{(q\circ\mathbf{X})(\omega)} = \frac{f(\mathbf{X}(\omega))}{q(\mathbf{X}(\omega))}$ is also measurable and the integral from (6.37) can be formulated as:

$$\begin{aligned}
\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \quad &= \quad \lambda(\mathbf{Q}^s) \int_{\mathbb{R}^s} \frac{f(\mathbf{x})}{q(\mathbf{x})}\, d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) && (6.71)\\
&\stackrel{\mathbf{x}=\mathbf{X}(\omega)}{=} \quad \lambda(\mathbf{Q}^s) \int_{\Omega} \frac{f(\mathbf{X}(\omega))}{q(\mathbf{X}(\omega))}\, d\left(\mathbb{P}\circ\mathbf{X}^{-1}\right)(\mathbf{X}(\omega)) && (6.72)\\
&= \quad \lambda(\mathbf{Q}^s) \int_{\Omega} \frac{f(\mathbf{X}(\omega))}{q(\mathbf{X}(\omega))}\, d\mathbb{P}(\omega) && (6.73)\\
&\stackrel{(2.732)}{=} \quad \lambda(\mathbf{Q}^s)\, E\left(\frac{f(\mathbf{X})}{q(\mathbf{X})}\right). && (6.74)
\end{aligned}$$

In this fashion, the original calculation of a high-dimensional integral is reduced to the evaluation of the expected value of the random variable $\frac{f\circ\mathbf{X}}{q\circ\mathbf{X}}$ with probability distribution $\mathbb{P}$ including the calculation of the volume $\lambda(\mathbf{Q}^s)$.

Measure (79)
Radon-Nikodým Derivative (176)
Probability Density Function (176)

REMARK 6.1 *Obviously the process of generating a probability measure by means of a measurable, non-negative function q, as described above, delivers according to Equation (2.585) a probability density function $p_{\mathbf{X}}$ defined by:*

$$p_{\mathbf{X}}(\mathbf{x}) \stackrel{def}{=} \frac{q(\mathbf{x})}{\lambda(\mathbf{Q}^s)}, \tag{6.75}$$

which satisfies the conditions required to a density function, that is,

i) $p_{\mathbf{X}}(\mathbf{x}) \geq 0,\ \forall \mathbf{x} \in \mathbb{R}^s$,

ii) $\int_{]-\infty,\infty[^s} p_{\mathbf{X}}(\mathbf{x})\, d\mu^s(\mathbf{x}) = \int_{\mathbf{Q}^s} \frac{q(\mathbf{x})}{\lambda(\mathbf{Q}^s)} d\mu^s(\mathbf{x}) \overset{(6.65)}{=} \frac{\int_{\mathbf{Q}^s} q(\mathbf{x}) d\mu^s(\mathbf{x})}{\int_{\mathbf{Q}^s} q(\mathbf{x})\, d\mu^s(\mathbf{x})} = 1.$

Using Relation (6.75) in Formula (6.74) then leads to:

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \quad = \quad E(f(\mathbf{X})), \tag{6.76}$$

now, the representation of the integral as expected value of a p-distributed random variable.

EXAMPLE 6.4 (The Integral as Expected Value of a Continuous Random Variable) *Let us show how Monte Carlo integration works for the following simple integral:*

$$\int_{[0,\frac{\pi}{2}]} \cos(x)\, d\mu(x). \tag{6.77}$$

As basic set of the desired probability space we choose the closed interval $\left[0, \frac{\pi}{2}\right]$ and as the associated σ-algebra, the Borel σ-algebra $\mathfrak{B}\left(\left[0, \frac{\pi}{2}\right]\right)$. For the construction of the image measure we can use any measurable, non-negative function. Let x be such a function then $\mathbb{P}_X$ is given via:

Borel σ-algebra (865)

$$\mathbb{P}_X(\mathbf{B}) \overset{def}{=} \frac{\lambda(\mathbf{B})}{\lambda\left(\left[0, \frac{\pi}{2}\right]\right)} = \frac{\int_{\mathbf{B}} x\, d\mu(x)}{\int_{[0,\frac{\pi}{2}]} x\, d\mu(x)} = \frac{\int_{\mathbf{B}} x\, d\mu(x)}{\frac{1}{2}x^2\Big|_0^{\frac{\pi}{2}}} = \frac{\int_{\mathbf{B}} x\, d\mu(x)}{\frac{1}{8}\pi^2} \tag{6.78}$$

or written as Radon-Nikodým derivative:

Radon-Nikodým Derivative (176)

$$x = \frac{\pi^2}{8}\frac{d\mathbb{P}_X}{d\mu} \quad \Leftrightarrow \quad d\mu = \frac{\pi^2}{8}\frac{d\mathbb{P}_X}{x}. \tag{6.79}$$

The integral from (6.77) can now be written as:

$$\begin{aligned} \int_{[0,\frac{\pi}{2}]} \cos(x)\, d\mu(x) &= \int_{[0,\frac{\pi}{2}]} \frac{\cos(x)}{\frac{8}{\pi^2}x} d\mathbb{P}_X(x) && (6.80)\\ &= \int_{[0,\frac{\pi}{2}]} \frac{\pi^2 \cos(X(\omega))}{8X(\omega)} d\mathbb{P}(\omega) && (6.81)\\ &= \frac{\pi^2}{8} E\left(\frac{\cos(X)}{X}\right), && (6.82) \end{aligned}$$

where the random variable $\frac{\cos(X)}{\frac{8X}{\pi^2}}$ is defined on $\left[0, \frac{\pi}{2}\right]$ and distributed with respect to the probability density function $p_X = \frac{8x}{\pi^2}$.

Another possible choice for q could be the constant function 1. According to Equation (6.75), then a probability density function p_X on $\left[0, \frac{\pi}{2}\right]$ is given by:

$$p_X(x) = \frac{2}{\pi} \tag{6.83}$$

and we deduce from Relations (6.67) and (6.68):

$$p_X = \frac{2}{\pi} = \frac{d\mathbb{P}_X}{d\mu} \quad or \quad d\mu = \frac{d\mathbb{P}_X}{p_X} = \frac{d\mathbb{P}_X}{\frac{2}{\pi}}. \tag{6.84}$$

Replacing the integrand from (6.77) by these relations, then the integral can be written as the expected value of the random variable X which is distributed according to the constant density $p_X = \frac{2}{\pi}$ on $\left[0, \frac{\pi}{2}\right]$, i.e.:

$$\begin{aligned} \int_{[0,\frac{\pi}{2}]} \cos(x)\, d\mu(x) &= \int_{[0,\frac{\pi}{2}]} \frac{\cos(x)}{\frac{2}{\pi}}\, d\mathbb{P}_X(x) && (6.85)\\ &= \int_{[0,\frac{\pi}{2}]} \frac{\pi \cos(X(\omega))}{2}\, d\mathbb{P}(\omega) && (6.86)\\ &= \frac{\pi}{2} E(\cos(X)), && (6.87) \end{aligned}$$

thus, the same formula as in Example 6.6.

EXAMPLE 6.5 (Representing the Measurement Equation as Expected Value of a Continuous Random Variable) *Let now us illustrate the idea behind Monte Carlo integration using the example of the measurement equation given by*

Measurement Equation (416)

$$M_j \stackrel{def}{=} \int_{\square_j \times S^2} W_e^j(\mathbf{r}) L_i(\mathbf{r})\, d\zeta^\perp(\mathbf{r}). \tag{6.88}$$

As base set of the underlying probability space we choose the integration domain $\square_j \times S^2$. Then, an associated probability measure $\mathbb{P}$ over the Borel σ-algebra of $\square_j \times S^2$ can be defined via the throughput measure $\zeta^\perp$ by:

Borel σ-algebra (865)
Throughput Measure (94)

$$\mathbb{P}(\mathbf{B}) \stackrel{def}{=} \frac{\zeta^\perp(\mathbf{B})}{\zeta^\perp(\square_j \times S^2)}. \tag{6.89}$$

Then, an associated probability space for the representing the measurement equation as expected value of a random variable is given by $\left(\square_j \times S^2, \mathfrak{B}(\square_j \times S^2), \mathbb{P}\right)$, and for a uniformly distributed random variable it holds:

$$\begin{aligned} M_j &\stackrel{def}{=} \int_{\square_j \times S^2} W_e^j(\mathbf{r}) L_i(\mathbf{r})\, d\zeta^\perp(\mathbf{r}) && (6.90)\\ &= \zeta^\perp(\square_j \times S^2)\, E\left(W_e^j(\mathbf{U}) L_i(\mathbf{U})\right). && (6.91) \end{aligned}$$

Unfortunately, for representing the integral from (6.37), Formula (6.74) is not always useful in practice as the calculation of the volume $\lambda(\mathbf{Q}^s)$ is often more difficult than the calculation of the entire integral. In order to avoid this problem the integral from (6.37) can also be transformed into one of the form:

$$\int_{\mathbf{I}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \tag{6.92}$$

over the s-dimensional unit cube $\mathbf{I}^s = [0,1]^s$ using change of variables and if necessary shifting and scaling operations. If we then choose a uniformly distributed random variable $\mathbf{U}$ on $\mathbf{I}^s$, distributed according to the probability density function $p_{\mathbf{U}} = 1$, we obtain with $\lambda(\mathbf{I}^s) = 1$ and the differential measure

Uniform Distribution (180)

$$d\mu^s = \lambda(\mathbf{I}^s)\frac{d\mathbb{P}_{\mathbf{U}}}{p_{\mathbf{U}}} = d\mathbb{P}_{\mathbf{U}} \tag{6.93}$$

for the above integral:

$$\begin{aligned}\int_{\mathbf{I}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) &= \int_{\mathbf{I}^s} f(\mathbf{x})\, d\mathbb{P}_{\mathbf{U}}(\mathbf{x}) && (6.94)\\ &\overset{\mathbf{x}=\mathbf{U}(\omega)=\omega}{=} \int_{\Omega} f(\omega)\, d\mathbb{P}(\omega) && (6.95)\\ &= E(f). && (6.96)\end{aligned}$$

Hence we can conclude that the application of this procedure with density $p_{\mathbf{U}} = 1$ clearly leads to the interpretation of the integral as the expected value of the random variable f defined on the probability space $(\mathbf{I}^s, \mathfrak{B}(\mathbf{I}^s), \mu^s)$.

EXAMPLE 6.6 (The Integral as Expected Value of a Uniformly Distributed Random Variable on the Canonical Probability Space) *Obviously, trivial integration by substitution leads to:*

$$\int_{[0,\frac{\pi}{2}]} \cos(x)\, d\mu(x) = \frac{\pi}{2}\int_{[0,1]} \cos\left(\frac{\pi}{2}x\right)\, d\mu(x). \tag{6.97}$$

Due to Relation (6.96), then it holds:

$$\begin{aligned}\frac{\pi}{2}\int_{[0,1]} \cos\left(\frac{\pi}{2}x\right)\, d\mu(x) &= \frac{\pi}{2}\int_{\mathbb{R}} \cos\left(\frac{\pi}{2}x\right)\chi_{[0,1]}(x)\, d\mathbb{P}_U(x) && (6.98)\\ &\overset{(2.732)}{=} \frac{\pi}{2}\int_{[0,1]} \cos\left(\frac{\pi}{2}\omega\right)\, d\mathbb{P}(\omega), && (6.99)\end{aligned}$$

where the random variable U is defined on:

$$U : \left(\left[0,\frac{\pi}{2}\right], \mathfrak{B}\left(\left[0,\frac{\pi}{2}\right]\right)\right) \to (\mathbb{R}, \mathfrak{B}(\mathbb{R})) \tag{6.100}$$

with $\omega \mapsto x = U(\omega) = \omega$ and f is a $\mathbb{R}$-$\mathbb{R}$-measurable function with $\omega \mapsto f(\omega) = \cos(\omega)$ and the density p_U corresponds to the characteristic function on the closed interval $[0,1]$.

REMARK 6.2 (Blind and Informed Monte Carlo Techniques) *As already mentioned above, in the literature, the above technique representing an* integral as expected value of a continuous random variable *based on uniform sampling is referred to as* basic Monte

Carlo integration. *Because the sample points used in a basic Monte Carlo integration scheme are generated according to a uniform PDF on the integration domain without looking at the function itself, basic Monte Carlo integration is also often called a* blind Monte Carlo technique. *Compared to blind Monte Carlo, an* informed Monte Carlo technique *makes use of some kind of information available about the function or its integration domain. Intuitively, one expects more exact results from informed Monte Carlo than from blind Monte Carlo techniques.*

Section 6.6

REMARK 6.3 *Normally, the principle of Monte Carlo integration is defined over the canonical probability space* $(\mathbf{I}^s, \mathfrak{B}(\mathbf{I}^s), \mu^s)$ *often further restricted to the case* $s = 1$. *The advantage of this is that we can obtain more simpler formulas than those developed above. Nevertheless, we have decided to take into account in our presentation of Monte Carlo integration general probability spaces based on the* σ*-algebra* $\mathfrak{B}([\mathbf{a},\mathbf{b}])$, $[\mathbf{a},\mathbf{b}] \in \mathbb{R}^s, s \geq 1$. *This is advantageous in particular when discussing variance reduction methods where probability spaces are required which are of more general nature than the canonical one. We also want to mention that the volume* $\lambda(\mathbf{Q}^s)$ *may be neglected in all existing formulas where it is possible to construct the underlying probability measure* $\mathbb{P}_{\mathbf{X}}$ *with the help of a probability density function* p. *Under these circumstances it becomes possible to express the integral from (6.37) as the expected value of a continuous random variable* $\mathbf{X}$ *that is distributed on* $\mathbf{Q}^s$ *according to the PDF* $p_{\mathbf{X}}$ *as it holds:*

$(\mathbf{I}^s, \mathfrak{B}(\mathbf{I}^s), \mu^s)$ (166)

$\mathfrak{B}([\mathbf{a},\mathbf{b}])$ (865)

Section 6.6

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \overset{d\mu^s = \frac{d\mathbb{P}_{\mathbf{X}}}{p_{\mathbf{X}}}}{=} \int_{\mathbb{R}^s} \frac{f(\mathbf{x}))}{p_{\mathbf{X}}(\mathbf{x}))} \chi_{\mathbf{Q}^s}(\mathbf{x})\, d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) \tag{6.101}$$

$$= \int_{\Omega} \frac{f(\mathbf{X}(\omega))}{p_{\mathbf{X}}(\mathbf{X}(\omega))}\, d\mathbb{P}(\omega) \tag{6.102}$$

$$\overset{(2.732)}{=} E\left(\frac{f(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})}\right). \tag{6.103}$$

This form is familiar from the literature. It can be used much easier in formulas rather than the expressions that were derived in the main part of our above discussion.

6.3 MONTE CARLO ESTIMATORS

Expected Value of a RV (197)

Probability Space (163)

As we have seen in the preceding section, an integral may be interpreted as the stochastic expected value of a continuous random variable given over a probability space. Now, we are interested in the numerical computation of this expected value leading to the highly important concept of the *Monte Carlo estimator*, a very efficient method for approximate solving multi-dimensional integrals.

Random Variable (168) DEFINITION 6.2 (Monte Carlo Estimator) *A* Monte Carlo estimator F_N *is defined as a function of* N *random variables or random vectors* $\mathbf{X}_1, \ldots, \mathbf{X}_N, N \in \mathbb{N}$ *of the form* Random Vectors (183)

$$F_N \stackrel{\text{def}}{=} F_N(\mathbf{X}_1, \ldots, \mathbf{X}_N), \tag{6.104}$$

where the random variables $\mathbf{X}_i$ *are distributed according to some convenient probability density function* p. *Commonly, these random variables are* independent and identically distributed, *also denoted by* i.i.d. *for short, but in general they can depend on each other and they can have different probability distributions.* PDF (176) Independent RV (204) Probability Distribution (80)

The aim of a Monte Carlo estimator is the approximation of some quantity $\mathcal{I}$, which has to be calculated. Of particular interest to our discussion is the case where $\mathcal{I}$ is the evaluation of the integral

$$\int_{\mathbf{Q}^s} f(\mathbf{x}) \, d\mu^s(\mathbf{x}), \tag{6.105}$$

which can be seen as the expected value of a continuous random variable $\mathbf{X}$.

REMARK 6.4 (Primary and Secondary Monte Carlo Estimators) *In particular in [190, Sillion & Puech 1994], [116, Lafortune 1996], and [232, Watt 1999] it is distinguished between a* primary, *in the case of* $N = 1$, *and a* secondary Monte Carlo estimator, *if* $N \gg 1$.

As the variance of a primary estimator, as we will see in the following section, is usually unacceptably large, we will reduce the uncertainty by taking more, say N, *samples* $\mathbf{X}_i$ *and averaging their corresponding primary estimators* $F_1(\mathbf{X}_i)$ *into a secondary estimator, that is,* Variance of a RV (201)

$$F_N(\mathbf{X}_1, \ldots, \mathbf{X}_N) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} F_1(\mathbf{X}_i). \tag{6.106}$$

For our further considerations let $\mathbf{X}_1, \ldots, \mathbf{X}_N$ be independent and identically, according to a probability density function p, distributed random variables and $\left(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s), \mathbb{P}_{\mathbf{X}}\right)$ be a probability space, where $\mathbf{X}$ is used as a synonym for the identically p-distributed random variables $\mathbf{X}_1, \ldots, \mathbf{X}_N$. Because the random variables $\mathbf{X}_i, 1 \leq i \leq N$ are identical it holds for the expected value:

$$E\left(\frac{f(\mathbf{X}_1)}{p_{\mathbf{X}}(\mathbf{X}_1)}\right) = \ldots = E\left(\frac{f(\mathbf{X}_N)}{p_{\mathbf{X}}(\mathbf{X}_N)}\right) = E\left(\frac{f(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})}\right). \tag{6.107}$$

Under these conditions, the *Strong Law of Large Numbers*—which says that if a sufficiently large number of random variables are given, then the arithmetic mean will Theorem of SLLN (216)

converge to the expected value of these random variables almost surely—yields with respect to the random variables $\frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}, 1 \leq i \leq N$:

$$\text{prob}\left\{\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)} = E\left(\frac{f(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})}\right)\right\} = 1. \tag{6.108}$$

From this, we conclude: If we take enough samples, i.e. $N \gg 1$, it is guaranteed that the arithmetic mean of $\frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}$ converges to the correct result. As already formulated in the remark above, this means: If we define a Monte Carlo estimator as the arithmetic mean of the random variables $\frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}$, or with other words as the arithmetic mean of N primary estimators $F_1(\mathbf{X}_i) = \frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}$ by

$$F_N = \frac{1}{N}\sum_{i=1}^{N} \frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}, \tag{6.109}$$

see Figure 6.5, then the sequence of the Monte Carlo estimators $(F_N)_{N\in\mathbb{N}}$ converges almost surely towards the expected value $E\left(\frac{f(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})}\right)$ for $N \to \infty$.

If we now replace in Equation (6.108) the expected value by the corresponding integral, then we get:

$$\text{prob}\left\{\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)} = \int_{\Omega} \frac{f(\mathbf{X}(\omega))}{p_{\mathbf{X}}(\mathbf{X}(\omega))}\, d\mathbb{P}(\omega)\right\} = 1. \tag{6.110}$$

This fact results in that a secondary Monte Carlo estimator F_N for approximating the value of the integral

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \overset{(6.73)}{=} \lambda(\mathbf{Q}^s) \int_{\Omega} \frac{f(\mathbf{X}(\omega))}{q(\mathbf{X}(\omega))}\, d\mathbb{P}(\omega) \tag{6.111}$$

has the form:

$$\begin{aligned} F_N &= \frac{\lambda(\mathbf{Q}^s)}{N}\sum_{i=1}^{N} F_1(\mathbf{X}_i) && (6.112)\\ &= \frac{\lambda(\mathbf{Q}^s)}{N}\sum_{i=1}^{N} \frac{f(\mathbf{X}_i)}{q(\mathbf{X}_i)}, && (6.113)\\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}, && (6.114) \end{aligned}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are independent and identically p-distributed random variables and $p_{\mathbf{X}}(\mathbf{X}_i) = \frac{q(\mathbf{X}_i)}{\lambda(\mathbf{Q}^s)}$ is a PDF defined on the probability space $(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s), \mathbb{P}_{\mathbf{X}})$.

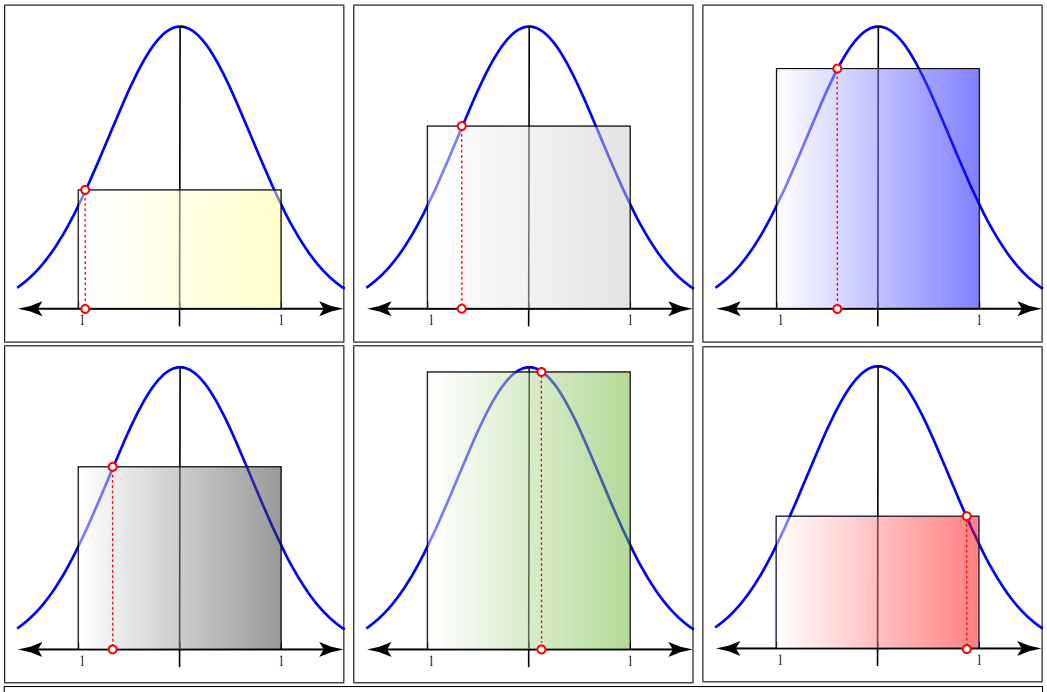


FIGURE 6.5: SECONDARY MONTE CARLO ESTIMATOR. The secondary Monte Carlo estimator $F_6 = \frac{1}{6}\sum_{i=1}^{6} \frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}$ from Relation (6.109) for evaluating the integral $\int_{[-1,1]^2} e^{-x^2}\, d\mu(x)$ using 6 independent and uniformly distributed random variables $\mathbf{X}_i$ drawn from integration domain $[-1,1]$. The value of the integral is then approximated by the average value of the 6 colored areas, constructed over $[-1,1]\times[0,f(\mathbf{X}_i)]$, for $1 \leq i \leq 6$.

EXAMPLE 6.7 (The Integral as Expected Value of a Continuous Random Variable, continued) *Let us consider once more the evaluation of the integral from (6.77). According to the Relations (6.82) and (6.87) then it holds:*

$$\int_{[0,\frac{\pi}{2}]} \cos(x)\, d\mu(x) = \frac{\pi}{8} E\left(\frac{\cos(X)}{X}\right) \tag{6.115}$$

respectively

$$\int_{[0,\frac{\pi}{2}]} \cos(x)\, d\mu(x) = \frac{\pi}{2} E(\cos(X)). \tag{6.116}$$

The corresponding Monte Carlo estimators for computing the value of the integral are given by

$$F_N = \frac{\pi^2}{8N} \sum_{i=1}^{N} \frac{\cos(X_i)}{X_i}, \tag{6.117}$$

with independent and according to the probability density function $p_X(x) = \frac{8x}{\pi^2}$ *identically distributed random variables* X_i *as well as*

$$F_N = \frac{\pi}{2N} \sum_{i=1}^{N} \cos(X_i), \tag{6.118}$$

where the independent random variables X_i *are distributed on* $\left[0, \frac{\pi}{2}\right]$ *according to* $p_X = \frac{2}{\pi}$. *Due to Equation (6.97), it should also be clear, that in the case of uniform sampling on* $[0, 1]$, *the corresponding estimator has the form*

$$F_N = \frac{\pi}{2N} \sum_{i=1}^{N} \cos\left(\frac{\pi}{2} X_i\right), \tag{6.119}$$

see Figure 6.6.

Theorem of SLLN (216) REMARK 6.5 *As a result of the Strong Law of Large Numbers, Monte Carlo integration will also converge if the variance of an estimator* F_N *is infinite, assumed that the expected value* $E(F_N)$ *exists.*

EXAMPLE 6.8 (Approximating the Form Factor Integral Using a Secondary Monte Carlo Estimator) *The approximation of an integral via a Monte Carlo estimator, as shown in the foregoing example for the one-dimensional case, can easily be extended to a multi-dimensional integral. For this, let us look at the two-dimensional integral for computing the differential-to-finite-area form factor between a differential surface patch and a surface patch* P_j *from Relation (2.207).*

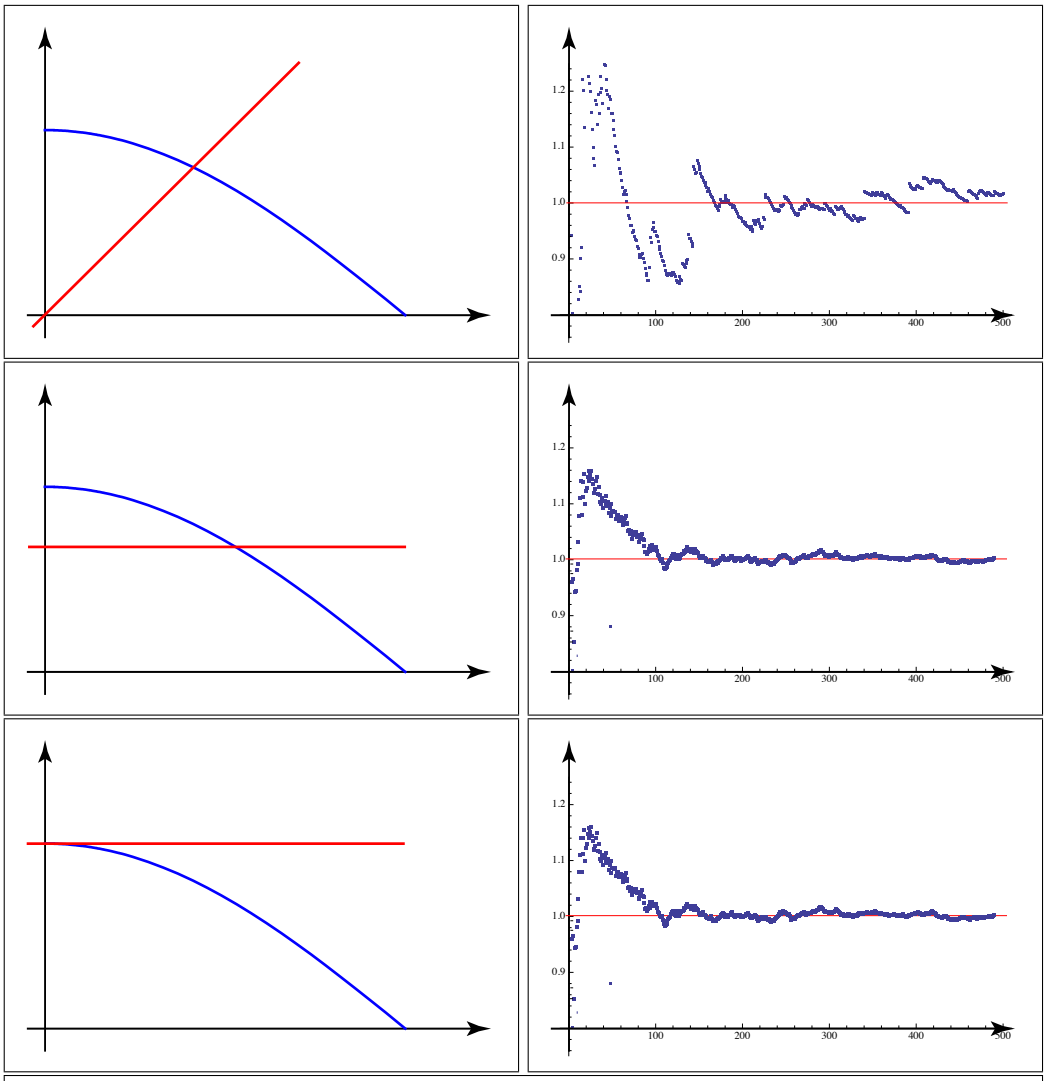


FIGURE 6.6: MONTE CARLO ESTIMATORS. The plots of Monte Carlo estimators for approximating the integral $\int_{[0,\frac{\pi}{2}]} \cos x \, d\mu(x)$. The two upper images illustrate the density $p_X(x) = \frac{8x}{\pi^2}$ and the associated estimator $F_N = \frac{\pi^2}{8N} \sum_{i=1}^{N} \frac{\cos(X_i)}{X_i}$ with $N = 500$ according to the density distributed random variables. The images in the middle show the uniform density $p_X(x) = \frac{2}{\pi}$ and the estimator $F_N = \frac{\pi}{2N} \sum_{i=1}^{N} \cos(X_i)$ based on 500 according to p_X-distributed random variables. The two lower images illustrate the transformation of uniformly distributed random variables from $[0,1]$ onto the interval $\left[0,\frac{\pi}{2}\right]$, where the corresponding estimator is given by $F_N = \frac{\pi}{2N} \sum_{i=1}^{N} \cos\left(\frac{\pi}{2} X_i\right)$. From the first two images on the right side we can conclude: The more a PDF approximates the form of the integrand, the better the estimator seems to approximate the integral.

Uniform sampling the density $p_{\mathbf{X}} = p_{X_1,X_2}$*, defined on the base set* P_j *of the probability space* $(P_j, \mathfrak{B}(P_j), \mathbb{P}_{\mathbf{X}})$*, where the density is given by*

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\mu^2(P_j)} \tag{6.120}$$

then leads to the following secondary Monte Carlo estimator for the integral from (2.207):

$$F_N^{F_{s_i P_j}} = \frac{\mu^2(P_j)}{N} \sum_{i=1}^{N} \frac{\left|\cos\theta_i^i \cos\theta_o^j\right| \; \mathcal{V}(\mathbf{X}_j \leftrightarrow \mathbf{s}_i)}{\pi \left\|\mathbf{X}_j - \mathbf{s}_i\right\|_2^2} \tag{6.121}$$

with $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *according to* $p_{\mathbf{X}}$ *chosen independent and on* P_j *uniformly distributed random variables.*

In Example 10.6, we will show that it is also straightforward to extend the above estimator $F_N^{F_{s_i P_j}}$ *to approximate the four-dimensional form factor integral from (10.26).*

EXAMPLE 6.9 (Trivial Pixel Sampling Using a Secondary Monte Carlo Estimator, Continued) *Let us consider the probability space* $(\square_j, \mathfrak{B}(\square_j), \mathbb{P}_{\mathbf{X}})$*, where the area of the pixel* $\square_j$ *is the base set,* $\mathfrak{B}(\square_j)$ *is as usual the Borel* σ*-algebra of measurable sets of* $\square_j$*, as well as* $\mathbb{P}_{\mathbf{X}}$ *is a probability measure. According to the Radon-Nikodým Theorem the probability measure* $\mathbb{P}_{\mathbf{X}}$ *can now be defined by means of the Lebesgue area measure as*

Borel σ-algebra (865)
Measurable Set (80)
Radon Nykodým Theorem (176)

$$\mathbb{P}_{\mathbf{X}}(\mathbf{B}) \stackrel{\text{def}}{=} \frac{\int_{\mathbf{B}} p_{\mathbf{X}}(\mathbf{x}) \, d\mu^2(\mathbf{x})}{\int_{\square_j} p_{\mathbf{X}}(\mathbf{x}) \, d\mu^2(\mathbf{x})} = \underbrace{\frac{\lambda^2(\mathbf{B})}{\lambda^2(\square_j)}}_{=1} = \lambda^2(\mathbf{B}), \qquad \forall\, \mathbf{B} \in \mathfrak{B}(\square_j), \tag{6.122}$$

where

$$\lambda^2(\mathbf{B}) = \int_{\mathbf{B}} p_{\mathbf{X}}(\mathbf{x}) \, d\mu^2(\mathbf{x}), \tag{6.123}$$

and $p_{\mathbf{X}}$ *is a probability density function on the area of the pixel* $\square_j$*. This means that if we choose a two-dimensional random variable* $\mathbf{X}$ *on* $\square_j$*, the probability, that* $\mathbf{X}$ *is drawn from a subset* $\mathbf{B}$ *of* $\square_j$*, is given by the ratio of the area of* $\mathbf{B}$ *and the area of the pixel* $\square_j$*.*

PDF (176)

This means: Replacing the integration measure μ^2 *within the integral from (??) with the help of Equations (6.122) and (6.123) by the Radon Nykodým derivative,*

Radon-Nikodým Derivative (176)

$$\mu^2 \stackrel{(6.68)}{=} \frac{d\mathbb{P}_{\mathbf{X}}}{p_{\mathbf{X}}}, \tag{6.124}$$

leads to:

$$\begin{aligned}\int_{\square_j} w(\mathbf{x})L(\mathbf{x})\,d\mu^2(\mathbf{x}) &= \int_{\square_j} \frac{w(\mathbf{x})L(\mathbf{x})}{p_{\mathbf{X}}(\mathbf{x})}\,d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) &(6.125)\\ &= \int_{\square_j} \frac{w(\mathbf{X}(\omega))L(\mathbf{X}(\omega))}{p_{\mathbf{X}}(\mathbf{X}(\omega))}\,d\mathbb{P}(\omega) &(6.126)\\ &= E\left(\frac{w(\mathbf{X})L(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})}\right). &(6.127)\end{aligned}$$

Following Relation (6.114), then for the secondary Monte Carlo estimator F_N*, approximating the color of a pixel, it holds:*

$$F_N = \frac{1}{N}\sum_{i=1}^{N} \frac{w(\mathbf{X}_i)L(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}, \tag{6.128}$$

with N *independent and identically, according to the probability density* $p_{\mathbf{X}}$*, over the area of the pixel* $\square_j$ *distributed samples* $\mathbf{X}_i = (X_{i1}, X_{i2})$*.*

Obviously, choosing the weighting function w *as probability density function* $p_{\mathbf{X}}$ *leads to:*

$$\begin{aligned}F_N &= \frac{1}{N}\sum_{i=1}^{N} \frac{w(\mathbf{X}_i)L(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}\\ &\overset{w=p_{\mathbf{X}}}{=} \frac{1}{N}\sum_{i=1}^{N} L(\mathbf{X}_i). &(6.129)\end{aligned}$$

This means that F_N *can simply be calculated as the sum of the radiance values at positions that are drawn independent and identically distributed according to the density* $p_{\mathbf{X}}$ *over the area of the pixel* $\square_j$*.*

The case, where a Monte Carlo estimator uses the density $p_{\mathbf{X}} = \frac{1}{\mu^2(\square_j)}$ *is called* box-filtering*, for details see [185, Shirley 2000].*

REMARK 6.6 (Advanced Filter Concepts for Pixel Sampling) *According to [184, Shirley & al. 1994], [185, Shirley 2000], the above described box-filtering is a less appropriate method with respect to shading a pixel. Applying* cubic B-spline blending functions *and* triangle filters*, separable and defined over the unit square* $\mathbf{I}^2$*, have been found to be far superior sampling strategies, the latter as they may be conceived of as the products of two one dimensional functions:* $1-|x|$ *and* $1-|y|$*.*

Another good choice of a filter for pixel filtering is the Gaussian filter*. It is based on a Gaussian bump symmetrical in* x *and* y*-direction with respect to the pixel center. Since a Gaussian function in two dimensions is separable into the product of two* 1D*-Gaussian functions, we can use*

$$f(x) = e^{-\alpha x^2} - e^{-\frac{\alpha e^2}{4}} \tag{6.130}$$

and

$$f(y) = e^{-\alpha y^2} - e^{-\frac{\alpha e^2}{4}}, \tag{6.131}$$

where the parameter α controls the falloff of the filter, and e is the size of the domain of the Gaussian where it takes on values unequal zero.

In [133, Mitchell & Netravali, 1988] a family of parametrized filter functions is presented, which solve some common artifacts from trivial reconstruction filters. Similar to the Gaussian filter, the Michell-Netravali filter *is also a product of* 1D*-filter functions in x and y direction, composed of two cubic polynomials valid in the closed interval $[-2, 2]$ and controlled by two parameters A, B. The Mitchell-Netravali filter is detailed discussed in [158, Pharr & Humphreys, 2004].*

Finally, we still mention the windowed sinc filter *composed of the sinc function and the* Lanczos windowing function*, w, defined by*

$$w(x) \stackrel{\text{def}}{=} \frac{\sin\left(\frac{\pi x}{\tau}\right)}{\frac{\pi x}{\tau}}, \tag{6.132}$$

where the parameter τ is used to control the cycles of the sinc function, for a detailed discussion see [158, Pharr & Humphreys, 2004].

To summarize, the following conclusions may be obtained from our first experiences with Monte Carlo methods applied to integrals:

REMARK 6.7 *Monte Carlo procedures used for evaluating the integral*

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \tag{6.133}$$

have been proven to be easily implementable algorithms. Apart from generating random variables $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \in \mathbf{Q}^s$, independent distributed according to some probability density functions, they require no more than the calculation and summing up of function values at the samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$.

Because they are independent on the dimension of the problem and the underlying topology of the integration domain, this method is by far superior to direct quadrature methods from numerical analysis. In particular, this holds for integrating functions with singularities and functions of dimension $s \geq 5$. Due to the fact that the method works with a much smaller set of samples than required by quadrature integration rules, Monte Carlo integration has become the most popular method for integrating multi-dimensional integrals. Unfortunately, this procedure has also a number of disadvantages, which have to be taken into account when developing efficient integration algorithms.

Section 6.1.1

PROPERTIES OF MONTE CARLO ESTIMATORS. Finally, we now define a number of important features of Monte Carlo estimators, which are useful for the successive derivation of a great variety of efficient Monte Carlo procedures.

DEFINITION 6.3 (Bias of a Monte Carlo Estimator) *Let $\mathcal{I}$ be some quantity of interest, normally the value of a given integral, and F_N be any Monte Carlo estimator for approximating the quantity $\mathcal{I}$. Then, we call the quantity $F_N - \mathcal{I}$ the* error, *and its expected value, thus,*

$$\begin{aligned} \beta(F_N) &\overset{\text{def}}{=} E(F_N - \mathcal{I}) && (6.134)\\ &= E(F_N) - E(\mathcal{I}) && (6.135)\\ &= E(F_N) - \mathcal{I}, && (6.136) \end{aligned}$$

the bias *of the estimator F_N. In other words, the bias is the difference between the estimator's expected value and the quantity which has to be estimated.*

The Monte Carlo estimator F_N is called unbiased *if for all sample sizes N it holds:*

$$\beta(F_N) = 0, \tag{6.137}$$

which is equivalent to

$$E(F_N) = \mathcal{I}. \tag{6.138}$$

If the bias of a Monte Carlo estimator F_N goes to zero with probability one, where the number of samples N increases then we call F_N consistent, *mathematically expressed as:*

$$\text{prob}\left\{\lim_{N\to\infty} \beta(F_N) = 0\right\} = 1 \quad \overset{(6.136)}{\Leftrightarrow} \quad \text{prob}\left\{\lim_{N\to\infty} E(F_N) = \mathcal{I}\right\} = 1. \tag{6.139}$$

Let F_N be an unbiased secondary Monte Carlo estimator then the following Lemma holds:

LEMMA 6.1 *The secondary Monte Carlo estimator F_N from Equation (6.114) for approximating the integral from (6.37) is an unbiased estimator.*

PROOF 6.1 *Due to the definition from above, F_N is unbiased, if the expected value of F_N is equal to the integral from (6.37) for all $N \in \mathbb{N}$. To show this, let N be a non-negative integer, then for $E(F_N)$ it holds:*

$$\begin{aligned} E(F_N) &= E\left(\frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{X}_i(\omega))}{p_{\mathbf{X}}(\mathbf{X}_i(\omega))}\right) && (6.140)\\ &\overset{p_{\mathbf{X}}=\frac{q}{\lambda(\mathbf{Q}^s)}}{=} E\left(\frac{\lambda(\mathbf{Q}^s)}{N}\sum_{i=1}^{N}\frac{f(\mathbf{X}_i(\omega))}{q(\mathbf{X}_i(\omega))}\right) && (6.141)\\ &\overset{(2.773)}{=} \frac{\lambda(\mathbf{Q}^s)}{N}\sum_{i=1}^{N} E\left(\frac{f(\mathbf{X}_i(\omega))}{q(\mathbf{X}_i(\omega))}\right). && (6.142) \end{aligned}$$

As the random variables $\mathbf{X}_1, \dots \mathbf{X}_N$ are independent and identically distributed according to the probabiliy denstity function $p_{\mathbf{X}}$, we can use $\mathbf{X}$ as a synonym for $\mathbf{X}_i$, i.e. we can write:

$$
\begin{aligned}
E(F_N) &\overset{\mathbf{x}_i = \mathbf{x}}{=} && \frac{\lambda(\mathbf{Q}^s)}{N} \sum_{i=1}^{N} E\left(\frac{f(\mathbf{X}(\omega))}{q(\mathbf{X}(\omega))} \right) && (6.143)\\
&= && \frac{\lambda(\mathbf{Q}^s)}{N} \sum_{i=1}^{N} \int_{\Omega} \frac{f(\mathbf{X}(\omega))}{q(\mathbf{X}(\omega))} \, d\mathbb{P}(\omega) && (6.144)\\
&= && \frac{\lambda(\mathbf{Q}^s)}{N} \sum_{i=1}^{N} \int_{\mathbf{Q}^s} \frac{f(\mathbf{x})}{q(\mathbf{x})} \, d\mathbb{P}_{\mathbf{X}}(\mathbf{x}) && (6.145)\\
&\overset{d\mu^s = \lambda(\mathbf{Q}^s)\frac{d\mathbb{P}_{\mathbf{X}}}{q}}{=} && \frac{1}{N} \sum_{i=1}^{N} \int_{\mathbf{Q}^s} f(\mathbf{x}) \, d\mu^s(\mathbf{x}) && (6.146)\\
&= && \int_{\mathbf{Q}^s} f(\mathbf{x}) \, d\mu^s(\mathbf{x}), && (6.147)
\end{aligned}
$$

which characterize F_N to be an unbiased Monte Carlo estimator to approximate the value of the integral $\int_{\mathbf{Q}^s} f(\mathbf{x}) \, d\mu^s(\mathbf{x})$.

REMARK 6.8 *From the definition above we conclude that any consistent estimator will ultimately converge towards the right answer if he uses more and more samples.*

We conclude furthermore from Relation (6.139) that an estimator is consistent, Variance (201) *if its bias and its variance both goes to zero, as the samples size N is increased, i.e.:*

$$\lim_{N \to \infty} \beta(F_N) = \lim_{N \to \infty} Var(F_N) = 0. \tag{6.148}$$

This means that an unbiased Monte Carlo estimator is consistent as long as its variance goes to zero as N goes to infinity.

Let us now demonstrate the concept of a biased estimator by means of an example from the field of computer graphics [220, Veach 1997].

EXAMPLE 6.10 (A Biased and Consistent Monte Carlo Estimator) *Suppose we are interested in antialiased samples on a pixel $\square_j$. For that, we have to estimate a quantity that is defined by an integral of the form*

$$\int_{\square_j} w(\mathbf{x}) f(\mathbf{x}) \, d\mu^2(\mathbf{x}), \tag{6.149}$$

where f is the image function on $\square_j$ and w is a filter function satisfying the normalization condition

$$\int_{\square_j} w(\mathbf{x}) \, d\mu^2(\mathbf{x}) = 1. \tag{6.150}$$

In sampling theory, a common way to compute the final value of a pixel is to use a weighted interpolation scheme. This results in the following estimator:

$$F_N = \frac{\sum_{i=1}^{N} w(\mathbf{X}_i) f(\mathbf{X}_i)}{\sum_{i=1}^{N} w(\mathbf{X}_i)}, \tag{6.151}$$

with uniformly and independent distributed random variables $\mathbf{X}_i$ *on* $\square_j$.

As is easily seen, the estimator F_N *is biased, since it holds:*

$$\begin{aligned}
E(F_N) &= E\left(\frac{\sum_{i=1}^{N} w(\mathbf{X}_i) f(\mathbf{X}_i)}{\sum_{i=1}^{N} w(\mathbf{X}_i)}\right) && (6.152)\\
&= \sum_{i=1}^{N} E\left(\frac{w(\mathbf{X}_i) f(\mathbf{X}_i)}{\sum_{i=1}^{N} w(\mathbf{X}_i)}\right) && (6.153)\\
&\overset{(2.732)}{=} \sum_{i=1}^{N} \int_{\square_j} \frac{w(\mathbf{x})}{\sum_{i=1}^{N} w(\mathbf{x})} f(\mathbf{x})\, d\mu^2(\mathbf{x}) && (6.154)\\
&= \frac{1}{N} \sum_{i=1}^{N} \int_{\square_j} \frac{w(\mathbf{x})}{w(\mathbf{x})} f(\mathbf{x})\, d\mu^2(\mathbf{x}) && (6.155)\\
&= \int_{\square_j} f(\mathbf{x})\, d\mu^2(\mathbf{x}) && (6.156)\\
&\neq \int_{\square_j} w(\mathbf{x}) f(\mathbf{x})\, d\mu^2(\mathbf{x}). && (6.157)
\end{aligned}$$

However, using an estimator F_N *of the form*

$$F_N = \frac{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{X}_i) f(\mathbf{X}_i)}{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{X}_i)}, \tag{6.158}$$

then due to the Strong Law of Large Numbers we get: SLLN (216)

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} w(\mathbf{X}_i) f(\mathbf{X}_i) \overset{(2.11)}{=} \int_{\square_j} w(\mathbf{x}) f(\mathbf{x})\, d\mu^2(\mathbf{x}). \tag{6.159}$$

and

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} w(\mathbf{X}_i) \overset{(2.11)}{=} \int_{\square_j} w(\mathbf{x})\, d\mu^2(\mathbf{x}). \tag{6.160}$$

Making use of these two identities in the definition of a consistent estimator,

then we get for F_N*:*

$$\lim_{N\to\infty} E(F_N) = \lim_{N\to\infty} E\left(\frac{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{X}_i) f(\mathbf{X}_i)}{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{X}_i)}\right) \tag{6.161}$$

$$= E\left(\lim_{N\to\infty} \frac{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{X}_i) f(\mathbf{X}_i)}{\frac{1}{N}\sum_{i=1}^{N} w(\mathbf{X}_i)}\right) \tag{6.162}$$

$$= E\left(\frac{\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} w(\mathbf{X}_i) f(\mathbf{X}_i)}{\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} w(\mathbf{X}_i)}\right) \tag{6.163}$$

$$= E\left(\frac{\int_{\square_j} w(\mathbf{x}) f(\mathbf{x})\, d\mu^2(x)}{\int_{\square_j} w(\mathbf{x})\, d\mu^2(\mathbf{x})}\right) \tag{6.164}$$

$$\overset{(6.150)}{=} \int_{\square_j} w(\mathbf{x}) f(\mathbf{x})\, d\mu^2(\mathbf{x}), \tag{6.165}$$

that is, the estimator F_N *is consistent.*

In [159, Pharr & Humphreys 2010] it is shown that in practice the biased estimator F_N *from above should be preferred over the unbiased estimator from Relation (6.128) from Example 6.9, since it results in less variance and the variance in the unbiased estimator leads to an undesirable result in the final image. So, if all radiance values* $L(\mathbf{X}_i)$ *are one, the above biased estimator leads, using* $f = L$ *in Formula (6.151), to pixel values which are also one, while the unbiased estimator*

$$F_N^{unbiased} = \frac{1}{N}\sum_{i=1}^{N} \frac{w(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)} \tag{6.166}$$

$$\overset{p_{\mathbf{X}}=\frac{1}{\mu^2(\square_j)}}{=} \frac{\mu^2(\square_j)}{N}\sum_{i=1}^{N} w(\mathbf{X}_i) \tag{6.167}$$

results—due to the variation in the filter function w*—in pixel values which are not all one, since the sum over the filter function will generally not be equal to the Lebesgue measure* $\mu^2(\square_j)$*. Obviously, in this specific case we will have undesirable variance in the image.*

Furthermore, [159, Pharr & Humphreys 2010] argue that in more complex images the variance that would be introduced by the unbiased estimator is a more objectionable artifact than the bias from Equation (6.151).

So far, we have only talked how an integral can be approximated by a finite sum of function values that are evaluated at randomly chosen points. We have not spoken about the convergence behavior of a Monte Carlo estimator. For that, we now introduce the concept of the *mean square error*, also called *MSE*. It measures the average of the square of the difference between an estimator and the true value of the quantity being estimated.

DEFINITION 6.4 (Mean Square Error of a Monte Carlo Estimator) *The mean square error of a Monte Carlo estimator* F_N *is defined as*

$$MSE(F_N) \overset{def}{=} E\left((F_N - \mathcal{I})^2\right). \tag{6.168}$$

As the 2^{nd}-moment of the error, the MSE apparently incorporates both the variance of the estimator and its bias. Let us show this in the following lemma.

n^{th}-moment (201)

LEMMA 6.2 *For the mean square error of any Monte Carlo estimator* F_N *it holds:*

$$MSE(F_N) = Var(F_N) + \beta^2(F_N). \tag{6.169}$$

PROOF 6.2 *Let* F_N *be any Monte Carlo estimator for approximating the quantity* $\mathcal{I}$*, then we have:*

$$\begin{aligned}
MSE(F_N) &= E\left((F_N - \mathcal{I})^2\right) && (6.170)\\
&= E\left((F_N - E(F_N) + E(F_N) - \mathcal{I})^2\right) && (6.171)\\
&= E\big((F_N - E(F_N))^2\big) + && (6.172)\\
&\quad 2E\left((F_N - E(F_N))\left(E(F_N) - \mathcal{I}\right)\right) + E\left((E(F_N) - \mathcal{I})^2\right) \\
&\overset{E(E(F_N))=E(F_N)}{=} E\left((F_N - E(F_N))^2\right) + E\left((E(F_N) - \mathcal{I})^2\right) && (6.173)\\
&\overset{(6.134)}{=} Var(F_N) + \beta^2(F_N). && (6.174)
\end{aligned}$$

From the lemma above we conclude: To estimate the error of any estimator, we need, in addition to the variance of the estimator, an upper bound on the possible bias. Now it is often very difficult to find such a suitable bound because it requires additional information about the estimate $\mathcal{I}$, which is mostly not available. It is much easier to estimate the error of an unbiased estimator F_N since the MSE then has the fortunate property to be identical to the variance, thus,

Upper Bound (862)

$$MSE(F_N) = Var(F_N). \tag{6.175}$$

This is also the main reason why we are interested in finding unbiased Monte Carlo estimators for approximating the integral from (6.37). Thus, error estimations with respect to unbiased estimators can be easily indicated via the variance of F_N by choosing independent samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ and defining

$$F_N = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i. \tag{6.176}$$

Let us now yet take a look at the variance of a secondary Monte Carlo estimator F_N. In the following lemma we will see that it can be expressed in terms of the variance of the primary estimator.

LEMMA 6.3 *Let* F_N*, given by*

$$F_N \stackrel{def}{=} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i), \tag{6.177}$$

be a secondary Monte Carlo estimator based on independent and identically distributed random variables $\mathbf{X}_i$ *defined on the probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$*. Then, the variance of* F_N *decreases linearly with* N*, i.e. it holds:*

$$Var(F_N) = \frac{1}{N} Var(F_1). \tag{6.178}$$

PROOF 6.3 *Let* F_N *be the secondary Monte Carlo estimator from Equation (6.177) then we obtain for it's variance:*

$$\begin{aligned} Var(F_N) &= Var\left(\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i)\right) \\ &\stackrel{(2.772),(2.791)}{=} \frac{1}{N^2}\sum_{i=1}^{N} Var(f(\mathbf{X}_i)). \end{aligned} \tag{6.179}$$

Since the random variables $\mathbf{X}_i, 1 \leq i \leq N$ *are independent and identically distributed, we can replace* $\mathbf{X}_i$ *in the above formula by the random variable* $\mathbf{X}$ *which results in*

$$Var(F_N) \stackrel{\mathbf{x}_i=\mathbf{x}}{=} \frac{1}{N^2} Var \sum_{i=1}^{N} (f(\mathbf{X})) \tag{6.180}$$

$$= \frac{1}{N} Var(f(\mathbf{X})) \tag{6.181}$$

$$= \frac{1}{N} Var(F_1), \tag{6.182}$$

where we have used the identity $F_1 = f(\mathbf{X})$*.*

As the variance decreases linearly with $N.$*, we can conclude that the error of an unbiased Monte Carlo estimator can be made as small as desired, provided, we draw a sufficient large number of samples.*

With the following example, let us illustrate how Monte Carlo integration works:

EXAMPLE 6.11 *Let us estimate the one-dimensional integral*

$$\mathcal{I} = \int_{[0,1]} x^2 \, d\mu(x) \tag{6.183}$$

using uniformly distributed random variables drawn from $[0,1]$ *via the PDF* $p_X = 1$. *Then, a secondary Monte Carlo estimator looks like:*

$$F_N = \frac{1}{N}\sum_{i=1}^{N} X_i^2. \tag{6.184}$$

Obviously, the estimator F_N *is unbiased—we leave the proof to the interested reader as a simple exercise. Since it holds* $\mathcal{I} = \frac{1}{3}$, *the variance of this function can be analytically computed as follows:*

$$\begin{aligned} \mathrm{Var}(F_N) &\overset{(6.182)}{=} \frac{1}{N}F_1 && (6.185)\\ &= \frac{1}{N}\left(\int_{[0,1]} x^4 \, d\mu(x) - \frac{1}{9}\right) && (6.186)\\ &= \frac{8}{9N}. && (6.187) \end{aligned}$$

As F_N *is unbiased, we have for the mean square error of* F_N*:*

$$\mathrm{MSE}(F_N) = \frac{8}{9N}. \tag{6.188}$$

This means: The estimator F_N *results in increasingly better approximations for* $\mathcal{I}$ *as* N *goes to infinity.*

Now, the above lemma confirms our first results from the foregoing discussion about the convergence of Monte Carlo estimators, namely, that the variance of an unbiased Monte Carlo estimator can be made as small as desired, provided, we take sufficiently many samples. According to [99, Kalos & Whitlock 1986], we can estimate the variance of any unbiased estimator by means of the following theorem.

THEOREM 6.1 *Let* F_N *be an unbiased estimator, given by:*

$$F_N \overset{\mathrm{def}}{=} \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i), \tag{6.189}$$

based on independent and identically distributed random variables $\mathbf{X}_i$ *defined on the probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$. *Then, the quantity* $\widehat{F}_N$ *defined by:*

$$\widehat{F}_N \overset{\mathrm{def}}{=} \frac{1}{N-1}\left\{\left(\frac{1}{N}\sum_{i=1}^{N} f^2(\mathbf{X}_i)\right) - \left(\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i)\right)^2\right\} \tag{6.190}$$

is an unbiased estimator for $\mathrm{Var}(F_N)$, *that is, it holds:*

$$\mathrm{E}(\widehat{F}_N) = \mathrm{Var}(F_N). \tag{6.191}$$

PROOF 6.1 *Let us consider the expected value of* $\mathsf{E}(\widehat{\mathsf{F}}_N)$ *then it holds:*

$$\begin{aligned}
\mathsf{E}(\widehat{\mathsf{F}}_N) &\overset{def}{=} \mathsf{E}\left(\frac{1}{N-1}\left\{\left(\frac{1}{N}\sum_{i=1}^{N} f^2(\mathbf{X}_i)\right) - \left(\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i)\right)^2\right\},\right) && (6.192)\\
&= \frac{1}{N-1}\left\{\mathsf{E}\left(\frac{1}{N}\sum_{i=1}^{N} f^2(\mathbf{X}_i)\right) - \mathsf{E}\left(\left(\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{X}_i)\right)^2\right)\right\}, && (6.193)\\
&= \frac{1}{N-1}\left\{\frac{1}{N}\mathsf{E}\left(\sum_{i=1}^{N} f^2(\mathbf{X}_i)\right) - \frac{1}{N^2}\mathsf{E}\left(\left(\sum_{i=1}^{N} f(\mathbf{X}_i)\right)^2\right)\right\} && (6.194)\\
&= \frac{1}{N-1}\frac{1}{N}\left\{\sum_{i=1}^{N}\mathsf{E}\left(f^2(\mathbf{X}_i)\right) - \frac{1}{N}\mathsf{E}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} f(\mathbf{X}_i)f(\mathbf{X}_j)\right)\right\}. && (6.195)
\end{aligned}$$

Now, the expected value over the double sum of the random variables $\mathbf{X}_i$ *can be written as:*

$$\begin{aligned}
\mathsf{E}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} f(\mathbf{X}_i)f(\mathbf{X}_j)\right) &= \sum_{i=1}^{N}\mathsf{E}(f^2(\mathbf{X}_i)) + \sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\mathsf{E}(f(\mathbf{X}_i)f(\mathbf{X}_j)) && (6.196)\\
&= \sum_{i=1}^{N}\mathsf{E}(f^2(\mathbf{X}_i)) + \sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\mathsf{E}(f(\mathbf{X}_i))\mathsf{E}(f(\mathbf{X}_j)), && (6.197)
\end{aligned}$$

where we have used the independence of the random variable $\mathbf{X}_i, \mathbf{X}_j$ *in the second step. Using this identity in the foregoing Equation then we get:*

$$\begin{aligned}
\mathsf{E}(\widehat{\mathsf{F}}_N) &= \frac{1}{N-1}\frac{1}{N}\left(\frac{N-1}{N}\sum_{i=1}^{N}\mathsf{E}\left(f^2(\mathbf{X}_i)\right) - \frac{1}{N}\sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\mathsf{E}(f(\mathbf{X}_i)\mathsf{E}(f(\mathbf{X}_j)\right) && (6.198)\\
&= \frac{1}{N-1}\frac{1}{N}\left(\frac{N-1}{N}\sum_{i=1}^{N}\mathsf{E}\left(f^2(\mathbf{X}_i)\right) - \frac{1}{N}\sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\mathsf{E}(f(\mathbf{X}_i)\mathsf{E}(f(\mathbf{X}_j)\right) && (6.199)\\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathsf{E}\left(f^2(\mathbf{X}_i)\right) - \frac{1}{N-1}\frac{1}{N^2}\sum_{\substack{i=1\\ i\neq j}}^{N}\sum_{j=1}^{N}\mathsf{E}(f(\mathbf{X}_i)\mathsf{E}(f(\mathbf{X}_j). && (6.200)
\end{aligned}$$

Now, as we have assumed that the random variables $\mathbf{X}_i$ *are identically distributed, it holds:* $\mathsf{E}(\mathbf{X}_i) = \mathsf{E}(\mathbf{X})$. *So, we can express the expected value of the random variable* $f(\mathbf{X}_i)$ *in Equation 6.200 via the expected value of the random variable* $\mathbf{X}$,

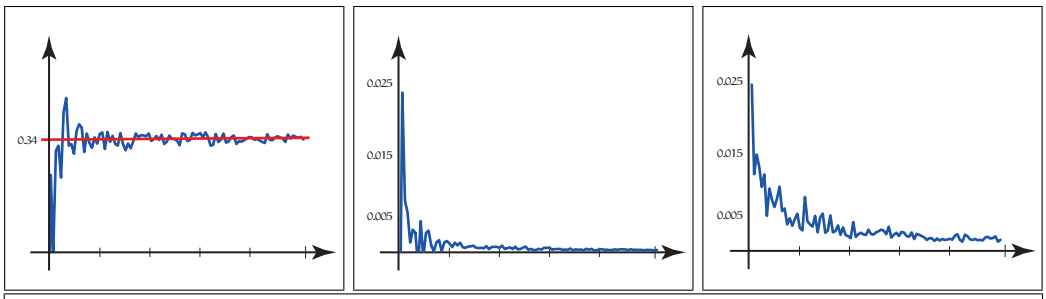


FIGURE 6.7: VARIANCE ANALYSIS OF THE SECONDARY MONTE CARLO ESTIMATOR $\sum_{i=1}^{N} U_i^2$. Error and variance estimations of the Monte Carlo estimator $F_N = \sum_{i=1}^{N} U_i^2$ assigned to the integral $\int_{[0,1]} x^2 \, d\mu(x)$ with the random variables $U_1, U_2, \ldots, U_N, N = 500$ independent and uniformly distributed on $I = [0, 1]$. The expected value $E(F_N)$, the MSE-Error $MSE(F_N)$, and the variance $Var(F_N)$ of $F_N = \sum_{i=1}^{N} U_i^2$.

namely:

$$
\begin{aligned}
E(\widehat{F}_N) &= \frac{1}{N} E(f^2(\mathbf{X})) - \frac{1}{N} E^2(f(\mathbf{X})) && (6.201)\\
&= \frac{1}{N} Var(f(\mathbf{X})) && (6.202)\\
&= \frac{1}{N} Var(F_1) && (6.203)\\
&\overset{\text{Lemma 6.3}}{=} Var(F_N). && (6.204)
\end{aligned}
$$

6.4 CONVERGENCE OF THE MONTE CARLO INTEGRATION

In the foregoing discussion, we have introduced the principle of Monte Carlo integration. As a consequence from the Strong Law of Large Numbers, we were able to make statements on the accuracy of any Monte Carlo computation for estimating an integral. However, until this point nothing could be said on the convergence rate of the method. In order to shine some light on this feature of Monte Carlo integration, let us apply the estimator F_N from Equation (6.114) in the following discussion on the integral from Equation (6.37). For that we use random variables $\mathbf{X}_i, 1 \leq i \leq N$, which are independent and identically distributed with respect to a probability density function p given on the probability space $(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s), \mathbb{P}_{\mathbf{X}})$. First weak statements about the convergence of the Monte Carlo method can be obtained using the Chebychev Inequality.

Theorem of SLLN (216)

Probability Space (163)

Chebychev Inequality (212)

Now Chebychev's Inequality requires that the integrand f is an element of the function space $\mathcal{L}^2(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s))$, which ensures that both $E(F_N)$ and $Var(F_N)$ exist and are finite. Then, applying Chebychev's Inequality to the estimator F_N leads to:

Monte Carlo Estimator (500)
Expeted Value of RV (196)
Variance of RV (201)

$$\text{prob}\left\{|F_N - E(F_N)| \geq \left(\frac{Var(F_N)}{\delta}\right)^{\frac{1}{2}}\right\} \leq \delta, \tag{6.205}$$

where $\delta \in [0,1]$. In [99, Kalos & Whitlock 1986], this inequality is called the *First Fundamental Theorem of Monte Carlo*.

Standard Deviation (213)

Expressed in terms of the standard deviation $\sigma(F_N)$, which also can be considered as a measurement of the error—namely, the *root mean square error*, or RMSE—Chebychev's inequality has the form:

$$\text{prob}\left\{|F_N - E(F_N)| \geq \frac{\sigma(F_N)}{\sqrt{\delta}}\right\} \leq \delta, \tag{6.206}$$

and can interpreted as follows: The probability that F_N differs from its expected value by more than $\frac{1}{\sqrt{\delta}}$ standard deviations is at most δ, i.e. for $\delta = \frac{1}{1000000}$, the probability that $F_N = E(F_N)$ is very large.

Now, applying Lemma 6.3 to the estimator F_N

$$Var(F_N) = \frac{(\lambda(\mathbf{Q}^s))^2}{N} Var(F_1). \tag{6.207}$$

and inserting this relation in Chebychev's Inequality, then we get:

$$\text{prob}\left\{|F_N - E(F_N)| \geq \frac{\lambda(\mathbf{Q}^s)}{\sqrt{N}}\left(\frac{Var(F_1)}{\delta}\right)^{\frac{1}{2}}\right\} \leq \delta \tag{6.208}$$

or with $F_1 = \frac{f(\mathbf{X})}{p(\mathbf{X})}$

$$\text{prob}\left\{|F_N - E(F_N)| \geq \frac{\lambda(\mathbf{Q}^s)}{\sqrt{N}}\left(\frac{Var\left(\frac{f(\mathbf{X})}{p(\mathbf{X})}\right)}{\delta}\right)^{\frac{1}{2}}\right\} \leq \delta. \tag{6.209}$$

As the variance of the estimator decreases with increasing sampling size N, the probability of getting a large deviation between the exact value and an estimate of the integral becomes very small. Relation (6.209) also allows to pull conclusions to the convergence behavior of a Monte Carlo algorithm. Thus, algorithms based on the principle of Monte Carlo integration for estimating an integral have a convergence rate of the order $O\left(\frac{1}{\sqrt{N}}\right)$, i.e. in order to halve the error of an estimate, the number of samples used must be quadrupled. This slow convergence behavior of a Monte Carlo algorithm is the classic result of the Monte Carlo method.

According to [197, Sobol 1985], stronger error boundaries to those received above
CLT (217) may be obtained by applying the results of the Central Limit Theorem. Thus, the CLT states that the values of an estimator F_N converges to a normal distribution as N goes to infinity. Therefore, the estimate lies in a close region around the expected value of the integral with higher probability.

Furthermore, we note, as N gets larger, the standard deviation decreases with $\frac{1}{\sqrt{N}}$ since it holds:

$$\sigma(F_N) \overset{def}{=} \sqrt{Var(F_N)} \tag{6.210}$$

$$\overset{(6.182)}{=} \frac{\lambda(\mathbf{Q}^s)}{\sqrt{N}}\sqrt{Var(F_1)} \tag{6.211}$$

$$= \frac{\lambda(\mathbf{Q}^s)}{\sqrt{N}}\sigma(F_1)\,. \tag{6.212}$$

Based on these fact the Central Limit Theorem then states that

$$\lim_{N\to\infty} prob\left\{F_N - E(F_N) \le t\frac{\lambda(\mathbf{Q}^s)\sigma(F_1)}{\sqrt{N}}\right\} = \frac{1}{\sqrt{2\pi}}\int_{]-\infty,t]} e^{\frac{-x^2}{2}}\,d\mu(x), \tag{6.213}$$

whereas the expression on the right-hand side is the well-known *nomal distribution*. Due to [221, Veach 1998] this equation can also be rearranged to give

$$prob\left\{\left|F_N - \int_{\mathbf{Q}^s} f(\mathbf{x})\,d\mu^s(\mathbf{x}))\right| \ge t\,\sigma(F_N)\right\} = \sqrt{\frac{2}{\pi}}\int_t^\infty e^{\frac{-x^2}{2}}\,dx. \tag{6.214}$$

Compared with Chebyshev's Inequality, the CLT leads to more powerful statements on the convergence behavior of the Monte Carlo method. It does not only allow to predict the probability of deviations measured in units of σ, but also allows a statement on the distribution of the values of the estimator F_N. Note: The CLT applies only asymptotically, that is, when N is large. How large N must be before the CLT applies is not clear. This often depends on the given problem. As we assume that N is large enough in our discussion, this restriction makes no problems for us. Care must be taken when small values of N are used.

EXAMPLE 6.12 *As the integral on the right decreases very fast, it can be derived for* $t = 3$ *that, if the sample size* N *is large enough, there is only about a* 0.3% *chance that* F_N *will differ from its expected value by more than three standard deviations.*

As a conclusion of this section we can note that Monte Carlo integration compared with other numerical integration methods not only has advantages. One of it's main disadvantages is that apart from the existence of merely probabilistic error boundaries, the method is strongly dependent on both, convergence results and error estimations of the involved random numbers. Another disadvantage is that even sufficiently smooth functions lead to the probabilistically slow convergence behavior, typical for Monte Carlo integration.

6.5 SAMPLING

Now, Monte Carlo estimators are based on the evaluation of an integral at a large number of points randomly chosen according to a given probability density function over its integration domain. In Example 6.7 we have seen that different PDFs can be used to generate samples from the same integration domain. In Figure 6.6, we have also seen, that different sets of samples, resulting from sampling different densities, lead to different convergence behavior of the corresponding estimators. Obviously, this then implies that the process of *sampling random numbers* has a crucial role when constructing of efficient Monte Carlo algorithms for evaluating integrals.

In the following sections we will present three different techniques for sampling random variables: The *transformation method*, the most well-known and most frequently used sampling procedure in the theory of Monte Carlo algorithms, *acceptance rejection sampling*, and the method of *Markov chain Monte Carlo*.

Section 6.5.1 The idea behind the transformation method is to map uniformly distributed random variables to random variables from a desired distribution, whose PDF can be integrated analytically. As we will see, often it is not possible to derive a formula for the cumulative distribution function via application of the transformation method, in this case the Section 6.5.2 last resort for independent sampling is *acceptance-rejection sampling*. Instead of sampling directly from a density, which is commonly difficult or even impossible to sample, in acceptance-rejection sampling another easily to sample density function—a so-called proposal density function, that approximates the desired density—is used. The algorithm then decides if a proposed sample is accepted, or if it should be rejected. Now, the main drawback of the acceptance-rejection method is that it is often very difficult to construct a suitable proposal distribution that leads to an efficient algorithm. One way to avoid this problem is to allow the proposed samples depend on the last accepted samples, which makes it easier to generate a suitable, but now conditional proposal distribution. The price, we pay for that is to generate samples from a sequence of dependent random variables instead of a sequence of independent random variables. Such procedures are known Section 6.5.3 under the notion of *Markov chain Monte Carlo algorithms*, a class of sampling techniques based on the Metropolis algorithm.

6.5.1 THE TRANSFORMATION METHOD

Let us consider the problem of generating a random variable $\mathbf{X}$, distributed according to a probability density function $p_{\mathbf{X}}$, where the associated cumulative distribution function $F_{\mathbf{X}}$ is known. The problem that we wish to solve is as follows: To a given random variable $\mathbf{X}$ with known probability density function $p_{\mathbf{X}}$ and a function $\mathbf{T}$ we are seeking the density

function of the random variable $\mathbf{Y} = \mathbf{T}(\mathbf{X})$ in terms of $p_{\mathbf{X}}$. This leads us to the *transformation method*, the most well-known and most frequently used sampling procedure in the theory of Monte Carlo algorithms. Chapter 9

Let us first discuss the simple one-dimensional case, afterwards we will present the more general, complex multivariate case.

THE ONE-DIMENSIONAL TRANSFORMATION METHOD. In the one-dimensional case, i.e., in the case where we want to sample a one-dimensional random variable X, the transformation method works as follows: Let $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ be a probability space, X a random variable defined on $(\Omega, \mathfrak{F}(\Omega))$ with cumulative distribution function $F_X(x)$, and $Y = T(X)$ is the image of the random variable X under a bijective, differentiable mapping T. Then, the following applies to the probability measure $\mathbb{P}_Y$: Random Variable (168) Probability Space (163) CDF (171)

$$\begin{aligned}
\mathbb{P}_{\mathbf{Y}}(Y \leq y) &\overset{(2.549)}{=} \left(\mathbb{P} \circ Y^{-1}\right)(Y \leq y) && (6.215)\\
&\overset{Y=T(X)}{=} \left(\mathbb{P} \circ (T \circ X)^{-1}\right)(Y \leq y) && (6.216)\\
&= \left(\mathbb{P} \circ X^{-1} \circ T^{-1}\right)(Y \leq y) && (6.217)\\
&= \left(\mathbb{P} \circ X^{-1}\right)(T^{-1}(Y) \leq T^{-1}(y)) && (6.218)\\
&= \left(\mathbb{P} \circ X^{-1}\right)(X \leq T^{-1}(y)) && (6.219)\\
&= \mathbb{P}_{\mathbf{X}}(X \leq T^{-1}(y)). && (6.220)
\end{aligned}$$

Based on this result, we conclude due to the definition of the cumulative distribution function for continuous random variables and $x = T^{-1}(y)$ that it holds: CDF (179)

$$\int_{(-\infty, y]} p_Y(y)\, d\mu(y) = \int_{(-\infty, T^{-1}(y)]} p_X(x)\, d\mu(x), \tag{6.221}$$

where p_X and p_Y are the probability density functions of the continuous random variables X and Y. Changing from variable x to $y = T^{-1}(x)$ in the second integral and applying the Theorem of Transformation for integrals yields: PDF (176) Theorem of Transformation (117)

$$\int_{(-\infty, T^{-1}(y)]} p_X(x)\, d\mu(x) = \int_{(-\infty, y]} p_X(T^{-1}(y)) \left| \frac{dT^{-1}(y)}{d\mu(y)} \right| d\mu(y). \tag{6.222}$$

Obviously it is possible to express the probability density function p_Y of the random variable Y in terms of p_X, namely by

$$p_Y(y) = p_X(T^{-1}(y)) \left| \frac{dT^{-1}(y)}{d\mu(y)} \right|. \tag{6.223}$$

Let us illustrate this result by means of an example, namely: Generating a uniformly distributed random variable on the interval $[a, b]$.

EXAMPLE 6.13 (Uniform Sampling on the Interval $[a, b]$) *Let U be the uniformly distributed random variable from part i) of Example (2.69) with probability density function $p_U = 1$. Furthermore, let T be a bijective, differentiable function and Y a random variable with $Y = T(U)$ given by $T(U) = a + U(b - a)$. Obvioulsy, then it holds:*

Uniform Distribution (180) PDF (176)

$$U = \frac{T(U) - a}{b - a} \tag{6.224}$$

$$\overset{Y=T(U)}{=} \frac{Y - a}{b - a}. \tag{6.225}$$

Since T is invertible we get:

$$T^{-1}(Y) = U = \frac{Y - a}{b - a}. \tag{6.226}$$

Due to the above results, the corresponding probability density function p_Y is given by

$$p_Y(y) = p_U(T^{-1}(y)) \left| \frac{dT^{-1}(y)}{d\mu(y)} \right|. \tag{6.227}$$

$$= p_U(u) \left| \frac{d\left(\frac{y-a}{b-a}\right)}{d\mu(y)} \right|. \tag{6.228}$$

$$\overset{p_U(u)=1}{=} \frac{1}{b - a}, \tag{6.229}$$

that is, the transformation $Y = a + U(b - a)$ of the uniformly distributed random variable U has the probability density function $p_Y = \frac{1}{b-a}$.

THE ONE-DIMENSIONAL INVERSION METHOD. An algorithm based on the transformation method used in many algorithms is the so-called *inversion method*. The goal of the inversion method is the generation of independent and according to a given density function p distributed random variables, see Figur 6.8. It supplies a rule for generating a $\mathbb{P}$-distributed random variable using a random variable U which is independently and uniformly distributed over the unit interval I.

Setting in the above derivation $X = U$ then we get with $Y = T(X) = T(U)$ for the probability measure $\mathbb{P}_Y$:

$$\mathbb{P}_Y(Y \leq y) = \mathbb{P}_U(U \leq T^{-1}(y)). \tag{6.230}$$

The corresponding CDFs then have form

$$F_Y(y) = F_U(T^{-1}(y)) \tag{6.231}$$

$$\overset{T^{-1}(y)=u}{=} F_U(u) = u, \tag{6.232}$$

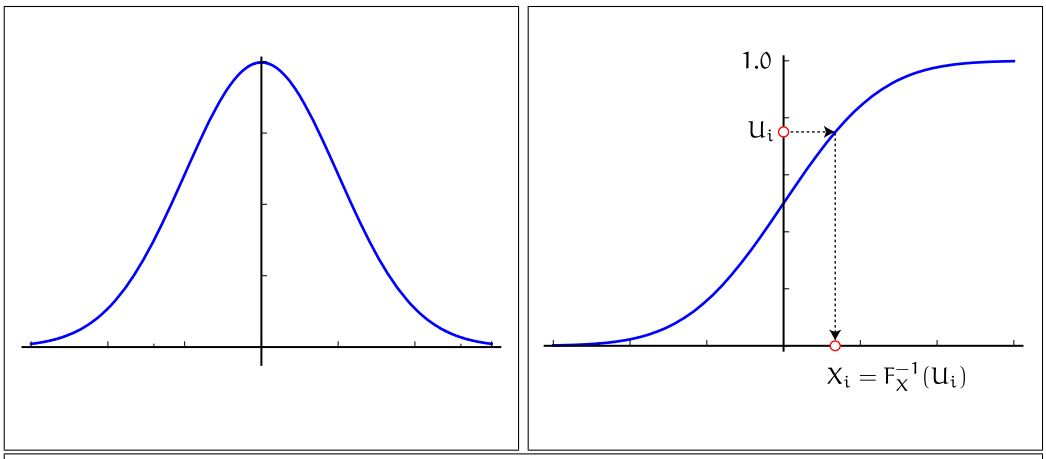


FIGURE 6.8: INVERSION METHOD. Generating exponential distributed random numbers with the help of uniformly distributed random variables on the unit interval $[0, 1]$. The left image shows the associated PDF. The image on the right-hand side illustrates the inversion method.

CDF (171) which, with an invertible CDF F_Y, can also be written as:

$$y = F_Y^{-1}(F_U(u)) = F_Y^{-1}(u). \tag{6.233}$$

From this discussion, we conclude that we can draw a sample Y from any arbitrary PDF $p_Y(y)$ via the following four steps:

i) compute the CDF $F_Y(y)$

ii) compute the inverse $F_Y^{-1}(y)$

iii) draw a uniformly distributed random number U

iv) compute $Y = F_Y^{-1}(U)$.

Let us show the use of the inversion method by means of a few interesting examples.

EXAMPLE 6.14 (Sampling from a Discrete Distribution) *Suppose X be a discrete random variable resulting from a Bernoulli random experiment with probability mass function $p_X(i) = \frac{1}{2^i}$ for $i \geq 1$. To draw a sample according to p_X via the inversion method, a canonical uniform random variable U can be plotted on the vertical axis of the plot of the appropriate CDF, see Figure 6.9. Now, the horizontal extension of U intersects the box representing the i^{th} outcome with probability $p_X(i)$, that is, the resulting distribution is distributed to the PMF, p_X.*

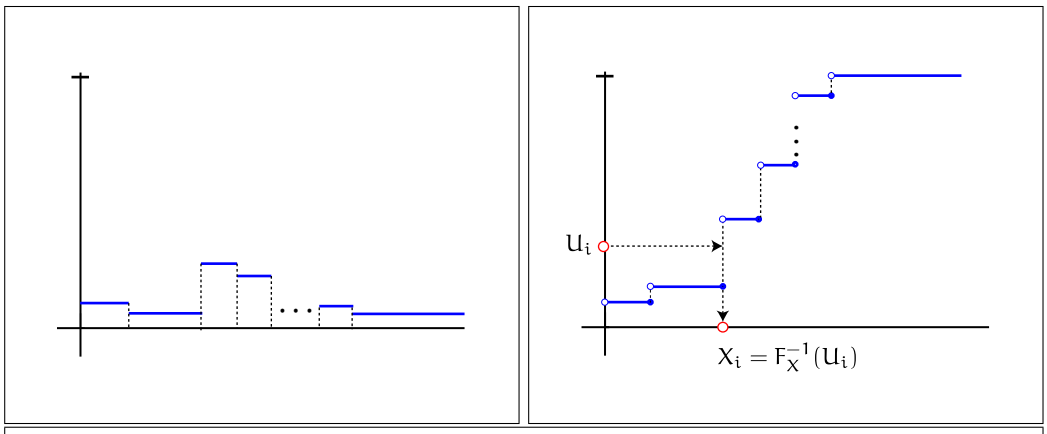


FIGURE 6.9: SAMPLING FROM A DISCRETE DISTRIBUTION VIA THE INVERSION METHOD.
Left the probability mass function, right the associated cumulative distribution function.

The inversion method is an important tool, since it allows in a simple way to sample from non-uniform distributions by applying transformations to uniform distributed random variables. Compared with the transformation method it has the disadvantage that it only works with uniform samples. Both methods permit to transform a stratification of $[0,1]$ generated with uniformly distributed random variables onto a stratification of the integration domain underlying an integral via a selected density, see Figure 6.8. While, as will been seen in the following paragraph, these methods yield more efficient sampling procedures, on the other hand they also have the disadvantage of assuming that the density is analytically integrable, a feature which in applications of computer graphics is not always given.

Uniform Distribution (180)

EXAMPLE 6.15 (Sampling According to a Power Distribution) *In the Blinn microfacet model a surface is statistically described by a distribution function* $D(\omega_h)$ *defined by*

$$D(\omega_h) \propto \langle \omega_h, \mathbf{N}(\mathbf{s}) \rangle^e, \tag{6.234}$$

where ω_h *is the half-angle vector between the incoming light direction* ω_i *and the exitant direction* ω_o*, and* $N(\mathbf{s})$ *is the averaged surface normal at point* $\mathbf{s}$*.*

Now, the dot product can also be expressed in terms of a cosine, namely by $\cos^e \theta_h$*, where* $\cos\theta_h = \langle \omega_h, \mathbf{N}(\mathbf{s}) \rangle$*, that is, the distribution* $D(\omega_h) \propto \cos^e \theta_h$ *can be interpreted as a power distribution of the form* $p_X(x) = Cx^n$*, for some constant* C*. To sample from* p_X*, we have firstly to find the proportionality constant* C*, that is, we have to evaluate the integral*

$$\int_{[0,1]} Cx^n \, d\mu(x) = C \frac{x^{n+1}}{n+1} \Bigg|_0^1 = 1, \tag{6.235}$$

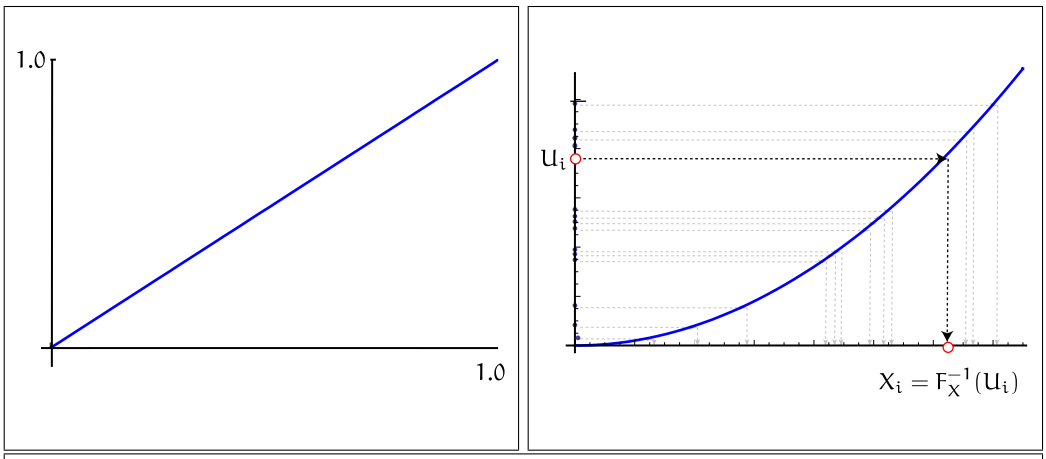


FIGURE 6.10: SAMPLING FROM A POWER DISTRIBUTION VIA THE INVERSION METHOD.
Left the probability density function, the right image illustrates the associated process of generating random variables via the inversion method.

which leads to $C = n + 1$.

Obviously, the associated CDF can be obtained by integrating p_X*, that is,*

$$F_X(x) = \int_{[0,x]} p_X(\xi)\, d\mu(\xi) = x^{n+1}. \tag{6.236}$$

As the inverse function of x^{n+1} *is given by* $\sqrt[n+1]{x}$*, samples* X_i *from the power distribution can be drawn by* $X_i = \sqrt[n+1]{U_i}$*, where* U_i *are uniformly distributed random variables form* $[0, 1]$*, see Figure 6.10.*

EXAMPLE 6.16 (Once more Uniform Sampling on the Interval $[a, b]$) *Compared to the foregoing example, we will now go the other way: Based on the uniformly distributed random variable* U *on* $[0, 1]$*, we want to generate a random variable* X *on* $[a, b]$ *that is distributed according to the probability density function* $p_X = \frac{1}{b-a}$.

It is easy to seen, that the CDF F_X *is given by*

$$F_X(x) \overset{\text{def}}{=} \int_{[a,x]} \frac{1}{b-a}\, d\mu(\xi) = \frac{x-a}{b-a}. \tag{6.237}$$

Obviously, now it holds:

$$u = F_U(u) \overset{(6.233)}{=} F_X(x) = \left(\frac{x-a}{b-a}\right) \tag{6.238}$$

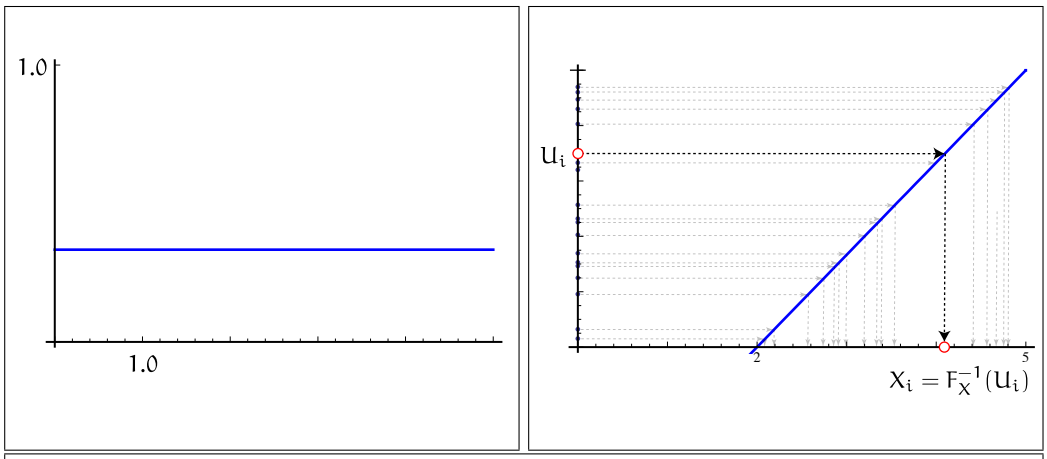


FIGURE 6.11: SAMPLING FROM A UNIFORM DISTRIBUTION VIA THE INVERSION METHOD. Left the probability density function, the right image illustrates the associated process of generating random variables via the inversion method.

leading to

$$X_i = a + U_i(b - a), \tag{6.239}$$

see Figure 6.11.

EXAMPLE 6.17 (Sampling the Attenuation Part of the Particle Transport Equation) *From Relation (4.54) we know, that the component of the particle transport equation which describes the attenuation of light as it travels through a medium of constant opacity is of the form*

$$\int_{[0,d_{\partial\mathcal{V}}(\mathbf{x},\omega)]} \beta(\mathbf{x}, \mathbf{x} - \alpha\omega)\, \mathbf{Q}(\mathbf{x} - \alpha\omega, \omega')\, d\mu(\alpha), \tag{6.240}$$

Path Absorption Function (292)

where $\beta(\mathbf{x}, \mathbf{x} - \alpha\omega)$ is the path absorption function and $\mathbf{Q}(\mathbf{x} - \alpha\omega, \omega')$ represents the amount of light sent toward the viewer at a distance $\mathbf{x} - \alpha\omega$.

$\tau(\mathbf{x}, \mathbf{y})$ (292)

Assuming that the participating medium is of constant opacity, then the optical distance function $\tau(\mathbf{x}, \mathbf{y})$, which occurs in the path absorption function $\beta(\mathbf{x}, \mathbf{x} - \alpha\omega)$, can be reduced to a linear function of the distance $\mathbf{x}$, i.e.,the integral from Equation (6.240) can be written as:

$$\int_{[0,d_{\partial\mathcal{V}}(\mathbf{x},\omega)]} e^{-C(\mathbf{x}-\alpha\omega)}(\mathbf{x}, \mathbf{x} - \alpha\omega)\, \mathbf{Q}(\mathbf{x} - \alpha\omega, \omega', \lambda')\, d\mu(\alpha). \tag{6.241}$$

To evaluate this integral, we first generate random samples α distributed according to the probability density

$$p_\alpha(\alpha) = C\, e^{-C\alpha}, \tag{6.242}$$

where the constant C is required for normalizing p_α. Obviously, the CDF of the random variable α is given by

$$\begin{aligned} F_\alpha(\alpha) &= \int_{[0,\alpha]} Ce^{-C\xi}\, d\mu(\xi) && (6.243)\\ &= -e^{-C\xi}\Big|_0^\alpha && (6.244)\\ &= 1 - e^{-C\alpha}. && (6.245) \end{aligned}$$

Applying the inversion method and using the fact, that with U also $1-U$ is uniformly distributed on $[0,1]$, then we obtain:

$$\begin{aligned} F_\alpha(\alpha) = U &\Rightarrow 1 - e^{-C\alpha} = U && (6.246)\\ &\Rightarrow -C\alpha = \ln(1-U) && (6.247)\\ &\Rightarrow \alpha = -\frac{1}{C}\ln(1-U) && (6.248)\\ &\Rightarrow \alpha = -\frac{1}{C}\ln(U), && (6.249) \end{aligned}$$

see Figure 6.12.

Afterwards, we transform the random variable α to $\mathbf{x} - \alpha\omega$. With N random variables α_i sampled according to the above probability density function, then a Monte Carlo estimator for Equation (6.240) has the form

$$\begin{aligned} F_N &= \frac{1}{N}\sum_{i=1}^{N} \frac{e^{-C(\mathbf{x}-\alpha_i\omega)}\mathbf{Q}(\mathbf{x}-\alpha_i\omega,\omega')}{Ce^{-C(\mathbf{x}-\alpha_i\omega)}} && (6.250)\\ &= \frac{1}{CN}\sum_{i=1}^{N} \mathbf{Q}(\mathbf{x}-\alpha_i\omega,\omega'). && (6.251) \end{aligned}$$

Let us now turn to the multi-dimensional case, where a random vector $\mathbf{X}$ is transformed via a function $\mathbf{T}$ to a random variable $\mathbf{T}(\mathbf{X})$ and we are interested in expressing the density function of $\mathbf{Y}$ in terms of the density of $\mathbf{X}$. Random Vector (183)

THE MULTI-DIMENSIONAL TRANSFORMATION METHOD. In the following $\mathbf{X}$ is assumed to represent an s-dimensional random variable and $\mathbf{Y} = \mathbf{T}(\mathbf{X})$ be the image of the random variable $\mathbf{X}$ under the diffeomorphism $\mathbf{T}$ on $\mathbb{R}^s$. In this case, the following applies to the

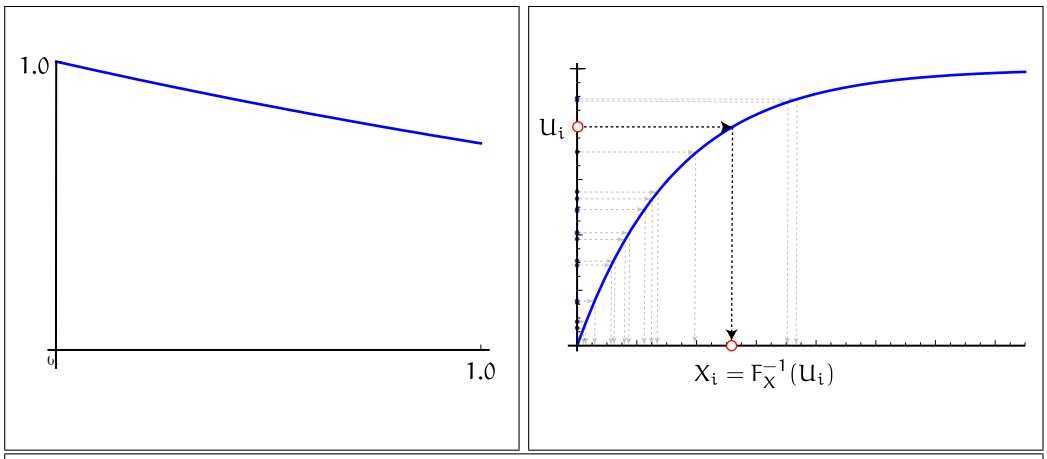


FIGURE 6.12: SAMPLING FROM AN EXPONENTIAL DISTRIBUTION VIA THE INVERSION METHOD. Left the probability density function, the right image illustrates the associated process of generating random variables via the inversion method.

probability measures $\mathbb{P}_{\mathbf{X}}$ and $\mathbb{P}_{\mathbf{Y}}$ induced by $\mathbf{X}$ and $\mathbf{Y}$:

$$\begin{aligned}
\mathbb{P}_{\mathbf{Y}}(\mathbf{Y} \leq \mathbf{y}) &\overset{(2.551)}{=} \left(\mathbb{P} \circ \mathbf{Y}^{-1}\right)(\mathbf{Y} \leq \mathbf{y}) && (6.252)\\
&\overset{\mathbf{Y}=\mathbf{T}(\mathbf{X})}{=} \left(\mathbb{P} \circ (\mathbf{T} \circ \mathbf{X})^{-1}\right)(\mathbf{Y} \leq \mathbf{y}) && (6.253)\\
&= \left(\mathbb{P} \circ \mathbf{X}^{-1} \circ \mathbf{T}^{-1}\right)(\mathbf{Y} \leq \mathbf{y}) && (6.254)\\
&= \left(\mathbb{P} \circ \mathbf{X}^{-1}\right)(\mathbf{T}^{-1}(\mathbf{Y}) \leq \mathbf{T}^{-1}(\mathbf{y})) && (6.255)\\
&= \left(\mathbb{P} \circ \mathbf{X}^{-1}\right)(\mathbf{X} \leq \mathbf{T}^{-1}(\mathbf{y})) && (6.256)\\
&= \mathbb{P}_{\mathbf{X}}(\mathbf{X} \leq \mathbf{T}^{-1}(\mathbf{y})). && (6.257)
\end{aligned}$$

Based on this result, we conclude due to the definition of the cumulative distribution function for continuous random variables that it holds:

CDF (179)

$$\int_{(-\infty,\mathbf{y}]} p_{\mathbf{Y}}(\mathbf{y})\, d\mu(\mathbf{y}) = \int_{(-\infty,\mathbf{T}^{-1}(\mathbf{y})]} p_{\mathbf{X}}(\mathbf{x})\, d\mu(\mathbf{x}) \tag{6.258}$$

Theorem of Transformation (117)

Changing from variable $\mathbf{x}$ to $\mathbf{T}^{-1}(\mathbf{y})$ in the second integral and applying the Theorem of Transformation for s-dimensional integrals yields:

$$\int_{(-\infty,\mathbf{y}]} p_{\mathbf{Y}}(\mathbf{y})\, d\mu(\mathbf{y}) = \int_{(-\infty,\mathbf{y}]} p_{\mathbf{X}}(\mathbf{T}^{-1}(\mathbf{y}))\left|\det\left(J_{\mathbf{T}^{-1}}(\mathbf{y})\right)\right| d\mu(\mathbf{y}). \tag{6.259}$$

Obviously it is possible to express the probability density function $p_{\mathbf{Y}}$ of the random variable $\mathbf{Y}$ in terms of $p_{\mathbf{X}}$, namely by

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(\mathbf{T}^{-1}(\mathbf{y}))\left|\det\left(J_{\mathbf{T}^{-1}}(\mathbf{y})\right)\right|, \tag{6.260}$$

which also can be written as:

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(\mathbf{x}) \frac{1}{\left|\det\left(J_{\mathbf{T}}(\mathbf{x})\right)\right|}. \tag{6.261}$$

In higher dimension, the transformation method can be used to map random samples from one domain to another, thus for example from unit circle to the hemisphere or to choose samples on a plane using polar coordinates. Let us consider some examples which should illustrate how the transformation method works.

EXAMPLE 6.18 (Sampling in Different 2-dimensional Coordinate Systems) *On and off, we are interested in sampling points from a plane according to a probability density function* $p_{R,\Theta}(r,\theta)$ *giving* $p_{X,Y}(x,y)$.

Due to the discussion from above, we need a transformation T*:*

$$T : \mathbb{R}^2 \to [0,1] \times [0,2\pi], \tag{6.262}$$

which maps a point, given in Cartesian coordinates x *and* y*, to a pair of polar coordinates* (r,θ)*. From Figure 6.13 it can easily be deduced, that* T *must have the form*

Polar Coordinates (832)

$$T(x,y) = \begin{pmatrix} T_1(x,y) \\ T_2(x,y) \end{pmatrix} = \begin{pmatrix} r \\ \theta \end{pmatrix} = \begin{pmatrix} \sqrt{x^2+y^2} \\ \arctan\frac{y}{x} \end{pmatrix}. \tag{6.263}$$

Due to Definition A.31, the associated Jacobian matrix $J_T(x,y)$ *is then given by*

$$J_T(x,y) = \begin{pmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{pmatrix}. \tag{6.264}$$

Based on Relation (6.261), the density function $p_{R,\Theta}(r,\theta)$ *can now be written as:*

$$\begin{aligned} p_{R,\Theta}(r,\theta) &= p_{X,Y}(x,y) \frac{1}{\left|\det\left(J_{\mathbf{T}}(X,Y)\right)\right|} && (6.265)\\ &\overset{(A.32)}{=} p_{X,Y}(x,y) \frac{1}{\frac{x^2}{\sqrt{x^2+y^2}\,(x^2+y^2)} + \frac{y^2}{\sqrt{x^2+y^2}\,(x^2+y^2)}} && (6.266)\\ &= p_{X,Y}(x,y)\sqrt{x^2+y^2} && (6.267)\\ &= p_{X,Y}(x,y)\, r. && (6.268) \end{aligned}$$

Replacing the variables x *and* y *in* $p_{X,Y}(x,y)$ *from Equation (6.268) by* $r\cos\theta$ *as well as* $r\sin\theta$*, then we get a probability density function expressed only in the variables* r *and* θ.

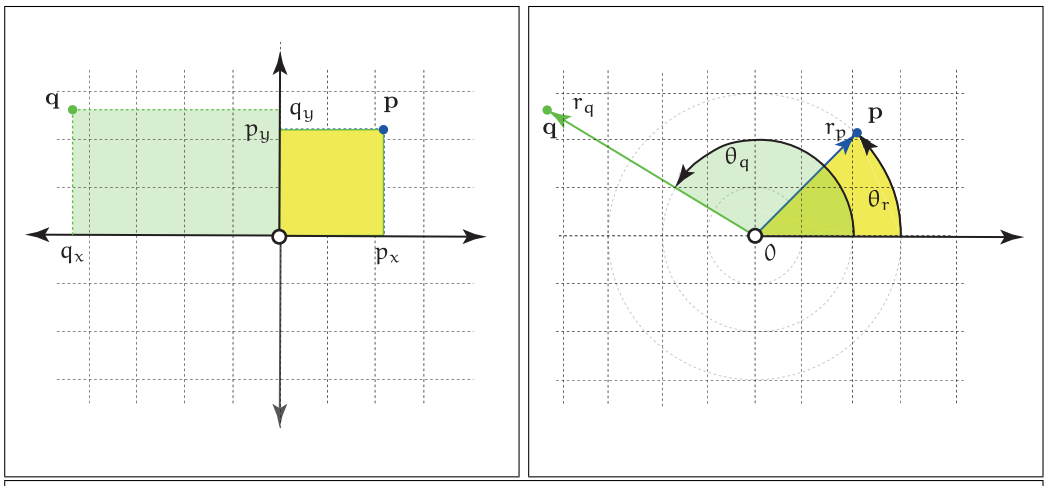


FIGURE 6.13: CONVERTING BETWEEN CARTESIAN AND POLAR COORDINATES. A point given in 2-dimensional Cartesian coordinates (x, y) is transformed via the mapping $T(x, y) = (T_1(x, y), T_2(x, y))^T = (r, \theta)^T = (\sqrt{x^2 + y^2}, \arctan \frac{y}{x})^T$. The inverse of T maps points given in polar coordinates r, θ to points in 2-dimensional Cartesian coordinates $(x = r \cos \theta, y = r \sin \theta)$

From Equation (6.268) we can also deduce, that it is possible to sample from $p_{X,Y}(x, y)$ *given* $p_{R,\Theta}(r, \theta)$*, since it holds:*

$$p_{X,Y}(x, y) = \frac{1}{r} p_{R,\Theta}(r, \theta) = \frac{1}{\sqrt{x^2 + y^2}} p_{R,\Theta}(r, \theta), \tag{6.269}$$

with $r = \sqrt{x^2 + y^2}$ *and* $\theta = \arctan \frac{y}{x}$.

To get a relation between the probability density functions $p_{X,Y}(x, y)$ *and a given PDF* $p_{R,\Theta}(r, \theta)$*, we have to go the reverse way by finding a transformation* T^{-1} *from* $[0, 1] \times [0, 2\pi]$ *to* $\mathbb{R}^2$*. Obviously,* T^{-1} *is defined as*

$$T^{-1}(r, \theta) = \begin{pmatrix} T_1^{-1}(r, \theta) \\ T_2^{-1}(r, \theta) \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \cos \theta \\ r \sin \theta \end{pmatrix} \tag{6.270}$$

with Jacobian matrix

$$J_{T^{-1}}(r, \theta) = \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix}. \tag{6.271}$$

Based on Relation (6.261), the density function $p_{X,Y}(x, y)$ *can now be written*

as:

$$\begin{aligned} p_{X,Y}(x,y) &= p_{R,\Theta}(r,\theta)\,\frac{1}{\left|\det\left(J_{\mathbf{T}}(r,\theta)\right)\right|} && (6.272)\\ &\overset{(A.32)}{=} p_{(R,\Theta)}(r,\theta)\,\frac{1}{r\cos^2\theta + r\sin^2\theta} && (6.273)\\ &= \frac{1}{r}p_{R,\Theta}(r,\theta) && (6.274)\\ &= \frac{1}{\sqrt{x^2+y^2}}p_{(R,\Theta)}(r,\theta). && (6.275) \end{aligned}$$

Finally, let us briefly show how the inversion method can be applied to generate a s-dimensional random vector $\mathbf{X}$ from a given CDF $F_{\mathbf{X}}$.

THE MULTI-DIMENSIONAL INVERSION METHOD. Let $\mathbf{X}$ be a random vector composed of the random variables $(X_1,\ldots,X_s)$. In case where all these random variables are independent, the joint probability density function $p_{\mathbf{X}}$ is given by the marginal densities $p_i(x_i)$ of the random variables X_i, that is,

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &= p_{X_1,\ldots,X_s}(x_1,\ldots,x_s) && (6.276)\\ &= \prod_{i=1}^{s} p_i(x_i). && (6.277) \end{aligned}$$

This then implies, that also the associated CDF ist separable, that is $F_{X_1,\ldots,X_s} = \prod_{i=1}^{s} F_{X_i}(x_i)$. Thus, the component X_i of the random vector $\mathbf{X}$ can be generated by applying the one-dimensional inversion method to each random variable X_i separately, that is,

$$X_i = F_{X_i}^{-1}(U_i) \tag{6.278}$$

for $i = 1,\ldots,n$.

Let us now illustrate the multi-dimensional inversion method with the help of the following simple example.

EXAMPLE 6.19 *Let us consider the 2-dimensional random vector $\mathbf{X} = (X_1, X_2)$ defined on the unit square $[0,1]^2$ with PDFs*

$$p_1(X_1) = x_1 \quad \textit{and} \quad p_2(X_2) = x_2^2, \tag{6.279}$$

thus, the joint probability density function $p_{\mathbf{X}} = p_{X_1,X_2}$ is given by

$$p(\mathbf{X}) = p_{X_1}p_{X_2} = x_1x_2^2. \tag{6.280}$$

Then, the CDF associated with $p_{\mathbf{X}}$ *has the form*

$$F_{X_1,X_2}(x_1,x_2) = \underbrace{\frac{1}{2}x_1^2}_{F_{X_1}(x_1)} \underbrace{\frac{1}{3}x_2^3}_{F_{X_2}(x_2)}. \tag{6.281}$$

Applying the one-dimensional inversion method to the function components $\frac{1}{2}x_1^2$ *and* $\frac{1}{3}x_2^3$ *then leads to:*

$$F_{X_1}(X_1) = U_1 \quad \Rightarrow \quad \frac{1}{2}x_1^2 = U_1 \tag{6.282}$$

$$\Rightarrow \quad X_1 = \sqrt{2U_1} \tag{6.283}$$

and

$$F_{X_2}(X_2) = U_2 \quad \Rightarrow \quad \frac{1}{3}x_2^3 = U_2 \tag{6.284}$$

$$\Rightarrow \quad X_2 = \sqrt[3]{3U_2}. \tag{6.285}$$

With the help of the following examples, we will now illustrate the manner in which the multi-dimensional transformation method may be applied to sample points on one of the hemispheres around a point at a surface. This method is an often used technique in ray tracing procedures for generating rays starting at a surface point.

Chapter 9

Reflectance Equation (321)

EXAMPLE 6.20 (cosine-weighted Hemisphere Sampling) *The reflectance equation*

$$L_o(\mathbf{s},\theta_o,\phi_o) \tag{6.286}$$
$$= \int_{[0,2\pi)}\int_{[0,\pi]} f_r(\mathbf{s},(\theta_i,\phi_i) \longrightarrow (\theta_o,\phi_o))\, L_i(\mathbf{s},\theta_i,\phi_i)\sin\theta_i\,|\cos\theta_i|\,d\mu(\theta_i)d\mu(\phi_i),$$

represented in spherical coordinates, has a cosine-term beneath the integral. So, if it is possible to sample a direction according to a cosine-weighted probability density function, a corresponding Monte Carlo estimator can easily be expressed as the sum of the radiance measured at the sampled points multiplied by the BRDF. That is, a clever choice of the PDF eliminates the cosine-term in Equation (6.286) and thus, reduces the computational costs. Let

BRDF (320)

$$p_{\Theta_i,\Phi_i}(\theta_i,\phi_i) = \frac{\sin\theta_i\,|\cos\theta_i|}{\pi} \tag{6.287}$$

be the PDF to be sampled from then we obtain the corresponding CDF by means of integration by parts

$$F_{\Theta_i,\Phi_i}(\theta_i,\phi_i) = \frac{1}{\pi}\int_{[0,\phi_i]}\int_{[0,\theta_i]} \sin\theta\,|\cos\theta|\,d\mu(\theta)d\mu(\phi) \tag{6.288}$$

$$= \frac{\phi_i}{\pi}\int_{[0,\theta_i]} \sin\theta\,|\cos\theta|\,d\mu(\theta) \tag{6.289}$$

$$= \frac{\phi_i}{2\pi}(1-\cos^2\theta_i). \tag{6.290}$$

Obviously, F_{Θ_i,Φ_i} is separable, that is, it can be splitted in two independent functions

$$F_{\Theta_i} = 1 - \cos^2 \theta_i \tag{6.291}$$

$$F_{\Phi_i} = \frac{\phi_i}{2\pi}. \tag{6.292}$$

Applying the one-dimensional inversion method to each single function provides us with

$$F_{\Theta_i}(\theta_i) = U_1 \Rightarrow 1 - \cos^2 \Theta_i = U_1 \tag{6.293}$$

$$\Rightarrow \Theta_i = \cos^{-1} \sqrt{1 - U_1} \tag{6.294}$$

and

$$F_{\Phi_i}(\phi_i) = U_2 \Rightarrow \frac{\Phi_i}{2\pi} = U_2 \tag{6.295}$$

$$\Rightarrow \Phi_i = 2\pi U_2, \tag{6.296}$$

see Figure 6.14.

Using the fact, that U_1, U_2 are independent and uniformly distributed random variables drawn from $[0,1]$, Equation (6.294) can also be written as:

$$\Theta_i = \cos^{-1} \sqrt{U_1}. \tag{6.297}$$

Thus, the random variables Θ_i and Φ_i are distributed according to the cosine weighted PDF from (6.287).

Using cosine-weighted hemisphere samples Θ_i, Φ_i, a secondary Monte Carlo estimator for the reflectance equation is then given by

$$F_N = \frac{\pi}{N} \sum_{i=1}^{N} f_r(\mathbf{s}, (\Theta_i, \Phi_i) \longrightarrow (\theta_o, \phi_o))\, L_i(\mathbf{s}, (\Theta_i, \Phi_i)). \tag{6.298}$$

EXAMPLE 6.21 (Uniform Sampling on the Hemisphere) *A possibility for uniform sampling on the hemisphere is to sample a direction ω_i uniformly according to solid angle, see Figure 6.15.*

From our discussions in Chapter 2 we know that the solid angle of a set of directions corresponds to the area of a point set on the unit sphere. Let us now consider the area of a cap of the hemisphere, defined by a set of directions, where the polar angle of direction ω_i lies between zero and a fixed angle θ_i. For the measure

Solid Angle (83)

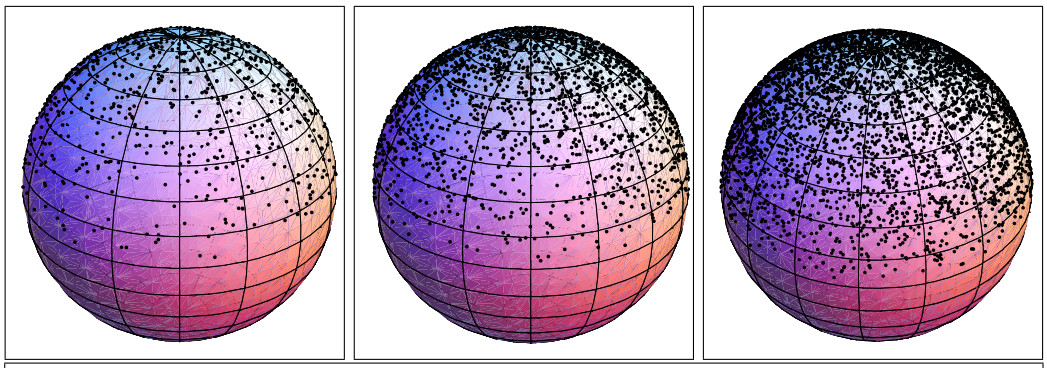


FIGURE 6.14: COSINE-WEIGHTED HEMISPHERE SAMPLING. Left, the hemisphere is sampled with 1000 points, in the center, the hemisphere is sampled with 2500 points, and the hemisphere on the right is sampled with 5000 points.

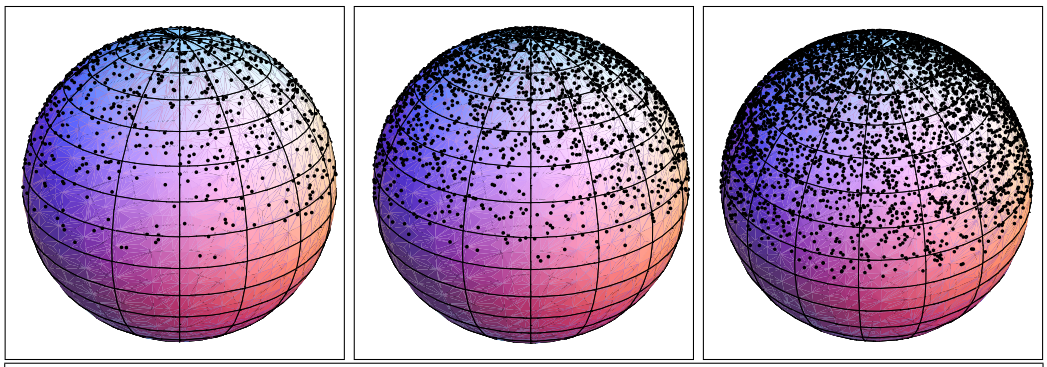


FIGURE 6.15: UNIFORM SAMPLING THE HEMISPHERE WITH RESPECT TO SOLID ANGLE. Left, the hemisphere is sampled with 1000 points, in the center, the hemisphere is sampled with 2500 points, and the hemisphere on the right, is sampled with 5000 points.

of this area, thus the restriction $\mathcal{H}^2_+\big|_{[0,2\pi)\times[0,\theta_i]}$*, holds:*

$$\begin{aligned}
\mu^2\left(\mathcal{H}^2_+\big|_{[0,2\pi)\times[0,\theta_i]}\right) &= \int_{[0,2\pi)}\int_{[0,\theta_i]} \sin(\theta)\, d\mu^2(\theta,\phi) && (6.299)\\
&= \left(\int_{[0,2\pi)} d\mu(\phi)\right)\left(\int_{[0,\theta_i]} \sin(\theta)\, d\mu(\theta)\right) && (6.300)\\
&= 2\pi\cdot(-\cos\theta)\Big|_0^{\theta_i} && (6.301)\\
&= 2\pi\cdot(1-\cos\theta_i). && (6.302)
\end{aligned}$$

Since we are interested in sampling a direction ω_i *uniformly with respect to solid angle, the probability of sampling an angle* θ_i *within the cap should be proportional to the area of the cap, thus* $\mu^2\left(\mathcal{H}^2_+\big|_{[0,2\pi)\times[0,\theta_i]}\right)$*. Then the desired cumulative distribution function for* Θ_i *can be obtained by normalizing, that is,*

$$\begin{aligned}
F_{\Theta_i}(\theta_i) &\stackrel{def}{=} \frac{\mu^2\left(\mathcal{H}^2_+\big|_{[0,2\pi)\times[0,\theta_i]}\right)}{\mu^2\left(\mathcal{H}^2_+\right)} && (6.303)\\
&= \frac{2\pi\cdot(1-\cos\theta_i)}{2\pi} && (6.304)\\
&= 1-\cos\theta_i. && (6.305)
\end{aligned}$$

Using the relation $z=\cos(\theta_i)$*, then the CDF* F_{Θ_i} *can be formulated as a function of the coordinate* z*, namely*

$$F_Z(z) = 1-z. \tag{6.306}$$

Relation (6.306) then implies, that the area of the hemisphere is uniformly distributed with respect to Z*, or in other words, any two horizontal slices with equal vertical thickness have the same surface area. Applying the inversion method on two uniform random variables* U_1 *and* U_2 *provides us*

$$\begin{aligned}
F_Z(z) = U_1 &\Rightarrow 1-Z = U_1 && (6.307)\\
&\Rightarrow Z = 1-U_1 && (6.308)\\
&\Rightarrow Z = U_1 && (6.309)
\end{aligned}$$

and with $F_{\Phi_i}(\phi_i)$ *from Example 6.20*

$$\begin{aligned}
F_{\Phi_i}(\phi_i) = U_2 &\Rightarrow \frac{\Phi_i}{2\pi} = U_2 && (6.310)\\
&\Rightarrow \Phi_i = 2\pi U_2. && (6.311)
\end{aligned}$$

We can then generate a uniform distribution of points (x, y, z) *in polar coordinate representation on the hemisphere with respect to solid angle choosing:*

$$X = \sin(\Theta_i)\cos(\Phi_i) = \sqrt{1-U_1^2}\cos(2\pi U_2), \tag{6.312}$$

$$Y = \sin(\Theta_i)\sin(\Phi_i) = \sqrt{1-U_1^2}\sin(2\pi U_2), \tag{6.313}$$

$$Z = \cos(\Theta_i) = U_1. \tag{6.314}$$

In order to complete the concept of the multi-dimensional inversion method we have still to consider the case where the random variables, from whose distributions we wish to sample, are dependent. For that purpose, let $\mathbf{X}$ be a random vector composed of dependent random variables $(X_1, \ldots, X_s)$. The joint probability density function $p_{\mathbf{X}}$ is then given by the marginal density p_{X_1} and the conditional PDFs, $p_{X_i|X_1 \ldots X_{i-1}}$, of the dependent random variables X_i for $2 \leq i \leq s$.

As the joint probability density function $p_{\mathbf{X}}$ is given by

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1,\ldots,X_s}(x_1, \ldots, x_s) \tag{6.315}$$

$$= p_1(x_1)\prod_{i=2}^{s} p_{X_i|X_1 \ldots X_{i-1}}(x_i|x_1 \ldots x_{i-1}), \tag{6.316}$$

the vector $\mathbf{X} = (X_1, \ldots, X_s)$, which is obtained from the solution of the following system of equations

$$\begin{cases} F_{X_1}(x_1) = U_1 \\ F_{X_2|X_1}(x_2|x_1) = U_2 \\ \vdots \\ F_{X_s|X_1 \ldots X_s}(x_s|x_1 \ldots x_s) = U_s, \end{cases} \tag{6.317}$$

is distributed according to the cumulative distribution function $F_{\mathbf{X}}$.

Thus, for generating a random vector, distributed according to the joint probability density $p_{\mathbf{X}}$ from Equation (6.316), first, we have to generate s uniformly independent distributed random variables U_i from $[0, 1]$ and then we have to solve the system of equations from Relation (6.317) with respect to $p_{\mathbf{X}} = (X_1, \ldots, X_s)$.

REMARK 6.9 *In [172, Rubinstein 1981] it is shown, that the efficiency of a simulation is dependent on the order in which random variables* $X_i, 1 \leq i \leq n$ *are taken while forming the random vector* $\mathbf{X}$*. Since there are* $s!$ *possibilities to represent the components* $X_1, \ldots, X_s$ *of* $p_{\mathbf{X}}$*, there are also* $s!$ *possibilities to generate the random vector* $\mathbf{X}$ *when solving the above system of equations. But a priori, there is no way to find the optimal order of components of the random vector* $\mathbf{X}$ *to minimize the run time.*

THE MULTI-DIMENSIONAL INVERSION METHOD, THE CASE $s = 2$. The idea behind the multi-dimensional inversion method is to isolate one particular variable via computing

the marginal density function and to use this density function for computing the required conditional density functions in the system of equations from Relation (6.317). As this technique can often be applied in graphics for sampling from 2-dimensional distributions, let us consider the inversion method for the case $s = 2$ a little bit more in detail.

Assuming, we have to sample from a probability density function $p_{X,Y}(x, y)$. The marginal probability density function $p_X(x)$ was defined as

Marginal Density Function (194)

$$p_X(x) = \int_\Omega p_{X,Y}(x, y) d\mu(y), \tag{6.318}$$

which means, that $p_X(x)$ is the PDF for the random variable X. The conditional PDF $p_{Y|X}(y|x)$ was given by

Conditional Density Function (209)

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}, \tag{6.319}$$

thus, the probability density function for the random variable Y, where we have fixed a particular value of x. Afterwards, we have to integrate the two PDF's, that is, we have to compute

$$F_X(x) \overset{def}{=} \int_\Omega p_X(x)\, d\mu(x) \tag{6.320}$$

$$F_{Y|X}(y|X) \overset{def}{=} \int_\Omega p_{Y|X}(y|x)\, d\mu(y) \tag{6.321}$$

Applying the inversion method to the uniformly distributed random variables U_1 and U_2 then leads to

$$F_X(x) = U_1 \quad \Rightarrow \quad X = F_X^{-1}(U_1) \tag{6.322}$$

$$F_{Y|X}(y|x) = U_2 \quad \Rightarrow \quad Y = F_{Y|X}^{-1}(U_2). \tag{6.323}$$

We now demonstrate this technique with the help of an interesting example useful for our further discussion in Chapter 9.

EXAMPLE 6.22 (Uniform Disk Sampling) *It is well known that contrary to a pinhole camera, which creates images where everything is in perfect focus, a thin lens camera model makes images with depth-of-field effects. In such a camera the pinhole is replaced with a disk-shaped thin lens, which has certain idealized behavior. To simulate depth-of-field effects we have to integrate over all rays passing through the area of the lens. For that purpose, we need a method for generating uniformly distributed samples in the unit circle, which then can be transformed on the camera lens, see Figure 6.16.*

Pinhole Camera (417)
Thin Lens Camera (686)

Representing the unit circle in porlar coordinates, then according to Relation (2.700) a PDF for sampling uniformly on the unit disk is given by

$$p_{R,\Theta}(r, \theta) = \frac{1}{\pi} r, \tag{6.324}$$

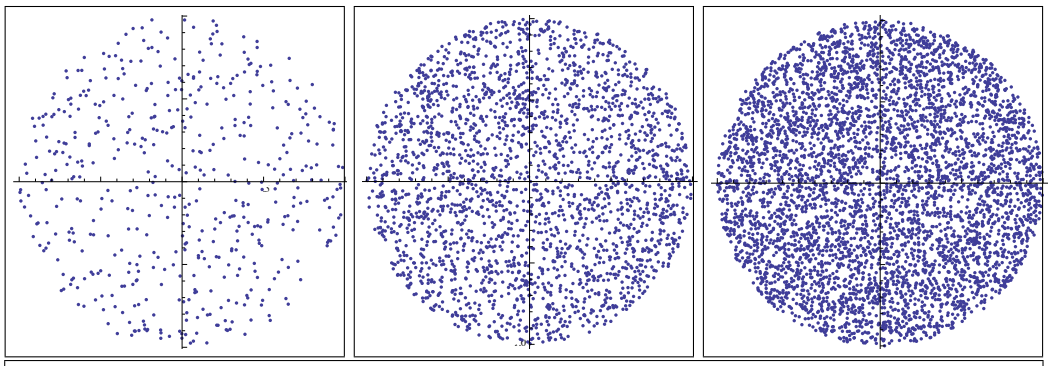


FIGURE 6.16: UNIFORM DISK SAMPLING. Left, the disk is sampled with 1000 points, in the center, the disk is sampled with 2500 points, and the disk on the right is sampled with 5000 points.

where $r \in [0,1]$ *and* $\theta \in [0, 2\pi)$.

Marginal Density Function (194) *In Example 2.78 we have already computed the marginal density* p_R*, it holds*

$$p_R(r) = \frac{1}{\pi} \int_{[0,2\pi)} r \, d\mu(\theta) = \frac{r}{\pi}\theta \Big|_0^{2\pi} = 2r. \tag{6.325}$$

Conditional PDF (209) *Now, the conditional density of* θ *can be computed via*

$$p_{\Theta|R}(\theta|r) = \frac{p_{R,\Theta}(r,\theta)}{p_R(r)} = \frac{r}{2r\pi} = \frac{1}{2\pi}. \tag{6.326}$$

Integrating $p_R(r)$ *over* $[0, R]$*, then leads to the CDF*

$$F_R(R) = 2 \int_{[0,R]} \xi \, d\mu(\xi) = 2 \frac{1}{2} \xi^2 \Big|_0^R = R^2 \tag{6.327}$$

CDF (179) *and the conditional cumulative distribution function of the random variable* Θ *given* R *has the form*

$$F_{\Theta|R}(\theta|r) = \frac{1}{2\pi} \int_{[0,\theta]} d\mu(\theta) = \frac{\theta}{2\pi}. \tag{6.328}$$

Applying the one-dimensional inversion method provides us with

$$F_R(R) = U_1 \quad \Rightarrow \quad R^2 = U_1 \tag{6.329}$$

$$\Rightarrow \quad R = \sqrt{U_1} \tag{6.330}$$

and

$$F_{\Theta|R}(\theta) = U_2 \quad \Rightarrow \quad \frac{\theta}{2\pi} = U_2 \tag{6.331}$$

$$\Rightarrow \quad \theta = 2\pi U_2 \tag{6.332}$$

REMARK 6.10 *In the above example it would also have been possible to write the PDF in the form*

$$p_{R,\Theta}(r,\theta) = \frac{1}{2\pi} 2r, \tag{6.333}$$

and to interpret the factor $\frac{1}{2\pi}$ *as the marginal density of the random variable* Θ*. The conditional probability density function* $p_{R|\Theta}$ *is then given by* $2r$*. This results in the same formulas for uniform sampling a disk as we them get from the example.*

6.5.2 ACCEPTANCE-REJECTION SAMPLING

In the last section, we presented with the transformation method, the most frequently used sampling procedure in the theory of Monte Carlo algorithms, where samples are generated via the CDF of a random variable. Now, in cases, where it is not possible to derive a formula for the cumulative distribution function of a random variable, the last resort for independent sampling is: *Acceptance-rejection Sampling*, [224, von Neumann 1951].

CDF (171)

Acceptance-rejection sampling, also known in the literature under the name of the *rejection method*, or the *hit-miss method*, is based on a Bernoulli experiment for simulating random variables which are distributed according to some arbitrary density function p. Instead to sample directly from the desired density, which is commonly difficult or even impossible, an easily to sample density function q is used. The method only requires to know the shape of p up to a multiplicative constant, no further information or deep analytical study of the density p is necessary.

PDF (176)

So, acceptance-rejection sampling is based on the idea of a proposal value and subjects this value to a special kind of test, where it may be either accepted or rejected. If the value is rejected another sample must be drawn and tested until an acceptable sample is drawn, see Figure 6.17. Closer scrutiny shows that acceptance-rejection sampling can be formulated as follows:

Instead of sampling from any given density p one uses a convenient density q as the upper boundary of p, that is, one chooses a density q with $p(\mathbf{x}) \leq Mq(\mathbf{x})$, $\forall\, \mathbf{x} \in \mathbf{Q}^s$, see Figure 6.18. Usually, M is greater than 1 and often q is chosen as the uniform density on $\mathbf{Q}^s$. Afterwards one generates, by means of uniformly distributed random variables $U_i \in [0,1]$, samples $\mathbf{X}_i$ according to the density q from $\mathbf{Q}^s$ until the following holds:

$$U_i \leq \frac{p(\mathbf{X}_i)}{M\, q(\mathbf{X}_i)}. \tag{6.334}$$

Let us show by means of the two following examples how acceptance-rejection sampling can be applied to sample from a one and a two-dimensional function.

```
ACCEPTANCE-REJECTION SAMPLING {
  for i = 1 to ∞ do {
    Sample X_i according to q
    Sample U_i uniformly on [0,1]
    if U_i ≤ p(X_i) / (M q(X_i)) {
      accept Y_i = X_i
      return Y_i
    }
  }
}
```

FIGURE 6.17: PSEUDOCODE FOR ACCEPTANCE-REJECTION SAMPLING.

EXAMPLE 6.23 *Given be the function* $p(x) = 3x^2$. *To generate samples from* p, *we use the uniform density* $q = 1$ *on* $[0,1]$ *with* $M = 3$ *and generate the random variable* X_i. *Due to* $p_u = 1$ *we then generate a uniformly distributed random variable* U_i *according to the density* $p_U = 1$ *from* $[0,1]$. *The sample* $Y_i = X_i$ *is accepted if* $U_i \leq \frac{3X_i^2}{3}$.

To generate samples from $p(x) = \frac{2}{\pi R^2}\sqrt{R^2 - x^2}$ *on* $-R \leq x \leq R$ *let us assume* $M = \frac{2}{\pi R}$. *We generate a sample* X_i *due to the density* $q = 1$ *with* $X_i = -R + 2RU_i = R(2U_i - 1)$ *and a uniformly distributed random variable* U_i *on* $[0,1]$. *The sample* $Y_i = X_i$ *is accepted if* $U_i \leq \frac{\frac{2}{\pi R^2}\sqrt{R^2 - X_i^2}}{M} = \sqrt{R^2 - X_i^2}$.

LEMMA 6.4 (Acceptance-rejection Sampling) *Let* p *be a density on* $\mathbf{Q}^s$, *acceptance-rejection sampling produces samples from a random variable* $\mathbf{Y}$ *that is distributed according to the density* p.

PROOF 6.4 *The distribution of* $\mathbf{Y}$ *is given by*

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{y}) = \mathbb{P}_{\mathbf{X}|U}\left(\mathbf{X} \leq \mathbf{y} | U \leq \frac{p(\mathbf{X})}{Mq(\mathbf{X})}\right) \tag{6.335}$$

$$= \frac{\mathbb{P}_{\mathbf{X}|U}\left(\mathbf{X} \leq \mathbf{y}, U \leq \frac{p(\mathbf{X})}{Mq(\mathbf{X})}\right)}{\mathbb{P}_U\left(U \leq \frac{p(\mathbf{X})}{Mq(\mathbf{X})}\right)}. \tag{6.336}$$

Expressing the probabilities $P_{\mathbf{X}|\mathbf{U}}$ *and* $P_{\mathbf{U}}$ *via the associated conditional and marginal densities then we get:*

Marginal Density (194)

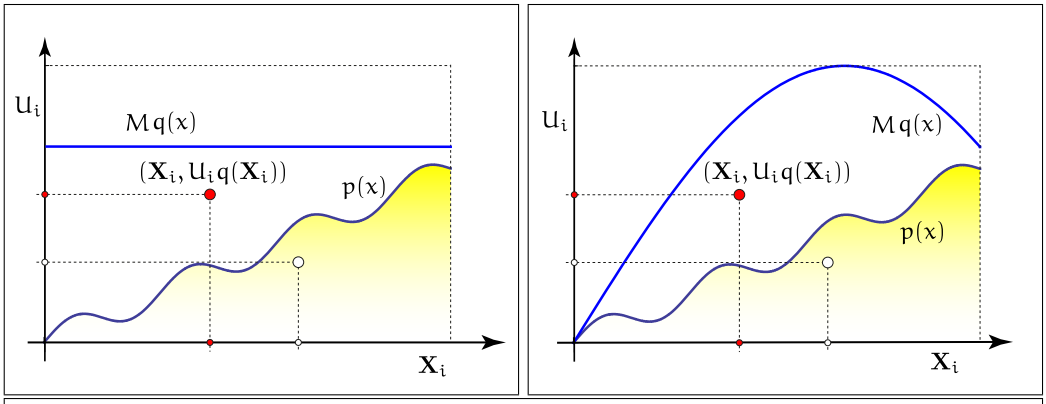


FIGURE 6.18: ILLUSTRATION OF ACCEPTANCE-REJECTION SAMPLING. At first, the algorithm generates a pair of random variables $(\mathbf{X}_i, U_i)$. Then, it tests whether the point $(\mathbf{X}_i, U_i\, Mq\mathbf{X}_i)$ lies under the curve $p(\mathbf{x})$. If this test is successful, open white circle, the sample $\mathbf{X}_i$ is accepted, otherwise, open red circle, a new pair of random variables $(\mathbf{X}_i, U_i)$ is generated and the test is repeated.

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{y}) \overset{(2.810)}{=} \frac{\int_{(-\infty^s, \mathbf{y}]} \left(\int_{[0, \frac{p(\mathbf{x})}{Mq(\mathbf{x})}]} d\mu(u) \right) q(\mathbf{x}) d\mu^s(\mathbf{x})}{\int_{(-\infty^s, -\infty^s]} \left(\int_{[0, \frac{p(\mathbf{x})}{Mq(\mathbf{x})}]} d\mu(u) \right) q(\mathbf{x}) d\mu^s(\mathbf{x})} \tag{6.337}$$

$$= \frac{\frac{1}{M} \int_{(-\infty^s, \mathbf{y}]} p(\mathbf{x}) d\mu^s(\mathbf{x})}{\frac{1}{M} \int_{(-\infty^s, -\infty^s]} p(\mathbf{x}) d\mu^s(\mathbf{x})} \tag{6.338}$$

$$= \int_{(-\infty^s, \mathbf{y}]} p(\mathbf{x}) d\mu^s(\mathbf{x}), \tag{6.339}$$

that is, the random variable is distributed according to the PDF p.

Obviously, acceptance-rejection sampling raises the dimension of the sampling procedure by one, since $s+1$-dimensional random variables must be drawn over $\mathbf{Q}^s \times [0, \{Mq(\mathbf{x})\}]$, $\mathbf{x} \in \mathbf{Q}^s$.

In the special case, where the probability density function p, defined on the domain $[a, b]$, satisfies the condition $p(x) \leq M$ for all $x \in [a, b]$, the function q can be chosen as 1, and the random variables $\mathbf{X}_i$ must be drawn from $[a, b] \times [0, M]$. The acceptance probability for $\mathbf{X}_i$ then becomes

$$\frac{p(\mathbf{X}_i)}{M\, q(\mathbf{X}_i)} = \frac{p(\mathbf{X}_i)}{M}. \tag{6.340}$$

Graphically, this can be interpreted as the ratio of the height of the curve p at the point $\mathbf{X}_i$ and the constant M, see the left image in Figure 6.18.

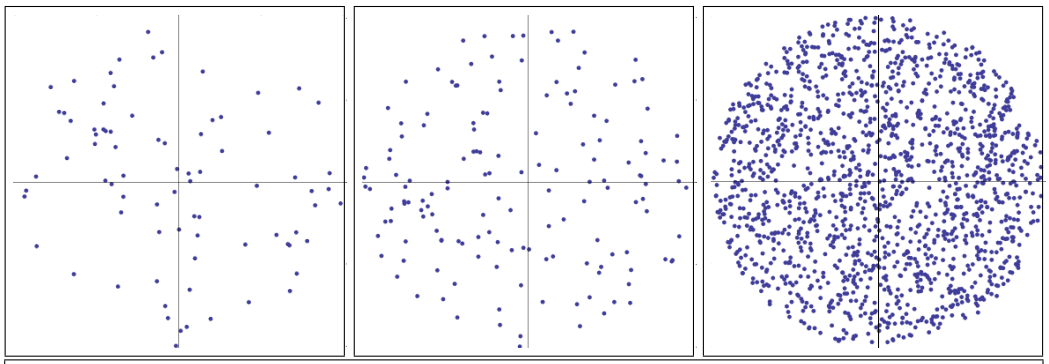


FIGURE 6.19: ACCEPTANCE-REJECTION SAMPLING. Approximating the surface area of the unit circle via acceptance-rejection sampling based on point sets of the sizes 100, 200 und 2000. Left, we have $\mu^2(o) = 0.981748\,\pi$, the image in the middle leads to the value $\mu^2(o) = 1.01342\,\pi$ and the right-hand image returns the value $\mu^2(o) = 1.00127\,\pi$.

Since it can be used with any density function, even those that cannot be integrated analytically, acceptance-rejection sampling can be seen as the *last resort independent sampling techniques*, which can be applied if all other sampling strategies fail. As the efficiency of acceptance-rejection sampling is strongly dependent on the choice of constant M, the form of the sampled density q should be adapted to the density to be sampled as much as possible. Naively applied, the rejection method is found to be not very effective, as the stratification of the integral domain, which can prevent sample-clumping and lead to a reduction of variance, is applied only with difficulty in the procedure.

Stratified Sampling (571)

Section 6.19

Acceptance-rejection sampling can also be applied to geometric based problems that do not correspond exactly to the machinery described above, see Figure 6.18. For it, let us consider a few examples from field of global illumination.

EXAMPLE 6.24 (Generating Cosine-weighted Rays on the Hemisphere by Acceptance-Rejection Sampling) *A typical application of acceptance-rejection sampling in field of global illumination is the generation of a ray $\mathbf{r}(\mathbf{s}, \omega_o)$ outgoing from point $\mathbf{s}$ on a surface patch in direction ω_o. To implement such a generation of rays in a rendering algorithm, first we sample points via the acceptance-rejection method within the unit circle, project these points orthogonal on the hemisphere and generate, starting at $\mathbf{s}$, rays through these points on the hemisphere.*

For that purpose, we generate 2-dimensional random variables $\mathbf{U}_i = (U_{i1}, U_{i2})$, uniformly distributed on $[0,1]^2$, and accept the sample $\mathbf{U}_i$ if it lies within the unit circle, i.e. if it holds $\|\mathbf{U}_i\| = \sqrt{U_{i1}^2 + U_{i2}^2} \leq 1$. Afterwards, we determine the function value $f(U_{i1}, U_{i2}) = \sqrt{1 - U_{i1}^2 - U_{i2}^2}$. It is obviously, that rays starting in $\mathbf{s}$ and passing trough the point $(U_{i1}, U_{i2}, f(U_{i1}, U_{i2}))$ are distributed on the hemisphere, see Figure

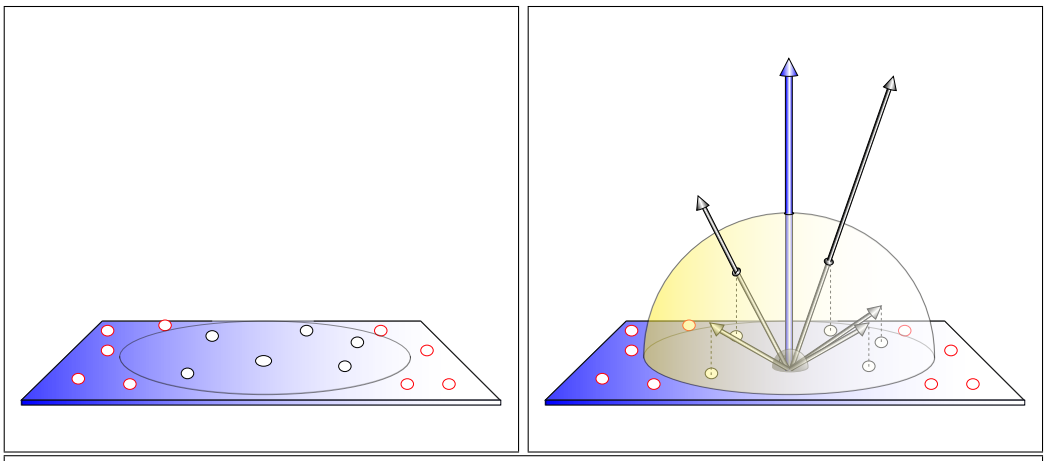


FIGURE 6.20: GENERATING RAYS OVER THE HEMISPHERE VIA ACCEPTANCE-REJECTION SAMPLING. A sample generated via acceptance-rejection sampling is accepted if it lies within the unit circle, otherwise it is rejected. Then, accepted samples are projected orthogonal on the hemisphere. Uniform sampling on the unit circle leads to a distribution of rays starting in the midpoint of the surrounding hemisphere and passing through the generated points on the hemisphere. This procedure is also known as *Malley's method*.

6.20.

EXAMPLE 6.25 (Poisson-disk Sampling) *Let* $\mathbf{I}^s$ *be the* s*-dimensional unit cube,* Poisson-disk sampling *pursuits the strategy to generate a set* $\mathbf{P}$ *of randomly chosen points* $\mathbf{p}_i, \mathbf{p}_j \in \mathbf{I}^s$ *which satisfy the* Poisson-disk criterion*: No two samples are closer together than some distance* d*. Mathematically, the Poisson-disk criterion can be formulated as:*

Poison-disk Sampling (648)

$$\forall\, \mathbf{p}_i, \mathbf{p}_j \in \mathbf{P} \quad \Rightarrow \quad \Delta(\mathbf{p}_i, \mathbf{p}_j) \geq \mathrm{d}, \tag{6.341}$$

with $\mathrm{d} > 0$*. Thus, the intersection of each two* s*-dimensional spheres with the centers* $\mathbf{p}_i$ *and* $\mathbf{p}_j$ *and a radius* d *is disjoint with respect to* $\mathbf{P}$*.*

A rudimentary procedure for constructing a Poisson-disk pattern $\mathbf{P}$*—the so-called* dart-throwing method*—is based on acceptance-rejection sampling. Here, a sample* $\mathbf{p}_i$ *randomly drawn from* $\mathbf{I}^s$ *is added to the set* $\mathbf{P}$ *if, and only if, it satisfies Relation (6.341) to all points* $\mathbf{p}_j$ *already contained in* $\mathbf{P}$*. Otherwise* $\mathbf{p}_i$ *is rejected.*

According to [67, Glassner 1995], for reasons of efficiency, the above described algorithm is recommended only for the construction of relatively small Poisson-disk patterns. It also has the disadvantage that so-called holes, i.e. sample-free areas may exist in $\mathbf{I}^s$*. In addition, because of the required distance which has to be maintained between points, difficulties may be encountered while generating patterns with*

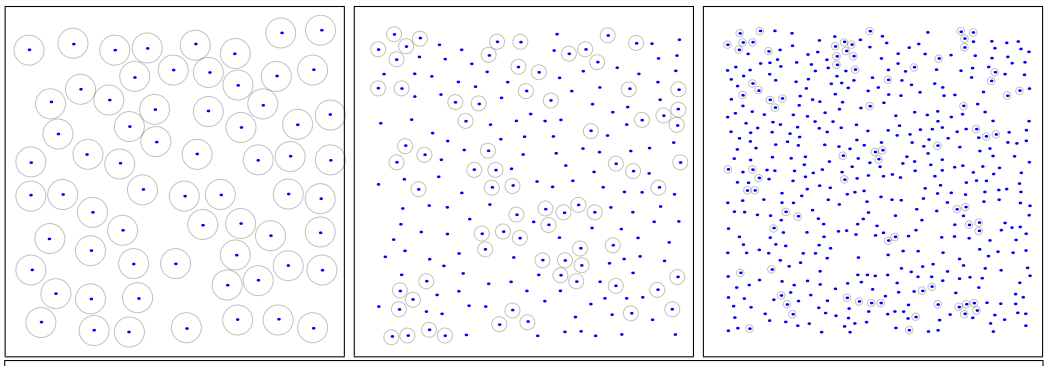


FIGURE 6.21: POISSON-DISK SAMPLING. Three Poisson-disk pattern generated via acceptance-rejection sampling. In the pattern on the left-hand side you can see a set of $N = 62$ points, where each point is surrounded by the circle with radius $d = 0.1$. The pattern in the center is a Poisson-disk pattern with $N = 211$ elements, also in this pattern some of the points are surrounded by circles with radius $d = 0.05$, indicating that they satisfy the Poisson-disk criterion. Right, a Poisson-disk pattern with $N = 473$ points and $d = 0.005$.

a predetermined number of samples. Nevertheless, Poisson-disk sampling has been found to be one of the most popular sampling techniques for generating point sets—especially in pixel sampling as the undesirable effect of aliasing is replaced by white noise, much more pleasant to the observer [104, Keller 1998], [67, Glassner 1995], [38, Cook 1986]. Figure 6.21 shows three Poisson-disk patterns generated due to the dart-throwing method using 1000 *samples, with the distances* $d \in \{0.1, 0.05, 0.025\}$.

For the above given reasons, the dart-throwing method [40, Cook 1984] described here for generating Poisson-disk patterns is thus hardly used anymore in general practice today. It has largely been replaced by two similar procedures, best-candidate- *and decreasing radius algorithm [67, Glassner 1995].*

EXAMPLE 6.26 (Poisson-disk-Hemisphere-Sampling) *Figure 6.22 illustrates a method for generating directions over the hemisphere by combining Poisson-disk sampling and acceptance-rejection sampling. A new point is added to the existing set if it is accepted, that is, if it lies within the unit circle around a pre-given center,* $\mathbf{s}$, *and if it satisfies the Poisson-disk criterion. Surface points are then projected onto the upper hemisphere* $\mathcal{H}^2_+(\mathbf{s})$, *where directions starting in* $\mathbf{s}$ *and passing through the projected points can be generated.*

6.5.3 MCMC - MARKOV CHAIN MONTE CARLO

All sampling techniques discussed until now produce independent samples from a given probability distribution. Now, one problem that comes with Monte Carlo integration is:

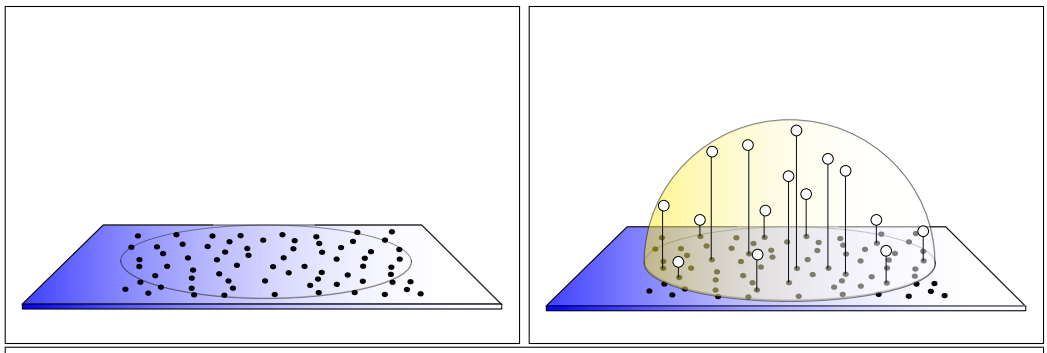


FIGURE 6.22: POISSON-DISK-HEMISPHERE-SAMPLING. Hemisphere sampling based on Poisson-disk Pattern of the size 64.

Generating independent samples from some complex probability distribution. As we have seen in the last section, often, the only sampling technique that remains is acceptance-rejection sampling. Now, the main drawback of the acceptance-rejection method is that it is often very difficult to construct a suitable proposal distribution that leads to an efficient algorithm. One way to avoid this problem is to allow the proposed value depend on the last accepted value, which makes it easier to generate a suitable, but now conditional proposal. The price, we pay for that, is to generate samples from a sequence $(\mathbf{X}_n)_{n\in\mathbb{N}}$ of dependent random variables instead of a sequence of independent random variables. Such for example, the Metropolis algorithm, which we will introduce in Section 6.5.3.2, generates, instead of independent, correlated variables from a discrete-time Markov process.

Section 6.5.2

Random Variable (168)

DT Markov Process (236)

Now, from the SLLN we know that the expected value of a random variable can be computed by averaging the outcomes of a sufficient large number of random experiments, whose associated random variables are all independent and identically distributed. Unfortunately, a stochastic process simulated by a discrete-time, continuous-space Markov chain represents a sequence of random variables that are not independent and identically distributed. That is, the SLLN can not be used to make statements on the convergence of dependent random variables. This means, that we need a new machinery of theorems and statements that characterizes the convergence of Markov chains and Markov processes.

SLLN (216)

IID Random Variable (499)

The present section can be considered as a quick overview of the mathematical foundations of Markov chain Monte Carlo that are required to understand the convergence of MCMC methods. As the whole field around the concept of the Markov process is often very difficult, we explain the mathematical foundations of MCMC methods mainly based on discrete-time, discrete-state Markov chains. Discrete-time, continuous-state Markov processes have analogous dependencies, but that are of a more technical nature and more difficult to capture. A good account of a generalization of all these concepts is given in [65, Gilks & al. 1996], full details appear in [130, Meyn & Tweedie 1993] or [170, Robert

Section 6.5.3.1

DT Markov Process (236)

Section 6.5.3.2

& Casella 1999]. Afterwards, we then introduce the *Metropolis algorithm*, due to the January-February 2000 issue of Computing in Science & Engineering—a joint publication of the American Institute of Physics and the IEEE Computer Society—awarded as one of the *Top Ten Algorithms of the Century, with the greatest influence on the development and practice of science and engineering in the* 20^{th} *century*.

6.5.3.1 MATHEMATICAL FOUNDATIONS OF MARKOV CHAIN MONTE CARLO

Importance sampling, introduced in Section 6.6.2, has shown that for evaluating the integral

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) = \int_{\mathbf{Q}^s} \frac{f(\mathbf{x})}{p(\mathbf{x})} p(\mathbf{x})\, d\mu^s(\mathbf{x}) \tag{6.342}$$

PDF (176)

it is absolutely not necessary to sample from the density p but that it is also possible to draw the samples from a function that is close to p on the whole integration domain. In this section, we will present a somewhat different strategy for evaluating integrals of the

Random Variables (168)

above type. Thus, we will develop a sequence of random variables $\mathbf{X}_n$ that are approximately distributed according to a probability density function without directly simulating the density. The idea underlying this strategy is the probability theoretical concept of the

DT Markov Process (236)

ergodic Markov process with *stationary distribution*.

Let us first define what we understand under the notion of a *Markov Chain Monte Carlo method*:

DEFINITION 6.5 (Markov Chain Monte Carlo Method) *A* Markov chain Monte Carlo method, *also called a* MCMC-method, *for simulating a probability distribution* p *on the probability space* $(\Omega, \mathfrak{F}(\Omega), \mathbb{P})$ *is any method that produces an* ergodic Markov process $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$, *whose stationary distribution corresponds to* p.

Section 2.4

Now, the concepts of the distribution and the Markov process from the above definition are already known to us, but what is mean with ergodic and a stationary distribution?

To illustrate these new concepts let us firstly restrict our focus to discrete-state spaces,

DT Markov Chain (226)

thus discrete-time, discrete-state Markov chains.

ERGODIC MARKOV CHAINS. As we will see in the following section, we are not so much interested in the initially dynamic of a process, simulated by a Markov chain, but rather in the state of the system after the chain has made a larger number of transitions. This will lead us to the investigation of the transition probabilities p_{ij}^n for sufficient large n, respectively the study of the limit

$$\lim_{n\to\infty} p_{ij}^n. \tag{6.343}$$

To discuss this behavior of a Markov chain, we have to introduce the concept of the stationary distribution of a Markov chain.

DEFINITION 6.6 (Stationary Distribution of a Discrete-time Markov Chain) *Let* $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ *be a discrete-time, discrete-state Markov chain with stochastic matrix* $\mathbf{M} = (p_{ij})_{i,j\in S}$ *as introduced in Definition 2.70. A probability distribution* $\boldsymbol{\pi}^* = (\pi_j^*)_{j\in S}$ *is called a* stationary distribution *for* $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$*, if it holds:*

Stochastic Matrix (229)

$$\pi_i^* \stackrel{def}{=} \sum_{j\in S} \pi_j^* p_{ij} \tag{6.344}$$

with

$$\sum_{i\in S} \pi_i^* = 1. \tag{6.345}$$

EXAMPLE 6.27 *Given be the stochastic matrix*

$$\mathbf{P} = \begin{pmatrix} 0.6 & 0.4 \\ 0.6 & 0.4 \end{pmatrix}. \tag{6.346}$$

To compute a stationary distribution for the associated Markov chain, Equation (6.344) implies that we have to find a solution for the following linear system:

$$\boldsymbol{\pi}^*\mathbf{P} = \boldsymbol{\pi}^*, \tag{6.347}$$

thus,

$$\boldsymbol{\pi}^*(\mathbf{P} - \mathbf{E}) = \mathbf{0}. \tag{6.348}$$

Two different stationary distributions are then given by $\boldsymbol{\pi}^* = (0.6, 0.4)$ *and* $\boldsymbol{\pi}^* = (0.2, 0.8)$ *with* $n \in \mathbb{N}$*. Obviously, the above Markov chain has infinitely many different stationary distributions that are depending on the initial distribution.*

For the distribution of a Markov chain $\mathbf{X}_n$ to converge towards a stationary distribution $\boldsymbol{\pi}^*$, the chain has to satisfy three important properties: First, the chain needs to be *irreducible*, which means that for any state of the chain, there is a positive probability of visiting all other states. Second, the chain should be aperiodic, that is, it should not get trapped in cycles, and finally, the chain must be positive recurrent, that is, if the initial value $\mathbf{X}_0$ is sampled from $\boldsymbol{\pi}^*$, then all subsequent values of $\mathbf{X}_n$ must also distributed according to $\boldsymbol{\pi}^*$. Let us now exactly define these properties of a stochastic process via the concept of the so-called *ergodic Markov chain*.

DEFINITION 6.7 (Ergodic Markov Chain) *A discrete-time, discrete-state Markov chain* $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ *is called* irreducible, *if any set of states can be reached from any other state in finite moves, that is, if it holds:*

$$p_{ij}^k > 0, \tag{6.349}$$

for some $k \geq 1$, *otherwise the process is called* reducible. *An irreducible stochastic process,* $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$, *is referred to as* positive recurrent, *if the expected value of the first return to some state* i *after* n *steps is finite, thus,*

$$\sum_{n=1}^{\infty} n\, p_{ii}^n < \infty. \tag{6.350}$$

Last but not least, we denote the irreducible stochastic process $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ *as* aperiodic, *if the greatest common divisor of return steps to some state is* 1, *thus,*

$$gcd\ \left\{k \geq 1 \,\middle|\, p_{ii}^k > 0\right\} = 1. \tag{6.351}$$

If a Markov chain satisfies the property of irreducibility, aperiodicity, *and* positive recurrence, *then it is also called an* ergodic Markov chain.

EXAMPLE 6.28 *Let* $\mathbf{M}$ *be a stochastic matrix associated with a Markov chain* $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ *given on a finite state-space. If* $\mathbf{M}$ *has at least one positive diagonal element, then* $\mathbf{M}$ *is obviously aperiodic. If all entries of* $\mathbf{M}$ *are positive then* $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ *is irreducible.*

In deed, the stochastic matrix from Example 6.27 is irreducible and aperiodic, but it is not positive recurrent. We leave the proof to the interested reader.

As already mentioned above, when considering a Markov chain starting at a given initial state, we are not interested in the initial dynamics of the chain, but rather in the state of the system after a large number of transitions, that is, we are interested in the limiting behavior of the chain. Obviously, the existence of a stationary distribution implies them to a first candidate for the limit distribution of a Markov chain.

As the following theorem shows, the stationary distribution of an ergodic Markov chain corresponds indeed also to the limiting distribution of successive iterates from the chain.

THEOREM 6.2 (Ergodic Theorem) *Let* $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ *be an ergodic, discrete-time, discrete-state Markov chain. Then, the stationary distribution* $\boldsymbol{\pi}^* = (\pi_j^*)_{j\in S}$ *of* $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ *is the unique probability distribution of the chain and the following consequences hold:*

i) The limiting behavior of the Markov chain converges to its stationary distribution $\boldsymbol{\pi}^*$, *that is,*

$$\lim_{n\to\infty} p_{ij}^n = \pi_j^* \qquad \forall i, j \in S. \tag{6.352}$$

ii) If $\mathrm{E}(\mathrm{f}(\mathbf{X}_n)) < \infty$, *then it holds almost surely:*

$$\mathrm{Prob}\left\{\frac{1}{N}\sum_{n=1}^{N}\mathrm{f}(\mathbf{X}_n) \to \mathrm{E}(\mathrm{f}(\mathbf{X}_n))\right\} = 1, \tag{6.353}$$

where $\mathrm{E}(\mathrm{f}(\mathbf{X}_n)) = \sum_{i\in S}\mathrm{f}(i)\pi_i^*$ *is the expected value of* $\mathrm{f}(\mathbf{X}_n)$ *with respect to* $\boldsymbol{\pi}^*$. Expected Value (196)

PROOF 6.2 *The convergence theorems of Markov chains and in particular of Markov processes are all together, not surprisingly, extremely difficult. The proof of the Ergodic Theorem is beyond the scope of this book. Because we need only the result of this theorem, we omit the proof and refer the interested reader to [130, Meyn & Tweedie 1993], [170, Robert & Casella, 1999], or [15, Ash & Doléans-Dade, 2000].*

Theorem 6.2 establishes a connection between the stationary distribution of a Markov chain to its asymptotic temporal development. The main consequence that we can draw from the Ergodic Theorem is, that the ergodicity of the chain guarantees the convergence of the chain to its stationary distribution, independent of the state, where the chain starts. Additionally, we have a tool for handling also sequences of random variables that are not necessarily independent and identically distributed. So, the Ergodic Theorem can be interpreted in some sense as a generalization of Kolmogorov's Strong Law of Large Numbers for dependent distributed random variables. SLLN (216)

After a sufficiently long *burn-in* of M iterations, then the samples $\mathbf{X}_n, n = M+1,\ldots,N$ will be dependent, approximately distributed according to $\boldsymbol{\pi}^*$ and they can be used to construct a Markov chain Monte Carlo estimator for the expected value $\mathrm{E}(\mathrm{f}(\mathbf{X}_n))$, namely:

$$\bar{\mathrm{F}}_N \overset{def}{=} \frac{1}{N-M}\sum_{n=M+1}^{N}\mathrm{f}(\mathbf{X}_n) \approx \mathrm{E}(\mathrm{f}(\mathbf{X}_n)). \tag{6.354}$$

REMARK 6.11 *As a consequence of the Ergodic Theorem we conclude, that it is a good practice to check the ergodicity property of a Markov chain, since neglecting this issue can lead to samples that do not converge to the desired stationary distribution. When a non-ergodic Markov chain starts in different initial states, it can converge to different stationary distributions due to the nonuniqueness of the solution to* $\boldsymbol{\pi}^*\mathbf{P} = \boldsymbol{\pi}^*$.

Let us now summarize some of the results of general continuous-state Markov chain theory as described in [170, Robert & Casella 1999] and [65, Gilks & al. 1996] as they apply to Markov chain Monte Carlo methods. The most results are analogous to the results for discrete-time, discrete-state Markov chains, but there are some differences.

ERGODIC MARKOV PROCESSES*. In the discrete case, a Markov chain was defined as irreducible, if all states communicate. In the continuous case, irreducibility must be defined with respect to a distribution ν.

DEFINITION 6.8 (ν-irreducibility) *Given a measure ν, then the discrete-time, continuous-state Markov process, $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$, with transition kernel, $\mathcal{K}(\mathbf{x},\mathbf{y})$, is called* ν-irreducible, *if for every set $A \in \mathfrak{B}(\Omega)$ with $\nu(A) > 0$, there exists $n \in \mathbb{N}$ such that $\mathcal{K}^n(\mathbf{x}, A) > 0$ for all $x \in \Omega$.*

Transition Kernel (234)

REMARK 6.12 *Based on the above definition, verifying the irreducibility of a Markov process is often straightforward, since we only have to check, if $\mathcal{K}^n$ has a positive density, f, such that $\mathcal{K}^n(\mathbf{x}, A) = \int_A f(\mathbf{x},\mathbf{y})d\nu(\mathbf{x})$ for all $\mathbf{x} \in \Omega$ and $A \subset \mathfrak{B}(\Omega)$. This is often the case for Metropolis samplers and as we will see for the transition kernel of Markov processes based rendering algorithms.*

Probability Density Function (176)

REMARK 6.13 *If a Markov chain is irreducible, then it has many different irreducibility distributions. However, it is possible to show that any irreducible chain has a maximal irreducibility distribution in the sense that all other irreducibility distributions are absolutely continuous with respect to the maximal irreducibility distribution [65, Gilks & al. 1996].*

Now, irreducibility is the property of a Markov process $\mathbf{X}_n$ that all interesting sets can be visited, but this property is too weak, it does not ensures that all these sets are visited often enough. Here, as known from the discrete case, we need the property of *recurrence*. The recurrence of a Markov process guarantees that all sets will be reached infinitely often, at least from almost all starting points.

DEFINITION 6.9 (Positive Recurrence) *A ν-irreducible Markov process, $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$, is* recurrent *if for any set $A \subset \mathfrak{B}(\Omega)$ with $\nu(A) > 0$ the following both conditions are satisfied:*

i) $\text{Prob}\,(\mathbf{X}_n \in A \text{ infinitely often}) > 0$ *for all x*

ii) $\text{Prob}\,(\mathbf{X}_n \in A \text{ infinitely often}) = 1$ *for ν-almost all x.*

An irreducible, recurrent Markov process is denoted as positive recurrent, *if it has an invariant probability distribution.*

REMARK 6.14 *Due to [65, Gilks & al. 1996], recurrence is sufficient to ensure that a law of large numbers holds for a Markov process, but to provide a central limit theorem, a discrete-time, continuous-state process must satisfy stronger conditions, such as the ergodicity.*

We will stop this incomplete discussion of ergodic Markov processes at this point, since a more detailed study of the topic, is out of the scope of this book. So, we refer the interested reader to the books by [65, Gilks & al. 1996], [170, Robert & Casella 1999], and in particular [130, Meyn and Tweedie 1993] as well as [204, Stroock 2005].

6.5.3.2 M(RT)^2 - METROPOLIS SAMPLING

Suppose we wish to sample from a non-negative function f, which indeed can be evaluated, but where we have no chance to directly generate a sample from. Acceptance-rejection sampling could be used if a trial density q can be found, where $\frac{f}{q}$ has a reasonable bound. Now, the main drawback of this sampling strategy is that it is often very difficult to construct a suitable proposal distribution, that leads to an efficient solution algorithm, in particular if f is high-dimensional. One way to solve this problem is to drop the strict requirement of generating independent samples and instead to generate a sequence $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ of dependent random variables such that each $\mathbf{X}_n$ is distributed according to the desired function f.

Section 6.5.2
Sequence of RVs (219)
Random Variable (168)

In this chapter we will introduce a first MCMC method: the Metropolis algorithm, [129, Metropolis & al. 1953], which goes back to Metropolis, Rosenbluth, Rosenbluth, Teller and Teller. It can be seen as the only known method of MCMC. All other MCMC methods, such as the Metropolis-Hastings algorithm, the Gibbs sampler etc. are mutations of the Metropolis algorithm. Introduced 1953 for handling difficult sampling problems in computational physics for obvious reasons it is also called the M(RT)^2 algorithm. It is an advanced sampling technique that can sample any density function in any number of dimensions.

Similar to the acceptance-rejection method it is based on proposing values sampled from an instrumental distribution. In an acceptance test then it is checked if the new tentative sample is kept —which reflects how likely it is that it is from the target distribution—or if the previous sample is furthermore used. So, the M(RT)^2 algorithm generates a sequence of correlated samples from a non-negative function f such, that the samples are distributed according to f. For that, the algorithms only needs to evaluate f at each generated sample, M(RT)^2 does not need any other information about f or its associated PDF, see Figure 6.23

DETAILED BALANCE AND THE TENTATIVE TRANSITION FUNCTION. In the following, let S be an uncountable state set, and f be a non-negative function defined on S with values in $\mathbb{R}$. We are also given some initial state $\mathbf{X}_0$. The goal of the M(RT)^2-algorithm is to generate a stochastic process $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ on the probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ independent on the initial state, such that the sequence of random variables $\mathbf{X}_n$ are approximately distributed proportional to f. In other words, we have to generate an ergodic, discrete-time Markov process $(\mathbf{X}_n)_{n\in\mathbb{N}_0}$ using a transition kernel $\mathcal{K}$ with stationary distribution f.The ergodicity of the process then ensures the convergence of $\mathbf{X}_n$ to the desired density f, that is, samples form this process are approximate simulations from f.

DT Markov Process (236)
Transition Kernel (234)
Ergodic Theorem (546)
Stationary Distribution (545)

The Metropolis algorithm precisely follows this idea by constructing a Markov process with a corresponding transition kernel. But the problem that arises is: How should we choose the transition kernel such that it fulfills the requirements from above? Here, the Metropolis algorithm makes use of an observation from physics, the so-called *detailed balance condition*.

```
M(RT)² {
    initialize X_0
    ∀ n ∈ {0, 1, ...} do {
        Y ← MUTATE(X_n)
        if RANDOM() < A(X_n → Y) {
            X_{n+1} ← Y
        } else {
            X_{n+1} ← X_n
        }
    }
}
```

FIGURE 6.23: PSEUDOCODE FOR THE METROPOLIS SAMPLING ALGORITHM.

Let us consider two boxes, one filled with gas, the other contains nothing, connected to each other via a clamped tube. If the tube is unclamped, gas begins to flow from the first box into the second box. After gas is flowed into the second box, some of that gas flows back into the first box. This process also continues even if an equilibrium between the two boxes is reached. That is, even if the system is in equilibrium, the chance that gas flows from the first box into the other is equal to the chance that gas flows from the second box into the first. This condition of a physical system, that guarantees that a system evolves toward equilibrium and stay there, is called *detailed balance*. Translated into the language of Markov processes, the detailed balance condition can be defined as follows:

Transition Kernel (234)

DEFINITION 6.10 (Detailed Balance Condition) *A Markov process with transition kernel* $\mathcal{K}$ *satisfies the* detailed balance condition *if there exists a function* f *satisfying*

$$\mathcal{K}(\mathbf{x}_{n+1} \to \mathbf{x}_n)\, f(\mathbf{x}_{n+1}) = \mathcal{K}(\mathbf{x}_n \to \mathbf{x}_{n+1})\, f(\mathbf{x}_n) \tag{6.355}$$

for all $\mathbf{x}_n, \mathbf{x}_{n+1} \in S$.

Stationary Distribution (545)

Now, in the language of Markov processes, Equation (6.355) can be interpreted in such a way, that, if a stationary distribution f is reached, then the probability of a transition from state $\mathbf{x}_{n+1}$ into state $\mathbf{x}_n$ is the same as being in state $\mathbf{x}_n$ and taking a transition into state $\mathbf{x}_{n+1}$.

REMARK 6.15 *To make our formulas easier readable, we have separate ourselves in the above definition from the commonly used notation for transition kernels, that is,*

in this section, we define:

$$\mathcal{K}(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1}) \stackrel{\text{def}}{=} \mathcal{K}(\mathbf{x}_n, \mathbf{x}_{n+1}), \tag{6.356}$$

which strongly clarifies the transition from $\mathbf{x}_n$ *into* $\mathbf{x}_{n+1}$.

Now, for constructing the transition kernel $\mathcal{K}$, M(RT)^2 uses a so-called *tentative transition function* T to propose a transition from the current state into some other state according to some chosen distribution. So, $T(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1})$ is a probability density function which gives the probability that the state $\mathbf{X}_{n+1} = \mathbf{x}_{n+1}$ given $\mathbf{X}_n = \mathbf{x}_n$.

THE ACCEPTANCE PROBABILITY FUNCTION. The tentative sample $\mathbf{X}_{n+1}$ is then either accepted or rejected according to a so-called *acceptance probability*. In order to reach the stationary distribution as quickly as possible, due to [221, Veach 1998], the best strategy should be to make $A(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1})$ and $A(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n)$ as large as possible. So, the *acceptance probability function*, A, is defined by:

$$A(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1}) \stackrel{\text{def}}{=} \min\left(1, \frac{f(\mathbf{x}_{n+1})\, T(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n)}{f(\mathbf{x}_n)\, T(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1})}\right), \tag{6.357}$$

where $A(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1})$ denotes the probability with which a move from $\mathbf{x}_n$ to $\mathbf{x}_{n+1}$ is accepted. This choice of the acceptance probability function guarantees, that transitions in one direction are always accepted, while in the other they are sometimes rejected, such that the expected number of moves each way is the same [221, Veach 1998].

REMARK 6.16 *Note: If the transition probability density is the same in both directions, the acceptance probability function simplifies to the division of the value of* f *at the tentative and the current sample, namely:*

$$A(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1}) = \min\left(1, \frac{f(\mathbf{x}_{n+1})}{f(\mathbf{x}_n)}\right). \tag{6.358}$$

M(RT)^2 then constructs the transition kernel, in this case also denoted as the *Metropolis kernel*, via the tentative function T and the acceptance function A by:

$$\mathcal{K}(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1}) \stackrel{\text{def}}{=} T(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1})\, A(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1}). \tag{6.359}$$

In the following lemma we will show that the Metropolis kernel satisfies the detailed balance condition.

LEMMA 6.5 *Let* $\mathcal{K}$ *be the Metropolis kernel from Equation (6.359), then it holds:* $\mathcal{K}$ *satisfies the detailed balance condition.*

PROOF 6.5 *Due to Definition 6.10 we have to show the validity of Equation (6.355) for a probability density function* f*. Let us first prove the case, where it holds:*

$$A(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1}) = 1, \tag{6.360}$$

which is equivalent to

$$\frac{f(\mathbf{x}_{n+1})\, T(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n)}{f(\mathbf{x}_n)\, T(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1})} \geq 1. \tag{6.361}$$

Due to Equation (6.357) then we get:

$$A(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n) = \frac{f(\mathbf{x}_n)\, T(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1})}{f(\mathbf{x}_{n+1})\, T(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n)}. \tag{6.362}$$

Using this relation in Equation (6.359) leads to:

$$\begin{aligned}
f(\mathbf{x}_{n+1})\, \mathcal{K}(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n) &\overset{def}{=} f(\mathbf{x}_{n+1})\, T(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n)\, A(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n) && (6.363)\\
&\overset{(6.362)}{=} f(\mathbf{x}_{n+1})\, T(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n)\, \frac{f(\mathbf{x}_n)\, T(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1})}{f(\mathbf{x}_{n+1})\, T(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n)} \\
&\overset{(6.360)}{=} f(\mathbf{x}_n)\, T(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1})\, A(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1}) && (6.364)\\
&\overset{def}{=} f(\mathbf{x}_n)\, \mathcal{K}(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1}). && (6.365)
\end{aligned}$$

The case where we have

$$A(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1}) = \frac{f(\mathbf{x}_{n+1})\, T(\mathbf{x}_{n+1} \rightarrow \mathbf{x}_n)}{f(\mathbf{x}_n)\, T(\mathbf{x}_n \rightarrow \mathbf{x}_{n+1})} \tag{6.366}$$

can similarly be proofed. We let the details to the interested reader as a simple exercise.

Ok, the Metropolis kernel satisfies the detailed balance condition. In the following lemma we will show that, due to the fact that it fulfills the detailed balance condition, the Metropolis kernel has a stationary distribution.

LEMMA 6.6 *Let* $(\mathbf{X}_n)_{n \in \mathbb{N}_0}$ *be a sequence of random variables constructed by the* $M(RT)^2$ *algorithm with the Metropolis kernel from Relation (6.359). Let us furthermore assume that* p_n *is the probability density function associated with the random variable* $\mathbf{X}_n$*, then the Metropolis kernel has a stationary distribution.*

PROOF 6.6 *Let* p_{n+1} *be the PDF associated with the random variable* $\mathbf{X}_{n+1}$*. For the transition from* $\mathbf{X}_n$ *to* $\mathbf{X}_{n+1}$*, there are three possibilities that have to be accounted for: Namely, either we are already in state* $\mathbf{x}_{n+1}$*, or we reach state* $\mathbf{x}_{n+1}$ *via an*

accepted transition from any $\mathbf{x}_n$*, or we are already in state* $\mathbf{x}_{n+1}$ *and a transition to any state* $\mathbf{x}_n$ *is rejected. So,* p_{n+1} *can easily be computed via:*

$$\begin{aligned} p_{n+1}(\mathbf{x}_{n+1}) &= p_n(\mathbf{x}_{n+1}) + \int_\Omega p_n(\mathbf{x}_n)\,\mathcal{K}(\mathbf{x}_n \to \mathbf{x}_{n+1})\mathrm{d}\mu(\mathbf{x}_n) - \\ &\quad \int_\Omega p_n(\mathbf{x}_{n+1})\,\mathcal{K}(\mathbf{x}_{n+1} \to \mathbf{x}_n)\mathrm{d}\mu(\mathbf{x}_n) \qquad (6.367) \\ &\overset{(6.355)}{=} p_n(\mathbf{x}_{n+1}) + \int_\Omega p_n(\mathbf{x}_n)\,\mathcal{K}(\mathbf{x}_n \to \mathbf{x}_{n+1})\mathrm{d}\mu(\mathbf{x}_n) - \\ &\quad \int_\Omega p_n(\mathbf{x}_n)\,\mathcal{K}(\mathbf{x}_n \to \mathbf{x}_{n+1})\mathrm{d}\mu(\mathbf{x}_n) \qquad (6.368) \\ &= p_n(\mathbf{x}_{n+1}), \qquad (6.369) \end{aligned}$$

but this means, that p_n *is a stationary distribution of the Markov process.*

From the last lemma we conclude: As the Metropolis algorithm is based on a Markov kernel that satisfies the detailed balance condition, the associated Markov process has a stationary distribution. Now, $\mathrm{M(RT)}^2$ does not only fulfills the requirement that it constructs a stochastic process satisfying the detailed balance condition, but it also guarantees that the process is ergodic—we omit the proof since it is outside the scope of this book, for a proof see [170, Robert & Casella 1999]. Then, the Ergodic Theorem ensures that the stationary distribution converge for any initial distribution to the unique equilibrium distribution of the process. Since the asymptotic distribution of the Markov process was assumed to corresponds to f, so, samples $\mathbf{X}_n$ from the process will be approximate simulations from f. This can easily bee see by interchanging p_n in Equation (6.368) with the target density, f:

$$\begin{aligned} p_{n+1}(\mathbf{x}_{n+1}) &= f(\mathbf{x}_{n+1}) + \int_\Omega f(\mathbf{x}_n)\,\mathcal{K}(\mathbf{x}_n \to \mathbf{x}_{n+1})\mathrm{d}\mu(\mathbf{x}_n) - \\ &\quad \int_\Omega f(\mathbf{x}_{n+1})\,\mathcal{K}(\mathbf{x}_{n+1} \to \mathbf{x}_n)\mathrm{d}\mu(\mathbf{x}_n) \qquad (6.370) \\ &\overset{(6.355)}{=} f(\mathbf{x}_{n+1}) + \int_\Omega f(\mathbf{x}_n)\,\mathcal{K}(\mathbf{x}_n \to \mathbf{x}_{n+1})\mathrm{d}\mu(\mathbf{x}_n) - \\ &\quad \int_\Omega f(\mathbf{x}_n)\,\mathcal{K}(\mathbf{x}_n \to \mathbf{x}_{n+1})\mathrm{d}\mu(\mathbf{x}_n) \qquad (6.371) \\ &= f(\mathbf{x}_{n+1}). \qquad (6.372) \end{aligned}$$

REMARK 6.17 *Let us summarize briefly: The* $\mathrm{M(RT)}^2$ *algorithm generates a sequence of correlated samples from a non-negative function* f *such, that the samples are distributed according to* f*. For that, the algorithms only needs to evaluate* f *at each generated sample,* $\mathrm{M(RT)}^2$ *does not need any other information on* f *or its associated PDF. Furthermore, we can say, that the asymptotic distribution of a Markov process requires to throw away the first* M *samples until the process approaches the limit*

distribution, where we can make statements about M *only in rare cases. Since the samples are also dependent, the variance of a Monte Carlo estimator using samples from the Metropolis algorithm will be larger than if the process would be independent. Finding the ideal proposal distribution is an art. Probably, this is the price we have to pay for the generality of the* $M(RT)^2$ *algorithm.*

As already mentioned above, the Metropolis sampling algorithm is the only known method of MCMC. There are a series of other algorithms such as the Metropolis-Hastings algorithm, or the Gibbs-Sampler. They are all special versions of the $M(RT)^2$ *algorithm. So, the Metropolis algorithm is the ideal algorithm to start into the theory of MCMC.*

6.6 VARIANCE REDUCTION TECHNIQUES

The foregoing discussions have shown that with every random variable also comes a variance which limits the desired precision of the result. We also discovered in Section 6.4 that with an increasing number of samples, i.e. a longer computation time for an estimator, more precise results may be expected. We asked now: It is possible to amortize such an increase in run time by a more efficient choice of the samples? A quantity that can be used to measure the quality of a Monte Carlo estimator is the *efficiency*. This concept is of particular importance when developing efficient algorithms for solving the rendering equation using Monte Carlo integration. Formally, the efficiency is defined as:

Chapter 9

Monte Carlo Estimator (499)

DEFINITION 6.11 (Efficiency of a Monte Carlo Estimator) *Let* F_N *be any Monte Carlo estimator, the* efficiency $\epsilon(F_N)$ *is defined as:*

$$\epsilon(F_N) \stackrel{\text{def}}{=} \frac{1}{\text{Var}(F_N) \cdot T(F_N)}, \tag{6.373}$$

where $T(F_N)$ *are the costs of the underlying algorithm, i.e. the time required to evaluate* F_N *and* $\text{Var}(F_N)$ *is the variance of the estimator* F_N.

Variance (201)

A main goal in the theory of Monte Carlo integration is to improve the efficiency of a Monte Carlo estimator. This can be reached by applying so-called *variance reduction techniques*.

Secondary Estimator (499)

Now, from Lemma 6.3 we know that the error of a secondary Monte Carlo estimator can be made as small as desired, assumed we take sufficiently many samples. However, with this linear decrease, by a factor of N, also a linear increase of the run time for evaluating a secondary Monte Carlo estimator is associated by the same factor.

Secondary Estimator (499)

For demonstrating this, let us now consider once more the estimator F_N from Equation (6.114) with independent and identically distributed samples $\mathbf{X}_1, \ldots, \mathbf{X}_N$ selected over $\mathbf{Q}^s$.

Independent RV (204)

Obviously the following applies to the variance of F_N:

$$Var(F_N) \overset{(6.182)}{=} \frac{1}{N} Var(F_1), \tag{6.374}$$

while the run time for evaluating F_N satisfies the equation

$$T(F_N) = N\,T(F_1). \tag{6.375}$$

From these both statements we conclude: Drawing N samples reduces the variance by a factor of N but increases the computation time by the same factor. So, the estimators F_N and F_1 can be regarded as equally efficient procedures, as it holds:

$$\epsilon(F_N) = \frac{1}{Var(F_N)T(F_N)} \tag{6.376}$$

$$\overset{(6.374),(6.375)}{=} \frac{1}{Var(F_1)T(F_1)} \tag{6.377}$$

$$= \epsilon(F_1)\,. \tag{6.378}$$

Now the main objective of Monte Carlo Integration is to maximize efficiency, i.e. to construct fast evaluable estimators with a variance as small as possible. For designing such efficient estimators it is required to obtain additional information from the problem to be solved. As already above-mentioned, the techniques with which this objective may be attained are called *variance reduction techniques*. Our following discussion will be limited to the presentation of those variance reduction techniques, which have been proven most useful for the solution of the global illumination problem.

6.6.1 USE OF EXPECTED VALUES

The first variance reduction procedure, which we will present here, *Use of Expected Values*, is based on the following unwritten law of Monte Carlo Integration [73, Hammersley & Handscomp 1964]:

... If there is anything in the integral to be determined that may be analytically integrated, it should be integrated.

The idea behind this technique lies in the reduction of the dimensionality of the original problem. For that purpose, let us write our original integration domain of the integral

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \tag{6.379}$$

as a Cartesian product of subspaces of dimension s_y and s_z, that is, we suppose that $\mathbf{Q}^s$ Subspace (855)

can be split into $\mathbf{Q}^s = \mathbf{Q}^{s_y} \times \mathbf{Q}^{s_z}$. Then, the integral from (6.379) can be rewritten as:

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) = \int_{\mathbf{Q}^{s_y}} \int_{\mathbf{Q}^{s_z}} f(\mathbf{y},\mathbf{z})\, d\mu^{s_z}(\mathbf{z})\, d\mu^{s_y}(\mathbf{y}) \tag{6.380}$$

$$= \int_{\mathbf{Q}^{s_y}} \int_{\mathbf{Q}^{s_z}} \frac{f(\mathbf{y},\mathbf{z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z})} p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z})\, d\mu^{s_z}(\mathbf{z}) d\mu^{s_y}(\mathbf{y}), \tag{6.381}$$

where we assume that $p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z})$ is any joint probability density function with respect to the random variables $\mathbf{Y}$ and $\mathbf{Z}$.

Monte Carlo Estimator (499) Random Variable (168) PDF (176)

Now, we are not interested in estimators of the form $F_N(\mathbf{Y},\mathbf{Z})$ but in estimators of the form $F_N(\mathbf{Y})$, thus, in estimators depending only on the random variable $\mathbf{Y}$. Reducing the dimensionality of the given problem means that we must be able to integrate both, the integrand $f(\mathbf{y},\mathbf{z})$ and the density function $p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z})$ with respect to $\mathbf{z}$. Integrating $f(\mathbf{y},\mathbf{z})$ with respect to the variable $\mathbf{y}$ leads to:

$$f(\mathbf{y}) = \int_{\mathbf{Q}^{s_z}} f(\mathbf{y},\mathbf{z})\, d\mu^{s_z}(\mathbf{z}) \tag{6.382}$$

Marginal Density Function (194)

and the marginal density function $p_{\mathbf{Y}}$ can be written as:

$$p_{\mathbf{Y}}(\mathbf{y}) = \int_{\mathbf{Q}^{s_z}} p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z})\, d\mu^{s_z}(\mathbf{z}). \tag{6.383}$$

Replacing on the right side of Equation (6.381) the probability density function $p_{\mathbf{Y},\mathbf{Z}}$ according to Definition 2.56 by the product of marginal density $p_{\mathbf{Y}}(\mathbf{y})$ with respect to the random variable $\mathbf{Y}$ and the conditional density $p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y})$, then we get:

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \overset{(6.381)}{=} \int_{\mathbf{Q}^{s_y}} \int_{\mathbf{Q}^{s_z}} \frac{f(\mathbf{y},\mathbf{z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z})} p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z})\, d\mu^{s_z}(\mathbf{z}) d\mu^{s_y}(\mathbf{y}) \tag{6.384}$$

$$= \int_{\mathbf{Q}^{s_y}} \left(\int_{\mathbf{Q}^{s_z}} \frac{f(\mathbf{y},\mathbf{z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z})} p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y})\, d\mu^{s_z}(\mathbf{z}) \right) p_{\mathbf{Y}}(\mathbf{y})\, d\mu^{s_y}(\mathbf{y}). \tag{6.385}$$

Due to Equation (2.821), we can consider the inner integral as the conditional expected value of the random variable $\mathbf{Z}$ given $\mathbf{Y}$, that is, as $E_{\mathbf{Z}}\left(\frac{f(\mathbf{Y},\mathbf{Z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z})} \middle| \mathbf{Y} \right)$. Using this identity, then we obtain for the integral from (6.379):

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) = \int_{\mathbf{Q}^{s_y}} E_{\mathbf{Z}}\left(\frac{f(\mathbf{Y},\mathbf{Z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z})} \middle| \mathbf{Y} \right) p_{\mathbf{Y}}(\mathbf{y})\, d\mu^{s_y}(\mathbf{y}). \tag{6.386}$$

$$\overset{(2.735)}{=} E_{\mathbf{Y}}\left(E_{\mathbf{Z}}\left(\frac{f(\mathbf{Y},\mathbf{Z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z})} \middle| \mathbf{Y} \right) \right). \tag{6.387}$$

Since the integral on the left hand side can be interpreted as the expected value of the random vector $\frac{f(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})}$, we get:

$$E_{\mathbf{X}}\left(\frac{f(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})} \right) = E_{\mathbf{Y},\mathbf{Z}}\left(\frac{f(\mathbf{Y},\mathbf{Z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z})} \right) = E_{\mathbf{Y}}\left(E_{\mathbf{Z}}\left(\frac{f(\mathbf{Y},\mathbf{Z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z})} \middle| \mathbf{Y} \right) \right), \tag{6.388}$$

where the indices emphasizes which density function is being integrated. That is, as an alternative to the above described solution via the determination of the expected value of $\frac{f(\mathbf{X})}{p(\mathbf{X})}$, the original integration problem may also be approached by calculating the conditional expected value of the random variable $\frac{f(\mathbf{Y},\mathbf{Z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z})}$ given $\mathbf{Y}$.

We now turn to the question: Why this technique may lead to variance reduction? For that purpose, let us take a look at the following estimators, where we use the abbreviations

$$F_N = \frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)} = \frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{Y}_i,\mathbf{Z}_i)}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y}_i,\mathbf{Z}_i)} \tag{6.389}$$

and

$$F_{N,\mathbf{Z}} = \frac{1}{N}\sum_{i=1}^{N}E_{\mathbf{Z}}\left(\frac{f(\mathbf{Y}_i,\mathbf{Z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y}_i,\mathbf{Z})}\middle|\mathbf{Y}_i\right). \tag{6.390}$$

Due to Lemma 2.4, for the variance of the estimator F_N it must hold:

$$Var(F_N) = E_{\mathbf{Y}}(Var_{\mathbf{Z}}(F_N)) + Var_{\mathbf{Y}}(E_{\mathbf{Z}}(F_N)), \tag{6.391}$$

which is equivalent to the equation

$$Var(F_N) - Var_{\mathbf{Y}}(E_{\mathbf{Z}}(F_N)) = E_{\mathbf{Y}}(Var_{\mathbf{Z}}(F_N)). \tag{6.392}$$

Due to its definition, the variance of a random variable is always non-negative, that is, also the expected value of the variance of the estimator F_N, thus the quantity $E_{\mathbf{Y}}(Var_{\mathbf{Z}}(F_N))$, is non-negative. Based on this result, the above equation implies to following inequality: Variance (201)

$$Var(F_N) - Var_{\mathbf{Y}}(E_{\mathbf{Z}}(F_N)) \geq 0. \tag{6.393}$$

To make a statement about the variance of the new estimator $F_{N,\mathbf{Z}}$, our goal is to express the quantity $Var_{\mathbf{Y}}(E_{\mathbf{Z}}(F))$ on the left hand side of the inequality in terms of the estimator $F_{N,\mathbf{Z}}$. For that purpose, let us consider the quantity $E_{\mathbf{Z}}(F_N)$, obviously it holds:

$$\begin{aligned}
E_{\mathbf{Z}}(F_N) &= E_{\mathbf{Z}}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{Y},\mathbf{Z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z})}\right) && (6.394)\\
&= \frac{1}{N}\sum_{i=1}^{N}E_{\mathbf{Z}}\left(\frac{f(\mathbf{Y},\mathbf{Z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y},\mathbf{Z})}\right) && (6.395)\\
&= \frac{1}{N}\sum_{i=1}^{N}\int_{\mathbf{Q}^{s_{\mathbf{Z}}}}\left(\frac{f(\mathbf{y},\mathbf{z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{y},\mathbf{z})}p_{\mathbf{Z}|\mathbf{Y}}(\mathbf{z}|\mathbf{y})\,d\mu^{s_{\mathbf{Z}}}(\mathbf{z})\right) && (6.396)\\
&\overset{(6.385)}{=} \frac{1}{N}\sum_{i=1}^{N}E_{\mathbf{Z}}\left(\frac{f(\mathbf{Y}_i,\mathbf{Z})}{p_{\mathbf{Y},\mathbf{Z}}(\mathbf{Y}_i,\mathbf{Z})}\middle|\mathbf{Y}_i\right) && (6.397)\\
&\overset{(6.390)}{=} F_{N,\mathbf{Z}}. && (6.398)
\end{aligned}$$

Using this identity, Inequality (6.393) can then be written as:

$$Var(F_N) - Var_{\mathbf{Y}}(F_{N,\mathbf{Z}}) \geq 0. \tag{6.399}$$

Since the estimator $F_{\mathbf{Z}}$ is independent on the random variable $\mathbf{Y}$, it holds:

$$Var_{\mathbf{Y}}(F_{N,\mathbf{Z}}) = Var(F_{N,\mathbf{Z}}), \tag{6.400}$$

which leads to

$$Var(F_N) - Var(F_{N,\mathbf{Z}}) \geq 0. \tag{6.401}$$

Obviously, the above inequality expresses the fact, that the variance of $F_{N,\mathbf{Z}}$ can never be greater than that of F_N, which in turn implies that all analytically integrable components of the original integrand are to be integrated.

REMARK 6.18 *Monte Carlo integration via the reduction of the dimension of the integration domain is one of the most powerful and frequently used Monte Carlo techniques, particularly in cases where the sampling and evaluation of the analytically integrated quantities is not overly time- and effort-consuming.*

6.6.2 IMPORTANCE SAMPLING

Now, we come to *Importance Sampling*, one of the most promising variance reduction procedures especially with respect to its applicability to the integral equations of global illumination theory.

As a motivation for importance sampling, let us consider the high frequency function f shown in Figure 6.24. We are interested in an estimate of the area under the graph of this function over the interval $[a, b]$. If the samples are chosen uniformly, the variance will obviously be high, since regions, which do not contribute to the estimate, are oversampled, while other important regions for estimating the integral are undersampled. An appropriate Monte Carlo estimator for approximating the area under the graph of f and limited by the interval boundaries a and b is based an a probability density function that takes many more samples in regions where f has high values. This should reduce the variance.

$p_{\mathbf{X}}$ (176)
Probability Space (163)

Importance sampling is based on the principle of selecting a probability density function $p_{\mathbf{X}}$ over the probability space $(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s), \mathbb{P}_{\mathbf{X}})$, which is similar to the integrand f, see Figure 6.25. In this way, the integral from (6.37) may be represented as stochastic expected value of the random variable $\frac{f(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})}$.

Expected Value of a RV (196)

DEFINITION 6.12 (Importance Sampling) *Let $(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s), \mathbb{P}_{\mathbf{X}})$ be a probability space and $\mathbf{X}$ a random variable defined on $\mathfrak{B}(\mathbf{Q}^s)$ with probability density, $p_{\mathbf{X}}$. Let us furthermore assume that f is a square Lebesgue-integrable function from $\mathcal{L}^2(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s))$. If*

$\mathcal{L}^2(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s))$ (107)

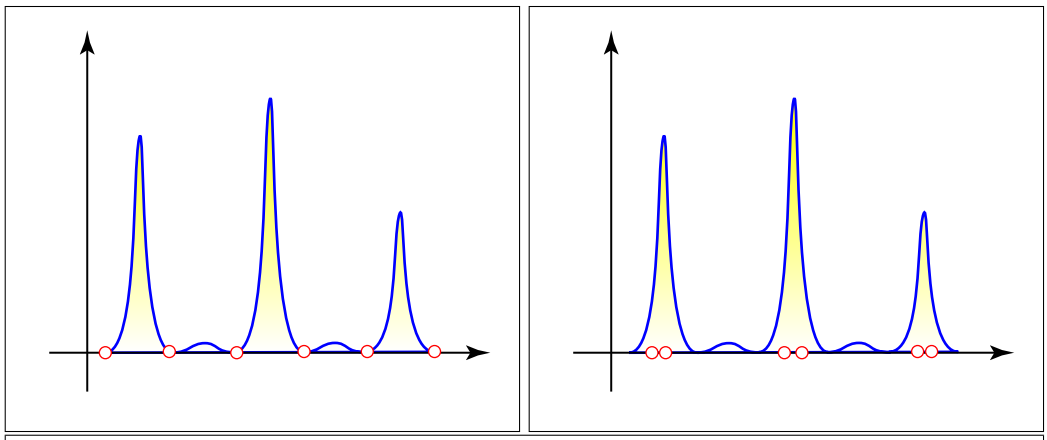


FIGURE 6.24: SAMPLING A HIGH FREQUENCY FUNCTION. We are interested in the area under the graph of a high frequency function over a given interval. If the samples are chosen uniformly, as in the image on the left, the variance will obviously be high, since regions, which do not contribute to the final value, are oversampled, while others important regions for the final value of the integral are undersampled. An appropriate probability density function for sampling has to take much more samples in regions where the function is large than in regions were the function has low values. This then leads to variance reduction.

we construct a random variable $\frac{f(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})}$ *on* $(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s))$ *then* importance sampling *is the method of evaluating the integral*

$$\mathcal{I} \stackrel{\text{def}}{=} \int_{\mathbf{Q}^s} f(\mathbf{x}) \, d\mu^s(\mathbf{x}) \tag{6.402}$$

via a secondary unbiased estimator F_N^{IS} *of the form*

$$F_N^{IS} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}, \tag{6.403}$$

where $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ *are i.i.d. random variables, which are sampled from parts of the integration domain that are of most importance to the estimate of the integral instead of spreading them out evenly in* $\mathbf{Q}^s$.

It can easily be shown that the estimator F_N^{IS} is unbiased, thus it holds: Unbiased MC Estimator (507)

$$E\left(F_N^{IS}\right) = \int_{\mathbf{Q}^s} f(\mathbf{x}) \, d\mu^s(\mathbf{x}), \tag{6.404}$$

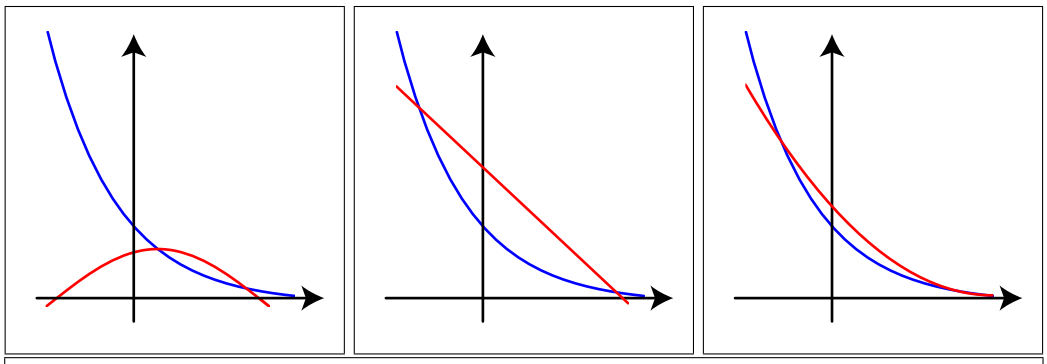


FIGURE 6.25: DIFFERENT DENSITIES FOR IMPORTANCE SAMPLING. Three density functions as possible candidates for importance sampling the function $f(x) = e^{-x}$. A bad choice of an importance sampling density is shown in the left image. This density does not match the shape of the function f we want to evaluate. A better choices is the density illustrated in the image in the center. Obviously, the best result can be expected by using the density on the right-hand side, since both graphs are equal in a large region of the integration domain. The variance of the three estimators clearly decreases from left to right.

and the variance of F_N^{IS} is given by:

$$Var\left(F_N^{IS}\right) = Var\left(\frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}\right) \tag{6.405}$$

$$\overset{(6.178)}{=} \frac{1}{N}\sum_{i=1}^{N} Var\left(F_1^{IS}\right). \tag{6.406}$$

As the variance depends on the PDF used, we now ask how $p_{\mathbf{X}}$ must be chosen to achieve an estimator with smaller variance. The choice of such a PDF is the difficulty in importance sampling. In the following lemma, we will show that the variance of the estimator F_N^{IS} will be minimal, if $p_{\mathbf{X}}$ is proportional to $|f(\mathbf{x})|$.

LEMMA 6.7 *Let F_N^{IS} be the secondary unbiased Monte Carlo estimator*

$$F_N^{IS} \overset{def}{=} \frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}, \tag{6.407}$$

then the minimal variance of F_N^{IS} is equal to

$$\min Var\left(F_N^{IS}\right) = \frac{1}{N}\left(\left(\int_{\mathbf{Q}^s} |f(\mathbf{x})|\, d\mu^s(\mathbf{x})\right)^2 - \mathcal{I}^2\right). \tag{6.408}$$

PROOF 6.7 *The proof of the above statement makes use of the fundamental Cauchy-Schwartz Inequality from functional analysis. Due to Relation (6.406) it holds for*

Cauchy-Schwartz Inequality (859)

the variance of the estimator F_N^{IS}*:*

$$\begin{aligned}
\mathrm{Var}\left(F_N^{IS}\right) &= \frac{1}{N}\mathrm{Var}\left(F_1^{IS}\right) && (6.409)\\
&= \frac{1}{N}\left(\int_{\mathbf{Q}^s}\frac{f^2(\mathbf{x})}{p_{\mathbf{X}}(\mathbf{x})}\,d\mu^s(\mathbf{x}) - \mathcal{I}^2\right) && (6.410)\\
&= \frac{1}{N}\left(\int_{\mathbf{Q}^s}\frac{f^2(\mathbf{x})}{p_{\mathbf{X}}(\mathbf{x})}\,d\mu^s(\mathbf{x})\int_{\mathbf{Q}^s}p_{\mathbf{X}}(\mathbf{x})\,d\mu^s(\mathbf{x}) - \mathcal{I}^2\right) && (6.411)\\
&\overset{(A.64)}{\geq} \frac{1}{N}\left(\left(\int_{\mathbf{Q}^s}\frac{|f(\mathbf{x})|}{(p_{\mathbf{X}}(\mathbf{x}))^{\frac{1}{2}}}(p_{\mathbf{X}}(\mathbf{x}))^{\frac{1}{2}}\,d\mu^s(\mathbf{x})\right)^2 - \mathcal{I}^2\right) && (6.412)\\
&= \frac{1}{N}\left(\left(\int_{\mathbf{Q}^s}|f(\mathbf{x})|\,d\mu^s(\mathbf{x})\right)^2 - \mathcal{I}^2\right). && (6.413)
\end{aligned}$$

Obvioulsy, the lower bound for the variance of the estimator F_N^{IS} occurs if the involved random variables are distributed according to

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{|f(\mathbf{x})|}{\int_{\mathbf{Q}^s}|f(\mathbf{x})|\,d\mu^s(\mathbf{x})}. \tag{6.414}$$

As a Monte Carlo estimator based on this PDF is unbiased, it has no variance, that is, the PDF $p_{\mathbf{X}}$ from Relation (6.414) is the best possible PDF for importance sampling. But it should also be clear, that this *best PDF* is of no practical use, since it requires the value of the integral, which we want to compute, in its denominator.

Monte Carlo Estimator (499)

Nevertheless, variance reduction may still be obtained by choosing appropriate density functions, which have similar shape as the integrand. With a probability density function $p_{\mathbf{X}}$ similar to the integrand, the above result implies variance reduction. Choosing $p_{\mathbf{X}} \propto f$, thus $p_{\mathbf{X}} = C\,f$, where C is a constant, then we can deduce:

$$\begin{aligned}
\mathrm{Var}(F_N^{IS}) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{C}\right) && (6.415)\\
&= \mathrm{Var}\left(\frac{1}{C}\right) && (6.416)\\
&= \mathrm{E}\left(\frac{1}{C^2}\right) - \mathrm{E}^2\left(\frac{1}{C}\right) \overset{(2.789)}{=} 0. && (6.417)
\end{aligned}$$

To achieve an efficient Monte Carlo strategy on this way, care must be taken to choose a PDF from where samples can easily be drawn.

When developing Monte Carlo algorithms for solving the light transport problem in Chapter 9, we will often encounter the problem that the integrands in the corresponding integrals are very complex. Since it is very difficult and often also impossible to construct a

PDF that is similar to the whole integrand, a popular method is to decompose an integrand f in a product of two or more measurable functions such as, for example: Measurable Function (98)

$$f(\mathbf{x}) = g(\mathbf{x})\, h(\mathbf{x}), \tag{6.418}$$

where g or h are easily to sample from. If, say, the density function $p_{\mathbf{X}}$ can be chosen as proportional to g, that is:

$$p_{\mathbf{X}}(\mathbf{x}) = C\, g(\mathbf{x}), \tag{6.419}$$

then this approach leads, based on i.i.d. $\mathbb{P}_{\mathbf{X}}$-distributed random samples $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_N$, to the following Monte Carlo approach:

$$\int_{\mathbf{Q}^s} g(\mathbf{x}) h(\mathbf{x})\, d\mu^s(\mathbf{x}) \overset{p_{\mathbf{X}} \propto g}{=} \frac{1}{C} \int_{\mathbf{Q}^s} p_{\mathbf{X}}(\mathbf{x})\, h(\mathbf{x})\, d\mu^s(\mathbf{x}) \tag{6.420}$$

$$\overset{d\mu^s = \frac{d\mathbb{P}}{p_{\mathbf{X}}}}{=} \frac{1}{C} \int_{\mathbf{Q}^s} h(\mathbf{X}(\omega))\, d\mathbb{P}(\omega) \tag{6.421}$$

$$\approx \frac{1}{CN} \sum_{i=1}^{N} h(\mathbf{X}_i). \tag{6.422}$$

Now, we will illustrate the technique of importance sampling by means of an example from theory of global illumination: Estimating the reflectance equation.

EXAMPLE 6.29 (Trivial Importance Sampling Applied to the Reflectance Equation) *A trivial importance sampling approach for estimating the reflectance equation* Reflectance Equation (321)

$$L_o(\mathbf{s}, \omega_o) \overset{def}{=} \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i), \tag{6.423}$$

is based on a probability density function $p^{\perp}_{\omega_i}$ *defined as*

$$p^{\perp}_{\omega_i}(\omega_i) = C\, f_r(\mathbf{s}, \omega_i \to \omega_o) \tag{6.424}$$

using i.i.d. random variables $(\omega_1, \ldots, \omega_N)$*, where* C *is the normalization constant of* $p^{\perp}_{\omega_i}$ *given by:*

$$C = \frac{1}{\int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s}, \omega_i \to \omega_o)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i)}. \tag{6.425}$$

Then, the associated secondary Monte Carlo estimator F^{IS}_N *is unbiased and has*

the form

$$F_N^{IS} = \frac{1}{N}\sum_{i=1}^{N}\frac{f_r(\mathbf{s},\boldsymbol{\omega}_i\rightarrow\omega_o)\,L_i(\mathbf{s},\boldsymbol{\omega}_i)}{p^{\perp}_{\boldsymbol{\omega}_i}(\boldsymbol{\omega}_i)} \tag{6.426}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{f_r(\mathbf{s},\boldsymbol{\omega}_i\rightarrow\omega_o)\,L_i(\mathbf{s},\boldsymbol{\omega}_i)}{C\,f_r(\mathbf{s},\boldsymbol{\omega}_i\rightarrow\omega_o)} \tag{6.427}$$

$$= \frac{1}{C\,N}\sum_{i=1}^{N}L_i(\mathbf{s},\boldsymbol{\omega}_i), \tag{6.428}$$

that is, an approximation for the value of the reflectance equation at surface point $\mathbf{s}$ *can be achieved by summing up the radiance incident at* $\mathbf{s}$ *via independent and identically according to the PDF* $p^{\perp}_{\boldsymbol{\omega}_i}$ *distributed direction samples.*

EXAMPLE 6.30 (Importance Sampling Applied to the Reflectance Equation, Cosine-weighted Hemisphere Sampling) *Let us now take a look at the reflectance equation*

Reflectance Equation (321)

$$L_o(\mathbf{s},\omega_o) \stackrel{def}{=} \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s},\omega_i\rightarrow\omega_o)\,L_i(\mathbf{s},\omega_i)\,|\cos\theta_i|\,d\sigma_{\mathbf{s}}(\omega_i), \tag{6.429}$$

where we have taken out the cosine-term from the projected solid angle measure. It is a multi-dimensional integral where the integrand is a product of three real-valued functions, namely, the BRDF f_r*, the incident radiance* L_i*, as well as a cosine-term.*

A naive Monte Carlo strategy for evaluating the reflectance equation could be to sample a direction uniformly over the hemisphere, but our discussion from above has shown, that an importance sampling strategy—such as those chosen in the foregoing example—promises less variance in a corresponding estimator.

Now, due to the cosine-term in the integrand, the contribution of radiance—incident at surface point $\mathbf{s}$ *over directions near the equator—to the final value is minuscule or perhaps even zero. Instead tracing such rays, a better strategy would be to sample directions where the cosine-term is large, that is, sampling directions near the surface normal. Now, an importance sampling strategy to estimate the reflectance equation could be to use a PDF that is a composition of the BRDF and the cosine-term, thus:*

$$p_{\boldsymbol{\omega}_i}(\boldsymbol{\omega}_i) = C\,f_r(\mathbf{s},\boldsymbol{\omega}_i\rightarrow\omega_o)\,|\cos\theta_i|. \tag{6.430}$$

where C *is the normalization constant of* $p_{\boldsymbol{\omega}_i}$ *given by:*

$$C = \frac{1}{\int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s},\boldsymbol{\omega}_i\rightarrow\omega_o)\,|\cos\theta_i|\,d\sigma_{\mathbf{s}}(\omega_i)}. \tag{6.431}$$

and it holds: $|\cos\theta_i| = \langle\mathbf{N}(\mathbf{s}),\cos\boldsymbol{\omega}_i\rangle$.

Based on identically and independent according to the PDF p_{ω_i} *distributed random samples* $\omega_1, \ldots, \omega_N$*, then the secondary unbiased Monte Carlo estimator* F_N^{IS} *has the form:*

$$F_N^{IS} = \frac{1}{N} \sum_{i=1}^{N} \frac{f_r(s, \omega_i \to \omega_o)\, L_i(s, \omega_i)\, |\cos \omega_i|}{p_{\omega_i}(\omega_i)} \tag{6.432}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \frac{f_r(s, \omega_i \to \omega_o)\, L_i(s, \omega_i)\, |\cos \omega_i|}{C\, f_r(s, \omega_i \to \omega_o)\, |\cos \omega_i|} \tag{6.433}$$

$$= \frac{1}{C\,N} \sum_{i=1}^{N} L_i(s, \omega_i). \tag{6.434}$$

In the following example we will pick up the technique of cosine-weighted hemisphere sampling as an importance sampling strategy for the reflectance equation, but now we will use a concrete BRDF.

EXAMPLE 6.31 (Importance Sampling Applied to the Reflectance Equation, Cosine-weighted Hemisphere Sampling) *Expressed in terms of spherical coordinate the reflectance equation has the form:*

Spherical Coordinates (832)

$$L_o(s, \phi_o, \theta_o) \tag{6.435}$$
$$= \int_{[0,2\pi)} \int_{[0,\frac{\pi}{2}]} f_r(s, (\phi_i, \theta_i) \longrightarrow (\phi_o, \theta_o))\, L_i(s, \phi_i, \theta_i) \sin\theta_i\, |\cos\theta_i|\, d\mu(\theta_i)\, d\mu(\phi_i),$$

thus, a two-dimensional integral where the integrand is a product of four real-valued functions, namely, the BRDF f_r*, the incident radiance* L_i*, as well as a sine and a cosine-term.*

Ideal Diffuse BRDF (339)

Let us further assume that the involved BRDF describes ideal diffuse reflection. Then, due to Relation (4.180) the BRDF corresponds to $f_r^o = \frac{\rho_{dh}}{\pi}$*, that is, the reflectance equation can be written as:*

$$L_o(s, \phi_o, \theta_o) = \int_{[0,2\pi)} \int_{[0,\frac{\pi}{2}]} \frac{\rho_{dh}}{\pi}\, L_i(s, \phi_i, \theta_i) \sin\theta_i\, |\cos\theta_i|\, d\mu(\theta_i)\, d\mu(\phi_i). \tag{6.436}$$

Due to the fact, that the function L_i *is unknown, the above approach recommends*

the choice of a density $p_{\Phi,\Theta} \propto f_r^o \sin\theta|\cos\theta|$ *leading to:*

$$p_{\Phi,\Theta}(\phi,\theta) = \frac{f_r^o \sin\theta\,|\cos\theta|}{\int_{[0,2\pi)}\int_{[0,\frac{\pi}{2}]} f_r^o \sin\theta\,|\cos\theta|\,d\mu(\theta)\,d\mu(\phi)} \tag{6.437}$$

$$\overset{f_r^o = \frac{\rho}{\pi}dh}{=} \frac{\sin\theta\,|\cos\theta|}{2\pi\int_{[0,\frac{\pi}{2}]}\sin\theta\,|\cos\theta|\,d\mu(\theta)} \tag{6.438}$$

$$\overset{\text{Int. by parts}}{=} \frac{2\sin\theta\,|\cos\theta|}{2\pi\left(\sin^2\theta\big|_0^{\frac{\pi}{2}}\right)} \tag{6.439}$$

$$= \frac{1}{2\pi}\left(2\sin\theta\,|\cos\theta|\right). \tag{6.440}$$

Obviously, the density $p_{\Phi,\Theta}(\phi,\theta)$ *allows the representation as a product of two density functions* $p_\Phi = \frac{1}{2\pi}$ *and* $p_\Theta = 2\sin\theta\,|\cos\theta|$ *of independent random variables* Φ,Θ. *With respect to their distribution functions we obtain* CDF (171)

$$p_\Phi(\phi_i) \overset{def}{=} \int_{[0,\phi_i]} \frac{1}{2\pi} d\mu(\phi) = \frac{\phi_i}{2\pi} \tag{6.441}$$

and

$$p_\Theta(\theta_i) \overset{def}{=} \int_{[0,\theta_i]} 2\sin\theta\,|\cos\theta|\,d\mu(\theta) = \sin^2\theta_i. \tag{6.442}$$

Applying the inversion method with respect to the random variables U_1 *and* U_2 *uniformly distributed on the unit interval* I *yields:* Inversion Method (520) Uniform Distribution (180)

$$p_\Phi(\phi_i) = U_1 \quad \Rightarrow \quad \frac{\phi_i}{2\pi} = U_1 \tag{6.443}$$

$$\Rightarrow \quad \phi_i = 2\pi U_1 \tag{6.444}$$

and

$$p_\Theta(\theta_i) = U_2 \quad \Rightarrow \quad \sin^2\theta_i = U_2 \tag{6.445}$$

$$\Rightarrow \quad \theta_i = \arcsin\sqrt{U_2}. \tag{6.446}$$

From all these considerations, we conclude that the radiance $L_o(s,\omega_o)$ *exitant from point* s *into the direction* ω_o *may be approached via the Monte Carlo estimator*

$$F_N^{f_r^o,IS} = \frac{1}{N}\sum_{i=1}^{N} L_i(s,\phi_i,\theta_i) \tag{6.447}$$

$$= \frac{1}{N}\sum_{i=1}^{N} L_i(s,2\pi U_1,\arcsin\sqrt{U_2}) \tag{6.448}$$

using the probability density functions p_Φ *and* p_Θ.

REMARK 6.19 *From a mathematical point of view importance sampling may be conceived of as a procedure based on the transformation method, if the underlying distribution is invertible. The selection of an independent random variable* $\mathbf{U}$ *uniformly distributed over* $\mathbf{I}^s$ *together with* $\mathbf{X}_i = \mathbb{P}_{\mathbf{U}}^{-1}(\mathbf{U})$ *yields:* Section 6.5.1

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \overset{d\mu^s = \frac{d\mathbb{P}}{p_{\mathbf{X}}}}{=} \int_{\mathbf{Q}^s} \frac{f(\mathbf{X}(\omega))}{p_{\mathbf{X}}(\mathbf{X}(\omega))}\, d\mathbb{P}(\omega) \tag{6.449}$$

$$\overset{\mathbf{X} = \mathbb{P}_{\mathbf{U}}^{-1}(\mathbf{U})}{=} \int_{\mathbf{I}^s} \frac{f\left(\mathbb{P}_{\mathbf{U}}^{-1}(\mathbf{U}(\omega))\right)}{p_{\mathbf{U}}\left(\mathbb{P}_{\mathbf{U}}^{-1}(\mathbf{U}(\omega))\right)}\, d\mathbb{P}(\omega), \tag{6.450}$$

which suggests a simple and appropriate rule for the implementation of importance sampling.

REMARK 6.20 *Importance Sampling has been found to be a highly efficient procedure, particularly with respect to many problem areas in global illumination theory, where it is the most frequently applied technique for variance reduction. So, importance sampling is not only used in pixel filtering [185, Shirley 2000] or for estimating the incident radiance at surface points emanating from light sources, [188, Shirley & al. 1996], but also when sampling special BRDFs and small light sources [181, Shirley 1990]. In all these cases it has been found to be far superior to other variance reduction procedures.* BRDF (320)

REMARK 6.21 *Ideal areas of application for importance sampling are* Monte Carlo path tracing *[116, Lafortune 1996], [50, Dutré 2003],* Monte Carlo light tracing *[116, Lafortune 1996], [47, Dutré 1996], and* bidirectional path tracing *[116, Lafortune 1996], [221, Veach 1998]. As we will see, these are rendering algorithms based on Monte Carlo methods applied for solving the stationary light transport equation in a vacuum.* Section 9.1 Section 9.2 Section 9.3 Section 4.4.2

6.6.3 CONTROL VARIATES

Another variance reduction technique for estimating integrals via Monte Carlo integration is *Control Variates*. The idea behind control variates is to find a function g similar to the integrand f that can be integrated analytically, and then subtract it. While g is integrated analytically, the difference between f and g is estimated via a Monte Carlo strategy, where instead to sample all points independently, control variates makes use of correlated points in the sampling [99, Kalos & Whitlock 1986].

The mathematical basis for control variates is the linearity property of the Lebesgue integral, i.e. one attemps to find an analytically square Lebesgue-integrable function g from $\mathcal{L}^2(\mathbf{Q}^s, \mu^s)$ similar to the integral Lebesgue Integral (105) $\mathcal{L}^1(\mathcal{R}, \mu^s)$ (107)

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}), \tag{6.451}$$

$\|\cdot\|_\infty$ (33) whereas for a small, non-negative $\tau \in \mathbb{R}$ it holds:

$$\|f - g\|_\infty < \tau. \tag{6.452}$$

DEFINITION 6.13 (Control Variates) *Let $(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s), \mathbb{P}_{\mathbf{X}})$ be a probability space and $\mathbf{X}$ a random variable defined on $(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s))$ with probability density $p_{\mathbf{X}}$. Let us furthermore assume that f and g are a square Lebesgue-integrable function that satisfies Condition (6.452), then* control variates *is the method of evaluating the integral*

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \tag{6.453}$$

via a secondary estimator F_N^{CV} of the form

$$F_N^{CV} \overset{def}{=} \int_{\mathbf{Q}^s} g(\mathbf{x})\, d\mu^s(\mathbf{x}) + \frac{1}{N}\sum_{i=1}^{N} \frac{f(\mathbf{X}_i) - g(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}, \tag{6.454}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are i.i.d. samples drawn from the probability density function $p_{\mathbf{X}}$. Note that we use the same samples in both function, f and g.

Obviously, the estimator F_N^{CV} is well-defined, as it holds:

$$\begin{aligned} \int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) &= \int_{\mathbf{Q}^s} g(\mathbf{x})\, d\mu^s(\mathbf{x}) + \int_{\mathbf{Q}^s} (f(\mathbf{x}) - g(\mathbf{x}))\, d\mu^s(\mathbf{x}) && (6.455)\\ &\overset{d\mu^s = \frac{d\mathbb{P}}{p}}{=} \int_{\mathbf{Q}^s} g(\mathbf{x})\, d\mu^s(\mathbf{x}) + \underbrace{\int_{\mathbf{Q}^s} \frac{f(\mathbf{X}(\omega)) - g(\mathbf{X}(\omega))}{p_{\mathbf{X}}(\mathbf{X}(\omega))} d\mathbb{P}(\omega)}_{E\left(\frac{f(\mathbf{X}) - g(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})}\right)}. && (6.456) \end{aligned}$$

As the function g is known, the integral on the right-hand side can be evaluated exactly, that is, the variance of F_N^{CV} is then given by:

$$Var\left(F_N^{CV}\right) = \frac{1}{N^2}\sum_{i=1}^{N} Var\left(\frac{f(\mathbf{X}_i) - g(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}\right). \tag{6.457}$$

Obviously, variance reduction can only be achieved if the following applies:

$$Var\left(\frac{f(\mathbf{X}_i) - g(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}\right) \leq Var\left(\frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}\right). \tag{6.458}$$

Restricting the integral domain in Equation (6.451) to the s-dimensional unit cube and assuming f be square Lebesgue-integrable on $[0,1]^s$, then we obtain via independent and uniformly distributed random variables $\mathbf{U}, \mathbf{U}_1, \ldots, \mathbf{U}_N$: Uniform Distribution (180)

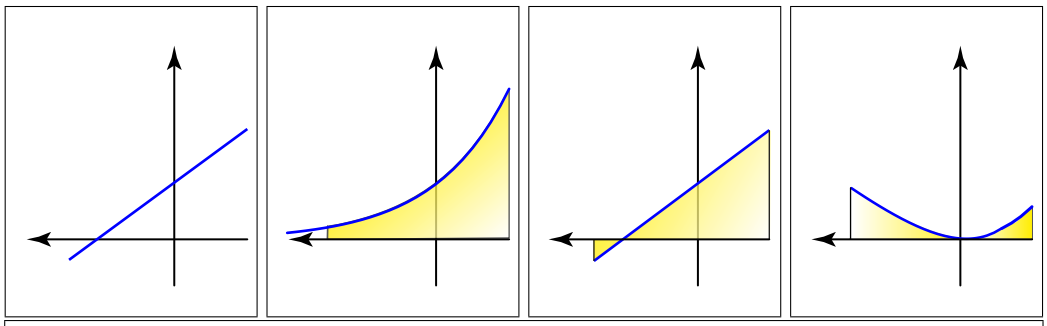


FIGURE 6.26: CONTROL VARIATES. Evaluation of the integral $\int_{[a,b]} e^x \, d\mu(x)$ with the control variate $1 + x$ according to Equation (6.455).

$$\begin{aligned}
\mathrm{Var}\left(f(\mathbf{U}_i) - g(\mathbf{U}_i)\right) &= \mathbb{E}\left(\left(f(\mathbf{U}_i) - g(\mathbf{U}_i)\right)^2\right) - \mathbb{E}^2\left(f(\mathbf{U}_i) - g(\mathbf{U}_i)\right) && (6.459)\\
&\leq \mathbb{E}\left(\left(f(\mathbf{U}_i) - g(\mathbf{U}_i)\right)^2\right) && (6.460)\\
&\overset{(2.732)}{=} \int_{\mathbf{I}^s} \left(f(\mathbf{U}(\omega)) - g(\mathbf{U}(\omega))\right)^2 \, d\mathbb{P}_{\mathbf{U}}(\omega) && (6.461)\\
&\overset{\|f-g\|_\infty < \tau}{<} \tau^2. && (6.462)
\end{aligned}$$

We conclude from this derivation that the variance of the estimator F_N^{CV} can be made arbitrary small by suited choice of the function g.

REMARK 6.22 *Another way to compute the estimator F_N^{CV} could be to construct two separate estimators*

$$F_N^{CV,1} = \frac{1}{N}\sum_{i=1}^{N} \frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)} \quad \textit{and} \quad F_N^{CV,2} = \frac{1}{N}\sum_{i=1}^{N} \frac{g(\mathbf{X}_i)}{p_{\mathbf{X}}} \tag{6.463}$$

and then to construct the estimator F_N^{CV} by:

$$\widehat{F}_N^{CV} \overset{(2.790)}{=} \int_{\mathbf{Q}^s} g(\mathbf{x}) \, d\mu^s(\mathbf{x}) + F_N^{CV,1} - F_N^{CV,2}. \tag{6.464}$$

Due to the correlated points in the sampling, the variance of F_N^{CV} is then given by:

$$\mathrm{Var}\left(\widehat{F}_N^{CV}\right) = \mathrm{Var}\left(F_N^{CV,1}\right) + \mathrm{Var}\left(F_N^{CV,2}\right) - 2\mathrm{Cov}\left(F_N^{CV,1}, F_N^{CV,2}\right), \tag{6.465}$$

which leads to variance reduction if it holds:

$$\mathrm{Cov}\left(F_N^{CV,1}, F_N^{CV,2}\right) > \frac{1}{2}\mathrm{Var}\left(F_N^{CV,2}\right). \tag{6.466}$$

Finally, let us shortly discuss the choice of the analytically square Lebesgue-integrable function g.

Let us assume g be proportional to the probability density function $p_{\mathbf{X}}$, that is $p_{\mathbf{X}}(\mathbf{x}) = Cg(\mathbf{x})$. Choosing C as the normalization constant for $p_{\mathbf{X}}$, thus

$$C = \frac{1}{\int_{\mathbf{Q}^s} g(\mathbf{x})\, d\mu^s(\mathbf{s})}, \tag{6.467}$$

then the function g is a good candidate for both importance sampling and as a control variate. As in this case the two estimators from Inequality (6.458) differ only by a constant, that is, their variance is the same, it makes no sense to use g as control variate if it was already used for importance sampling.

From another point of view, let us assume g be a good approximation to f. In this case, we must to decide whether to use it as a density function for importance sampling or as a control variate. [99, Kalos & Whitlock 1986] have shown that for a nearly constant quotient $\frac{f}{g}$, the function g should be used as a density for importance, while if $f - g$ is a nearly constant function, then g should be used as a control variate.

Remember, in the last section we have mentioned that the integrands in the light transport equation are algebraic terms of more than one function, often of the form $f(\mathbf{x}) = f_1(\mathbf{x}) \cdots f_3(\mathbf{x})$ or $f(\mathbf{x}) = f_1(\mathbf{x})(f_2(\mathbf{x}) + f_3(\mathbf{x})$ as well as $f(\mathbf{x}) = f_1(\mathbf{x})f_2(\mathbf{x}) + f_3(\mathbf{x}))$. We have also seen that it could be a good importance sampling strategy to generate a PDF that is proportional to a part of the integrand f, such as the BRDF, the cosine-factor, or the incident radiance distribution. Unfortunately, this strategy can not be transformed to the technique of control variates. Thus, it would not make sense in the above representations of the integrand f to use the function $f_1(\mathbf{x})$ alone as a control variate for f, since it is multiplied at least by a factor f_2, and the function f_1 must be replaced by $f_1(x) - 1$. Thus, a function g is only useful as a control variate if it takes into account all significant factors of f.

EXAMPLE 6.32 (Ambient Illumination) *Let us consider the stationary light transport equation in terms of exitant and incident radiance and assume that a known amount of ambient light* L_a *exists in the underlying scene. Due to [118, Lafortune & Willems 1994], [116, Lafortune 1996], it becomes possible to regard* L_a *as a control variate in the reflectance part of the SLTEV. This in turn implies:*

SLTEV (398) Radiance (250) Reflection Equation (321)

$$L_o(s,\omega_o) \overset{(4.89)}{=} L_e(s,\omega_o) + \int_{\mathcal{H}_i^2(s)} f_r(s,\omega_i \to \omega_o)\, L_i(s,\omega_i)\, d\sigma_s^\perp(\omega_i) \quad (6.468)$$

$$= L_e(s,\omega_o) + \int_{\mathcal{H}_i^2(s)} f_r(s,\omega_i \to \omega_o)\, L_a\, d\sigma_s^\perp(\omega_i) + \quad (6.469)$$

$$\int_{\mathcal{H}_i^2(s)} f_r(s,\omega_i \to \omega_o)\, \big(L_i(s,\omega_i) - L_a\big)\, d\sigma_s^\perp(\omega_i)$$

$$\overset{(4.162)}{=} L_e(s,\omega_o) + \pi\, \rho_{hd}(s)\, L_a + \quad (6.470)$$

$$\int_{\mathcal{H}_i^2(s)} f_r(s,\omega_i \to \omega_o)\, \big(L_i(s,\omega_i) - L_a\big)\, d\sigma_s^\perp(\omega_i).$$

Then, an associated secondary Monte Carlo estimator $F_N^{IMP,CV}$ *with* N*, according to the density* $f_r(s,\boldsymbol{\omega}_i \to \omega_o)\langle \mathbf{N}(s), \boldsymbol{\omega}_i^j\rangle$ *via importance sampling drawn samples* $\boldsymbol{\omega}_i^j$*, has the form:*

$$F_N^{IMP,CV} = L_e(s,\omega_o) + \pi\, \rho_{hd}(s)\, L_a + \frac{1}{N}\sum_{j=1}^{N} \big(L_i(s,\boldsymbol{\omega}_i^j) - L_a\big). \quad (6.471)$$

Restricting to the reflectance equation the estimator $F_N^{IMP,CV}$ *can be also used to estimate the direct illumination at surface point* $\mathbf{s}$ *by:*

$$F_N^{IMP,CV} = \pi\, \rho_{hd}(s)\, L_a + \frac{1}{N}\sum_{j=1}^{N} \big(L_e(s,\boldsymbol{\omega}_i^j) - L_a\big), \quad (6.472)$$

where L_e *is the known emitted radiance, that is, the sum of the constant ambient term multiplied by* π *and the reflectance* rho_{hd} *and an averaged sum over the difference of the emitted radiance and the constant ambient illumination.*

6.6.4 STRATIFIED SAMPLING

In Section 6.6.2 we have see that an importance sampling strategy for estimating the integral

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}), \quad (6.473)$$

results in a Monte Carlo estimator

$$F_N^{IS} = \frac{1}{N}\sum_{I=1}^{N} \frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}, \quad (6.474)$$

where the samples $\mathbf{X}_i$ are independent and identically distributed random variables drawn according to a well chosen probability density function p. We have also learned that

importance sampling can lead to Monte Carlo estimators with small variance and which are fast to evaluate. But, the strategy of importance sampling, namely to choose random samples where the integrand is large, can also lead to badly covered domains of integration, such as for example the clumping of samples in a domain or regions that are not sampled. Increasing the number of samples can eventually solve this problems but amortize the efficiency of the estimator $\epsilon(F_N^{IS})$ at the expense of longer run times. An interesting technique that avoids these drawbacks of importance sampling and which also guarantees variance reduction in the resulting Monte Carlo estimators is *Stratified Sampling*.

$\epsilon(F_N)$ (554)

The idea of stratified sampling is to ensure that the chosen samples are well distributed over the integration domain, with no two sample points too close together and where no excessively large regions are not sampled. This reduces the clumping of samples in the integration domain, see Figure 6.27. This goal will be achieved by decomposing the integration domain $\mathbf{Q}^s$ of the above integral in n disjoint subdomains $\mathbf{Q}_i^s$, so-called *strata*. The additivity of the Lebesgue integral then allows with

$$\mathbf{Q}^s = \bigcup_{i=1}^{n} \mathbf{Q}_i^s, \tag{6.475}$$

the following representation of the integral from (6.473):

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \quad = \quad \sum_{i=1}^{n} \int_{\mathbf{Q}_i^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}). \tag{6.476}$$

Now, we can construct probability spaces $(\mathbf{Q}_i^s, \mathfrak{B}(\mathbf{Q}_i^s), \mathbb{P}_{\mathbf{X}_i})$ over the n strata $\mathbf{Q}_i^s$, similar to the method from Section 6.2, where the associated probability measures $\mathbb{P}_{\mathbf{X}_i}$ can be defined based on the PDF $p_{\mathbf{X}}$ via

Probability Space (163)

$$\frac{d\mathbb{P}_{\mathbf{X}_i}}{p_i} = \mu^s(\mathbf{Q}_i^s) \frac{d\mathbb{P}_{\mathbf{X}_i}}{p_{\mathbf{X}}}, \tag{6.477}$$

Replacing the Lebesgue measure in Equation (6.476) by these probability measures leads to:

Lebesgue Measure (75)
Probability Measure (80)

$$\begin{aligned}
\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \quad &= \quad \sum_{i=1}^{n} \int_{\mathbf{Q}_i^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) && (6.478)\\
&\overset{d\mu^s = \frac{d\mathbb{P}_{\mathbf{X}_i}}{p_i}}{=} \quad \sum_{i=1}^{n} \int_{\mathbf{Q}_i^s} \frac{f(\mathbf{X})}{p_i(\mathbf{X})}\, d\mathbb{P}_{\mathbf{X}_i}(\omega) && (6.479)\\
&\overset{p_i = \frac{p_{\mathbf{X}}}{\mu^s(\mathbf{Q}_i^s)}}{=} \quad \sum_{i=1}^{n} \mu^s(\mathbf{Q}_i^s) \int_{\mathbf{Q}_i^s} \frac{f(\mathbf{X})}{p_{\mathbf{X}}(\mathbf{X})}\, d\mathbb{P}_{\mathbf{X}}(\omega) && (6.480)\\
&= \quad \sum_{i=1}^{n} \mu^s(\mathbf{Q}_i^s) E\left(\frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}\right). && (6.481)
\end{aligned}$$

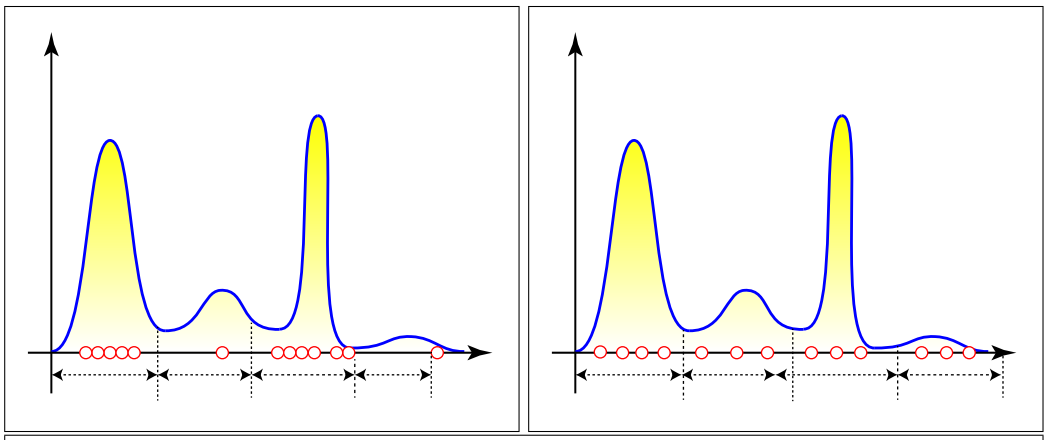


FIGURE 6.27: CLUMPING OF SAMPLES. The images show a high frequency function given over a set of four intervals of almost the same length. On the left, the samples are not well-distributed, since some of the intervals contain many more samples than others. Additionally, in intervals that contain many samples, these are clumped. In the right image the domain is better stratified. Here all intervals contain almost the same number of samples and the samples are even well-deistributed.

This shows, that the integral may be represented as sum of expected values of functions of independent and identically according to probability density $p_{\mathbf{X}}$ distributed random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$. If, in accordance with our preceding discussions, we now replace the expected values in Equation (6.481) by the secondary Monte Carlo estimators $F_N^{Q_i^s,\text{STRAT}}$, constructed on the strata $\mathbf{Q}_i^s$, then we get:

Expected Value of RV (196)
Probability Density (176)
Monte Carlo Estimator (499)

$$F_N^{Q_i^s,\text{STRAT}} = \frac{\mu^s(\mathbf{Q}_i^s)}{n_i} \sum_{j=1}^{n_i} \frac{f(\mathbf{X}_{i,j})}{p_{\mathbf{X}}(\mathbf{X}_{i,j})}, \tag{6.482}$$

where $\mathbf{X}_{i,j}$ are n_i samples from the strata $\mathbf{Q}_i^s$. That is, a Monte Carlo estimator F_N^{STRAT} defined on $\mathbf{Q}^s$ can be defined as follows:

DEFINITION 6.14 (Stratified Sampling) *Let $\bigcup_{i=1}^n \mathbf{Q}_i^s$ be the decomposition of the integration domain $\mathbf{Q}^s$, let furthermore $(\mathbf{Q}_i^s, \mathfrak{B}(\mathbf{Q}_i^s), \mathbb{P}_{\mathbf{X}_i})$ be probability spaces over the n strata $\mathbf{Q}_i^s$, then* stratified sampling *is the evaluation of the integral*

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \tag{6.483}$$

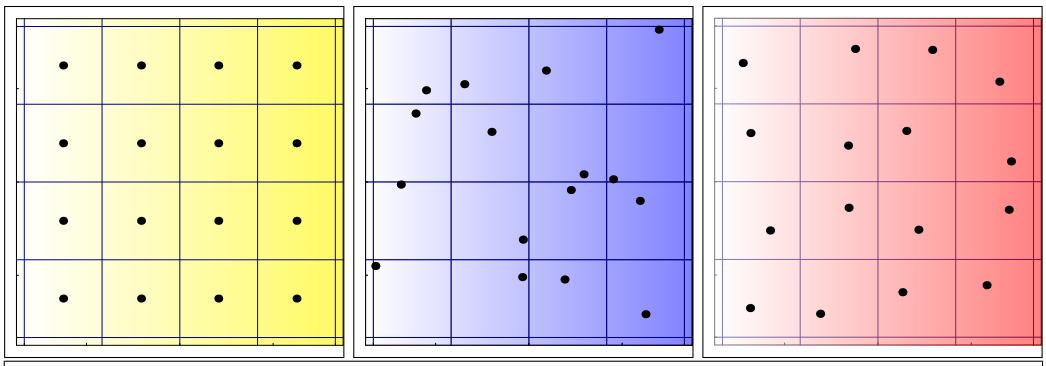


FIGURE 6.28: SUPERSAMPLING A PIXEL. A pixel is broken into a $n \times n$ grid, where a random point is chosen from each of these n^2 strata. The samples within the left image are chosen according to a regular grid at the midpoints within the strata. The samples within the image in the center result from uniformly distributed random variables on the strata. The pattern within the right image is a Poisson pattern, that is, a point is randomly drawn within a single strata only if its distance to the boundaries of the strata is greater than a predefined value.

via a secondary estimator F_N^{STRAT} *of the form*

$$F_N^{\text{STRAT}} = \sum_{i=1}^{n} F_N^{Q_i^s,\text{STRAT}} \tag{6.484}$$

$$= \sum_{i=1}^{n} \frac{\mu^s(Q_i^s)}{n_i} \sum_{j=1}^{n_i} \frac{f(X_{i,j})}{p_X(X_{i,j})}, \tag{6.485}$$

where $X_{i,j}, 1 \le i \le n, 1 \le j \le n_i$ *are identical and independent according to the probability density function* p_X *distributed random samples.*

Let us show by means of the following example, how this technique can be used in computer graphics.

EXAMPLE 6.33 (Supersampling a Pixel) *This procedure describes exactly the technique of supersampling a pixel in computer graphics. Here, the pixel* $\square_j$ *is broken into a* $n \times n$ *grid, where a random point is chosen from each of the* n^2 *strata, see Figure 6.28. Summing up the radiance values at these samples and subsequent averaging provides the irradiance at the corresponding pixel.* Irradiance (257)

Based on the Monte Carlo estimator from Equation (6.485) this trivial form of supersampling a pixel—broken in $N = n \times n$ *strata—can be estimated by the following*

stratified estimator:

$$F_N^{STRAT} = \sum_{i=1}^{N} \frac{f(\mathbf{X}_{i,1})}{p_i(\mathbf{X}_{i,1})} \tag{6.486}$$

$$= \sum_{i=1}^{N} \mu^2 \left(\frac{\square_j}{N} \right) \frac{f(\mathbf{X}_{i,1})}{p_X(\mathbf{X}_{i,1})}, \tag{6.487}$$

where the random variables $\mathbf{X}_{i,1}$ are independent and identically in each strata according to the probability density function p_X uniformly distributed samples.

Uniform Distribution (180)

Variance of a RV (201)

Let us now make a first statement on the variance of the stratified Monte Carlo estimator F_N^{STRAT} using n_i according to the PDF p_X-distributed random samples in each stratum. Obviously, it holds:

$$Var\left(F_N^{STRAT}\right) \overset{(2.772)}{=} \sum_{i=1}^{n} \left(\frac{\mu^s(\mathbf{Q}_i^s)}{n_i} \right)^2 \sum_{j=1}^{n_i} Var\left(\frac{f(\mathbf{X}_{i,j})}{p_X(\mathbf{X}_{i,j})} \right) \tag{6.488}$$

$$\overset{(2.772)}{=} \sum_{i=1}^{n} \frac{\mu^s(\mathbf{Q}_i^s)^2}{n_i} Var\left(\frac{f(\mathbf{X}_{i,1})}{p_X(\mathbf{X}_{i,1})} \right). \tag{6.489}$$

Note that in we did not take into account the size of the strata relative to each other as well as the number of samples per stratum in our formula for computing the variance of F_N^{STRAT}. It is also not easy to determine these degrees of freedom, such that the final variance is the smallest possible. But in the following theorem we will prove that the optimal number of samples in one stratum should be proportional to the variance of the estimator in the observed stratum relative to the variance of the estimator in all strata.

THEOREM 6.3 *Let us consider the stratified secondary Monte Carlo estimator F_N^{STRAT} given*

$$F_N^{STRAT} = \sum_{i=1}^{n} \frac{\mu^s(\mathbf{Q}_i^s)}{n_i} \sum_{j=1}^{n_i} \frac{f(\mathbf{X}_{i,j})}{p_X(\mathbf{X}_{i,j})} \tag{6.490}$$

based on the decomposition (6.475) of the integral from (6.473). Then, for the minimal variance of F_N^{STRAT} it holds:

$$\min Var\left(F_N^{STRAT}\right) = \frac{1}{N} \left(\sum_{i=1}^{n} \mu^s(\mathbf{Q}_i^s) \sqrt{Var\left(\frac{f(\mathbf{X}_{i,1})}{p_X(\mathbf{X}_{i,1})} \right)} \right)^2 \tag{6.491}$$

where, based on the assumption $\sum_{i=1}^{n} n_i = N$, the minimal variance occurs for

$$n_i = N \frac{\mu^s(\mathbf{Q}_i^s) \sqrt{Var\left(\frac{f(\mathbf{X}_{i,1})}{p_X(\mathbf{X}_{i,1})} \right)}}{\sum_{i=1}^{n} \mu^s(\mathbf{Q}_i^s) \sqrt{Var\left(\frac{f(\mathbf{X}_{i,1})}{p_X(\mathbf{X}_{i,1})} \right)}}. \tag{6.492}$$

PROOF 6.3 *The method of Lagrange multipliers allows us to maximize or minimize functions subject to a constraint, by solving a system of simultaneous equations*

$$\nabla f(x_1,\ldots,x_n) = \lambda\nabla^2 g(x_1,\ldots,x_n) \tag{6.493}$$
$$g(x_1,\ldots,x_n) = C, \tag{6.494}$$

where the variable λ is a dummy variable, called a Lagrange multiplier.

Setting f identical to the variance of $\mathrm{Var}\left(F_N^{STRAT}\right)$, thus

$$f(n_1,\ldots,n_n) \overset{(6.489)}{=} \sum_{i=1}^{n} \frac{(\mu^s(\mathbf{Q}_i^s))^2\ \mathrm{Var}\left(\frac{f(\mathbf{X}_{i,1})}{p_{\mathbf{X}}(\mathbf{X}_{i,1})}\right)}{n_i} \tag{6.495}$$

and choosing the constraint $g(n_1,\ldots,n_n)$ as

$$\sum_{i=1}^{n} n_i = N, \tag{6.496}$$

then differentiation with respect to variable n_i leads to the following system of equations:

$$-\frac{|\mu^s(\mathbf{Q}_i^s)|^2\ \mathrm{Var}\left(\frac{f(\mathbf{X}_{i,1})}{p_{\mathbf{X}}(\mathbf{X}_{i,1})}\right)}{n_i^2} = \lambda^2 \tag{6.497}$$

for $i = 1,\ldots,n$.

This equation can now be solved for the variable n_i resulting in

$$n_i = \frac{|\mu^s(\mathbf{Q}_i^s)|\ \sqrt{\mathrm{Var}\left(\frac{f(\mathbf{X}_{i,1})}{p_{\mathbf{X}}(\mathbf{X}_{i,1})}\right)}}{\lambda} \qquad \text{for } i = 1,\ldots,n. \tag{6.498}$$

Using the constraint $N = \sum_{i=1}^{n} n_i$ then we get:

$$N = \sum_{i=1}^{n} \frac{|\mu^s(\mathbf{Q}_i^s)|\ \sqrt{\mathrm{Var}\left(\frac{f(\mathbf{X}_{i,1})}{p_{\mathbf{X}}(\mathbf{X}_{i,1})}\right)}}{\lambda} \tag{6.499}$$

and for the Lagrange multiplier λ it holds:

$$\lambda = \sum_{i=1}^{n} \frac{|\mu^s(\mathbf{Q}_i^s)|\ \sqrt{\mathrm{Var}\left(\frac{f(\mathbf{X}_{i,1})}{p_{\mathbf{X}}(\mathbf{X}_{i,1})}\right)}}{N}. \tag{6.500}$$

Inserting the expression for λ in Equation (6.498) leads to

$$n_i = N\frac{|\mu^s(\mathbf{Q}_i^s)|\ \sqrt{\mathrm{Var}\left(\frac{f(\mathbf{X}_{i,1})}{p_{\mathbf{X}}(\mathbf{X}_{i,1})}\right)}}{\sum_{i=1}^{n} |\mu^s(\mathbf{Q}_i^s)|\ \sqrt{\mathrm{Var}\left(\frac{f(\mathbf{X}_{i,1})}{p_{\mathbf{X}}(\mathbf{X}_{i,1})}\right)}} \tag{6.501}$$

Via Relation (6.501) then it is straightforward to proof Formula (6.491) for $\min \mathrm{Var}\left(F_N^{STRAT}\right)$. *We leave this easy exercise to the interested reader.*

The statement of the above theorem can be interpreted in such a way that for prescribed strata the variance of the estimator is minimal when the number of samples in strata $\mathbf{Q}_i^s$ is proportional to $\mu^s(\mathbf{Q}_i^s)\sqrt{\mathrm{Var}\left(F_N^{STRAT}\right)}$.

Let us now compare the variance of the stratified Monte Carlo estimator $\mathrm{Var}\left(F_N^{STRAT}\right)$ against its unstratified version. For that purpose, let us suppose the sample size n_i in $\mathbf{Q}_i^s$ is proportional to the volume of the corresponding stratum, i.e., $n_i = \mu^s(\mathbf{Q}_i^s)\,N$, where N is the total number of samples. Then, we obtain:

$$\mathrm{Var}\left(F_N^{STRAT}\right) = \frac{1}{N}\sum_{i=1}^{n}\mu^s(\mathbf{Q}_i^s)\mathrm{Var}\left(\frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}\right), \tag{6.502}$$

where $\mathbf{X}_i$ are independent and identically $p_{\mathbf{X}}$-distributed random variables.

In [220, Veach 1997] it is shown that the variance of the unstratified Monte Carlo estimator can be represented as the mean of the individual variances plus the variance of the means, that is it holds:

$$\begin{aligned}&\mathrm{Var}\left(F_N^{UNSTRAT}\right) \qquad (6.503)\\ &= \frac{1}{N}\left(\sum_{i=1}^{n}\mu^s(\mathbf{Q}_i^s)\mathrm{Var}\left(\frac{f(\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}\right) + \sum_{i=1}^{n}\mu^s(\mathbf{Q}_i^s)\left(\bar{m}_{\mathbf{Q}_i^s} - \bar{m}_{\mathbf{Q}^s}\right)^2\right),\end{aligned}$$

where $\mathbf{X}_{i,j}$ are independent and identically $p_{\mathbf{X}}$-distributed random samples on $\mathbf{Q}^s$, $\bar{m}_{\mathbf{Q}_i^s}$ is the mean value of f in the strata $\mathbf{Q}_i^s$, thus,

$$\bar{m}_{\mathbf{Q}_i^s} \stackrel{def}{=} \frac{1}{\mu^s(\mathbf{Q}_i^s)}\int_{\mathbf{Q}_i^s} f(\mathbf{x})\,d\mu^s(\mathbf{x}) \tag{6.504}$$

and $\bar{m}_{\mathbf{Q}^s}$ the mean value of f over the whole domain of integration, that is,

$$\bar{m}_{\mathbf{Q}^s} \stackrel{def}{=} \frac{1}{\mu^s(\mathbf{Q}^s)}\int_{\mathbf{Q}^s} f(\mathbf{x})\,d\mu^s(\mathbf{x}). \tag{6.505}$$

Now, from Equation (6.503) we can conclude that—under the assumption that the sample size n_i is proportional to the volume of the corresponding stratum, i.e., $n_i = \mu^s(\mathbf{Q}_i^s)\,N$—the variance of the non-stratified estimator can never be smaller than the one constructed with the help of stratified sampling. However, in stratified sampling variance reduction can only be achieved, if the strata $\mathbf{Q}_i^s$ have different means, that is, in a stratified sampling algorithm, the strata should be chosen such that their means are different as possible, at least when the number of samples in each stratum is proportional to its volume. In [158, Pharr & Humphreys 2004] it is suggested to choose the strata as

compact as possible if nothing is known about the integrand f. If the strata are wide, they will contain more variation and will have means $\bar{m}_{\mathbf{Q}_i^s}$ closer to the true mean $\bar{m}_{\mathbf{Q}^s}$.

We can undoubtedly say that stratified sampling his one of the most promising variance reduction technique in Monte Carlo integration which is also of great importance in the field of Monte Carlo rendering. The method works very well when the number of samples required is known in advance and the dimension of the underlying problem is relatively low, that is, $s < 20$.

REMARK 6.23 *A problem that comes with stratified sampling is the partition of the domain of integration into strata. Thus, the decomposition of* $\mathbf{Q}^s$ *into strata of equal size results in a minimum of* 2^s *strata—one split in each dimension provided. This corresponds to an enormous number of strata, and thus to an enormous number of samples which must be drawn. Here, several techniques—such as* latin hypercube sampling, orthogonal array sampling, *and several* quasi-Monte Carlo methods*—exist that can remedy this problem.*

Section 6.6.5
Section 6.6.7
Chapter 7

Let us finally present some different techniques that can be helpful for stratifying the domains of integration which are subject to the light transport equations of global illumination theory.

EXAMPLE 6.34 (Voronoi-Diagrams) *Let us consider domains of integration of the form* $\mathbf{Q}^2 = \mathbf{I}^2$ *as they occur in pixel sampling. If we construct the so-called* Hammersley point set*, that is, an* N*-element set of 2-dimensional points* $\mathbf{P}_N = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$*, then we can observe that the distance between two element* $\mathbf{x}_i, \mathbf{x}_j$ *from* $\mathbf{P}_N$ *satisfies the condition*

Hammersley point set (634)

$$\|\mathbf{x}_i - \mathbf{x}_j\|_2 \geq \frac{1}{b^n}, \tag{6.506}$$

whereas $N = b^n$ *with* $b \in \mathbb{N}, b \geq 2$.

Now, the Voronoi area of a point $\mathbf{x}_i$ *in the unit square, denoted as* $\mathrm{Vor}(\mathbf{x}_i)$*, is defined by*

$$\mathrm{Vor}(\mathbf{x}_i) = \{\mathbf{x} \in \mathbf{I}^2 \,|\, \|\mathbf{x}_i - \mathbf{x}\|_2 \leq \|\mathbf{x}_k - \mathbf{x}\|_2, i \neq k, 1 \leq k \leq N\}. \tag{6.507}$$

Thus, the set $\mathbf{P}_N$ *implies a disjoint partition of* $\mathbf{I}^2$ *in the* N *convex Vornoi areas* $\mathrm{Vor}(\mathbf{x}_i), 1 \leq i \leq N$*, namely:*

$$\mathbf{I}^2 = \bigcup_{i=1}^{N} \mathrm{Vor}(\mathbf{x}_i), \tag{6.508}$$

which yields an implicit stratification of roughly equally sized integration domains [67, Glassner 1995]. Figure (6.29) illustrates the above partition for the 4*-element, 2-dimensional Hammersley point set* $\mathbf{P}_4$ *whose characteristics will be discussed in more detail when dealing with quasi-Monte Carlo algorithms.*

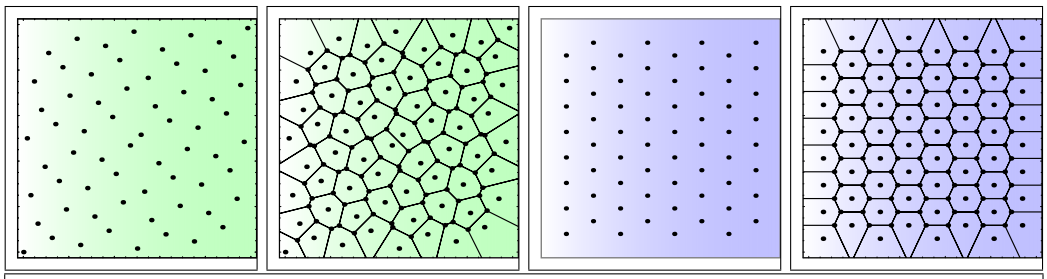


FIGURE 6.29: STRATIFICATION OF I^2 WITH VORONOI DIAGRAMS. Left, a 64-element Hammersley point set. The second image shows a Voronoi diagram implied from the Hammersley point set. Next, a 64-element hexagonal grid with the associated Voronoi diagram.

EXAMPLE 6.35 (Stratification of $[0,1)^2$) *Let us consider the half-open unit square* $[0,1)^2$, *then the set* $\mathcal{E} = \prod_{j=1}^{2} \left[\frac{a_j}{b^{l_j}}, \frac{a_j+1}{b^{l_j}}\right)$, *with* $l_j \geq 0$ *and* $0 \leq a_j < b^{l_j}$ *describes the* 2-dimensional elementary intervals *of* $[0,1)$ *with respect to the base* b *with Lebesgue measure* $\mu^s(\mathcal{E}) = \prod_{j=1}^{2} \frac{1}{b^{l_j}} = \frac{1}{b^{\sum_{j=1}^{2} l_j}}$. *Figure (6.30) demonstrates possible stratifications of* $[0,1)^2$ *into all elementary intervals with Lebesgue measure* $\frac{1}{32}$.

Lebesgue Measure (82)
Elementary Intervals (641)
(t,m,s)-net (641)

EXAMPLE 6.36 (Polar and Concentric Map) *Normally, implementations for sampling camera lenses are based on naive transformations, which maps points sampled within the unit square* $\mathbf{I}^2$ *onto points within the unit circle* $\mathbf{C}$. *Indeed, these mappings lead to useable sampling patterns, but the transformation distorts several beautiful properties of the original sampling patterns. For example, in [158, Pharr & Humphreys 2004] it is shown that a stratified sampling pattern on the unit square is mapped to an unstratified pattern on the unit circle with less compact strata away from the center. To preserve* good *sampling properties, we should be interested in constructing mappings that are robust with respect to the characteristics of the pattern which should be transformed.*

Thus for example, in [107, Kolb 1995 & al.] and [185, Shirley 2000] it is required that a mapping $\mathbf{T} : \mathbf{I}^2 \to \mathbf{C}$ *should not only fulfill the condition of area preservation, i.e.,*

$$\frac{\mu^2(\mathbf{A})}{\mu^2(\mathbf{I}^2)} = \frac{\mu^2(\mathbf{T}(\mathbf{A}))}{\mu^2(\mathbf{D})}, \tag{6.509}$$

with $\mathbf{A} \subset \mathbf{I}^2$, *as well as the continuity of* $\mathbf{T}$ *and* $\mathbf{T}^{-1}$, *but the mapping should also satisfy the property of form retention. In [185, Shirley 2000] it is shown that an area preserved transformation maps a point set drawn on* $\mathbf{I}^2$ *onto a similar one on* $\mathbf{C}$ *and that the continuity- and form-preserving characteristics will lead to hold the neighborhood relations between point pairs. Thus, Figure 6.31 demonstrates stratifications*

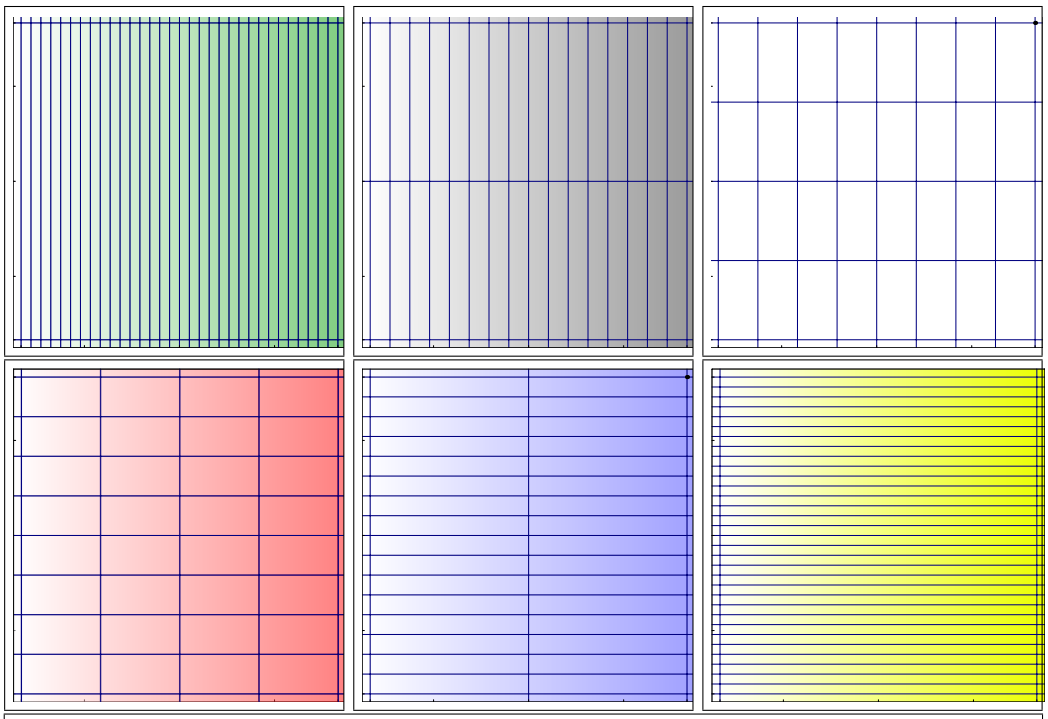


FIGURE 6.30: STRATIFICATION OF I^2 USING ELEMENTARY INTERVALS. The unit square stratified with elemantary intervals in intervals with Lebesgue measure $\frac{1}{32}$.

of the unit circle incurred by different parameterizations, where the concentric map is formed through the rotation and reflection of the mapping $\begin{pmatrix} x \\ \frac{\pi y}{4x} \end{pmatrix}, x, y \in \mathbf{I}^2$.

REMARK 6.24 (Stratified Sampling on $\mathbf{I}^s$) *Applying the technique of stratified sampling to the integral given in (6.473), where the integration domain is the s-dimensional unit cube, leads with independent, identically, and uniformly distributed samples* Uniform Distribution (180)
$\mathbf{U}_{i,1}, \ldots, \mathbf{U}_{i,N}$, $1 \leq i \leq N$ *from* $\mathbf{I}^s = \bigcup_{i=1}^{N} \mathbf{Q}_i^s$ *to the following form of the secondary Monte Carlo estimator* $F_2^{I^s,STRAT}$, *well-known from a number of works:*

$$F_N^{STRAT} = \sum_{i=1}^{n} F_N^{Q_i^s,STRAT} = \sum_{i=1}^{n} \frac{\mu^s(\mathbf{Q}_i^s)}{n_i} \sum_{j=1}^{n_i} f(\mathbf{U}_{i,j}). \tag{6.510}$$

6.6.5 LATIN HYPERCUBE SAMPLING

As already mentioned above, stratified sampling is mainly effective for low-dimensional Section 6.6.4

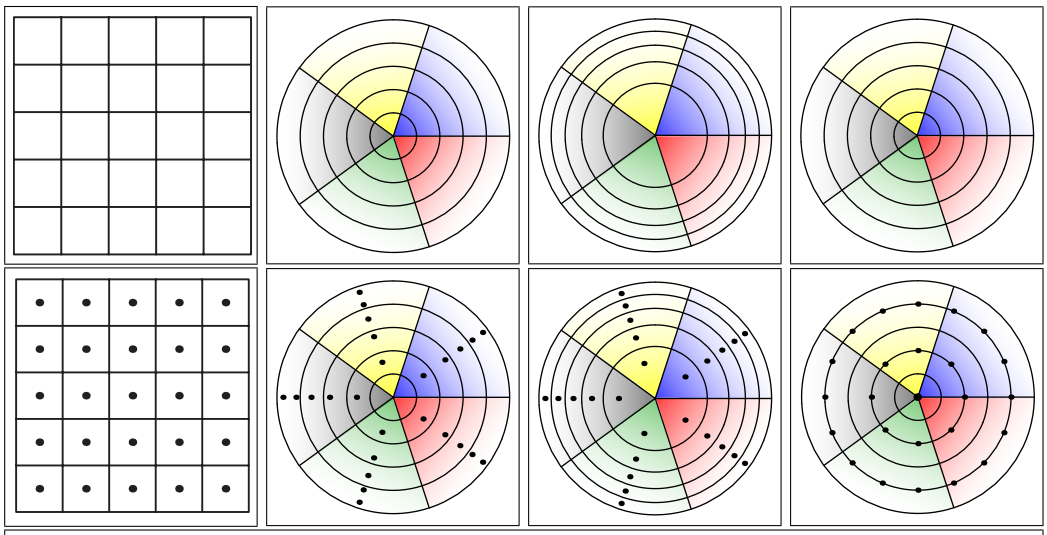


FIGURE 6.31: CARTESIAN, POLAR AND CONCENTRIC MAPS. Cartesian maps and their polar and concentric maps according to the parameterizations $\begin{pmatrix} r \\ \phi \end{pmatrix} = \begin{pmatrix} u_1 \\ 2\pi u_2 \end{pmatrix}$, $\begin{pmatrix} \sqrt{u_1} \\ 2\pi u_2 \end{pmatrix}$, with $u_1, u_2 \in I$, and $\begin{pmatrix} v_1 \\ \frac{\pi v_2}{4v_1} \end{pmatrix}$, $v_1, v_2 \in [-1, 1]$, as examples for the stratification of the unit circle.

integration problems, where the integrand is well-behaved. Naively extended to high-dimensional integration domains, stratified sampling is, because of strong increases in the sampling rate, not very helpful in the avoidance of clumping—the phenomenon that a large number of points belonging to a set may be packed into little balls. For a s-dimensional function the number of samples required is N^s, which can prohibitive for large values of s. Here a procedure related to stratified sampling is more useful: *Latin Hypercube Sampling*.

Latin hypercube sampling, or briefly *LHS*, slightly modified also known as N-*rooks sampling* [181, Shirley 1990], is based on solving a problem in chess, namely, that N rooks placed on a checker board must not defeat each other in one move.

Extended to s dimensions, this method suggests to transform the original integration domain $\mathbf{Q}^s$ onto the unit cube $\mathbf{I}^s$ and then to split $\mathbf{I}^s$ into N subintervals along the s coordinate axes, where each of these N intervals has Lebesgue measure $\frac{1}{N}$ and contains exactly one sample.

Lebesgue Measure (82)

DEFINITION 6.15 (Latin Hypercube Sampling, LHS) *Let $\mathbf{I}^s$ be the s-dimensional unit cube. Then,* Latin hypercube sampling, *also called as* LHS, *is denoted as the sampling method, where, based on* s N*-element permutations $\pi_1, \ldots, \pi_s$ and uniformly*

Uniform Distribution (180)

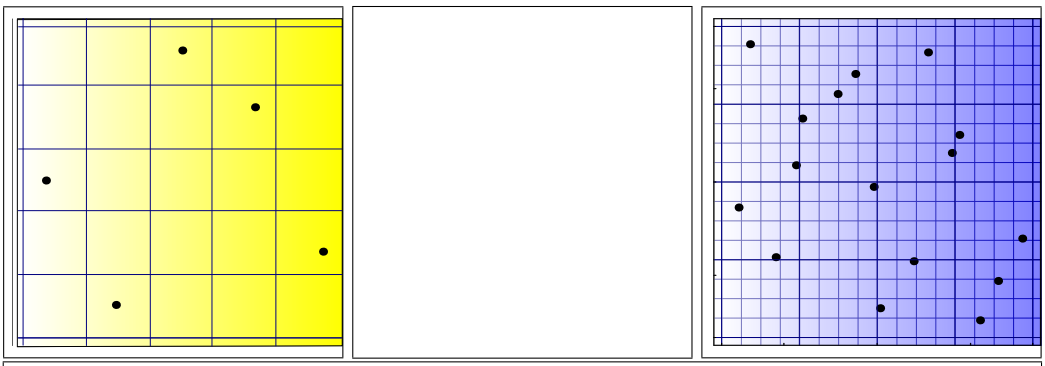


FIGURE 6.32: LATIN HYPERCUBE SAMPLING. On the left, a 5-element point set sampled with Latin Hypercube sampling with $s = 2, N = 5$ the associated permutations are given by: $\pi_1 = (1,2,3,4,5)$ and $\pi_2 = (3,1,5,4,2)$. The right image shows a 15-element point set sampled with Latin Hypercube sampling.

distributed random variables U_{ij} *on* $[0,1]$, *sample locations* $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,s})$ *given by*

$$X_{i,j} \stackrel{def}{=} \frac{\pi_j(i) + U_{ij}}{N}, \qquad 1 \leq j \leq s \tag{6.511}$$

where $X_{i,j}$ *denotes the* j^{th} *coordinate of sample* $\mathbf{X}_i$.

This construction results in a uniform probability distribution in each of the subintervals, and therefore in each s-dimensional subcube. For more information on LHS, in particular a discussion on variance analysis of LHS, see [220, Veach 1997].

EXAMPLE 6.37 *Visualization of 5-rooks sampling in the square* $\mathbf{Q}^2 = [2,3] \times [1,2]$ *is shown in part a) of Figure 6.32. The chosen permutations are* $\pi_1 = (1,2,3,4,5)$ *and* $\pi_2 = (3,1,5,4,2)$. *Based on these choice, the resulting samples lie in the subsquares* $(1,3), (2,1), (3,5), (4,4), (5,2)$.

EXAMPLE 6.38 (Raytracer Conception) *One typical application area of Latin hypercube sampling is found in the implementation of modern ray tracers. Apart from conventional ray distribution over pixels, the tracing of reflective and refractive rays, as well as the sampling of light sources, modern ray tracers also solve problems such as aliasing, depth of field, and motion blur. This means that in modern ray tracer there must exist routines for stratifying pixels, camera lenses, as well as, the time. Now, sampling in higher dimensions requires generating* s^k *rays per pixel for* k*-stratified dimensions. One possibility to avoid this undesired effect is to use the technique of LHS, where only* s *of the originally* s^k *rays must be generated to shade a pixel, see*

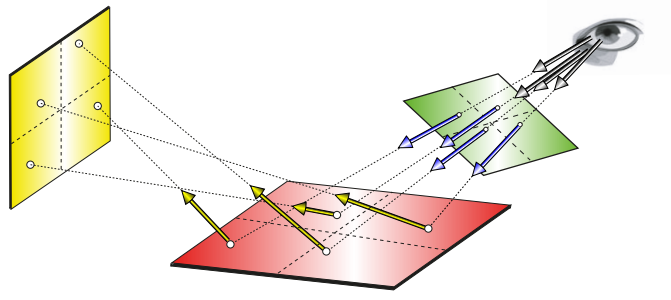


FIGURE 6.33: LATIN HYPERCUBE SAMPLING FOR RAY TRACER CONCEPTION. Latin hypercube sampling for pixel-, lens- and light source sampling used in the conception of a simple ray tracer.

Figure 6.33.

REMARK 6.25 *LHS requires storing the involved permutations, which in s-dimensional cases implies a storage capacity of* $O(sN)$. *It also has the disadvantage, that the quality of the generated pattern is highly dependent on the choice of the permutations used. This is the reason why Latin hypercube sampling is not appropriate for the solution of a number of integrals given here, under the condition of highly sampled integration domains [104, Keller 1998].*

6.6.6 JITTERED SAMPLING

Let us now take a look at another method of variance reduction: *Jittered Sampling*. Jittered sampling may be regarded as a two-step sampling technique. In the first step, jittered sampling constructs a uniform stratification of the underlying integration domain, resulting in a regular or hexagonal grid, a N-rooks pattern, respectively a Voronoi diagram. Then, the method chooses a single sample in every strata and moves this sample with the help of a randomly selected displacement [183, Shirley 1991]. Figure 6.34 shows a point set based on a regular grid with the associated pattern generated by jittering.

n-rooks Sampling (580)
Voronoi Diagram (637)

[183, Shirley 1991] presents a variant of jittered sampling, *half jittered sampling*, see the right image of Figure 6.34. Instead of moving samples in the Voronoi area allocated to the grid point a sample is placed only in a square selected around the grid point.

According to [104, Keller 1998]], due to its dimension-dependent characteristics, jittered sampling has been found to be a less appropriate sampling technique for larger dimensions s with $s > 6$.

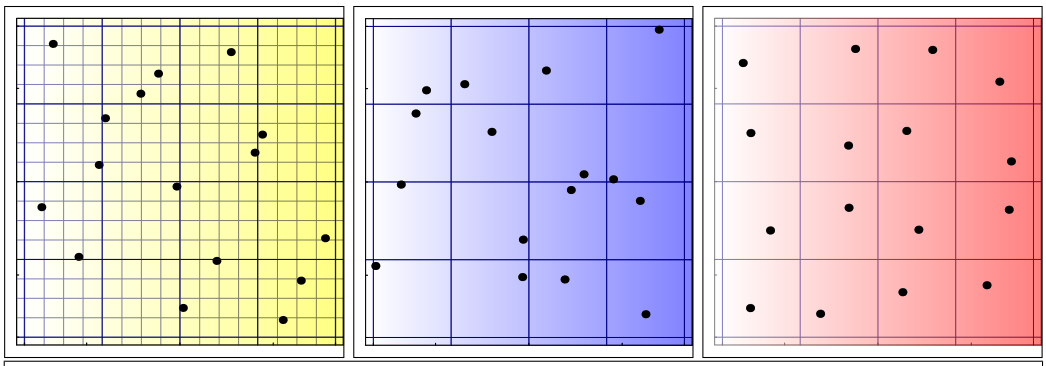


FIGURE 6.34: JITTERED SAMPLING. First, the underlying domain is stratified. Then the algorithm chooses a single sample in every strata and moves this sample with the help of a randomly selected displacement [183, Shirley 1991]. Figure 6.34 shows a point set based on a regular grid together with its pattern generated with jittering. Instead of moving samples in the Voronoi area allocated to the grid point a sample is placed only in a square selected around the grid point

REMARK 6.26 *If we turn once more to the estimator* $F_N^{\mathbf{I}_i^s,\mathrm{STRAT}}$ *from Equation (6.510). Due to the decomposition of* $\mathbf{I}^s = \bigcup_{i=1}^N \mathbf{I}_i^s$*, with Lebesgue measure* $\mu^2(\mathbf{I}_i^s) = \frac{1}{N}$*, a jittered-Monte Carlo estimator based on stratified sampling is given by:*

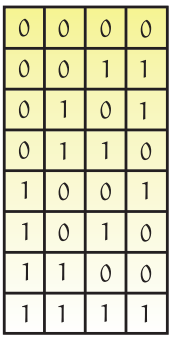
$$F_N^{\mathrm{JIT}} = \sum_{i=1}^{N} F_N^{\mathbf{I}_i^s,\mathrm{STRAT}}. \tag{6.512}$$

6.6.7 ORTHOGONAL ARRAY SAMPLING

Orthogonal Array Sampling [143, Owen 1992] is a generalized type of Latin hypercube sampling that shares certain desirable properties with LHS and but additionally possess desirable statistical features. It returns samples that are well-distributed with respect to any combination of two, three, or more dimensions. Section 6.6.5

DEFINITION 6.16 (Orthogonal Array) *An* orthogonal array $\mathrm{OA}(N, s, b, t)$ *is a* $N \times s$ *matrix, whose coefficients are drawn from an alphabet of size* b*, such that every* $N \times t$ *submatrix contains exactly the same number of rows with the same permutation of elements. Let* λ *be the number of times that each row appears in the submatrix* $N \times t$*, then it is obviously to see that the total number of rows in an orthogonal array is* $N = \lambda b^t$*. The parameter* λ *is often called the* index *of* OA *and* t *is denoted as the* strength *of the orthogonal array.* Matrix (853)

EXAMPLE 6.39 *Figure 6.35 shows an orthogonal array* $\mathrm{OA} = (8, 4, 2, 3)$*. Obviously, it*



0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1

FIGURE 6.35: ORTHOGONAL ARRAY SAMPLING. $OA = (8, 4, 2, 3)$ constructed over the alphabet $\{0, 1\}$. Every permutation of length 3 consisting of coefficients of $\{0, 1\}$ is contained exactly once in a 8×3 submatrix.

is a 8×4 *matrix with coefficients from the alphabet* $\{0, 1\}$ *where every permutation of* $\{0, 1\}$ *appears* $\lambda = 1$ *times in every submatrix constructed from* 8 *rows and* 3 *columns.*

How can we use such an array for sampling?

Now [220, Veach 1997] argues as follows: Let OA be an $N \times s$ orthogonal array of strength t, whose coefficients are drawn from the alphabet $\{0, 1, \ldots, b-1\}$. In a first step, we randomize the array OA using a permutation π_j of the given alphabet to each column, i.e., we obtain a new orthogonal array

$$\widehat{OA}_{i,j} \stackrel{def}{=} \pi_j(OA_{i,j}) \quad \forall\, i, j, \tag{6.513}$$

with the same parameters as the original one, where $\pi_1, \ldots, \pi_s$ are random permutations of the symbols $\{0, 1, \ldots, b-1\}$. Obviously this initial step guarantees, that each of the b^s possible rows occurs in $\widehat{OA}$ with equal probability.

Let us assume that the domain is the unit cube $[0, 1]^s$. By splitting each of its edges into b intervals with Lebesgue-measure $\frac{1}{b^s}$, we obtain a set of b^s s-dimensional subcubes, where each $1 \times s$ row of $\widehat{OA}$ can be interpreted as an index into this set of subcubes. The idea behind orthogonal array sampling is to generate one sample $\mathbf{X}_i = (X_{i_1}, \ldots, X_{i_s})$ in each of the N subcubes specified by the rows of $\widehat{OA}$, where the j-th coordinate of $\mathbf{X}_i$ is given by

Lebesgue Measure (75)

$$X_{i,j} \stackrel{def}{=} \frac{\widehat{OA}_{ij} + U_{i,j}}{b}, \tag{6.514}$$

and $U_{i,j}$ are uniformly distributed random variables on $[0, 1]$. As the sample $\mathbf{X}_i$ is uniformly distributed in $[0, 1]^s$, an unbiased estimator for the integral from (6.92) is given

Uniform Distribution (180)
Unbiased (507)
Monte Carlo Estimator (499)

by

$$F_N^{OA} = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{X}_i). \tag{6.515}$$

If we then consider the projection of these samples onto the subspace spanned by t coordinate axes, the main property of orthogonal array sampling ensures that these samples will be uniformly distributed over the b^t subcubes which are obtained by stratifying each of the t axes into intervals of Lebesgue-measure $\frac{1}{b^t}$. To see this, observe that the coordinates of the projected samples are specified by the rows of a particular $N \times t$ submatrix. Due to the definition of an orthogonal array, each of the possible b^t rows occurs λ times in this submatrix, so that each of the b^t subcubes occurs exactly λ times. Therefore, orthogonal array sampling generates samples that are stratified with respect to every possible subset of t coordinates.

Due to [143, Owen 1992], [144, Owen 1994], the variance of orthogonal array sampling can be estimated by

$$Var\left(F_N^{OA}\right) = \frac{1}{N} \sum_{|u|>t} \int f_u^2(x^u)\, d\mu(x) + o\left(\frac{1}{N}\right) \quad \text{with } u \subseteq \mathfrak{P}(\{1,\ldots,s\}), \tag{6.516}$$

and $f(x) = \sum_u f_u(x^u)$, i.e., the convergence rate is improved with respect to all components of the integrand that depend on t coordinates or less.

Of special interest for graphics is the case $t = 2$. Applied to distribution ray tracing orthogonal array sampling ensures that all 2-dimensional projections are well stratified over the pixel, lens aperture, light source, etc..

Distribution Ray Tracing (672)

6.6.8 ANTITHETIC VARIATES

Usually Monte Carlo methods use random points which are drawn independent of each other. The procedure, which we now introduce, deliberately makes use of correlated samples taking advantage of the fact that such a correlation may be negative. The idea of *Antithetic Variates*, one of the simplest and widely used methods of reducing variance, is to find an estimator F'_N having the same expected value as the estimator F_N from Equation (6.109) but which is strongly negatively correlated.

Independent RV (204)

Let us consider the Monte Carlo estimator

Monte Carlo Estimator (499)

$$F_N^{AV} \stackrel{def}{=} \frac{1}{2}\left(F_N + F'_N\right) \tag{6.517}$$

for approximating the integral

$$\int_{\mathbf{I}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) = \int_{\mathbf{I}^s} \frac{1}{2}\left(f(\mathbf{x}) - f(1-\mathbf{x})\right)\, d\mu^s(\mathbf{x}), \tag{6.518}$$

where f is a linear, real-valued function.

Obviously, F_N^{AV} has the same expected value as F_N and F'_N, i.e., if F_N is unbiased, then F_N^{AV} is also unbiased. We leave the proof of this statement to the interested reader as an exercise.

Expected Value (196)
Unbiasedness (507)

Now, our focus is on the variance of the estimator F_N^{AV}. If we can arrange to sample points such that F_N and F'_N are sufficiently negatively correlated, then it holds:

Variance (201)
Covariance (203)

$$\mathrm{Var}\left(F_N^{AV}\right) \overset{(2.785)}{=} \frac{1}{4}\mathrm{Var}\left(F_N\right) + \frac{1}{4}\mathrm{Var}(F'_N) + \frac{1}{2}\mathrm{Cov}(F_N, F'_N), \tag{6.519}$$

i.e., the estimator F_N^{AV} will have a lower variance as F_N. Assuming that F'_N has the same variance as F_N then we may write

$$\mathrm{Var}\left(F_N^{AV}\right) = \frac{1}{2}\left(\mathrm{Var}(F_N) + \mathrm{Cov}(F_N, F'_N)\right), \tag{6.520}$$

which—under the condition that $\mathrm{Cov}(F_N, F'_N)$ is strongly negative correlated—implies, that the combined estimator $\mathrm{Var}\left(F_N^{AV}\right)$ has lower variance as the original one.

EXAMPLE 6.40 (Sampling on the Unit Cube) *The simplest way to construct the estimator F'_N on the unit cube $[0,1]^s$ is to use pairs of sample points $(\mathbf{X},\mathbf{Y})$ of the form $\mathbf{X} = (U_1,\ldots,U_s)$ and $\mathbf{Y} = (1-U_1,\ldots,1-U_s)$. If the integrand of the Integral from (6.518) is a monotone function, then $f(\mathbf{X})$ is large where $f(\mathbf{Y})$ is small and vice versa, i.e., the variations will largely cancel each other and the integral from Relation (6.518) can be written as*

$$\int_{[0,1]^s} \frac{1}{2}\big(f(\mathbf{x}) + f(1-\mathbf{x})\big)\, d\mu^s(\mathbf{x}). \tag{6.521}$$

Obviously, an associated secondary Monte Carlo estimator has the simple form

$$F_N^{AV} = \frac{1}{2N}\sum_{i=1}^{N}\left(f(\mathbf{X}_i) + f(1-\mathbf{X}_i)\right). \tag{6.522}$$

REMARK 6.27 *Antithetic variates can also easily be combined with other methods for reducing the variance of a Monte Carlo estimator such as stratified or importance sampling. Combined with importance sampling, the estimator F_N^{AV} looks like this:*

$$F_N^{IS,AV} = \frac{1}{2N}\sum_{i=1}^{N}\frac{f(\mathbf{X}_i) + f(1-\mathbf{X}_i)}{p_{\mathbf{X}}(\mathbf{X}_i)}, \tag{6.523}$$

where $p_{\mathbf{X}}$ is a well chosen PDF used for importance sampling, and applied to stratified sampling, where the domain of integration $\mathbf{I}^s$ is decomposed in the union of the n disjoint $\mathbf{I}_1^s,\ldots\mathbf{I}_n^s$, we get:

$$F_N^{STRAT,AV} = \sum_{i=1}^{n}\frac{\mu^s(\mathbf{I}_i^s)}{2n_i}\sum_{j=1}^{n_i}\frac{f(\mathbf{X}_{i,j}) - f(1-\mathbf{X}_{i,j})}{p_{\mathbf{X}}(\mathbf{X}_{i,j})}. \tag{6.524}$$

REMARK 6.28 *With respect to integral equations underlying the global illumination problem the technique of antithetic variates is of limited usefulness, since variance comes mainly from the discontinuities and singularities of the underlying integrands, so that variance improvements on smooth regions of the integrands are rejected.*

6.6.9 MULTIPLE IMPORTANCE SAMPLING

All of our sampling strategies in Monte Carlo integration, introduced until to now, use a single probability density function to generate samples from given integrands. Now, the integrands involved in the SLTEV are usually complex, as they depend on material properties of the objects in the scene, the scene geometry, or since they are often mathematical expressions of the form

Probability Density Function (176)
SLTEV (398)

$$f(\mathbf{x}) = f_1(\mathbf{x})f_2(\mathbf{x}) + f_3(\mathbf{x}) \quad (6.525)$$
$$f(\mathbf{x}) = (f_1(\mathbf{x}) + f_2(\mathbf{x}))f_3(\mathbf{x}) \quad (6.526)$$
$$f(\mathbf{x}) = f_1(\mathbf{x})f_2(\mathbf{x})f_3(\mathbf{x}), \quad (6.527)$$

which depend on parameters whose values are not known in advance. In all those circumstances, it is difficult to design a single efficient sampling strategy that works well.

One hitherto common approach to solve the problem was, to partition the domain of integration into several regions, and to design separate sampling strategy for each region, or simply to ignore some of the unknown components and to design sampling strategies only for known components. Such an approach will be presented in Section 9.1 when introducing *Monte Carlo path tracing*, where we use different sampling techniques for the evaluation of direct and indirect lighting, as well as glossy, diffuse, or specular reflections.

Glossy Reflection (304)

Now, since a best sampling strategy cannot be chosen, in [222, Veach and Guibas 1995] and [221, Veach 1998] the proposal is made, instead to concoct a good sampling strategy from several bad sampling strategies, to combine the sample values in a provably good way, so that the strengths of each sampling strategy is preserved. The idea behind this strategy is to draw samples from multiple distributions in the hope that at least one of these distributions will match the shape of the integrand in a reasonable manner.

6.6.9.1 THE GLOSSY HIGHLIGHTS PROBLEM

Let us consider the rendering of a glossy surface illuminated by a nearby area light source $\text{☼} \in \partial\mathcal{V}$. Since our scene only exists of a light source and a surface, the scattering part in the stationary light transport equation, integrated over the unit sphere S^2, has the form

Scattering Equation (374)

$$\int_{S^2(\mathbf{s}_j)} f_s(\mathbf{s}_j, \omega_i^j \to \omega_o^j)\, L_{e,i}(\mathbf{s}_j, \omega_i^j)\, d\sigma_{\mathbf{s}_j}^{\perp}(\omega_i^j) \quad (6.528)$$

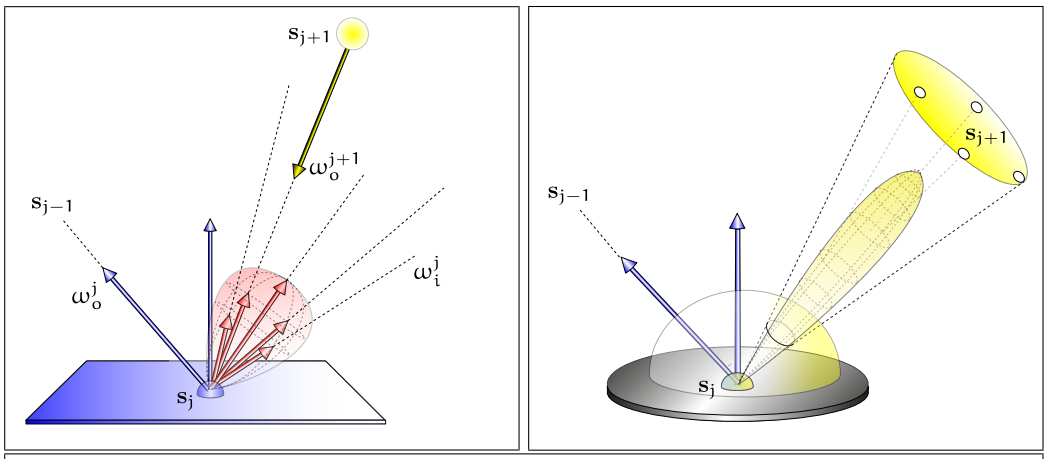


FIGURE 6.36: GEOMETRY OF THE GLOSSY HIGHLIGHTS PROBLEM. The integral within the scattering equation can be evaluated using directional sampling, also called sampling the BSDF, or integration can be done by *area sampling* of the light source.

where, integrated via the area of the light source ☼, it looks like this

$$\int_{\partial\mathcal{V}} f_s(s_{j+1} \to s_j \to s_{j-1})\, L_e(s_{j+1} \to s_j)\, \mathcal{G}(s_{j+1} \leftrightarrow s_j)\, d\mu^2(s_{j+1}) \tag{6.529}$$

with $s_{j+1} \in ☼$.

Now, there are two different sampling strategies for evaluating these integrals, see Figure 6.36: The first integral can be evaluated using *directional sampling*, also called *sampling the BSDF*, or *BRDF-sampling*, and the second integral can be evaluated by *area sampling* of the light source.

Probability Density Function (176)
BSDF (371)

With directional sampling, we have to sample incident directions $\omega_i^{j_k}$ according to a probability density function p_σ usually chosen to be proportional to the BSDF f_s thus:

$$p_\sigma(\omega_i^{j_k}) \propto f_s(s_j, \omega_i^{j_k} \to \omega_o^j) \tag{6.530}$$

or

$$p_\sigma(\omega_i^{j_k}) \propto f_s(s_j, \omega_i^{j_k} \to \omega_o^j) \left|\cos\theta_i^{j_k}\right|. \tag{6.531}$$

Then, an associated secondary Monte Carlo estimator for approximating Equation (6.528) has the form

$$\frac{1}{N}\sum_{k=1}^{N} \frac{f_s(s_j, \omega_i^{j_k} \to \omega_o^j)\, L_e(s_j, \omega_i^{j_k})}{p_\sigma(\omega_i^{j_k})}, \tag{6.532}$$

PDF (176) where $\boldsymbol{\omega}_i^{j_k}$ are N according to the pdf p_σ distributed independent random variables and $L_e(\mathbf{s}_j, \boldsymbol{\omega}_i^{j_k})$ is the radiance exitant from point $\mathbf{s}_j$ in direction $\boldsymbol{\omega}_i^{j_k}$. Note, not all samples $\boldsymbol{\omega}_i^{j_k}$ are important for the computation of the highlight, only those samples that hit the light source are relevant.

The integral from Equation (6.529) is usually evaluated by a sampling strategy, which is known as *area sampling*. A typical strategy with area sampling is to randomly sample points on $\mathfrak{A}$ according to a PDF which is uniformly distributed with respect to surface area $\mu^2(\mathfrak{A})$ or the emitted power $\mu(\Phi_{\mathfrak{A}})$, that is, Uniform Distribution (180)

$$p_{\mu^2}(\mathbf{x}) = \frac{1}{\mu^2(\mathfrak{A})} \tag{6.533}$$

or

$$p_\mu(\mathbf{x}) = \frac{1}{\mu(\Phi_{\mathfrak{A}})}. \tag{6.534}$$

Then, an associated secondary Monte Carlo estimator for approximating Equation (6.529) has the form Monte Carlo Estimator (499)

$$\frac{1}{N}\sum_{k=1}^{N} \frac{f_s(\mathbf{X}_{j+1,k} \to \mathbf{s}_j \to \mathbf{s}_{j-1})\, L_e(\mathbf{X}_{j+1,k} \to \mathbf{s}_j)\, \mathcal{G}(\mathbf{X}_{j+1,k} \leftrightarrow \mathbf{s}_j)}{p_{\mu^2}(\mathbf{X}_{j+1,k})}, \tag{6.535}$$

or

$$\frac{1}{N}\sum_{k=1}^{N} \frac{f_s(\mathbf{X}_{j+1,k} \to \mathbf{s}_j \to \mathbf{s}_{j-1})\, L_e(\mathbf{X}_{j+1,k} \to \mathbf{s}_j)\, \mathcal{G}(\mathbf{X}_{j+1,k} \leftrightarrow \mathbf{s}_j)}{p_{\mu}(\mathbf{X}_{j+1,k})}, \tag{6.536}$$

where $\mathbf{X}_{j+1,k}$ are N according to the PDFs p_{μ^2} or p_μ distributed random variables.

Due to the more or less recognizable noise in the reflections of the light sources on the glossy plates in Figure 6.37, we can conclude that depending on the size of the light source and the roughness of the surface, one of these sampling strategies solves the glossy highlight problem much better than the other.

Obviously, it is relatively unlikely that rays sampled according to the BSDF hit a very small light source, that is, the second sampling strategy delivers better results in those cases where the light sources are very small and the material is more diffuse, see the lower left portions of the images in Figure 6.37. In the opposite case, where the light source is large and the material is highly smooth, sampling the BSDF is far superior than area sampling, compare the upper right portions of Figure 6.37. The reason for this is that points, randomly chosen on the light source according to area sampling, will probably not contribute significantly to the radiance reflected along the viewing ray. Radiance (250)

Now, this result should not be surprisingly, as the integrand in the scattering equation is a product of various unrelated factors—the BSDF f_s, the emitted radiance L_e, and several geometric quantities like the visibility and the geometry term—but both sampling

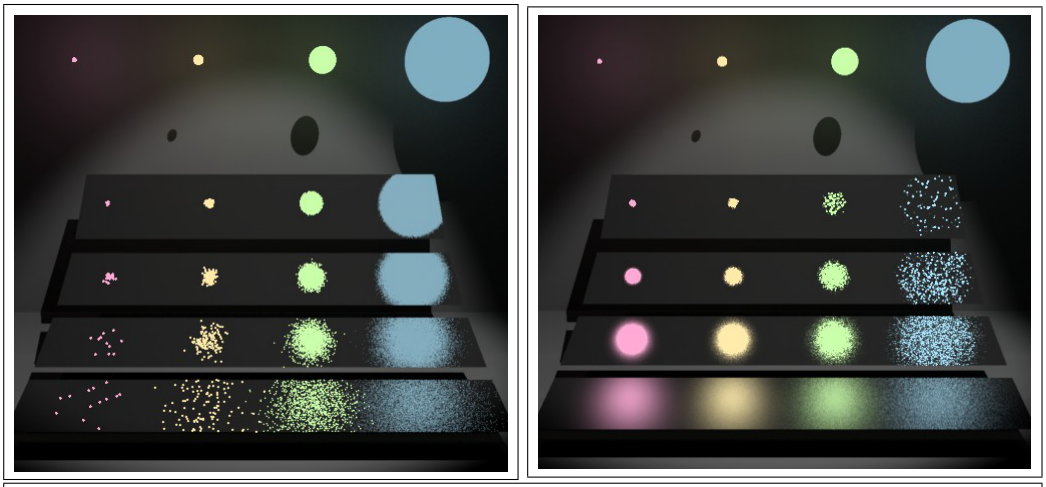


FIGURE 6.37: THE GLOSSY HIGHLIGHTS PROBLEM. A comparison of two sampling strategies for glossy highlights from area light sources. The images show a scene model composed of four rectangular plates of different degrees of surface roughness, and illuminated by four spherical light sources of varying radii and color that emit the same total power. Obviously, the different surface roughness controls how sharp or fuzzy the reflections of the light sources at the plates are. The images are rendered using different sampling techniques: In the left image, incident directions are sampled proportional to the BSDF, on the right-hand side, the light sources are sampled and shadow rays are fired in direction to the light sources. Image courtesy of Eric Veach, Stanford University.

strategies does not take into account all of these factors. While BSDF sampling does not account for the radiance emitted from the light source, area sampling does not take into account the BSDF. This can be interpreted in such a way, that a disregarded dominant Visibility Function (45)
Geometry Term (129) factor leads to a weak sampling technique. Thus, an ideal probability density function should be proportional to the product of all of these factors.

REMARK 6.29 *Let us recall: Both sampling strategies work on the same domain of integration, which can be interpreted as a set of directions or as a set of surface points. Therefore, we can also express one of our sampling strategy in terms of the*

other using the relationships

$$p_{\mu^2}(s_{j+1}) = \frac{d\mathbb{P}_{\mu^2}}{d\mu^2}(s_{j+1}) \tag{6.537}$$

$$= \frac{d\mathbb{P}_\sigma}{d\sigma}(\omega_i^j)\frac{d\sigma(\omega_i^j)}{d\mu^2(s_{j+1})} \tag{6.538}$$

$$\overset{(2.196)}{=} p_\sigma(\omega_i^j)\frac{d\mu^2(s_{j+1})\left|\cos\theta_o^{j+1}\right|}{d\mu^2(s_{j+1})\,\|s_{j+1}-s_j\|_2^2} \tag{6.539}$$

$$= p_\sigma(\omega_i^j)\frac{\left|\cos\theta_o^{j+1}\right|}{\|s_{j+1}-s_j\|_2^2} \tag{6.540}$$

as well as

$$p_\sigma(\omega_i^j) = \frac{d\mathbb{P}_\sigma}{d\sigma}(\omega_i^j) \tag{6.541}$$

$$= \frac{d\mathbb{P}_{\mu^2}}{d\mu^2}(s_{j+1})\frac{d\mu^2(s_{j+1})}{d\sigma(\omega_i^j)} \tag{6.542}$$

$$\overset{(2.196)}{=} p_{\mu^2}(s_{j+1})\frac{d\mu^2(s_{j+1})\|s_{j+1}-s_j\|_2^2}{d\mu^2(s_{j+1})\left|\cos\theta_o^{j+1}\right|} \tag{6.543}$$

$$= p_{\mu^2}(s_{j+1})\frac{\|s_{j+1}-s_j\|_2^2}{|\cos\theta_o^{j+1}|}. \tag{6.544}$$

With the help of these two formulas, we can then convert a directional density into an area density and vice versa, resulting in two different sampling strategies given over the same integration domain.

6.6.9.2 COMBINING SAMPLING TECHNIQUES

When designing a Monte Carlo technique for evaluating the SLTEV, we have rarely accurate information about the look and the shape of the integrand. The only thing we know about the integrand are a few parameters, such as the BSDF, the scene, and light source geometry, etc., that describe them. As these parameters can vary, it is difficult to design a single sampling strategy that works reliably in all situations.

SLTEV (398)

BSDF (371)

Now, our goal is to design sampling strategies that guarantee low-variance results for the whole range of parameter values, thus, for all possible integrands resulting from changes in the parameters. But this can not be achieved with the hitherto methods as the integrand is usually a sum or product of many different factors, which can not be sampled directly. One tries to solve this problem by choosing samples from PDFs that are proportional to some subsets of factors of the intergrand. But as the glossy highlights problem shows, this can lead to high variance, when a dominant factor is unconsidered.

Variance of a RV (201)

PDF (176)

The key for solving this problem is to avoid insufficient sampling of the integrand f where its values are large.

To achieve this goal, that is, for solving the integral,

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}), \tag{6.545}$$

by constructing estimators that have low variance for a broad class of integrands, in [222, Veach and Guibas 1995], [221, Veach 1998], the following approach is made:

Design a set of importance sampling techniques on the domain $\mathbf{Q}^s$—whose corresponding PDFs are denoted by $p_1, \ldots, p_n$—where at least one of those sampling techniques must be able to generate samples where the integrand is large, but not every p_i must be able to sample from the entire domain. Finally, we then estimate the integral as a weighted combination of all the samples. Due to [221, Veach 1998], this approach is called *Multiple Importance Sampling*, or *MIS*.

Let us now introduce a combined estimator that assigns an appropriate weight to each individual sample, the so-called *the multi-sample estimator*. It is defined as follows:

DEFINITION 6.17 (The Multiple-sample Estimator) *Let us consider the integral from Equation (6.545) and let us assume* $p_1, \ldots, p_n$ *be probability density functions defined on the probability space* $(\mathbf{Q}^s, \mathfrak{B}(\mathbf{Q}^s), \mu^s)$. *Let us furthermore assume that* $w_i, 1 \leq i \leq n$ *are chosen weight functions, which apart from*

Probability Density Function (176)

$$\sum_{i=1}^{n} w_i(\mathbf{x}) = 1 \text{ whenever } f(\mathbf{x}) \neq 0 \tag{6.546}$$

also must satisfy the condition:

$$w_i(\mathbf{x}) = 0 \text{ for } p_i(\mathbf{x}) = 0. \tag{6.547}$$

Let $\mathbf{X}_{i,j}, 1 \leq i \leq n, 1 \leq j \leq n_i$ *with* $\sum_{i=1}^{n} n_i = N$ *according to the PDF* p_i *distributed independent random variables on* $\mathbf{Q}^s$*, then we call the expression*

Random Variable (168)

$$F_N^{MIS} \stackrel{def}{=} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} w_i(\mathbf{X}_{i,j}) \frac{f(\mathbf{X}_{i,j})}{p_i(\mathbf{X}_{i,j})}, \tag{6.548}$$

the multiple-sample estimator.

As already mentioned above, multiple importance sampling is a technique, that attempts to find weighting functions w_i and PDFs p_i that satisfy the condition, that at least one sampling techniques must be able to generate samples in those regions, where the integrand is not equal to zero. Contrary to importance sampling, such a technique must not

be able to draw samples on the whole domain of integration, but it can concentrate on important regions for the integrand. That is, the condition from Equation (6.546) ensures, that a sample, drawn from several PDFs, is accounted for exactly once while the second condition says, that a sample, which cannot be drawn by a certain PDF should evaluate to zero.

REMARK 6.30 (Simple Examples of Weighting Functions) *By choosing the weighting functions w_i as constant over the whole domain of integration, we get a multiple-sample estimator, which can be seen as a weighted combination of separate estimators. On the other hand, setting w_i to one for specific subdomains $\mathbf{Q}_i^s$ of $\mathbf{Q}^s$ leads to a separation of the original domain of integration.*

The multiple-sample estimator F_N^{MIS} from Definition 6.17 satisfies the following theorem:

THEOREM 6.4 *The multiple-sample estimator F_N^{MIS} is unbiased, that is, it holds:*

$$E\left(F_N^{MIS}\right) = \int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}). \tag{6.549}$$

PROOF 6.4 *Obviously it holds for the expected value of the multiple-sample estimator:*

$$\begin{aligned}
E\left(F_N^{MIS}\right) &= E\left(\sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{w_i(\mathbf{X}_{i,j}) f(\mathbf{X}_{i,j})}{p_i(\mathbf{X}_{i,j})}\right) && (6.550)\\
&\overset{(2.732)}{=} \int_{\mathbf{Q}^s} \left(\sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{w_i(\mathbf{x}) f(\mathbf{x})}{p_i(\mathbf{x})}\right) d\mathbb{P}(\mathbf{x}) && (6.551)\\
&= \int_{\mathbf{Q}^s} \left(\sum_{i=1}^{n} \frac{w_i(\mathbf{x}) f(\mathbf{x})}{p_i(\mathbf{x})}\right) d\mathbb{P}(\mathbf{x}) && (6.552)\\
&\overset{d\mu^s = \frac{d\mathbb{P}}{p_i}}{=} \int_{\mathbf{Q}^s} \left(\sum_{i=1}^{n} w_i(\mathbf{x}) f(\mathbf{x})\right) d\mu^s(\mathbf{x}) && (6.553)\\
&\overset{(6.546)}{=} \int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}). && (6.554)
\end{aligned}$$

Let us finally consider an example for the construction of a multiple-sample estimator for the glossy highlights problem:

EXAMPLE 6.41 (A Multiple-sample Estimator for the Glossy Highlights Problem) *Based on our discussion in Section 6.6.9.1, we can combine the two sampling strategies, sampling the BSDF and sampling the area light source, to achieve a multiple sample estimator.*

Let p_σ and p_{μ^2} be a BSDF-sampling, respectively, area light source sampling technique. If we now choose the weighting functions as constant, such as e.g. $w_i = \frac{1}{2}$ for $1 \leq i \leq 2$, then a multiple-sample estimator can be written as:

$$F_N^{MIS} = \frac{1}{n_i}\sum_{k=1}^{n_i}\frac{1}{2}\frac{f_s(\mathbf{s}_j, \boldsymbol{\omega}_i^{j_k} \rightarrow \omega_o^j)\, L_{e,i}(\mathbf{s}_j, \boldsymbol{\omega}_i^{j_k})}{p_\sigma(\boldsymbol{\omega}_i^{j_k})} + \frac{1}{n_i}\sum_{k=1}^{n_i}\frac{1}{2}\frac{f_s(\mathbf{X}_{j+1,k} \rightarrow \mathbf{s}_j \rightarrow \mathbf{s}_{j-1})\, L_e(\mathbf{X}_{j+1,k} \rightarrow \mathbf{s}_j)\, G(\mathbf{X}_{j+1,k} \leftrightarrow \mathbf{s}_j)}{p_{\mu^2}(\mathbf{X}_{j+1,k})}, \tag{6.555}$$

where $\boldsymbol{\omega}_i^j$ and $\mathbf{X}_{j+1,k}$ are independent according to the probability density functions p_σ and p_{μ^2} independent distributed random variables.

With $n_i = 1$ the multiple-sample estimator F_2^{MIS} can simplified be written as:

$$F_2^{MIS} = \frac{1}{2}\left(\frac{f_s(\mathbf{s}_j, \boldsymbol{\omega}_i^j \rightarrow \omega_o^j)\, L_{e,i}(\mathbf{s}_j, \boldsymbol{\omega}_i^j)}{p_\sigma(\boldsymbol{\omega}_i^j)} + \frac{f_s(\mathbf{X}_{j+1} \rightarrow \mathbf{s}_j \rightarrow \mathbf{s}_{j-1})\, L_e(\mathbf{X}_{j+1} \rightarrow \mathbf{s}_j)\, G(\mathbf{X}_{j+1} \leftrightarrow \mathbf{s}_j)}{p_{\mu^2}(\mathbf{X}_{j+1})}\right). \tag{6.556}$$

It should be clear, that the above multiple-sample estimator F_N^{MIS}—by using the Formulas (6.540) and (6.544)—can also be completely expressed either with directional or spatial random variables.

The above choice of the weighting functions $w_i = \frac{1}{2}$ is not a really good choice. In the next section, we discuss weighting heuristics that work very well.

EXAMPLE 6.42 (A Multiple-sample Estimator for the Glossy Highlights Problem, continued)
The above multiple-sample estimator F_N^{MIS} can be drilled, by using the fact that a BSDF is usually decomposed of a diffuse, specular, and a glossy component. Based on this decomposition, the integrand of the stationary light transport equation can be written in the form

$$f_s(\mathbf{s}_j, \omega_i^j \rightarrow \omega_o^j)\, L_i(\mathbf{s}_j, \omega_i^j) = \left(f_s^o(\mathbf{s}_j, \omega_i^j \rightarrow \omega_o^j) + f_s^\vee(\mathbf{s}_j, \omega_i^j \rightarrow \omega_o^j) + f_s^{gl}(\mathbf{s}_j, \omega_i^j \rightarrow \omega_o^j)\right) L_i(\mathbf{s}_j, \omega_i^j). \tag{6.557}$$

With respect to the glossy highlights problem, we then use three importance-based BSDF-sampling strategies for sampling the different components of the BSDF and combine these techniques with our area-sampling strategy for sampling the involved light source.

Let $p_o, p_\vee$ and p_{gl} denote the corresponding BSDF-sampling techniques and $p_\square$ the area-sampling technique. With constant weighting functions $w_i = \frac{1}{4}$ for $l \in$

$\{o, \vee, gl, \square\}$, *then the above multiple-sample estimator can be written as:*

$$F_N^{MIS} = \sum_{l \in \{o,\vee,gl\}} \frac{1}{4n_l} \sum_{k=1}^{n_i} \frac{f_s^l(\mathbf{s}_j, \boldsymbol{\omega}_i^{j_k} \rightarrow \omega_o)\, L_i(\mathbf{s}_j, \boldsymbol{\omega}_i^{j_k})}{p_{\sigma_l}(\boldsymbol{\omega}_i^{j_k})} + \frac{1}{4n_l} \frac{f_s(\mathbf{X}_{j+1,k} \rightarrow \mathbf{s}_j \rightarrow \mathbf{s}_{j-1})\, L(\mathbf{X}_{j+1,k} \rightarrow \mathbf{s}_j)\, \mathcal{G}(\mathbf{X}_{j+1,k} \leftrightarrow \mathbf{s}_j)}{p_{\mu^2}(\mathbf{X}_{j+1,k})}, \tag{6.558}$$

where $\boldsymbol{\omega}_i^{j_k}$ *and* $\mathbf{X}_{j+1,k}$ *are independent according to the probability density functions* p_{σ_l} *and* p_{μ^2} *distributed random variables.*

With $n_l = 1$ *the multiple-sample estimator* F_4^{MIS} *can simplified written as:*

$$F_4^{MIS} = \frac{1}{4} \sum_{l \in \{o,\vee,gl\}} \frac{f_s^l(\mathbf{s}_j, \boldsymbol{\omega}_i^j \rightarrow \omega_o)\, L_i(\mathbf{s}_j, \boldsymbol{\omega}_i^j)}{p_{\sigma_l}(\boldsymbol{\omega}_i^j)} + \frac{f_s(\mathbf{X}_{j+1} \rightarrow \mathbf{s}_j \rightarrow \mathbf{s}_{j-1})\, L(\mathbf{X}_{j+1} \rightarrow \mathbf{s}_j)\, \mathcal{G}(\mathbf{X}_{j+1} \leftrightarrow \mathbf{s}_j)}{p_{\mu^2}(\mathbf{X}_{j+1})}. \tag{6.559}$$

As already mentioned in Remark 6.30, the partition of the integration domain in subdomains, as known from stratified sampling, can also be modeled via multiple importance sampling. For this, let us consider the following example.

Section 6.6.4

EXAMPLE 6.43 (Partition of the Integral Domain) *By constructing weighting functions* w_i*, which are defined only over subdomains of the original domain of integration, multiple importance sampling allows the representation of the original integral as the sum of integrals over disjoint subdomains of* $\mathbf{Q}^s$*, that is, the integral from Equation (6.545) can be expressed as:*

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) = \sum_{i=1}^{n} \int_{\mathbf{Q}_i^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}) \tag{6.560}$$

where it holds $\mathbf{Q}_i^s \cap \mathbf{Q}_j^s = \emptyset$ *for* $i \neq j, 1 \leq i, j \leq n$ *and*

$$w_i(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathbf{Q}_i^s \\ 0 & \text{otherwise.} \end{cases} \tag{6.561}$$

The subdomains $\mathbf{Q}_i^s$ *can then be sampled separately by techniques* p_i*. This leads to a multiple-sample estimator of the form*

$$F_N^{MIS} = \sum_{i=1}^{n} \frac{1}{n_i} w_i(\mathbf{X}_{i,k}) \frac{f(\mathbf{X}_{i,k})}{p_i(\mathbf{X}_{i,k})} \tag{6.562}$$

with n_i *according to the probability density functions* p_i *distributed independent random variables drawn from* $\mathbf{Q}_i^s$*.*

A typical application of this technique is the partitioning of a scene into light source regions and non-light source region when evaluating the scattering equation, as we will encounter them when discussing Monte Carlo path tracing.

Scattering Equation (374)

Section 9.1

6.6.9.3 WEIGHTING HEURISTICS

For constructing an efficient Monte Carlo estimator F_N^{MIS} simple combining the weighting functions as we did it in Example (6.41) with

$$F_N^{MIS} = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} w_i(\mathbf{X}_{i,k}) \frac{f(\mathbf{X}_{i,k})}{p_i(\mathbf{X}_{i,k})} \tag{6.563}$$

$$\overset{w_i=\frac{1}{n}}{=} \frac{1}{n} \sum_{i=1}^{n} F_{N,i}^{IS} \tag{6.564}$$

Variance (201) and $F_{N,i}^{IS} = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{f(\mathbf{X}_{i,k})}{p_i(\mathbf{X}_{i,k})}$ is not a good choice, since the variance of F_N^{MIS} is depending on the variance of the estimators $F_{N,i}^{IS}$. That is, high variance in any of the estimators $F_{N,i}^{IS}$ leads to high variance of F_N^{MIS}.

THE BALANCE HEURISTIC. Our goal is now to find an estimator F_N^{MIS} with minimum variance, where the weighting functions w_i are chosen appropriately.

DEFINITION 6.18 (Balance Heuristic) *Let $p_1, \ldots, p_n$ be strategies for sampling the integral from Equation (6.545). The* balance heuristic *is given by the following weighting functions:*

$$w_i(\mathbf{x}) \overset{def}{=} \frac{n_i p_i(\mathbf{x})}{\sum_k n_k p_k(\mathbf{x})} \tag{6.565}$$

where $, 1 \leq i, k \leq n$. Due to [221, Veach 1998], there is no other combination which is much better than the balance heuristic.

Based on the balance heuristic, the multiple-sample estimator can be written as:

$$F_N^{MIS} = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \left(\frac{n_i p_i(\mathbf{X}_{i,k})}{\sum_k n_k p_k(\mathbf{X}_{i,k})} \right) \frac{f(\mathbf{X}_{i,k})}{p_i(\mathbf{X}_{i,k})} \tag{6.566}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{f(\mathbf{X}_{i,k})}{\sum_k n_k p_k(\mathbf{X}_{i,k})} \tag{6.567}$$

$$= \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{f(\mathbf{X}_{i,k})}{\sum_k c_k p_k(\mathbf{X}_{i,k})}, \tag{6.568}$$

where $N = \sum_{i=1}^{n} n_i$ is the total number of samples and $c_k = \frac{n_k}{N}$ corresponds to the fraction of samples chosen via p_k. Expressing the sum in the denominator in terms of a so-called *combined sample density* $\hat{p}$, defined by:

$$\hat{p}(\mathbf{x}) \overset{def}{=} \sum_{k=1}^{n} c_k p_k(\mathbf{x}), \tag{6.569}$$

the multiple-sample estimator takes on the form of a standard Monte Carlo estimator.

Due to its definition, the balance heuristic depends on both, the PDF as well as the number of samples used. Obviously, the contribution $f(\mathbf{X}_{i,k})$ of an *improbable sample* $\mathbf{X}_{i,k}$, thus a sample which is drawn with very small probability, to the estimator is large. As such samples increase the variance of an unweighted estimator, the associated PDF is not suitable for sampling. In multiple importance sampling, the contributions of these samples are compensated by the weights w_i. That is, if there is another sampling strategy p_i that draws the sample $\mathbf{X}_{i,k}$ with high probability, the sum of the weights $w_i(\mathbf{X}_{i,k})$ in the denominator becomes large compared to the nominator, and the small weight of this sample compensates the large unweighted contribution of $\frac{f(\mathbf{X}_{i,k})}{p_i(\mathbf{X}_{i,k})}$. Apart from the PDF also the number of samples have an influence on the weights w_i, as a sampling strategy, that uses more samples—and which has therefore a lower variance—works with a larger weight than a sampling strategy, that uses fewer samples.

ADVANCED HEURISTICS. When no information about the integrand is available, the balance heuristic is a good choice for constructing a multiple-sample estimator. But in [221, Veach 1998] it is also shown that there are other heuristics, variations of the balance heuristic, that have better performance in low-variance problems. These techniques use a so-called *sharpening strategy*. Roughly formulated this means that the weighting functions are modified in such a way that large weights are made closer to one and small weights are made closer to zero in a certain part of the integration domain.

DEFINITION 6.19 (The Cutoff Heuristic) *The* cutoff heuristic *modifies the weighting functions by discarding samples with low weight, according to a cutoff threshold* $\alpha \in [0, 1]$*:*

$$w_i(\mathbf{x}) = \begin{cases} 0 & \text{if } n_i p_i(\mathbf{x}) < \alpha \max_k(n_k p_k(\mathbf{x})) \\ \frac{n_i p_i(\mathbf{x})}{\sum_k \{n_k p_k(\mathbf{x}) | n_k p_k(\mathbf{x}) \geq \alpha \max_k(n_k p_k(\mathbf{x}))\}} & \text{otherwise.} \end{cases} \tag{6.570}$$

The threshold α *determines how small* $n_i p_i(\mathbf{x})$ *must be before it is thrown away [221, Veach 1998].*

DEFINITION 6.20 (The Power Heuristic) *The* power heuristic*, [221, Veach 1998], modifies the weighting functions in a different way, by raising all of the weights to an exponent* β*, and then renormalizing:*

$$w_i(\mathbf{x}) = \frac{(n_i p_i(\mathbf{x}))^\beta}{\sum_k (n_k p_k(\mathbf{x}))^\beta}. \tag{6.571}$$

An exponent $\beta = 2$ works well in practice and is most often used. The exponentiation increases the weight of a sample for those techniques that drawn the sample with a high probability $p_i(\mathbf{x})$.

DEFINITION 6.21 (The Maximum Heuristic) *The* maximum heuristic *partitions the domain into* n *regions, according to which function* q_i *is largest at each point* $\mathbf{x}$*:*

$$w_i(\mathbf{x}) = \begin{cases} 1 & \text{if } n_i p_i(\mathbf{x}) = \max_k(n_k p_k(\mathbf{x}))\} \\ 0 & \text{otherwise.} \end{cases} \tag{6.572}$$

REMARK 6.31 *Obviously, the balance heuristic is a special case of the cutoff as well as the power heuristic since due to the choice of* $\alpha = 0$*, respectively,* $\beta = 1$ *the definitions of the cutoff and power heuristic lead to the balance heuristic. The same holds for the maximum heuristic by choosing* $\alpha = 1$*, respectively,* $\beta = \infty$*. While the power heuristic is of a great practical interest, the two other heuristics are rather of theoretical interest.*

REMARK 6.32 *The discussion above shows that multiple importance sampling is a great tool for constructing robust Monte Carlo estimators. Instead of trying to find an ideal PDF that matches the shape of the integrand over the whole domain of integration, now specific PDFs, adapted to the underlying problem, can be used. Due to the combination strategy of MIS we can then expect good results in the process of sampling. But it should also be clear that this fine sampling strategy does not come for free. Compared with one of our standard variance reduction techniques, multiple importance sampling requires additional cost for computing the weights* w_i *and the evaluation of* n *PDFs at a sample* $\mathbf{X}_{i,k}$*.*

In Section 9.3, we present bidirectional path tracing*, a Monte Carlo rendering algorithms that makes use of MIS as sampling strategy.*

6.7 MONTE CARLO INTEGRATION AND FREDHOLM INTEGRAL EQUATIONS OF THE 2nd KIND

Fredholm Integral Equations (127)

As already noted, the linear integral equations underlying the global illumination problem are Fredholm equations of the 2^{nd} kind, that is, they are all of the form

$$f(\mathbf{x}) = g(\mathbf{x}) + \int_{\mathcal{R}} \mathbf{k}(\mathbf{x}, \mathbf{y})\, f(\mathbf{y})\, d\mu(\mathbf{y}), \tag{6.573}$$

Measurable Set (80)
$\mathcal{L}^2(\mathcal{R}, \mu)$ (110)
Integral Kernel (127)

where $\mathcal{R}$ is a measurable set, f, g are real-valued functions from $\mathcal{L}^2(\mathcal{R}, \mu)$, and the kernel $\mathbf{k}$ is a real-valued function of the two variables $\mathbf{x}$ and $\mathbf{y}$ from $\mathcal{L}^2(\mathcal{R} \times \mathcal{R}, \mu \times \mu)$.

In Section 2.3.3 we have already presented deterministic, numerical methods for solving Fredholm integral equations of the 2^{nd} kind, which are based on functional analytical concepts. Now, we are interested in solution methods that are subjected to the principles of probability theory, i.e. we are interested in solving Fredholm equations via Monte Carlo methods. In this context, we discuss four approaches:

i) the approach of *successive integral substitution*,

ii) an approach based on *the Neumann series*,

iii) an approach based on *discrete-time Markov processes*, and

iv) *next event estimation*, a method where the original integral domain is split up into two separate integration domains.

All four approaches lead to the same result, but are based on different mathematical constructs, from which a variety to different solution algorithms can be generated.

6.7.1 A MONTE CARLO APPROACH BASED ON THE METHOD OF SUCCESSIVE INTEGRAL SUBSTITUTION

Our goal in this section is the derivation of a secondary Monte Carlo estimator—using a finite number of samples drawn from integration domain $\mathcal{R}$—that serves as an approximate solution to a Fredholm integral equation of the 2^{nd} kind. For that purpose, we have to generate according to a probability density function distributed random variables or random vectors from probability space $(\mathcal{R}, \mathfrak{B}(\mathcal{R}), \mathbb{P})$.

Monte Carlo Estimator (499)
Linear Integral Equation (127)
Independence of RV (204)

For the following discussion, let $\mathbf{X}_0$ and $\mathbf{X}_{0i_1}, 1 \leq i_1 \leq N_1$ be $N_1 + 1$ independent random variables or random vectors from $(\mathcal{R}, \mathfrak{B}(\mathcal{R}), \mathbb{P})$, distributed according to probability densities p_0 and p_1. Obviously, a first simple formula for a coarse approximation of

Random Variable (168)
Probability Space (163)
Probability Density Function (176)

$$f(\mathbf{x}) = g(\mathbf{x}) + \int_{\mathcal{R}} \mathbf{k}(\mathbf{x}, \mathbf{y})\, f(\mathbf{y})\, d\mu(\mathbf{y}) \tag{6.574}$$

is given by the secondary Monte Carlo estimator $F_N^{f(\mathbf{x})}$:

Monte Carlo Estimator (499)

$$F_N^{f(\mathbf{x})} = \frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} + \frac{1}{N_1} \sum_{i_1=1}^{N_1} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0) p_1(\mathbf{X}_{0i_1} | \mathbf{X}_0)} f(\mathbf{X}_{0i_1}) \tag{6.575}$$

with $\mathbf{x} = \mathbf{X}_0$ and $p_0(\mathbf{X}_0) = 1$.

As the function f in the above estimator is unknown, only the source function g contributes to our approximation $F_N^{f(\mathbf{x})}$, i.e. the value returned by $F_N^{f(\mathbf{x})}$ is not really usable.

Source Function (127)

Now, to make use of $F_N^{f(\mathbf{x})}$, we have to estimate the values $f(\mathbf{X}_{0i_1})$ from Equation (6.575) by additional N_1 secondary Monte Carlo estimators. Under the condition, that the samples $\mathbf{X}_{0i_1}$ are already drawn, each of these estimators uses N_2 according to a

conditional density $p_2(\mathbf{X}_{0i_1i_2}|\mathbf{X}_{0i_1})$ chosen new random variables $\mathbf{X}_{0i_1i_2}, 1 \leq i_2 \leq N_2$ from $(\mathcal{R}, \mathfrak{B}(\mathcal{R}), \mathbb{P})$. This then results in the estimator $F_N^{f(\mathbf{x})}$ given by:

$$\begin{aligned} F_N^{f(\mathbf{x})} &= \frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} + \\ &\quad \frac{1}{N_1}\sum_{i_1=1}^{N_1} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)} \left(g(\mathbf{X}_{0i_1}) + \frac{1}{N_2}\sum_{i_2=1}^{N_2} \frac{\mathbf{k}(\mathbf{X}_{0i_1}, \mathbf{X}_{0i_1i_2})}{p_2(\mathbf{X}_{0i_1i_2}|\mathbf{X}_{0i_1})} f(\mathbf{X}_{0i_1i_2}) \right) \end{aligned} \tag{6.576}$$

$$\begin{aligned} &= \frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} + \\ &\quad \frac{1}{N_1}\sum_{i_1=1}^{N_1} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)} g(\mathbf{X}_{0i_1}) + \\ &\quad \frac{1}{N_1}\frac{1}{N_2}\sum_{i_1=1}^{N_1}\sum_{i_2=1}^{N_2} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)} \frac{\mathbf{k}(\mathbf{X}_{0i_1}, \mathbf{X}_{0i_1i_2})}{p_2(\mathbf{X}_{0i_1i_2}|\mathbf{X}_{0i_1})} f(\mathbf{X}_{0i_1i_2}). \end{aligned} \tag{6.577}$$

With the same argumentation from above, this estimator is also not really usable, as the function f must be estimated by N_2 other secondary Monte Carlo estimators for $f(\mathbf{X}_{0i_1i_2})$, each of those is in turn using N_3 new samples.

The whole process can then be repeated by using further random variables $\mathbf{X}_{0i_1i_2i_3}$, $\mathbf{X}_{0i_1i_2i_3i_4}, \ldots$ and so on—for an illustration of the construction of these random variables see Figure 6.38.

If we stop after M steps, then we get a secondary Monte Carlo estimator for Equation(6.574) given by

$$\begin{aligned} &F_N^{f(\mathbf{x})} \\ &= \frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} + \\ &\quad \frac{1}{N_1}\sum_{i_1=1}^{N_1} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)} g(\mathbf{X}_{0i_1}) + \\ &\quad \frac{1}{N_1}\frac{1}{N_2}\sum_{i_1=1}^{N_1}\sum_{i_2=1}^{N_2} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)} \frac{\mathbf{k}(\mathbf{X}_{0i_1}, \mathbf{X}_{0i_1i_2})}{p_2(\mathbf{X}_{0i_1i_2}|\mathbf{X}_{0i_1})} g(\mathbf{X}_{0i_1i_2}) + \\ &\quad \ldots + \\ &\quad \prod_{j=1}^{M} \frac{1}{N_j}\sum_{i_1=1}^{N_1}\cdots\sum_{i_M=1}^{N_M} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)} \cdots \frac{\mathbf{k}(\mathbf{X}_{0i_1\ldots i_{M-1}}, \mathbf{X}_{0i_1\ldots i_M})}{p_M(\mathbf{X}_{0i_1\ldots i_M}|\mathbf{X}_{0i_1\ldots i_{M-1}})} g(\mathbf{X}_{0i_1\ldots i_M}), \end{aligned} \tag{6.578}$$

using $N = \prod_{j=1}^{M} N_j$ samples in the M^{th} estimation with $\mathbf{x} = \mathbf{X}_0$ and $p_0(\mathbf{X}_0) = 1$.

We call this method for approximating the solution of a Fredholm type integral equation the *method of successive integral substitution*, see Figure 6.39.

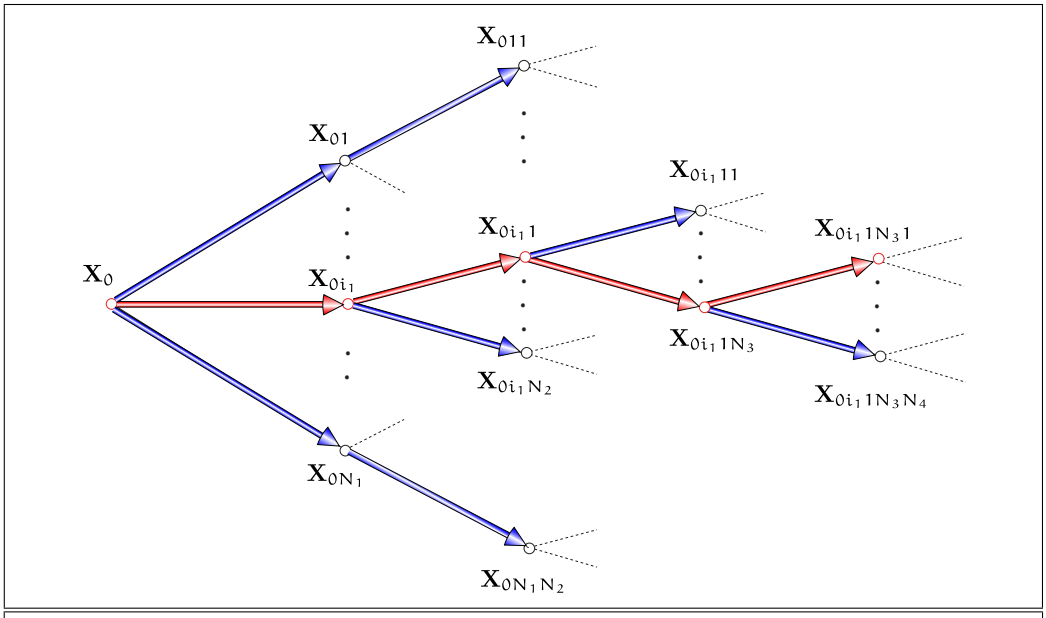


FIGURE 6.38: CONSTRUCTION OF THE SAMPLES $\mathbf{X}_{0i_1\ldots i_j}$. A sample $\mathbf{X}_{0i_1\ldots i_j}$ is starting point of N_{j+1} lines leading to further samples $\mathbf{X}_{0i_1\ldots i_j i_{j+1}}$ but only a single line goes from point $\mathbf{X}_{0i_1\ldots i_j}$ to its predecessor $\mathbf{X}_{0i_1\ldots i_{j-1}}$.

Obviously, a closed formula for the method of successive integral substitution is then given by:

$$\begin{aligned} F_N^{f(x)} &= \frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} + \\ &+ \sum_{l=1}^{M} \left\{ \sum_{i_1=1}^{N_1} \cdots \sum_{i_l=1}^{N_l} \left(\prod_{j=1}^{l} \frac{1}{N_j} \frac{k(\mathbf{X}_{0i_1\ldots i_{j-1}}, \mathbf{X}_{0i_1\ldots i_j})}{p_0(\mathbf{X}_0) p_j(\mathbf{X}_{0i_1\ldots i_j} | \mathbf{X}_{0i_1\ldots i_{j-1}})} \right) g(\mathbf{X}_{0i_1\ldots i_l}) \right\}, \end{aligned} \tag{6.579}$$

where we have used the identities:

$$\mathbf{X}_{0i_1 i_0} = \mathbf{X}_0 \quad \text{and} \quad \mathbf{X}_{0i_1 i_1} = \mathbf{X}_{0i_1}. \tag{6.580}$$

Let us illustrate, how these method can be applied to derive an algorithm for solving the global illumination problem.

EXAMPLE 6.44 (A First Example of a Naive Monte Carlo Rendering Algorithm) *Let us suppose the SLTEV is given in its spherical form, where the radiance is expressed in terms of exitant quantities, namely:* Section 8.4 SLTEV (403)

$$L_o(s, \omega_o) = L_e(s, \omega_o) + \int_{S^2} f_s(s, \omega_i \to \omega_o)\, L_o(\gamma(s, \omega_i), -\omega_i) d\sigma_s^{\perp}(\omega_i). \tag{6.581}$$

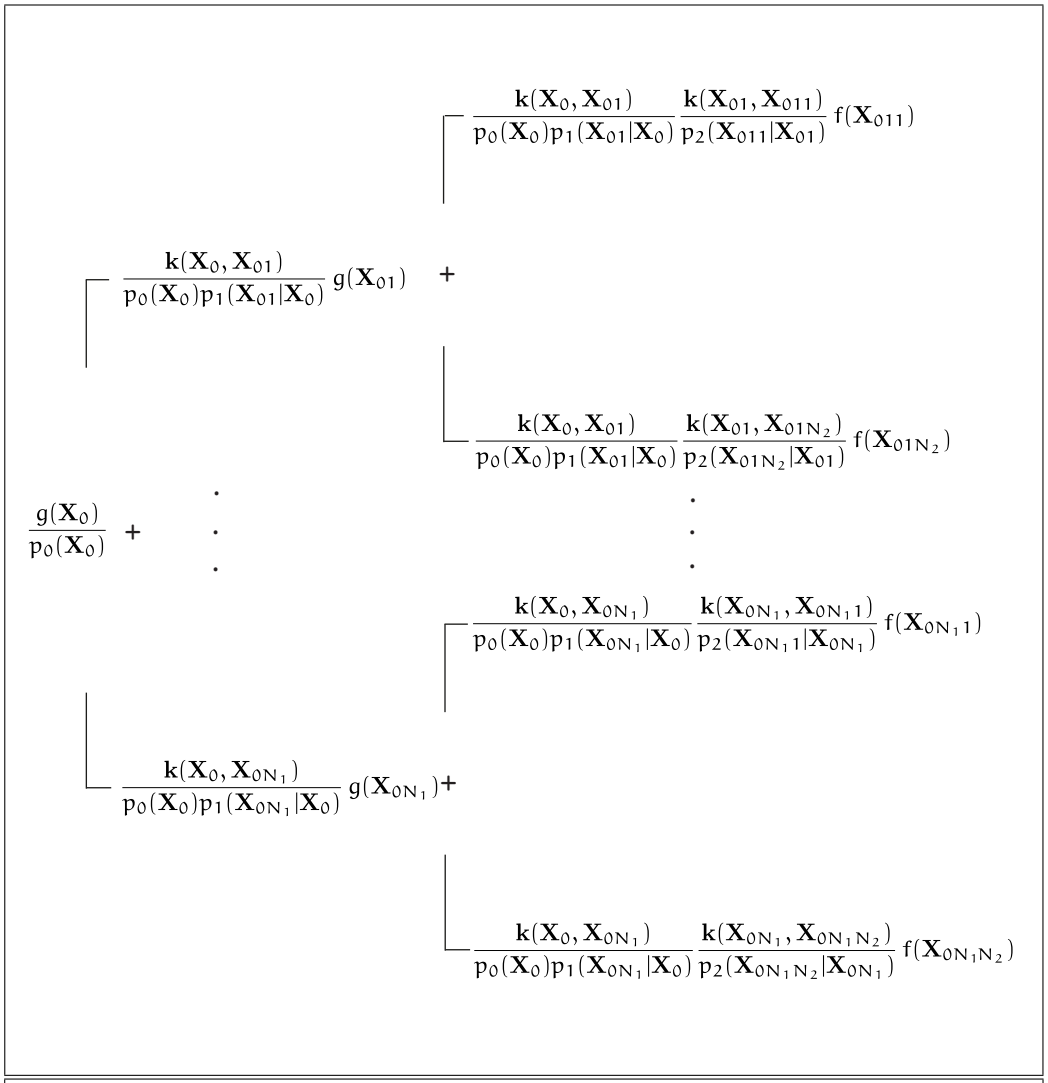


FIGURE 6.39: THE METHOD OF SUCCESSIVE SUBSTITUTION. For solving a Fredholm integral equation of the 2^{nd} kind, the method of successive integral substitution must evaluate the above computation tree. Due to the contracting property of the kernel function $\mathbf{k}$ and the increasing number of multiplications of $\mathbf{k}$, the nodes, lying on lower levels of the tree, yield less significant contributions to the overall result than the nodes lying on levels in the upper region of the tree. Since there are much more nodes deep inside of the tree, we cannot ignore the contributions induced by these nodes, because they guarantee the unbiasednes of the estimator. Here, a compromise would be desirable, namely to concentrate more work in the higher branches of the tree, without ignoring possible contributions from the lower levels of the tree: Russian roulette. Note, the tree, as shown above, only represents the first two levels of integral substitution in a Fredholm equation.

Obviously, this then suggest the following approach for estimating the SLTEV by the secondary Monte Carlo estimator $F_N^{f(\mathbf{x})}$ *from Equation (6.579):*

0) Choose $\mathbf{s} = \mathbf{X}_0$, *with* $p_0(\mathbf{s}) = p_0(\mathbf{X}_0) = 1$.

i) Evaluate the emitted radiance $L_e(\mathbf{s}, \omega_o)$ *at sample* $\mathbf{X}_0$ *in direction* ω_o *via*

$$\frac{L_e(\mathbf{X}_0, \omega_o)}{p_0(\mathbf{X}_0)} = L_e(\mathbf{X}_0, \omega_o). \tag{6.582}$$

ii) To estimate the integral kernel of the SLTEV via the multidimensional sum of Formula (6.579) do the following:

Generate, according to appropriate probability density functions p_j, *a tree of direction samples* $\boldsymbol{\omega}_i^{i_1 \dots i_j}, 1 \le j \le M, 1 \le i_j \le N_j$ *with root* $\mathbf{X}_0 = \mathbf{s}$, *depending on the surface and material properties of the objects, where it holds:*

$$\gamma(\mathbf{X}_{0 i_1 \dots i_{j-1}}, \boldsymbol{\omega}_i^{i_1 \dots i_j}) = \mathbf{X}_{0 i_1 \dots i_j}, \tag{6.583}$$

see Figure 6.40. Estimate the BSDF at point $\mathbf{X}_{0 i_1 \dots i_{j-1}}$ *in direction* $\omega_o^{i_1 \dots i_{j-1}}$ *by the fraction of light that arrives from* N_j *neighboring points via the directions* $\omega_i^{i_1 \dots i_j}$, *that is, by*

$$f_s(\mathbf{X}_{0 i_1 \dots i_{j-1}}, \omega_i^{i_1 \dots i_j} \to \omega_o^{i_1 \dots i_{j-1}}). \tag{6.584}$$

iii) Use the emitted radiance at $\mathbf{X}_{0 i_1 \dots i_l}$ *in direction* $\omega_o^{i_1 \dots i_l}$ *as the source function* g.

Then, a secondary Monte Carlo estimator for estimating $L_o(\mathbf{s}, \omega_o)$ *after* M *successive substitutions has the form*

$$F_N^{L_o(\mathbf{s}, \omega_o)} = \frac{L_e(\mathbf{X}_0, \omega_o)}{p_0(\mathbf{X}_0)} + \tag{6.585}$$
$$\sum_{l=1}^{M} \left\{ \sum_{i_1=1}^{N_1} \cdots \sum_{i_l=1}^{N_l} \left(\prod_{j=1}^{l} \frac{1}{N_j} \frac{f_s(\mathbf{X}_{0 i_1 \dots i_{j-1}}, \boldsymbol{\omega}_i^{i_1 \dots i_j} \to \omega_o^{i_1 \dots i_{j-1}}) \left| \cos \boldsymbol{\omega}_i^{i_1 \dots i_j} \right|}{p_0(\mathbf{X}_0)\, p_j(\boldsymbol{\omega}_i^{i_1 \dots i_j} | \boldsymbol{\omega}_i^{i_1 \dots i_{j-1}})} \right) \right.$$
$$\left. L_e(\mathbf{X}_{0 i_1 \dots i_l}, \omega_o^{i_1 \dots i_l}) \right\},$$

where we have used apart from the identities from Relation (6.580) also

$$\omega_o^{i_1 i_0} = \omega_o, \quad \omega_o^{i_1 i_1} = \omega_o^{i_1} \quad \text{as well as} \quad \omega_i^{i_1 i_1} = \omega_i^{i_1} \tag{6.586}$$

with $p_0(\mathbf{X}_0) = 1$ *and* $p_1(\boldsymbol{\omega}_i^{i_1 i_1} | \boldsymbol{\omega}_i^{i_1 i_0}) = p_1(\boldsymbol{\omega}_i^{i_1})$.

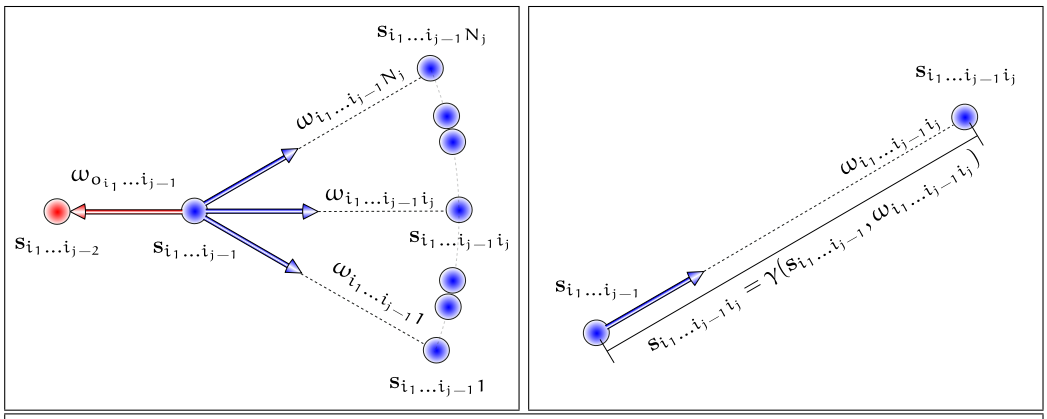


FIGURE 6.40: GEOMETRY OF THE BSDF USED IN THE NAIVE MONTE CARLO RENDERING ALGORITHM. At point $s_{i_1 \dots i_{j-1}}$ the BSDF gathers light contributions from N_j incident directions $\omega_{i_1 \dots i_{j-1} i_j}$ and forward these contributions, depending on the scattering behavior of the surface, in the outgoing direction $\omega_{o_{i_1 \dots i_{j-1}}}$ to point $s_{i_1 \dots i_{j-2}}$.

Obviously, our naive ray tracing algorithm fires, starting at point $\mathbf{s}$*, in a first step rays in a scene. At the hit points of these primary rays with the objects, the algorithm gathers the light that comes from these points, generates new rays, and also fires these rays into the scene. Repeated application of this approach then results in a tree of paths with root at point* $\mathbf{s}$ *that can be explored to compute the light arriving at* $\mathbf{s}$ *from points within the scene. The entire process will be repeated again and again, until a ray does not hit an object or the recursion depth of the method is exceeded, see Figure 6.41.*

Usually, the first term in Equation (6.585) is zero, since it represents the light emitted from point $\mathbf{s}$ *located at a light source. The sum in Formula (6.585) covers the direct and indirect illumination at point* $\mathbf{s}$*. While the direct illumination at* $\mathbf{s}$ *is evaluated via the first summation about the index* i_1*, the indirect illumination component is described by all other summations.*

Section 8.4 Section 9.1 SITEV (413)

REMARK 6.33 *The estimator* $F_N^{f(\mathbf{x})}$ *will be used to derive* distribution ray tracing *and* Monte Carlo path tracing*. If we also substitute the emitted radiance by the emitted importance, then* $F_N^{f(\mathbf{x})}$ *also leads to a Monte Carlo estimator for the* SITEV*.*

$O\left(\frac{1}{\sqrt{N}}\right)$ (516)

As you can see from Equation (6.579), the $O\left(\frac{1}{\sqrt{N}}\right)$ convergence rate of a Monte Carlo algorithm leads to high costs for the generation of the samples needed. The number of samples for evaluating a significant approximation increases according to the dimension of the involved integral. So, we need $\prod_{j=1}^{k} N_j$ samples when computing the k^{th} term of

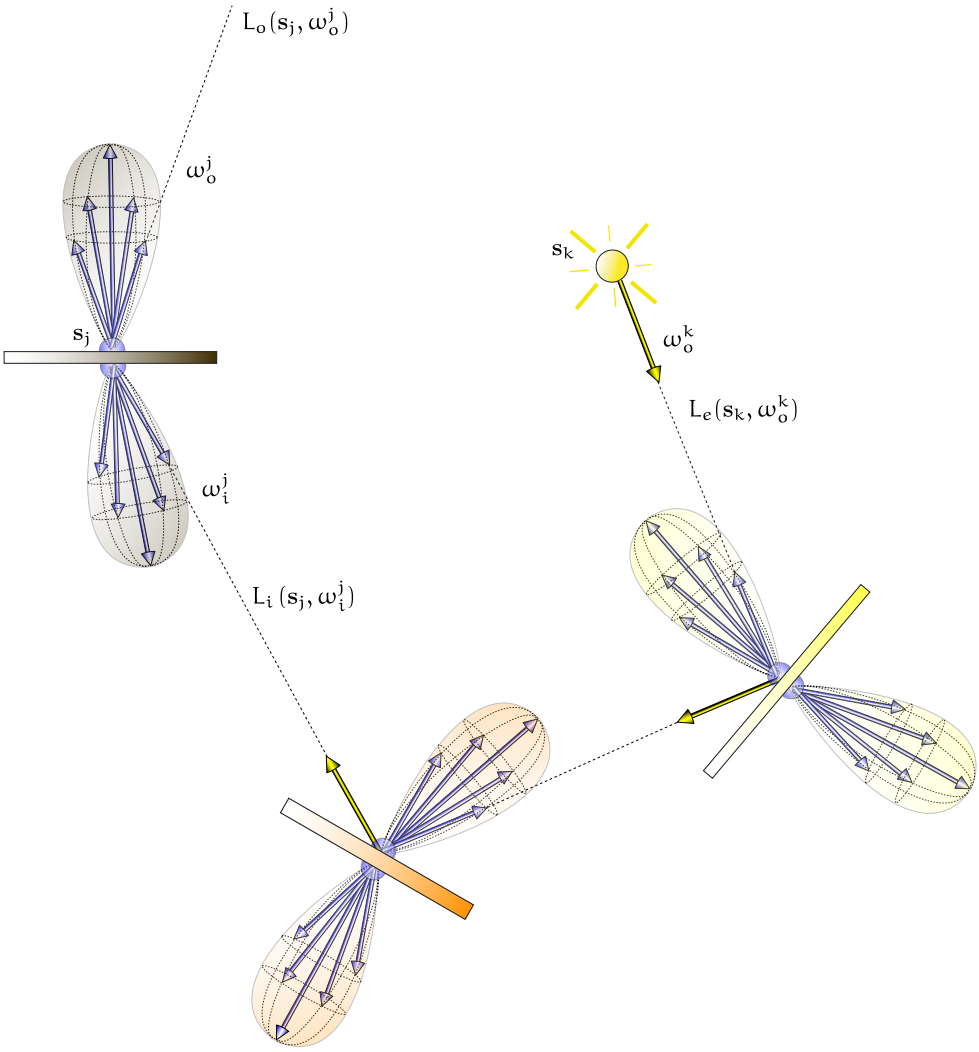


FIGURE 6.41: A NAIVE MONTE CARLO RENDERING ALGORITHM. Starting at point s_j the algorithm generates a large number of rays depending on the material and surface properties. At the hit points of these rays with objects of the scene, new rays are generated recursively until a predefined depth of recursion is achieved. The exitant radiance at point s_j in direction ω_o^j is the result of incident light coming from visible points on the surfaces, attenuated by reflection and/or refraction processes on their way to s_j.

the estimator from Equation (6.579). This leads to an evaluation of the integral kernel $\mathbf{k}$ at $\sum_{l=1}^{M}\prod_{j=1}^{l} N_j = O(N^M)$ sampling points for computing a partial sum of M members, where we assume $N_j = N$. Due to these high costs we conclude that Monte Carlo methods based on this naive successive substitution converge only very slowly. Without applying variance reduction techniques they are proper for sampling of only small or very small samples sizes. This fact gives us reason to reconsider our strategy for computing a solution of a Fredholm integral equations, which will lead to a new more simpler and more efficient strategy in Section 6.7.3.

Section 6.6

Expected Value (196)

Now, let us take a look at the expected value of our Monte Carlo estimator $F_N^{f(\mathbf{x})}$, obviously it holds:

$$
\begin{aligned}
E\left(F_N^{f(\mathbf{x})}\right) &= E\left(\frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)}\right) + \qquad (6.587)\\
&\quad E\left(\sum_{l=1}^{M}\left\{\sum_{i_1=1}^{N_1}\cdots\sum_{i_l=1}^{N_l}\left(\prod_{j=1}^{l}\frac{1}{N_j}\frac{\mathbf{k}(\mathbf{X}_{0i_1\ldots i_{j-1}},\mathbf{X}_{0i_1\ldots i_j})}{p_j(\mathbf{X}_{0i_1\ldots i_j}|\mathbf{X}_{0i_1\ldots i_{j-1}})}\right) g(\mathbf{X}_{0i_1\ldots i_l})\right\}\right)\\
&= \underbrace{E\left(\frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)}\right)}_{g(\mathbf{x}_0)} +\\
&\quad \underbrace{E\left(\frac{1}{N_1}\sum_{i_1=1}^{N_1}\frac{\mathbf{k}(\mathbf{X}_0,\mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)}\,g(\mathbf{X}_{0i_1})\right)}_{\int_{\mathcal{R}}\mathbf{k}(\mathbf{x}_0,\mathbf{x}_1)g(\mathbf{x}_1)\,d\mu(\mathbf{x}_1)} +\\
&\quad \underbrace{E\left(\sum_{i_1=1}^{N_1}\sum_{i_2=1}^{N_2}\left(\prod_{j=1}^{2}\frac{1}{N_j}\frac{\mathbf{k}(\mathbf{X}_{0i_1\ldots i_{j-1}},\mathbf{X}_{0i_1\ldots i_j})}{p_j(\mathbf{X}_{0i_1\ldots i_j}|\mathbf{X}_{0i_1\ldots i_{j-1}})}\right)\,g(\mathbf{X}_{0i_1i_2})\right)}_{\int_{\mathcal{R}}\int_{\mathcal{R}}\mathbf{k}(\mathbf{x}_0,\mathbf{x}_1)\mathbf{k}(\mathbf{x}_1,\mathbf{x}_2)g(\mathbf{x}_2)\,d\mu(\mathbf{x}_1)\,d\mu(\mathbf{x}_2)} + \qquad (6.588)\\
&\quad \ldots +\\
&\quad \underbrace{E\left(\sum_{i_1=1}^{N_1}\cdots\sum_{i_2=1}^{N_M}\left(\prod_{j=1}^{M}\frac{1}{N_j}\frac{\mathbf{k}(\mathbf{X}_{0i_1\ldots i_{j-1}},\mathbf{X}_{0i_1\ldots i_j})}{p_j(\mathbf{X}_{0i_1\ldots i_j}|\mathbf{X}_{0i_1\ldots i_{j-1}})}\right)\,g(\mathbf{X}_{0\ldots i_M})\right)}_{\int_{\mathcal{R}}\ldots\int_{\mathcal{R}}\mathbf{k}(\mathbf{x}_0,\mathbf{x}_1)\ldots\mathbf{k}(\mathbf{x}_{M-1},\mathbf{x}_M)g(\mathbf{x}_M)\,d\mu(\mathbf{x}_1)\ldots d\mu(\mathbf{x}_M)}\\
&= g(\mathbf{x}_0) + \qquad (6.589)\\
&\quad \sum_{j=1}^{M}\int_{\mathcal{R}}\ldots\int_{\mathcal{R}}\left(\prod_{k=1}^{j}\mathbf{k}(\mathbf{x}_{k-1},\mathbf{x}_k)\right) g(\mathbf{x}_j)\,d\mu(\mathbf{x}_1)\ldots d\mu(\mathbf{x}_j).
\end{aligned}
$$

As is easily seen from this result, the estimator $F_N^{f(\mathbf{x})}$ is unbiased only in the case where $M \to \infty$. But since we neglect an infinite number of terms from $F_N^{f(\mathbf{x})}$, it can not be unbiased, that is, it holds:

Unbiased (507)

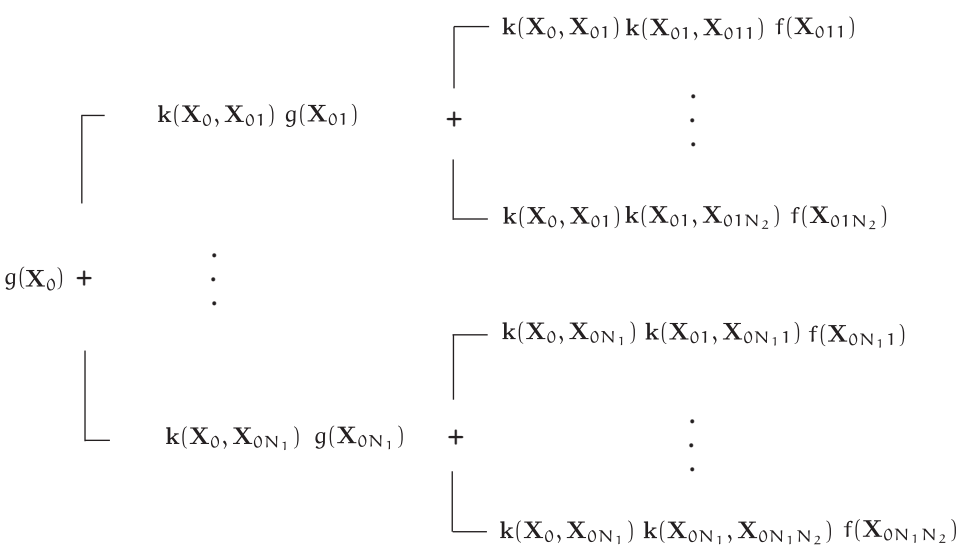


FIGURE 6.42: COMPUTATION TREE OF $f(\mathbf{x})$. Due to the multiplicity of the kernels, nodes that lie deep in the interior of the tree for computing $f(\mathbf{x})$ contribute only less to the result since it holds: $\prod_{j=1}^{l} \mathbf{k}(\mathbf{X}_{0i_1\ldots i_{j-1}}, \mathbf{X}_{0i_1\ldots i_j}) < \|\mathbf{k}\|^l \ll 1$.

$$\mathrm{E}\left(F_N^{f(\mathbf{x})}\right) \neq g(\mathbf{x}) + \int_{\mathcal{R}} \mathbf{k}(\mathbf{x},\mathbf{y})\, f(\mathbf{y})\, d\mu(\mathbf{y}) \tag{6.590}$$

$$= f(\mathbf{x}). \tag{6.591}$$

Evidently, the contributions of the terms that are not represented in $F_N^{f(\mathbf{x})}$ hurt the unbiased property of the algorithm. So, $F_N^{f(\mathbf{x})}$ is of limited use for the implementation of an approximate solution of the global illumination problem. Now, from Figure 6.42 you can see, that nodes which lie deep in the interior of the tree for computing f contribute—due to the multiplicity of the kernels, which we assumed as contracting, thus $\|\mathbf{k}\| < 1$—only less to the final result. Fortunately we conclude, due to our discussions in Section 6.3, that $F_N^{f(\mathbf{x})}$ is at least consistent if we revert to a sufficient number of terms.

Unbias (507)

Operator Norm (56)

Consistent (507)

The question that now arises: How can we avoid the bias, caused due to truncating the recursion? Now, a solution to this question can be found by applying the technique of Russian roulette as a stopping condition for evaluating terms of $F_N^{f(\mathbf{x})}$.

Russian Roulette (200)

For that purpose, we involve into the above process of successive substitution the sampling of independent on $[0,1]$ uniformly distributed random variables $U_i, i \geq 1$, and stop the recursion for computing new terms, if the random variable U_i does exceeds a predefined value α_i, $0 < \alpha_i < 1$. Now, in the case where we stop the recursion, infinitely

Uniform Distribution (180)

many terms are neglected, that is, we have to give the non-absorbed terms a higher weight. For this, we weight the term, evaluated in the i^{th} step of our process, with the probability of the drawn random variable U_i. As seen in Example 2.82, the resulting estimator is then unbiased, that is, it can be used to implement an approximate solver in any rendering algorithm.

Chapter 6

REMARK 6.34 *Due to [47, Dutre 1996], the choice of the parameter α_i should be taken carefully. A value of α_i close to one implies Monte Carlo estimators with large number of terms, which must be evaluated, that is, the result shows to be more reliable. If we choose α_i rather small, then the process of naive successive substitution will terminate rapidly, but due to this fact we must expect a higher variance on the final image.*

Variance (201)

In Section 9.1 we will encounter the principle of Russian roulette once more when developing Monte Carlo path tracing.

6.7.2 A MONTE CARLO APPROACH BASED ON THE NEUMANN SERIES APPROACH

Section 2.3

An alternative method for solving a Fredholm integral equation of the 2^{nd} kind is the Neumann series approach introduced in Section 2.3.3.1.1. Contrary to the technique of naive successive substitution from the preceding section, where we replaced the estimated solution at each step, we will now iterate replacements on the operator. This leads to a more elegant method leading to the same result, but from a different point of view.

Linear Operator (53)

Equation (2.387) of our discussion about the Neumann series approach for solving Fredholm integral equation of the 2^{nd} kind shows that a solution of a Fredholm type integral equation is given by:

$$f(x) = \sum_{j=0}^{\infty} \mathbf{K}^j g(x) \tag{6.592}$$

$$= g(\mathbf{x}) + \mathbf{K}\, g(\mathbf{x}) + \mathbf{K}^2\, g(\mathbf{x}) + \mathbf{K}^3\, g(\mathbf{x}) + \ldots, \tag{6.593}$$

Source Function (127)

where g is the source function, $\mathbf{K}$ is a contracting, linear integral operator, and for $l > 0$ the l^{th}-term in Equation (6.593) represents the integral

Linear Integral Operator (130)

$$\int_{\mathcal{R}} \ldots \int_{\mathcal{R}} \prod_{k=1}^{l} \mathbf{k}(\mathbf{x}_{k-1}, \mathbf{x}_k)\, g(\mathbf{x}_k)\, d\mu(\mathbf{x}_1) \cdot \ldots \cdot d\mu(\mathbf{x}_k) \tag{6.594}$$

with $g(\mathbf{x})$ as the 0^{th}-term.

Monte Carlo Estimator (499)

Now, any of these integrals can be estimated by a Monte Carlo estimator $F_l^{\mathbf{K}^l\, g(\mathbf{x})}$

using independent pairs $(\mathbf{X}_{0i_1\dots i_l})$ of random variables distributed according to probability densities p_l given by:

$$F_l^{\mathbf{K}^l g(\mathbf{x})} = \sum_{i_1=1}^{N_1} \cdots \sum_{i_l=1}^{N_l} \left(\prod_{j=1}^{l} \frac{1}{N_j} \frac{\mathbf{k}(\mathbf{X}_{0i_1\dots i_{j-1}}, \mathbf{X}_{0i_1\dots i_j})}{p_0(\mathbf{X}_0) p_j(\mathbf{X}_{0i_1\dots i_j})} \right) g(\mathbf{X}_{0i_1\dots i_l}), \tag{6.595}$$

with $p_0(\mathbf{X}_0) = 1$. PDF (176)

A secondary Monte Carlo estimator for approximating a Fredholm integral equation of the 2^{nd} kind based on M terms of the Neumann series is then given by:

$$\begin{aligned} F_N^{f(\mathbf{x})} &= \frac{g(\mathbf{X})}{p_0(\mathbf{X}_0)} + \sum_{l=1}^{M} F_l^{\mathbf{K}^l g(\mathbf{x})} && (6.596) \\ &= \frac{g(\mathbf{X})}{p_0(\mathbf{X}_0)} + && (6.597) \\ &\quad \sum_{l=1}^{M} \left\{ \sum_{i_1=1}^{N_1} \cdots \sum_{i_l=1}^{N_l} \left(\prod_{j=1}^{l} \frac{1}{N_l} \frac{\mathbf{k}(\mathbf{X}_{0i_1\dots i_{j-1}}, \mathbf{X}_{0i_1\dots i_j})}{p_0(\mathbf{X}_0) p_j(\mathbf{X}_{0i_1\dots i_j})} \right) g(\mathbf{X}_{0i_1\dots i_l}) \right\}. \end{aligned}$$

From this formula, we can conclude that both approaches, the successive integral substitution as well as the Neumann series approach, lead to the same result. The difference between these two approaches is that successive integral substitution can be interpreted to have a rather intuitive and practical background, while the Neumann series approach is based on theoretical, functional analytical concepts. So, the Neumann series approach, as a more elegant method, requires also knowledge about linear operator theory from functional analysis, which is not needed in the successive integral substitution. Here, we have only to estimate the unknown function g by a simple Monte Carlo scheme using furthermore estimations at many other points. Section 2.1.4

6.7.3 A Probabilistic Approach Based on a Discrete Markov Process

In Example 2.98 we have shown that a linear system of equations, under certain conditions, can be solved via a discrete-time Markov chain. In this example, we rephrased the solution of a linear system of type DT Markov Chain (226)

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{6.598}$$

into a Neumann series Neumann Series (135)

$$\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}, \tag{6.599}$$

where $\mathbf{I}$ and $\mathbf{A}$ were finite-dimensional operators, i.e. matrices, and $\mathbf{x}$ is a vector, which was stochastically computed via a discrete-time Markov chain.

Now, since a Fredholm type integral equation can be written as

$$f(x) = g(x) + \mathbf{K}f(x), \tag{6.600}$$

it looks just like a linear system of equations, where a solution to this operator equation is also given in form of a Neumann series. This then suggests to apply the principle of the discrete-time stochastic process to operator equations for solving Fredholm integral equations of the 2^{nd} kind.

Neumann Series (135)
Fredholm Integral Equations (127)

For the following discussion, let us assume that $(\mathcal{R}, \mathfrak{R}(\mathcal{R}), \mathbb{P})$ is a probability space with uncountable base set $\mathcal{R}$. In analogy to the initial distribution p_0 and the stochastic matrix $(p_{ij})_{1 \leq i,j \leq s}$ from Example 2.98, we now define an *initial probability density function* p_0 and a *transition kernel* p, since we are working on a continuous probability space.

Probability Space (163)
Uncountable Set (827)
Stochastic Matrix (229)
Transition Kernel (234)

We define the *initial probability density function* p_0 such that it holds:

DT Markov Process (236)

$$p_0(\mathbf{x}) > 0 \tag{6.601}$$

with $\int_{\mathcal{R}} p_0(\mathbf{x}) d\mu(\mathbf{x}) = 1$, and the *probability transition kernel* p by:

$$p(\mathbf{y}|\mathbf{x}) > 0, \qquad \text{if } \mathbf{k}(\mathbf{x}, \mathbf{y}) \neq 0, \tag{6.602}$$

with $\int_{\mathcal{R}} p(\mathbf{y}|\mathbf{x}) d\mu(\mathbf{y}) = 1$ for $k > 1$.

In analogy to Example 2.98, let us then furthermore generate a random walk $\mathbf{X}_0 \to \mathbf{X}_1 \to \mathbf{X}_2 \ldots \to \mathbf{X}_m$ with associated random variable $\mathbf{Y}_m$ via a discrete-time Markov process defined over the continuous state set $\mathcal{R}$ by:

DT Markov Process (236)

$$\mathbf{Y}_m \overset{def}{=} \frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} + \sum_{i=1}^{m} \left(\frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_1)}{p_0(\mathbf{X}_0)\, p(\mathbf{X}_1|\mathbf{X}_0)} \prod_{k=1}^{i-1} \frac{\mathbf{k}(\mathbf{X}_k, \mathbf{X}_{k+1})}{p(\mathbf{X}_{k+1}|\mathbf{X}_k)} \right) g(\mathbf{X}_i), \tag{6.603}$$

where $p(\mathbf{X}_{k+1}|\mathbf{X}_k)$ is the conditional probability density that $\mathbf{X}_{k+1}$ is sampled in step $k+1$ under the condition that $\mathbf{X}_k$ was sampled in step k, see Figure 6.43.

Since this random walk starts at point $\mathbf{X}_0$, chosen according to the initial PDF p_0, and $p(\mathbf{X}_{k+1}|\mathbf{X}_k)$ is the probability for a transition from point $\mathbf{X}_k$ to point $\mathbf{X}_{k+1}$, the probability for generating the random walk $\mathbf{X}_0 \to \mathbf{X}_1 \to \ldots \to \mathbf{X}_i$ corresponds to the product of the initial probability for choosing $\mathbf{X}_0$ and the transition probabilities $p(\mathbf{X}_{k+1}|\mathbf{X}_k), 1 \leq k \leq i-1$, namely:

$$p_0(\mathbf{X}_0)\, p(\mathbf{X}_1|\mathbf{X}_0) \prod_{k=1}^{i-1} p(\mathbf{X}_{k+1}|\mathbf{X}_k). \tag{6.604}$$

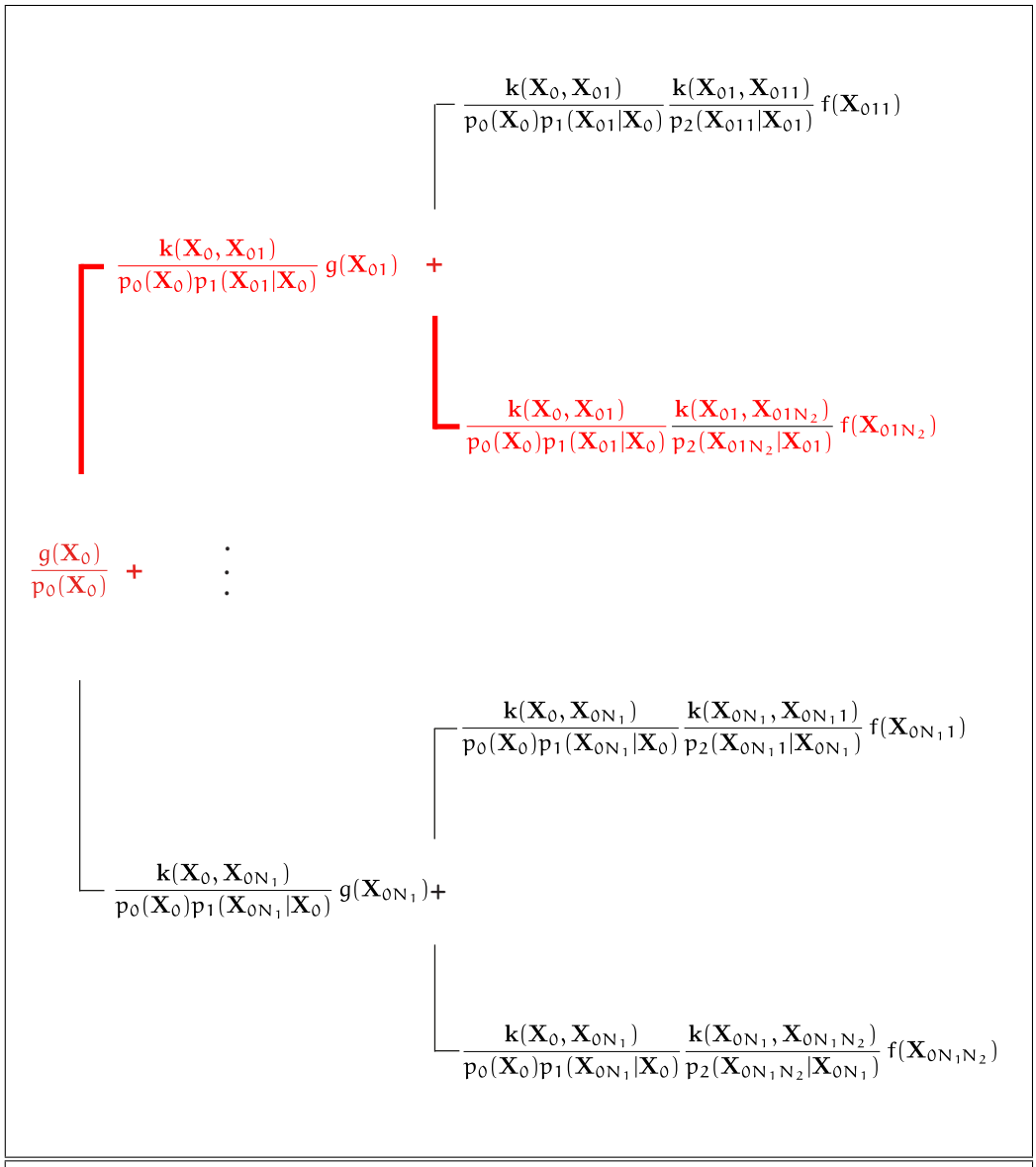


FIGURE 6.43: DISCRETE MARKOV PROCESS FOR APPROXIMATING A FREDHOLM INTEGRAL EQUATION OF THE 2^{nd} KIND. For evaluating a Fredholm integral equation of the 2^{nd} kind, a Markov process evaluates only a single path of the computation tree starting at the root. As a single path only contributes very little information to a solution, a Markov process computes an average value over the information returned by a great number of such paths. Note, the tree, as shown above, only represents the first two levels of the integral substitution in a Fredholm equation.

Now, we can show, that the expected value of $\mathbf{Y}_m$ for $m \to \infty$ converges towards the proper value of the Neumann series since it holds:

Neumann Series (135)

$$
\begin{aligned}
&\mathrm{E}\left(\lim_{m\to\infty}\mathbf{Y}_m\right)\\
&= \mathrm{E}\left(\frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)}+\sum_{i=1}^{\infty}\left(\frac{k(\mathbf{X}_0,\mathbf{X}_1)}{p_0(\mathbf{X}_0)\,p(\mathbf{X}_1|\mathbf{X}_0)}\prod_{k=1}^{i-1}\frac{k(\mathbf{X}_k,\mathbf{X}_{k+1})}{p(\mathbf{X}_{k+1}|\mathbf{X}_k)}\right)g(\mathbf{X}_i)\right) && (6.605)\\
&= g(\mathbf{X}_0)+\sum_{i=1}^{\infty}\mathrm{E}\left(\left(\frac{k(\mathbf{X}_0,\mathbf{X}_1)}{p_0(\mathbf{X}_0)\,p_k(\mathbf{X}_1|\mathbf{X}_0)}\prod_{k=1}^{i-1}\frac{k(\mathbf{X}_k,\mathbf{X}_{k+1})}{p_k(\mathbf{X}_{k+1}|\mathbf{X}_k)}\right)g(\mathbf{X}_i)\right) && (6.606)\\
&\overset{(2.735)}{=} g(\mathbf{X}_0)+\\
&\quad\sum_{i=1}^{\infty}\underbrace{\int_{\mathcal{R}}\dots\int_{\mathcal{R}}\left(k(\mathbf{x}_0,\mathbf{x}_1)\prod_{k=1}^{i-1}k(\mathbf{x}_k,\mathbf{x}_{k+1})\,g(\mathbf{x}_i)\,d\mu(\mathbf{x}_2)\dots d\mu(\mathbf{x}_{k+1})\right)}_{\mathbf{K}^i g(\mathbf{x})} && (6.607)\\
&= g(\mathbf{x})+\sum_{i=1}^{\infty}\mathbf{K}^i g(\mathbf{x}) && (6.608)\\
&= \sum_{i=0}^{\infty}\mathbf{K}^i g(\mathbf{x}). && (6.609)
\end{aligned}
$$

So, the approach for solving a Fredholm integral equation by an algorithm based on the construction of a discrete-time Markov process is identically to the procedure of solving the corresponding integral operator equation via the Neumann series approach. Due to the SLLN we can approximate $f(x)$ by generating and averaging N independent random walks $\mathbf{Y}_m^{(k)}, 1 \le k \le N$ resulting from a discrete-time Markov process, that is:

SLLN (216)

$$
f(x) \approx \frac{1}{N}\sum_{k=1}^{N}\mathbf{Y}_m^{(k)}. \tag{6.610}
$$

REMARK 6.35 *Such a random walk solution method follows the approach to construct a path, starting at the root of the computation tree of* f *where only those terms of* $(\mathbf{I}-\mathbf{K})^{-1}g(\mathbf{x})$ *are computed that are associated with the currently visited nodes, see Figure 6.43. Compared to our naive approach for evaluating a Fredholm equation of the* 2^{nd} *kind from Section 6.7, a discrete-time Markov process requires the evaluation of a random variable once only per iteration. This means, that a solution can be achieved relatively fast, but the quality of a solution is rarely satisfactory. Evidently, the advantage due to the run time is then compensated by generating and averaging the contributions from many random walks.*

REMARK 6.36 *The above random walk technique is not computationally affordable, since infinitely many terms have to be evaluated. A more efficient method could be*

to focus the evaluation on terms with small indices in the Neumann series, since these terms contribute a large amount to the final result, while terms with large indices within the Neumann series contribute relatively small. To guarantee that the method is at least consistent, these terms can then be weighted more heavily via Russian roulette.

Russian Roulette (200)

For that purpose, the state space of the Markov process must be expanded by a new element, the so-called absorption state*, $\dagger$, such that the transition from $\dagger$ to any other state is zero and it must hold:* $p(\dagger,\dagger) = 1$ *and* $P_0(\dagger) = p_{\dagger j} = 0$ *for any state* $j \neq \dagger$.

6.7.4 NEXT EVENT ESTIMATION

According to our analysis of Monte Carlo estimators, and especially from our consideration with respect to variance reduction techniques from Section 6.6, we know that the probability density, involved in a Monte Carlo estimator, has a strong influence on the variance of the procedure. Hence, to obtain estimators that guarantee small variance, it is required to choose densities according to variance reduction techniques. With respect to linear integral equations this means that due to their recursive structure and the fact, that the integrands contain unknown functions, informed Monte Carlo methods are not really available as efficient variance reduction techniques. But there is an other technique that can help us: *Next Event Estimation*

Monte Carlo Estimators (499)
Probability Density Function (176)
Variance (201)
Linear Integral Equations (126)

Let us consider a Fredholm integral equation of the 2^{nd} kind a little bit closer. Then, we see that it is composed of two parts: the known source function g and an integral over the unknown function f, namely:

Fredholm Integral Equations (127)

$$f(\mathbf{x}) = g(\mathbf{x}) + \underbrace{\int_{\mathcal{R}} \mathbf{k}(\mathbf{x},\mathbf{x}_1)\, f(\mathbf{x}_1)\, d\mu(\mathbf{x}_1)}_{h(\mathbf{x})}. \tag{6.611}$$

Writing the integral as a new function h in the variable $\mathbf{x}$, then f is the composition of a known and an unknown function, that is,

$$f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}). \tag{6.612}$$

We can now use the information stored in the driving function g for evaluating the integral in the Fredholm equation, namely:

$$\begin{aligned} h(\mathbf{x}) &= \int_{\mathcal{R}} \mathbf{k}(\mathbf{x},\mathbf{x}_1)\, f(\mathbf{x}_1)\, d\mu(\mathbf{x}_1) && (6.613)\\ &= \int_{\mathcal{R}} \mathbf{k}(\mathbf{x},\mathbf{x}_1)\, \left(g(\mathbf{x}_1) + h(\mathbf{x}_1)\right)\, d\mu(\mathbf{x}_1) && (6.614)\\ &= \int_{\mathcal{R}} \mathbf{k}(\mathbf{x},\mathbf{x}_1)\, g(\mathbf{x}_1)\, d\mu(\mathbf{x}_1) + \int_{\mathcal{R}} \mathbf{k}(\mathbf{x},\mathbf{x}_1)\, h(\mathbf{x}_1)\, d\mu(\mathbf{x}_1). && (6.615) \end{aligned}$$

EXAMPLE 6.45 (Direct and Indirect Illumination) *Let us consider the SLTEV in spherical form based on exitant radiance. With respect to the surface point* $\mathbf{s}$ *it can be written as sum of the emitted radiance at point* $\mathbf{s}$ *in direction* ω_o *and the scattering equation from Relation (4.311), namely:*

SLTEV (404)

$$L_o(\mathbf{s},\omega_o) = L_e(\mathbf{s},\omega_o) + \underbrace{\int_{S^2(\mathbf{s})} f_s(\mathbf{s},\omega_i \to \omega_o)\, L_o(\gamma(\mathbf{s},\omega_i),-\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i)}_{h_o(\mathbf{s},\omega_o)} \quad (6.616)$$

$$= L_e(\mathbf{s},\omega_o) + h_o(\mathbf{s},\omega_o), \quad (6.617)$$

where the scattering equation is denoted by the exitant radiance function h_o.

Replacing the exitant radiance $L_o(\gamma(\mathbf{s},\omega_i),-\omega_i)$ *in* h_o *by the exitant SLTEV in spherical form at point* $\gamma(\mathbf{s},\omega_i)$ *in direction* $-\omega_i$, *that is,*

$$L_o(\gamma(\mathbf{s},\omega_i),-\omega_i) = L_e(\gamma(\mathbf{s},\omega_i),-\omega_i) + h_o(\gamma(\mathbf{s},\omega_i),-\omega_i), \quad (6.618)$$

then $h(\mathbf{s},\omega_o)$ *can be written as:*

$$h_o(\mathbf{s},\omega_o) = \int_{S^2(\mathbf{s})} f_s(\mathbf{s},\omega_i \to \omega_o)\, \big(L_e(\gamma(\mathbf{s},\omega_i),-\omega_i) + h_o(\gamma(\mathbf{s},\omega_i),-\omega_i)\big)\, d\sigma^\perp_{\mathbf{s}}(\omega_i) \quad (6.619)$$

$$= \int_{S^2(\mathbf{s})} f_s(\mathbf{s},\omega_i \to \omega_o)\, L_e(\gamma(\mathbf{s},\omega_i),-\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s},\omega_i \to \omega_o)\, h_o(\gamma(\mathbf{s},\omega_i),-\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i). \quad (6.620)$$

In Example 4.15 we discussed direct and indirect illumination at a surface point $\mathbf{s}$ *by splitting the unit sphere around* $\mathbf{s}$ *in a disjoint union of two sets: The projection of all regions of light sources, visible from the center of the unit sphere, onto the unit sphere, thus the set* $☼^\perp$, *and its complement, the set* $\overline{☼^\perp} = S^2 \setminus ☼^\perp$.

Since the integrand of the first integral is nonzero only for light sources, the integration domain of the first integral can be reduced to the set $☼^\perp$. *The integrand of the second integral describes light that, scattered at least once at surfaces within the scene, arrives at point* $\mathbf{s}$ *from where it is scattered in direction* ω_o, *see Figure 6.44. Relevant for the evaluation of this integral are only the non-emitting surfaces, that is, the integration domain of the second integral can be reduced* $\overline{☼^\perp}$. *Applying these fact to the above two integrals, then the function* h_o *can be written as*

$$h_o(\mathbf{s},\omega_o) = \int_{☼^\perp} f_s(\mathbf{s},\omega_i \to \omega_o)\, L_e(\gamma(\mathbf{s},\omega_i),-\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i) + \int_{\overline{☼^\perp}} f_s(\mathbf{s},\omega_i \to \omega_o)\, h_o(\gamma(\mathbf{s},\omega_i),-\omega_i)\, d\sigma^\perp_{\mathbf{s}}(\omega_i). \quad (6.621)$$

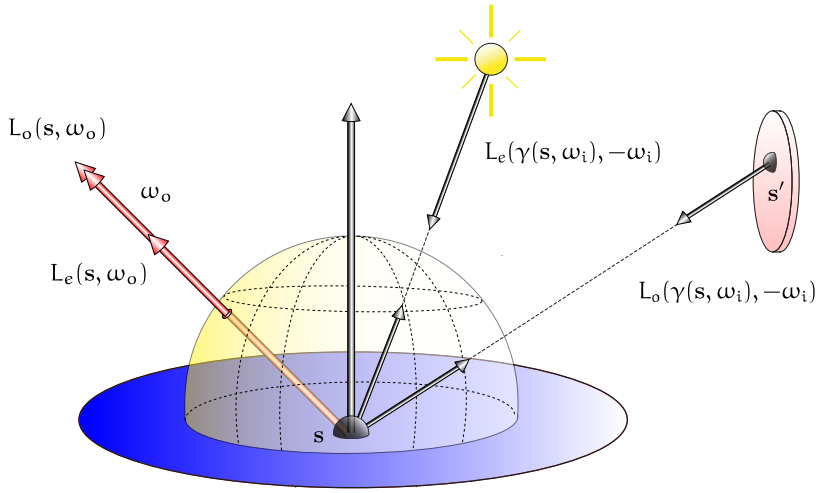


FIGURE 6.44: DIRECT AND INDIRECT ILLUMINATION AT A SURFACE POINT. Point s is directly illuminated by a point light source, additionally it receives also light that is scattered at point $\gamma(s, \omega_i)$ in direction $-\omega_i$, thus light that comes from surfaces that are not light sources.

Due to the principle of radiance invariance then the exitant function h_o *can be expressed in terms of incident radiance, thus,*

$$h_o(\gamma(s, \omega_i), -\omega_i) = h_i(s, \omega_i). \tag{6.622}$$

Using this relation in the above representation of the function h_o *then we get:* Radiance Invariance (253)

$$h_o(s, \omega_o) = \int_{\overset{\perp}{\odot}} f_s(s, \omega_i \to \omega_o)\, L_e(\gamma(s, \omega_i), -\omega_i)\, d\sigma^{\perp}_s(\omega_i) + \tag{6.623}$$

$$\int_{\overline{\overset{\perp}{\odot}}} f_s(s, \omega_i \to \omega_o)\, h_i(s, \omega_i)\, d\sigma^{\perp}_s(\omega_i). \tag{6.624}$$

We know these two integrals from our discussion about the direct and indirect illumination formulation of the stationary light transport equation in a vacuum from Section 4.4.2.2. They correspond to the direct illumination, $L^{\leftarrow}(s, \omega^i_o)$*, and the indirect illumination,* $L^{\Leftarrow}(s, \omega^i_o)$*, from Relation (4.407), that is, the technique of next event estimation is equivalent to the stratification of the integration domain of the SLTEV.*

Provided that it is not analytically solvable, the decomposition of the function h in the two integrals from Equation (6.615) then suggests to apply a variance reduction technique to the first integral while the second part of Equation (6.615) must be solved by one of the methods introduced in the previous sections. Both possibilities for solving the first integral in Equation (6.615), analytical solution or variance reduction techniques, lead to variance reduction in the result. A secondary Monte Carlo estimator for $f(\mathbf{x})$ is then given by:

$$F_N^{f(\mathbf{x})} = g(\mathbf{x}) + F_N^{h(\mathbf{x})} \tag{6.625}$$

with $\mathbf{X}_0 = \mathbf{x}$ and

$$F_N^{h(\mathbf{x})} = \frac{1}{N_1} \sum_{i_1=1}^{N_1} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)\, p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)} \left(g(\mathbf{X}_{0i_1}) + F_N^{h(\mathbf{x}_{0i_1})} \right) \tag{6.626}$$

$$= \frac{1}{N_1} \sum_{i_1=1}^{N_1} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)\, p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)} g(\mathbf{X}_{0i_1}) + \tag{6.627}$$

$$\frac{1}{N_1} \sum_{i_1=1}^{N_1} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)\, p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)} F_N^{h(\mathbf{x}_{0i_1})} \tag{6.628}$$

$$F_N^{h(\mathbf{x}_{0i_1})} = \frac{1}{N_2} \sum_{i_2=1}^{N_2} \frac{\mathbf{k}(\mathbf{X}_1, \mathbf{X}_{0i_1i_2})}{p_2(\mathbf{X}_{0i_1i_2}|\mathbf{X}_{0i_1})} \left(g(\mathbf{X}_{0i_1i_2}) + F_N^{h(\mathbf{x}_{0i_1i_2})} \right) \tag{6.629}$$

$$\vdots$$

$$F_N^{h(\mathbf{x}_{0i_1\ldots i_M})} = \frac{1}{N_M} \sum_{i_1=1}^{N_M} \frac{\mathbf{k}(\mathbf{X}_{0i_1\ldots M-1}, \mathbf{X}_{0i_1\ldots i_M})}{p_M(\mathbf{X}_{0i_1\ldots i_M}|\mathbf{X}_{0i_1\ldots i_{M-1}})} \left(g(\mathbf{X}_{0i_1\ldots i_M}) + \ldots \right. \tag{6.630}$$

Going back to the nomenclature of the previous sections a secondary Monte Carlo estimator for the integral equation from Equation (6.611), based on next event estimation, is then given by:

$$F_N^{f(\mathbf{x})} = g(\mathbf{x}) + \frac{1}{N_1} \sum_{i_1=1}^{N_1} \frac{\mathbf{k}(\mathbf{X}_0, \mathbf{X}_{0i_1})}{p_0(\mathbf{X}_0)\, p_1(\mathbf{X}_{0i_1}|\mathbf{X}_0)} g(\mathbf{X}_{0i_1}) + \tag{6.631}$$

$$\sum_{l=2}^{M} \left\{ \sum_{i_1=1}^{N_1} \cdots \sum_{i_l=1}^{N_l} \left(\prod_{j=1}^{l} \frac{1}{N_j} \frac{\mathbf{k}(\mathbf{X}_{0i_1\ldots i_{j-1}}, \mathbf{X}_{0i_1\ldots i_j})}{p_0(\mathbf{X}_0) p_j(\mathbf{X}_{0i_1\ldots i_j}|\mathbf{X}_{0i_1\ldots i_{j-1}})} \right) g(\mathbf{X}_{0i_1\ldots i_l}) \right\}.$$

Next event estimation is an extreme interesting technique that plays an important role in field of global illumination algorithms. We will encounter this technique in Chapter 9 again and again. So, we will show, how we get better, and more efficient variants of these algorithms by extending the classic *gathering* and *shooting algorithms* by the technique of next event estimation. A typical example for such an extension of a rendering algorithm

Section 4.4.2.2

is to involve direct illumination into pure-Monte Carlo path tracing. This leads to more realistic images with no much more effort.

EXAMPLE 6.46 (Sampling Direct Illumination) *Let us now turn to the problem of computing the direct illumination at a surface point, that is, we are interested in the evaluation of the integral*

$$L_o^{\leftarrow}(\mathbf{s}_j, \omega_o^j) \overset{(6.623)}{=} \int_{\mathfrak{S}^\perp} f_s(\mathbf{s}_j, \omega_i^j \to \omega_o^j)\, L_e(\gamma(\mathbf{s}_j, \omega_i^j), -\omega_i^j)\, d\sigma_{\mathbf{s}_j}^{\perp}(\omega_i^j). \tag{6.632}$$

For evaluating $L_o^{\leftarrow}(\mathbf{s}_j, \omega_o^j)$ via a Monte Carlo strategy, we have to sample the integration domain $\mathfrak{S}^\perp$ in a clever and efficient manner. Now, developing such a sampling strategy requires a closed formulation of the integration domain, which due to [10, Arvo 1995] can be a very tricky and complicated task.

Since the position, orientation, and the shape of light sources in a scene are always known, instead to compute the projections of the light sources onto the unit sphere, we can also use the set of all surfaces as domain of integration. This then requires the representation of the direct illumination as a surface integral, that is, we have to transform the spherical integral, defined by $L_o^{\leftarrow}(\mathbf{s}_j, \omega_o^j)$, into its 3-point form. Replacing the projected solid angle measure in Equation (6.632) by the Lebesgue area measure—for a detailed derivation see Example 2.51—and using the standard notation for describing integral equations in 3-point form, the direct illumination at point $\mathbf{s}_j$ in direction ω_o^j can then be written as:

$$L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1}) = \int_{\mathfrak{S}} f_s(\mathbf{l} \to \mathbf{s}_j \to \mathbf{s}_{j-1})\, L_e(\mathbf{l} \to \mathbf{s}_j)\, \mathcal{G}(\mathbf{l} \leftrightarrow \mathbf{s}_j),\, d\mu^2(\mathbf{l}), \tag{6.633}$$

where $\gamma(\mathbf{s}_j, \omega_o^j) = \mathbf{s}_{j-1}$, while the indirect illumination is still written as a spherical integral.

Based on this integral formulation, we can sample so-called shadow-rays *by choosing points $\mathbf{l}_i$ according to a probability density $p_{\mathfrak{S}}$ on the probability$(\mathfrak{S}, \mathfrak{B}(\mathfrak{S}), \mu^2)$ space, where, in the simplest case, $p_{\mathfrak{S}}$ corresponds to uniform sampling of light source area. An associated secondary Monte Carlo estimator for approximating $L_o^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1})$ then has the form:*

PDF (176)

Probability Space (163)

$$F_N^{L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1})} = \sum_{i=1}^{N} \frac{f_s(\mathbf{l}_i \to \mathbf{s}_j \to \mathbf{s}_{j-1})\, L_e(\mathbf{l}_i \to \mathbf{s}_j)\, \mathcal{G}(\mathbf{l}_i \leftrightarrow \mathbf{s}_j)}{p_{\mathfrak{S}}(\mathbf{l}_i)}. \tag{6.634}$$

It should be clear, that the above estimator contains several sources for noise in the resulting images. Mainly responsible for this is the geometry term $\mathcal{G}$, which is composed of the visibility function $\mathcal{V}$, two cosine-terms, and the distance of two surface points within the denominator. So, we get noise in an image at points lying in the penumbra region of a shadow areas since some of the shadow rays can hit a

$\mathcal{G}$ (129)

light source while others does not hit a light source. For points outside of shadow areas, noise can also be arise due to one of the cosine-terms, or the denominator of the geometry term.

6.8 REFERENCE LITERATURE AND FURTHER READING

This chapter about Monte Carlo procedures is mainly based on the known standard works [73, Hammersley & Handscomp 1964], [172, Rubinstein 1981], [99, Kalos & Whitlock 1986], and [63, Gentle 1998]. For those with a good background knowledge in measure and probability theory we recommend the books by [60, Fishman 1996] and [170, Robert & Casella 1999], for newcomers, which are more interested in practical application, [197, Sobol 1985] should be an ideal starting point. A good source of background information specially for globillumers are the dissertations of [47, Dutré 1996], [116, Lafortune 1996], [221, Veach 1998], [209, Szirmay-Kalos 2000], and [194, Slusallek 2000]. Many interesting examples for demonstrating the concept of sampling can be found in Eric Veach's course *CS 448: Topics in Computer Graphics* [220, Veach 1997], Philippe Dutré's *Global Illumination Compendium* [50, Dutré 2003], [51, Dutré & al. 2006], [185, Shirley 2000], [187, Shirley and Morley 2003], and in the PBRT-book by [158, Pharr & Humphreys 2004], [159, Pharr & Humphreys 2010]. All three references were of great help for us. For the studies of other sampling strategies, not related to the field of global illumination, we recommend [172, Rubinstein 1981], [99, Kalos & Whitlock 1986], and [60, Fishman 1996] to the interested reader.

We recommend [179, Schmeisser and Schirmeier 1976] and [202, Stoer 1979] as textbooks on numerical mathematics, which treat numerical integration on a high-level. For a rather advanced insight into the theory of numerical integration, see [57, Evans & Swartz 2000]. This book requires knowledge from measure theory.

Section 6.5.3 is intended as a minimalist refresher on Markov chain Monte Carlo methods. If you have doubts or want more details about this method, you are strongly advised to check a more thorough treatment such as [170, Robert & Casella 1999], [65, Gilks & al. 1996], and in particular [130, Meyn and Tweedie 1993] as well as [204, Stroock 2005], since no theory of convergence is provided in our book.

A beautiful article that describes random walk solutions to Fredholm integral equations is [41, Daucet & al. 2010]. Random walk solutions to integral equations are also treated in [172, Rubinstein 1981], [99, Kalos & Whitlock 1986], [68, Glassner 1995], and [60, Fishman 1996].

Multiple importance sampling is described in detail in [221, Veach 1998].

CHAPTER SEVEN

QUASI-MONTE CARLO INTEGRATION

As observed in the foregoing chapter Monte Carlo integration suffers—besides their slow convergence rate—from the disadvantages that only probabilistic statements on convergence and error boundaries are possible. Additionally, the *quality* of the process depends on the random numbers used. So, the success of any Monte Carlo procedure stands or falls with the quality of these random samples, where we use the notion of quality to make a statement about the true randomness of the random samples. Now, it is not only this disadvantage which is inherent in the stochastic nature of the method but also a detailed analysis in [138, Niederreiter 1992] has shown that it is not the true randomness of the samples which is so relevant, but rather the uniform distribution of the random samples over the integration domain. In addition, this analysis shows that deterministic error bounds can be formulated if the samples are selected in a deterministic way. In principle this implies, that it is possible in advance, to generate an integration rule that yields a given accuracy. This then suggest the idea of generating samples in a deterministic way with error bounds as small as possible: The fundamental principle of *quasi-Monte Carlo Integration*. While ordinary Monte Carlo integration, using n random variables, yields a convergence rate of $O\left(\frac{1}{\sqrt{n}}\right)$ in any dimension, regardless of the smoothness of the integrand, quasi-Monte Carlo integration, applied to smooth functions and using low-discrepancy point sequences, yields convergence rates on the scale of $O\left(\frac{\log^{s-1} n}{n}\right)$, where s denotes the dimension of the involved integrand.

Chapter 6
Section 6.4
Section 6.6
Uniform Distribution (180)

Simply spoken, a quasi-Monte Carlo method can be considered as a Monte Carlo procedure, where the random samples are replaced by well-chosen deterministic points. Instead to draw sequences of random numbers on the integration domain $\mathbf{Q}^s$, deterministic sequences $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathbf{Q}^s$ are chosen, that cover the integration domain more uniformly than the randomly selected samples in Monte Carlo Integration. So, this process decreases the effect of clumping in samples, see Figure 7.1, and it delivers deterministic error bounds that are simpler to interpret, and in addition, smaller than the stochastic error bounds from random sampling.

Random Variable (168)

As in Monte Carlo integration, also in quasi-Monte Carlo integration, the integral

Lebesgue Integral (105)

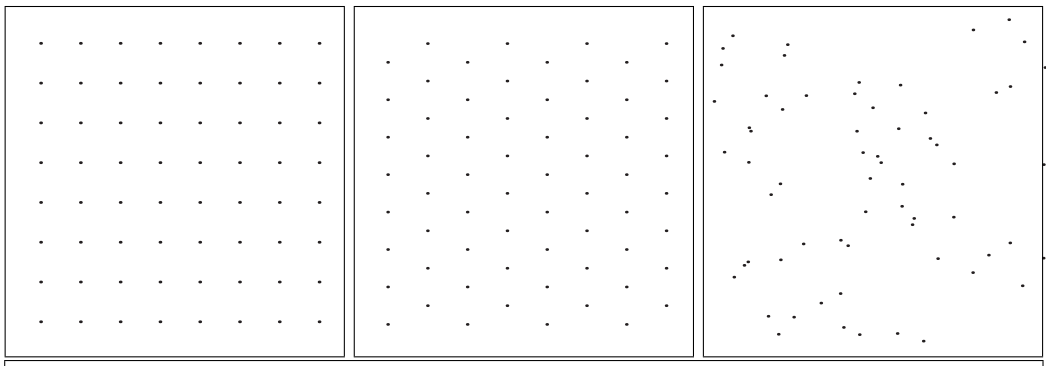


FIGURE 7.1: REGULAR AND HEXAGONAL GRID, AS WELL AS A POISSON PATTERN ON THE UNIT INTERVAL $\mathbf{I}^2 = [0,1] \times [0,1]$. 64 element, 2-dimensional point sets: A regular grid, a hexagonal grid, and a pattern generated via Poisson sampling.

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}), \tag{7.1}$$

μ^s (82) must be evaluated, where, as usual, μ^s denotes the Lebesgue measure on $\mathbb{R}^s$ and the integration domain $\mathbf{Q}^s$ is a finite subset of $\mathbb{R}^s$, that is: $0 < \mu^s(\mathbf{Q}^s) < \infty$.

Now, this integral can be computed either via a sequence of deterministic points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathbf{I}^s$, which has to be transformed into a sequence $T(\mathbf{x}_1), T(\mathbf{x}_2), \ldots, T(\mathbf{x}_N) \in \mathbf{Q}^s$ for evaluating the sum

$$\frac{1}{N} \sum_{i=1}^{N} f(T(\mathbf{x}_i)) \tag{7.2}$$

or by transforming the Integral (7.30) into an integral of type

$$\int_{\mathbf{I}^s} \bar{f}(\mathbf{x})\, d\mu^s(\mathbf{x}) \tag{7.3}$$

and evaluating the sum

$$\frac{1}{N} \sum_{i=1}^{N} \bar{f}(\mathbf{x}_i) \tag{7.4}$$

at the deterministically determined points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathbf{I}^s$.

OVERVIEW OF THIS CHAPTER. The point of departure of the following excursion into quasi-
Section 7.1 Monte Carlo integration is the concept of *discrepancy*, which can be interpreted as a quantitative measure for the deviation of a given point set from its uniform distribution. Then, we introduce various types of discrepancy, discuss their basic properties, and present the

Koksma-Hlawka Inequality, the central result of quasi-Monte Carlo integration. It shows the path to be taken for the construction of so-called *low descrepancy point sequences* as well as *low-descrepancy point sets*, which, used in quasi-Monte Carlo procedures, guarantee deterministic and small error bounds. Thereafter, we discuss the construct of *low-discrepancy sequences* and present with the s-*dimensional Halton Sequence* and the s-*dimensional Hammersley Point Set* first examples of low-discrepancy point sets and low-discrepancy point sequences. Then, we introduce further examples of low-discrepancy sequences and we show how it is possible by means of *scrambling procedures* to remove the regular structures in low-discrepancy sequences of higher dimensions. Following this, the theory of the currently most promising low-discrepancy sequences will be introduced: (t, m, s)-*nets* and (t, s)-*sequences* constructions, which were already shortly mentioned in the discussion of stratified sampling. We also throw a glance at the construction of *randomised* (t, m, s)-*nets* and *randomised* (t, s)-*sequences*. The chapter will be concluded by demonstrating the importance of the concept of *Fourier analysis* as a method for analyzing and interpreting the most important sampling processes applied in Monte Carlo and quasi-Monte Carlo procedures.

Section 7.2

Section 7.3

7.1 DISCREPANCY

In any procedure for numerical integrating a function, the error between the exact and the approximated solution should be minimized by a good choice of the used sample points. Intuitively, this error is based on two independent factors: the choice of the samples in the integration domain and how quickly the function changes its values between these sample points.

DISCREPANCY. If the distribution of the sample points is not uniform, then there are large regions where there are no sample points at all, which can increases the error, see Figure 7.1. Closely related to this is the fact that a smooth function is evaluated at unnecessary many locations if samples are clumped. So, discussions on uniformly distributed random numbers in [138, Niederreiter 1992] have shown that it is not randomness but the uniform distribution of a sequence of samples on an interval, which is necessary for the convergence of the estimated value. This leads to the idea of selecting samples in a deterministic way under the condition that the error bound can be made as small as possible. In this context, uniformly distributed refers to the fact that in a subvolume of the s-dimensional unit cube, $\mathbf{I}^s$, the relation between the samples, contained in this subvolume, and the total number of samples in $\mathbf{I}^s$ differs only slightly from the fraction of the volume of the sub-domain and the volume of $\mathbf{I}^s$. This condition may be expressed via the concept of *discrepancy*.

DEFINITION 7.1 (Discrepancy) *Let* $\mathbf{P} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ *with* $\mathbf{x}_i \in \mathbf{I}^s, i = 1, \ldots, N$ *be a point set. The* discrepancy *of* $\mathbf{P}$*, denoted as* $D_N(\mathbf{P})$*, is a measure for the deviation*

of a point set from its ideal uniform distribution. The discrepancy of $\mathbf{P}$ *is defined as:*

$$D_N(\mathbf{P}) \equiv D_N(\mathbf{P}, \mathcal{B}) \tag{7.5}$$

$$\stackrel{def}{=} \sup_{\mathbf{B}\in\mathcal{B}} \left| \frac{\#(\mathbf{P}\cap\mathbf{B})}{N} - \mu^s(\mathbf{B}) \right|, \tag{7.6}$$

Lebesgue Measurable Set (75) #(·) (81) μ^s (82)

where $\mathcal{B}$ *corresponds to a Lebesgue measurable family of subsets of* $\mathbf{I}^s$, # *corresponds to the counting measure over* $\mathcal{B}$ *with respect to* $\mathbf{P}$, μ^s *is, as usual, the Lebesgue measure, and* $\mathbf{B}$ *refers to a non empty subset of* $\mathcal{B}$.

Obviously, the discrepancy gives the maximum difference between the fraction of points from a point set which lie inside subvolumes of the s-dimensional unit cube and the volume of these subvolumes. Let us demonstrate the concept of discrepancy by means of a simple example.

EXAMPLE 7.1 *Let us consider the regular grid generated over the interval* $\mathbf{I}^2$ *from Figure 7.1. Obvioulsy, the interval* $\mathbf{B} = [0, \frac{3}{16}) \times [0, \frac{3}{16})$ *has Lebesgue measure* $\frac{9}{256}$. *As the number of points within* $\mathbf{B}$ *is* $\#(\mathbf{P}\cap\mathbf{B}) = 1$ *we conclude that* $\left|\frac{\#(\mathbf{P}\cap\mathbf{B})}{64} - \mu^2(\mathbf{B})\right| = \left|\frac{4}{256} - \frac{9}{256}\right| = \frac{5}{256} > \frac{1}{64}$. *Since* $\mathbf{B}$ *is only a single element of the family* $\mathcal{B}$ *of subsets of* $\mathbf{I}^2$, *we can conclude that* $\frac{1}{64}$ *is a lower bound of the discrepancy of this regular grid.*

Lower Bound (862)

In Remark 7.1 it is shown that the discrepancy of point sets located on regular s-dimensional grids is of order $O\left(\frac{1}{\sqrt[s]{N}}\right)$, *where* N *is the number of samples.*

The basic idea behind the concept of discrepancy is to minimize the effect of clumping in samples. This is done by considering various regions of the domain and comparing the volume of these regions to the number of samples inside them. Due to the fact, that our definition of discrepancy does not make any statement about the choice of these regions, we are free in our decision about the choice of $\mathcal{B}$. So, it should also be clear, that the choice of different families of Lebesgue measurable sets in the definition above leads to different concepts of discrepancy. From the multiplicity of possible discrepancy concepts, two are important for us: the *star discrepancy* and the *extreme discrepancy*.

Lebesgue Measurable Set (75)

DEFINITION 7.2 (Star Discrepancy and Extreme Discrepancy, $D_N^*(\mathbf{P})$ and $D_N(\mathbf{P})$) *Let us replace the Lebesgue measurable set family* $\mathcal{B}$ *given over* $\mathbf{I}^s$ *in the above definition by:*

$$\mathcal{I}^* \stackrel{def}{=} \left\{ \mathbf{B} \middle| \mathbf{B} = \prod_{i=1}^{s} [0, u_i) \subset \mathbf{I}^s \right\}, \tag{7.7}$$

that is, the axis-aligned s-dimensional subvolumes of $\mathbf{I}^s$ *attached to the origin. Then, we obtain the most important discrepancy concept in the theory of quasi-Monte Carlo*

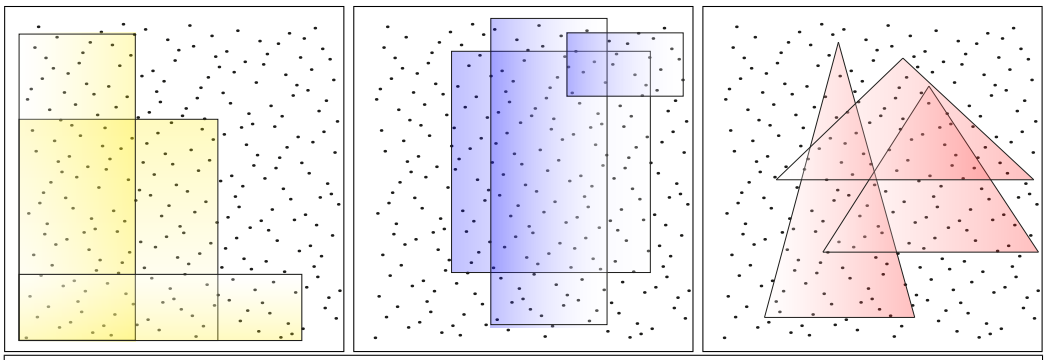


FIGURE 7.2: STAR DISCREPANCY AND EXTREME DISCREPANCY. Visualization of the discrepancy concepts—case s=2—introduced in Definition 7.2. The star discrepancy based on axis-aligned 2-dimensional subareas of $\mathbf{I}^2$ attached at the origin, and the extreme discrepancy based on the choice of arbitrary 2-dimensional subvolumes of $\mathbf{I}^2$. Further discrepancy concepts can be consider such as the *triangle discrepancy* in the right figure, where the set $\mathbf{B}$ are axis-aligned triangle within $\mathbf{I}^2$.

integration: the star disrepancy, $D_N^*(\mathbf{P})$, *defined as:*

$$D_N^*(\mathbf{P}) \stackrel{\text{def}}{=} D_N(\mathbf{P}, \mathcal{I}^*). \tag{7.8}$$

Another important concept of discrepancy, the extreme discrepancy, $D_N(\mathbf{P})$, *is given by:*

$$D_N(\mathbf{P}) \stackrel{\text{def}}{=} D_N(\mathbf{P}, \mathcal{I}), \tag{7.9}$$

where the set $\mathcal{B}$ is based on the choice of arbitrary s-dimensional subvolumes of $\mathbf{I}^s$, thus, $\mathcal{B}$ corresponds to $\mathcal{I}$ with

$$\mathcal{I} \stackrel{\text{def}}{=} \left\{ \mathbf{B} | \mathbf{B} = \prod_{i=1}^{s} [u_i, v_i) \ \subset \mathbf{I}^s \right\}. \tag{7.10}$$

Obviously, these discrepancies compute how much a N-element point set $\mathbf{P}$ deviates from its ideal distribution, that is, a quasi-Monte Carlo technique attempts to distribute samples in such a way that every box of size $\mu(B)$ contains $\mu(B) \cdot N$ points of $\mathbf{P}$.

EXAMPLE 7.2 *Let us consider a one-dimensional point set* $P = x_1, \ldots, x_N$ *with* $x_i = \frac{i}{N}, i = 1, \ldots, N$ *as illustrated in Figure 7.3, [220, Veach 1997]. Now, the following applies to all intervals of the type* $[0, \frac{i}{N})$*:* $\#\left(P \cap [0, \frac{i}{N})\right) = i-1$*, with* $\mu\left([0, \frac{i}{N})\right) = \frac{i}{N}$*. So, it follows:* μ (75)

$$D_N^*(P) = \frac{1}{N}. \tag{7.11}$$

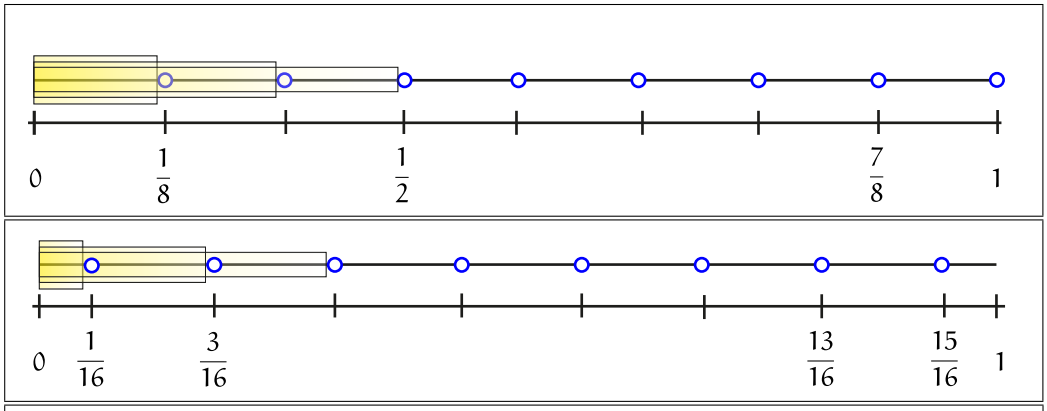


FIGURE 7.3: STAR DISCREPANCY OF A POINT SET. The point set $\mathbf{P} = \{x_1, \ldots, x_8\}$ with $x_i = \frac{i}{8}, i = 1, \ldots, 8$ and the associated modification $\mathbf{P}'$, with $x'_i = \frac{i-\frac{1}{2}}{8}$, resulting from a left-shift of $\mathbf{P}$ by $\frac{1}{16}$. As one can easily see, the pattern of the sets are equal, but $\mathbf{P}'$ has a lower discrepancy than the original set $\mathbf{P}$.

Discussing the following slight modification to the above point set where $x'_i = \frac{i-\frac{1}{2}}{N}$, *then it holds to all intervals of the type* $\left[0, \frac{i-\frac{1}{2}}{N}\right)$: $\#\left(P \cap \left[0, \frac{i-\frac{1}{2}}{N}\right)\right) = i - 1$, *with* $\mu\left(\left[0, \frac{i-\frac{1}{2}}{N}\right)\right) = \frac{i-\frac{1}{2}}{N}$, *that is,*

$$D^*_N(x'_1, \ldots, x'_N) = \sup_{[0,u)\in\mathcal{I}^*} \left| \frac{i-1}{N} - \frac{i-\frac{1}{2}}{N} \right| = \frac{1}{2N}. \tag{7.12}$$

Since the following holds for the star discrepancy of one-dimensional point sets

$$D^*_N(x'_1, \ldots, x'_N) = \frac{1}{2N} + \max_{1\le i\le N} \left| x'_i - \frac{2i-1}{2N} \right|, \tag{7.13}$$

the point set $x'_1, \ldots, x'_N$ *has the smallest possible discrepancy.*

The point set $\mathbf{P}$ *from above is one of the few point sets, where we can compute the discrepancy analytically. For the most higher dimensional point sets or sequences, we have no chance to compute the discrepancy in such a simple way. Here the discrepancy must be estimated rather numerically via the construction of a large number of boxes, computing their discrepancy, and reporting the maximum.*

REMARK 7.1 *The following bounds apply to the star discrepancy of randomly and independently selected samples involved in ordinary Monte Carlo procedures:*

$$D^*_N(\mathbf{P}) \in O\left(\sqrt{\frac{\log\log N}{N}}\right), \tag{7.14}$$

while the star discrepancy of point sequences located on regular s-dimensional grids implies:

$$D_N^*(\mathbf{P}) \in O\left(\frac{1}{\sqrt[s]{N}}\right). \tag{7.15}$$

REMARK 7.2 (Further Discrepancy Concepts) *Apart of the already mentioned Lebesgue measurable set families $\mathcal{I}^*$ and $\mathcal{I}$ in the definition above, it is also possible—depending on the characteristics of the underlying problem—to choose other types for the set $\mathcal{B}$. Thus, [104, Keller 1998] recommends—especially for the analysis of Monte Carlo methods involving pixel supersampling—to consider classes formed of sets which reflects the basic geometry occurring most frequently in pixels, such as the set of all triangles contained in $\mathbf{I}^2$ or, alternatively, the set of lines obtained by intersections of half planes with $\mathbf{I}^2$.*

Lebesgue Measurable Set (75)

The definitions of all hitherto discussed discrepancies were based on the supremum norm. As alternatives to this, other well known norms may also be applied. A discrepancy concept based on the $\mathcal{L}^2$-norm, for example, would take the following form:

Supremum Norm (33)
$\|\cdot\|_{\mathcal{L}^2}$ (110)

$$D_2^*(P_N) \stackrel{def}{=} \left\| \frac{\#(\mathbf{P} \cap \mathbf{B})}{N} - \mu^s\big(\mathbf{B}(\mathbf{x})\big) \right\|_{\mathcal{L}^2} \tag{7.16}$$

whereby $\mathbf{B}(\mathbf{x}) \subset \mathbf{I}^s$ would correspond to the subcube $\prod_{j=1}^s [0, \mathbf{x}_j)$ with $\mu^s\big(\mathbf{B}(\mathbf{x})\big) = \prod_{j=1}^s \mathbf{x}_j$.

μ^s (82)

We can conclude from the definition of the discrepancy that we are interested in generating sequences of samples with low-discrepancy, that is, sequences $\mathbf{P}$ where $D_N(\mathbf{P})$ goes to zero as N goes to to infinity.

Now, the goal of any quasi-Monte Carlo method is to minimize the irregularity of distribution of the samples with respect to some measure. This means: If we wish to construct sequences for which the quasi-Monte Carlo estimated value converges towards the integral, the aspect of discrepancy must be taken into account in their construction. Since the star discrepancy and the extreme discrepancy are equivalent discrepancy concepts, that is,

$$D_N^*(\mathbf{P}) \leq D_N(\mathbf{P}) \leq 2^s\, D_N^*(\mathbf{P}), \quad \mathbf{P} \subset \mathbf{I}^s \tag{7.17}$$

it suffices in our discussions to work with any of these norms, so, we will usually use the star discrepancy D_N^*.

UNIFORMLY DISTRIBUTED SEQUENCE. To generate sequences of low-discrepancy, we need the notion of the *uniformly distributed sequence*, which is based on the concept of discrepancy.

DEFINITION 7.3 (Uniformly Distributed Sequence) *Let* $(\mathbf{x}_i)_{i\in\mathbb{N}}$ *be a sequence of numbers with values inside the unit cube* $\mathbf{I}^s$. $(\mathbf{x}_i)_{i\in\mathbb{N}}$ *is referred to as* uniformly distributed *if the following holds:*

$$\lim_{N\to\infty} D_N^*(\mathbf{x}_i)_{i\in\mathbb{N}} = 0. \tag{7.18}$$

REMARK 7.3 *Due to the equivalence property (7.17), it should also be clear, that the star discrepancy as well as the extreme discrepancy can be viewed as a qualification of the definition of uniformly distributed sequences in* $\mathbf{I}^n$.

As already mentioned at the beginning of this section, the error of a Monte Carlo integration scheme is also depending on how quickly a function changes its values between sample points. So, it is obviously that a function, which changes significantly in small integration regions, involves a quite large error. If the function is smooth between sample points, then the error will be small. A mathematical concept that provides us information about changes of a function is the *variation* of a function. In the case of one-dimensional functions, the *variation in sense of Vitali*—defined as the sum of the differences of function values at neighboring samples—does a good job. In our case—we are interested mainly in functions of high dimensions—the variation of Vitali is not a good choice. For measuring the changes of functions of more than a single variable we need the concept of the *variation in sense of Hardy and Krause.*

DEFINITION 7.4 (Variation in the Sense of Hardy and Krause) *Let* f *be a real-valued, s-dimensional function. The* variation in the sense of Hardy and Krause *is then given by:*

$$V_{HK} \stackrel{def}{=} \sum_{k=1}^{s} \sum_{1\le i_1<\ldots<i_k\le s} V^{(k)}(f; i_1, \ldots, i_k), \tag{7.19}$$

with $(f; i_1, \ldots, i_k) = f\big|_{(u_1,\ldots,u_s)}$ *with* $u_j = 1$ *for* $j \neq i_1, \ldots, i_k$, *where*

$$V^{(s)}(f) \stackrel{def}{=} \sup_{\mathcal{P}} \sum_{A\in\mathcal{P}} |\Delta(f, A)| \tag{7.20}$$

$$= \sup_{\mathcal{P}} \sum_{i=1}^{n} |f(\mathbf{x}_i) - f(\mathbf{x}_{i-1})| \tag{7.21}$$

is the variation in the sense of Vitali *for* $\mathcal{P} = \{\mathbf{x}_0 < \mathbf{x}_0 < \cdots \mathbf{x}_{n-1} < \mathbf{x}_n\}$. *The supremum is taken over all partitions* $\mathcal{P}$ *of* $\mathbf{I}^s$ *and*

$$\Delta(f, A) \stackrel{def}{=} \sum_{i_1=0}^{1} \cdots \sum_{i_s=0}^{1} (-1)^{\sum_{k=1}^{s} i_k} f(u_1^{(i_1)}, \ldots, u_s^{(i_s)}) \tag{7.22}$$

for $A = \prod_{k=1}^{s} [u_k, v_k)$, *that is,* $\Delta(f, A)$ *denotes an alternating sum of the values of* f *at vertices of* A.

We say, f *is of* bounded variation in the sense of Hardy and Krause, *if* $V_{HK}(f) < \infty$, *[57, Evans & Swartz 2000].*

REMARK 7.4 (A Restricted Definition of the Variation in the Sense of Hardy and Krause) *Often, the definition of the variation in the sense of Hardy and Krause is too weaken, and can be restricted to apply to a much smaller, but simpler to handle class of functions. As this version of the variation in the sense of Hardy and Krause is not so complicated as those given in Definition 7.4, we formulate them for the sake of completness:*

Let f *be a real-valued,* s*-dimensional function, where it holds:* f *is a differentiable function from* $C^s(\mathbf{I}^s)$. *The variation in the sense of Hardy and Krause, used in Definition 7.4, can then be replaced by:*

$C^s(\cdot)$ (28)

$$V_{HK}^{(s)}(f) = \int_{\mathbf{I}^s} \left| \frac{\partial^s f}{\partial u_1 \dots \partial u_s} \right| du_1 \dots du_s. \tag{7.23}$$

EXAMPLE 7.3 *For a one-dimensional differentiable function, the variation in the sense of Hardy and Krause is simply defined as:*

$$V_{HK}(f) \stackrel{def}{=} \int_{[0,1)} |f'(x)| \, d\mu(x). \tag{7.24}$$

For differentiable functions of more than a single variable, induction is used to determine the variation, that is, for a 2-dimensional function $f(x,y)$ *it holds:*

$$V_{HK}(f) = \int_{\mathbf{I}^2} \left| \frac{\partial^2 f(x,y)}{\partial x \partial y} \right| d\mu^2(x,y) + \tag{7.25}$$

$$\int_{[0,1)} \left| \frac{\partial f(x,1)}{\partial x} \right| d\mu(x) + \int_{[0,1)} \left| \frac{\partial f(1,y)}{\partial y} \right| d\mu(y). \tag{7.26}$$

REMARK 7.5 *Due to [131, Jiang & McNamara 2002], a piecewise differentiable function is of bounded variation in the sense of Hardy and Krause, if the discontinuity is only located on finitely many hyperplanes parallel to the sides of the hypercube and each differentiable piece is of bounded variation in the sense of Hardy-Krause. But discontinuous functions of dimension* ≥ 2 *are usually not of bounded variation in the sense of Hardy-Krause.*

THE KOKSMA-HLAWKA INEQUALITY. Based on the concept of the variation in the sense of Hardy and Krause, it is now possible to bound the error of integration by the *Koksma-Hlawka Inequality.*

THEOREM 7.1 (The Koksma-Hlawka Inequality) *Let* $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ *be a point set of numbers from* $\mathbf{I}^s$, f *be a Lebesgue measurable function on* $\mathbf{I}^s$, *and* $D_N^*(\mathbf{x}_1, \dots, \mathbf{x}_N)$ *the star*

Measurable Function (98)

discrepancy of the given point set, then it holds:

$$\left|\frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n) - \int_{\mathbf{I}^s} f(\mathbf{x})\,d\mu^s(\mathbf{x})\right| \leq V_{HK}(f)\, D_N^*(\mathbf{x}_1,\ldots,\mathbf{x}_n), \tag{7.27}$$

where $V_{HK}(f)$ is the variation in the sense of Hardy and Krause.

PROOF 7.1 *For a proof of the Koksma-Hlawka Inequality see [138, Niederreiter 1992].*

Section (6.4)

Compared to the error boundaries from theory of Monte Carlo integration, which were all probabilistic, the Koksma-Hlawka inequality provides us with a deterministic bound. It describes the convergence of any quasi-Monte Carlo procedure depending on the variation of the integrand, but also influenced by the choice of the point sequences involved. As long as the variation of f can be kept within boundaries, i.e. if it is smaller than a constant, it is the construction of suitable point sequences which provides for a correspondingly rapid convergence of the method. In one dimension piecewise continuous function satisfy this condition, that is $V_{HK}(f) < \infty$. So, the Koksma-Hlawka inequality indicates the path to be taken for the construction of point sequences whose quasi-Monte Carlo estimated value will be very good.

REMARK 7.6 *Unfortunately, the Koksma-Hlawka inequality is of little support in estimating low-discrepancy point sets used to integrate s-dimensional discontinuous functions ($s \geq 2$), since the variation of such a function in the sense of Hardy and Krause is infinite. Now, functions of this type appear frequently in problems of computer graphics, e.g. we imagine a diagonal line from $(0,0)$ to $(1,1)$ with $f = 1$ above and $f = 0$ below that line. Although the variance is quart, the variation of f in the sense of Hardy and Krause is infinite, [104, Keller 1998].*

Variance (201)

Note: Even if a function has infinite variation, we can use quasi-Monte Carlo integration. If the samples are based on a uniformly distributed sequence, we can assume, that the approximated quasi-Monte Carlo value converges to the exact value of the integral, although the Koksma-Hlawka can not be applied [210, Szirmay-Kalos 1999].

Due to the equivalence property of the star discrepancy and the extreme discrepancy it will suffice to focus our interest with respect to generating uniformly distributed sequences on sequences with a low-discrepancy, so-called *low-discrepancy sequences*.

7.2 LOW-DISCREPANCY POINT SETS AND LOW-DISCREPANCY SEQUENCES

The deterministic mode of computers usually does not permit the writing of algorithms which generate real random numbers, but only allows for the generation of approximations

thereof, referred to as *pseudo random numbers*. Discrepancy statements relating to pseudo random numbers were already introduced in the last section, see Remark 7.1, so that the question now arises: Can we do it better? Exist point sequences with even lower discrepancy?

Yes, there are such sequences, so-called *low-discrepancy sequences*.

DEFINITION 7.5 (Low-discrepancy Point Set) *Let $\mathbf{P} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a point set of numbers from $\mathbf{I}^s$. $\mathbf{P}$ is referred to as a* low-discrepancy point set *if it holds:*

$$D_N^*(\mathbf{P}) = O\left(\frac{\log^{s-1} N}{N}\right). \tag{7.28}$$

Now, in practice it is often convenient to be able to change the value N of a low-discrepancy point set without loosing the previously computed sample values. For this reason, it should be clever to work with sequences of points and then to take the first N numbers of such a sequence whenever the value of N has been selected. So, N can be increased and the data from earlier computations can furthermore used. This idea implies the construction of so-called *low-discrepancy sequences*.

DEFINITION 7.6 (Low-discrepancy Sequence) *Let $(\mathbf{x}_i)_{i\in\mathbb{N}}$ be a s-dimensional sequence of numbers from $\mathbf{I}^s$. $(\mathbf{x}_i)_{i\in\mathbb{N}}$ is referred to as a* low-discrepancy sequence, *if for each prefix $\mathbf{P} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ it holds:*

$$D_N^*(\mathbf{P}) = O\left(\frac{\log^{s} N}{N}\right). \tag{7.29}$$

REMARK 7.7 *In [138, Niederreiter 1992] it is shown that the above boundaries are the best which may be expected for s-dimensional point sets and sequences.*

DEFINITION 7.7 (Quasi-Monte Carlo Integration) *The construction of a low-discrepancy point set, $\mathbf{P} \subset \mathbf{I}^s$, and the subsequent evaluation of the integral*

$$\int_{\mathbf{Q}^s} f(\mathbf{x})\, d\mu^s(\mathbf{x}), \tag{7.30}$$

from Equation (7.30) as the arithmetic mean of the function values at points of $\mathbf{P}$ is referred to as quasi-Monte Carlo integration.

REMARK 7.8 *The elements of a low-discrepancy sequence $\mathbf{P}$ are also referred to as quasi-random numbers, because they have some statistical qualities that make them acceptable substitutes for real random numbers. Although they are not completely uniform distributed, however they approximate the property of uniform distribution of a given point set in an optimal way, and are additionally, due to their deterministic origin, devoid of probabilistic features.*

UD Sequence (626)

Let us now insert the error boundary from Equation (7.29) into the Koksma-Hlawka Inequality, then we obtain:

Koksma-Hlawka inequality (627)

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n) - \int_{\mathbf{I}^s} f(\mathbf{x})\,d\mu^s(\mathbf{x})\right| &\leq V_{HK}(f)\,D_N^*(\mathbf{x}_1,\ldots,\mathbf{x}_N)\\ &= O\left(\frac{\log^s N}{N}\right).\end{aligned}$$

This means, that the error of any quasi-Monte Carlo algorithm is at most $O\left(\frac{\log^{s-1} N}{N}\right)$ using a low-discrepancy point set, or $O\left(\frac{\log^s N}{N}\right)$ using a prefix of a low-discrepancy sequence, provided that the variation of the function f in sense of Hardy and Krause may be limited by a constant. As already mentioned in the previous section this holds for continuous functions in one-dimension. In cases, in which the functions to be integrated are high-dimensional and discontinuous, as encountered in integral equations of global illumination, practicable convergence statements cannot be obtained with the Koksma-Hlawka inequality. In such cases, considerably weaker error boundaries must be applied.

Hardy and Krause (626)

In the above discussion, the discrepancy boundaries of low-discrepancy point sequences and point sets were outlined. We now turn to the question how low-discrepancy sequences and point sets may be generated.

There are many different low-discrepancy sequences used in quasi-Monte Carlo techniques. We present a few of these sequences in this and the following subsections, where we will start with low-discrepancy sequences based on the concept of the *radical-inverse function*.

DEFINITION 7.8 (Radical-inverse Function) *Assuming* $b \geq 2, b \in \mathbb{N}$, *then the* radical-inverse function *is defined as:*

$$\Phi_b : \mathbb{N}_0 \times \mathfrak{S}^b \longrightarrow [0,1]$$

with

$$\Phi_b(i,\pi) \stackrel{def}{=} \sum_{j=0}^{\infty} \pi\big(a_j(i)\big) b^{-j-1} \qquad \textit{with} \qquad i = \sum_{j=0}^{\infty} a_j(i) b^j, \tag{7.31}$$

whereby $(a_j)_{j\in\mathbb{N}_0}$ *corresponds to the representation of the number* $i \in \mathbb{N}_0$ *with respect to the basis* b *and* π *is a permutation from the symmetric group* $\mathfrak{S}^b$ *on* $\{0,\ldots,b-1\} \subset \mathbb{N}_0$.

Based on the radical-inverse function, now a very large number of well-known low-discrepancy sequences can be generated. As a first example, let us consider the *van der Corput sequence*.

EXAMPLE 7.4 (Van der Corput Sequences) *The sequence* $(\Phi_b(i,\pi))_{i\in\mathbb{N}_0}$ *is termed the* general van der Corput sequence. *It is a one-dimensional low-discrepancy sequence of order* $O(\frac{\log N}{N})$. *In particular, if the permutation* π *is the identity, one obtains the* van der Corput sequence. *For*

$$x_i = \Phi_2(i) \equiv \Phi_2(i, \mathrm{id}), \tag{7.32}$$

it corresponds to the binary representation of i *reflected at the decimal point, that is,*

$$0.0_2, 0.1_2, 0.01_2, 0.11_2, 0.001_2, 0.101_2, 0.011_2, \ldots \tag{7.33}$$

$$0, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{3}{4}, \tfrac{1}{8}, \tfrac{5}{8}, \tfrac{3}{8}, \ldots. \tag{7.34}$$

To generate a low-discrepancy sequence in several dimensions, we must use different radical inverse sequences with different bases in each dimension. Thus, a s-dimensional low-discrepancy sequence $(x_i^s)_{i\in\mathbb{N}_0}$ with relatively prime bases b_i has the form

$$x_i^s \stackrel{def}{=} \big(\Phi_{b_1}(i,\pi), \ldots, \Phi_{b_s}(i,\pi)\big). \tag{7.35}$$

REMARK 7.9 *Note: Our Monte Carlo estimators are defined as sums of the type*

$$F_N = \sum_{i=1}^{N} f(X_i) \tag{7.36}$$

where $\{X_1, \ldots, X_N\}$ *is a* N*-element set of random variables.*

Due to the definition of the radical inverse function, we have and will construct, in this and the following sections, our low-discrepancy sequences as functions defined on the domain $\mathbb{N}_0$ *instead of* $\mathbb{N}$. *This is not mandatory but, as the construction of low-discrepancy sequences is strongly coupled to the index set of the radical inverse function, this will lead to the known formulas for low-discrepancy sequences from the literature.*

7.2.1 THE CLASSICAL CONSTRUCTS: HALTON SEQUENCE AND HAMMERSLEY POINT SET

In addition to the above introduced van der Corput sequences further low-discrepancy sequences may be generated on the basis of the radical-inverse function. In the following, we will focus in more detail on the construction of the most well-known low-discrepancy sequences, also of greatest importance to the goals of our discussion: the *Halton sequence*, the *Hammersley point set*, and the *Zaremba sequence*.

van der Corput Sequence (631)
Radical-inverse Function (630)

HALTON SEQUENCE. The Halton sequence is one of the most easily computable low-discrepancy sequences.

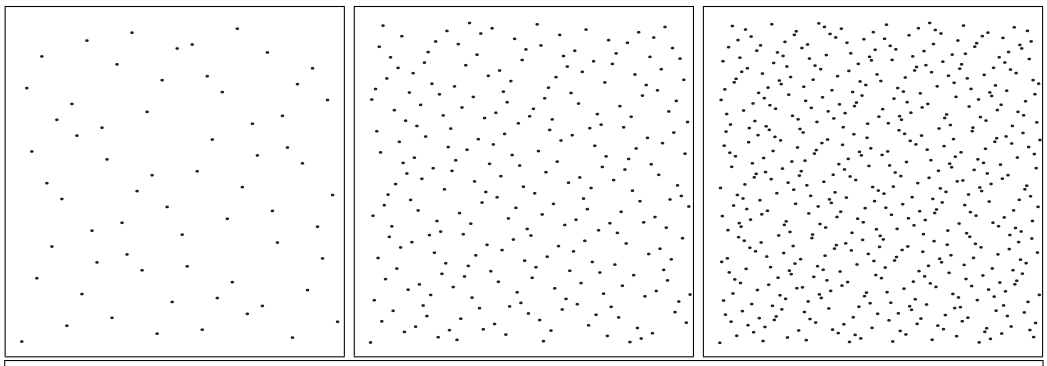


FIGURE 7.4: HALTON SEQUENCE. The first 64, 256, and 512 points of the 2-dimensional Halton sequence $\mathbf{P}^2_{\text{HAL}} = (\Phi_2(i), \Phi_3(i))_{i \in \mathbb{N}_0}$.

DEFINITION 7.9 (Halton Sequence) *Let the bases of the radical-inverse function be given by the s prime numbers* $p_1, p_2, \ldots, p_s$. *Let furthermore* π *be the identical permutation, then with*

$$\Phi_b(i) \stackrel{\text{def}}{=} \sum_{j=0}^{\infty} a_j(i) b^{-j-1} \tag{7.37}$$

Radical Inverse Function (630) *and*

$$i = \sum_{j=0}^{\infty} a_j(i) b^j, \tag{7.38}$$

the s-dimensional Halton Sequence, $\mathbf{P}^s_{\text{HAL}} = (\mathbf{x}^s_i)_{i \in \mathbb{N}_0}$, *is given by:*

$$\mathbf{x}^s_i \stackrel{\text{def}}{=} (\Phi_{p_1}(i), \Phi_{p_2}(i), \ldots, \Phi_{p_s}(i)). \tag{7.39}$$

REMARK 7.10 (Construction of Halton Sequences) *The s-dimensional Halton sequence can be created from s one-dimensional Halton sequences using different bases for each dimension. For the construction of a one-dimensional Halton sequence see Figure 7.5. Table 7.1 shows the first elements of a 2-dimensional Halton sequence in the bases* $p_1 = 2$ *and* $p_2 = 3$, *computed via the procedure in Figure 7.5.*

Obviously, the single components of this s-dimensional sequence can be constructed by repeatedly dividing the unit interval $[0, 1]$ by the bases, that is—if the bases corresponds to the first prime numbers—the first component in halves, fourths, eights, the second in thirds, ninths, twenty-seventh, the third in fifths, twenty-fifths and so on. With other words, a Halton sequence in more dimensions is created from the one-dimensional Halton sequences of the bases $p_1, p_2, \ldots, p_s$.

One-dimensional Halton Sequence in Base p {

$\forall\ i \in \{0, \dots, N-1\}$ {

compute i in base p thus $i_p = \alpha_{n-1} \dots \alpha_0$ via

$i = \sum_{i=0}^{n-1} \alpha_i p^i$, with $\alpha_i \in \{0, \dots, p-1\}$

compute $\bar{i}_p = 0.\alpha_0 \cdots \alpha_{n-1}$ by reversing i_p around the decimal point

compute $\Phi_p(i) = \sum_{i=0}^{n-1} \alpha_i p^{-(i+1)}$

}

}

FIGURE 7.5: ONE-DIMENSIONAL HALTON SEQUENCE IN BASE p. The i^{th} element of a one-dimensional Halton sequence can be computed in a three step process. First, every number from $\{0, \dots, N-1\}$ is represented as a number in the base p. In the 2^{nd} step the representation in base p is reversed and attached after the decimal point. In the last step this member is then converted to base 10.

i	$p_1 = 2$	$p_2 = 3$
0	$0_2 \to .0_2 \to 0$	$0_3 \to .0_3 \to 0$
1	$1_2 \to .1_2 \to 1 \cdot 2^{-1} = \frac{1}{2}$	$1_3 \to .1_3 \to 1 \cdot 3^{-1} = \frac{1}{3}$
2	$10_2 \to .01_2 \to 1 \cdot 2^{-2} = \frac{1}{4}$	$2_3 \to .2_3 \to 2 \cdot 3^{-1} = \frac{2}{3}$
3	$11_2 \to .11_2 \to 1 \cdot 2^{-1} + 1 \cdot 2^{-2} = \frac{3}{4}$	$10_3 \to .01_3 \to 1 \cdot 3^{-2} = \frac{1}{9}$
4	$100_2 \to .001_2 \to 1 \cdot 2^{-3} = \frac{1}{8}$	$11_3 \to .11_3 \to 1 \cdot 3^{-1} + 1 \cdot 3^{-2} = \frac{4}{9}$
5	$101_2 \to .101_2 \to 1 \cdot 2^{-1} + 1 \cdot 2^{-3} = \frac{5}{8}$	$12_3 \to .21_3 \to 2 \cdot 3^{-1} + 1 \cdot 3^{-2} = \frac{7}{9}$

TABLE 7.1: COMPUTATION OF $\mathbf{P}^2_{HAL} = \left(\Phi_2(i-1), \Phi_3(i-1)\right)_{i \in \mathbb{N}}$. The first six elements $0 \le i \le 5$ of the 2-dimensional Halton sequence with bases $p_1 = 2$ and $p_2 = 3$.

REMARK 7.11 *The following upper bound applies to the star-discrepancy of the s-dimensional Halton sequence $\mathbf{P}^s_{HAL}$:*

$$D^*_N(P^s_{HAL}) < \frac{s}{N} + \frac{1}{N} \prod_{j=1}^{s} \left(\frac{b_j - 1}{2 \log b_j} \log N + \frac{b_j + 1}{2} \right) \in O\left(\frac{\log^s N}{N} \right). \tag{7.40}$$

The Halton sequence constructed in this manner is an incremental pattern which implies that it may be expanded if necessary without discarding the samples already drawn. New samples $\mathbf{x}^s_i$ may be added at incremental cost.

HAMMERSLEY POINT SET. On the other hand, if the number of samples which need to be calculated for the solution of the problem at hand is known in advance, the discrepancy of a predetermined sequence is easily improved by focusing exclusively on a finite number of

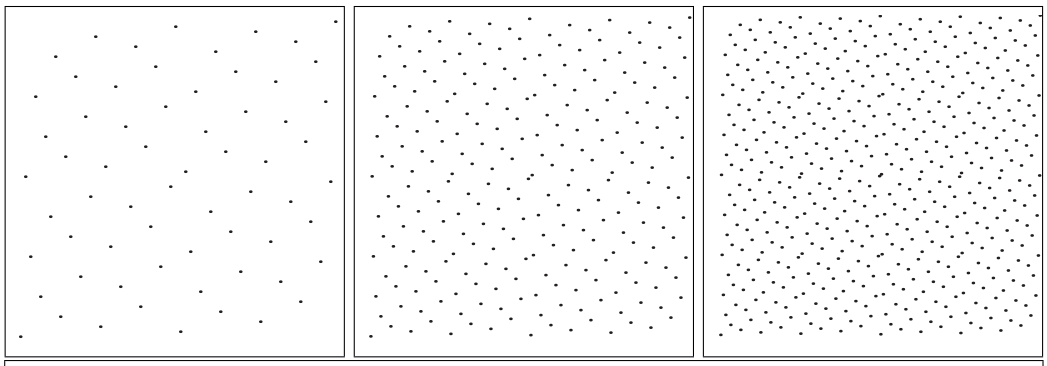


FIGURE 7.6: HAMMERSLEY POINT SET. Three 2-dimensional Hammersley point sets $\mathbf{P}^2_{\text{HAM}} = \left(\frac{i}{N}, \Phi_2(i)\right)_{i\in(0,\ldots,N-1)}$ of sizes $N = 64, 256$, and $N = 512$.

sequence members. If this method is applied to the above constructed Halton sequence, one obtains the s-dimensional *Hammersley point set*.

DEFINITION 7.10 (Hammersley Point Set) *The* s-dimensional Hammersley point set, $\mathbf{P}^s_{\text{HAM}} = (\mathbf{x}^s_i)_{i\in(0,\ldots,N-1)}$, *is defined by:*

$$\mathbf{x}^s_i \stackrel{\text{def}}{=} \left(\frac{i}{N}, \Phi_{p_1}(i), \Phi_{p_2}(i), \Phi_{p_3}(i), \ldots, \Phi_{p_{s-1}}(i)\right), \tag{7.41}$$

Radical Inverse Function (630) *where the bases of the radical-inverse function are given by the* $s-1$ *prime numbers* $p_1, p_2, \ldots, p_{s-1}$.

EXAMPLE 7.5 (The Hammersley Point Set on the Euclidean Plane) *On the Euclidean plane, a* N*-element Hammersley point set can be generated by choosing* $p_1 = 2$, *that is,*

$$\mathbf{x}^2_i = \left(\frac{i}{N}, \Phi_2(i)\right) \tag{7.42}$$

$$= (0,0), \left(\frac{1}{N}, \frac{1}{2}\right), \left(\frac{2}{N}, \frac{1}{4}\right), \left(\frac{3}{N}, \frac{3}{4}\right), \ldots \tag{7.43}$$

with $i = 0, \ldots, N-1$, *see Figure 7.6.*

As shown in Example 7.4, the Hammersley point set $(\mathbf{x}^2_i)_{i\in(0,\ldots,N-1)}$ *can be computed by taking all numbers in the range from* 0 *to* $N-1$ *and interpreting them as binary fractions. With reference to Example 7.4 this means: for* $N = 64, i = 6$, *we first compute* $\text{bin}(6) = 000110_2$. *Reversing the binary digits results in* $0.011000_2 = \frac{3}{8}$, *thus,* $(x_6, y_6) = \left(\frac{6}{64}, \frac{3}{8}\right) = \left(\frac{3}{32}, \frac{3}{8}\right)$.

EXAMPLE 7.6 (The Hammersley Point Set on the Unit Sphere) *To generate directions over the unit sphere, due to [198, Spanier & Gelbard 1969], we can map in a first step the Hammersley point set* $\mathbf{P}^2_{\mathrm{HAM}} = (\mathbf{x}^2_i)_{i\in(0,\ldots,N-1)}$ *linearly to the cylindrical domain* $[0,2\pi)\times[-1,1]$*, thus,*

$$\mathbf{x}^2_i = \left(\frac{i}{N}, \Phi_p(i)\right) \mapsto (\phi, t) = \left(\frac{i+0.5}{N}\,2\pi, 2\Phi_p(i)-1\right). \tag{7.44}$$

Then, we use a z-preserving radial projection from the unit cylinder

$$C = \{(x,y,z)|x^2+y^2=1, |z|\le 1\} \tag{7.45}$$

to the unit sphere

$$(\phi, t) \mapsto \left(\sqrt{1-t^2}\cos\phi, \sqrt{1-t^2}\sin\phi, t\right). \tag{7.46}$$

The result of this procedure is visualized in Figure 7.7.

REMARK 7.12 *The following upper bound applies to the star-discrepancy of the s-dimensional Hammersley point set* $\mathbf{P}^s_{\mathrm{HAM}}$

$$D^*_N(P^s_{\mathrm{HAM}}) < \frac{s}{N} + \frac{1}{N}\prod_{j=1}^{s-1}\left(\frac{b_j-1}{2\log b_j}\log N + \frac{b_j+1}{2}\right) \in O\left(\frac{(\log^{s-1} N}{N}\right). \tag{7.47}$$

REMARK 7.13 (Adaptive Sampling a Pixel) *A similar sampling strategy to supersampling a pixel is* adaptive sampling. *In adaptive sampling, a ray is traced through each corners of a pixel. If the intensity of the four corners varies significantly from the others, then the pixel is split into four rectangular subdivision. This process of subdivision is repeated to an arbitrary level until the intensity of the four corners of a subdivision are not varies significantly.*

Obviously, the number of samples needed for adaptive sampling is not known in advanced. So, the restriction of the Hammersley point set to a pregiven fixed number N *of samples makes it not really usable for adaptive sampling, since all samples that are generated in a step of the process must be discarded if a further subdivision is required.*

ZAREMBA SEQUENCE. To conclude, we will take a look at a further low-discrepancy sequence based on the radical inverse function: the *Zaremba sequence*.

DEFINITION 7.11 (Zaremba-sequence) *Replacing* a_j *in the radical-inverse function by* $(a_j+j) \mod b$*, that is, defining*

$$\Psi_b(i,\pi) \stackrel{\mathrm{def}}{=} \sum_{j=0}^{\infty} \pi\big((a_j+j) \mod b(i)\big) b^{-j-1}, \tag{7.48}$$

Radical Inverse Function (630)

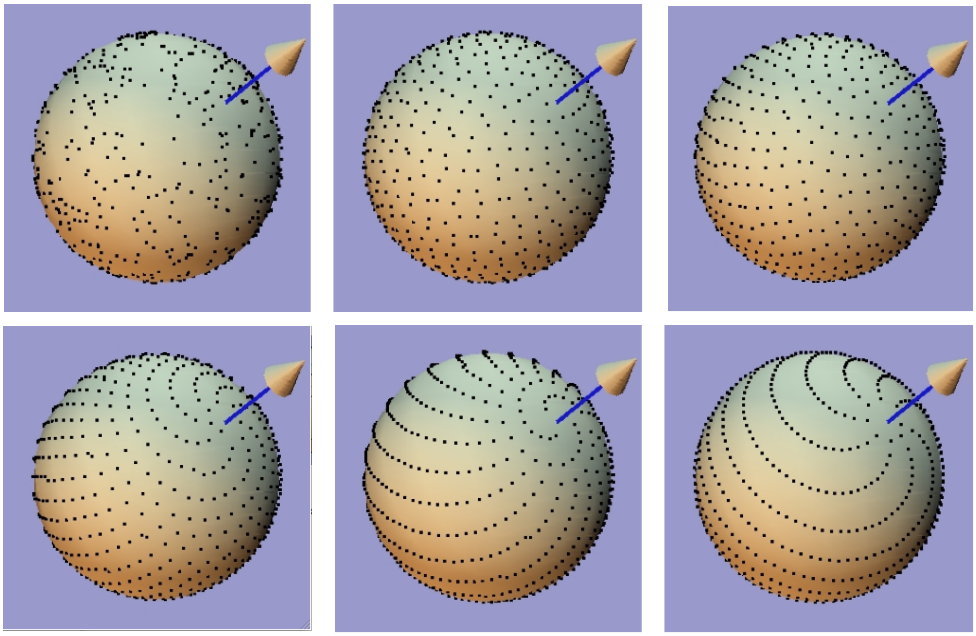


FIGURE 7.7: HAMMERSLEY POINT SET ON THE UNIT SPHERE. Comparison of a random pattern on the unit sphere with five Hammersley point set generated with different bases $p_1 = 2, p_1 = 3, p_1 = 5, p_1 = 7$ and $p_1 = 11$, $N = 1000$. The Hammersley point set with $p_1 = 2$ gives a pleasant, less clumped pattern. The points are uniformly distributed without a perceptible pattern, it gives the best uniformly distribution on the sphere. As p_1 increases (from upper right to lower right), points start to line up and form regular lines on the sphere. The position of the pole, marked with an arrow, becomes distinguishable from the pattern. Image courtesey of Tien-Tsin Wong, Pheng-Ann Heng, The Chinese University of Hong Kong and Wai-Shing Luk, Katholieke Universiteit Leuven.

then we call the sequence $\mathbf{P}^s_{ZAR} = (\mathbf{x}^s_i)_{i \in \mathbb{N}_0}$ *defined by:*

$$\mathbf{x}^s_i \stackrel{def}{=} \left(\Psi_2(i), \Psi_3(i), \Psi_5(i), \ldots, \Psi_{p_s}(i)\right) \tag{7.49}$$

the s-dimensional Zaremba sequence. *A visualization of the Zaremba sequence is shown in Figure 7.8.*

REMARK 7.14 *The Zaremba sequence may be regarded as a bridge to the scrambled versions of the above discussed low-discrepancy sequences in so far as it represents a first attempt to avoid the correlations which occur in the elements generated in a low-discrepancy sequence.*

Section 7.2.2

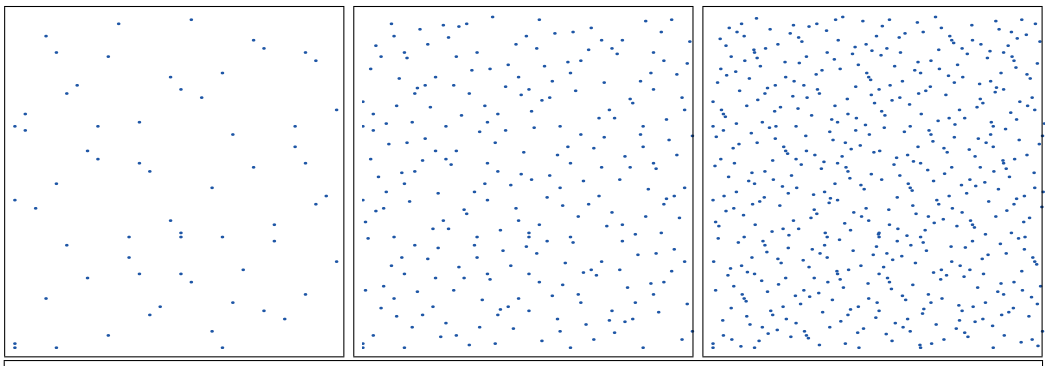


FIGURE 7.8: ZAREMBA SEQUENCE. The first 64, 256, and 512 points of the 2-dimensional Zaremba sequence $\left(\Psi_2(i), \Psi_3(i)\right)_{i \in \mathbb{N}_0}$.

Let us now consider a few application areas of the above introduced low-discrepancy sequence with respect to the stratification and sampling of integration domains.

EXAMPLE 7.7 (Stratification of $\mathbf{I}^s$) *Let us recall Example 6.34, where we have introduced the construct of the Voronoi diagram as a method for stratifying the unit square. Obviously, the set $\mathbf{P}_N$, which we observed there, corresponds to the above constructed N-element 2-dimensional Hammersley point set. The question that takes now is: Comes with any s-dimensional Hammersley point set also a Voronoi diagram?*

Voronoi Diagram (577)

Let us contemplate the general case of an s-dimensional low-discrepancy sequence. According to [104, Keller 1998] the following applies to the distance of two points $\mathbf{p}_i, \mathbf{p}_j, i \neq j$, calculated with the radical-inverse function: $\inf_{i \neq j}(\phi_b(i,\pi) - \phi_b(j,\pi)) = \frac{1}{N}$. If we now determine to a given N all $n_j \in \mathbb{N}$, with $b_j^{n_j - 1} < N \leq b_j^{n_j}$, then the radical-inverse function places the members of such an s-dimensional low-discrepancy sequence in the lower left corners of a regular $\frac{1}{b_j^{n_j}}$-grid. In the case of $N = b_j^{n_j}, 1 \leq j \leq s$ it also fills out the j-th dimension of the grid. That is, the grid structure of the Hammersley point set guarantees a minimal distance of the samples, which implies a s-dimensional Voronoi diagram as a stratification of $\mathbf{I}^s$.

LD Sequence (629)

Radical Inverse Function (630)

EXAMPLE 7.8 (Jittered Low-discrepancy Point Sets) *From Figure 7.4 and Figure 7.6, it can be seen that 2-dimensional low-discrepancy samples are always aligned in the lower left corner of their intervals. Applied to neighboring pixels this may result in aliasing effects if the sample rate is too low to satisfies Shannon's sampling theorem—for a detailed discussion see [68, Glassner 1995]. Therefore it is recommended to combine the deterministic construction of low-discrepancy samples with a jittered step. In this operation especially for the avoidance of potential aliasing effects, the involved points are displaced inside their intervals by an amount resulting from draw-*

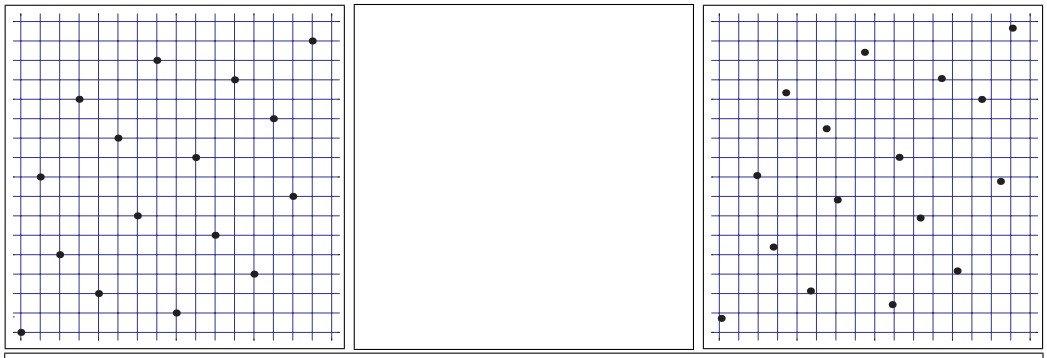


FIGURE 7.9: HAMMERSLEY AND JITTERED HAMMERSLEY POINT SETS. The first 16 elements of the 2-dimensional Hammersley point set $\mathbf{P}_{\text{HAM}}^{2} = \left(\frac{i}{N}, \Phi_2(i)\right)$, $0 \leq i \leq 15$ and the jittered Hammersley point set, which simulates 2^4-rooks sampling.

ing a random variable.

The s-dimensional jittered Hammersley point set $\mathbf{P}_{\text{HAM}}^{s,\text{jit}} = (\mathbf{x}_i^{s,\text{jit}})_{i\in(0,\ldots,N-1)}$ *is thus defined as:*

$$\mathbf{x}_i^{s,\text{jit}} \stackrel{\text{def}}{=} \left(\frac{i+u_1}{N}, \Phi_{b_1}(i) + \frac{u_2}{b_i^{n_1}}, \ldots, \Phi_{b_{s-1}}(i) + \frac{u_s}{b_{s-1}^{n_{s-1}}}\right), \tag{7.50}$$

Random Variable (168) Uniform Distribution (180) 2^n-rooks sampling (579)

whereby $u_i, 1 \leq i \leq s$ *represent random numbers uniformly distributed within* $\mathbf{I}^s$. *If in the 2-dimensional case we choose* $b = 2$ *and* $N = 2^n$, *then* $\mathbf{P}_{\text{HAM}}^{2,\text{jit}}$ *corresponds to* 2^n*-rooks sampling with deterministically determined permutations* π_1, π_2.

As briefly noted above, the jittering of low-discrepancy points forestalls the occurrence of aliasing effects, representing them following [40, Cook 1984], [67, Glassner 1995], onto the effect of high frequency noise, which is markedly more pleasant for an observer.

7.2.2 SCRAMBLING

As may be clearly seen in Figure 7.10, the choice of $\pi = \text{id}$ in the radical-inverse function of the above generated 2-dimensional Halton sequences leads to the generation of points arranged on lines. With respect to the algorithms used in computer graphics these in turn can lead to aliasing effects. In order to avoid this, the corresponding permutations in the radical-inverse function are chosen according to a procedure introduced by Faure.

DEFINITION 7.12 (Faure's Permutation) *Begin with the permutation* $\pi_2 = (0, 1)$ *for* $b = 2$ *and build the permutation* π_b *by:*

i) taking the values of $2\pi_{\frac{b}{2}}$ *and appending the values from* $2\pi_{\frac{b}{2}} + 1$ *if* b *is even*

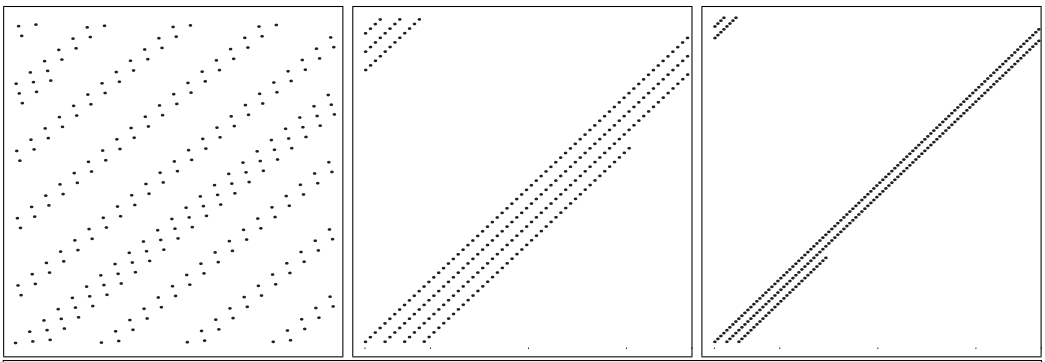


FIGURE 7.10: 2-DIMENSIONAL HALTON SEQUENCES. The first 256 elements of the 2-dimensional Halton sequence with $\mathbf{P}^2_{\mathrm{HAL}} = \left(\Phi_9(i), \Phi_{10}(i)\right)$, $\mathbf{P}^2_{\mathrm{HAL}} = \left(\Phi_{19}(i), \Phi_{20}(i)\right)$ and $\mathbf{P}^2_{\mathrm{HAL}} = \left(\Phi_{29}(i), \Phi_{30}(i)\right), 0 \leq i \leq 255$. The three patterns illustrate the problem with Halton sequences: Points in successive dimensions are highly correlated, which can lead to bad integral estimates. So, Halton sequence of dimension 14 and more are unsatisfactory. Due the correlation, many people avoid the use of Halton sequence for more than 6 or 8 dimensions in practice.

ii) taking the values of π_{b-1}*, inserting the value* $\frac{b-1}{2}$ *in the middle of the values and incrementing each value greater or equal than* $\frac{b-1}{2}$ *by one.*

The *scrambled versions* of the 2-dimensional Halton sequence $\mathbf{P}^{s,scr}_{\mathrm{HAL}}$ and the 2-dimensional Hammersley point set $\mathbf{P}^{s,scr}_{\mathrm{HAM}}$ may now be formulated as follows:

$$\mathbf{x}^s_i \stackrel{def}{=} \left(\Phi_2(i,\pi_2), \Phi_3(i,\pi_3), \Phi_5(i,\pi_5), \ldots, \Phi_{p_s}(i,\pi_{p_s})\right), \quad i \in \mathbb{N}_0 \tag{7.51}$$

and

$$\mathbf{x}^s_i = \left(\frac{i}{N}, \Phi_2(i,\pi_2), \Phi_3(i,\pi_3), \ldots, \Phi_{p_s}(i,\pi_{p_{s-1}})\right), \quad i \in (0,\ldots,N-1) \tag{7.52}$$

with

$$\begin{aligned}
\pi_2 &= (0,1)\\
\pi_3 &= (0,1,2)\\
\pi_4 &= (0,2,1,3)\\
\pi_5 &= (0,3,2,1,4)\\
\pi_6 &= (0,2,4,1,3,5)\\
\pi_7 &= (0,2,5,3,1,4,6)\\
\pi_8 &= (0,4,2,6,1,5,3,7)\\
&\vdots
\end{aligned}$$

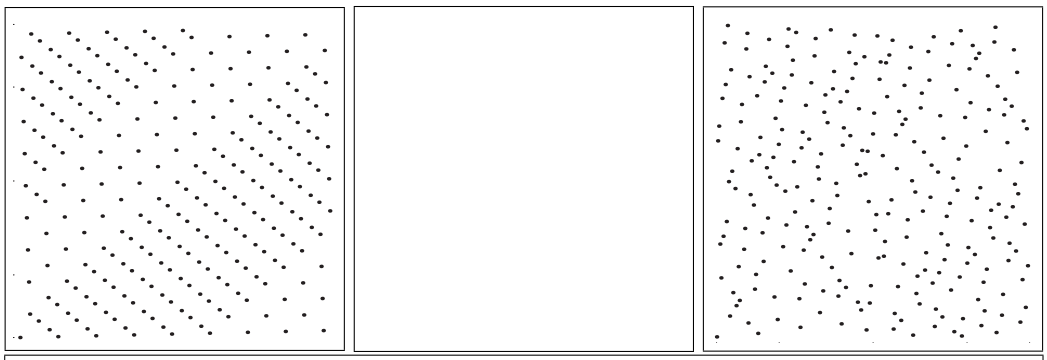


FIGURE 7.11: HALTON SEQUENCE AND SCRAMBLED HALTON SEQUENCE, DIMENSIONS 7 AND 8. **(a)** The first 256 elements of the 2-dimensional Halton sequence $\mathbf{P}^2_{\mathrm{HAL}} = \left(\Phi_7(i), \Phi_8(i)\right)$ and the scrambled versions of dimension 7 and 8 generated according to the procedure of Faure.

EXAMPLE 7.9 (Halton and Scrambled Halton Sequence) *Figure 7.11 illustrates the effect of scrambling applied to dimensions 7 and 8 of the 2-dimensional Halton sequence* $\mathbf{P}^2_{\mathrm{HAL}} = \left(\Phi_7(i), \Phi_8(i)\right)_{i\in\mathbb{N}_0}$ *with* $N = 256$.

In [138, Niederreiter 1992] it is shown that the coefficients in the error boundaries of Halton sequences and Hammersley point sets, as the product of the first s and $s-1$ factors $\frac{b_i-1}{2\log b_i}$, grow super-exponentially with respect to s. As this in turn means that the indicated boundaries are of use only for small s, it becomes necessary to search for procedures and methods able to deliver significantly smaller boundaries than those indicated in the Equations (7.40) and (7.47).

7.2.3 (t, m, s)-NETS AND (t, s)-SEQUENCES

The currently most promising theory underlying the construction of sequences with much smaller error bounds than those indicated in the previous section is that of the (t, m, s)-*nets* and the (t, s)-*sequences*.

(t, m, s)-NETS. The defining terms of a (t, m, s)-net are best understood in connection with a practical example from numerical integration. For that purpose, let us assume, we are interested in evaluating a 2-dimensional integral on the unit interval $\mathbf{I}^2$, where the samples come from a (t, m, s)-net in base $b = 2$. Since the parameter s in the triple (t, m, s) stands for the dimension of the space, s is in our case 2, that is, our net is of the form $(t, m, 2)$. Now, as the length N of the sequence of samples is depending on the parameter m, we can not choice the number of samples as an arbitrary integer. Due to

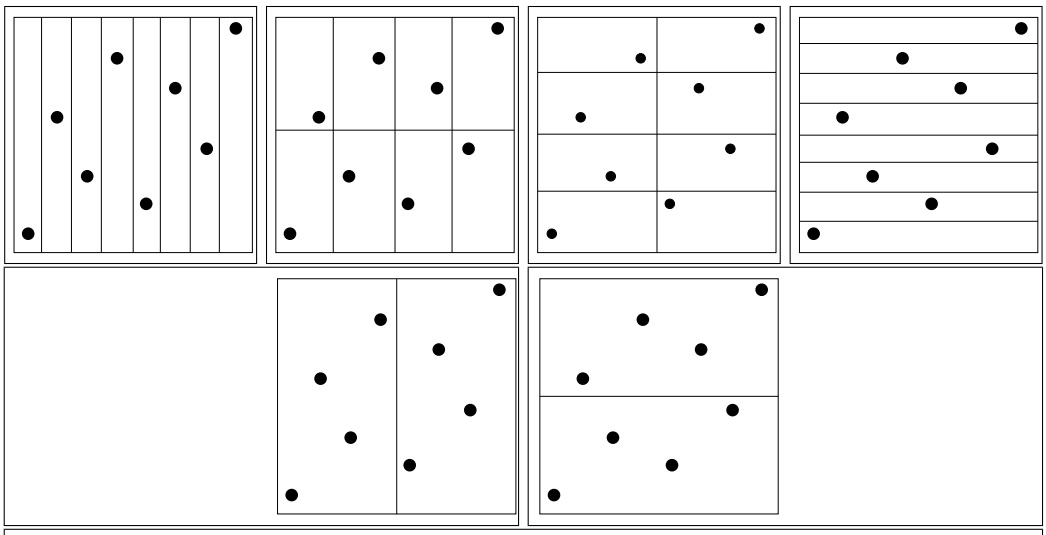


FIGURE 7.12: ELEMENTARY INTERVALS OF A $(1,3,2)$-NET AND $(2,3,2)$-NET. The 4 elementary intervals of the $(1,3,2)$-net and the 2 elementary intervals of the $(2,3,2)$-net associated with a given point set.

its definition, the length of a (t,m,s)-net is prescribed by the relation $N = b^m$, i.e., only sequences of length of powers of 2 are possible, e.g. $2^3 = 8$, which implies: $m = 3$. This results in a net of the form $(t,3,2)$ in base 2. Now, we come to the choice of the parameter t, which is a little bit tricky. A (t,m,s)-net in base b must contain b^t points in each s-dimensional subspace—under the condition, that the subspaces are created by splitting each dimension into b^k segments of length $\frac{1}{b^k}$ with volume b^{t-m} for some $k = 0,1,\ldots,m$. This means, that we must find a partition of each dimension in k elementary intervals that all contains only b^t elements given that the length of an elementary intervals is $\frac{1}{b^k}$ and its volume is b^{t-m}. With respect to the point set shown in Figure 7.12, it cannot be a $(0,3,2)$-net as the first elementary interval in the left pattern contains 2 points. This point set can only be a $(1,3,2)$-net or a $(2,3,2)$-net.

DEFINITION 7.13 ((t,m,s)-Net) *Let the* dimension $s \geq 1$, *the base* $b \geq 2$, *then for* $l_j \geq 0$ *and* $0 \leq a_j \leq b^{l_j}$

$$\mathcal{E} \stackrel{\text{def}}{=} \prod_{j=1}^{s} \left[\frac{a_j}{b^{l_j}}, \frac{a_j+1}{b^{l_j}} \right) \subseteq [0,1)^s$$

is called an elementary interval $\mathcal{E}$ *in the base* b *with Lebesgue measure* μ^s (82)

$$\mu^s(\mathcal{E}) = \frac{1}{b^{\sum_{j=1}^{s} l_j}}. \tag{7.53}$$

Based on this definition, a finite, b^m-element, s-dimensional point set is termed a (t, m, s)-net in the base b if every elementary interval $\mathcal{E}$ with the Lebesgue measure $\mu^s(\mathcal{E}) = b^{t-m}$ contains exactly b^t points for $0 \leq t \leq m$, $t, m \in \mathbb{N}$.

In this connection, the term s relates to the dimension *of the net, m is the power to which the base b is raised to obtain the length of the net, thus b^m, and t is a so-called* quality parameter*, where smaller values of t guarantee a better uniform distribution, i.e. a better stratification of the unit cube.*

Uniform Distribution (180)
Section (6.6.4)

Due to this definition, an elementary interval in base b is an axis aligned box where each dimension of the box must be a negative power of b, and the box must be aligned to an integer multiple of its size in each dimension. Further, we conclude that a (t, m, s)-net can only be constructed for lengths that are some power of the net's base, thus in the case of base 2, a (t, m, s)-net can only have lengths $2, 4, 8, ...$; it cannot have a length of, say 100 or 1000. In this sense (t, m, s)-nets are less flexible than Halton sequences.

Halton Sequence (632)

EXAMPLE 7.10 *According to the above definition a $(0, m, s)$-net with respect to the base b corresponds to a b^m element point set, whereby every elementary interval with the Lebesgue measure $\mu(\mathcal{E}) = \frac{1}{b^m}$ contains exactly one point out of $\mathbf{P}$.*

This means that the $(0, 3, 2)$-net with respect to the base 3 defines a $3^3 = 27$-element point set in the unit square. Since $t = 0$, the number of samples in each elementary interval is $b^t = 3^0 = 1$. Obviously, there are $b^{m-t} = 3^{3-0} = 27$ elementary intervals of Lebesgue measure $\frac{1}{27}$. These elementary intervals can be constructed by partitioning $\mathbf{I}^2$ in 27 segment in direction of the x-axis, or 9 segment in x and 3 segment in y-direction, 3 segments in x and 9 segments in y-direction, or 27 segments in direction to the y-axis.

EXAMPLE 7.11 *Let us consider the $(0, 2n, 2)$-net, then it can be shown that this is identical to the Hammersley points $\mathbf{P} = \left(\frac{i}{N}, \phi_2(i)\right)_{i \in (0,\dots,N-1)}$ with $N = (2^n)^2$, which were introduced in the last section. Here each elementary interval with Lebesgue measure $\frac{1}{2^{2n}} = \frac{1}{N}$ contains exactly one point out of $\mathbf{P}$. Figure 7.13 illustrates the above statement for $n = 2$ with respect the base $b = 2$.*

Hammersley Point Set (634)

CONSTRUCTION OF A (t, m, s)-NET IN BASE p. A $(t, m, 1)$-net, thus a one-dimensional (t, m, s)-net, can simply be created via the algorithm for computing a one-dimensional Halton sequence. In this algorithm, as we presented them in Figure 7.5, only an additionally step has to be inserted after reversing the representation $\bar{i}_p = 0.\alpha_0\alpha_1 \dots \alpha_{m-1}$ of a number $i \in \{0, 1, \dots, N-1\}$ in base p around the decimal point. This new step is a matrix-vector multiplication, where the m-dimensional vector ν is given by the digits of the representation $\bar{i}_p$ occurring after the decimal point, thus $(\alpha_0, \alpha_1, \dots, \alpha_{m-1})^T$ and the matrix M, with coefficients of $\{0, 1, \dots, p-1\}$ has to be chosen appropriately. Then, the result of the matrix-vector multiplication, where the arithmetic must be performed modulo p, has to be attached after the decimal point. Converted in base 10 we get the position

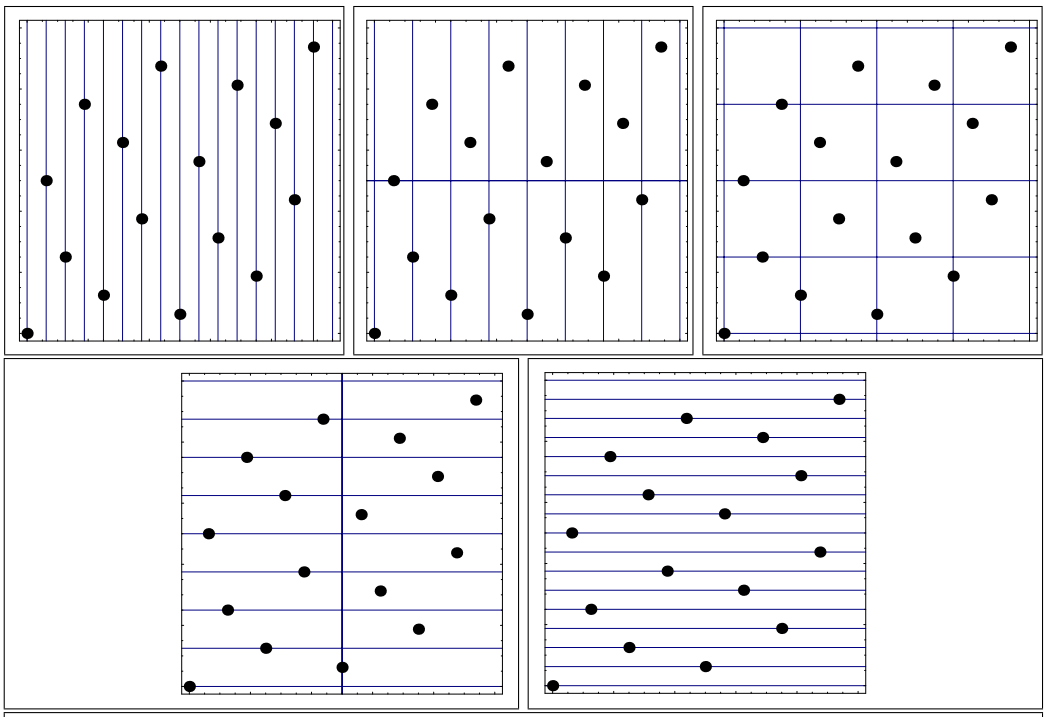


FIGURE 7.13: ELEMENTARY INTERVALS OF A $(0,4,2)$-NET. The 5 elementary intervals of the $(0,4,2)$-net defined in the unit cube with respect to the base $b = 2$ and a possible distribution of $16 = 2^4$ points that covers all cells of the net. The goal is to place points in a way that gives the best cover. Note, all points lie on boundaries of elementary intervals. A slight shift, the same in both directions and applied to all points places the points in the interior of the elementary intervals.

of the number i in the unit interval $[0,1]$, The algorithm for computing a $(t,m,1)$-net is shown in Figure 7.14.

Nets in s dimensions are created by using the same base p for all dimensions but applying different generating matrices $M_1,\ldots,M_s$. For the construction of a two-dimensional $(0,3,2)$-net and the computation of its elements, see Table 7.2 and Figure 7.15.

REMARK 7.15 *Due to the fact that all dimensions of a (t,m,s)-net are defined in the same base—but a Halton sequence uses different bases—we can conclude, from the above construction of (t,m,s)-nets with generating matrices, that at Halton sequence can not be a (t,m,s)-net.*

REMARK 7.16 *Until now, we considered almost exclusively $(0,m,s)$-nets. Compared to this simplified case, where every elementary interval contains only a single point, in*

One-dimensional (t, m, s)-Net in Base p {
  $\forall\ i \in \{0, \dots, p^m - 1\}$ {
    compute $i_p = \alpha_{m-1} \dots \alpha_0$ with $\alpha_i \in \{0, \dots, p-1\}$ and $i = \sum_{i=0}^{m-1} \alpha_i p^i$
    compute $\bar{i}_p = 0.\alpha_0 \cdots \alpha_{m-1}$ by reversing i_p around the decimal point
    compute $(\alpha'_0, \cdots, \alpha'_{m-1}) = M \cdot (\alpha_0, \dots, \alpha_{m-1})^T$
    convert $\bar{i}'_p = 0.\alpha'_0 \cdots \alpha'_{m-1}$ in base 10
  }
}

FIGURE 7.14: ONE-DIMENSIONAL (t, m, s)-NET IN BASE p. The i^{th} element of a one -dimensional (t, m, s)-net can be computed in a four step process. First, every number from $\{0, \dots, p^m - 1\}$ is represented as a number in base p. In the 2^{nd} step the representation the digits of the number, computed in base p, are reversed and attached after the decimal point. In the third step, the reversed digits are used to form a m-dimensional vector, which is multiplied by an appropriated chosen matrix, where the arithmetic is performed modulo p. The components of the resulting vector then forms the new digits after the decimal point. Converting the number in base 10 delivers the position of point p in the unit interval $[0, 1]$.

a general (t, m, s)-net, with $t > 0$, every box of size b^{t-m} must contain $b^t > 1$ points, see Figure 7.16.

(t, s)-SEQUENCES. Let us conclude the present discussion on the construction of low-discrepancy sequences with a short overview of the so-called (t, s)-sequences and their randomized *versions* in the following subsection..

Section 7.2

DEFINITION 7.14 ((t, s)-Sequences) *An infinite sequence of numbers $(\mathbf{x}_n)_{n \in \mathbb{N}}$ is referred to as a* (t, s)-sequence with respect to the base b *with $t \geq 0$, if the partial sequence*

$$\mathbf{x}_{kb^m+1}, \dots, \mathbf{x}_{(k+1)b^m} \in I^s \tag{7.55}$$

forms a (t, m, s)-net for all $k \geq 0$ and $m \geq t$ in the base b.

Star Discrepancy (622)

So, we get for the star-discrepancy of a (t, m, s)-net

$$D_N^*(\mathbf{P}) \leq B(s, b) b^t \frac{\log^{s-1} N}{N} + O\left(b^t \frac{\log^{s-2} N}{N}\right), \tag{7.56}$$

and for a (t, s)-sequence

$$D_N^*(\mathbf{P}) \leq C(s, b) b^t \frac{\log^s N}{N} + O\left(b^t \frac{\log^{s-1} N}{N}\right) \tag{7.57}$$

i	$p_1 = 2$	$\bar{i}_2$	M_1	M_2	(x, y)
0	000_2	0.000_2	0.000_2	0.000_2	$(0, 0)$
1	001_2	0.100_2	0.100_2	0.001_2	$\left(\frac{1}{2}, \frac{1}{8}\right)$
2	010_2	0.010_2	0.010_2	0.010_2	$\left(\frac{1}{4}, \frac{1}{4}\right)$
3	011_2	0.110_2	0.110_2	0.011_2	$\left(\frac{3}{4}, \frac{3}{8}\right)$
4	100_2	0.001_2	0.001_2	0.100_2	$\left(\frac{1}{8}, \frac{1}{2}\right)$
5	101_2	0.101_2	0.101_2	0.101_2	$\left(\frac{5}{8}, \frac{5}{8}\right)$
6	110_2	0.011_2	0.011_2	0.110_2	$\left(\frac{3}{8}, \frac{6}{8}\right)$
7	111_2	0.111_2	0.111_2	0.111_2	$\left(\frac{7}{8}, \frac{7}{8}\right)$

TABLE 7.2: COMPUTATION OF A $(0, 3, 2)$-NET IN BASE 2. In a first step, the algorithm from Figure 7.14 computes the binary representation i_2 of a number $i \in \{0, \ldots, N-1\}$. Then, the algorithms reverses all bits of i_2 around the decimal point. The digits after the decimal point in column 4 are then written as a 3-dimensional vector, e.g., 0.101_2 is written as $(1, 0, 1)^T$. This vector is multiplied with matrix M_1, and M_2, where arithmetic is performed modulo 2 and it holds:

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad M_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{7.54}$$

The result, expressed in base 10, then corresponds to the x-component of the point $p \in \mathbf{I}^2$.

given the corresponding constants $B(s, b)$ und $C(s, b)$, see [138, Niederreiter 1992].

EXAMPLE 7.12 *(i) Van der Corput sequences with respect to the base b correspond to $(0, 1)$-sequences in base b; (ii) If one expands a (t, s)-sequence by the component $\frac{i}{N}$, with $N = b^m$, one receives a $(0, m, 2)$-net.*

Van der Corput Sequence (631)

7.2.4 RANDOMIZED (t, m, s)-NETS AND (t, s)-SEQUENCES

A significant disadvantage of low-discrepancy sequences is its deterministically distribution of samples. In computer graphics, this leads to aliasing artifacts in particular in the application to pixel sampling. Since, in contrast to sequences of random numbers, one does not try to reproduce features, such as the mutual independence of individual members, estimations of error sizes based on independent individual members are also not possible. Randomized quasi-Monte Carlo procedures, on the other hand, combine low-discrepancy sequences with random numbers in order to generate independent estimations on the value of the integral. With the help of these estimations, empirical standard deviations may be calculated.

LD Sequence (629)
Random Variable (168)
LD Sequence (629)

In the following, we will briefly present two approaches of randomized (t, m, s)-nets

(t, m, s)-net (641)

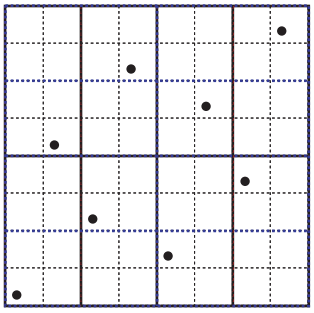


FIGURE 7.15: A $(0, 3, 2)$-NET CONSTRUCTED VIA GENERATING MATRICES. A 2-dimensional (t, m, s)-net with 8 points based on the algorithm for generating one-dimensional (t, m, s)-nets via generating matrices from Figure 7.14.

and (t, s)-sequences: *Cranley-Patterson rotation* and *Owen scrambling*.

In the procedure of Cranley and Patterson, the elements $\mathbf{x}_i$ of an s-dimensional point set are shifted by the amount $\xi^s = (\xi_1, \ldots, \xi_s)$ in the corresponding dimension $\mod 1$. This may lead to the loss of the original structure of a (t, m, s)-net, see Figure 7.17.

In contrast to *Cranley-Patterson rotation*, *Owen scrambling* largely retains the structure of (t, m, s)-nets of the base b. This procedure begins by decomposing the interval $[0, 1)^s$ along each of its s coordinates in b equal size parallelepipeds $\mathbf{Q}_1, \ldots, \mathbf{Q}_b$. These are permuted in a random and independent manner where this process is then applied recursively to $\mathbf{Q}_i, 1 \leq i \leq b$.

7.3 FOURIER ANALYSIS

In the previous sections a number of low-discrepancy sequences have been presented, which may be used for sampling domains of integral equations. In this section we now present a method, which makes it possible to formulate statements on the quality of patterns created by low-discrepancy sequences on $\mathbf{I}^s$: the *Fourier analysis*.

Fourier analysis provides us with a tool for the graphic interpretation of low sampling patterns resulting from various different sampling processes: the *Fourier-transform*.-

Fourier Transform (113)

Transformed in the Fourier domain, an image is represented as a weighted set of spatial frequencies. In this manner, it is very easy to draw conclusions from composition and character of an image, whereas high frequencies corresponds to fine structures in the image and low frequencies reflect slower changes in the structure of the image.

With reference to the definition of convolution given in Example 2.43, the quasi-Monte

quasi-Monte Carlo estimator (500)

Carlo estimator $F_N^{\mathbf{Q}^s}$ approximating the expected value of the integral

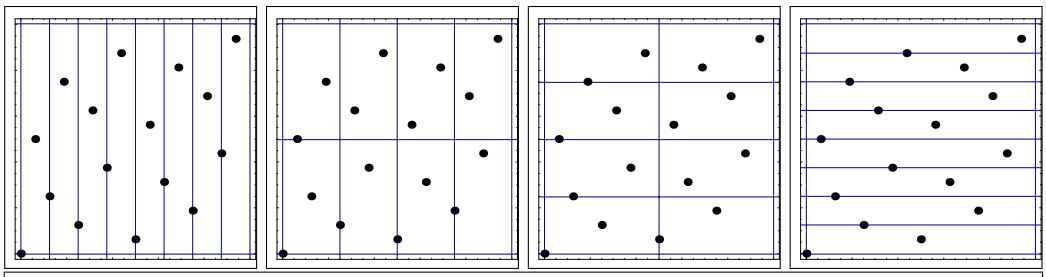


FIGURE 7.16: ELEMENTARY INTERVALS OF A $(1,4,2)$-NET. The 5 elementary intervals of the $(1,4,2)$-net defined in the unit cube with respect to the base $b = 2$ and a possible distribution of $16 = 2^4$ points that covers all cells of the net. Note, some points lie on boundaries of elementary intervals. A slight shift, the same in both directions and applied to all points places the points in the interior of the elementary intervals.

$$\int_{\mathbf{Q}^s} f(\mathbf{x}) d\mu^s(\mathbf{x})$$

may be formulated as a convolution of the integrand f and a sampling function s

$$s : \mathbb{R}^s \to \mathbb{R}$$

given by

Dirac δ-Distribution (118)

$$s(\mathbf{x}) \stackrel{\text{def}}{=} \frac{\lambda(\mathbf{Q}^s)}{N} \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i), \quad \mathbf{x}_i \in \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \tag{7.58}$$

whereas $\delta(\mathbf{x})$ corresponds to the Dirac δ-distribution. For the quasi-Monte Carlo estimator $F_N^{\mathbf{Q}^s}$ then it holds

$$\begin{aligned} F_N^{\mathbf{Q}^s} &\stackrel{(6.112)}{=} \frac{\lambda(\mathbf{Q}^s)}{N} \sum_{i=1}^{N} f(\mathbf{X}_i) \\ &\stackrel{(2.302)}{=} \frac{\lambda(\mathbf{Q}^s)}{N} \sum_{i=1}^{N} \int_{[-\infty,\infty]^s} f(\mathbf{y}) \delta(\mathbf{y} - \mathbf{x}_i) d\mu^s(\mathbf{y}) \\ &= \int_{[-\infty,\infty]^s} f(\mathbf{y}) \left(\frac{\lambda(\mathbf{Q}^s)}{N} \sum_{i=1}^{N} \delta(\mathbf{y} - \mathbf{x}_i) \right) d\mu^s(\mathbf{y}). \end{aligned}$$

Considering the manner in which the above sampling function works, it becomes clear that important information may be derived from the analysis of the Fourier spectrum of the pattern resulting from the sampling strategy applied. Thus, if $\widehat{s}(\mathbf{t}_1, \mathbf{t}_2) \stackrel{\text{def}}{=} \mathcal{F}s$ is taken to define the Fourier-transform of the sampling function s, defined with respect to $\mathbf{I}^2$, then

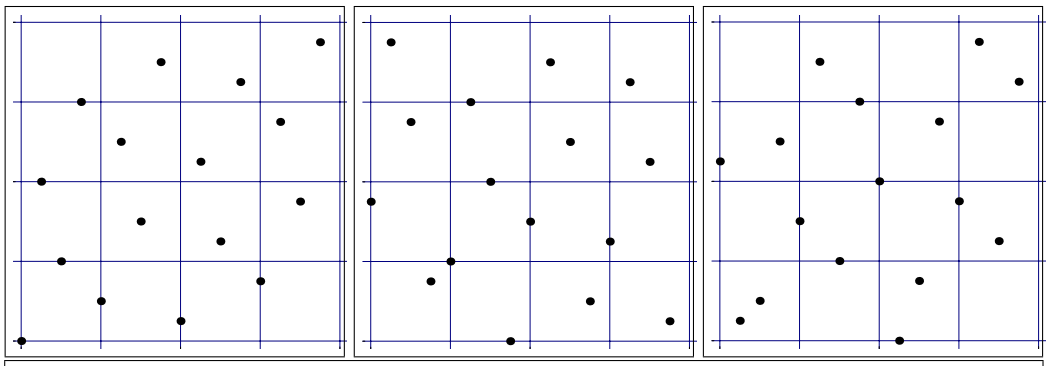


FIGURE 7.17: RANDOMIZED (t, m, s)-NETS. A $(0, 4, 2)$-net and its randomized version shifted by $\xi = (\frac{1}{4}, \frac{1}{4})$. Right, a randomized version of the left pattern violating the (t, m, s)-structure shifted by $\xi = (\frac{1}{8}, \frac{1}{8})$.

according to the above made statements, the following applies

$$\begin{aligned}
\mathcal{F}s \overset{\text{def}}{=} \widehat{s}(\mathbf{t}_1, \mathbf{t}_2) &\overset{(2.278)}{=} \frac{1}{(2\pi)} \int_{\mathbf{I}^2} \exp^{-i\langle \mathbf{t}, \mathbf{x}\rangle} s(\mathbf{x}) d\mu^2(\mathbf{x}) \\
&\overset{(7.58)}{=} \frac{1}{(2\pi)} \int_{\mathbf{I}^2} \exp^{-i\langle \mathbf{t}, \mathbf{x}\rangle} \left(\frac{\lambda(\mathbf{Q}^s)}{N} \sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i) \right) d\mu^2(\mathbf{x}) \\
&\approx \frac{\lambda(\mathbf{Q}^s)}{(2\pi)\, N} \sum_{i=1}^{N} \exp^{-i(\mathbf{t}_1 \mathbf{x}_{1i} + \mathbf{t}_2 \mathbf{x}_{2i})}, \quad \mathbf{x}_{1i}, \mathbf{x}_{2i} \in \mathbb{R}^2.
\end{aligned}$$

If we choose the samples of s as points of a regular grid, from the Halton sequence or the Hammersley point set, then, as may be clearly seen in Fig. 7.18, the high frequencies in their Fourier spectra indicate fine structures in the underlying pattern. Used in sampling strategies for pixel filtering, this results in aliasing effects.

Halton sequence (632)
Hammersley Point Set (634)

Poison-disk Sampling (541)

REMARK 7.17 (Poisson-Disk Sampling) *In the fovea of the human eye the photoreceptors are arranged according to a regular structure, a fact that corrects the aliasing effects created by the sampling process in the eye via the low-pass filtering features of the lens. Outside the fovea these are arranged according to Poisson-disk patterns [40, Cook 1984]. Here, the distance property of the samples replace the aliasing effects created by high frequencies with image noise, which is strictly more pleasant for the human eye. This clearly makes Poisson-disk sampling the most important sampling strategy for the objectives aimed at here. It follows, therefore, that in the generation of suitable sampling procedures efforts should focus on the construction of patterns similar to Poisson-disk patterns – providing, of course, that these are easily and efficiently calculable.*

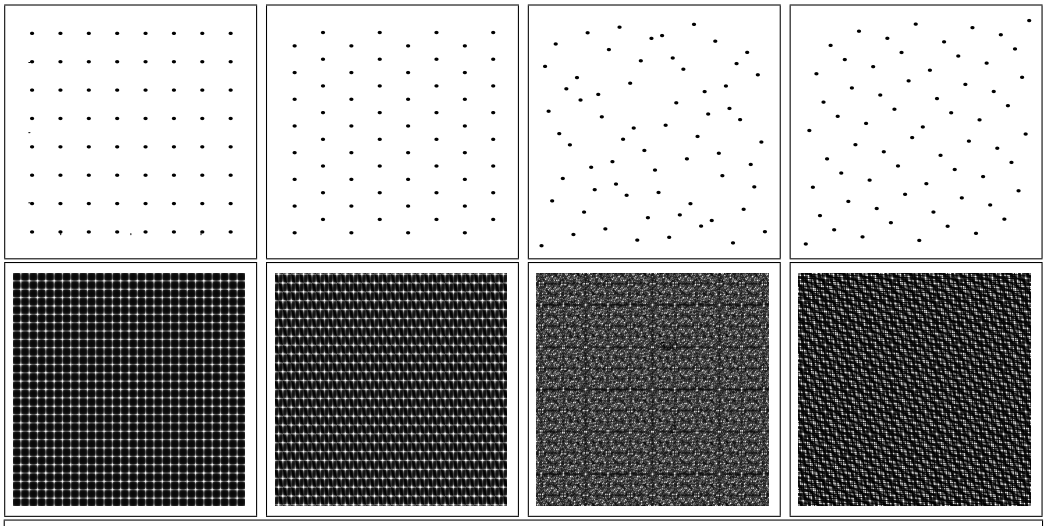


FIGURE 7.18: 64-ELEMENT POINT SETS WITH CORRESPONDING FOURIER SPECTRA. Regular grid, hexagonal-grid, Halton sequence, and Hammersley point set with associated Fourier spectra.

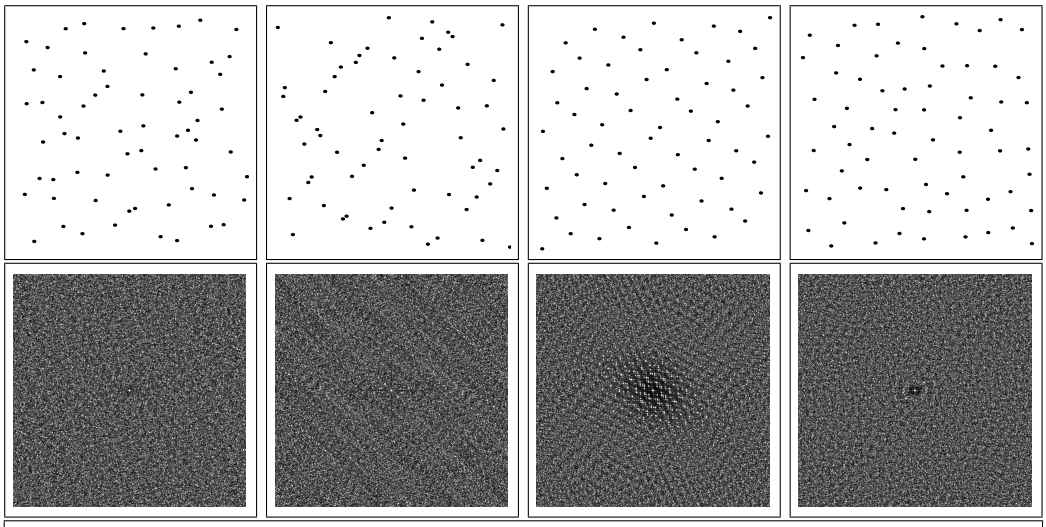


FIGURE 7.19: 64-ELEMENT POINT SETS WITH CORRESPONDING FOURIER SPECTRA. Jittered, N-rooks, jittered-Hammersley, and via Poisson-disk sampling generated patterns with associated Fourier spectra.

As the final point, let us contemplate some of the patterns generated according to jittering or on the basis of the Latin hypercube, shown in 7.19. It is clearly to be seen that their Fourier-spectra show similarities with the Poisson-disk pattern. In fact, with respect to the frequency domain, they come relatively close to the ideal Poisson-disk pattern. Apart from the δ-peak typically found in the centre where it occurs particularly pronounced, Fig. 7.19 also shows the strong high-frequency noise and the relatively weak intensities in the low frequency area, phenomena to which the human eye is very sensitive [67, Glassner 1995].

REMARK 7.18 *In computer graphics, a long time low discrepancy sampling only was applied for pixel supersampling, but our discussions in the last sections have shown, that, the inherent stratification property makes low discrepancy sequences and point sets also usable in sampling strategies for computing direct or indirect illumination, the construction of so-called* quasi-random walks*, or multiple importance sampling, [106, Keller and Heinrich 1996], [102, Keller 1996], and [109, Kollig and Keller 2002].*

7.4 REFERENCE LITERATURE AND FURTHER READING

Our discussion about quasi-Monte Carlo techniques is mainly based on [138, Niederreiter 1992], the standard work on quasi-Monte Carlo integration and—with respect to the results of greatest interest to computer graphics—from [104, Keller 1998] and [100, Keller 2002]. Although [67, Glassner 1995], [50, Dutré &. al 2003], and [158, Pharr and Humphreys 2004] discuss the topic only incomplete and on a very high mathematical level, all three books were useful references for our presentation of the quasi-Monte Carlo integration in Chapter 7.

The team around Keller also wrote a series of papers which show that it can be of great advantage to use quasi-Monte Carlo integration for different subjects of global illumination. Here, we mention in particular [105, Keller and Heinrich 1994] and [106, Keller and Heinrich 1996], where quasi-Monte Carlo techniques are proposed and investigated for solving the radiance equation and the global illumination problem. Also [109, Kollig and Keller 2002] contains a variety information about the topic. The usage of randomized quasi-Monte Carlo integration resulting in an efficient bidirectional path tracing algorithm is discussed in [108, Kollig & Keller 2002] and quasi-random walk techniques for the approximation of functionals for solving second kind Fredholm integral equations can be found in [102, Keller 1996]. With respect to radiosity methods, we mention [101, Keller 1996], where a fast algorithm is presented for computing form factors, which is based on low discrepancy samples and which is superior to random sampling.

An excellent and easily understandable article on quasi-Monte Carlo integration are the SIGGRAPH 2003 course notes, [146, Owen 2003]. This paper can be considered as a useful tutorial, that describes in a short and simple manner the way from ordinary Monte Carlo sampling via stratification, jittering, and LHS to different QMC sampling techniques like digital nets, integration lattices, and randomized quasi-Monte Carlo sampling. The entire band width of sampling and reconstruction techniques in theory and practice is also covered in [67, Glassner 1995]. Additionally, we recommend [210, Szirmay-Kalos & Purgathofer 1999], where the integration of discontinuous functions is examined and it is explained what kind of improvements can be expected from quasi-Monte Carlo techniques.

The concept of the discrepancy is discussed in many papers, in particular with respect to sample uniform distributions and patterns. So, in [182, Shirley 1991] it is shown that the discrepancy concept is a useful metric for examining equidistribution sampling issues in computer graphics. An arbitrary-edge discrepancy measure, motivated by the edge aliasing problem from computer graphics, is introduced in [45, Dopkin & al. 1993]. A comparision of sampling patterns is given in [237, Wong & al. 1997].

Readers interested in the generation of further low-discrepancy sequences are referred in particular to [138, Niederreiter 1992], [104, Keller 1998] und [197, Sobol 1985]. A comparison of the Halton, Sobol, and Faure quasi-random sequences with respect to effects on convergence of certain properties of integrals is made in [134, Morokoff and Calfish 1995].

A quasi-Monte Carlo Method for integral equations that has the potential to improve upon the convergence rate of conventional Monte Carlo and quasi-Monte Carlo simulations is presented in [199, Spanier & Li, 1998] and an interested article that describes the evaluation of high-dimensional integrals with quasi-Monte Carlo principles is [58, Faure 2009].

The standard reference for useful information on (t, m, s)-nets and (t, s)-sequences is [138, Niederreiter 1992]. The technique for generating (t, m, s)-nets via generating matrices, presented in Section 7.2.3, is from [242, Sándor & Train, 2004]. In [212, Tan & Boyle 2000] two novel techniques for speeding up the generation of digital (t,s)-sequences are introduced. Based on these results a new algorithm for the construction of Owens randomly permuted (t, s)-sequences is developed and analyzed. In [23, Bierbauer & Edel 1997] a series of generating matrices for good ternary (t, m, s)-nets are presented. As the description of explicit constructions of (t, s)-sequences, such as Faure-, Niederreiter- and Sobol sequences would exceed the scope of this excursion into the generation of low-discrepancy sequences, the interested reader is referred to Niederreiters standard work on quasi-Monte Carlo methods [138, Niederreiter 1992]. A beautiful discussion on randomized quasi-Monte Carlo methods is presented in [145, Owen 1998].

The mathematical principle of Fourier analysis is often described in many books on Functional Analysis, for this see the *Reference Literature and Further Reading Section* in Chapter 2. For the less mathematical oriented reader to this topic, we recommend [62, Foley & al. 1987], [233, Watt 1992], [67, Glassner 1995], and [55, Encarnacao 1997 & al.] where Fourier analysis is discussed on high-level introductions.

CHAPTER EIGHT

THE CLASSIC RENDERING ALGORITHMS BASED ON THE PRINCIPLE OF RAY TRACING

In the previous sections we have studied the most important techniques for generating algorithms based on Monte Carlo methods for solving the light transport as well as the importance and the measurement equation. Before we discuss in the next chapter concrete approaches used in practice and discuss them in the mathematical framework built so far, it is imperative to give us an overview of the most relevant classic rendering algorithms based on the principle of ray tracing.

All ray tracing based rendering procedures pursue the same idea: They shoot from certain sources, such as importance or light emitters, rays into the scene and trace these rays on their travel over the scene objects. Depending on the starting points of the generated paths, we distinguish between two great classes of rendering algorithms: *shooting* and *gathering algorithms*. If a path has its origin at the observer, the algorithm is a *gathering algorithm*, see Figure 8.1. If it has its origin in one of the light sources of the scene, we speak of a *shooting algorithm*, see Figure 8.2. Shooting algorithms approximate solutions of the importance equation, while gathering methods lead to solutions for the different light transport equations.

OVERVIEW OF THIS CHAPTER. We begin our summary about the classic rendering algorithms based on the principle of ray tracing by introducing Heckbert's *regular expression notations* for paths. It gives us a convenient way for describing what happens if light travels through a scene. The first ray tracing algorithm, which we introduce, is *Ray Casting*, a technique for fast rendering a 3D scene. Afterwards, we discuss the *classic Whitted-style Ray Tracing*. It can be considered as the foundation of all ray-based procedures used in computer graphics. Finally, we then present the first ray tracing algorithm based on probabilistic approaches for solving the light transport equation in free space: *Distribution Ray Tracing*. As an early, general, but inefficient prototype of a Monte Carlo rendering algorithm, it includes many interesting techniques based on stochastic principles, which

Section 8.1
Section 8.2
Section 8.3
Section 8.4

GATHERING ALGORITHMS {
  $\forall$ pixel $\square_i \in (\square_1, \ldots, \square_M)$ do {
    $L = 0$
    $\forall$ $j \in (1, \ldots, N)$ {
      sample a point $\mathbf{p}_j \in \square_i$ and a direction $\omega_j \in S^2$
      generate an eye-ray $\mathbf{r} = \mathbf{p}_j + \alpha\omega_j$
      $L +=$ GatherRadiance($\mathbf{r}$) in scene
    }
    $L = \frac{L}{N}$
  }
}

FIGURE 8.1: GATHERING ALGORITHMS. Gathering algorithms correspond to solutions of the light transport equations. They start at the eye or at a virtual camera and gather via the function GatherRadiance($\mathbf{r}$) the contributions of light at intersection points of $\mathbf{r}$ with objects surfaces during its travel through the scene. $\square_i$ corresponds to a pixel on the image plane, ω_j is as usual a direction starting at the eye through pixel $\square_i$, and L is the radiance.

SHOOTING ALGORITHMS {
  $\forall$ emitters $\sun_i \in (\sun, \ldots, \sun_M)$ do {
    $L = 0$
    $\forall$ $j \in (1, \ldots, N)$ {
      sample a point $\mathbf{l}_j \in \sun_i$ and a direction $\omega_j \in S^2$
      generate an light-ray $\mathbf{r} = \mathbf{l}_j + \alpha\omega_j$
      $W +=$ DepositImportance($\mathbf{r}$) in scene
    }
    $W = \frac{W}{N}$
  }
}

FIGURE 8.2: SHOOTING ALGORITHM. Shooting algorithms correspond to solutions of the adjoint light transport equation. They start at a light source and deposit via the function DepositImportance($\mathbf{r}$) the contributions of importance at intersection points of $\mathbf{r}$ with objects surfaces during its travel through the scene. $\sun_i$ stands for a light source within the scene, ω_j is as usual a direction starting at the light source $\sun_i$, and W is the importance.

we will use in our further discussions in Chapter 9.

8.1 HECKBERT'S PATH NOTATION BASED ON REGULAR EXPRESSIONS

Recall, the Neumann series approach, considered as a mathematical model of light and importance transport, can be physically interpreted as follows: Each application of the exitant light transport operator $\mathbf{T}_{L_o}^{\partial\mathcal{V}}$ corresponds to the interaction of light at a surface along a path starting at a light source within a scene. So, L_e, $\mathbf{T}_{L_o}^{\partial\mathcal{V}} L_e$ and $\mathbf{T}_{L_o}^{\partial\mathcal{V}^2} L_e$ are three paths, where light comes directly, via one, as well as two bounces from a light source to a sensor, see Figure 8.3.

Neumann Series Approach (608)

$\mathbf{T}_{L_o}^{\partial\mathcal{V}}$ (457)

As we know, the reflection or refraction behavior of light at a surface can be described by a BSDF, f_s, which himself can be split into an ideal diffuse, f_s^o, an ideal specular, $f_s^\vee$, and a glossy component, f_s^{gl}, that is, the BSDF can be written as:

BSDF (371)

Composition of BSDF (375)

$$f_s = f_s^o + f_s^\vee + f_s^{gl}. \tag{8.1}$$

This splitting then implies also a splitting of the light transport operator $\mathbf{T}_{L_o}^{\partial\mathcal{V}}$ given by:

$$\mathbf{T}_{L_o}^{\partial\mathcal{V}} \stackrel{def}{=} \mathbf{T}_o^{\partial\mathcal{V}} + \mathbf{T}_\vee^{\partial\mathcal{V}} + \mathbf{T}_{gl}^{\partial\mathcal{V}}. \tag{8.2}$$

Applied to the Neumann series representation of the light transport from Equation (5.33) then it holds:

$$\begin{aligned} L_o &= \sum_{j=0}^{\infty} \left(\mathbf{T}_o^{\partial\mathcal{V}} + \mathbf{T}_\vee^{\partial\mathcal{V}} + \mathbf{T}_{gl}^{\partial\mathcal{V}}\right)^j L_e && (8.3)\\ &= L_e + \left(\mathbf{T}_o^{\partial\mathcal{V}} + \mathbf{T}_\vee^{\partial\mathcal{V}} + \mathbf{T}_{gl}^{\partial\mathcal{V}}\right) L_e + \left(\mathbf{T}_o^{\partial\mathcal{V}} + \mathbf{T}_\vee^{\partial\mathcal{V}} + \mathbf{T}_{gl}^{\partial\mathcal{V}}\right)^2 L_e + \dots && (8.4)\\ &= L_e + \mathbf{T}_o^{\partial\mathcal{V}} L_e + \mathbf{T}_\vee^{\partial\mathcal{V}} L_e + \mathbf{T}_{gl}^{\partial\mathcal{V}} L_e + \mathbf{T}_o^{\partial\mathcal{V}^2} L_e + \mathbf{T}_o^{\partial\mathcal{V}} \mathbf{T}_\vee^{\partial\mathcal{V}} L_e + \dots && (8.5) \end{aligned}$$

with

$$\begin{aligned} \left(\mathbf{T}_o^{\partial\mathcal{V}} L_o\right)(\mathbf{s}, \omega_o) &\stackrel{def}{=} \int_{S^2(\mathbf{s})} f_s^o(\mathbf{s}, \omega_i \to \omega_o)\, L_o(\mathbf{s}, \omega_o)\, d\sigma_{\mathbf{s}}^\perp(\omega_i) && (8.6)\\ \left(\mathbf{T}_\vee^{\partial\mathcal{V}} L_o\right)(\mathbf{s}, \omega_o) &\stackrel{def}{=} \int_{S^2(\mathbf{s})} f_s^\vee(\mathbf{s}, \omega_i \to \omega_o)\, L_o(\mathbf{s}, \omega_o)\, d\sigma_{\mathbf{s}}^\perp(\omega_i) && (8.7)\\ \left(\mathbf{T}_{gl}^{\partial\mathcal{V}} L_o\right)(\mathbf{s}, \omega_o) &\stackrel{def}{=} \int_{S^2(\mathbf{s})} f_s^{gl}(\mathbf{s}, \omega_i \to \omega_o)\, L_o(\mathbf{s}, \omega_o)\, d\sigma_{\mathbf{s}}^\perp(\omega_i) && (8.8) \end{aligned}$$

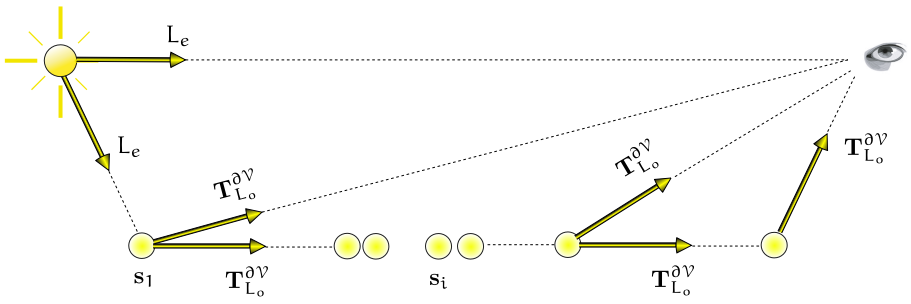


FIGURE 8.3: LIGHT PATHS OF DIFFERENT LENGTH. Light emitted by a source can directly arrive at the eye, this corresponds to a light path of length one. A light path of length two goes over the reflective surface $\mathbf{s}_1$ to the eye. Light paths of length ≥ 2 arrive at the eye by reflection at more than a single surface.

for any exitant function $L_o \in \mathcal{R}^{\partial\mathcal{V}}$.

Each term in this representation of the Neumann series corresponds to a path through the scene starting at a light source and ending at a sensor with different types of scattering in between. Since the analysis and the comparison of rendering algorithms often requires information about the chain of events occurring on such paths, the Equations (8.3) - (8.5) are not suitable to deliver this information.

A convenient way for describing what happens if light travels through a scene is given in [81, Heckbert 1990]. Heckbert describes a path trough a scene by a short string, which corresponds to a *regular expression* generated over a *finite alphabet* $\mathfrak{A}$ as known from automata theory, see [87, Hopcroft & al. 1979].

The alphabet $\mathfrak{A} = \{E, D, G, L, S\}$, over which we generate our regular expressions, is predefined by the physical phenomena occurring at a surface. Thus,

- L describes the emission of photons from a light source and
- E stands for absorption of photons at a sensor.

The reflection and transmission behavior at surfaces is abbreviated by

- D for ideal diffuse,
- S for ideal specular, and
- G for glossy reflection or refraction,

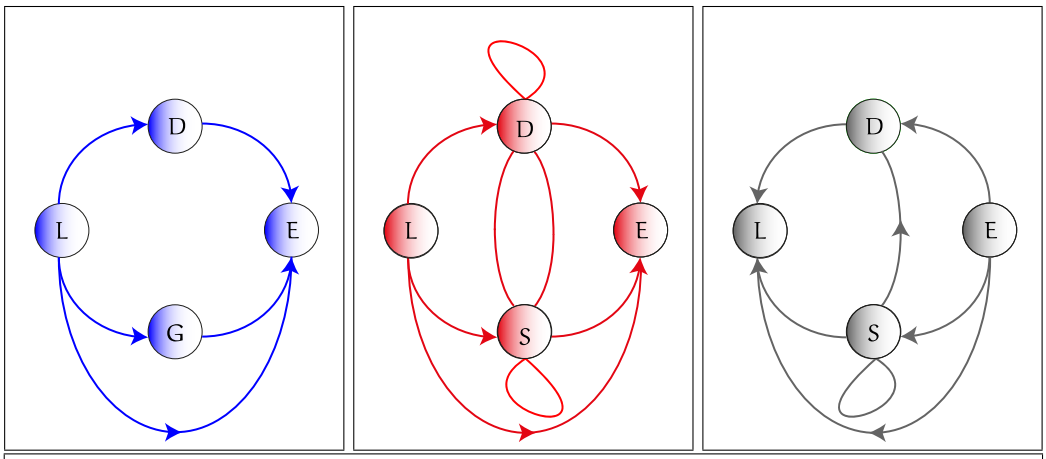


FIGURE 8.4: EXTENDED HECKBERT PATH NOTATION. On the left, one light path has length one, the others are light paths of length 2 over a diffuse or a glossy surface. The corresponding graph represents paths of characteristic $\overleftarrow{E[D|G]L}$, thus a local illumination model. The graph in the center describes light paths of characteristic $\overleftarrow{E(D|S)^*L}$, that is, he simulates global illumination in a scene consisting of diffuse and specular surfaces. As we will see in Section 8.3, the graph on the right hand side, simulates classic Whited-style ray tracing in scenes with specular and diffuse surfaces, thus eye paths of length ≥ 1 of type $\overrightarrow{ES^*[D]L}$.

that is, the abbreviations for the individual components of a BSDF.

Furthermore, we use standard regular expression notation as in [87, Hopcroft & al. 1979], that is: Subexpressions may be grouped in parentheses, the superscripts $*$ and $+$ correspond to 0 or more repetitions, respectively 1 or more repetitions of the substring, which they superscribe. A term in square brackets is optional and the vertical bar $|$ between two members indicates a selection among them. We expand Heckbert's original notation by an arrow over the expression, indicating the direction of computation, that is: gathering $\rightarrow$ and shooting $\leftarrow$.

Let us consider some simple regular expressions that classify various transport paths.

EXAMPLE 8.1 (Important Transport Paths Expressed in Expanded Heckbert Notation) *i) The simplest light path in a scene can be described by* $\overleftarrow{EL}$*, that is, light emitted from a source is transported directly—without interaction with objects in a scene—into the eye.* $\overleftarrow{EL}$ *is a light path of length* 1*, see Figure 8.4. A scene only rendered with these paths would be black except for visible light sources.*

ii) Obviously, local reflection models simulate paths of characteristic $\overleftarrow{E[D|G]L}$ *or* $\overrightarrow{E[D|G]L}$*, since local reflection models does not account for indirect illumination. Paths in local reflection models always have length* ≤ 2*, while paths in global illu-*

mination models are of characteristic $\overrightarrow{E(D|S|G)^*L}$ *or* $\overleftarrow{E(D|S|G)^*L}$ *with length* ≥ 1*, see Figure 8.4.*

iii) As we will show in Section 8.3, classic ray tracing *generates eye paths of characteristic* $\overrightarrow{ES^*[D|G]L}$*. That is, the algorithm generates paths starting at the eye and continuing over none, one, or more specular surfaces until a diffuse or a glossy surface is hit from where they are connected to a light source, see Figure 8.4.*

iv) The set of all paths starting at the eye of an observer and ending in a light source can be described by the regular expression $\overrightarrow{E(D|G|S)^*L}$*. This set consists of an eye path of length* 1 *or eye paths of length greater than* 1 *via diffuse, specular, or glossy reflection/refraction at object surfaces within the scene.*

v) In Chapter 10, we discuss radiosity algorithms*. These are procedures based on finite element methods for solving the light transport equation in scenes existing of purely diffuse surfaces. Hence, radiosity algorithms generate paths of characteristic* $\overleftrightarrow{ED^*L}$*, that is, they only consider diffuse reflections. So, we have an eye-light or light-eye path* $\overleftrightarrow{EL}$*, or eye respectively light paths of length greater than* 1 *over the diffuse object surfaces within the scene.*

vi) A caustic *is a light pattern generated by light that is reflected or transmitted at a number of subsequent specular surfaces before interacting with a diffuse surface, such as it occurs onto a table when light passing through a glass filled with water,* Monte Carlo Light Tracing (710) *wine, or so. For a long time, caustics were the most difficult to simulate light phenomena in computer graphics. Such a light pattern is based on eye subpaths of characteristic* $\overrightarrow{EDS^+L}$ *or light paths of type* $\overleftarrow{EDS^+L}$*.*

REMARK 8.1 *Since all paths must involve a light source L as well as the eye E, they have length at least equal to 1. A nice thing about this notation is that it is clear when certain types of paths are not traced, that is, when certain types of light transport are not considered by the algorithm. For example, the ray casting algorithm, introduced in the next section, only traces paths of length* ≤ 2*, namely* $\overrightarrow{E[D|G]L}$*, ignoring longer paths; thus, only direct lighting is considered.*

As we will show in Section 8.3, classic Whitted-style ray tracing *traces paths of any length, but all those paths begin with a sequence of zero or more specular reflection and refraction steps. Thus, Whitted's technique accepts paths of characteristic* Monte Carlo Path Tracing (692) $\overrightarrow{ES^*[D|G]L}$ *but ignores paths like* $\overrightarrow{EDSDSL}$ *or* $\overrightarrow{E(D|G)^*L}$*.* Monte Carlo path tracing *and* Distribution Ray Tracing (672) distribution ray tracing *can simulate paths where light bounces between non-specular surfaces such as* $\overrightarrow{E(D|G|S)^*L}$*. However, these methods have difficulties to accept paths of the form* $\overrightarrow{E(D|G)S^*L}$*—that is, multiple specular bounces from the light source as in a caustic—since it is often very unlikely to meet a small light source after a specular reflection. Obviously, any technique that ignores whole classes of paths will not correctly compute the solution to the light transport equation [74, Hanrahan 2001].*

8.2 RAY CASTING

From our considerations concerning the transport of light in free space we know that on the one hand radiance along a ray is invariant and, on the other hand physically defined BRDFs underly the Helmholz principle of reciprocity. Referring to the transport of light this means that it can be described in two ways: First, naturally by photon emission, where photons are emitted from light sources and arrive at a sensor, but also in the reverse direction, i.e. starting from a sensor and finding its way to a light source. In the introductory section to this chapter, we called those algorithms gathering algorithms, as they make use of the reversibility of the propagation of light.

Radiance (250)
BRDF (320)
Helmholz Reciprocity (331)
Gathering Algorithms (653)

CLASSIC RAY CASTING. Let us start to consider a very simple rendering technique of purely geometric nature where illumination plays no role: *Ray Casting*. Ray casting, a so-called *image-precision algorithm*, works as a *visibility detection tool* similar to *depth buffering*, see [62, Foley & al. 1987] and [78, Hearn & Backer 1994]. The algorithm follows the idea of shooting a mathematical ray $\mathbf{r} = \mathbf{e} + \alpha\omega, \alpha > 0$, starting at the camera or the observer's eye $\mathbf{e}$ and passing through a pixel of the image plane in direction ω into the scene to find the closest point on an object visible along this ray. Via the ray casting function γ the nearest intersection point of such a ray with an object can be computed. If the specified object is a light source, the current pixel is painted with the color of the light source, otherwise the pixel remains black, see Figure 8.5 and the left image in Figure 8.6. Since the light sources in ray casting algorithms are commonly assumed to be point light sources, an image rendered via classic ray casting is usually completely black.

Ray (11)
γ (47)

RAY CASTING EQUIPPED WITH A LOCAL ILLUMINATION MODEL. A slightly modified version of ray casting, above described as a purely visibility detection tool, includes a local illumination model into the process of visibility detection. The algorithm experiences a change in such a way that after the collision of a ray with an object, new rays are fired in direction to the existing light sources. Regardless whether these rays are blocked by other objects on its ways to the light sources, the corresponding pixel is always colored with the color of the closest object depending on the contribution of light that comes from the light sources, see Figure 8.7 and the right image in Figure 8.6. In this modified version ray casting can also simulate paths of length two, that is, paths of characteristic $\overrightarrow{E[D|G]L}$, which enables ray casting to produce a very fast preview of a scene.

Based on these considerations, ray casting can be considered as an approximate solver of a very simple version of the stationary light transport equation in vacuum, thus,

SLTEV (398)

$$L_o(\mathbf{s}, \omega_o) \quad = \quad L_e(\mathbf{s}, \omega_o) + \int_{\mathcal{S}^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i), \qquad (8.9)$$

where we assume that the BSDF is a composition of a diffuse f_s^o, and a glossy component f_s^{gl}.

```
CLASSIC RAY CASTING {
    ∀ pixel □_i ∈ {□_1, ..., □_{s_x · s_y}} do {
        choose center p ∈ □_i
        generate an eye-ray r = e → p
        compute closest hit point s of r with objects ∂V in scene
        if s ∈ {☼_1, ..., ☼_M} do {
            L(s → e) = L_e(s → e)
        } else {
            L(s → e) = 0
        }
    }
}
```

FIGURE 8.5: PSEUDOCODE FOR CLASSIC RAY CASTING. The classic ray casting algorithm shoots an eye-ray through the center of every pixel. If the closest object that has been hit by this ray is a light source, then the corresponding pixel is shaded with the color of the light source, otherwise, the pixel remains black.

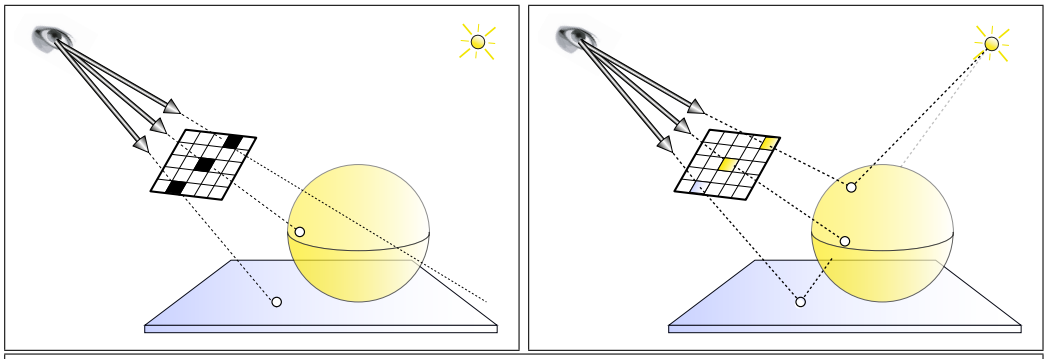


FIGURE 8.6: CLASSIC RAY CASTING AND RAY CASTING EQUIPPED WITH A LOCAL ILLUMINATION MODEL. In its classic version, ray casting shoots a ray through the center of a pixel and determines the hit point of this ray with an object in the scene. If the ray hits a light source, then the pixel is shaded with the color of the light source, in all other cases it is colored black. The modified version of ray casting uses a local illumination model but without to take into account the computation of shadows. If a ray hits an object in the scene, the algorithms computes shadow rays in direction to all light sources, which are assumed to be point light sources. The pixel associated with the primary ray gets the color of the object depending on the contribution of light coming from the light sources.

```
RAY CASTING with LOCAL ILLUMINATION {
    ∀ pixel □_i ∈ {□_1, ..., □_{s_x · s_y}} do {
        choose center p ∈ □_i
        generate an eye-ray r = e → p
        compute closest hit point s of r with objects ∂V in scene
        L(s → e) = 0
        ∀ light sources ☼_j ∈ (☼_1, ..., ☼_M) do {
            sample point l ∈ ☼_j
            L(s → e) += f_s(s, l → s → e) L_e(l → s)
        }
    }
}
```

FIGURE 8.7: PSEUDOCODE FOR RAY CASTING WITH LOCAL ILLUMINATION. If an eye-ray hits a scene object at point $\mathbf{s}$, the algorithm shoots shadow-rays in direction to all light sources and shades $\mathbf{s}$ depending on the contributions of the light sources and the reflection behavior of light at hit point $\mathbf{s}$.

The incident radiance L_i under the integral is simply replaced by the radiance L_e emitted from a light source, that is, the associated, simplified SLTEV has the form

Incident & emitted Function (48)

$$\begin{aligned} L_o(\mathbf{s}, \omega_o) &= L_e(\mathbf{s}, \omega_o) \\ &+ \int_{\mathcal{S}^2(\mathbf{s})} f_s^o(\mathbf{s}, \omega_i \to \omega_o)\, L_e(\gamma(\mathbf{s}, \omega_i), -\omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i) \qquad (8.10) \\ &+ \int_{\mathcal{S}^2(\mathbf{s})} f_s^{gl}(\mathbf{s}, \omega_i \to \omega_o)\, L_e(\gamma(\mathbf{s}, \omega_i), -\omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i), \end{aligned}$$

where $\gamma(\mathbf{s}, \omega_i)$ are points on the light sources in direction to $\mathbf{s}$.

Using the Relations (8.6) - (8.8), then Equation (8.10) corresponds to eye paths of characteristic

$$L_o = L_e + \underbrace{\left(\mathbf{T}_o^{\partial\mathcal{V}} + \mathbf{T}_{gl}^{\partial\mathcal{V}}\right)}_{[D|G]} L_e \qquad (8.11)$$

$$\equiv \overrightarrow{E[D|G]L}. \qquad (8.12)$$

Evidently, the integrand in Equation (8.10) is determined by the radiance emitted from the light sources. Instead to integrate over the entire unit sphere, a more efficient strategy could be to integrate only over the solid angles of all light sources projected onto the unit sphere.

Now, the projection of a countable set of point light sources on the unit sphere results in a null set, thus a set of measure zero, and the projection of an area light source onto the sphere is, due to [10, Arvo 1995], a very complex task. But both problems can be solved in a simple way: The problem of point light sources can be solved by using Dirac δ-distributions, and the projection of area light sources onto the unit sphere can be circumvented by formulating the above modified variant of the SLTEV in its 3-point form, where the integration domain is given via the area light sources existing in the scene.

Null Set (80)
Dirac δ-distribution (117)
3-point form (403)

RAY CASTING WITH POINT LIGHT SOURCES. In its classic version, ray casting assumes that the scene to be rendered is illuminated by point light sources. Therefore, let us firstly devote to the problem where a scene is only illuminated by m point light sources $*_j$, located at points $\mathbf{x}_j \in \mathbb{R}^3$ with $1 \leq j \leq M$.

Due to our derivation in Example 4.6, the factor $L_e(\gamma(\mathbf{s}, \omega_i), -\omega_i)$ in Equation (8.10) is then replaced by the product of the Dirac δ-distribution $\delta(\omega_i - \omega_i^{l_j})$, where $\omega_i^{l_j} \equiv \frac{\mathbf{s} \rightarrow *_j}{\|\mathbf{s} \rightarrow *_j\|_2}$ is the shadow ray between surface point $\mathbf{s}$ and the light source $*_j$, and the irradiance at points $\mathbf{s}$ given by

Dirac δ-distribution (117)
Irradiance (257)

$$\mathbf{E}(\mathbf{s}) = \frac{\Phi_{e_j}(*_j)}{4\pi} \frac{\left|\cos\theta_i^{l_j}\right|}{\|*_j - \mathbf{s}\|_2^2}, \tag{8.13}$$

where $\cos\theta_i^{l_j}$ is the angle between the surface normal at $\mathbf{s}$ and the direction $\omega_i^{l_j}$ towards the light source $*_j$. Using this construct in Equation (8.10), then we get:

$$\begin{aligned} L_o(\mathbf{s}, \omega_o) &= L_e(\mathbf{s}, \omega_o) + \qquad (8.14) \\ &\quad \int_{S^2(\mathbf{s})} \left(f_s^o(\mathbf{s}, \omega_i \longrightarrow \omega_o) + f_s^{gl}(\mathbf{s}, \omega_i \longrightarrow \omega_o)\right) \\ &\quad\quad \delta(\omega_i - \omega_i^{l_j})\, L_e(\gamma(\mathbf{s}, \omega_i), -\omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i) \\ &= L_e(\mathbf{s}, \omega_o) + \qquad (8.15) \\ &\quad \sum_{j=1}^{M} \left(f_s^o(\mathbf{s}, \omega_i^{l_j} \longrightarrow \omega_o) + f_s^{gl}(\mathbf{s}, \omega_i^{l_j} \longrightarrow \omega_o)\right) \frac{\Phi_{e_j}(*_j)}{4\pi} \frac{\left|\cos\theta_i^{l_j}\right|}{\|*_j - \mathbf{s}\|_2^2}. \end{aligned}$$

RAY CASTING WITH AREA LIGHT SOURCES. For simulating area light sources in ray casting we simply transform the spherical form of the above modified SLTEV into its 3-point representation. As already mentioned above, this transformation has the advantage that the integration domain is given via the surface areas of all area light sources $\mathbin{\rotatebox{45}{$\boxtimes$}}_j$, $1 \leq j \leq M$. In 3-point form, Equation (8.10) then looks like:

SLTEV in 3-point Form (402)

$$
\begin{aligned}
L(\mathbf{s} \rightarrow \mathbf{e}) &= L_e(\mathbf{s} \rightarrow \mathbf{e}) + \qquad (8.16)\\
&\quad \int_{\mathfrak{S}} \left(f_s^o(\mathbf{l} \rightarrow \mathbf{s} \rightarrow \mathbf{e}) + f_s^{gl}(\mathbf{l} \rightarrow \mathbf{s} \rightarrow \mathbf{e}) \right) L_e(\mathbf{l} \rightarrow \mathbf{s}) \widehat{\mathcal{G}}(\mathbf{l} \leftrightarrow \mathbf{s}) \, d\mu^2(\mathbf{l}),\\
&= L_e(\mathbf{s} \rightarrow \mathbf{e}) +\\
&\quad \sum_{j=1}^{M} \int_{\mathfrak{S}_j} f_s^o(\mathbf{l} \rightarrow \mathbf{s} \rightarrow \mathbf{e}) \, L_e(\mathbf{l} \rightarrow \mathbf{s}) \widehat{\mathcal{G}}(\mathbf{l} \leftrightarrow \mathbf{s}) \, d\mu^2(\mathbf{l}) + \qquad (8.17)\\
&\quad \sum_{j=1}^{M} \int_{\mathfrak{S}_j} f_s^{gl}(\mathbf{l} \rightarrow \mathbf{s} \rightarrow \mathbf{e}) \, L_e(\mathbf{l} \rightarrow \mathbf{s}) \widehat{\mathcal{G}}(\mathbf{l} \leftrightarrow \mathbf{s}) \, d\mu^2(\mathbf{l}),
\end{aligned}
$$

where $\mathbf{s} \rightarrow \mathbf{e} = \omega_o$ and $\mathbf{l}$ are points on the area light sources $\mathfrak{S}_j$.

REMARK 8.2 *Note, in the above both equations we use a modified geometry term* $\widehat{\mathcal{G}}$ *instead of the geometry term* $\mathcal{G}$*, which we usually use in our 3-point formulations of the SLTEV. The reason for this is the fact that a ray casting algorithm does not account for, whether rays, fired in direction to light sources, are blocked by other objects. This means that the visibility term in* $\mathcal{G}$ *can be neglected, that is, we can define:* $\mathcal{G}$ (129)

$$
\widehat{\mathcal{G}}(\mathbf{l} \leftrightarrow \mathbf{s}) \stackrel{def}{=} \frac{\mathcal{G}(\mathbf{l} \leftrightarrow \mathbf{s})}{\mathcal{V}(\mathbf{l} \leftrightarrow \mathbf{s})}. \qquad (8.18)
$$

An efficient and simple Monte Carlo sampling strategy for estimating $L(\mathbf{s} \rightarrow \mathbf{e})$ chooses a point $\mathbf{l}_j$ on each light source, generates a shadow ray from $\mathbf{s}$ in direction to $\mathbf{l}_j$, and computes the radiance arriving at $\mathbf{s}$ along this ray. For that purpose, we have to draw a sample from each of the light sources according to probability density functions $p_{\mathfrak{S}_j}$ on the probability spaces $(\mathfrak{S}_j, \mathfrak{B}(\mathfrak{S}_j), \mu^2)$. This then leads to the primary Monte Carlo estimator F_1^{RC} for approximating $L(\mathbf{s} \rightarrow \mathbf{e})$, given by: PDF (176) Probability Space (163)

$$
\begin{aligned}
F_1^{RC} &= L_e(\mathbf{s} \rightarrow \mathbf{e}) +\\
&\quad \frac{1}{M} \sum_{j=1}^{M} \frac{f_s^o(\mathbf{l}_j \rightarrow \mathbf{s} \rightarrow \mathbf{e}) \, L_e(\mathbf{l}_j \rightarrow \mathbf{s}) \, \widehat{\mathcal{G}}(\mathbf{l}_j \leftrightarrow \mathbf{s})}{p_{\mathfrak{S}_j}(\mathbf{l}_j)} + \qquad (8.19)\\
&\quad \frac{1}{M} \sum_{j=1}^{M} \frac{f_s^{gl}(\mathbf{l}_j \rightarrow \mathbf{s} \rightarrow \mathbf{e}) \, L_e(\mathbf{l}_j \rightarrow \mathbf{s}) \, \widehat{\mathcal{G}}(\mathbf{l}_j \leftrightarrow \mathbf{s})}{p_{\mathfrak{S}_j}(\mathbf{l}_j)},
\end{aligned}
$$

where the samples $\mathbf{l}_j$ are generated according to the probability densities $p_{\mathfrak{S}_j}$ on the areas of the light sources, and $L_e(\mathbf{s} \rightarrow \mathbf{e}) \neq 0$ if and only if $\mathbf{s} \in \mathfrak{S}_j$ for $1 \leq j \leq M$.

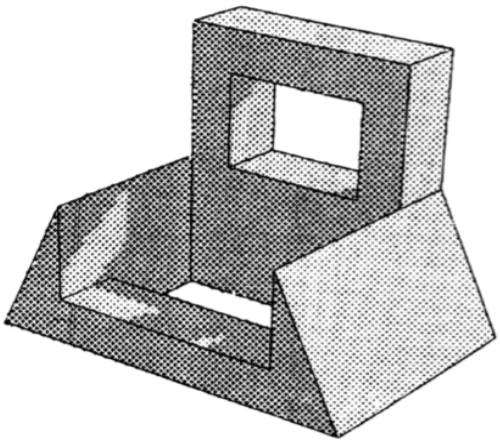


FIGURE 8.8: ONE OF THE FIRST IMAGES RENDERED WITH RAY TRACING. The image is rendered using a pen plotter, where a halftone pattern is simulated by locating a different sized +-sign at each pixel of the image plane depending on the pixel's intensity. Image courtesy of Arthur Apple.

8.3 CLASSIC WHITTED-STYLE RAY TRACING

We have seen that a ray casting algorithm only generates primary rays, and shadow rays in its extended version. It does not generate secondary rays after a primary ray has hit an object within a scene, nor it does account for whether shadow rays, fired in direction to a light source, are blocked by other objects. As such an algorithm only simulates light paths of characteristic $\overrightarrow{E[D|G]L}$, it can not reproduces simple light effects like shadows let alone indirect illumination. That is, images generated via a ray casting procedure are very sterile and unrealistic.

In [5, Apple 1968], the ray casting algorithm was extended to simulate shadows by taking into account whether a shadow ray is blocked by other objects on its way to one of the light sources. Therefore, Apple's algorithm can be seen as the first ray tracing algorithm, see Figure 8.8. Based on this technique, in [236, Whitted 1980] then a process is introduced, which repairs all the cons of ray casting, listed above. Additionally, this new approach also traces recursive rays in the reflection and/or refraction direction for reflective and refractive materials: *Classic Whitted-style Ray Tracing* was born.

THE CLASSIC WHITTED-STYLE RAY TRACING ALGORITHM. In analogy to ray casting, classic Whitted-style ray tracing also starts with shooting a primary ray from a sensor—typically the eye of an observer or a virtual camera—through a pixel of the image plane into the scene to be rendered. At the first hit point of the primary ray with the closest scene object,

classic Whitted-style ray tracing can—depending on the properties of the material of the concerned surface—generate up to three new types of rays: a reflection ray, a refraction ray, and a shadow ray. The algorithm then estimates the incoming light at the intersection point of the primary ray with the object, examines the material properties of the object, and combines this information to a light contribution for the final color of the pixel. The computation of the light contributions of the reflected as well as the refracted ray then takes place exactly in the same way as for the primary ray until a diffuse surface is hit, the ray does not intersect an object within the scene, or the intensity of the ray is below a threshold value, respectively, the recursive depth of the ray generation exceeds a predefined value. This is the reason why the method is also called *recursive ray tracing*, see Figure 8.9.

Reflection Ray (300)
Refraction Ray (305)
Shadow Ray (14)

Now, let us check which types of light transport paths in free space are simulated by classic Whitted-style ray tracing. As the most part of objects in a scene are assumed to be specular reflective or specular refractive, the BSDF f_s, involved in the SLTEV is supposed to be composed of a BRDF f_r, and a BTDF f_t. Then, the SLTEV simulated by classic Whitted-style ray tracing is of the form:

f_s (371)
f_t (330)
f_r (320)

$$\begin{aligned} L_o(\mathbf{s},\omega_o) &= L_e(\mathbf{s},\omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s},\omega_i \to \omega_o)\, L_i(\mathbf{s},\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) \qquad (8.20)\\ &= L_e(\mathbf{s},\omega_o) + \\ &\quad \int_{\mathcal{H}^2_i(\mathbf{s})} f_r(\mathbf{s},\omega_i \to \omega_o)\, L_i(\mathbf{s},\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) + \qquad (8.21)\\ &\quad \int_{\mathcal{H}^2_i(\mathbf{s})} f_t(\mathbf{s},\omega_i \to \omega_o)\, L_i(\mathbf{s},\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i). \end{aligned}$$

As the BRDF as well as the BTDF can be written as the sum of a specular, a diffuse, and a glossy component, the SLTEV can also be written as:

$$\begin{aligned} L_o(\mathbf{s},\omega_o) &\overset{(4.198)}{=} L_e(\mathbf{s},\omega_o) + \\ &\quad \int_{\mathcal{H}^2_i(\mathbf{s})} \left(f^o_r(\mathbf{s},\omega_i \to \omega_o) + f^{\vee}_r(\mathbf{s},\omega_i \to \omega_o) + f^{gl}_r(\mathbf{s},\omega_i \to \omega_o)\right) \\ &\qquad L_i(\mathbf{s},\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) + \qquad (8.22)\\ &\quad \int_{\mathcal{H}^2_i(\mathbf{s})} \left(f^o_t(\mathbf{s},\omega_i \to \omega_o) + f^{\vee}_t(\mathbf{s},\omega_i \to \omega_o) + f^{gl}_t(\mathbf{s},\omega_i \to \omega_o)\right) \\ &\qquad L_i(\mathbf{s},\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i). \end{aligned}$$

Now, except of the last interaction of light at an object surface, ray tracing only accounts for ideal specular reflection as well as ideal refraction, that is, the specular component of the BRDF and the BTDF can be replaced by Dirac δ-distributions in the mirrored direction $\omega_r = M_{\mathbf{N}}(\omega_o)$ as well as the refracted direction $\omega_t = R(\omega_o)$. Using the representations of the ideal specular BRDF from Equation (4.104) and the ideal transmitted

Dirac δ-distribution (117)

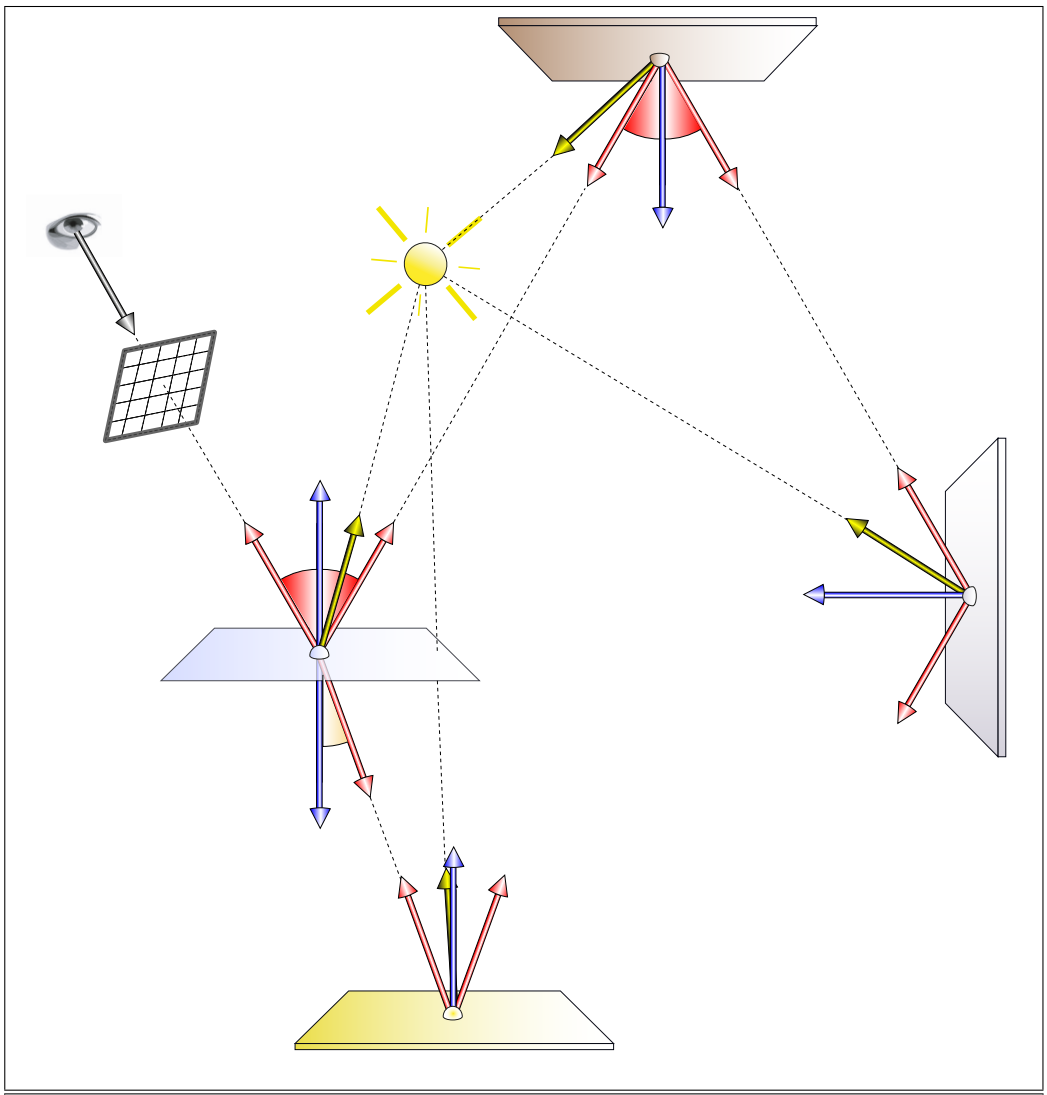


FIGURE 8.9: CLASSIC WHITTED-STYLE RAY TRACING. The algorithm starts with generating a primary ray from a sensor, typically the eye of an observer or a virtual camera, through a pixel of the image plane. At the first hit point of this ray with the closest scene object, the algorithm can generate, depending on the properties of the material of the concerned surface, up to three new types of rays: a reflection ray, a refraction ray, and a shadow ray. The algorithm estimates the incoming light at the intersection point of the primary ray with an object and combines this information to a contribution to the final color of the pixel. The computation of the light contributions of the reflected as well as the refracted ray are taken recursively until a diffuse surface is hit, the ray doesn't intersect an object within the scene, or the intensity of the ray is below a threshold value respectively the recursive depth of ray generation exceeds a predefined value.

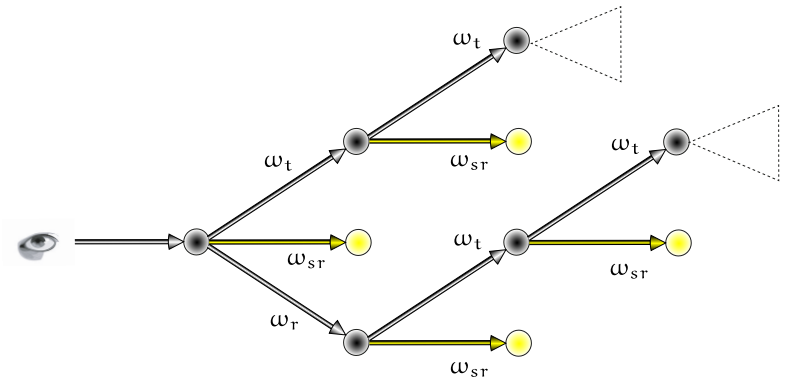


FIGURE 8.10: RAY TREE FOR LIGHT PATHS. Often, the ray tracing algorithm is represented in form of a ternary tree. The root of this tree represents the eye or a virtual camera and each inner node of the tree corresponds to a point where the surface is hit. At each such point, the algorithm then generates up to three new rays, a reflected and/or refracted ray, and a shadow ray.

BTDF from Equation (4.133) then implies:

$$\begin{aligned} L_o(\mathbf{s},\omega_o) \;=\; & L_e(\mathbf{s},\omega_o) + \\ & \rho_{dd}(\mathbf{s},\omega_r \to \omega_o)\, L_i(\mathbf{s},\omega_r) + \\ & \tau_{dd}(\mathbf{s},\omega_t \to \omega_o)\, L_i(\mathbf{s},\omega_t) + \\ & \int_{S^2(\mathbf{s})} \left(f_s^o(\mathbf{s},\omega_i \to \omega_o) + f_s^{gl}(\mathbf{s},\omega_i \to \omega_o)\right) L_i(\mathbf{s},\omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i), \end{aligned} \tag{8.23}$$

where ρ_{dd} is the directional-directional reflectance as defined in Equation (4.163) and τ_{dd} is the directional-directional transmittance from Equation (4.194). Note: The diffuse and glossy components of the BRDF and the BTDF in the above equation are expressed via the concept of the BSDF.

As in any ray tracing algorithm the recursive ray generation is stopped if a not too diffuse or a glossy surface is hit, the incident radiance in the scattering equation from Equation (8.23) is completely determined by the radiance emitted from sources illuminating the scene. That is, for the incident radiance it must hold: $L_i(\mathbf{s},\omega_i) = L_e(\gamma(\mathbf{s},\omega_i),-\omega_i)$, where $\gamma(\mathbf{s},\omega_i)$ is a point on any of the light sources. Obviously, this then implies that classic ray tracing can not compute indirect illumination on diffuse or glossy surfaces.

Scattering Equation (374)

γ (47)

Under these conditions, classic Whitted-style ray tracing only delivers an approximate

```
CLASSIC WHITTED-STYLE RAY TRACING {
    ∀ pixel □_i ∈ (□_1, ..., □_{s_x · s_y}) do {
        sample point p ∈ □_i
        generate an eye-ray r = e → p
        L(s → e) = TRACE(r)
}

TRACE(r) {
        compute hit point s of r with closest object ∂V in scene
        compute normal N(s) at point s
        return SHADE(s, N(s))
}

SHADE(s, N(s)) {
        L = 0
        ∀ light sources ☼_i ∈ (☼_1, ..., ☼_M) do {
            sample point l ∈ ☼_i
            if V(s ↔ l) = 1 {
                L+ = f_s(s, l → s → s') L_e(l → s) where s' = pred(s) in ray tree
            }
        }
        if ∂V is specular {
            generate secondary reflected and/or refracted ray r'
            L+ = TRACE(r')
        }
        return L
}
```

FIGURE 8.11: PSEUDOCODE FOR CLASSIC WHITTED-STYLE RAY TRACING. The classic Whitted-style ray tracing algorithm is only based on two simple methods: After generating primary rays through pixels on the image plane, the algorithm uses the method TRACE() recursively—for tracing primary rays through the scene to be rendered—and the method SHADE(), for coloring the associated pixels.

solution to the following model of the SLTEV:

$$\begin{aligned} L_o(s,\omega_o) &= L_e(s,\omega_o) + \\ &\int_{S^2(s)} \left(f_s^o(s,\omega_i \to \omega_o) + f_s^{gl}(s,\omega_i \to \omega_o)\right) L_e(\gamma(s,\omega_i),-\omega_i)\, d\sigma_s^{\perp}(\omega_i) + \\ &\rho_{dd}(s,\omega_r \to \omega_o)\, L_o(\gamma(s,\omega_r),-\omega_r) + \\ &\tau_{dd}(s,\omega_t \to \omega_o)\, L_o(\gamma(s,\omega_t),-\omega_t), \end{aligned} \tag{8.24}$$

where $\gamma(s,\omega_r)$ and $\gamma(s,\omega_t)$ are surface points in direction to the reflected as well as the refracted ray, and the incident radiance is expressed in terms of exitant radiance from emitters within the scene.

With the Relations (8.6) - (8.8), Equation (8.24) then simulates light transport paths of characteristic:

$$L_o = L_e + \sum_{i=1}^{\infty} \mathbf{T}_{\vee}^{\partial\mathcal{V}^i} L_e + \left(\mathbf{T}_o^{\partial\mathcal{V}} + \mathbf{T}_{gl}^{\partial\mathcal{V}}\right) L_e \tag{8.25}$$

$$= \left(\underbrace{\sum_{i=0}^{\infty} \mathbf{T}_{\vee}^{\partial\mathcal{V}^i}}_{ES^*} + \underbrace{\left(\mathbf{T}_o^{\partial\mathcal{V}} + \mathbf{T}_{gl}^{\partial\mathcal{V}}\right)}_{[D|G]} \right) L_e \tag{8.26}$$

$$\equiv \overrightarrow{ES^*[D|G]L}, \tag{8.27}$$

where the second term corresponds to the direct illumination at diffuse or glossy surfaces and the infinite sum represents the indirect illumination via ideal specular reflective as well as ideal refractive surfaces.

REMARK 8.3 *Obviously, classic Whitted-style ray tracing does not compute any indirect illumination than via specular paths, in particular, it can not compute indirect illumination via diffuse or glossy surfaces, nor caustics, which would be paths of characteristic* $\overrightarrow{EDS^+L}$. *The fraction of light coming directly from light sources and calculated by shadow rays is contained within the total energy as a local, diffuse-glossy component. All this means that ray tracing only computes a very coarse approximation of the original SLTEV.*

Caustic (658)

SLTEV (398)

Now, in classic Whitted-style ray tracing, all light sources in a scene are assumed to be point light sources. Therefore, let us assume that M point light sources $*_j \in \mathbb{R}^3$ with $1 \le j \le M$ illuminate the scene. Due to our derivation in Example 4.6, the factor $L_e(\gamma(s,\omega_i),-\omega_i)$ in Equation (8.24) is then replaced by the product of the Dirac δ-distribution $\delta(\omega_i - \omega_i^{*_j})$, where $\omega_i^{*_j} \equiv \frac{s \to *_j}{\|s \to *_j\|_2}$ is the shadow ray between surface point s and the light source $*_j$, and the irradiance at points s is given by:

Dirac δ-distribution (117)

Irradiance (257)

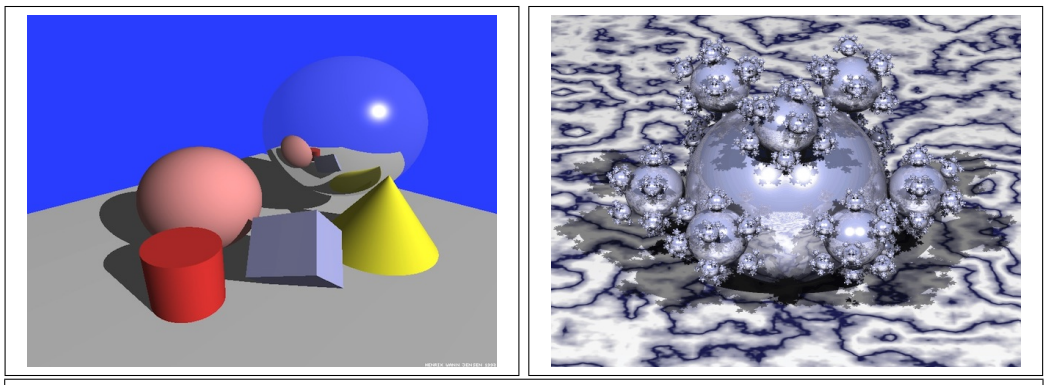


FIGURE 8.12: IMAGES RENDERED WITH CLASSIC WHITTED-STYLE RAY TRACING. The algorithm can render shadows and specular reflections as well as specular refractions, but it does not simulate indirect illumination of diffuse surfaces. Image courtesy of Henrik Wann Jensen, UCSD.

$$\mathbf{E}(\mathbf{s}) = \frac{\Phi_{e_j}(*_j)}{4\pi} \frac{\left|\cos\theta_i^{*_j}\right|}{\|*_j - \mathbf{s}\|_2^2}. \tag{8.28}$$

Using this result in Equation (8.24), then we get:

$$\begin{aligned} L_o(\mathbf{s},\omega_o) &= L_e(\mathbf{s},\omega_o) + \\ &\quad \rho_{dd}(\mathbf{s},\omega_r \rightarrow \omega_o)\, L_o(\gamma(\mathbf{s},\omega_r),-\omega_r) + \\ &\quad \rho_{dd}(\mathbf{s},\omega_t \rightarrow \omega_o)\, L_o(\gamma(\mathbf{s},\omega_t),-\omega_t) + \\ &\quad \int_{S^2(\mathbf{s})} \left(f_s^o(\mathbf{s},\omega_i \longrightarrow \omega_o) + f_s^{gl}(\mathbf{s},\omega_i \longrightarrow \omega_o)\right) \\ &\qquad \delta(\omega_i - \omega_i^{*_j})\, L_e(\gamma(\mathbf{s},\omega_i),-\omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i) \end{aligned} \tag{8.29}$$

$$\begin{aligned} &= L_e(\mathbf{s},\omega_o) + \\ &\quad \rho_{dd}(\mathbf{s},\omega_r \rightarrow \omega_o)\, L_o(\gamma(\mathbf{s},\omega_r),-\omega_r) + \\ &\quad \rho_{dd}(\mathbf{s},\omega_t \rightarrow \omega_o)\, L_o(\gamma(\mathbf{s},\omega_t),-\omega_t) + \\ &\quad \sum_{j=1}^{M} \left(f_s^o(\mathbf{s},\omega_i^{*_j} \longrightarrow \omega_o) + f_s^{gl}(\mathbf{s},\omega_i^{*_j} \longrightarrow \omega_o)\right) \frac{\Phi_{e_j}(*_j)}{4\pi} \frac{\left|\cos\theta_i^{*_j}\right|}{\|*_j - \mathbf{s}\|_2^2}. \end{aligned} \tag{8.30}$$

EXTENSIONS OF CLASSIC WHITTED-STYLE RAY TRACING. Due to the fact that classic ray tracing only accounts for point light sources and reflections and/or refractions at specular surfaces, naively implemented classic Whitted-style ray tracing generates only little realistic images. This is particularly obvious since the algorithms simulates only hard shadows and ideal reflections on surfaces of brilliant objects, see Figure 8.12.

The problem of hard shadows in classic Whitted-style ray tracing can easily be solved by using area light sources. For that purpose, let $\text{☼} = \{\text{☼}_1, \ldots, \text{☼}_M\}$ be M light sources existing in our scene. As the integral in Equation (8.24) is determined by the radiance emitted from light sources, instead to integrate over the entire unit sphere we integrate only over the set of surface areas of light sources in the scene. For that, we express Equation (8.24) in its 3-point form, where the integration domain is given by the area light sources, that is, we reformulate the spherical form of the SLTEV where we can write:

$$\begin{aligned} L(\mathbf{s} \to \mathbf{e}) &= L_e(\mathbf{s} \to \mathbf{e}) + \\ &\quad \rho_{dd}(\mathbf{s}_r \to \mathbf{s} \to \mathbf{e}) L(\mathbf{s}_r \to \mathbf{s}) + \\ &\quad \rho_{dd}(\mathbf{s}_t \to \mathbf{s} \to \mathbf{e}) L(\mathbf{s}_t \to \mathbf{s}) + \\ &\quad \int_{\text{☼}} \left(f_s^o(\mathbf{l} \to \mathbf{s} \to \mathbf{e}) + f_s^{gl}(\mathbf{l} \to \mathbf{s} \to \mathbf{e}) \right) L_e(\mathbf{l} \to \mathbf{s}) \mathcal{G}(\mathbf{l} \leftrightarrow \mathbf{s}) \, d\mu^2(\mathbf{l}), \end{aligned} \tag{8.31}$$

$$\begin{aligned} &= L_e(\mathbf{s} \to \mathbf{e}) + \\ &\quad \rho_{dd}(\mathbf{s}_r \to \mathbf{s} \to \mathbf{e}) L(\mathbf{s}_r \to \mathbf{s}) + \\ &\quad \rho_{dd}(\mathbf{s}_t \to \mathbf{s} \to \mathbf{e}) L(\mathbf{s}_t \to \mathbf{s}) + \\ &\quad \sum_{j=1}^{M} \int_{\text{☼}_j} \left(f_s^o(\mathbf{l} \to \mathbf{s} \to \mathbf{e}) + f_s^{gl}(\mathbf{l} \to \mathbf{s} \to \mathbf{e}) \right) L_e(\mathbf{l} \to \mathbf{s}) \mathcal{G}(\mathbf{l} \leftrightarrow \mathbf{s}) \, d\mu^2(\mathbf{l}), \end{aligned} \tag{8.32}$$

where $\mathbf{s}_r = \gamma(\mathbf{s}, \omega_r) \in \partial\mathcal{V}$, $\mathbf{s}_t = \gamma(\mathbf{s}, \omega_t) \in \partial\mathcal{V}$ and it holds: $\gamma(\mathbf{s}, \omega_o) = \mathbf{e}$, see Figure 8.13

The choice of a sample on a light source according to the probability density functions $p_{\text{☼}_i}$ over the probability spaces $(\text{☼}_i, \mathfrak{B}(\text{☼}_i), \mu^2)$ then leads to the following primary Monte Carlo estimator for approximating Formula (8.32):

PDF (176)

Probability Space (163)

$$\begin{aligned} F_1^{cWRT} &= L_e(\mathbf{s} \to \mathbf{e}) + \\ &\quad \rho_{dd}(\mathbf{s}_r \to \mathbf{s} \to \mathbf{e}) L(\mathbf{s}_r \to \mathbf{s}) + \rho_{dd}(\mathbf{s}_t \to \mathbf{s} \to \mathbf{e}) L(\mathbf{s}_t \to \mathbf{s}) + \\ &\quad \frac{1}{M} \sum_{j=1}^{M} \frac{\left(f_s^o(\mathbf{l}_j \to \mathbf{s} \to \mathbf{e}) + f_s^{gl}(\mathbf{l}_j \to \mathbf{s} \to \mathbf{e}) \right) L_e(\mathbf{l}_j \to \mathbf{s}) \, \mathcal{G}(\mathbf{l}_j \leftrightarrow \mathbf{s})}{p_{\text{☼}_j}(\mathbf{l}_j)}, \end{aligned} \tag{8.33}$$

where $\mathbf{l}_j$ are samples chosen at the light sources, and the evaluation of f_s^o, respectively f_s^{gl} is depending on the material of the specified object. This then solves the problem of hard shadows in the resulting images.

REMARK 8.4 *However, images rendered with ray tracing algorithms often appear artificial and unrealistic as the underlying procedures miss many important aspects of light, which limits the realism that could be normally achieved. In view to their authenticity, these effects can be improved by slight modifications in the basic ray tracing algorithm and the associated parameters. Examples for such modifications will be presented in the following section. So, we can, apart from soft shadows, also simulate depth of field, and motion blur effects. The undesired aliasing phenomenon*

Depth of Field (685)

Motion Blur (688)

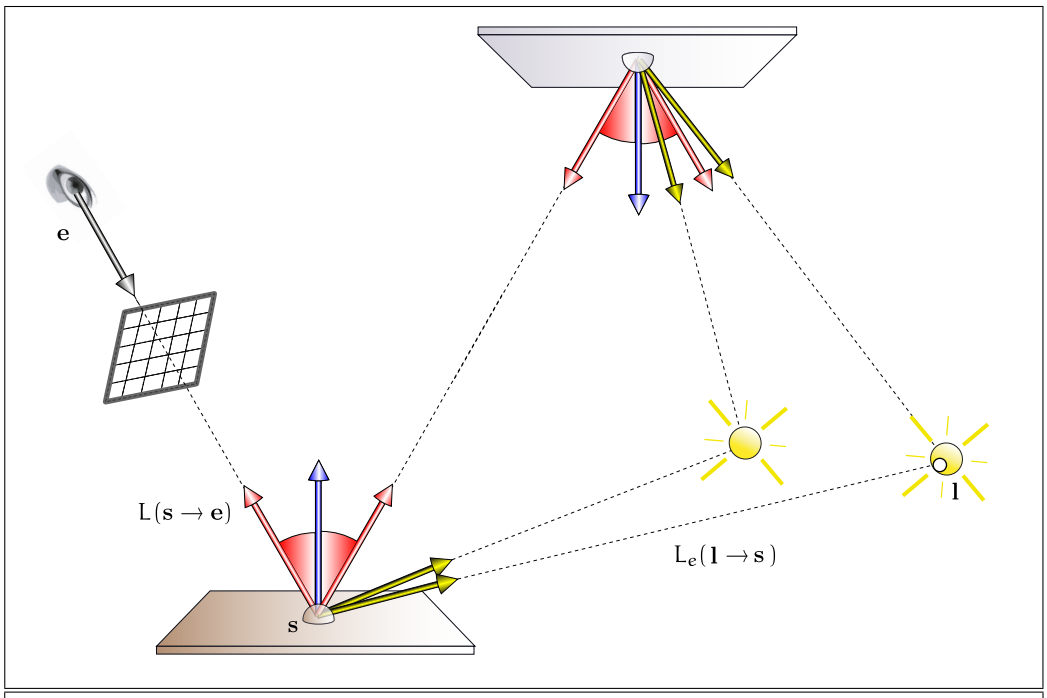


FIGURE 8.13: CLASSIC WHITTED-STYLE RAY TRACING USING AREA LIGHT SOURCES. The problem of hard shadows in classic Whitted-style ray tracing can easily be solved using area light sources. For that purpose, the algorithm generates at a hit point of a ray with any surface of the scene shadow rays in direction to the existing light sources for computing the direct illumination at the point to be shaded.

can also be adjusted by simple extension of the algorithm by firing more than a single ray into the scene and averaging the radiance which flows along the rays in order to shade the pixel.

8.4 DISTRIBUTION RAY TRACING

Classic Whitted-style ray tracing, as introduced in the last section, is not a full global illumination algorithm. Since the algorithm does not samples directions other than the perfectly specular reflected or the perfectly specular refracted directions, it cannot compute indirect illumination via diffuse or glossy surfaces. So, classic Whitted-style ray tracing cannot simulate all the interesting light effects, which occur at a point of interest, that is, the algorithm delivers only a coarse approximate solution to the stationary light transport equation in vacuum.

Glossy Reflection (304)
SLTEV (398)

In the present section we introduce a more powerful rendering algorithms: *Distribution Ray Tracing*. Distribution ray tracing is a global illumination algorithms, which can simulate all possible light effects that can occur within a scene. First off all, we illuminate the idea behind distribution ray tracing and show, how it can be applied to solve the stationary light transport equation within a vacuum. Afterwards, we knit around this method an algorithm that can be used to render a given scene: the *Classic Distribution Ray Tracing* algorithm. Finally, we extend the classic distribution ray tracing algorithm with some features to generate more realistic images, that is, we extend it by strategies for sampling more dimensions, such as the pixels, the lens of the involved camera system, and last, but not least, also the time.

Section 8.4.1
Section 8.4.2
Section 8.4.3

8.4.1 SOLVING THE SLTEV VIA DISTRIBUTING RAYS

Recall our first naive Monte Carlo rendering algorithm introduced in Example 6.44. It is based on the method of successive integral substitution from Section 6.7.1 for solving Fredholm integral equations of the 2^{nd} kind, and was applied to the SLTEV, given in the form:

SLTEV (398)

$$L_o(\mathbf{s}, \omega_o) = L_e(\mathbf{s}, \omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i). \tag{8.34}$$

This naive Monte Carlo ray tracing algorithm generates in a first step a large number of rays and shoots these rays, starting at point $\mathbf{s}$, into the scene. At the hit points of these primary rays with objects of the scene, the algorithm gathers the light that comes from these points, generates new rays, and shoots also these rays into the scene. Repeated application of this approach then results in a tree of paths with root at point $\mathbf{s}$ which can be used to compute the light arriving at $\mathbf{s}$ from points within the scene reachable via paths over scene objects. The entire process will be repeated again and again, until a ray does not hit an object or the recursion depth of the algorithm is exceeded, see Figure 8.14.

As shown in Example 6.44, our naive Monte Carlo algorithm uses the principle of invariance of radiance and expresses the incident radiance L_i in Equation (8.34) in terms of exitant radiance L_o. So, it delivers an approximate solution of the SLTEV expressed in terms of exitant radiance, namely:

Radiance Invariance (253)

$$\begin{aligned} & L_o(\mathbf{s}, \omega_o) \\ &= L_e(\mathbf{s}, \omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_i(\mathbf{s}, \omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i), && (8.35) \\ &= L_e(\mathbf{s}, \omega_o) + \int_{S^2(\mathbf{s})} f_s(\mathbf{s}, \omega_i \to \omega_o)\, L_o(\gamma(\mathbf{s}, \omega_i), -\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i). && (8.36) \end{aligned}$$

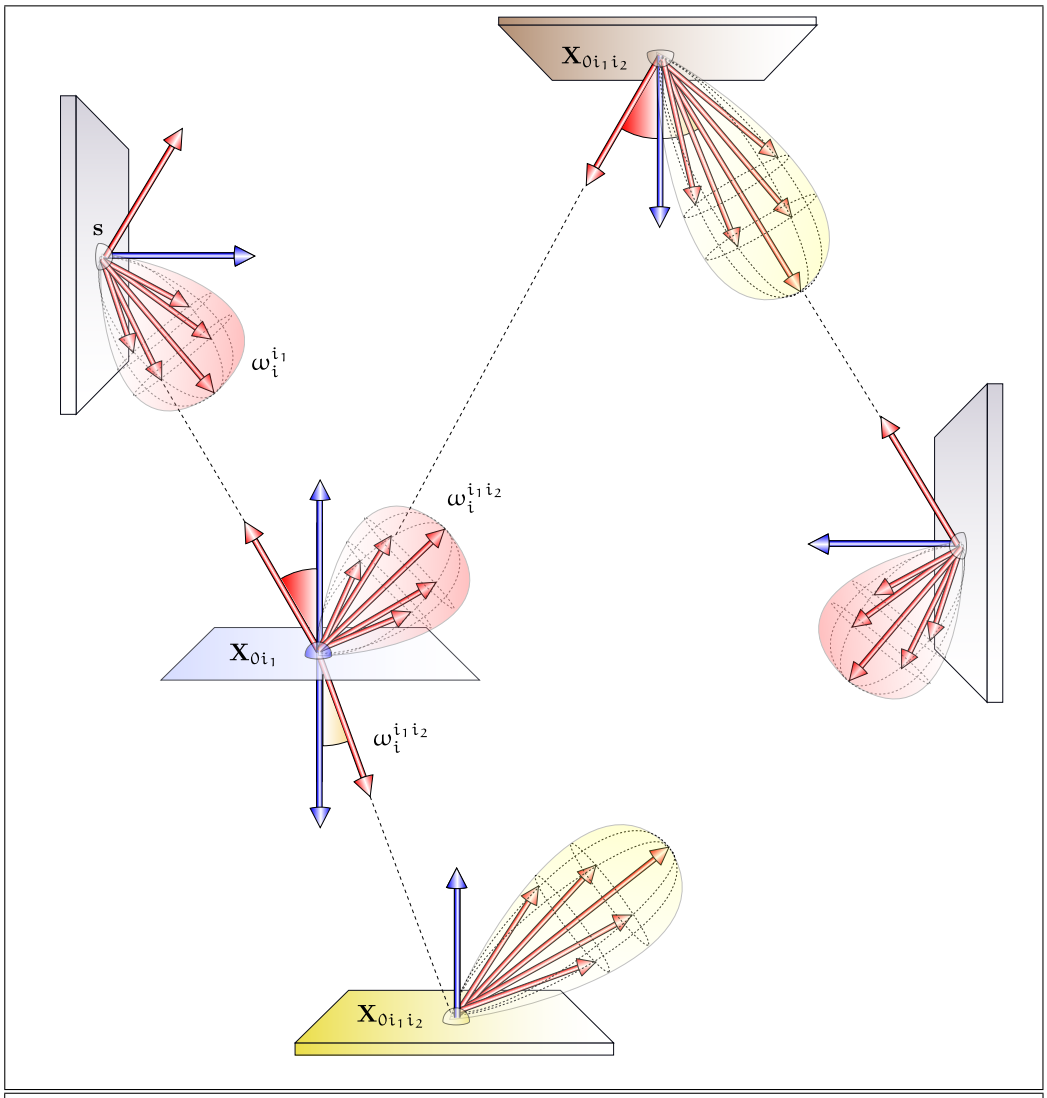


FIGURE 8.14: SOLVING THE SLTEV VIA DISTRIBUTING RAYS. Starting from point $\mathbf{s}$ the algorithm generates N_1 rays in directions $\boldsymbol{\omega}_i^{i_1}$. At the intersection points $\mathbf{X}_{0i_1}$ of these rays with object surfaces within the scene the algorithm gathers the light that comes from these points and generates N_2 new rays in directions $\boldsymbol{\omega}_i^{i_1i_2}$. These directions can intersect the scene objects in further points $\mathbf{X}_{0i_1i_2}$. Repeated application of this approach then results in a tree of paths with root at point $\mathbf{s}$, which can be explored to compute the light arriving at $\mathbf{s}$ from points on object surfaces of the scene. The entire process will be repeated again and again, until a ray does not hit an object or the recursion depth of the algorithm is exceeded. The exitant radiance at point $\mathbf{s}$ in direction ω_o is the result of incident light that flows along all paths, originated at $\mathbf{s}$, attenuated by reflection and/or refraction processes.

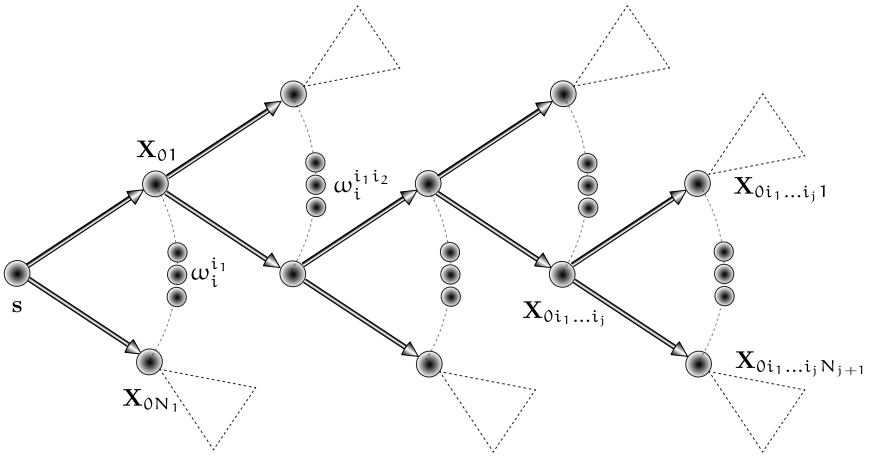


FIGURE 8.15: RAY DISTRIBUTION AT HIT POINTS WITHIN A SCENE. The algorithm can be visualized in form of a multi-branched tree. The root of this tree corresponds to the point $\mathbf{s}$ which we are interested in to compute the exitant radiance in direction ω_o, thus, $L_o(\mathbf{s}, \omega_o)$. From this point, N_1 rays are generated resulting in potentially new points $\mathbf{X}_{01}, \dots, \mathbf{X}_{0N_1}$ on object surfaces. At each such point, the algorithm can then generate many new rays. Note the way we labeled the nodes of the tree: At node $\mathbf{X}_{0i_1 \dots i_j}$, the algorithm generates N_{j+1} new rays to nodes $\mathbf{X}_{0i_1 \dots i_j 1}, \dots, \mathbf{X}_{0i_1 \dots i_j N_{J+1}}$.

The unknown exitant function L_o, that is, the outgoing radiance and the integral on the right-hand side can now be approximated by a Monte Carlo estimator—composed of the known emitted radiance L_e, the BSDF f_s, and the furthermore unknown integrand L_o, for a detailed description see Example 6.44 of Section 6.7.1. Using the identities from the Relations (6.580) and (6.586), a secondary Monte Carlo estimator $F_N^{DRT, L_o(\mathbf{s}, \omega_o)}$ for approximating the exitant radiance $L_o(\mathbf{s}, \omega_o)$ is then given by:

$$F_N^{DRT, L_o(\mathbf{s}, \omega_o)} = \frac{L_e(\mathbf{X}_0, \omega_o)}{p_0(\mathbf{X}_0)} + \sum_{l=1}^{M} \left\{ \sum_{i_1=1}^{N_1} \cdots \sum_{i_l=1}^{N_l} \left(\prod_{j=1}^{l} \frac{1}{N_j} \frac{f_s(\mathbf{X}_{0i_1 \dots i_{j-1}}, \boldsymbol{\omega}_i^{i_1 \dots i_j} \rightarrow \omega_o^{i_1 \dots i_{j-1}}) \left| \cos \boldsymbol{\omega}_i^{i_1 \dots i_j} \right|}{p_0(\mathbf{X}_0)\, p_j(\boldsymbol{\omega}_i^{i_1 \dots i_j} | \boldsymbol{\omega}_i^{i_1 \dots i_{j-1}})} \right) L_e(\mathbf{X}_{0i_1 \dots i_l}, \omega_o^{i_1 \dots i_l}) \right\}. \tag{8.37}$$

In the above equation, the number M corresponds to the maximal recursion depth of the algorithm, thus, the maximal recursive ray generation, N_j is the number of rays that are fired from sample points $\mathbf{X}_{0i_1 \dots i_{j-1}}$ in new directions $\boldsymbol{\omega}_i^{i_1 \dots i_j}$ and it holds: $p(\mathbf{X}_0) = 1$

as well as $p_1(\boldsymbol{\omega}_i^{i_1 i_1} | \boldsymbol{\omega}_i^{i_1 i_0}) = p_1(\boldsymbol{\omega}_i^{i_1})$.

REMARK 8.5 *Let us compare the estimator* $F_N^{DRT,L_o(\mathbf{s},\omega_o)}$ *with the estimator derived from classic Whitted-style ray tracing from Section 8.3: As the new algorithm always generates a large number of rays at an intersection point of a ray with a scene object, depending on the scattering properties of the surface, the estimator* $F_N^{DRT,L_o(\mathbf{s},\omega_o)}$ *can simulate much more light effects than the estimator corresponding to Whitted-style ray tracing. This is particularly evident when approximating of the glossy part of the BSDF via appropriate probability density functions* p_j*. Additionally, the new estimator can also take into account that at ideal specular reflective or refractive as well as at diffuse surfaces the corresponding components of the involved BSDFs has to be integrated. All these cases are already covered by* $F_N^{DRT,L_o(\mathbf{s},\omega_o)}$*. In the first case the BSDFs are* δ*-distributions and the associated terms in* $F_N^{DRT,L_o(\mathbf{s},\omega_o)}$ *must be replaced by the product of the directional-directional reflectance* ρ_{dd} *respectively the directional-directional transmittance* τ_{dd} *and the radiance. That is, the corresponding sums reduce to single terms in the estimator. For the case where a ray hits a diffuse surface, the estimator approximates the integration over the whole sphere at the hit point* $\mathbf{s}$ *by the choice of an associated probability density functions* p_j *according to the diffuse component of the BSDFs. Last, but not least,* $F_N^{DRT,L_o(\mathbf{s},\omega_o)}$ *can also solve the problem of hard shadows, known from classic Whitted-style ray tracing, by choosing PDFs proportional to the solid angles of the visible parts of the light sources in the scene.*

ρ_{dd} (338)
τ_{dd} (342)
Section 8.4.3

Using the Relations (8.6) - (8.8) then the SLTEV from Equation (8.34) corresponds to transport paths of characteristic

$$L_o \overset{(5.135)}{=} L_e + \sum_{i=1}^{\infty} \mathbf{T}_{L_o}^{\partial\mathcal{V}} L_e \tag{8.38}$$

$$\overset{(8.6)-(8.8)}{=} \left(\underbrace{\sum_{i=0}^{\infty} \left(\mathbf{T}_o^{\partial\mathcal{V}} + \mathbf{T}_{\vee}^{\partial\mathcal{V}} + \mathbf{T}_{gl}^{\partial\mathcal{V}} \right)}_{(D|S|G)^*} \right) L_e \tag{8.39}$$

$$\equiv \overrightarrow{E(D|S|G)^* L}, \tag{8.40}$$

that is, the algorithm which computes an approximate solution of the SLTEV based on distributing rays can generate paths of characteristic $\overrightarrow{E(D|G|S)^* L}$. Note, this algorithm can be used as the core of a full global illumination algorithm.

8.4.2 CLASSIC DISTRIBUTION RAY TRACING

Due to Definition 1.3, the global illumination problem consists in evaluating the measurement equation

$$\mathcal{M}_j \overset{(4.429)}{=} \int_{\partial\mathcal{V}} \int_{S^2} W_e^j(\mathbf{s},\omega)\, L_i(\mathbf{s},\omega)\, d\sigma_{\mathbf{s}}^{\perp}(\omega) d\mu^2(\mathbf{s}) \tag{8.41}$$

for all pixels $\square_j$ of the image plane, that is, a full global illumination rendering algorithm must solve the SLTEV at points within regions, which are visible through the pixel $\square_j$. Additionally, it must combine the solution of the SLTEV at these points with the emitted importance.

Due to the principle of radiance invariance in a vacuum, the incident radiance L_i from the measurement equation can then be written in terms of exitant radiance, namely as:

Radiance Invariance (253)

$$L_i(\mathbf{s},\omega) = L_o(\gamma(\mathbf{s},\omega),-\omega), \tag{8.42}$$

where $\gamma(\mathbf{s},\omega)$ are points within the scene visible from pixel $\square_j$.

As the exitant radiance $L_o(\gamma(\mathbf{s},\omega),-\omega)$ can be approximated via the secondary Monte Carlo estimator $F_N^{DRT,L_o(\gamma(\mathbf{s},\omega),-\omega)}$, for shading the pixel $\square_j$ we only have to estimate Equation (8.41).

Using a pinhole camera model, the measurement equation can obviously be written as:

Pinhole Camera (417)

$$\mathcal{M}_j \overset{(4.432)}{=} \int_{\square_j} f_j(\mathbf{s}) L_o(\gamma(\mathbf{s},\omega_e),-\omega_e)\langle \mathbf{N}(\mathbf{s}),\omega_e\rangle\, d\mu^2(\mathbf{s}), \tag{8.43}$$

with $\omega_e = \frac{\mathbf{e}\to\mathbf{s}}{\|\mathbf{e}\to\mathbf{s}\|_2^2}$, for details see Example 4.16.

A simple approach to approximate this type of measurement equation is then given via a primary Monte Carlo estimator $F_1^{\mathcal{M}_j,DRT}$ using a random variable $\mathbf{X}_0$ defined on the pixel area $\square_j$, thus:

$$F_1^{\mathcal{M}_j,DRT} = \frac{f_j(\mathbf{X}_0)}{p(\mathbf{X}_0)}\langle \mathbf{N}(\mathbf{X}_0),\omega_e\rangle\, F_N^{DRT,L_o(\gamma(\mathbf{X}_0,\omega_e),-\omega_e))}, \tag{8.44}$$

where $F_N^{DRT,L_o(\gamma(\mathbf{X}_0,\omega_e),-\omega_e)}$ is the secondary estimator for estimating the exitant radiance at sample point $\gamma(\mathbf{X}_0,\omega_e)$ in direction $-\omega_e$ from the previous section, see Figure 8.16.

This naive Monte Carlo rendering algorithm, firstly introduced in Example 6.44 and detailed discussed in the last section, can be seen as the basis of any ray tracing algorithm based on the distribution of rays. Extended by sampling a point on the pixel $\square_j$, and the construction of a primary ray, starting at the eye $\mathbf{e}$ and passing through the sample $\mathbf{X}_0$, it is basis of the *classic distribution ray tracing algorithm*, firstly presented in [40, Cook & al. 1984] under the name *distributed ray tracing* often also called *stochastic ray tracing*, see Figure 8.17. The pseudo-code of classic distribution ray tracing is shown in Figure 8.18.

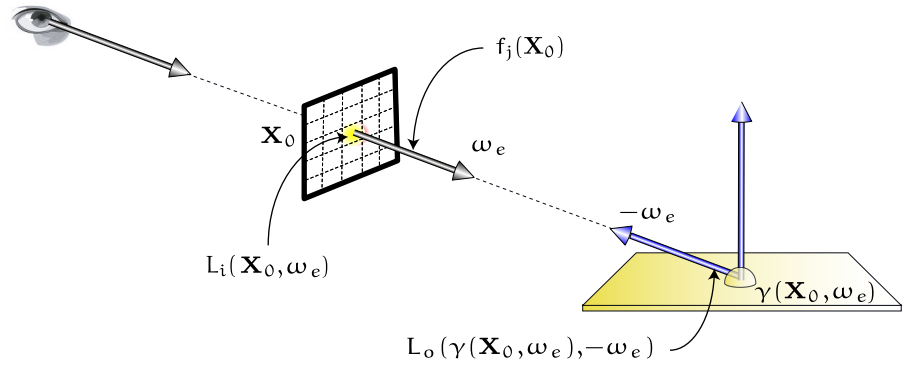


FIGURE 8.16: VISUALIZATION OF $F_1^{\mathcal{M}_j,\mathrm{DRT}}$. For estimating the measurement equation via a primary Monte Carlo estimator, first, a point $\mathbf{X}_0$ is sampled at a pixel. Due to the principle of radiance invariance in free space, the importance at $\mathbf{X}_0$ can easily be multiplied with the estimate of the incident radiance $L_i(\mathbf{X}_0,\omega_e)$, approximated via the reflected radiance $L_o(\gamma(\mathbf{X}_0,\omega_e),-\omega_e)$, computed with DRT.

REMARK 8.6 *The name distributed ray tracing is based on the fact that the algorithm distributes rays in a probabilistic way to sample quantities that produce effects like soft shadows, glossy reflections, and refractions, as well as depth of field and motion blur. In order to avoid confusion with distributed computing, the algorithm is named distribution ray tracing today. It was the first ray tracing algorithm that makes use of Monte Carlo techniques for solving the light transport equation in free space.*

Section 8.4.3

8.4.3 SAMPLING MORE DIMENSIONS: PIXELS, LENS AND TIME

Let us consider Figure 8.19, where we recognize effects such as blurred refection, soft shadows, penumbras, depth of field, and motion blur. As the radiance value, measured at a pixel on the image plane, is a function depending on time, pixel region, and lens optics, all these effects can be integrated as additional dimensions in the measurement equation. The measurement equation itself then mutates to a high-dimensional integral—one dimension for time, two dimensions for pixel area, respectively, lens aperture, and area light sources. Although this integral can be tremendously complicated, we can estimate it with the help of Monte Carlo techniques regardless how complicated it is.

Measurement Equation (416)

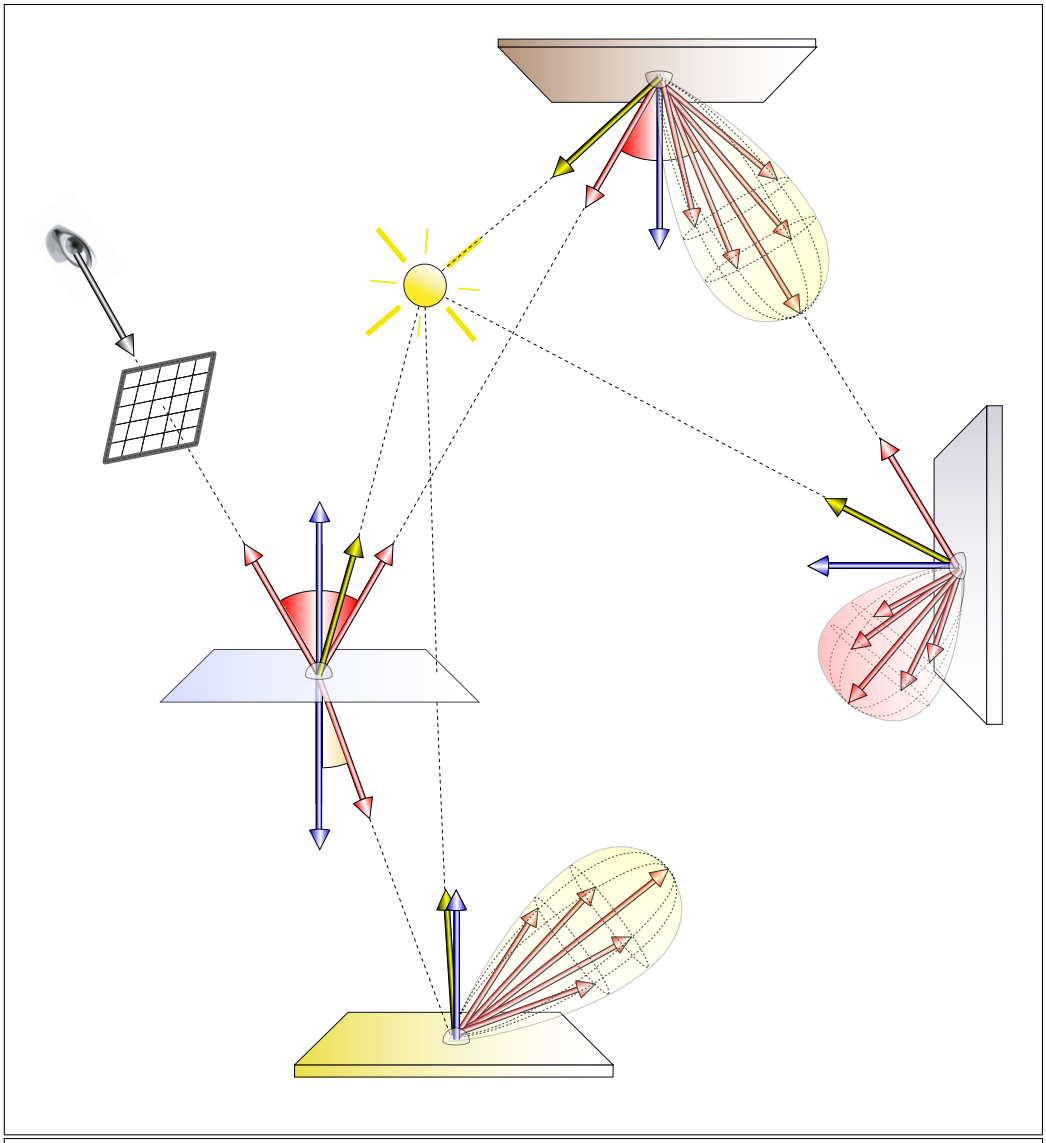


FIGURE 8.17: CLASSIC DISTRIBUTION RAY TRACING. The algorithm starts with generating a primary ray via the eye and a point sampled within a pixel $\square_j$. At the first hit point of this ray with the closest scene object, the algorithm can generate a large number of rays, depending on the properties of the material of the concerned surface. Additionally, a shadow ray a is also constructed. The algorithm estimates the incoming light at the intersection point of the primary ray with the object and combines this information to a contribution to the final color of the pixel. The computation of the light contributions of the distributed rays are taken recursively until the recursion depth of ray generation exceeds a predefined value.

```
CLASSIC DISTRIBUTION RAY TRACING {
    ∀ pixel □_i ∈ (□_1, ..., □_{s_x · s_y}) do {
        sample point X_0 ∈ □_i
        generate an eye-ray r = e → X_0
        L(s → e) = TRACE(r)
}

TRACE(r) {
        compute hit point s of r with closest object ∂V in scene
        compute normal N(s) at point s
        return SHADE(s, N(s))
}

SHADE(s, N(s)) {
        L = 0
        ∀ light sources ☼_i ∈ (☼_1, ..., ☼_M) do {
            sample point l ∈ ☼_i
            if V(s ↔ l) = 1 {
                L+ = f_s(s, l → s → s') L_e(l → s) where s' = pred(s) in ray tree
            }
        }
        if ∂V_c is specular {
            generate secondary reflected and/or refracted ray r'
            L+ = TRACE(r')
        } else {
            sample directions ω_i due to the reflection and/or
            refraction behavior of surface ∂V
            generate N secondary rays r_{2_i} = s + αω_i according to
            return TRACE(r_{2_i})
        }
        return L
}
```

FIGURE 8.18: PSEUDOCODE FOR CLASSIC DISTRIBUTION RAY TRACING. A coarse framework of distribution ray tracing consist of only three simple methods: one for generating and tracing of primary rays through pixels of the image plane, the method TRACE(), which is called for all primary rays, and the method SHADE() for coloring the corresponding pixels.

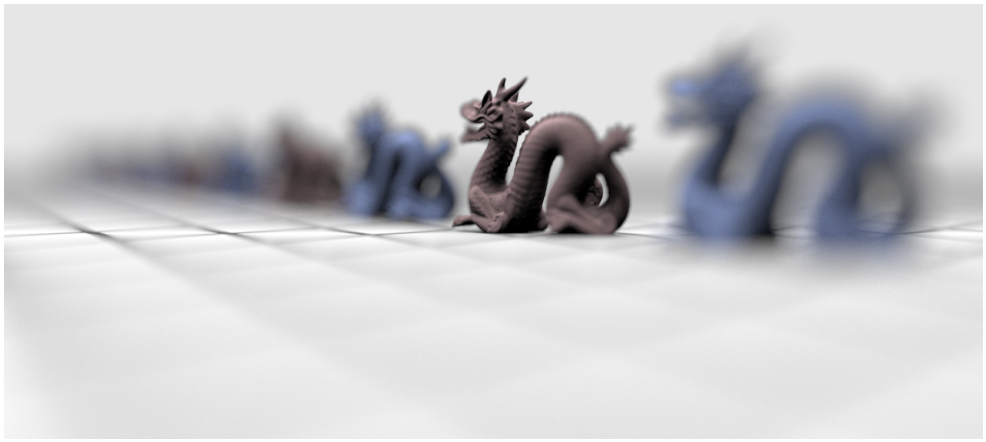


FIGURE 8.19: THE RENDERING OF FUZZY LIGHT PHENOMENA. Within the images, we recognize effects caused by fuzzy light phenomena, thus, the reflections of the billiard balls and the room are motion blurred, as are the penumbras. Image courtesy of Robert L. Cook, Thomas Porter and Loren Carpenter from LucasFilm.
The dragon picture illustrates the effect of depth of fields, where the camera is focused on the 2^{nd} dragon of the right. Images courtesy of Math Pharr and Greg Humphreys.

Recall, the goal of any rendering procedure based on principles of ray tracing is the
Flux (249) computation of the flux vector $(\mathcal{M}_1, \ldots, \mathcal{M}_n)$, where $\mathcal{M}_j$ is the solution of the measurement equation for pixel j, $1 \leq j \leq n$, i.e.

$$\mathcal{M}_j \quad \stackrel{def}{=} \quad \int_{\partial\mathcal{V}} \int_{S^2} W_e^j(s,\omega)\, L_i(s,\omega)\, d\sigma_s^{\perp}(\omega) d\mu^2(s). \tag{8.45}$$

Monte Carlo Estimator (499) A more advanced Monte Carlo estimation for computing $\mathcal{M}_j$ based on a distribution ray tracing strategy can now be performed by means of the following two step procedure:

i) Since the incident radiance in the measurement equation is unknown, we estimate $\mathcal{M}_j$
Monte Carlo Integration (477) by a straightforward Monte Carlo method, using a large number of random variables $\mathbf{X}_{0_k}$ and $\boldsymbol{\omega}_k$, and

ii) under the condition, that the samples $\mathbf{X}_{0_k}$ and $\boldsymbol{\omega}_k$ are chosen for estimating $\mathcal{M}_j$, we can start for every pair $(\mathbf{X}_{0_k}, \boldsymbol{\omega}_k)$ an estimation of $L_i(\mathbf{X}_{0_k}, \boldsymbol{\omega}_k)$, where we make
Radiance Invariance (253) use of the principle of radiance invariance in a vacuum and compute an estimation of $L_o(\gamma(\mathbf{X}_{0_k}, \boldsymbol{\omega}_k), -\omega_k)$ instead of $L_i(\mathbf{X}_{0_k}, \boldsymbol{\omega}_k)$ via distributing rays.

Mathematically, this idea can be converted as follows: Let us choose N identically
Probability Density Function (176) and independent, according to the probability density p distributed, random variables
Probability Space (163) $\mathbf{X}_{0_k}, \boldsymbol{\omega}_k$ from probability space $(\partial\mathcal{V} \times S^2, \mathfrak{B}(\partial\mathcal{V} \times S^2), \mathbb{P})$. A secondary Monte Carlo
Monte Carlo Estimator (499) estimator $F_N^{\mathcal{M}_j, DRT}$ for approximating the measurement equation is then given by:

$$F_N^{\mathcal{M}_j, DRT} \quad = \quad \frac{1}{N} \sum_{k=1}^{N} \frac{W_e^j(\mathbf{X}_{0_k}, \boldsymbol{\omega}_k) \langle \mathbf{N}(\mathbf{X}_{0_k}), \boldsymbol{\omega}_k \rangle}{p(\mathbf{X}_{0_k}, \boldsymbol{\omega}_k)}\, F_N^{DRT, L_o(\gamma(\mathbf{X}_{0_k}, \boldsymbol{\omega}_k), -\boldsymbol{\omega}_k)}, \tag{8.46}$$

where $F_N^{DRT, L_o(\gamma(\mathbf{X}_{0_k}, \boldsymbol{\omega}_k), -\boldsymbol{\omega}_k)}$ is the secondary Monte Carlo estimator for estimating $L_o(\gamma(\mathbf{X}_{0_k}, \boldsymbol{\omega}_k), -\boldsymbol{\omega}_k$ from the previous section, see also Figure 8.20.

8.4.3.1 PIXEL SAMPLING: ANTIALIASING

Computers are not only deterministic machines but also discrete devices that can only control displays with a finite number of pixels and a finite number of colors. Now, our rendering procedures are also of discrete nature and sample scenes only at a finite number of discrete points. So, the images produced are subject to aliasing, which is reflected as *jaggies* at the edges of objects, jagged highlights, or Moiré pattern, see Figure 8.21.

There are many different techniques to tackle this problem, but in most cases—see [62, Foley & al. 1987], [67, Glassner 1995], [55, Encarnacao & al. 1997], or [158, Pharr & Humphrys 2004]—the problem of aliasing cannot be avoided with limited frequencies and

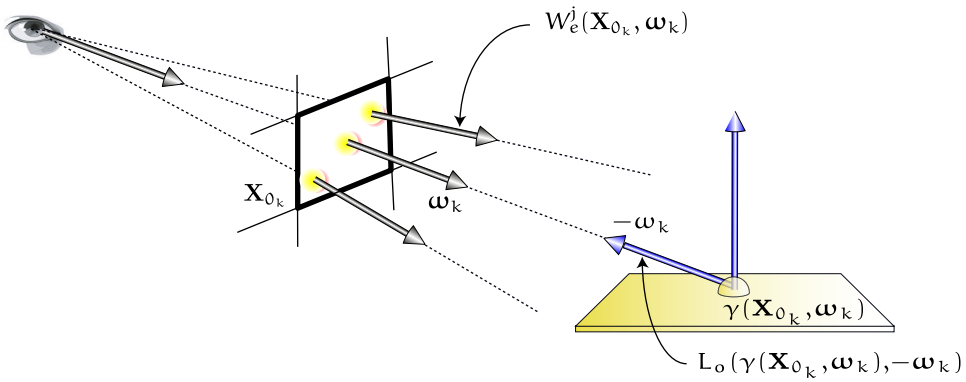
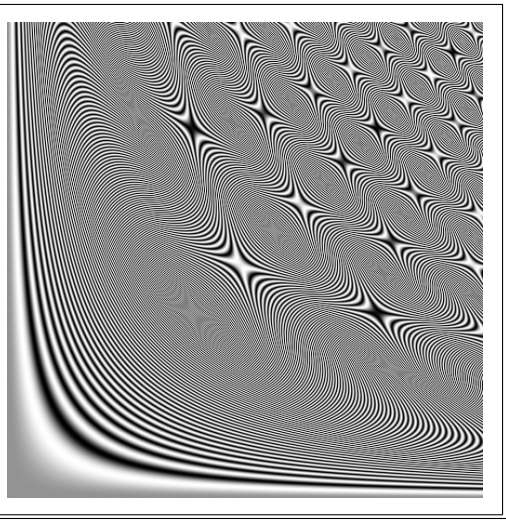


FIGURE 8.20: VISUALIZATION OF $F_N^{\mathcal{M}_j, \mathrm{DRT}}$. A secondary Monte Carlo estimator for the measurement equation averages—based on the principle of radiance invariance—the incident radiance, computed at $\gamma(\mathbf{X}_{0_k}, \omega_k)$ via DRT, with the exitant importance.



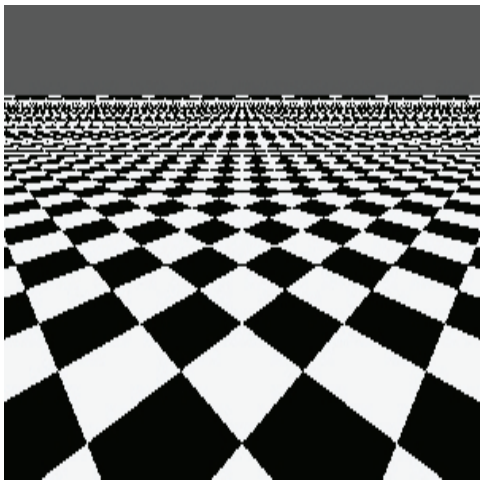


FIGURE 8.21: ALIASING EFFECTS. Left, the plot of the function $f(x, y) = \frac{1}{2}(1 + \sin(x^2 y^2))$, rendered in the range $[0, 10.83]^2$ at 512×512 pixels with one ray per pixel. The image is littered with so-called *Moiré pattern*. The checkerboard is also rendered with one ray per pixel. You can see how the checkers break up as they approach the horizon. Image Courtesy of Kevin Suffern, University of Technology, Sydney.

finite sampling. Indeed, aliasing artifacts can be reduced via the techniques of *supersampling* and *adaptive sampling*, but the problem can not be solved with the help of these techniques. A very efficient approach to reduce aliasing bases on stochastic principles: Convert it to less perceptually disturbing noise, [38, Cook 1986].

Now, it was empirically shown that better images can be achieved by interpreting a pixel as the average color of a sampled continuous region around the pixel center. With respect to the measurement equation $\mathcal{M}_j$ this means that we can define a *filter function* $f_j(\mathbf{s})$ by:

$$\int_{\square_j} f_j(\mathbf{s})\, d\mu^2(\mathbf{s}) = 1, \tag{8.47}$$

which serves to weight the incident radiance value at points $\mathbf{s}$ within the pixel $\square_j$.

This filter function is then embedded into the importance function W_e^j, that is, the *anti-aliasing measurement equation* has the form

$$\mathcal{M}_j^{AA} \stackrel{def}{=} \int_{\square_j}\int_{S^2} W_e^j(\mathbf{s},\omega)\, L_i(\mathbf{s},\omega)\, d\sigma_{\mathbf{s}}^{\perp}(\omega) d\mu^2(\mathbf{s}). \tag{8.48}$$

Due to its similarity with the measurement equation from Equation (8.45), an associated estimation is given by the estimator $F_N^{\mathcal{M}_j,DRT}$ from above, where the samples $(\mathbf{s}_k, \boldsymbol{\omega}_k)$ are drawn from $\square_j \times S^2$.

REMARK 8.7 *Sampling a continuous image function converts the image into a discrete set of values, one for each pixel. The goal of the weighting function W_e^j is the conversion in a such way that, combined with a reconstruction filter, it leads to the best possible reconstruction of the original image function. A major issue here is the exact shape of the reconstruction filter, which is often not or only roughly known at rendering time. This is because it includes all effects between display of the image and its perception by the user, such as blurring by a projector, the RGB, subpixel arrangement on the screen, etc..*

EXAMPLE 8.2 (Pixel Sampling with a Box Filter) *One of the most commonly used filters in CG is the box filter as introduced in Example 6.9, thus:*

$$f_j(\mathbf{s}) \stackrel{def}{=} \begin{cases} \frac{1}{\mu^2(\square_j)} & \text{if } \mathbf{s} \in \square_j \\ 0 & \text{otherwise.} \end{cases} \tag{8.49}$$

With a box filter all samples within a pixel are weighted by a constant, that is, the resulting pixel value is simply the average of the continuous image over the domain $\square_j$. Based on f_j, we then define the importance function W_e^j as:

$$W_e^j(\mathbf{s},\omega) \stackrel{(4.430)}{=} f_j(\mathbf{s})\delta(\omega - \omega_e), \tag{8.50}$$

where $\omega_e = \frac{\mathbf{e}\to\mathbf{s}}{\|\mathbf{e}\to\mathbf{s}\|_2^2}$ *is the direction of a ray* $\mathbf{r}$ *starting at the eye* $\mathbf{e}$ *through point* $\mathbf{s}$ *within pixel* $\square_j$*. As we have already shown in Equation (4.432), the measurement equation then reduces to the form*

$$\mathcal{M}_j^{AA} = \int_{\square_j} f_j(\mathbf{s})\, L_i(\mathbf{s}, \omega_e)\, |\cos\theta_{\mathbf{s}}|\, d\mu^2(\mathbf{s}). \tag{8.51}$$

Now, in Example 6.9 we have derived a secondary Monte Carlo estimator for trivial pixel filtering. With N *according to the probability density* $p(\mathbf{s}_k) = \frac{1}{\mu^2(\square_j)}$ *distributed, independent random variables* $\mathbf{s}_k$ *from probability space* $(\square_j, \mathfrak{B}(\square_j), \mathbb{P})$ *with* $\mathbb{P} = \mu^2$*, Equation (6.128) then implies the following estimator for a measurement* $\mathcal{M}_j^{AA}$*:*

$$F_N^{\mathcal{M}_j^{AA}} = \frac{1}{N}\sum_{k=1}^{N} \frac{f^j(\mathbf{s}_k)\langle \mathbf{N}(\mathbf{s}_k), \boldsymbol{\omega}_e\rangle}{p(\mathbf{s}_k)}\, F_N^{DRT, L_o(\gamma(\mathbf{s}_k, \omega_e), -\omega_e)} \tag{8.52}$$

$$\overset{p=f^j}{=} \frac{1}{N}\sum_{k=1}^{N} \langle \mathbf{N}(\mathbf{s}_k), \boldsymbol{\omega}_e\rangle\, F_N^{DRT, L_o(\gamma(\mathbf{s}_k, \omega_e), -\omega_e)}. \tag{8.53}$$

In Accordance with Equation (6.129) we then get:

$$F_N^{\mathcal{M}_j^{AA}} = \frac{1}{N}\sum_{k=1}^{N} \langle \mathbf{N}(\mathbf{s}_k), \boldsymbol{\omega}_k\rangle\, F_N^{L_i(\mathbf{s}_k, \omega_k)}. \tag{8.54}$$

As shown in Equation (8.53), the choice of a box filter as sampling strategy is good with respect to the implementation, but computationally, the choice of a box filter is not efficient, as it allows high-frequency sample data to leak into the reconstructed values [158, Pharr & Humphreys, 2004].

REMARK 8.8 *Advanced filter concepts, as shortly introduced in Remark 6.6, lead to less aliasing artifacts. Over and above that, variance reduction methods, such as LHS,* LHS (580) *promise considerably increase in convergence speed with respect to the computation of* $\mathcal{M}_j^{AA}$*, since the generation of random variables is independent on the dimension of the measurement equation.*

8.4.3.2 SAMPLING THE LENS OF A CAMERA: DEPTH OF FIELD

In all of our previous rendering algorithms, a virtual *pinhole camera system* is used as viewing device. Because the lens of a pinhole camera is infinitely small, every point on the image plane gets also light from only a single point within the scene. Due to the fact that the exposure of a point on the image plane is proportional to the light arriving from a single direction, images, which are made with a pinhole camera, are indeed perfectly sharp

Pinhole Camera System (417)

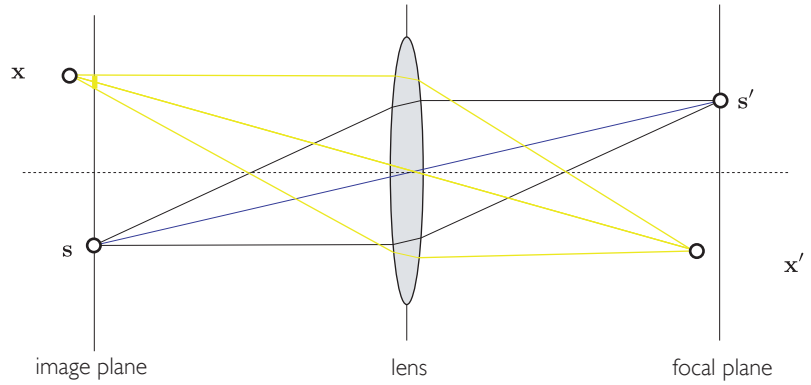


FIGURE 8.22: A THIN LENS CAMERA SYSTEM. Cross section through a thin lens camera system with focal and associated image plane. Rays starting at points on the focal plane intersect the image plane at the same point, thus they are in focus, while rays starting at points not lying on the focal plane intersect the image plane at different location. That is, they map the point on a region of the image plane, the circle of confusion. The above lens camera system has zero DOF.

but underexposed. To get really good pictures, the exposure of an image requires very long time. This is one of the reasons why pinhole cameras are rarely used in real world.

In real life, we usually use so-called *thin lens camera systems*.These are camera systems, similar to the pinhole camera model, but where the infinitely thin pinhole is replaced by a large aperture and a lens is put in it. This means, that a point on the image plane is not longer illuminated by a single light ray, but by a cone of light rays, see Figure 8.22.

Now, the use of a thin lens camera model instead of a pinhole camera has a big disadvantage: the generated images are not longer razor sharp. The reason for that is, that with a thin lens camera model only points lying on the focal plane are mapped to points on the image plane. All other points are mapped to small circles on the image plane, the so-called *circle of confusion*, whose size depends on the distance of the scene point to the focal plane and on the lens optics, see Figure 8.22. This then results in images, where only a central part of the image is in focus—we say also, this part is in *depth of field, DOF*—while the rest of the image appears very blurry. The depth of field of a camera is the range of distances parallel to the focal plane where the scene is in focus.

Now, DOF effects can be an unwanted artifact, or it can also be a desirable effect. In a rendering algorithm, DOF effects can easily be achieved by a simple modification in the above measurement equation.

For that, we change the integration domain $\square_j \times S^2$ of the measurement equation to $\square_j \times \Gamma_j$, where Γ_j is the solid angle subtended by the lens of the camera as seen from pixel

$\square_j$. With this modification, the measurement equation takes on the form

$$\mathcal{M}_j^{AA,DOF} \stackrel{def}{=} \int_{\square_j} \int_{\Gamma_j} W_e^j(\mathbf{s},\omega)\, L_i(\mathbf{s},\omega)\, d\sigma_{\mathbf{s}}^{\perp}(\omega) d\mu^2(\mathbf{s}). \tag{8.55}$$

Usually, a thin lens has the shape of circle. Because we do not want to limit to a circular form of a lens, the solid angle subtended by the lens as seen from a pixel can have any complicated form, that is, sampling from such a solid angle can be difficult. Therefore, we choose N identically and independent, according to a probability density $p_{\mathbf{X},\mathbf{l}}$ distributed, pairs of random variables $(\mathbf{X}_{0_k}, \mathbf{l}_{0_k})$ over the probability space $(\square_j \times \odot, \mathfrak{B}(\square_j \times \odot), \mathbb{P})$, where $\odot$ is the area of the lens, and $\mathbb{P}$ corresponds to the Lebesgue area measure $\mu^2 \times \mu^2$.

Probability Space (163)

Now, an estimator for approximating the measurement equation of pixel j based on the samples $(\mathbf{X}_{0_k}, \mathbf{l}_{0_k})$ obviously requires to express the direction samples $\boldsymbol{\omega}_k$—used for estimating the incident radiance $L_i(\mathbf{s},\omega)$—in terms of the samples $(\mathbf{X}_{0_k}, \mathbf{l}_{0_k})$. This can be done via a measure transform from the Lebesgue area measure to the solid angle measures, as described in Section 2.2.2.

Let $p_{\mathbf{X},\mathbf{l}}$ be the PDF which we sample from. Since $p_{\mathbf{X},\mathbf{l}}$ is separable, we can write

$$p_{\mathbf{X},\mathbf{l}}(\mathbf{X}_{0_k}, \mathbf{l}_{0_k}) = p_{\mathbf{X}}(\mathbf{X}_{0_k}) p_{\mathbf{l}}(\mathbf{l}_{0_k}) \tag{8.56}$$

Estimating the incident radiance within the measurement equation then requires that the PDF $p_{\mathbf{l}}(\mathbf{l}_{0_k})$ has to be replaced by a PDF in terms of a directional quantity. Let us denote this PDF as p_σ. To transform the spatial PDF $p_{\mathbf{l}}(\mathbf{l}_{0_k})$ into a directional PDF, we use the measure transformation from Equation (2.196) and get:

$$p_\sigma(\omega) \stackrel{(2.47)}{=} \frac{d\mathbb{P}(\omega)}{d\sigma(\omega)} \tag{8.57}$$

$$\stackrel{(2.196)}{=} \frac{d\mathbb{P}(\mathbf{l})}{d\mu^2(\mathbf{l})} \frac{\|\mathbf{s}-\mathbf{l}\|_2^2}{|\cos\theta|} \tag{8.58}$$

$$\stackrel{(2.47)}{=} p_{\mathbf{l}}(\mathbf{l}) \frac{\|\mathbf{s}-\mathbf{l}\|_2^2}{|\cos\theta|}. \tag{8.59}$$

That is, from the samples $(\mathbf{X}_{0_k}, \mathbf{l}_{0_k})$ we generate direction samples $\boldsymbol{\omega}_{0_k}$ by shooting a ray starting from pixel sample $\mathbf{X}_{0_k}$ through the lens sample $\mathbf{l}_{0_k}$ for $1 \le k \le N$. An associated Monte Carlo estimator for $\mathcal{M}_j^{AA,DOF}$ is then given by

$$F_N^{\mathcal{M}_j^{AA,DOF}} = \frac{1}{N} \sum_{k=1}^{N} \frac{W_e^j(\mathbf{X}_{0_k}, \boldsymbol{\omega}_{0_k}) \langle \mathbf{N}(\mathbf{X}_{0_k}), \boldsymbol{\omega}_{0_k} \rangle}{p_{\mathbf{X}}(\mathbf{X}_{0_k}) p_\sigma(\boldsymbol{\omega}_{0_k} | \mathbf{X}_{0_k})} F_N^{L_o(\gamma(\mathbf{X}_{0_k}, \boldsymbol{\omega}_{0_k}), -\boldsymbol{\omega}_{0_k})}, \tag{8.60}$$

where $F_N^{L_o(\gamma(\mathbf{X}_{0_k}, \boldsymbol{\omega}_{0_k}), -\boldsymbol{\omega}_{0_k})}$ is an estimation for the radiance incident at lens point $\mathbf{l}_{0_k}$ coming from sample point $\mathbf{X}_{0_k}$ from direction $-\boldsymbol{\omega}_{0_k}$, thus, the exitant radiance coming from the hit point, where the ray $\boldsymbol{\omega}_{0_k} = \frac{\mathbf{X}_{0_k} \to \mathbf{l}_{0_k}}{\|\mathbf{X}_{0_k} \to \mathbf{l}_{0_k}\|_2}$ intersects the focal plane.

REMARK 8.9 *As the direction samples $\boldsymbol{\omega}_{0_k}$ are constructed from area samples $\mathbf{X}_{0_k}$ on the pixel and samples $\mathbf{l}_{0_k}$ on the lens, a more better sampling strategy would be to transform the measurement equation into a surface integral, as we did it for sampling of shadow rays in Example 6.46. The integration domain is then given by the Cartesian product of $\partial\mathcal{V} \times \odot$, where $\odot$ corresponds to the area of the lens. This transformation is possible since W_e^j is unequal zero only for rays starting at the pixel and going through the lens.*

REMARK 8.10 *Depending on the type of lens used and the desired method for pixel filtering, a series of alternatives are available for the choice of W_e^j as well as the procedures for pixel and direction sampling which vary on the efficiency of the estimator and the accuracy of the result.*

REMARK 8.11 *Note, in rendering algorithms we assume, that the lens is infinitely thin, that is, rays passing the lens are not refracted.*

8.4.3.3 SAMPLING THE SHUTTER OPEN TIME: MOTION BLUR

Now, since physical sensors need a finite size and need to be integrated over a finite solid angle, they measure energy and not power. So, we need also to integrate over a finite period of time. This can result in *motion blur*. Motion blur appears in an image, when objects or the camera itself move during the exposure of an image. Particularly, this phenomenon occurs when the recorded objects are moving fast or when the exposure time is very long, see the upper image in Figure 8.19.

Motion blur effects can be simulated by extending the measurement equation by an additional integral over time. The associated measurement equation then has the form

$$\mathcal{M}_j^{AA,DOF,MB} \stackrel{def}{=} \int_T \int_{\square_j} \int_{\Gamma_j} W_e^j(\mathbf{x},\omega,t)\, L_i(\mathbf{x},\omega,t)\, d\sigma_{\mathbf{x}}^{\perp}(\omega) d\mu^2(\mathbf{x}) d\mu(t) \quad (8.61)$$

where the associated secondary Monte Carlo estimator is of the form:

$$F_N^{\mathcal{M}_j^{AA,DOF,MB}} = \frac{1}{N}\sum_{k=1}^{N} \frac{W_e^j(\mathbf{s}_k,\boldsymbol{\omega}_k,t_k)\langle \mathbf{N}(\mathbf{s}_k),\boldsymbol{\omega}_k\rangle}{p(\mathbf{s}_k,\mathbf{l}_k,t_k)} F_N^{L_o(\gamma(\mathbf{s}_k,\boldsymbol{\omega}_k),-\boldsymbol{\omega}_k,t_k)}. \quad (8.62)$$

In this formula, $F_N^{L_o(\gamma(\mathbf{s}_k,\boldsymbol{\omega}_k),-\boldsymbol{\omega}_k,t_k)}$ is an estimate for the radiance incident at lens point $\mathbf{l}_k$ coming at time t_k from direction $\boldsymbol{\omega}_k$, that is, the exitant radiance coming from hit point of $\boldsymbol{\omega}_k = \frac{(\mathbf{l}_k-\mathbf{s}_k)}{\|\mathbf{l}_k-\mathbf{s}_k\|_2^2}$ with the focal plane, and $(\mathbf{s}_k,\boldsymbol{\omega}_k,t_k)$ are independent, and identically distributed random variables drawn from $\square_j \times \odot \times T$ according to the PDF p, where it holds: $T = [t_0,t_1]$.

REMARK 8.12 *Instead of independent sampling each variable, in [185, Shirley 2000] it is suggested to choose a strategy that leads to a better stratification of the samples. Thus, the pixel, the lens, and the time should be stratified into the same number of strata, from where a triple of samples* $(s_{k_1}, \omega_{k_2}, t_{k_3})$, $1 \leq k_1, k_2, k_3 \leq n$ *should randomly be generated.*

REMARK 8.13 *By temporally as well as spatially distributing rays, we can easily simulate motion blur effects with distribution ray tracing. For that, we sample a time for a ray, move the objects accordingly before the ray is shot into the scene, and average the rays to compute the final radiance value. With this technique, the path of motion doesnt play a role, it can be arbitrarily complex, the only requirement is the ability to determine the position of an object at a specific time.*

THE PROBLEM OF EXPONENTIALLY INCREASING NUMBER OF RAYS. Distribution ray tracing is indeed a full global illumination algorithm, but it has a significant drawback: The exponentially increasing number of rays. Due to the fact that every ray underlies a series of processing steps, this leads to huge computation costs. So for example, at level l of the ray tree the algorithm shoots $\prod_{j=1}^{l} N_j$ rays into the scene, which corresponds to an enormous number of rays. As we also know from our discussion of the method of successive integral substitution for solving Fredholm integral equations of the 2^{nd}, nodes lying deep in the interior of the computation tree contribute, due to the contracting property of the kernels, only less to the final result. That is, the exponential effort for generating new rays leads to exponentially less contribution to the image.

8.5 REFERENCE LITERATURE AND FURTHER READING

Even though there are already a series of different notations for transport paths of light and importance, see [220, Veach 1997], [68, Glassner 1995], or [95, Jensen 2001], all based on Heckbert's path notation presented firstly in [81, Heckbert 1990], we have introduced a new notation for transport paths, that indicates the flow of stuff by an arrow in the direction of the flow. Specifically for capture transport paths in the photon mapping algorithm, we have also extended the alphabet, used to formulate corresponding transport paths, by two letters, D_G and S^G, describing the reflection at slightly glossy and high glossy surfaces. We expect from these two extensions more transparency with respect to the characteristic of a transport path.

Compared with ray tracing, the number of papers dealing with ray casting is comprehensible. The classical ray casting algorithm has its origin in [5, Apple 1968]. In most books about computer graphics, such as [62, Foley & al. 1987] and [78, Hearn & Backer

1994], ray casting is primarily presented as a visible detection tool. So, it is often ignored, that the idea behind ray casting is the real groundwork of all ray tracing based algorithms.

Ray tracing was firstly introduced in [236, Whitted 1980]. In contrast to ray casting, there is an enormous list of papers and textbooks dealing with ray tracing and extensions of the classic ray tracing algorithm. Here, we emphasize firstly the textbooks by [66, Glassner 1989] as well as [185, Shirley 2000], [187, Shirley & Morley 2003], and [205, Suffern 2007]. Glassner's book was the first to cover the whole research area of ray tracing from beginning to the nineties of the 20th century. Contrary to [185, Shirley 2000], [187, Shirley & Morley 2003], which also present results of the research until today in a more formal way—for example the stochastic methods of path tracing—Glassner describes the ray tracing procedure rather intuitively than formally. Similar to the books by Shirley, also the textbook by [205, Suffern 2007] is aimed at students that are interested in implementing a ray tracer.

Apart from these books, that deal almost exclusively with ray tracing, we suggest the reading of [67, Glassner 1995], [68, Glassner 1995], [95, Jensen 2001], [50, Dutré & al. 2003], [51, Dutré & al. 2006], and [158, Pharr & Humphreys 2004], [159, Pharr & Humphreys 2010]. Except for the typos, Glassner's two volume work is a brilliant reference for the study of ray tracing. He discusses nearly all interesting problems, from sampling theory via radiometry to integral equations and their solution methods. [95, Jensen 2001] and [50, Dutré 2003], [51, Dutré & al. 2006] are rather tailored to particular techniques for solving the global illumination equations. While Jensen presents its photon mapping algorithm, in [50, Dutré 2003], [51, Dutré & al. 2006] the Monte Carlo based methods of path- and light tracing are discussed in depth. [158, Pharr & Humphreys 2004], [159, Pharr & Humphreys 2010] is a wonderful work, since, as Per Christensen says, it covers all the marvelous math, fascinating physics, practical software engineering, and last not but least an enormous set of clever tricks that are necessary to write state-of-the-art photorealistic renderer. It is a long literate program, that is, reading this book means reading the full implementation of the pbrt rendering system, not just a high level description. Therefore, we necessarily recommended this book due their detailed implementations of the techniques which form part of ray tracing and, in particular, the interesting sources they provide for application programmers.

There is also a series of beautiful textbooks on computer graphics, which discuss ray tracing shortly and on a high level. Here we allude [62, Foley & al. 1987], [233, Watt & Watt 1992], [78, Hearn & Baker 1994], [232, Watt 1999], and [186, Shirley 2002].

The classic distribution ray tracing algorithm was introduced in [40, Cook & al. 1984] under the name distributed ray tracing also called stochastic ray tracing. Similar to ray casting, even distribution ray tracing, the mother of all stochastic ray tracing algorithms, is rather seldom discussed in literature. The reason for that is obviously the very restricted applicability of distribution ray tracing in practice. Thus, only in introductory textbooks, such as [62, Foley & al. 1987], [233, Watt & Watt 1992], [78, Hearn & Baker 1994], [232, Watt 1999], and [186, Shirley 2002], are devoted a few pages to distribution ray tracing.

CHAPTER NINE

MARKOV PROCESS BASED RENDERING ALGORITHMS

In the last chapter, we have presented the classic rendering algorithms based on the principle of ray tracing. With distribution ray tracing we have met a first full global illumination algorithm that makes use of concepts from probability theory. We have also seen that this technique indeed solves the global illumination problem, but in practice it is not fully usable due to the problem of exponentially increasing number of rays.

In this chapter, we now present advanced modern rendering algorithms, simple variants of distribution ray tracing in some sense, that are based on the stochastic concept of the discrete Markov process. They all make use of the property that the expected value of a family of random variables can be approximated by the expected value of a large number of random walks, so-called Markov processes, as introduced in Section 2.4.7.2. This technique then solves the problem of exponentially increasing number of rays and makes the algorithms well usable in practice.

OVERVIEW OF THIS CHAPTER. We begin with *Monte Carlo Path Tracing*, a very general and powerful method for solving the stationary light transport equation in free space. It is the standard global illumination algorithm used in computer graphics, which solves the global illumination problem by generating a large set of random walks starting at the eye. Afterwards, we present *Monte Carlo Light Tracing*. As dual to path tracing, Monte Carlo light tracing solves the global illumination problem via the construction of light paths that start at the light sources in a scene. We also discuss and analyze *Bidirectional Path Tracing*, a technique that combines path and light tracing, up to date, one of the most powerful methods for generating photorealistic images. With the *Metropolis Light Transport* algorithm, based on a Markov chain Monte Carlo approach, we then present a very powerful rendering algorithm for difficult sampling problems in high-dimensional spaces. An other nowadays widely used rendering technique is the *Photon-mapping Concept*. It belongs to a class of highly efficient two-pass algorithm that can be used to solve the problem that comes often with path algorithms when rendering specular surfaces in scenes with luminaries that are not large in area. We conclude this chapter with a short

Section 9.1
Section 9.2
Section 9.3
Section 9.4
Section 9.5

overview of *Instant Global Illumination*, a rendering algorithm that allows to simulate the most important illumination effects at realtime rates. Section 9.6

9.1 MONTE CARLO PATH TRACING

Distribution ray tracing, as introduced in the last section of the previous chapter, is indeed a full global illumination algorithm. But DRT has a significant drawback: The algorithm needs, coupled to the recursion depth, an exponentially increasing number of rays. So, DRT shoots during the l^{th} recursion step $\prod_{k=1}^{l} N_k$ rays into the scene, which results in a very large number of rays and an enormous amount of computational costs. Since all these rays, in particular, if they are generated deep in the ray tree, does not contribute much to the shading of a pixel of the final image, the strategy to generate many new rays at a hit point of a ray with a scene object is not a good idea. But, this obviously drawback of distribution ray tracing can easily be corrected with the help of an elegant probability theoretical model: a discrete-time Markov process. The underlying algorithm is Jim Kajiya's *Monte Carlo Path Tracing*, [98, Kajiya 1986].

DT Markov Process (236)

We begin this section with the classic Monte Carlo ray tracing algorithm: *pure-Monte Carlo Path Tracing*. It can be considered as a special variant of DRT, where, instead of many new rays, only a single ray is traced through the scene. This simplification then reflects in the quality of the generated images. So, images generated with pure-Monte Carlo path tracing are very noisy, which makes pure-Monte Carlo path tracing not really usable in practice. Therefore, we will extend this algorithms by a technique, which helps to decrease the noise in the rendered images: *Monte Carlo Path Tracing with Next Event Estimation*. One can say that this algorithm is *the standard rendering procedure* for solving the global illumination problem.

Section 9.1.1

Section 9.1.2

9.1.1 PURE-MONTE CARLO PATH TRACING

Let us consider the ray tree from Figure 9.1 as it is typically constructed by distribution ray tracing. Instead to generate N_j new rays at each inner node of tree level $k, k \geq 1$, we can also generate only a single ray. That is, on each level k, we choose $N_k = 1$ and trace this single ray on its travel through the scene. Then you can see, that the original ray tree shrinks to a path, a so-called *random walk*, whose state set is given by the objects surfaces of the scene that are visited by the path, see Figure 9.2.

Random Walk (233)

State Set (219)

PURE-MONTE CARLO PATH TRACING. The construction of a random walk, generated over the hit points of a ray with objects during its travel through the scene, can be simulated via the probability theoretical model of the discrete-time Markov process. Let $\overline{\mathbf{X}} = \mathbf{X}_0\mathbf{X}_1 \ldots \mathbf{X}_M$ be a random walk starting at the eye of an observer or at a virtual

DT Markov Process (236)

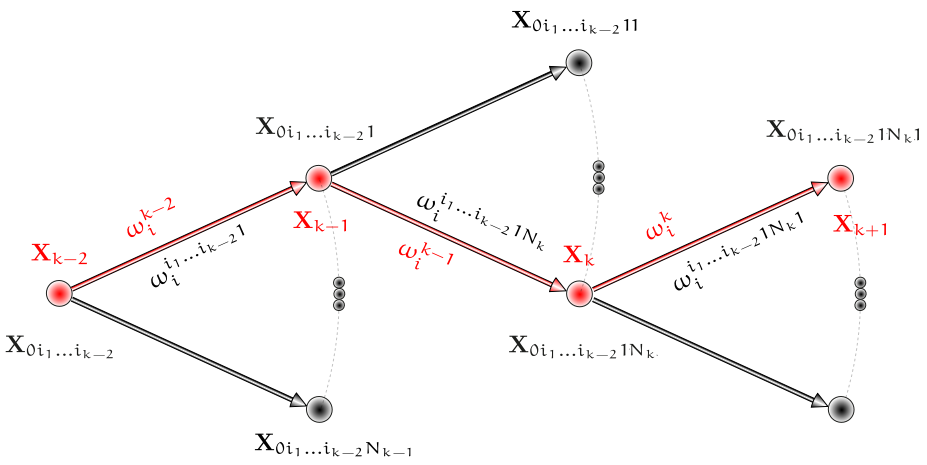


FIGURE 9.1: PATH NOTATION IN PURE-MONTE CARLO PATH TRACING. Compared to distribution ray tracing, pure-Monte Carlo path tracing generates at each level of the ray tree only a single ray and traces this ray on its travel through the scene. The original ray tree then shrinks to a random walk over the objects of the scene. In the above figure, a part of a random walk is shown whose nodes are marked in red. The starting point of a path always corresponds to the node $\mathbf{X} = \mathbf{X}_0$, and the successor of node $\mathbf{X}_{k-2}$ is $\mathbf{X}_{k-1}$ for $k \geq 2$. The incident direction at node $\mathbf{X}_{k-1}$ is ω_i^{k-1}, the corresponding outgoing direction towards $\mathbf{X}_{k-2}$ is denoted as ω_o^{k-1}.

camera that was generated via a Markov process. Let us furthermore assume, that the process stops if the path $\overline{\mathbf{X}}$ goes over a light source, a ray does not hit any object, or if the travel of the ray through the scene is stopped via Russian roulette, respectively, if the length of the path exceeds a predetermined default value. Under the assumption that the path ends at a light source, then we can gather the amount of light coming from the source, and can give back this light contribution over the path to the sensor, where the path comes from. Obviously, this algorithm only contributes for shading a pixel if the generated random walk ends at a light source, that is, if the node $\mathbf{X}_M$ is chosen at a light source. Since all other light contributions—whether implied by direct or indirect illumination—at other nodes of the path are neglected, we call this rendering technique, *pure-Monte Carlo path tracing*.

Russian Roulette (200)

Direct Illumination (410)

Indirect Illumination (410)

Pure-Monte Carlo path tracing, abbreviated *pMCPT*, often also simply denoted as *path tracing* or *Monte Carlo ray tracing*, was firstly presented in [98, Kajiya 1986]. As already mentioned above, we can interpret this algorithm as a discrete-time Markov process solution of the stationary vacuum light transport equation. This stochastic process can be simulated via a random walk whose associated random variable is defined as the sum of random variables given on the measurable spaces $(S^2, \mathfrak{B}(S^2))$, if a direction has

DT Markov Process (236)

Section 6.7.3

Measurable Space (80)

to be sampled, or $(\sphericalangle, \mathfrak{B}(\sphericalangle))$ respectively $(\square, \mathfrak{B}(\square))$, if point sampling is done on a light source or a pixel, for a detailed discussion on continuous random walks see Section 6.7.3. The pseudocode for pure-Monte Carlo path tracing is shown in Figure 9.3.

A Monte Carlo estimator that computes an approximate solution to the SLTEV based on pMCPT can now be derived from the secondary estimator $F_N^{DRT,L_o(s_j,\omega_o)}$ from Equation (8.37) by a few simple modifications: First, we set the number of generated rays $N_k = 1$ for $k \geq 1$. This transforms the secondary estimator $F_N^{DRT,L_o(s_j,\omega_o^j)}$ for $L_o(s_j,\omega_o^j)$ to the primary estimator $F_1^{pMCPT,L_o(s_j,\omega_o^j)}$ of the following form:

$$F_1^{pMCPT,L_o(s_j,\omega_o^j)} = \frac{L_e(\mathbf{X}_0,\omega_o)}{p_0(\mathbf{X}_0)} + \sum_{l=1}^{M}\left(\prod_{k=1}^{l}\frac{f_s(\mathbf{X}_{0i_1\ldots i_{k-1}},\boldsymbol{\omega}_i^{i_1\ldots i_k}\rightarrow\omega_o^{i_1\ldots i_{k-1}})\left|\cos\boldsymbol{\omega}_i^{i_1\ldots i_k}\right|}{p_0(\mathbf{X}_0)\,p_k(\boldsymbol{\omega}_i^{i_1\ldots i_k}|\boldsymbol{\omega}_i^{i_1\ldots i_{k-1}})}\right) L_e(\mathbf{X}_{0i_1\ldots i_l},\omega_o^{i_1\ldots i_l}). \tag{9.1}$$

Choosing $\mathbf{X}_0$ as starting point s_j and renaming the path nodes $\mathbf{X}_{0i_1\ldots i_{k-1}}$ into $\mathbf{X}_{k-1}$, the directions $\boldsymbol{\omega}_i^{i_1\ldots i_k}$ and $\omega_o^{i_1\ldots i_{k-1}}$ into $\boldsymbol{\omega}_i^{k-1}$, respectively, $\boldsymbol{\omega}_o^{k-1}$ then leads to a more simpler expression for the estimator $F_1^{pMCPT,L_o(s_j,\omega_o^j)}$, namely:

$$\begin{aligned} F_1^{pMCPT,L_o(s_j,\omega_o^j)} &= \frac{L_e(\mathbf{X}_0,\omega_o)}{p_0(\mathbf{X}_0)} + \sum_{l=1}^{M}\left(\prod_{k=1}^{l}\frac{f_s(\mathbf{X}_{k-1},\boldsymbol{\omega}_i^{k-1}\rightarrow\omega_o^{k-1})\left|\cos\boldsymbol{\omega}_i^{k-1}\right|}{p_k(\boldsymbol{\omega}_i^{k-1}|\boldsymbol{\omega}_i^{k-2})}\right) L_e(\mathbf{X}_l,\omega_o^l) && (9.2)\\ &= \frac{L_e(\mathbf{X}_0,\omega_o)}{p_0(\mathbf{X}_0)} + \sum_{l=0}^{M-1}\left(\prod_{k=0}^{l}\frac{f_s(\mathbf{X}_k,\boldsymbol{\omega}_i^{k}\rightarrow\omega_o^{k})\left|\cos\boldsymbol{\omega}_i^{k}\right|}{p_k(\boldsymbol{\omega}_i^{k}|\boldsymbol{\omega}_i^{k-1})}\right) L_e(\mathbf{X}_{l+1},\omega_o^{l+1}) && (9.3)\end{aligned}$$

with $p_0(\boldsymbol{\omega}_i^0|\boldsymbol{\omega}_i^{-1}) = p_0(\boldsymbol{\omega}_i^0)$ and $p_0(\mathbf{X}_0) = 1$.

REMARK 9.1 *Let us consider the random variable from Equation (6.603) associated with the random path* $\overline{\mathbf{X}} = \mathbf{X}_0\mathbf{X}_1\ldots\mathbf{X}_M$ *for estimating the Neumann series associated with a Fredholm integral equation of the* 2^{nd} *kind at sample* $\mathbf{X}_0$*, obviously it holds:*

$$\mathbf{Y}_M \overset{def}{=} \frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} + \sum_{l=1}^{M}\left(\frac{k(\mathbf{X}_0,\mathbf{X}_1)}{p_0(\mathbf{X}_0)\,p(\mathbf{X}_1|\mathbf{X}_0)}\prod_{k=1}^{l-1}\frac{k(\mathbf{X}_k,\mathbf{X}_{k+1})}{p(\mathbf{X}_{k+1}|\mathbf{X}_k)}\right) g(\mathbf{X}_l) \tag{9.4}$$

$$\overset{p_0(\mathbf{X}_0)=1}{=} \frac{g(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} + \sum_{l=0}^{M-1}\left(\prod_{k=0}^{l}\frac{k(\mathbf{X}_k,\mathbf{X}_{k+1})}{p(\mathbf{X}_{k+1}|\mathbf{X}_k)}\right) g(\mathbf{X}_{l+1}) \tag{9.5}$$

Now, replacing the source term g *by the emitted radiance* L_e *at the point* $\mathbf{X}_0$*—sampled according to* p_0 *on a light source—in direction* ω_o *and the transition kernel*

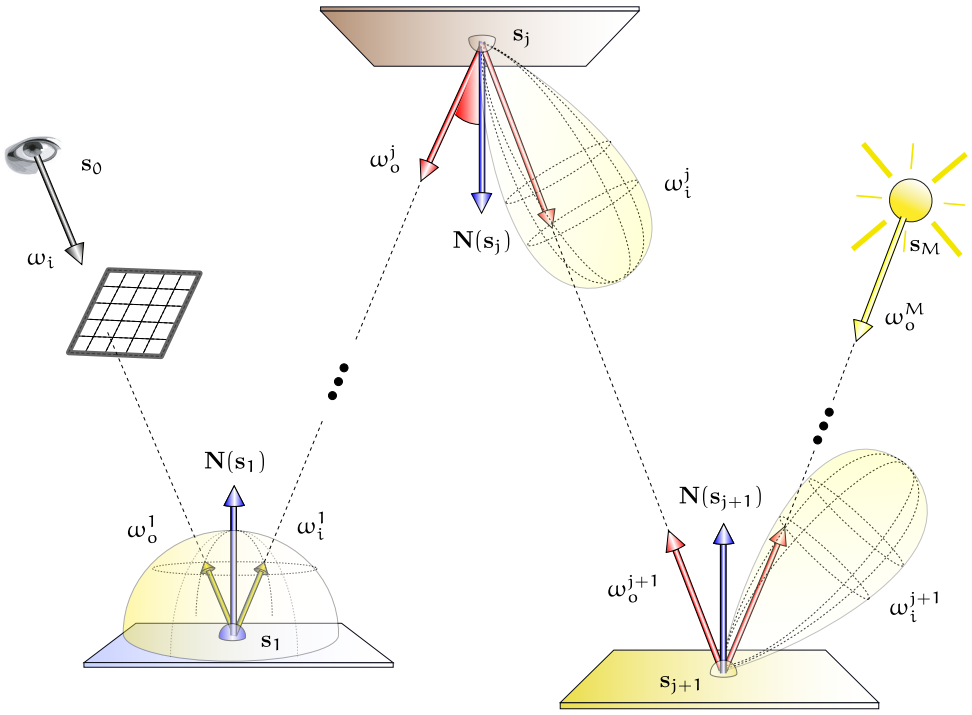


FIGURE 9.2: PURE-MONTE CARLO PATH TRACING. Starting from point s_0, usually the eye of an observer or a virtual camera, pure-Monte Carlo path tracing generates only one single ray and traces this ray trough the scene to be rendered. At the first intersection point, s_1, of this ray with an object in the scene, pMCPT also generates only a single new ray in direction ω_i^1 over the unit or the hemisphere about s_1 depending on the material and the surface properties of the object that has been hit. The algorithm then repeats this step recursively until a ray hits a light source in the scene, a ray does not hit any object, or if the travel of the ray through the scene is stopped via Russian roulette, respectively, if the length of the path exceeds a predetermined default value. The incident radiance at point s_{j+1} from direction ω_i^{j+1} then corresponds to the light emitted from a light source, that has been hit by the path. As the scattering behavior of an object in a scene is usually not ideal, the energy of light at the interaction with a surface is attenuated depending on the material and the surface properties of the object. So, under certain circumstances, only a small fraction of light emitted by a light source arrives at a sensor.To show how pure-Monte Carlo path tracing works, we have visualized a path, $\bar{s} = s_0 \dots s_M$, as it could be generated by pMCPT. The sphere around point s_1 indicates diffuse reflection at the surface and the cosine lobes around the points s_j and s_{j+1} stand for specular and gloss interaction of light with surfaces. Also note the upper index at the incident and exitant direction at a surface point.

```
PURE-MONTE CARLO PATH TRACING {
    ∀ pixel □_j ∈ (□_1, ..., □_{s_x · s_y}) do {
        sample point p ∈ □_j
        generate an eye-ray r = e → p
        L(s_1 → e) = TRACE(r)
}

TRACE(r) {
        compute hit point s_1 of r with object ∂V_c closest in scene
        compute normal N(s_1) at point s_1
        return SHADE(s_1, N(s_1))
}

SHADE(s, N(s)) {
        L = 0
        if ∂V_c is specular {
            generate secondary reflected or refracted ray r^∨
            L+ = TRACE(r^∨)
        } else {
            sample directions ω_i ∈ S^2
            generate a secondary ray r_2 = s + αω_i according to the interaction at ∂V_c
            L+ = TRACE(r_2)
        }
        return L
}
```

FIGURE 9.3: PSEUDOCODE FOR PURE-MONTE CARLO PATH TRACING. A coarse framework of pure-Monte Carlo path tracing consist of only three simple methods: one for generating and tracing primary rays through pixels of the image plane, the method TRACE(), which is called for all primary rays, and the method SHADE() for coloring the corresponding pixels.

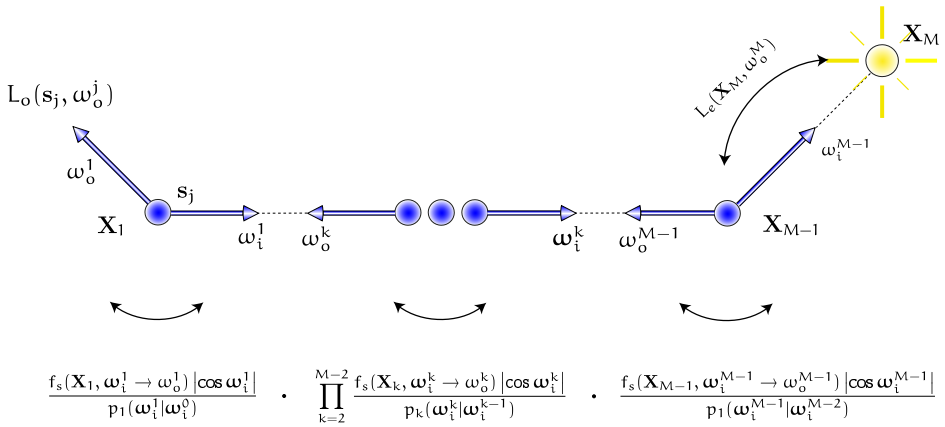


FIGURE 9.4: THE PRIMARY ESTIMATOR $F_1^{pMCPT,L_o(s_j,\omega_o^j)}$. An approximate of the exitant radiance at point s_j in direction ω_o^j can be computed via the simulation of a random walk $\overline{X} = X_1 \ldots X_M$, starting at the eye and ending at one of the light sources. Then, the direct illumination at the sample X_M can be used for shading point s_j.

k *by the BSDF at corresponding samples* ω_i^k *chosen due to probability distributions* $p_k(\omega_i^k|\omega_i^{k-1})$*, then the random variable* Y_M *is of the same type as the estimator* $F_1^{pMCPT,L_o(s_j,\omega_o^j)}$*. That is, the algorithm underlying the principle of pure-Monte Carlo path tracing is based on the stochastic model of the discrete-time, continuous-state Markov process, where the state space is given over all directions about the unit sphere centered at an interested surface point.*

Due to the emitted radiance terms in Formula (9.3), the estimator $F_1^{pMCPT,L_o(s_j,\omega_o^j)}$ not only accounts for the light contribution of the whole path $\overline{X}$, but also the light contributions from subpaths $\overline{X}$. As pure-Monte Carlo path tracing stops, if a ray hits a light source, the only path node that can end at a light source is the node X_M, that is, for all other nodes $X_k, 0 \leq k < M$ it must hold: $L_e(X_k, \omega_o^k) = 0$. Using this fact and assuming the point s_j corresponds to the sample X_1 then we get the estimator $F_1^{pMCPT,L_o(s_j,\omega_o^j)}$ in its final formulation, namely as:

$$F_1^{pMCPT,L_o(s_j,\omega_o^j)} = \prod_{k=1}^{M-1} \frac{f_s(X_k, \omega_i^k \to \omega_o^k)\,|\cos\omega_i^k|}{p_k(\omega_i^k|\omega_i^{k-1})}\, L_e(X_M, \omega_o^M), \tag{9.6}$$

for an illustration of the estimator $F_1^{pMCPT,L_o(s_j,\omega_o^j)}$, see Figure 9.4.

Due to this modification, the original drawback of our distribution ray tracing al- Section 8.4

gorithm is canceled. The construction of a random walk via the probability theoretical model of the discrete-time Markov process solves the problem of the exponentially increasing number of secondary rays, which lie deep in the ray tree and which, due to the multitude of the involved BSDFs, only contribute a small fraction of light to the final shading of point $\mathbf{s}_j$.

DT Markov Process (236)

BSDF (371)

Now, to solve the global illumination problem via pure-Monte Carlo path tracing, we have to evaluate the measurement equation in a similar way as we did it with distribution ray tracing, that is, we solve the measurement equation expressed in terms of exitant radiance, thus:

Measurement Equation (416)

$$\mathcal{M}_j = \int_{\square_j} \int_{S^2(\mathbf{s}_0)} W_e^j(\mathbf{s}_0, \omega) L_o(\gamma(\mathbf{s}_0, \omega), -\omega) \, d\sigma^{\perp}_{\mathbf{s}_0}(\omega) d\mu^2(\mathbf{s}_0), \tag{9.7}$$

where we have replaced the incident radiance by its equivalent exitant analogue, and assumed that $\mathbf{s}_0$ is a point within pixel $\square_j$ visible from the eye or a virtual camera.

Pinhole Camera (417)

Using a pinhole camera model—for details see Example 4.16—then a primary Monte Carlo estimator for pure-Monte Carlo path tracing is given by:

$$\begin{aligned} & F_1^{\mathcal{M}_j, pMCPT} \\ &= \frac{f_j(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} \langle \mathbf{N}(\mathbf{X}_0), \omega_e \rangle \, F_1^{pMCPT, L_o(\gamma(\mathbf{X}_0, \omega_e), -\omega_e)} && (9.8) \\ &= \frac{f_j(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} \left| \cos\theta_e \right| \cdot \prod_{k=1}^{M-1} \frac{f_s(\mathbf{X}_k, \boldsymbol{\omega}_i^k \to \omega_o^k) \left| \cos \boldsymbol{\omega}_i^k \right|}{p_k(\boldsymbol{\omega}_i^k | \boldsymbol{\omega}_i^{k-1})} L_e(\mathbf{X}_M, \omega_o^M), && (9.9) \end{aligned}$$

where the sample $\mathbf{X}_0$ is chosen within the pixel $\square_j$ according to the density function, p_0, $F_1^{pMCPT, L_o(\gamma(\mathbf{X}_0, \omega_e), -\omega_e)}$ is the primary estimator for evaluating the exitant radiance at point $\gamma(\mathbf{X}_0, \omega_e)$ in direction $\omega_o^1 = -\omega_e$, and ω_e corresponds to the ray starting at the eye $\mathbf{e}$ and passing through the pixel sample $\mathbf{X}_0$. A visualization of the estimator $F_1^{\mathcal{M}_j, pMCPT}$ is shown in Figure 9.5.

REMARK 9.2 *It should be clear, that pure-Monte Carlo path tracing can easily be adapted to other camera models than the pinhole camera system. Furthermore, the problem of aliasing can also be reduced and effects like depth of field or motion blur can be simulated with pMCPT. The only thing we have to do is to adapt the estimator* $F_1^{\mathcal{M}_j, pMCPT}$ *to the camera model used.*

REMARK 9.3 (Pure-Monte Carlo Path Tracing with Russian Roulette) *The primary Monte Carlo estimator* $F_1^{\mathcal{M}_j, pMCPT}$ *from above needs a stopping condition to prevent a path being of infinite length. Obviously, simply cutting off a path leads to bias into the process of image generation, even if we neglect only small contributions to the shading of a pixel. An unbiased image can be produced by using the technique of Russian Roulette. As we have seen in Section 2.4.4, we can handle the problem of keeping*

Russian Roulette (200)

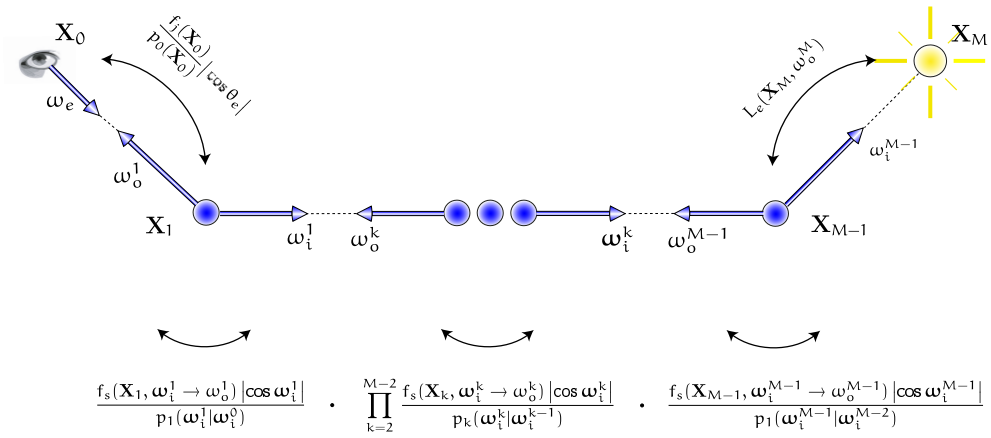


FIGURE 9.5: THE PRIMARY ESTIMATOR $F_1^{\mathcal{M}_j,\mathrm{pMCPT}}$. Pure-Monte Carlo path tracing estimates the measurement equation at pixel $\square_j$ via the simulation of a random walk starting at the first hit point, $\mathbf{s}_1$, of an eye ray with a surface object. If the associated random walk ends in one of the light sources, a contribution to the shading of the pixel is made via the exitant radiance at sample $\mathbf{X}_1$ in direction $\omega_o^1 = -\omega_e$

the lengths of a path manageable with the help of this technique, and additionally, it does not restrict us to explore all possible paths of any length.

Let $\alpha_k, k \geq 1$, *be continuous random variables defined on the canonical probability space* $([0,1], \mathfrak{B}([0,1]), \mathbb{P})$. *If we stop the process of recursive path tracing with the so-called* absorption probability $(1-\alpha_k)$ *at node* $\mathbf{X}_k$, *then we have to multiply the corresponding terms in* $F_1^{\mathcal{M}_j,\mathrm{pMCPT}}$ *with the weight* $\frac{1}{\alpha_k}$, *that is, we get:* $([0,1], \mathfrak{B}([0,1]), \mathbb{P})$ (166)

$$F_1^{\mathcal{M}_j,\mathrm{pMCPT,RR}} = \frac{f_j(\mathbf{X}_0)}{p_0(\mathbf{X}_0)}\left|\cos\theta_e\right| \cdot \prod_{k=1}^{\infty} \frac{f_s(\mathbf{X}_k, \boldsymbol{\omega}_i^k \to \omega_o^k)\,|\cos\boldsymbol{\omega}_i^k|}{\alpha_k p_k(\boldsymbol{\omega}_i^k|\boldsymbol{\omega}_i^{k-1})} L_e(\mathbf{X}_\infty, \omega_o^\infty). \tag{9.10}$$

As we have already mentioned, if the absorption probability is large, the recursion will stop sooner, which leads to a higher variance in the estimator $F_1^{\mathcal{M}_j,\mathrm{pMCPT,RR}}$. *If it is small, the recursion will continue many times, and* $F_1^{\mathcal{M}_j,\mathrm{pMCPT,RR}}$ *will be more accurate. Linked to our path tracing algorithm, this means that we get either accurate estimates if paths of a long length are generated, or less accurate estimates, if very short paths are generated. In principle any value for* α_k *can be picked for controlling the recursive depth and execution time of the algorithm.*

In rendering, the reflectance of the material of a surface is often used in connection with Russian Roulette, that is, a path is more easily absorbed at dark surfaces,

while at lighter surfaces the probability is larger that the path will be continued

REMARK 9.4 *The estimator* $F_1^{\mathcal{M}_j,pMCPT}$ *computes the radiance arriving at the pixel* $\square_j$ *after performing* M *steps via pMCPT. Only in the case where the path node* $\mathbf{X}_M$ *corresponds to the intersection of a ray with a light source, the estimator is not equals zero. In all other cases, that is, if the path generation stops due to Russian roulette, or if the maximal default length of the path is extended, the pixel remains black, since it does not get a contribution from any of the existing emitters.*

In the following example, we show the derivation of a primary Monte Carlo estimator for pure-Monte Carlo path tracing in an idealized scene only consisting of diffuse surfaces.

EXAMPLE 9.1 (Pure-Monte Carlo Path Tracing with Diffuse Surfaces) *In scenes only modeled of opaque, diffuse surfaces, the Monte Carlo estimator* $F_1^{pMCPT,L_o(\mathbf{s}_j,\omega_o^j)}$ *has a simple form. Using a BRDF that simulates ideal diffuse reflection, then this BRDF can be expressed in terms of the directional-hemispherical reflectance* ρ_{dh} *by:*

Ideal Diffuse BRDF (325)

$$f_r^o(\mathbf{s}_j, \omega_i^j \to \omega_o^j) \overset{(4.161)}{=} \frac{\rho_{dh}}{\pi}. \tag{9.11}$$

Then the estimator $F_1^{pMCPT,L_o(\mathbf{s}_j,\omega_o^j)}$ *can be formulated as:*

$$F_1^{pMCPT,L_o(\mathbf{s}_j,\omega_o^j)} = \prod_{k=1}^{M-1} \frac{\rho_{dh}}{\pi} \frac{|\cos\omega_i^k|}{p_k(\omega_i^k|\omega_i^{k-1})} L_e(\mathbf{X}_M, \omega_o^M) \tag{9.12}$$

$$= \left(\frac{\rho_{dh}}{\pi}\right)^{M-1} \prod_{k=1}^{M-1} \frac{|\cos\omega_i^k|}{p_k(\omega_i^k|\omega_i{}^{k-1})} L_e(\mathbf{X}_M, \omega_o^M), \tag{9.13}$$

where $\mathbf{X}_1$ *corresponds to point* $\mathbf{s}_j$, ω_i^k *are according to the PDF* p_k *distributed random variables drawn over the upper hemisphere* $\mathcal{H}_i^2$ *about the sample* $\mathbf{X}_k$, *and* $p_0(\omega_i^1|\omega_i^0) = p_0(\omega_i^1)$.

Using this estimator in the measurement equation then we get:

$$F_1^{\mathcal{M}_j,pMCPT} = \frac{f_j(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} |\cos\theta_e| \cdot \left(\frac{\rho_{dh}}{\pi}\right)^{M-1} \prod_{k=1}^{M-1} \frac{|\cos\omega_i^k|}{p_k(\omega_i^k|\omega_i{}^{k-1})} L_e(\mathbf{X}_M, \omega_o^M). \tag{9.14}$$

For an illustration of the estimator $F_1^{\mathcal{M}_j,pMCPT}$, *see Figure 9.6.*

Applying a cosine-weighted hemisphere sampling strategy—directions near to the surface normal are favored over those at oblique angles to the surface—as presented in Example 6.20, where the involved PDF is of type

$$p_k(\omega_i^{k-1}|\omega_i^{k-2}) = \frac{\cos\omega_i^{k-1}}{\pi} \tag{9.15}$$

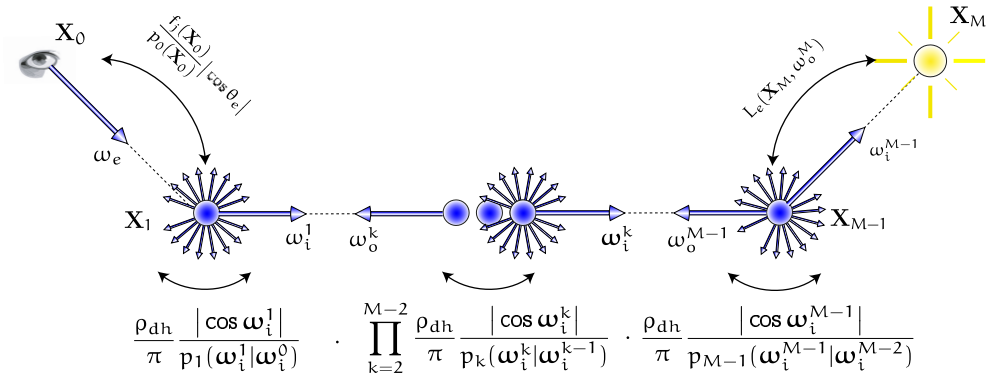


FIGURE 9.6: PURE-MONTE CARLO PATH TRACING WITH DIFFUSE SURFACES. The path $\overline{\mathbf{X}} = \mathbf{X}_0\mathbf{X}_1\ldots\mathbf{X}_M$, generated via pure-Monte Carlo path tracing, goes over diffuse surfaces. Since the ideal diffuse BRDF is constant, the estimator $F_1^{\mathcal{M}_j,\mathrm{pMCPT}}$ can be expressed in a particularly simple form.

then leads to the following form of the estimator:

$$F_1^{\mathcal{M}_j,\mathrm{pMCPT}} = \frac{f_j(\mathbf{X}_0)}{p_0(\mathbf{X}_0)}\langle \mathbf{N}(\mathbf{X}_0), \omega_e\rangle \cdot \left(\prod_{k=1}^{M-1}\rho_{dh}\right) L_e(\mathbf{X}_M, \omega_o^M) \tag{9.16}$$

$$= \frac{f_j(\mathbf{X}_0)}{p_0(\mathbf{X}_0)}|\cos\theta_e| \cdot \rho_{dh}^{M-1} L_e(\mathbf{X}_M, \omega_o^M). \tag{9.17}$$

REMARK 9.5 *As pure-Monte Carlo path tracing is based on distribution ray tracing, the algorithm can also simulate random walks of characteristic* $\overrightarrow{E(D|G|S)^*L}$. *So, in principle, it provides us a complete solution of the stationary light transport equation in vacuum, but it suffers from the fact that the resulting images are very noisy, see Figure 9.7.* SLTEV (398)

THE PROBLEM OF HITTING A LIGHT SOURCE. Let us consider the images in Figure 9.7, rendered with pure-Monte Carlo path tracing, a little bit closely. They should show the Cornell box only modeled with diffuse surfaces illuminated by a single spherical light source. Within the left image, we do not recognize the Cornell box, since this image is mostly black with some very bright pixels.

Pure-Monte Carlo path tracing renders this image by shooting a single ray from the eye through a pixel and traces this ray through the scene. As light sources, compared to other objects within a scene, are mostly small, the probability of hitting a light source is also very small. This means that most pixels within the image are black, and those pixels that are starting points of a path ending in a light source gets a large radiance value from

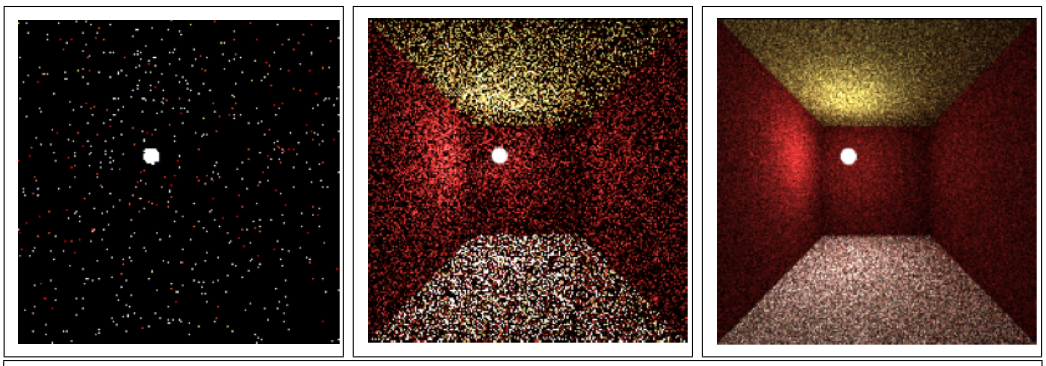


FIGURE 9.7: IMAGES RENDERED WITH PURE-MONTE CARLO PATH TRACING. The above images are rendered with pMCPT using eye paths of length ≤ 5. They show the Cornell box, illuminated by a spherical light source, where all surfaces corresponds to diffuse objects. The algorithm uses $1, 64$ and 1024 samples per pixel. Increasing the number of samples per pixel also leads to an increasing of the probability that eye-paths starting at the corresponding pixel hit the light source. This then results in less dark pixels within the image, and the brightness of colored pixels is scaled down due to averaging the light contributions of several paths ending at the light source, that is, we perceive colors. Note: Also the first two images, although different in the quality, are, due to pixels being very bright, perfectly valid secondary estimators for the correct solution of the SLTEV. Image courtesy of Hugh McCabe, Department of Computer Science, Trinity College, Dublin.

the light source, attenuated only by a factor smaller than one at the surfaces induced to the non-ideal diffuse BRDFs. This is the reason why the resulting images are often dark and very noisy.

Now, from probability theory it is known, that the variance—in our case, shown as noise in an image—can be reduced by taking more samples. With respect to pMCPT this mean: Generating more random walks starting at a pixel improves the quality of our images considerably. This can easily be seen at the right image in Figure 9.7, where the Cornell box is sampled with 1024 paths per pixel. But also this image is still very noisy, and thus not satisfactory. That is, we must look for another method that can help us further to reduce the noise in an image. A technique that can be used to approach this goal is to account for the local, directly incident illumination at a path node. This method is called *Monte Carlo path tracing with next event estimation.*

9.1.2 MONTE CARLO PATH TRACING WITH NEXT EVENT ESTIMATION

Let $\overline{\mathbf{X}} = \mathbf{X}_0\mathbf{X}_1 \ldots \mathbf{X}_M$ be a path generated with pure-Monte Carlo path tracing, where the sample $\mathbf{X}_0$ is chosen in the pixel $\square_j$. As the estimator $F_1^{\mathcal{M}_j,\mathrm{pMCPT}}$ from Equation (9.8) shows, this path only contributes to the shading of the pixel if $\overline{\mathbf{X}}$ ends in one of the light

sources, that is, if it holds: $\mathbf{X}_M \in \text{☼}$. In the case where $\mathbf{X}_M \notin \text{☼}$ the considered pixel remains black.

Now, in Remark 9.1, we have noticed that the foundation of the pMCPT algorithm can be found in a discrete-time, continuous-state Markov Process associated with a random variable $\mathbf{Y}_m$ which was used for estimating a Fredholm integral equation of the 2^{nd} kind via the Neumann series approach, for details see Section 6.7.3.

Now, the Monte Carlo estimator $F_1^{pMCPT,L_o(s_j,\omega_o^j)}$ for pure-Monte Carlo path tracing from Equation 9.6 only accounts for a single transport path, namely the path that ends at a light source, i.e. pMCPT computes even only a single contribution of the random variable $\mathbf{Y}_m$ to the shading of a surface point. This also means, that only a single addendum of the m^{th} term of the Neumann series approach is estimated with pMCPT. Obviously, this is not a good approach. A much better idea for estimating the illumination of a surface point is, to account for also the contributions of all subpaths of $\overline{\mathbf{X}}$. This then corresponds to the evaluation of a single term of all sums of the Neumann series approach.

Transferred to Monte Carlo path tracing, our discussion from above shows: pMCPT only accounts for the direct illumination at the last node of the path $\overline{\mathbf{X}}$, if $\mathbf{X}_M \in \text{☼}$. As the direct illumination represents the largest contribution for shading a surface point, the direct illumination should be accounted for at every path node of $\overline{\mathbf{X}}$. This idea then leads to the following modification of pMCPT: At every node of a random walk $\overline{\mathbf{X}} = \mathbf{X}_0 \dots \mathbf{X}_1 \dots \mathbf{X}_M$, generated via pMCPT, our new algorithm shoots one or more shadow rays in direction to the light sources. The light that directly arrives at a path node via a shadow ray can then flow back over the path $\overline{\mathbf{X}}$ to its origin $\mathbf{X}_0$ where it is accumulated for shading a pixel, see Figure 9.8.

Direct Illumination (410)

Shadow Ray (14)

MONTE CARLO PATH TRACING WITH NEXT EVENT ESTIMATION. From our discussion in Section 4.4.2.2 it is known that, the integration domain of the SLTEV, thus the unit sphere around surface point $\mathbf{s}_j$ can be split into two disjoint, Lebesgue measurable sets: The projection of all visible light sources onto the unit sphere around surface point $\mathbf{s}_j$, denoted by the set $\text{☼}^\perp$, and the complement of this set $\overline{\text{☼}^\perp} = S^2 \setminus \text{☼}^\perp$. The linearity property of the Lebesgue integral with respect to the integration domain then allows to write the SLTEV as composed of three different types of exitant radiance at point $\mathbf{s}_j$, namely:

SLTEV (398)

Lebesgue Measurable Set (80)

Lebesgue Integral (105)

SLTEV (398)

$$L_o(\mathbf{s}_j, \omega_o^j) \overset{(4.407)}{=} L_e(\mathbf{s}_j, \omega_o^j) + L^{\leftarrow}(\mathbf{s}_j, \omega_o^j) + L^{\leftleftarrows}(\mathbf{s}_j, \omega_o^j), \tag{9.18}$$

where $L_e(\mathbf{s}_j, \omega_o^j)$ corresponds to the self-emitted radiance at $\mathbf{s}_j$ in direction ω_o^j, $L^{\leftarrow}(\mathbf{s}, \omega_o^j)$ represents the direct illumination at point $\mathbf{s}_j$ reflected in direction ω_o^j, given by:

Section 4.4.2.2

$$L^{\leftarrow}(\mathbf{s}_j, \omega_o^j) \overset{(4.406)}{=} \int_{\text{☼}^\perp} f_s(\mathbf{s}_j, \omega_i^j \to \omega_o^j)\, L_e(\gamma(\mathbf{s}_j, \omega_i^j), -\omega_i^j)\, d\sigma^\perp_{\mathbf{s}_j}(\omega_i^j), \tag{9.19}$$

and $L^{\leftleftarrows}(\mathbf{s}_j, \omega_o^j)$, defined by:

$$L^{\leftleftarrows}(\mathbf{s}_j, \omega_o^j) \overset{(4.406)}{=} \int_{\overline{\text{☼}^\perp}} f_s(\mathbf{s}_j, \omega_i^j \to \omega_o^j)\, L_i(\mathbf{s}_j, \omega_i^j)\, d\sigma^\perp_{\mathbf{s}_j}(\omega_i^j), \tag{9.20}$$

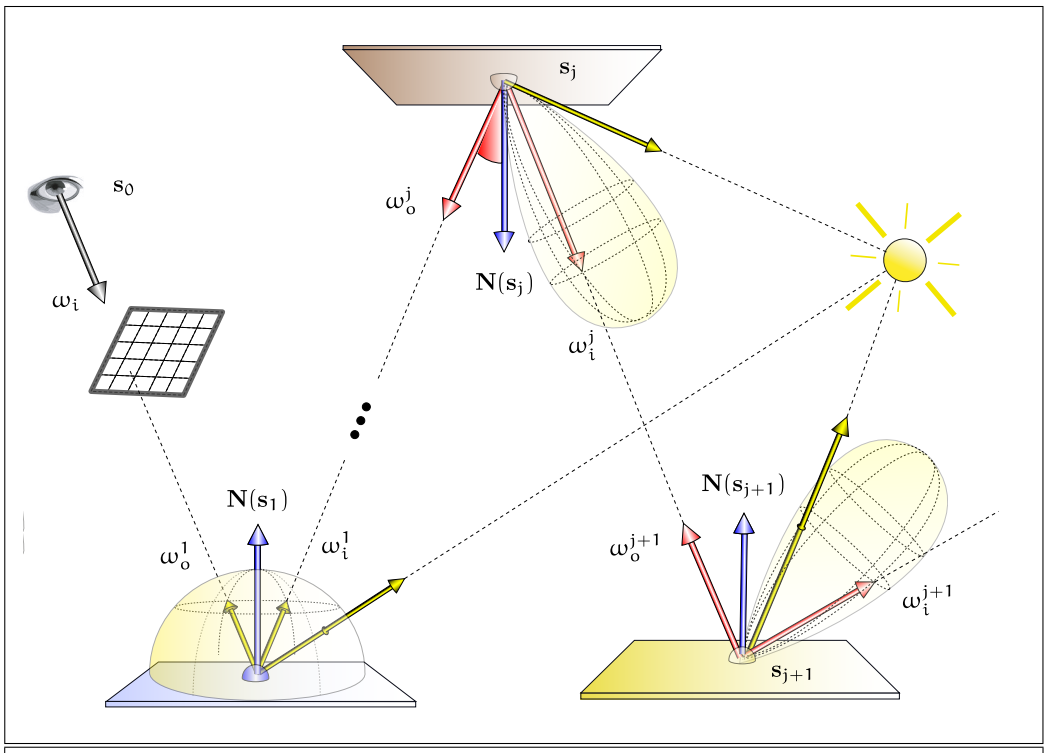


FIGURE 9.8: MONTE CARLO PATH TRACING WITH NEXT EVENT ESTIMATION, MCPT. Starting from the eye of an observer or the virtual camera, the algorithm generates a ray and shoots this ray through a pixel into the scene to be rendered. At the first intersection point, s_1, of this ray with an object in the scene, Monte Carlo path tracing with next event estimation generates—depending on the material and the surface properties of the object that has been hit—a new ray in direction ω_i^1 over the unit sphere S^2, or the upper respectively the lower hemisphere about s_1 . Additionally, the algorithm generates a so-called shadow ray in direction to one of the light sources in the scene for computing the direct illumination at point s_1. The algorithm then repeats this step recursively until a ray hits a light source of the scene, a ray does not hit any object of scene, or the process is stopped via Russian roulette respectively the length of the path exceeds a predetermined default value. The incident radiance $L_i(s_j, \omega_i^j)$ then corresponds to the contributions due to direct illumination along the path nodes. Attenuated by the product of the factors induced by the BSDFs at the intersection points of a path with the objects of scene, it is accumulated at the origin of the path for shading the considered pixel. If a path ends in one of the light sources, we must be careful not to account for this light contribution twice.

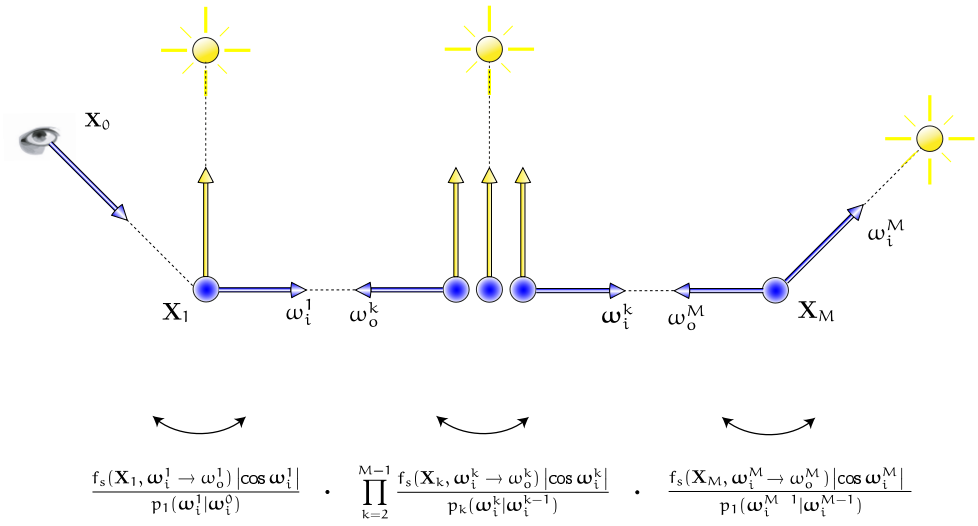


FIGURE 9.9: CONTRIBUTIONS TO THE SHADING OF A PIXEL IN MCPT WITH NEXT EVENT ESTIMATION. At every node of a random walk generated via pMCPT the algorithm shoots one or more shadow rays in direction to the light sources. The light that directly arrives at a path node via a shadow ray can then flow back over the path to its origin where it delivers a contribution for shading a pixel.

corresponds to the indirect illumination component of the SLTEV at point $\mathbf{s}_j$ from direction ω_i^j.

Now, let $\overline{\mathbf{X}} = \mathbf{X}_0\mathbf{X}_1 \ldots \mathbf{X}_M$ be a random walk generated by pMCPT, where $\mathbf{X}_0$ is a sample on pixel $\square_j$. Regardless of whether the last path node $\mathbf{X}_M$ was sampled on a light source, MCPT with next event estimation always determines the direct illumination at each path node $\mathbf{X}_k, 1 \leq k \leq M$ of the path $\overline{\mathbf{X}}$. For this, the algorithm samples one or more shadow rays in direction to the light sources, captures the light contribution of the source and gives it back to the origin of the path for shading the corresponding pixel, see Figure 9.9. Note, the indirect illumination at node $\mathbf{X}_k$ corresponds to the accumulated light contributions from all successor $\mathbf{X}_{k+1}, 1 \leq k \leq M-1$ of path node $\mathbf{X}_k$.

As shown in Example 6.46, so it is more efficient to compute the direct illumination at a path node using the 3-point formulation of the SLTEV instead of to sample shadow rays over the hemisphere. Thus, we transform the spherical integral from Equation (9.19) into an area integral, where the light surfaces are used as the domain of integration. Then, the direct illumination at point $\mathbf{s}_j$ reflected in direction ω_o^j can be formulated as:

SLTEV (398)

$$L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1}) = \int_{\mathbb{S}} f_s(\mathbf{l} \to \mathbf{s}_j \to \mathbf{s}_{j-1})\, L_e(\mathbf{l} \to \mathbf{s}_j)\, \mathcal{G}(\mathbf{l} \leftrightarrow \mathbf{s}_j)\, d\mu^2(\mathbf{l}), \tag{9.21}$$

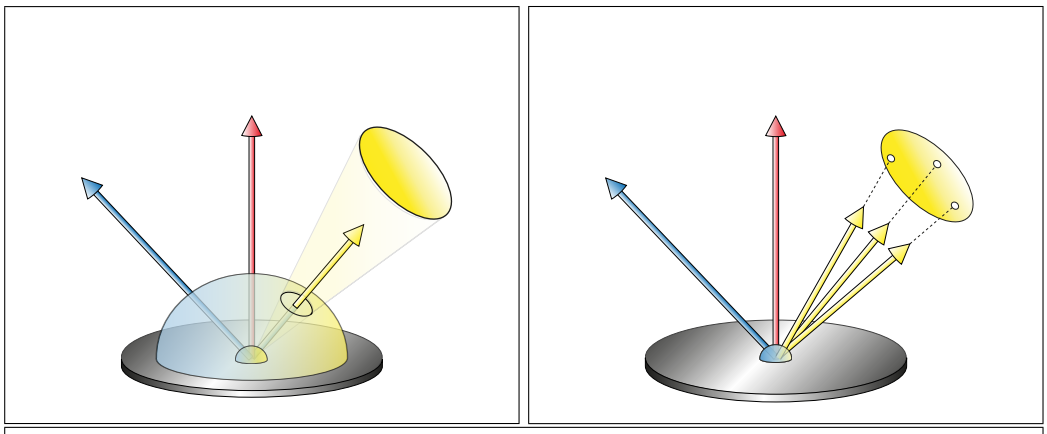


FIGURE 9.10: THE GEOMETRY FOR SAMPLING THE DIRECT ILLUMINATION IN MONTE CARLO PATH TRACING. Since it is more efficient to sample direct illumination using the 3-point formulation of the SLTEV, we transform the spherical integral which describes the direct illumination at point $\mathbf{s}_j$ into an area integral.

where $\mathbf{l}$ are points at the light sources and $\mathbf{s}_{j-1} = \gamma(\mathbf{s}_j, \omega_o^j)$ is a surface point visible from $\mathbf{s}_j$ in direction ω_o^j. For an illustration, see Figure 9.10.

PDF (176) Probability Space (163) $\mathfrak{B}(\cdot)$ (865)

Based on this integral formulation, then we sample shadow rays by choosing points $\mathbf{l}_j$ according to a probability density $p_{\sphericalangle}$ on the probability space $(\sphericalangle, \mathfrak{B}(\sphericalangle), \mu^2)$, as we did it in Example 6.46. An associated primary Monte Carlo estimator for approximating $L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1})$ then has the form

$$F_1^{L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1})} = \frac{f_s(\mathbf{l}_j \to \mathbf{s}_j \to \mathbf{s}_{j-1})\, L_e(\mathbf{l}_j \to \mathbf{s}_j)\, \mathcal{G}(\mathbf{l}_j \leftrightarrow \mathbf{s}_j)}{p_{\sphericalangle}(\mathbf{l}_j)}; \tag{9.22}$$

for an implementation of the estimator $F_1^{L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1})}$ see Figure 9.11.

EXAMPLE 9.2 (Uniform Sampling of Light Source Area) *The simplest and in praxis mostly used method for generating shadow rays bases on uniform area sampling of the light sources existing in a scene. Let us now assume that all those light sources $\sphericalangle_1 \cup \ldots \cup \sphericalangle_N$ are composed to a single big, fat light source $\sphericalangle = \sphericalangle_1 \cup \ldots \cup \sphericalangle_N$. Then, $\mu^2(\sphericalangle)$* μ^2 (82) *corresponds to the Lebesgue area measure of this emitter.*

Sampling N points $\mathbf{l}_{j_k}$ according to the probability density function $p_{\sphericalangle} = \frac{1}{\mu^2(\sphericalangle)}$ $\mathfrak{B}(\cdot)$ (865) *on the probability space $(\sphericalangle, \mathfrak{B}(\sphericalangle), \mu^2)$ leads to a secondary estimator $F_N^{L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1})}$ of the form:*

$$F_N^{L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1})} = \frac{\mu^2(\sphericalangle)}{N} \sum_{k=1}^{N} f_s(\mathbf{l}_{j_k} \to \mathbf{s}_j \to \mathbf{s}_{j-1})\, L_e(\mathbf{l}_{j_k} \to \mathbf{s}_j)\, \mathcal{G}(\mathbf{l}_{j_k} \leftrightarrow \mathbf{s}_j). \tag{9.23}$$

$F_1^{L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1})}() \{$
  $L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1}) = 0$
  sample $\mathbf{l}_j \in \overleftarrow{\Omega} = \bigcup_{j=1}^{N} \overleftarrow{\Omega}_j$ according the PDF $p_{\overleftarrow{\Omega}}$
  $L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1}) += \frac{f_s(\mathbf{l}_j \to \mathbf{s}_j \to \mathbf{s}_{j-1})\, L_e(\mathbf{l}_j \to \mathbf{s}_j)\, G(\mathbf{l}_j \leftrightarrow \mathbf{s}_j)}{p_{\overleftarrow{\Omega}}(\mathbf{l}_j)}$
  return $L^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1})$
$\}$

FIGURE 9.11: IMPLEMENTATION OF DIRECT ILLUMINATION IN MONTE CARLO PATH TRACING. The secondary estimator $F_N^{L_o^{\leftarrow}(\mathbf{s}_j \to \mathbf{s}_{j-1})}$ approximated via Monte Carlo path tracing with next event estimation. In implementations of MCPT with next event estimation, the number of shadow rays is commonly chosen as $N = 1$.

As already mentioned in Example 6.46, this naive sampling strategy can lead to noise in the resulting images, which can easily be reduced using variance reduction techniques. Section 6.6

If the problem of direct illumination is solved, the problem of sampling directions over $\overleftarrow{\Omega^{\perp}}$ for computing the indirect illumination can easily be circumvented. Instead of to sample a direction over the possible complex, Lebesgue measurable set $\overleftarrow{\Omega^{\perp}}$, we sample a direction over the whole unit sphere according to the involved BSDF. Here, we have to take care—under the premise that the corresponding path ends at a light source—that we does not take into account the contribution of a light source, which would mean that we would count light sources twice. That is, the indirect illumination component $L^{\Leftarrow}$ can be computed via pure-Monte Carlo path tracing.

Then, a Monte Carlo estimator based on path tracing with next event estimation, that approximates a solution to the measurement equation, has the following form:

$$\mathcal{M}_j = \int_{\square_j} \int_{S^2(\mathbf{s}_0)} W_e^j(\mathbf{s}_0, \omega) L_i(\mathbf{s}_0, \omega)\, d\sigma^{\perp}_{\mathbf{s}_0}(\omega) d\mu^2(\mathbf{s}_0) \tag{9.24}$$

$$\overset{(9.18)}{=} \int_{\square_j} \int_{S^2(\mathbf{s}_0)} W_e^j(\mathbf{s}_0, \omega) \left(L_e(\gamma(\mathbf{s}_0, \omega), -\omega) + \right. \tag{9.25}$$
$$\left. L^{\leftarrow}(\gamma(\mathbf{s}_0, \omega), -\omega) + L^{\Leftarrow}(\gamma(\mathbf{s}_0, \omega), -\omega)\right) d\sigma^{\perp}_{\mathbf{s}_0}(\omega) d\mu^2(\mathbf{s}_0),$$

where the incident radiance L_i was replaced by the equivalent exitant representation of the SLTEV at point $\gamma(\mathbf{s}_0, \omega)$ expressed in terms of emitted, direct, and indirect radiance from Equation (9.18). Furthermore, we have assumed that the first hit point of the primary ray, starting at the eye and going in direction ω, with a scene object corresponds to the point $\mathbf{s}_1 = \gamma(\mathbf{s}, \omega)$ and the outgoing direction at $\mathbf{s}_1$ is given via $\omega_o^1 = -\omega$.

Pinhole Camera (417)

Using a pinhole camera—for details see Example 4.16—then a primary Monte Carlo estimator based on Monte Carlo path tracing with next event estimation can easily be derived by combining the estimator from pMCPT, thus $F_1^{\mathcal{M}_j,pMCPT}$ from Formula (9.8), and the estimator that approximates the direct illumination at the observation point from Equation (9.22). Applied to the path $\overline{\mathbf{X}} = \mathbf{X}_0, \mathbf{X}_1 \ldots \mathbf{X}_M$, we get:

$$
\begin{aligned}
F_1^{\mathcal{M}_j,MCPT} = & \\
& \frac{f_j(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} \langle \mathbf{N}(\mathbf{X}_0), \omega_e \rangle \cdot \\
& \Bigg(F_1^{L^{\leftarrow}(\mathbf{X}_1 \to \mathbf{X}_0)} + \\
& \underbrace{\frac{f_s(\mathbf{X}_1, \omega_i^1 \to \omega_o^1) \left|\cos \omega_i^1\right|}{p_1(\omega_i^1)} F_1^{L^{\leftarrow}(\mathbf{X}_2 \to \mathbf{X}_1)}}_{F_1^{L^{\Leftarrow}(\mathbf{X}_1 \to \mathbf{X}_0)}} + \\
& \underbrace{\frac{f_s(\mathbf{X}_1, \omega_i^1 \to \omega_o^1) \left|\cos \omega_i^1\right|}{p_1(\omega_i^1)} \frac{f_s(\mathbf{X}_2, \omega_i^2 \to \omega_o^2) \left|\cos \omega_i^2\right|}{p_2(\omega_i^2 | \omega_i^0)} F_1^{L^{\leftarrow}(\mathbf{X}_3 \to \mathbf{X}_2)}}_{F_1^{L^{\Leftarrow}(\mathbf{X}_2 \to \mathbf{X}_0)}} + \ldots + \\
& \prod_{k=1}^{M-1} \frac{f_s(\mathbf{X}_k, \omega_i^k \to \omega_o^k) \left|\cos \omega_i^k\right|}{p_k(\omega_i^k | \omega_i^{k-1})} F_1^{L^{\leftarrow}(\mathbf{X}_{k+1} \to \mathbf{X}_k)} \Bigg),
\end{aligned}
\tag{9.26}
$$

where $\mathbf{X}_0$ is a point chosen on the pixel $\square_j$ according to the PDF p_0, $F_1^{L^{\leftarrow}(\mathbf{X}_k \to \mathbf{X}_{k-1})}$ is the primary estimator for estimating the direct illumination at point $\mathbf{X}_k$ in direction to $\mathbf{X}_{k-1}$, and ω_e is as usually the direction from eye point $\mathbf{e}$ through the pixel sample $\mathbf{X}_0$.

Using

$$
F_1^{L^{\leftarrow}(\mathbf{X}_k \to \mathbf{X}_{k-1})} = \frac{f_s(\mathbf{l}_k \to \mathbf{X}_k \to \mathbf{X}_{k-1}) \, L_e(\mathbf{l}_k \to \mathbf{X}_k) \, \mathcal{G}(\mathbf{l}_k \leftrightarrow \mathbf{X}_k)}{p_{\text{☼}}(\mathbf{l}_k)},
\tag{9.27}
$$

where the sample $\mathbf{l}_k$ is chosen from one of the existing light sources according to the PDF

$p_{\sphericalangle}$, then we get:

$$
\begin{aligned}
F_1^{\mathcal{M}_j,MCPT} = & \\
& \frac{f_j(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} \langle \mathbf{N}(\mathbf{X}_0), \omega_e \rangle \cdot \\
& \left(f_s(\mathbf{l}_1 \to \mathbf{X}_1 \to \mathbf{X}_0)\, \mathcal{G}(\mathbf{l}_1 \leftrightarrow \mathbf{X}_1)\, \frac{L_e(\mathbf{l}_1 \to \mathbf{X}_1)}{p_{\sphericalangle}(\mathbf{l}_1)} + \right. \\
& \frac{f_s(\mathbf{X}_1, \omega_i^1 \to \omega_o^1) \left| \cos \omega_i^1 \right|}{p_1(\omega_i^1 | \omega_i^0)}\, f_s(\mathbf{l}_2 \to \mathbf{X}_2 \to \mathbf{X}_1)\, \mathcal{G}(\mathbf{l}_2 \leftrightarrow \mathbf{X}_2)\, \frac{L_e(\mathbf{l}_2 \to \mathbf{X}_2)}{p_{\sphericalangle}(\mathbf{l}_2)} + \ldots + \\
& \prod_{k=1}^{M-1} \frac{f_s(\mathbf{X}_k, \omega_i^k \to \omega_o^k) \left| \cos \omega_i^k \right|}{p_k(\omega_i^k | \omega_i^{k-1})}\, f_s(\mathbf{l}_{M-1} \to \mathbf{X}_{M-1} \to \mathbf{X}_{M-2})\, \mathcal{G}(\mathbf{l}_M \leftrightarrow \mathbf{X}_{M-1}) \cdot \\
& \left. \frac{L_e(\mathbf{l}_M \to \mathbf{X}_M)}{p_{\sphericalangle}(\mathbf{l}_M)} \right),
\end{aligned}
\tag{9.28}
$$

that is, a closed formula for $F_1^{\mathcal{M}_j,MCPT}$ then looks like this:

$$
\begin{aligned}
F_1^{\mathcal{M}_j,MCPT} = & \frac{f_j(\mathbf{X}_0)}{p_0(\mathbf{X}_0)} \left| \cos \theta_e \right| \cdot \\
& \left(\sum_{j=1}^{M+1} \prod_{k=1}^{j-1} \frac{f_s(\mathbf{X}_k, \omega_i^k \to \omega_o^k) \left| \cos \omega_i^k \right|}{p_k(\omega_i^k | \omega_i^{k-1})} f_s(\mathbf{l}_j \to \mathbf{X}_j \to \mathbf{X}_{j-1})\, \mathcal{G}(\mathbf{l}_j \leftrightarrow \mathbf{X}_j)\, \frac{L_e(\mathbf{l}_j \to \mathbf{X}_j)}{p_{\sphericalangle}(\mathbf{l}_j)} \right),
\end{aligned}
\tag{9.29}
$$

where we assume that it holds: $\prod_{k=1}^{0} \frac{f_s(\mathbf{X}_k, \omega_i^k \to \omega_o^k) \left| \cos \omega_i^k \right|}{p_j(\omega_i^k | \omega_i^{k-1})} = 1$, see Figure 9.12.

REMARK 9.6 *Note, when evaluating this estimator we must be careful when the chosen directional sample hits a light source. Since this sample may be associated with a shadow ray, we may not take into account the contribution of this sample to the final radiance value of the pixel, otherwise we could account for light sources twice.*

Care must also be taken when computing direct illumination at specular objects. Due to the laws of reflection and refraction, only in the case where the BSDF associated with a surface has also a small diffuse or a glossy component, the direct illumination can contribute to the shading of a pixel. This then prevents the usually failures in many ray tracing images, where reflection of light sources is fuzzy while reflection of specular objects is hard.

Law of Reflection (300)

Law of Refraction (305)

REMARK 9.7 *Note: Monte Carlo Path tracing can be considered as the* standard algorithm *in the field of realistic rendering.*

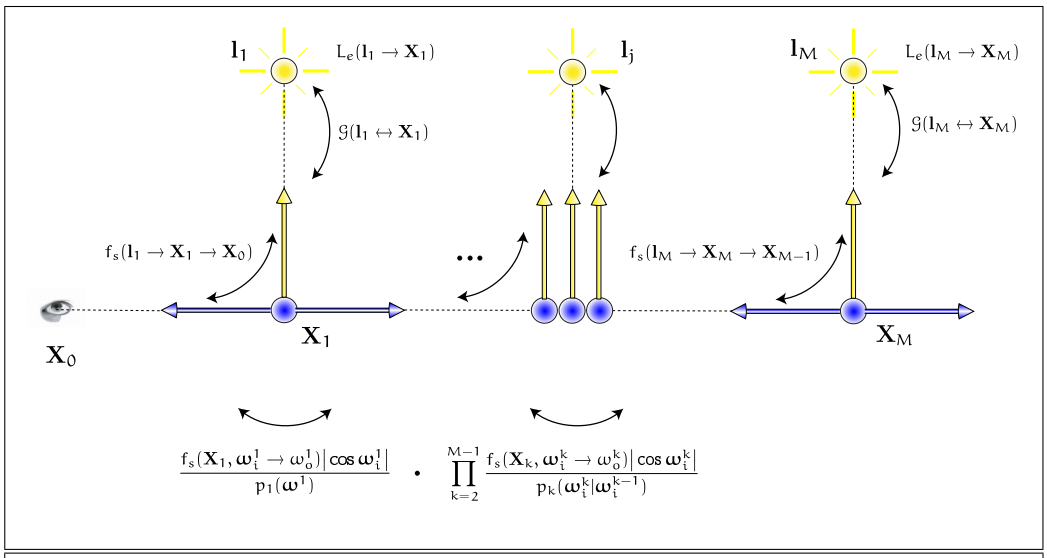


FIGURE 9.12: THE PRIMARY ESTIMATOR $F_1^{\mathcal{M}_j,MCPT}$. The j^{th} term of $F_1^{\mathcal{M}_j,MCPT}$ corresponds to the product of the labels of the path starting at the eye and ending at the light source labeled with j.

9.2 MONTE CARLO LIGHT TRACING

Let us consider Figure 9.13, where the left image is rendered with Monte Carlo path tracing. Compared to the image on the right, it is very noisy, in particular the region, where you can see the caustic. For simulating caustics, MCPT has to trace paths of characteristic $\overrightarrow{EDS^+L}$. That is, a ray starting a the eye and reflected by a diffuse surface has to be scattered at least at one specular surface, before it ends in one of the light sources. Obviously, paths of such type deliver a hight contribution to the final image, but the probability that MCPT generates such paths is very small. One hand, this is connected with the fact that rays are not necessarily scattered at diffuse surfaces in directions to specular objects, and on the other hand the probability that a specular reflected or scattered ray hits a light source is very small, since light sources in a scene are usually small. Indeed, path tracing is able to simulate eye paths of characteristic $\overrightarrow{EDS^+L}$, but it has great problems to generate a large number of such paths for the pixel to be considered.

Section 9.1
Caustics (658)
Heckbert's Path Notation (655)

We can solve this problem by changing our rendering strategy: Instead of tracing eye paths in directions to the light sources, we simulate the natural propagation of light, that is, we trace light on its natural way from light sources to the eye of an observer or a virtual camera. All rendering algorithms that pursue this strategy are summarized under the term *Monte Carlo light tracing*, [52, Dutré 1993], [116, Lafortune 1996], also shortly

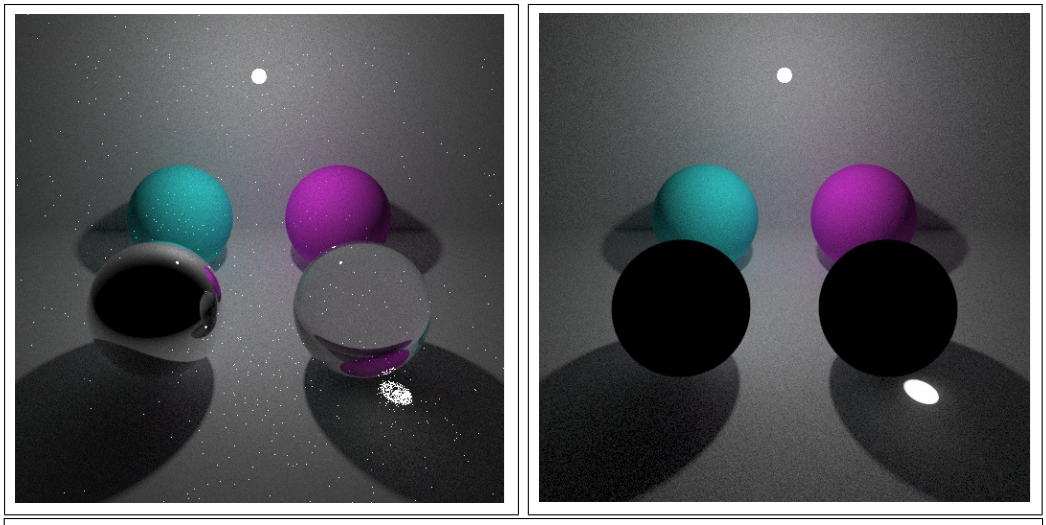


FIGURE 9.13: CAUSTICS. The image on the left is rendered with Monte Carlo path tracing. It is pretty noisy due to caustic paths that are poorly sampled by path tracing. Image courtesy of Simon Brown.

denoted as *MCLT*, or simply *light tracing*—often also known in the literature as *particle tracing*, [150, Pattanaik and Mudur 1993] and [149, Pattanaik and Mudur 1993].

Now, *Monte Carlo path tracing*, as discussed in the last section, evaluates the measurement equation, Measurement Equation (416)

$$\mathcal{M}_j = \langle W_{e,o}, L_i \rangle, \tag{9.30}$$

via a straightforward Monte Carlo approach by solving the SLTEV at points visible by the eye and combines the radiance incident at the eye with the exitant importance through the pixel $\square_j$. SLTEV (398)

As we know from Chapter 5, this is not the only possible way to evaluate the measurement equation. As the Relations (5.157) and (5.159) show, the measurement equation can also be considered in the dual forms

$$\mathcal{M}_j = \langle W_o, L_{e,i} \rangle, \tag{9.31}$$

$$\mathcal{M}_j = \langle W_i, L_{e,o} \rangle, \tag{9.32}$$

that is, it can also be evaluated by solving the importance transport equation at points within a scene combined with the light emitted from light sources. *Monte Carlo light tracing* follows exactly this approach, that is, the algorithm tries to solve the dual formulation of the global illumination problem by means of a Markov process via a procedure simulating the natural propagation of light. Discrete Markov Process (236)

9.2.1 PURE-MONTE CARLO LIGHT TRACING

Since Monte Carlo light tracing can be interpreted as the dual algorithm to Monte Carlo path tracing, the only thing we have to do to develop a light tracing algorithm, is to change the kind of rays and the strategy how we have to trace these rays through a scene. Instead of generating eye paths, a light tracing algorithm generates so-called *light paths*, starting from points at the light sources.

PURE-MONTE CARLO LIGHT TRACING. Generating light paths from emitters and tracing these paths through the scene until a sensor is hit is called *pure-Monte Carlo light tracing*, see Figure 9.14. Mathematically, these light paths can be interpreted as random walks from a discrete-time Markov process for solving the stationary vacuum importance transport equation. Whenever such a random walk passes through the frustum of a pixel and ends at a sensor, light is shot from the origin of the path to the pixel and a light contribution to that pixel can be added. The algorithm stops, if a path ends either at a sensor, a ray does not hit any object, or if the travel of the ray through the scene is stopped via Russian roulette, respectively the length of the random walk exceeds a predetermined length.

Markov Process (236)
SITEV (413)
Russian Roulette (200)

The SITEV can now easily be estimated via the following primary Monte Carlo estimator based on pMCLT, namely:

$$F_1^{pMCLT,W_o(s,\omega_o)} = \prod_{k=0}^{M-1} \frac{f_s^*(\mathbf{X}_k, \boldsymbol{\omega}_i^k \to \omega_o^k)\left|\cos \boldsymbol{\omega}_i^k\right|}{p_k(\boldsymbol{\omega}_i^k|\boldsymbol{\omega}_i^{k-1})} W_e(\mathbf{X}_M, \omega_o^M). \tag{9.33}$$

Based on the estimator $F_1^{pMCLT,W_o(s,\omega_o)}$, a primary estimator for pure-Monte Carlo light tracing is then given by:

$$\begin{aligned} &F_1^{M_j,pMCPT} \\ &= \frac{L_e(\mathbf{X}_0, \boldsymbol{\omega}_i)}{p_0(\mathbf{X}_0, \boldsymbol{\omega}_i)} \langle \mathbf{N}(\mathbf{X}_0), \omega_l \rangle F_1^{pMCLT,W_o(\gamma(\mathbf{X}_0,\omega_l),-\omega_l)} && (9.34) \\ &= \frac{L_e(\mathbf{X}_0, \boldsymbol{\omega}_i)}{p_0(\mathbf{X}_0, \boldsymbol{\omega}_i)} \left|\cos \theta_l\right| \cdot \prod_{k=1}^{M} \frac{f_s^*(\mathbf{X}_k, \boldsymbol{\omega}_i^k \to \omega_o^k)\left|\cos \boldsymbol{\omega}_i^k\right|}{p_k(\boldsymbol{\omega}_i^k|\boldsymbol{\omega}_i^{k-1})} W_e(\mathbf{X}_{M+1}, \omega_o^{M+1}), && (9.35) \end{aligned}$$

where $\mathbf{X}_0$ is a sample chosen at a light source, ω_i is the incident direction sampled on a light source due to the PDF p_0, as well as $F_1^{pMCLT,W_o(\gamma(\mathbf{X}_0,\omega_i),-\omega_i}$ is the primary estimator for estimating the importance incident at the sample $\mathbf{X}_0$ from direction ω_i.

REMARK 9.8 *There are a few essential differences between pure-Monte Carlo light tracing and pure-Monte Carlo path tracing: Thus, a random walk in pMCLT is independent of a pixel. Instead of considering each pixel in turn in pure-Monte Carlo light tracing pixels can therefore all be dealt with at the same time. Whenever a random walk passes through the frustum of a pixel a contribution can be added to the estimate of any of the pixels.*

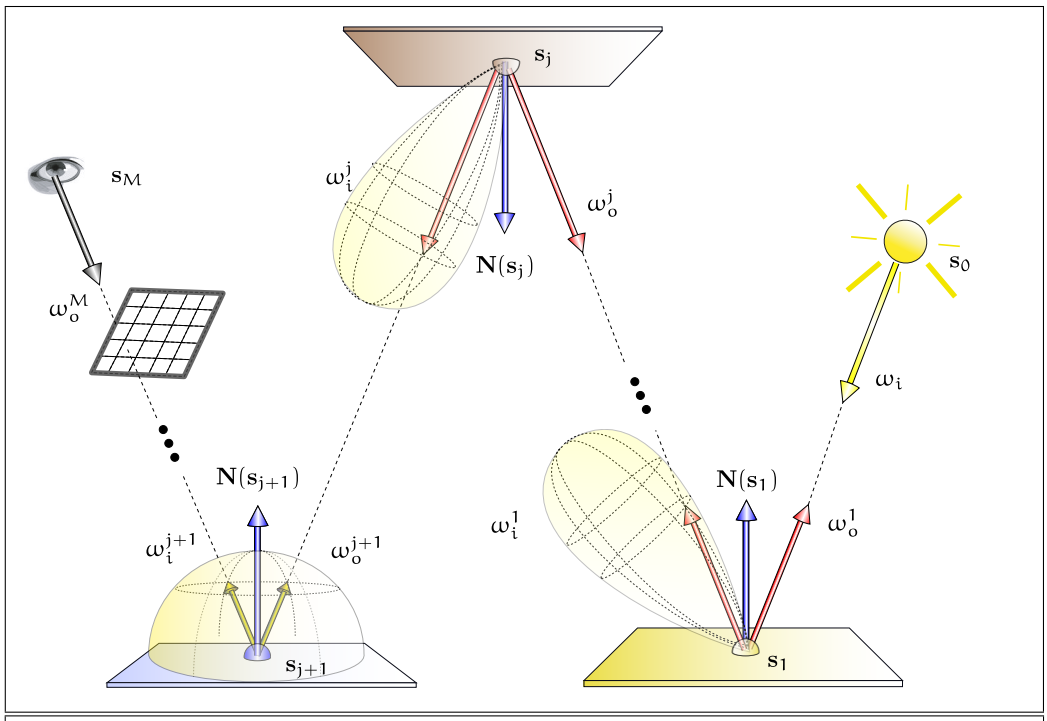


FIGURE 9.14: PURE-MONTE CARLO LIGHT TRACING. Starting from point s_0 at a light source, the algorithm generates a single primary ray and shoots this ray into the scene. Then, MCLT traces this ray through the scene until a pixel of the image plane is hit, the predefined default length of the random path is exceeded, or the ray does not hit any object in the scene, respectively the travel of the ray through the scene is stopped via Russian roulette.

REMARK 9.9 *Obviously, pure-Monte Carlo light tracing simulates random walks of characteristic* $\overrightarrow{L(D|G|S)^*E}$. *Thus, in principle, it provides us a complete solution of the stationary light transport equation in vacuum, but it suffers from the fact that the resulting images are usually rather noisy, see Figure 9.16.*

Heckberts Path Notation (655)

THE PROBLEM OF HITTING A SENSOR. Recall, pure-Monte Carlo light tracing is dual to pure-Monte Carlo path tracing, that is, the problem of hitting a light source in pMCPT is equivalent to the problem of hitting a sensor in pMCLT. Since sensors compared to other objects within a scene, are mostly small, the probability of hitting a sensor is even very small. This means, that due to the attenuation of light at object surfaces due to scattering effects, if any, only a vanishing small fraction of light is contributed to the final color of the pixel, thus, the resulting images are often dark and very noisy.

9.2.2 MONTE CARLO LIGHT TRACING WITH NEXT EVENT ESTIMATION

As the estimator $F_N^{\mathcal{M}_j,pMCLT}$ from Equation (9.8) shows, a random walk, generated by pure-Monte Carlo light tracing only contributes to the shading of a pixel, if it finally reaches the eye of an observer or a virtual camera within the scene. Except of the last node of a light path, pMCLT ignores the importance of a sensor. As the sensor also has a direct importance to all other nodes of a light path, Monte Carlo light tracing has to involve the sensor more strictly in its process of importance evaluation. This can be done by combining pure-Monte Carlo light tracing also with next event estimation.

The idea behind it: We extend pMCLT in such way, that the algorithm—during the travel of a ray through the scene—generates so-called *contribution rays* at the nodes of a light path, thus rays in direction to the sensor, computes the importance that directly influences these surface points and combines the corresponding amounts of importance with the indirect importance flowing along the path back to a light source. As we will see below, this leads to a powerful variant of pMCLT, called: *Monte Carlo light tracing with next event estimation*, see Figure 9.15.

MONTE CARLO LIGHT TRACING WITH NEXT EVENT ESTIMATION. As the importance equation is dual to the stationary light transport equation, Monte Carlo light tracing with next event estimation can easily be derived from the Monte Carlo path tracing with next event estimation algorithm, introduced in the last section. By exchanging the quantities radiance and importance, as well as the notions of the pixel with that of the light source, a primary estimator for MCLT with next event estimation can easily be derived from the estimator $F_1^{\mathcal{M}_j,MCPT}$ know from Equation (9.28). We leave the derivation of the estimator $F_1^{\mathcal{M}_j,MCLT}$ to the interested reader as an exercise.

REMARK 9.10 *Note, when estimating the SITEV, we must be care wether the chosen sample has already hit the sensor. Since this sample may be associated with a contribution ray, we must not take into account the contribution of that sample to the final importance, because it contributes perhaps already to the direct importance at this point.*

REMARK 9.11 *It is easily seen that pure-Monte Carlo light tracing, under the condition that a simple pinhole camera model is used, does not provides us with a reasonably well solution of the importance equation, since the probability that the corresponding random walk goes through the pinhole is equal zero.*

Due to the fact, that MCPT and MCLT solve the global illumination problem in different ways, we can also expect that images, rendered with these methods, have different properties. So, MCPT is the more efficient method for rendering images that only shows little sections of an illuminated scene, as the camera determines

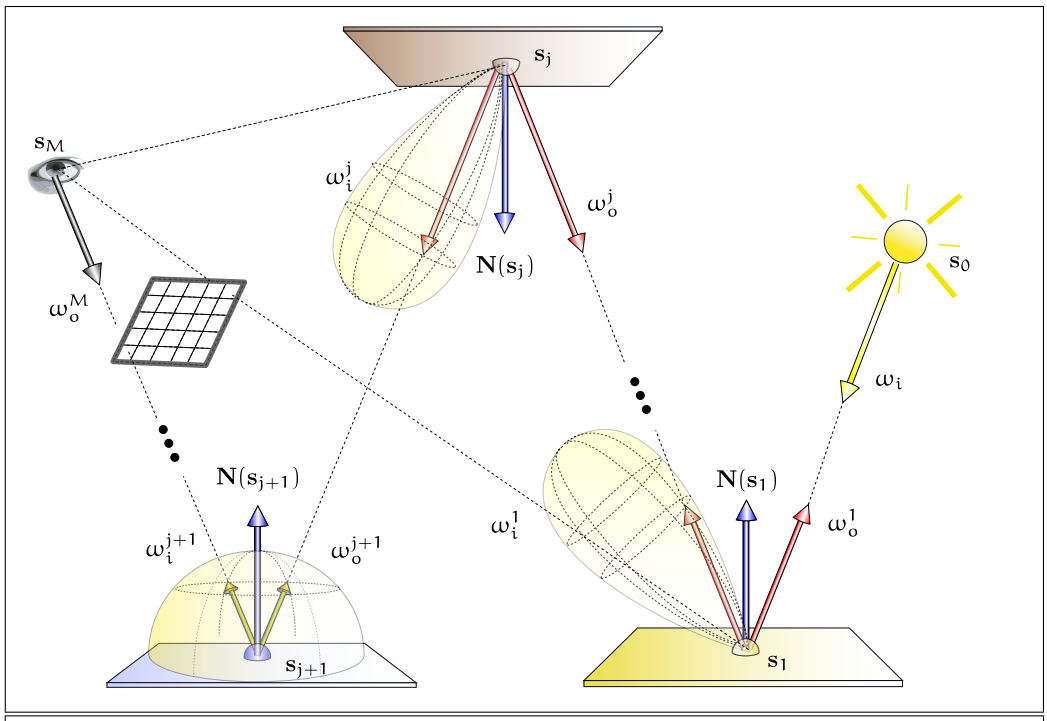


FIGURE 9.15: MONTE CARLO LIGHT TRACING WITH NEXT EVENT ESTIMATION. Compared to pMCLT, Monte Carlo light tracing generates at each hit point of a ray with an object of the scene a so-called contribution ray in direction to the sensor.

the view volume, and the algorithm generates random walks depending on the view volume. On the other hand, light tracing is the better method for simulating caustics.

REMARK 9.12 *As already mentioned above, Monte Carlo path tracing and Monte Carlo light tracing are dual algorithms. While in path tracing the light sources act as sources of radiance, in light tracing, the image plane serves as an emitter for importance. The passivity of the light sources in Monte Carlo path tracing confronts to the passivity of the image plane in Monte Carlo light tracing. Furthermore, the concept of the shadow ray is dual to the concept of the contribution ray. Thus, all variance reduction techniques—as the computation of flux via direct or indirect importance or the embedding of various sampling strategies for approximating a solution of the importance equation—presented in the previous section for path tracing, can also be used in an algorithm based on light tracing.*

The images in Figure 9.16 are rendered with Monte Carlo light tracing. They show the Cornell box consisting of only diffuse surfaces, illuminated by a single spherical light

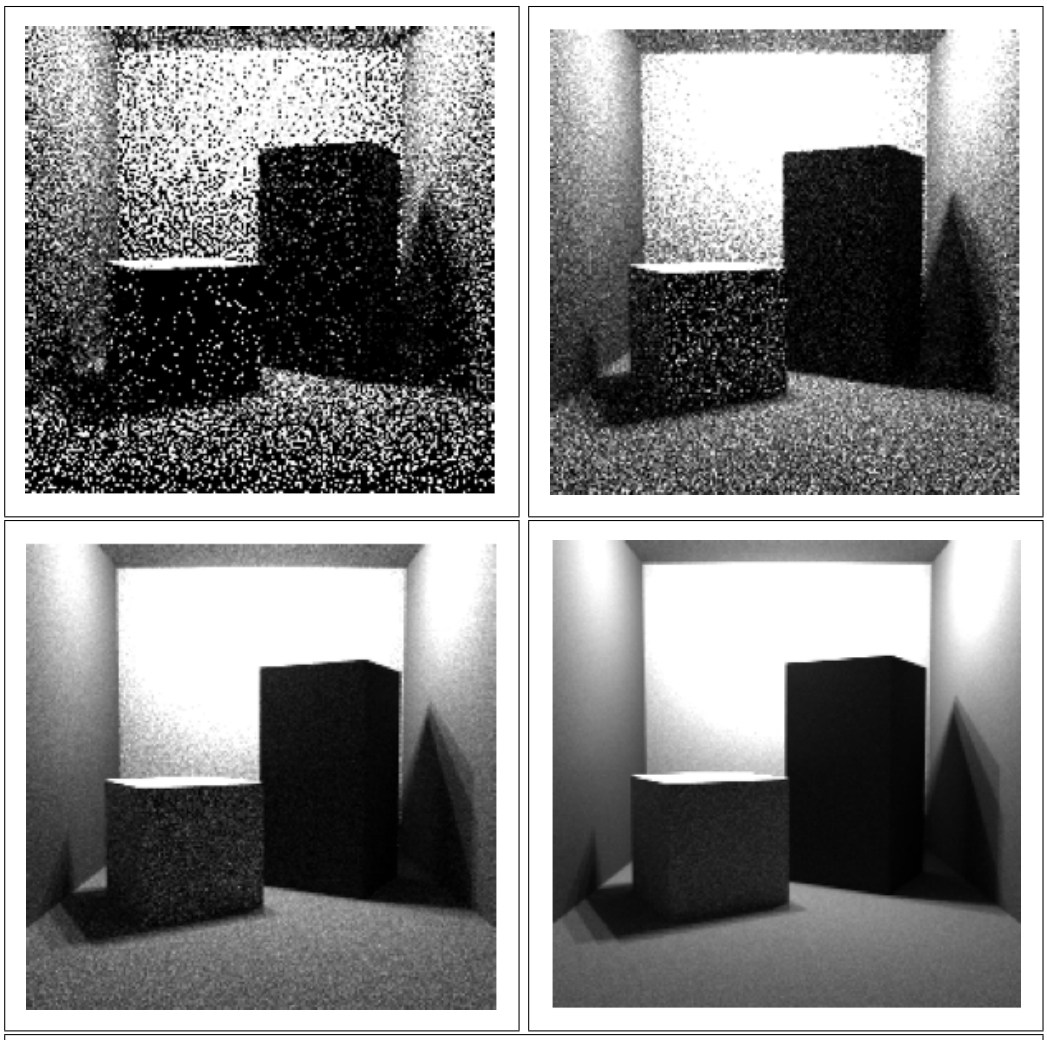
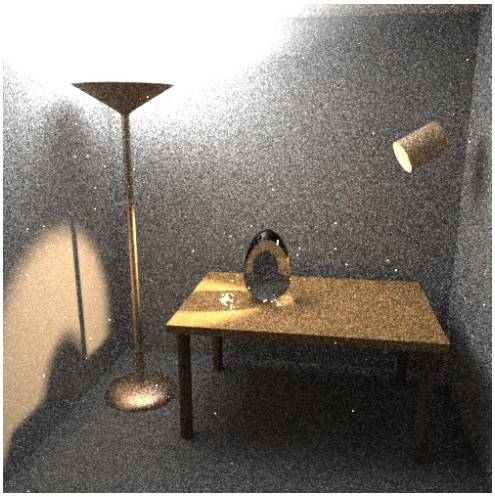


FIGURE 9.16: IMAGES RENDERED WITH PURE-MONTE CARLO LIGHT TRACING. The images are rendered via pure-Monte Carlo light tracing using $100,000$, $1,000,000$, $10,000,000$ and $100,000,000$ rays. They show the Cornell box consisting of only diffuse surfaces, illuminated by a single spherical light source. The first two images are mostly black, with some very brightly pixels. Since sensors compared to other objects within a scene, are mostly small, the probability of hitting a sensor is also very small. This means, that due to the attenuation of light at object surfaces due to scattering effects, if any, only a vanishing small fraction of light is contributed to the final color of the pixel. That is, the resulting images are often dark and very noisy. Using more light-paths leads to better the images. Image courtesy of Philippe Dutré, Department of Computer Science, K. U. Leuven.



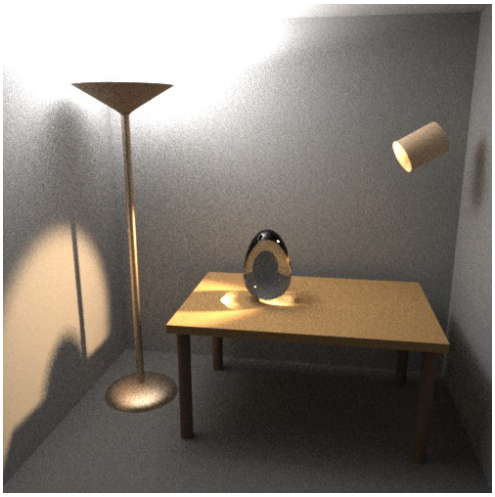


FIGURE 9.17: A COMPARISON OF BIDIRECTIONAL PATH TRACING AND MONTE CARLO PATH TRACING. The scene contains a spotlight, a floor lamp, a table, and a large glass egg. The left image is 500 by 500 and was rendered via Monte Carlo path tracing with 56 samples per pixel. The image on the right was computed with bidirectional path tracing, using the power heuristic with $\beta = 2$ to combine the samples for each path length. The image is also 500 by 500 with 25 samples per pixel. Both images are rendered in the same amount of time. Image courtesy of Eric Veach.

source. The first two images are mostly black. Note: Although different in quality, even the first two images, are, due to the bright pixels, perfectly valid secondary estimators for the correct solution of the SITEV. That variance can be reduced by taking more samples, this can be observed in the lower two images. Here, more random walks are generated, resulting in images with less noise. But also these results are still noisy, and thus not satisfactory.

9.3 BIDIRECTIONAL PATH TRACING

Let us consider the scene in the images of Figure 9.17 where a small area on the ceiling of a room is illuminated by a floor lamp and a spotlight illuminates a small region at the left wall of the scene, consisting of diffuse, gloss, and pure specular surfaces. The rest of the scene is illuminated indirectly by light reflecting from these areas. The image rendered with Monte Carlo path tracing is very noisy, which is not surprisingly due to the manner a path tracer works. Obviously, the probability that a path, starting from the camera, hits the illuminated regions before a shadow ray is generated, is very small. Thus, all paths, except of a small number, namely those that hit the illuminated regions, does not

contribute to the shading of a pixel. As a result the image has a hight variance. Scenes with similar difficult lighting conditions can be handled more effectively by a rendering method called *bidirectional path tracing*, or as we will say for short BDPT.

Bidirectional path tracing was at first introduced in [119, Lafortune & Willems 1993] and a year later independently developed in [222, Veach & Guibas 1995]. Although both variants of bidirectional path tracing lead to similar results, they are based on two different mathematical frameworks. While Eric Lafortune's variant of bidirectional path tracing starts from the formulation of the global reflectance distribution function, Eric Veach's variant of bidirectional path tracing is based on the path integral formulation of the light transport problem. Both algorithms have its origin in [6, Arvo 1986], where the rendering of caustics was described by means of a ray tracing algorithm, that takes its starting point in one of the light sources of the environment but not in the eye of the observer or a virtual camera, as done with path tracing.

GRDF (473)
Path Integral Formulation (466)

We will now present bidirectional path tracing in the variant of Eric Veach and Leonidas Guibas, that is, based on the formulation of the measurement equation as a path integral over all paths generated on the surfaces within a scene. So, we will show how the basic bidirectional path tracing algorithm constructs light and eye paths, and how the algorithm can uses these paths to generate so-called *transport paths* between an emitter and a sensor for rendering. We discuss the mathematical framework behind this process and analyze the technique of transport path construction with respect to their usage as a Monte Carlo rendering algorithm. To improve the performance of the basic algorithm with respect to the transport paths and the contributions of these paths to an estimator, we also discuss a refinement of the basic algorithm: the idea of using a family of different sampling techniques for transport paths, and combining them using the principle of multiple importance sampling.

Measurement Equation (416)
Section 9.3.1
Section 9.3.2

REMARK 9.13 *A detailed description of Eric Lafortunes variant of bidirectional path tracing can be found in [119, Lafortune & Willems 1993] and [50, Dutré & al. 2003].*

9.3.1 GENERATING AND ESTIMATING TRANSPORT PATHS

As we know from the last two sections, some light phenomena can more easily be simulated via path tracing while others can rather be represented via tracing rays from light sources. The idea behind bidirectional path tracing is to exploit this fact. Thus, a basic bidirectional path tracing algorithm combines Monte Carlo path tracing with Monte Carlo light tracing, that is, instead of just tracing a random walk starting at the eye, a so-called *eye path*, additionally a *light path*, starting at the eye, is also traced into the scene. Both paths are then joined together at their ends resulting in a set of transport paths.

Let us now discuss in detail how BDPT generates transport paths, and how these paths can be used in a Monte Carlo estimator for evaluating the path integral, that de-

scribes the light transport problem in a vacuum.

GENERATING TRANSPORT PATHS. At first, bidirectional path tracing generates, in a similar manner to Monte Carlo light tracing, a light subpath

$$\bar{\mathbf{y}} = \mathbf{y}_0 \dots \mathbf{y}_{n_L-1} \tag{9.36}$$

with n_L vertices, by choosing a random point $\mathbf{y}_0$ on a light source and finding the points $\mathbf{y}_{j+1}, 0 \leq j \leq n_L - 2$ via casting a ray randomly with respect to the BSDF from $\mathbf{y}_j$. By a similar process, the algorithm then constructs an eye subpath

$$\bar{\mathbf{z}} = \mathbf{z}_{n_E-1} \dots \mathbf{z}_0 \tag{9.37}$$

with n_E vertices starting from a random point $\mathbf{z}_0$ on the camera lens, where the length of each subpath is determined by a form of Russian roulette.

Russian Roulette (200)

Via the visibility function $\mathcal{V}$ the algorithm then determines the visibility of the two endpoints $\mathbf{y}_{n_L-1}$ and $\mathbf{z}_{n_E-1}$ and concatenates these two vertices if they are visible to each other. Thus, we get a complete transport path $\bar{\mathbf{x}}$ of length $n_L + n_E - 1$ defined by:

$\mathcal{V}$ (45)

$$\begin{aligned} \bar{\mathbf{x}} &\overset{def}{=} \bar{\mathbf{y}}\,\bar{\mathbf{z}} && (9.38)\\ &\overset{(9.36),(9.37)}{=} \mathbf{y}_0 \dots \mathbf{y}_{n_L-1}\,\mathbf{z}_{n_E-1} \dots \mathbf{z}_0 && (9.39)\\ &= \mathbf{x}_0 \dots \mathbf{x}_{n_L+n_E-1} && (9.40) \end{aligned}$$

see Figure 9.18. In this case, the vertices $\mathbf{y}_{n_L-1}$ and $\mathbf{z}_{n_E-1}$ are called the *connecting vertices*, and the edge between them is denoted as the *connecting edge*. If the two points are not visible to each other, or if the BSDF at the connecting vertices does not scatter light towards the other, then we define the contribution for that path to be zero.

Obviously, paths of this type can be used to estimate the path integral

Path Integral (466)

$$\mathcal{M}_j = \int_{\mathbf{P}^\infty} f_j(\bar{\mathbf{x}})\, d\mu_\infty(\bar{\mathbf{x}}), \tag{9.41}$$

representing the stationary light transport in a vacuum. Due to Definition 5.19, the path integral can be written as infinite sum of integrals over paths of all finite length, thus,

$$\begin{aligned} \mathcal{M}_j &= \int_{\mathbf{P}^\infty} f_j(\bar{\mathbf{x}})\, d\mu_\infty(\bar{\mathbf{x}}) && (9.42)\\ &= \sum_{k=1}^{\infty} \int_{\mathbf{P}^k} f_j(\bar{\mathbf{x}}_k)\, d\mu_k(\bar{\mathbf{x}}_k), && (9.43) \end{aligned}$$

where $\bar{\mathbf{x}}_k \in \mathbf{P}^k$ is a transport path of length k.

Let $\bar{\mathbf{x}}_k$ be a transport path of length $n_L + n_E - 1$ generated with bidirectional path tracing, then, a secondary Monte Carlo estimator based on the path integral of the light

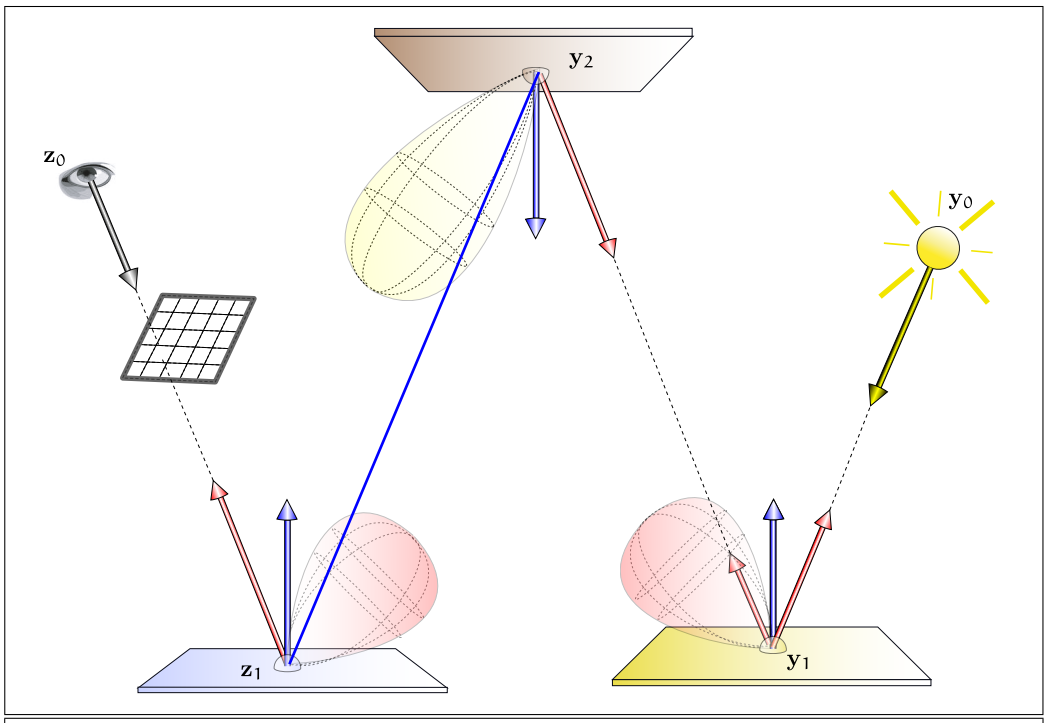


FIGURE 9.18: A TRANSPORT PATH FROM A LIGHT SOURCE TO THE CAMERA LENS.
A light path $\bar{y} = y_0 y_1 y_2$ of length 2 and an eye path $\bar{z} = z_0 z_1$ of length 1 constructed via bidirectional path tracing. Both paths are joined together at their ends, y_2 and z_1, to a transport path $\bar{x} = y_0 y_1 y_2 z_1 z_0$ of length 4.

transport and the basic algorithm of BDPT, denoted as bBDPT, is given by:

$$F_N^{\mathcal{M}_j,\mathrm{bBDPT}} \frac{1}{N} \sum_{k=1}^{N} \frac{f_j(\bar{x}_k)}{p(\bar{x}_k)}, \tag{9.44}$$

f_j (463) where f_j is the measurement contribution function, $\bar{x}_k$ are N paths from $\mathbf{P}^\infty$, and p is the
$\mathbf{P}^\infty$ (461) probability density function with which the paths $\bar{x}_k$ are sampled.
PDF (176)

REMARK 9.14 *In [222, Veach & Guibas 1995] the quantities $\frac{f_j(\bar{x}_k)}{p(\bar{x}_k)}$ in the above estimator are denoted as the unweighted contributions. In the following section we will endow the unweighted contributions with weighting functions w_k. Using these*
Multiple Sample Estimator (592) *weighted contributions, we are then able to construct a multiple sample estimator for the path integral formulation of the light transport problem.*

Obviously, the evaluation of the estimator $F_N^{\mathcal{M}_j,\mathrm{bBDPT}}$ requires the computation of the probabilities p with which the paths $\bar{x}_k$ are sampled and the evaluation of the mea-

surement contribution function f_j applied to the transport paths $\overline{\mathbf{x}}_k$. Let us now show, how bidirectional path tracing efficiently mastered this task.

ESTIMATING TRANSPORT PATHS. As already mentioned above, bidirectional path tracing generates the vertices $\mathbf{y}_0$ and $\mathbf{z}_0$ of a transport path directly on the surface of a light source, respectively the lens of the camera, while all other nodes of the path are generated via tracing a ray through the scene.

Let $p_{\mu^2}(\mathbf{y}_0)$ and $p_{\mu^2}(\mathbf{z}_0)$ denote the PDFs for sampling the first vertices of the corresponding subpaths of the transport $\overline{\mathbf{x}}_k$, measured with respect to the Lebesgue area measure μ^2. Bidirectional path tracing then samples the successor $\mathbf{y}_j$ of a path node $\mathbf{y}_{j-1}$, respectively the successor $\mathbf{z}_j$ of a path node $\mathbf{z}_{j-1}$, of the transport path

PDF (176)

μ^2 (82)

$$\overline{\mathbf{x}}_k = \underbrace{\mathbf{y}_0 \dots \mathbf{y}_{s-1}}_{\overline{\mathbf{y}}} \underbrace{\mathbf{z}_{t-1} \dots \mathbf{z}_0}_{\overline{\mathbf{z}}} \tag{9.45}$$

by choosing a direction and casting a ray from the current subpath node to the new sampled vertex with respect to the projected solid angle measure $\sigma^\perp$.

$\sigma^\perp$ (88)

Then, the density for sampling the $(j+1)^{\text{th}}$ vertex of one of the subpaths $\overline{\mathbf{y}}$ or $\overline{\mathbf{z}}$ is given by the conditional density that $\mathbf{y}_j$, respectively $\mathbf{z}_j$, is chosen given $\mathbf{y}_{j-1}$, respectively $\mathbf{z}_{j-1}$, multiplied with the PDF for generating the associated subpath $\mathbf{y}_0 \dots \mathbf{y}_{j-1}$ or $\mathbf{z}_0 \dots \mathbf{z}_{j-1}$, that is,

$$p(\overline{\mathbf{y}}) = p_{\mu^2}(\mathbf{y}_0) \prod_{j=1}^{n_L-1} p_{\mu^2}(\mathbf{y}_j|\mathbf{y}_{j-1}) \tag{9.46}$$

$$= p_{\mu^2}(\mathbf{y}_0) \prod_{j=1}^{n_L-1} \left(p_{\sigma^\perp}(\mathbf{y}_{j-1} \to \mathbf{y}_j|\mathbf{y}_{j-2} \to \mathbf{y}_{j-1})\, G(\mathbf{y}_{j-1} \leftrightarrow \mathbf{y}_j)\right) \tag{9.47}$$

and

$$p(\overline{\mathbf{z}}) = p_{\mu^2}(\mathbf{z}_0) \prod_{j=1}^{n_E-1} p_{\mu^2}(\mathbf{z}_j|\mathbf{z}_{j-1}) \tag{9.48}$$

$$= \prod_{j=1}^{n_E-1} \left(p_{\sigma^\perp}(\mathbf{z}_{j-1} \to \mathbf{z}_j|\mathbf{z}_{j-2} \to \mathbf{z}_{j-1})\, G(\mathbf{z}_{j-1} \leftrightarrow \mathbf{z}_j)\right) \cdot p_{\mu^2}(\mathbf{z}_0), \tag{9.49}$$

where we have expressed the PDF p_{μ^2} in terms of the probability density function $p_{\sigma^\perp}$

according to

$$
\begin{aligned}
p_{\mu^2}(\mathbf{s}_j) &= \frac{d\mathbb{P}_{\mu^2}}{d\mu^2}(\mathbf{s}_j) && (9.50)\\
&= \frac{d\mathbb{P}_{\sigma}}{d\sigma^{\perp}}(\omega_i^{j-1})\frac{d\sigma^{\perp}(\omega_i^{j-1})}{d\mu^2(\mathbf{s}_j)} && (9.51)\\
&\overset{(2.196)}{=} p_{\sigma^{\perp}}(\omega_i^{j-1})\frac{d\mu^2(\mathbf{s}_j)\left|\cos\theta_i^{j-1}\cos\theta_o^{j}\right|}{d\mu^2(\mathbf{s}_j)\,\|\mathbf{s}_j-\mathbf{s}_{j-1}\|_2^2} && (9.52)\\
&= p_{\sigma^{\perp}}(\omega_i^{j-1})\frac{\left|\cos\theta_i^{j-1}\cos\theta_o^{j}\right|}{\|\mathbf{s}_j-\mathbf{s}_{j-1}\|_2^2} && (9.53)\\
&= p_{\sigma^{\perp}}(\omega_i^{j-1})G(\mathbf{s}_{j-1}\leftrightarrow\mathbf{s}_j), && (9.54)
\end{aligned}
$$

where $\mathbf{s}_j$ can be identified as the path nodes $\mathbf{y}_{j-1}$ or $\mathbf{z}_{j-1}$ and the direction $\omega_i^{j-1}=\mathbf{y}_{i-1}\rightarrow\mathbf{y}_i$ respectively $\omega_i^{j-1}=\mathbf{z}_{i-1}\rightarrow\mathbf{z}_i$. An illustration for computing the path probability of a transport path is shown in Figure 9.19.

As the current endpoints $\mathbf{y}_{n_L-1}$ and $\mathbf{z}_{n_E-1}$ are connected if they are visible to each other, the probability for generating the connecting edge $\mathbf{y}_{n_L-1}\mathbf{z}_{n_E-1}$ is one, thus, the PDF for generating the path $\overline{\mathbf{x}}_k$ is then given by the product of the densities for generating the subpaths $\overline{\mathbf{y}}$ and $\overline{\mathbf{z}}$, and the probability for generating the connecting edge $\mathbf{y}_{n_L-1}\mathbf{z}_{n_E-1}$. That is, the probability density function for sampling $\overline{\mathbf{x}}_k$ is given by:

$$
\begin{aligned}
p(\overline{\mathbf{x}}_k) &= p(\overline{\mathbf{y}})\cdot\underbrace{\mathbb{P}_{\sigma^{\perp}}(\mathbf{y}_{n_L-1}\rightarrow\mathbf{z}_{n_E-1})}_{=1}\cdot p(\overline{\mathbf{z}}) && (9.55)\\
&= p_{\mu^2}(\mathbf{y}_0)\prod_{j=1}^{n_L-1}\left(p_{\sigma^{\perp}}(\mathbf{y}_{j-1}\rightarrow\mathbf{y}_j|\mathbf{y}_{j-2}\rightarrow\mathbf{y}_{j-1})\,G(\mathbf{y}_{j-1}\leftrightarrow\mathbf{y}_j)\right)\cdot && (9.56)\\
&\quad\prod_{j=1}^{n_E-1}\left(p_{\sigma^{\perp}}(\mathbf{z}_{j-1}\rightarrow\mathbf{z}_j|\mathbf{z}_{j-2}\rightarrow\mathbf{z}_{j-1})\,G(\mathbf{z}_{j-1}\leftrightarrow\mathbf{z}_j)\right)\cdot p_{\mu^2}(\mathbf{z}_0).
\end{aligned}
$$

EXAMPLE 9.3 *Let us consider the transport path generated via bidirectional path tracing passing through an ideal diffuse scene with light subpath $\overline{\mathbf{y}}=\mathbf{y}_0\mathbf{y}_1\mathbf{y}_2$ and eye subpath $\overline{\mathbf{z}}=\mathbf{z}_1\mathbf{z}_0$ from Figure 9.20. Then, $\overline{\mathbf{x}}_4$ is composed of a light path of length two and an eye path of length one, that is, it is of the form*

$$\overline{\mathbf{x}}_4=\mathbf{y}_0\mathbf{y}_1\mathbf{y}_2\mathbf{z}_1\mathbf{z}_0. \tag{9.57}$$

As, the starting nodes $\mathbf{y}_0$ and $\mathbf{z}_0$ are sampled with respect to the Lebesgue area measure on a light source ☼ respectively, at the lens ⊙ of the camera, and the points $\mathbf{y}_1,\mathbf{y}_2$ and $\mathbf{z}_1$ are sampled with respect to projected solid angle, the probability with

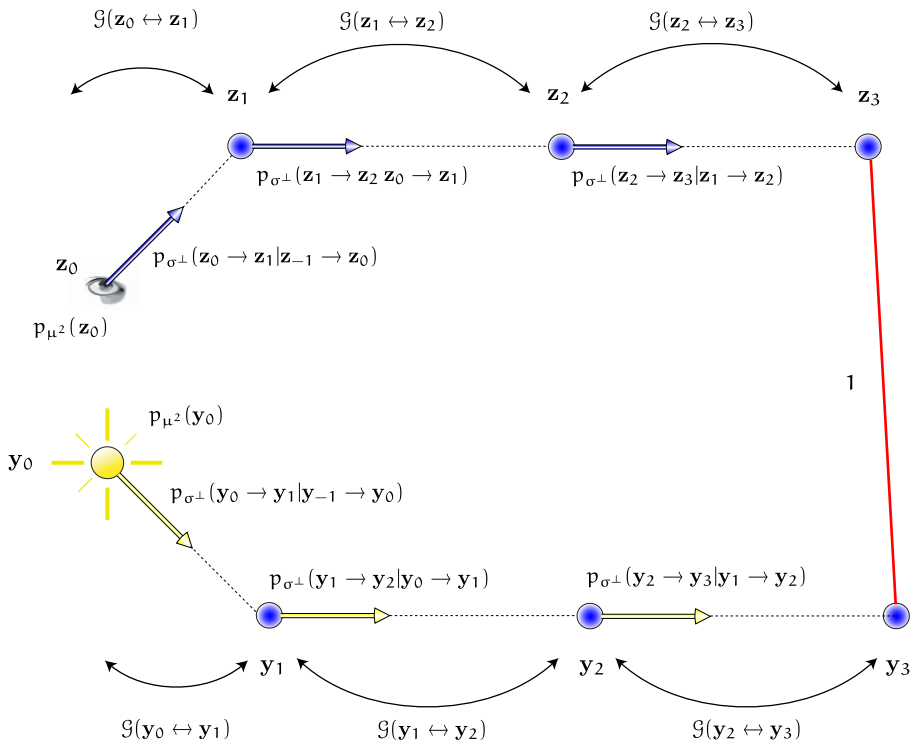


FIGURE 9.19: THE PROBABILITY DENSITY FUNCTION FOR GENERATING A TRANSPORT PATH. Shown is a transport $\overline{\mathbf{x}}_7 = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2\mathbf{y}_3\mathbf{z}_3\mathbf{z}_2\mathbf{z}_1\mathbf{z}_0$ path of length 7. The probability of computing the light path $\overline{\mathbf{y}} = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2\mathbf{y}_3$ is given by the product of the probabilities for sampling the point $\mathbf{y}_0$ and the direction ω_i, ω_i^1 and ω_i^2. The same holds for the eye path $\overline{\mathbf{x}} = \mathbf{x}_0\mathbf{x}_1\mathbf{x}_2\mathbf{z}_3$. The probability for computing the connecting edge $\mathbf{y}_3 \leftrightarrow \mathbf{x}_3$ is one.

which BDPT generates the corresponding transport path is given by:

$$p(\overline{\mathbf{y}}) = \underbrace{p_{\mu^2}(\mathbf{y}_0)}_{\frac{1}{\mu^2(\text{☼})}} \cdot \underbrace{p_{\sigma^\perp}(\mathbf{y}_0 \to \mathbf{y}_1 | \mathbf{y}_0)}_{\frac{|\cos\theta_i^0|}{\pi}} \mathcal{G}(\mathbf{y}_0 \leftrightarrow \mathbf{y}_1) \underbrace{p_{\sigma^\perp}(\mathbf{y}_1 \to \mathbf{y}_2 | \mathbf{y}_0 \to \mathbf{y}_1)}_{\frac{|\cos\theta_i^1|}{\pi}} \mathcal{G}(\mathbf{y}_1 \leftrightarrow \mathbf{y}_2) \tag{9.58}$$

$$= \frac{1}{\mu^2(\text{☼})} \cdot \frac{|\cos\theta_i^0|}{\pi} \frac{|\cos\theta_i^0 \cos\theta_o^1|}{\|\mathbf{y}_0 - \mathbf{y}_1\|_2^2} \frac{|\cos\theta_i^1|}{\pi} \frac{|\cos\theta_i^1 \cos\theta_o^2|}{\|\mathbf{y}_1 - \mathbf{y}_2\|_2^2} \tag{9.59}$$

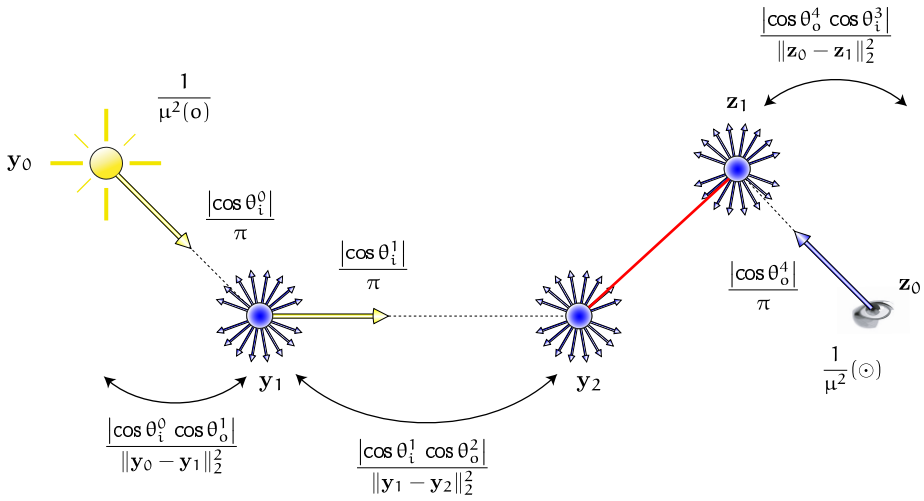


FIGURE 9.20: SAMPLING A TRANSPORT PATH WITH BIDIRECTIONAL PATH TRACING. The path $\overline{x} = y_0y_1y_2z_1z_0$ is generated via concatenating the light path $\overline{y} = y_0y_1y_2$ and the eye path $\overline{z} = z_1z_0$. The subpaths theirselves are generated by sampling the starting nodes y_0 and z_0 on the corresponding surfaces, and the path segments are generated via sampling a direction with respect to the BSDF valid at the associated surface.

as well as:

$$p(\overline{z}) = \underbrace{p_{\sigma^\perp}(z_0 \to z_1|z_0)}_{\frac{|\cos\theta_o^3|}{\pi}} G(z_0 \leftrightarrow z_1) \cdot \underbrace{p_{\mu^2}(z_0)}_{\frac{1}{\mu^2(\odot)}} \tag{9.60}$$

$$= \frac{|\cos\theta_o^4|}{\pi} \frac{|\cos\theta_o^4\ \cos\theta_i^3|}{\|z_0 - z_1\|_2^2} \cdot \frac{1}{\mu^2(\odot)}, \tag{9.61}$$

where we have sampled the light source as well as the lens with respect to their surface areas. That is, BDPT generates the path $\overline{x}_k$ with probability

$$p(\overline{x}_k) = p(\overline{y}) \cdot p(\overline{z}) \tag{9.62}$$

$$= \frac{1}{\mu^2(\sun)} \cdot \frac{|\cos\theta_i^0|}{\pi} \frac{|\cos\theta_i^0\ \cos\theta_o^1|}{\|y_0 - y_1\|_2^2} \frac{|\cos\theta_i^1|}{\pi} \frac{|\cos\theta_i^1\ \cos\theta_o^2|}{\|y_1 - y_2\|_2^2} \cdot \frac{|\cos\theta_o^4|}{\pi} \frac{|\cos\theta_o^4\ \cos\theta_i^3|}{\|z_0 - z_1\|_2^2} \cdot \frac{1}{\mu^2}(\odot). \tag{9.63}$$

For evaluating the estimator $F_N^{\mathcal{M}_j,\mathrm{bDPT}}$, apart from computing the densities $p(\overline{x}_k)$ for generating the transport paths $\overline{x}_k$, the measurement contribution function f_j has also

f_j (463)

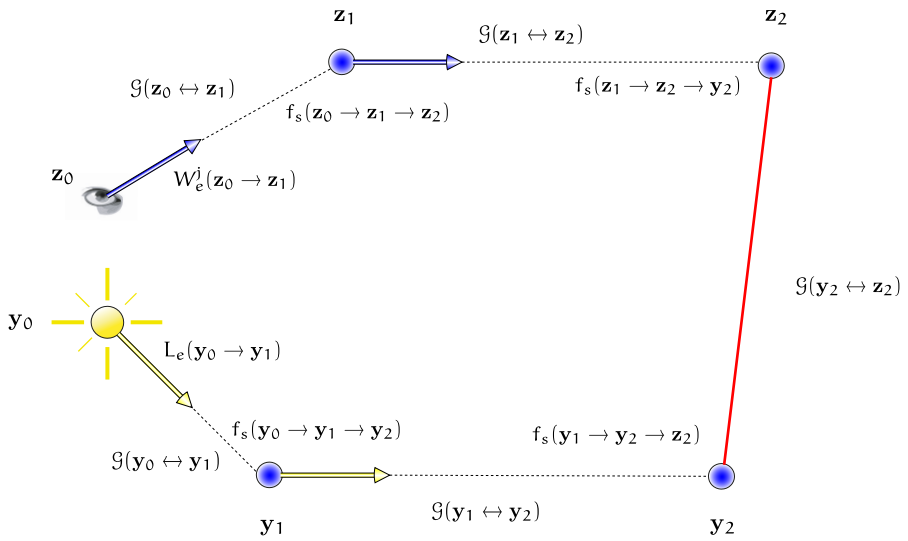


FIGURE 9.21: COMPUTING THE MEASUREMENT CONTRIBUTION FUNCTION FOR A TRANSPORT PATH. The transport path $\overline{\mathbf{x}}_5 = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2\mathbf{z}_2\mathbf{z}_1\mathbf{z}_0$ is composed of a light and an eye path of length 2. The measurement contribution function for the path $\overline{\mathbf{x}}_5$ can be computed via the product of the geometry terms, the BSDFs as well as the emitted radiance and importance along the transport path.

to be evaluated for the transport path $\overline{\mathbf{x}}_k$. To do this in an efficient way, let us consider f_j a little bit more closely. Obviously, it holds:

$$\begin{aligned} f_j(\overline{\mathbf{x}}_k) \quad = \quad & L_e(\mathbf{y}_0 \to \mathbf{y}_1) \cdot \mathcal{G}(\mathbf{y}_0 \leftrightarrow \mathbf{y}_1) \cdot \prod_{j=1}^{n_L-2} \left(f_s(\mathbf{y}_{j-1} \to \mathbf{y}_j \to \mathbf{y}_{j+1}) \cdot \mathcal{G}(\mathbf{y}_j \leftrightarrow \mathbf{y}_{j+1}) \right) \cdot \\ & c_k \cdot \\ & \prod_{j=1}^{n_E-2} \left(\mathcal{G}(\mathbf{z}_{j+1} \leftrightarrow \mathbf{z}_j) f_s(\mathbf{z}_{j+1} \to \mathbf{z}_j \to \mathbf{z}_{j-1}) \right) \cdot \mathcal{G}(\mathbf{z}_0 \leftrightarrow \mathbf{z}_1) \cdot W_e^j(\mathbf{z}_0 \to \mathbf{z}_1), \end{aligned} \tag{9.64}$$

where c_k depends on the connecting edge $\mathbf{y}_{n_L-1}\mathbf{z}_{n_E-1}$,

$$\begin{aligned} c_k \quad \stackrel{\text{def}}{=} \quad & f_s(\mathbf{y}_{n_L-2} \to \mathbf{y}_{n_L-1} \to \mathbf{z}_{n_E-1}) \cdot \\ & \mathcal{G}(\mathbf{y}_{n_L-1} \leftrightarrow \mathbf{z}_{n_E-1}) \cdot \\ & f_s(\mathbf{z}_{n_E-2} \to \mathbf{z}_{n_E-1} \to \mathbf{y}_{n_L-1}), \end{aligned} \tag{9.65}$$

see Figure 9.21.

Now, evaluating the emitted radiance $L_e(\mathbf{y}_0 \to \mathbf{y}_1)$ requires to sample the point $\mathbf{y}_0$ on a light source followed by a directional sampling, that is, the emitted radiance $L_e(\mathbf{y}_0 \to \mathbf{y}_1)$ can be split into a product

$$L_e(\mathbf{y}_0 \to \mathbf{y}_1) = L_e^0(\mathbf{y}_0) L_e^1(\mathbf{y}_0 \to \mathbf{y}_1). \tag{9.66}$$

Similar to this decomposition we can also split the quantity $W_e(\mathbf{y}_0 \to \mathbf{y}_1)$ into the product

$$W_e(\mathbf{z}_0 \to \mathbf{z}_1) = W_e^0(\mathbf{z}_0) W_e^1(\mathbf{z}_0 \to \mathbf{z}_1). \tag{9.67}$$

Using the conventions

$$L_e^1(\mathbf{y}_0 \to \mathbf{y}_1) \stackrel{def}{=} f_s(\mathbf{y}_{-1} \to \mathbf{y}_0 \to \mathbf{y}_1) \tag{9.68}$$

$$W_e^1(\mathbf{z}_0 \to \mathbf{z}_1) \stackrel{def}{=} f_s(\mathbf{z}_{-1} \to \mathbf{z}_0 \to \mathbf{z}_1) \tag{9.69}$$

and the probabilities for generating the paths $\overline{\mathbf{x}}_k$ from above, then we get for the unweighted contributions C_k^* of the paths $\overline{\mathbf{x}}_k$ the following formula:

$$C_k^* \stackrel{def}{=} \frac{f_j(\overline{\mathbf{x}}_k)}{p(\overline{\mathbf{x}}_k)} \tag{9.70}$$

$$\begin{aligned} = \; & \frac{L_e^0(\mathbf{y}_0)}{p_{\mu^2}(\mathbf{y}_0)} \frac{f_s(\mathbf{y}_{-1} \to \mathbf{y}_0 \to \mathbf{y}_1)}{p_{\sigma^\perp}(\mathbf{y}_0 \to \mathbf{y}_1 | \mathbf{y}_0 \to \mathbf{y}_{-1}) \mathcal{G}(\mathbf{y}_0 \leftrightarrow \mathbf{y}_1)} \cdot \mathcal{G}(\mathbf{y}_0 \leftrightarrow \mathbf{y}_1) \cdot \\ & \frac{\prod_{j=1}^{n_L-2} \left(f_s(\mathbf{y}_{j-1} \to \mathbf{y}_j \to \mathbf{y}_{j+1}) \, \mathcal{G}(\mathbf{y}_j \leftrightarrow \mathbf{y}_{j+1}) \right)}{\prod_{j=1}^{n_L-2} \left(p_{\sigma^\perp}(\mathbf{y}_j \to \mathbf{y}_{j+1} | \mathbf{y}_{j-1} \to \mathbf{y}_j) \mathcal{G}(\mathbf{y}_j \leftrightarrow \mathbf{y}_{j+1}) \right)} \cdot \end{aligned} \tag{9.71}$$

$$\begin{aligned} & c_k \cdot \\ & \frac{\prod_{j=1}^{n_E-2} \left(\mathcal{G}(\mathbf{z}_{j+1} \leftrightarrow \mathbf{z}_j) \, f_s(\mathbf{z}_{j+1} \to \mathbf{z}_j \to \mathbf{z}_{j-1}) \right)}{\prod_{j=1}^{n_E-2} \left(p_{\sigma^\perp}(\mathbf{z}_j \to \mathbf{z}_{j+1} | \mathbf{z}_{j-1} \to \mathbf{z}_j) \, \mathcal{G}(\mathbf{z}_j \leftrightarrow \mathbf{z}_{j+1}) \right)} \cdot \\ & \frac{W_e^0(\mathbf{z}_0)}{p_{\mu^2}(\mathbf{z}_0)} \frac{f_s(\mathbf{z}_{-1} \to \mathbf{z}_0 \to \mathbf{z}_1)}{p_{\sigma^\perp}(\mathbf{z}_0 \to \mathbf{z}_1 | \mathbf{z}_0 \to \mathbf{z}_{-1}) \mathcal{G}(\mathbf{z}_0 \leftrightarrow \mathbf{z}_1)} \cdot \mathcal{G}(\mathbf{z}_0 \leftrightarrow \mathbf{z}_1) \cdot \end{aligned}$$

Evidently, the geometry terms occurring in the nominator as well as in the denominator can be canceled, that is, the formula for the unweighted contributions C_k^* can be simplified to:

$$\begin{aligned} C_k^* \stackrel{def}{=} & \frac{L_e^0(\mathbf{y}_0)}{p_{\mu^2}(\mathbf{y}_0)} \cdot \prod_{j=0}^{n_L-2} \frac{f_s(\mathbf{y}_{j-1} \to \mathbf{y}_j \to \mathbf{y}_{j+1})}{p_{\sigma^\perp}(\mathbf{y}_j \to \mathbf{y}_{j+1} | \mathbf{y}_{j-1} \to \mathbf{y}_j)} \cdot \\ & c_k \cdot \\ & \prod_{j=0}^{n_E-2} \frac{f_s(\mathbf{z}_{j+1} \to \mathbf{z}_j \to \mathbf{z}_{j-1})}{p_{\sigma^\perp}(\mathbf{z}_j \to \mathbf{z}_{j+1} | \mathbf{z}_{j-1} \to \mathbf{z}_j)} \cdot \frac{W_e^0(\mathbf{z}_0)}{p_{\mu^2}(\mathbf{z}_0)}. \end{aligned} \tag{9.72}$$

Using these formulae within the estimator $F_N^{\mathcal{M}_j,\text{bBDPT}}$ then we get

$$F_N^{\mathcal{M}_j,\text{bBDPT}} = \frac{1}{N}\sum_{k=1}^{N}\frac{f_j(\overline{x}_k)}{p(\overline{x}_k)} \tag{9.73}$$

$$= \frac{1}{N}\sum_{k=1}^{N} C_k^* \tag{9.74}$$

$$= \frac{1}{N}\sum_{k=1}^{N}\left(\frac{L_e^0(y_0)}{p_{\mu^2}(y_0)}\cdot\prod_{j=0}^{n_L-2}\frac{f_s(y_{j-1}\to y_j\to y_{j+1})}{p_{\sigma^\perp}(y_j\to y_{j+1}|y_{j-1}\to y_j)}\right)\cdot c_k\cdot \left(\prod_{j=0}^{n_E-2}\frac{f_s(z_{j+1}\to z_j\to z_{j-1})}{p_{\sigma^\perp}(z_j\to z_{j+1}|z_{j-1}\to z_j)}\cdot\frac{W_e^0(z_0)}{p_{\mu^2}(z_0)}\right), \tag{9.75}$$

where $\overline{x}_k$ is transport path of length $k = n_L + n_E - 1$.

REMARK 9.15 (pure-Monte Carlo Path Tracing Based on the Path Integral Formulation) *Setting the length of a light path in the basic bidirectional path tracing algorithm to zero, i.e. choosing $n_L = 0$, then BDPT generates only eye paths using n_E path nodes. This means, that BDPT simulates pure-Monte Carlo path tracing, as introduced in Section 9.1.1. A corresponding primary estimator, $F_1^{\mathcal{M}_j,\text{pMCPT}}$, for solving the measurement equation via pMCPT based on the path integral formulation, has the following form:* Section 5.4.1

$$F_1^{\mathcal{M}_j,\text{pMCPT}} = \frac{f_j(\overline{z}_k)}{p(\overline{z}_k)} \tag{9.76}$$

$$= \frac{W_e^0(z_0\to z_1)}{p_{\mu^2}(z_0)}\prod_{j=0}^{n_E-1}\frac{f_s(z_{j-1}\to z_j\to z_{j+1})}{p_{\sigma^\perp}(z_{j-1}\to z_j|z_{j+1}\to z_j)}L_e(z_{n_E}\to z_{n_E-1}), \tag{9.77}$$

where $\overline{x}_k = z_0\ldots z_{N_E-1}$ is a transport path of length $k = n_E - 1$, and a pinhole camera model was used.

Obviously, setting the length of an eye path to zero leads to pure-Monte Carlo light tracing. We leave a detailed derivation to the interested reader as a simple exercise.

9.3.2 THE PATH REUSE STAGE AND THE MULTIPLE SAMPLE ESTIMATOR

Except for the type of transport paths that the basic bidirectional path tracing algorithm generates, the algorithm is not fundamentally different from pure-Monte Carlo path tracing or pure-Monte Carlo light tracing. But also this basic form of BDPT can be extended:

Namely, by varying the length of the light and the eye paths. The algorithm then provides a set of techniques for sampling different kind of paths that are responsible for a variety of lighting effects in the resulting image. By combining samples from all of these techniques via multiple importance sampling we then get a very powerful rendering technique for solving the global illumination problem.

Let us consider a transport path

$$\bar{\mathbf{x}}_k = \underbrace{\mathbf{y}_0 \dots \mathbf{y}_{n_L-1}}_{\bar{\mathbf{y}}} \underbrace{\mathbf{z}_{n_E-1} \dots \mathbf{z}_0}_{\bar{\mathbf{z}}}, \tag{9.78}$$

generated by concatenating an existing light subpath $\bar{\mathbf{y}}$ and an eye path $\bar{\mathbf{z}}$ at its endpoints. By varying the number of vertices from each side of the subpaths $\bar{\mathbf{y}}$ and $\bar{\mathbf{z}}$, the transport path $\bar{\mathbf{x}}_k$ can be reused to construct a variety of unique paths

$$\bar{\mathbf{x}}_{s,t} = \mathbf{y}_0 \dots \mathbf{y}_{s-1}\, \mathbf{z}_{t-1} \dots \mathbf{z}_0 \tag{9.79}$$

from a light source to the camera, with $0 \le s \le n_L$, and $0 \le t \le n_E$.

Then, these paths $\bar{\mathbf{x}}_{s,t}$ can be interpreted as the results of different sampling strategies, where each of these sampling techniques correspond to a different probability density function $p_{s,t}$ on the space of paths $\mathbf{P}^\infty$. Since all density functions take into account different factors of the measurement contribution function f_j, they are all good candidates for importance sampling, that is, each technique can efficiently sample a variety of lighting effects. By using multiple importance sampling, samples from all of these techniques can be combined in a so-called multiple sample estimator $F_N^{\mathcal{M}_j,\mathrm{MIS}}$ of the form

PDF (176) $\mathbf{P}^\infty$ (461) f_j (463) Multiple Sample Estimator (592)

$$\begin{aligned} F_N^{\mathcal{M}_j,\mathrm{MIS,BDPT}} &= \sum_{s\ge 0}\sum_{t\ge 0} w_{s,t}(\bar{\mathbf{x}}_{s,t}) \underbrace{\frac{f_j(\bar{\mathbf{x}}_{s,t})}{p_{s,t}(\bar{\mathbf{x}}_{s,t})}}_{C^*_{s,t}} && (9.80)\\ &= \sum_{s\ge 0}\sum_{t\ge 0} \underbrace{w_{s,t}(\bar{\mathbf{x}}_{s,t}) C^*_{s,t}}_{C_{s,t}} && (9.81)\\ &= \sum_{s\ge 0}\sum_{t\ge 0} C_{s,t} && (9.82) \end{aligned}$$

$C^*_{s,t}$ (726) where $w_{s,t}(\bar{\mathbf{x}}_{s,t})$ are weighting functions, $C^*_{s,t}$ are the unweighted contributions of the path $\bar{\mathbf{x}}_{s,t}$ to $F_N^{\mathcal{M}_j,\mathrm{MIS}}$, and $C_{s,t}$ are the weighted contributions of $\bar{\mathbf{x}}_{s,t}$ to the multiple sample estimator. That is, bidirectional path tracing evaluates the estimator $F_N^{\mathcal{M}_j,\mathrm{MIS}}$ by computing the weighted contributions of all paths $\bar{\mathbf{x}}_{s,t}$ that can be generated from a given transport path composed of a light path $\bar{\mathbf{y}}$ and an eye path $\bar{\mathbf{z}}$.

REMARK 9.16 *Note: Described via Equation (9.82), the estimator* $F_N^{\mathcal{M}_j,\mathrm{MIS,BDPT}}$ *corresponds to the sum of contributions from paths sampled from an infinite number of*

techniques. By defining the sample $\overline{\mathbf{x}}_{s,t} = \epsilon$ *whenever* $s > n_L$ *or* $t > n_E$ *all paths, except from a finite number, have contribution zero, that is, their contribution to the shading of a pixel can be ignored.*

Let us now show how bidirectional path tracing determines in an efficient manner the components used in the above multiple sample estimator.

THE PATH REUSE STAGE. Let $\overline{\mathbf{y}} = \mathbf{y}_0 \dots \mathbf{y}_{n_L-1}$ and $\overline{\mathbf{z}} = \mathbf{z}_0 \dots \mathbf{z}_{n_E-1}$ be a light, respectively, an eye subpath generated via bBDPT. Then, extended bidirectional path tracing does not only connect the two endpoints of the subpaths $\overline{\mathbf{y}}$ and $\overline{\mathbf{z}}$ to get a transport path from a light source to a sensor, but also all endpoints of subpaths $\mathbf{y}_0 \dots \mathbf{y}_{s-1}, 0 \le s \le n_L$ of $\overline{\mathbf{y}}$ with subpaths $\mathbf{z}_{t-1} \dots \mathbf{z}_0, 0 \le t \le n_E$ of $\overline{\mathbf{z}}$. This then results in $s+t+1$ paths

$$\overline{\mathbf{x}}_{s,t} = \mathbf{y}_0 \dots \mathbf{y}_{s-1}\, \mathbf{z}_{t-1} \dots \mathbf{z}_0 \tag{9.83}$$

of length $s+t-1$, consisting of $s+t$ vertices and $s+t-1$ edges, see Figure 9.22.

EXAMPLE 9.4 (Paths Generated via Bidirectional Path Tracing) *Let us consider the light path* $\overline{\mathbf{y}} = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2$ *consisting of* $n_L = 3$ *vertices and the eye path* $\overline{\mathbf{z}} = \mathbf{z}_1\mathbf{z}_0$ *with* $n_E = 2$ *nodes. Then, the transport path* $\overline{\mathbf{x}}_{s,t}$*, composed of* $\overline{\mathbf{y}}$ *and* $\overline{\mathbf{z}}$*, has length* $n_L + n_E - 1 = 4$ *and is of the form*

$$\overline{\mathbf{x}}_{s,t} = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2\mathbf{z}_1\mathbf{z}_0. \tag{9.84}$$

Based on $\overline{\mathbf{x}}_{s,t}$*, the bidirectional path tracing algorithm then generates* $k+2$ *transport paths of lengths* k*:*

$$\begin{aligned} k = 1 \quad \rightarrow \quad & \overline{\mathbf{x}}_{0,2} = \mathbf{z}_1\mathbf{z}_0 \\ & \overline{\mathbf{x}}_{1,1} = \mathbf{y}_0\mathbf{z}_0 \\ & \overline{\mathbf{x}}_{2,0} = \mathbf{y}_0\mathbf{y}_1 \end{aligned} \tag{9.85}$$

and

$$\begin{aligned} k = 2 \quad \rightarrow \quad & \overline{\mathbf{x}}_{0,3} = \epsilon \\ & \overline{\mathbf{x}}_{1,2} = \mathbf{y}_0\mathbf{z}_1\mathbf{z}_0 \\ & \overline{\mathbf{x}}_{2,1} = \mathbf{y}_0\mathbf{y}_1\mathbf{z}_0 \\ & \overline{\mathbf{x}}_{3,0} = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2 \end{aligned} \tag{9.86}$$

$$\begin{aligned} k = 3 \quad \rightarrow \quad & \overline{\mathbf{x}}_{0,4} = \epsilon \\ & \overline{\mathbf{x}}_{1,3} = \epsilon \\ & \overline{\mathbf{x}}_{2,2} = \mathbf{y}_0\mathbf{y}_1\mathbf{z}_1\mathbf{z}_0 \\ & \overline{\mathbf{x}}_{3,1} = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2\mathbf{z}_0 \\ & \overline{\mathbf{x}}_{4,0} = \epsilon, \end{aligned} \tag{9.87}$$

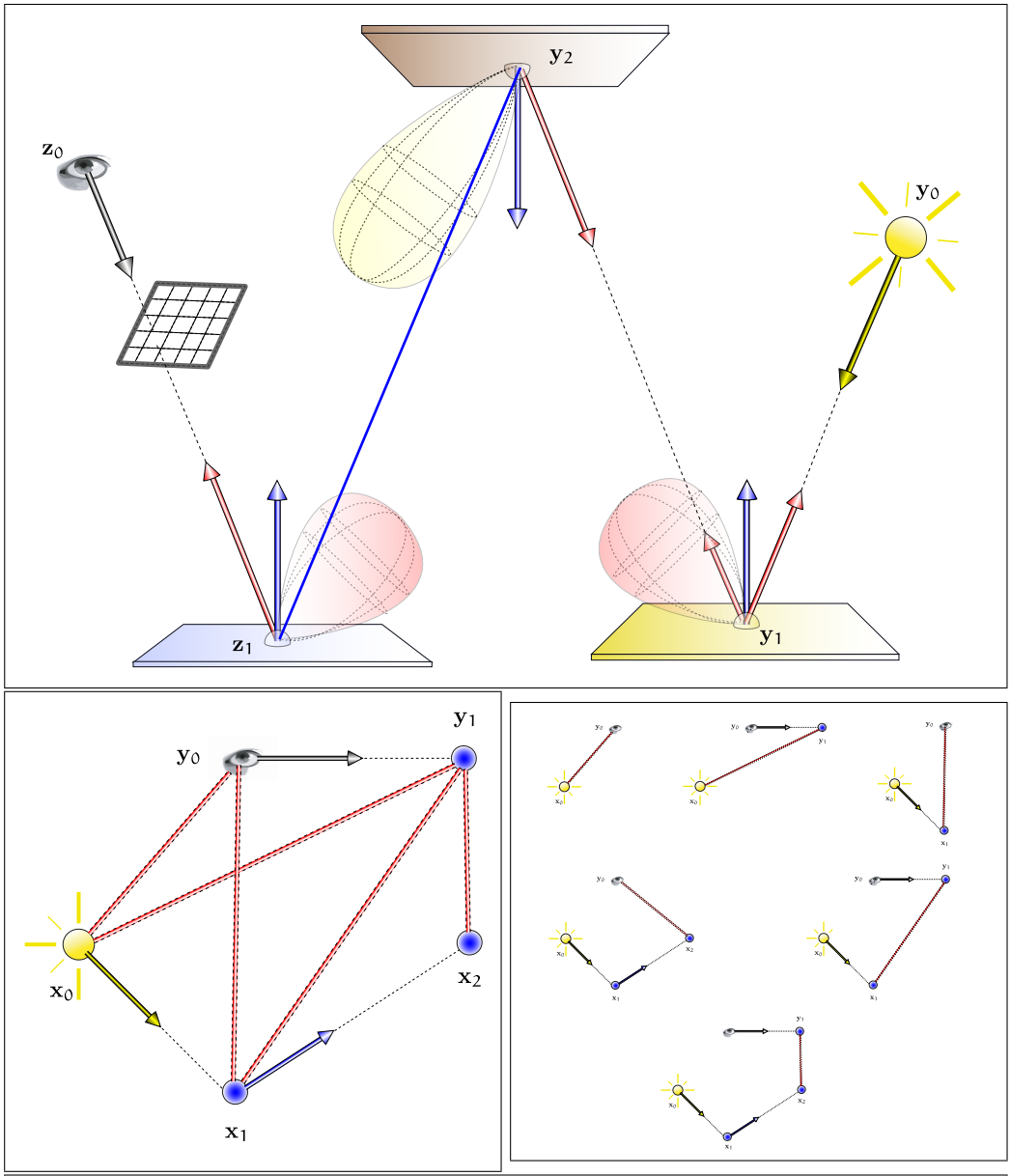


FIGURE 9.22: A TRANSPORT PATH WITH ASSOCIATED SUBPATHS. By varying the number of vertices in the light path $\overline{y} = y_0 y_1 y_2$ and in the eye path $\overline{z} = z_1 z_0$ bidirectional path tracing generates a large class of new transport paths. Thus, $\overline{x}_{3,1} = y_0 y_1 y_2 z_0$ is a path of length three connecting the light path $\overline{y}$ with the point z_0 on the light source. Another path of length three is given by connecting the eye path $\overline{x}$ with the light subpath $y_0 y_1$ resulting in $\overline{x}_{2,2} = y_0 y_1 z_1 z_0$. Note: All paths that can be generated by varying in the number of vertices of $\overline{y}$ and $\overline{z}$ have length k, with $1 \leq k \leq 4$.

where $\overline{\mathbf{x}}_{s,t} \stackrel{\text{def}}{=} \epsilon$ *is the empty path whenever* $s > n_L$ *or* $t > n_E$*, see Figure 9.23.*

It should be clear, that for paths of length, $k = 4$*, only the path* $\overline{\mathbf{x}}_{3,2}$ *has to be accounted for.*

As already mentioned above, each of the transport paths $\overline{\mathbf{x}}_{s,t}$ can now be interpreted as a sample generated from a sampling strategy $p_{s,t}$ that corresponds to a different density function on the space of paths of all length $\mathbf{P}^\infty$. Thus, $p_{s,t}$ is a sampling technique that performs s steps via Monte Carlo light tracing and t steps via Monte Carlo path tracing. Path Space (461)

In the same way as the basic variant of bidirectional path tracing, the algorithm samples the subpaths $\overline{\mathbf{y}} = \mathbf{y}_0 \dots \mathbf{y}_{n_L-1}$ and $\overline{\mathbf{z}} = \mathbf{z}_0 \dots \mathbf{z}_{n_E-1}$, of the transport path

$$\overline{\mathbf{x}}_{s,t} = \mathbf{y}_0 \dots \mathbf{y}_{s-1}\mathbf{z}_{t-1} \dots \mathbf{z}_0. \tag{9.88}$$

Using the abbreviations

$$p_1^L = p_{\mu^2}(\mathbf{y}_0) \tag{9.89}$$

$$p_1^E = p_{\mu^2}(\mathbf{z}_0) \tag{9.90}$$

and

$$p_{j+1}^L = p_{\sigma^\perp}(\mathbf{y}_{j-1} \to \mathbf{y}_j | \mathbf{y}_{j-2} \to \mathbf{y}_{j-1})\, G(\mathbf{y}_{j-1} \leftrightarrow \mathbf{y}_j)\, p_j^L \tag{9.91}$$

as well as

$$p_{j+1}^E = p_{\sigma^\perp}(\mathbf{z}_0)\mathbb{P}_{\sigma^\perp}(\mathbf{z}_{j-1} \to \mathbf{z}_j | \mathbf{z}_{j-2} \to \mathbf{z}_{j-1})\, G(\mathbf{z}_{j-1} \leftrightarrow \mathbf{z}_j)\, p_j^E \tag{9.92}$$

for $j \geq 1$, then the density for generating the path $\overline{\mathbf{x}}_{s,t}$ is given by:

$$p_{s,t}(\overline{\mathbf{x}}_{s,t}) = \underbrace{p_{\mu^2}(\mathbf{y}_0)}_{p_1^L} \prod_{j=1}^{s-1} \underbrace{p_{\sigma^\perp}(\mathbf{y}_{j-1} \to \mathbf{y}_j | \mathbf{y}_{j-2} \to \mathbf{y}_{j-1})\, G(\mathbf{y}_{j-1} \leftrightarrow \mathbf{y}_j)}_{\frac{p_{j+1}^L}{p_j^L}} \cdot \tag{9.93}$$

$$\underbrace{p_{\mu^2}(\mathbf{z}_0)}_{p_1^E} \prod_{j=1}^{t-1} \underbrace{p_{\sigma^\perp}(\mathbf{z}_{j-1} \to \mathbf{z}_j | \mathbf{z}_{j-2} \to \mathbf{z}_{j-1})\, G(\mathbf{z}_{j-1} \leftrightarrow \mathbf{z}_j)}_{\frac{p_{j+1}^E}{p_j^E}}$$

$$= \underbrace{p_1^L \prod_{j=1}^{s-1} \frac{p_{j+1}^L}{p_j^L}}_{p_s^L} \cdot \underbrace{p_1^E \prod_{k=1}^{t-1} \frac{p_{j+1}^E}{p_j^E}}_{p_t^E} \tag{9.94}$$

$$= p_s^L \cdot p_t^E. \tag{9.95}$$

Here, we also used the formulas

$$p_{\sigma^\perp}(\mathbf{y}_0 \to \mathbf{y}_1 | \mathbf{y}_{-1} \to \mathbf{y}_0) = p_{\sigma^\perp}(\mathbf{y}_0 \to \mathbf{y}_1 | \mathbf{y}_0) \tag{9.96}$$

$$p_{\sigma^\perp}(\mathbf{z}_0 \to \mathbf{z}_1 | \mathbf{z}_{-1} \to \mathbf{z}_0) = p_{\sigma^\perp}(\mathbf{z}_0 \to \mathbf{z}_1 | \mathbf{z}_0) \tag{9.97}$$

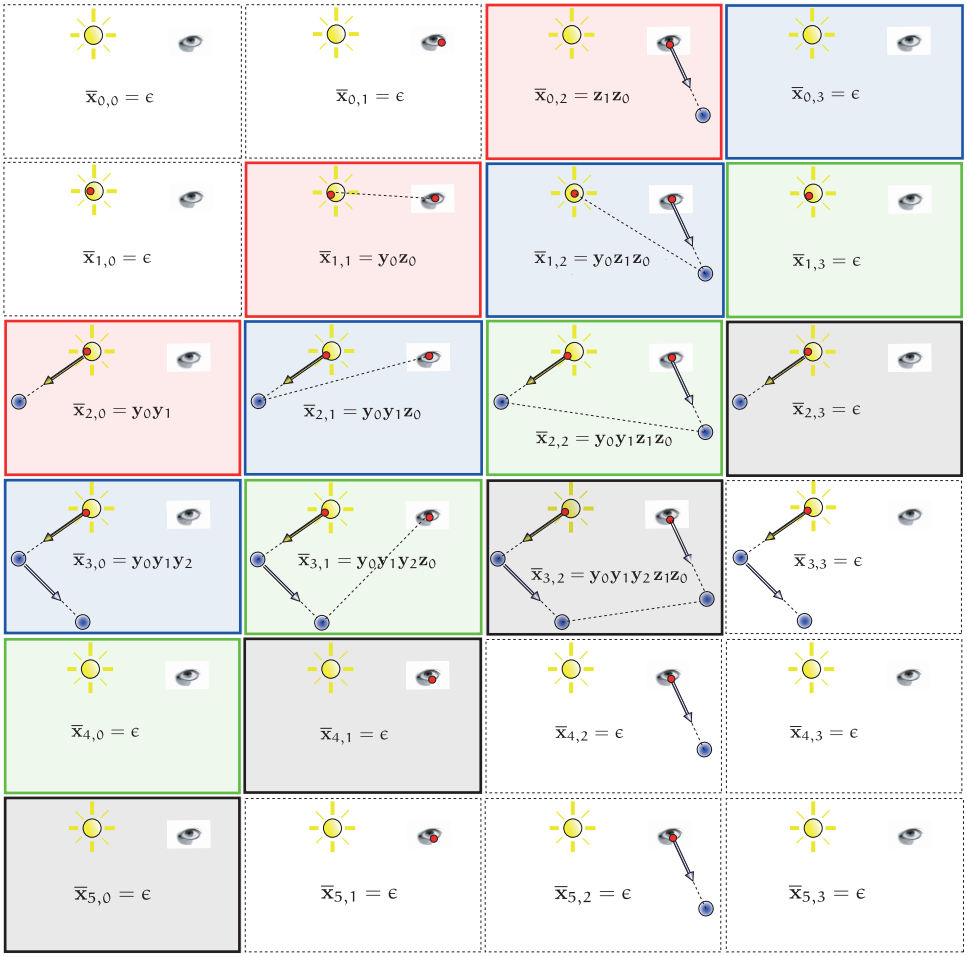


FIGURE 9.23: GENERATING SAMPLES $\overline{x}_{s,t}$ FROM A TRANSPORT PATH. The transport path $\overline{x}_4 = y_0 y_1 y_2 z_1 z_0$ implies the construction of subpaths of length $k = 1, 2, 3$ and $k = 4$. Note, paths of type $\overline{x}_{s,t}$ with $s > n_L$ or $t > n_E$ as well as paths of lengths $k = -1$ and $k = 0$ are empty paths. They deliver no contribution to the multiple sample estimator.

and assume that it holds: $p_0^L = p_0^E = 1$, see Figure 9.24.

As, the geometry terms occurring within the nominator and the denominator of the unweighted contributions $C^*_{s,t}$ can then be canceled, the formula for $C^*_{s,t}$ can be simplified in accordance with Equation (9.72) to:

$$\begin{aligned} C^*_{s,t} &\stackrel{def}{=} \underbrace{\frac{L_e^0(\mathbf{y}_0)}{p_{\mu^2}(\mathbf{y}_0)}}_{\alpha_1^L} \cdot \prod_{j=0}^{s-2} \underbrace{\frac{f_s(\mathbf{y}_{j-1} \to \mathbf{y}_j \to \mathbf{y}_{j+1})}{p_{\sigma^\perp}(\mathbf{y}_j \to \mathbf{y}_{j+1} | \mathbf{y}_{j-1} \to \mathbf{y}_j)}}_{\frac{\alpha_{j+2}^L}{\alpha_{j+1}^L}} \cdot \\ & \quad c_{s,t} \cdot \\ & \quad \prod_{j=0}^{t-2} \underbrace{\frac{f_s(\mathbf{z}_{j+1} \to \mathbf{z}_j \to \mathbf{z}_{j-1})}{p_{\sigma^\perp}(\mathbf{z}_j \to \mathbf{z}_{j+1} | \mathbf{z}_{j-1} \to \mathbf{z}_j)}}_{\frac{\alpha_{j+2}^E}{\alpha_{j+1}^E}} \cdot \underbrace{\frac{W_e^0(\mathbf{z}_0)}{p_{\mu^2}(\mathbf{z}_0)}}_{\alpha_1^E} \end{aligned} \tag{9.98}$$

$$= \alpha_s^L \cdot c_{s,t} \cdot \alpha_t^E \tag{9.99}$$

with

$$c_{s,t} \stackrel{def}{=} \begin{cases} L_e(\mathbf{z}_{t-1} \to \mathbf{z}_{t-2}) & \text{if } s = 0, t > 0 \\ W_e^j(\mathbf{y}_{s-2} \to \mathbf{y}_{s-1}) & \text{if } s > 0, t = 0 \\ f_s(\mathbf{y}_{s-2} \to \mathbf{y}_{s-1} \to \mathbf{z}_{t-1}) \cdot \mathcal{G}(\mathbf{y}_{s-1} \leftrightarrow \mathbf{z}_{t-1}) \cdot & \text{if } s, t > 0 \\ f_s(\mathbf{y}_{s-1} \to \mathbf{z}_{t-1} \to \mathbf{y}_{t-2}). & \end{cases} \tag{9.100}$$

THE MULTIPLE SAMPLE ESTIMATOR. Finally , we consider the computation of the weighting functions $w_{s,t}$. For that purpose, let us consider the path $\overline{\mathbf{x}}_{i,(s+t)-i}$ using a light subpath with i vertices and an eye subpath with $s+t-i$ vertices, given by:

$$\overline{\mathbf{x}}_{i,(s+t)-i} = \underbrace{\mathbf{y}_0 \dots \mathbf{y}_{i-1}}_{\mathbf{x}_0 \dots \mathbf{x}_{i-1}} \underbrace{\mathbf{z}_{(s+t)-i-1} \mathbf{z}_{(s+t)-i-2} \dots \mathbf{z}_0}_{\mathbf{x}_i \mathbf{x}_{i+1} \dots \mathbf{x}_{s+t}} \tag{9.101}$$

as well as the path $\overline{\mathbf{x}}_{i+1,(s+t)-i-1}$ using a light subpath with $i+1$ vertices and an eye subpath with $s+t-i-1$ vertices, given by:

$$\overline{\mathbf{x}}_{i+1,(s+t)-i-1} = \underbrace{\mathbf{y}_0 \dots \mathbf{y}_{i-1} \mathbf{y}_i}_{\mathbf{x}_0 \dots \mathbf{x}_{i-1} \mathbf{x}_i} \underbrace{\mathbf{z}_{(s+t)-i-2} \mathbf{z}_{(s+t)-i-3} \dots \mathbf{z}_0}_{\mathbf{x}_{i+1} \dots \mathbf{x}_{s+t}}. \tag{9.102}$$

Let furthermore p_i and p_{i+1} denote the probabilities for generating these paths, then the ratio $\frac{p_{i+1}}{p_i}$ can be expressed in terms of the probabilities for generating the light and

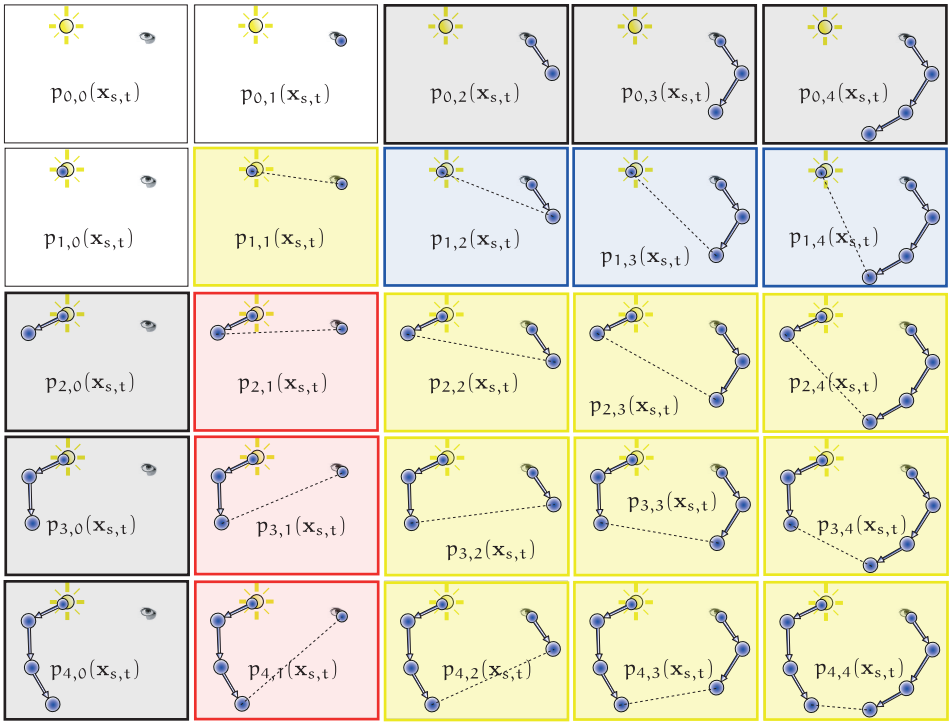


FIGURE 9.24: A TRANSPORT PATH WITH ASSOCIATED SAMPLING STRATEGIES. Except for the first two contributions, the left-most column corresponds to pure-Monte Carlo light tracing. The second column corresponds to Monte Carlo light tracing with next event estimation. On the other side, the first row corresponds to pure-Monte Carlo light tracing. The second row corresponds to Monte Carlo light tracing with next event estimation. All other contributions are unique to bidirectional path tracing.

eye subpaths of lengths i and $i+1$ from Equations (9.91) and (9.92) given by:

$$\frac{p_{i+1}}{p_i} = \frac{p^L_{i+1} p^E_{s+t-i-1}}{p^L_i p^E_{s+t-i}} \tag{9.103}$$

$$= \frac{p^L_i \left(p_{\sigma^\perp}(x_{i-1} \to x_i | x_{i-2} \to x_{i-1})\, \mathcal{G}(x_{i-1} \leftrightarrow x_i)\right) p^E_{s+t-i-1}}{p^L_i \left(p_{\sigma^\perp}(x_{i+1} \to x_i | x_i \to x_{i-1})\, \mathcal{G}(x_{i+1} \leftrightarrow x_i)\right) p^E_{s+t-i-1}} \tag{9.104}$$

$$= \frac{p_{\sigma^\perp}(x_{i-1} \to x_i | x_{i-2} \to x_{i-1})\, \mathcal{G}(x_{i-1} \leftrightarrow x_i)}{p_{\sigma^\perp}(x_{i+1} \to x_i | x_i \to x_{i-1})\, \mathcal{G}(x_{i+1} \leftrightarrow x_i)} \tag{9.105}$$

with

$$p_1 \stackrel{def}{=} p_{\mu^2}(x_0) \tag{9.106}$$

and

$$p_{s+t} \stackrel{def}{=} p_{\mu^2}(\mathbf{x}_{s+t}). \tag{9.107}$$

Obviously, the only difference between p_i and p_{i+1} lies in how the vertex $\mathbf{x}_i$ is chosen: With respect to sampling strategy p_i, it is generated as part of the eye subpath $\mathbf{x}_i \dots \mathbf{x}_{s+t}$, while for p_{i+1} it is generated via the light subpath $\mathbf{x}_0 \dots \mathbf{x}_i$. Bidirectional path tracing can now uses this result in its combination strategies for computing the weighting functions w_i. Thus, if the samples are combined using the power heuristic with $\beta = 2$, we have to compute:

Power Heuristic (597)

$$w_{s,t} \stackrel{def}{=} \frac{p_s^2}{\sum_{i=0}^{s+t} p_i^2} = \frac{1}{\sum_{i=0}^{s+t} \left(\frac{p_i}{p_s}\right)^2}. \tag{9.108}$$

For evaluating the weighting functions $w_{s,t}$, BDT requires the probability p_s of the currently generated path $\overline{\mathbf{x}}_{s,t}$ and the probabilities $p_0, p_1, \dots, p_{s-1}, p_{s+1}, \dots, p_{s+t}$ with which all other sampling strategies would generate this path. Now, these probabilities can easily computed via the ratio $\frac{p_{s+1}}{p_s}$, respectively, the reciprocal ratio $\frac{p_s}{p_{s+1}}$. Based on these quantities, then the required probability $\frac{p_i}{p_s}$ for evaluating the weighting functions $w_{s,t}$ can be expressed in terms of $\frac{p_j}{p_{j+1}}$. So, we get for $i < s$:

$$\frac{p_i}{p_s} = \prod_{j=i}^{s-1} \frac{p_j}{p_{j+1}}, \tag{9.109}$$

respectively, for $i > s$

$$\frac{p_i}{p_s} = \prod_{j=s}^{i-1} \frac{p_{j+1}}{p_j}. \tag{9.110}$$

REMARK 9.17 *If we compare bidirectional path tracing with Monte Carlo path tracing, then we can say, that BDPT performs a lot better. The algorithm is powerful in particular for many kinds of indoor scenes, with or without strong indirect lighting, and for scenes containing caustics. A weakness of BDPT arises when rendering outdoor scene, or scenes where the light sources and the viewer are separated by difficult geometry, such as a single room in a large building where the light sources are far away. In this case, another rendering techniques, should be used: the* Metropolis light transport *algorithm.*

Section 9.4

In [222, Veach & Guibas 1995] bidirectional path tracing was compared against ordinary path tracing using some test scenes shown in Figure 9.25.

FIGURE 9.25: THE WEIGHTED CONTRIBUTIONS OF BIDIRECTIONAL SAMPLING TECHNIQUES $\overline{\mathbf{p}}_{s,t}$. The pyramid of images has to be seen in connection with the image in Figure 9.17 rendered with bidirectional path tracing. Except of the first row—which we have missed, thus the light that is directly visible to eye—each row of the pyramid contains contributions, weighted via MIS, to the final image from paths with the same path length. Each row r of the pyramid shows the contributions of the sampling techniques $p_{1,r+1}, \ldots, p_{r+1,1}$ for paths of length $k = r + 1$, where the position of an image in the row indicates how the associated path were generated. Obviously, the images on the left can interpreted as rendered by Monte Carlo path tracing with next event estimation, while the images on the right are considered to be rendered via Monte Carlo light tracing with next event estimation. The s^{th} image from left is constructed via a light subpath with s vertices, while the t^{th} image from the right uses an eye subpath with t vertices. Thus, the lower left image is constructed via paths that uses only a single light vertex but six eye vertices, while the upper right image uses two light vertices but only a single eye vertex. Note: The images have been over-exposed so that their details can be seen. See this image in connection with Figure 9.26. Image courtesy of Eric Veach.

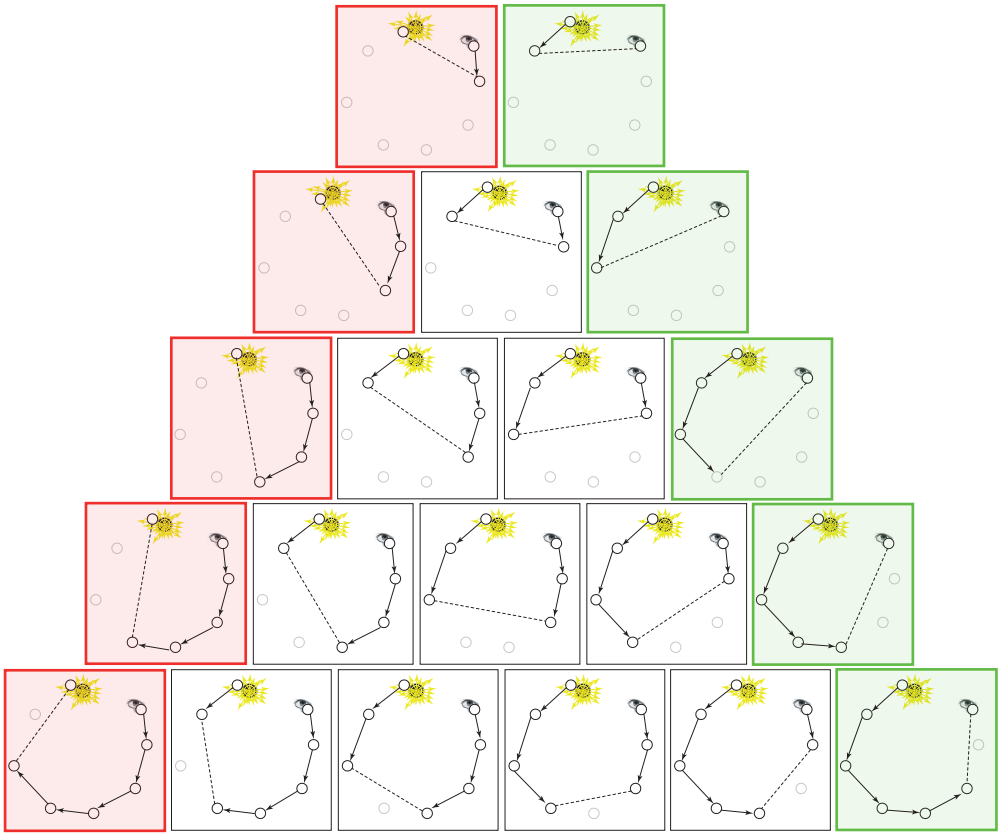


FIGURE 9.26: THE WEIGHTED CONTRIBUTIONS OF BIDIRECTIONAL SAMPLING TECHNIQUES $\overline{\mathbf{p}}_{s,t}$. Visualization of the transport paths associated with the rendered images in Figure 9.25.

9.4 METROPOLIS LIGHT TRANSPORT

As we have seen in the previous sections of this chapter, so, it is very difficult to design a path-based light transport algorithm that is general valid, efficient, and that simulates all light effects artifact-free. Such an algorithm has to sample all kinds of transport paths between light sources and a virtual camera or the eye of an observer. One problem that often arises is: Based on the underlying scene model, many paths do not contribute significantly to the image to be rendered. If we consider for example a scene where a brightly illuminated room is connected with a dark room via a slightly opened door, then path tracing, as well as light tracing, have problems to sample paths that contribute significantly to the final image. With the idea of connecting a light and an eye path, bidirectional path tracing

Section 9.1

Section 9.2

Section 9.3 comes close to this goal, but BDPT suffers from, that the algorithm can not guarantee to generate a large number of such paths that are required to simulate this special illumination situation. Here, there exists another more efficient method, based on a MCMC approach, that handles such difficult sampling problems more efficiently: the *Metropolis light transport* algorithm.

The Metropolis light transport, also briefly denoted as *MLT*, firstly presented in [223, Veach & Guibas 1997], is an unbiased algorithm, handles general geometric and scattering models and uses little storage. It is based on the Metropolis sampling algorithm, $M(RT)^2$. As shown in Section 6.5.3.2, $M(RT)^2$ generates a sequence of correlated samples from a non-negative function f such that the samples are distributed according to f. For that, the algorithms only needs to evaluate f at each generated sample, $M(RT)^2$ does not require any other information about f or its associated PDF. Thus, this approach is quite different from the sampling strategies of MCPT, MCLT, or BDPT, where samples are chosen PDF (176) according to a PDF and a function is evaluated at the generated samples.

In the following, we will describe the Metropolis light transport algorithm, as firstly introduced in [223, Veach & Guibas 1997] and [221, Veach 1998], where the Metropolis algorithm is applied to the path integral formulation introduced in Section 5.4.

Section 9.4.1 So, we show how the light transport problem must be formulated that it fits the $M(RT)^2$ framework and discuss how an image can be computed by sampling a finite number of random paths according to some density function. We also present the idea, how the MLT algorithm can be initialized to avoid the start-up bias that comes with a Markov process constructed via Metropolis sampling. Finally, we talk about mutation Section 9.4.2 strategies, the heart of the MLT algorithm, that should help to minimize the error in the final image rendered with MLT.

9.4.1 THE METROPOLIS LIGHT TRANSPORT ALGORITHM

Recall from Section 5.4, where we have presented the path integral formulation of light transport. Based on the path integral formulation, the flux $\mathcal{M}_j$ through the pixel $\square_j$ can be written as an integral of the form

$$\mathcal{M}_j \stackrel{def}{=} \int_{\mathbf{P}^\infty} f_j(\overline{\mathbf{x}})\, d\mu_\infty(\overline{\mathbf{x}}), \tag{9.111}$$

$\mathbf{P}^\infty$ (461) where the path space $\mathbf{P}^\infty$ is the integration domain, the measurement contribution function f_j (463) f_j corresponds to the integrand, and μ_∞ is the path measure defined on the path space μ_∞ (461) $\mathbf{P}^\infty$.

In Section 9.3 we have estimated this integral by a secondary Monte Carlo estimator associated with the basic bidirectional path tracing algorithm of the form

$$F_N^{\mathcal{M}_j,bBDPT} = \frac{1}{N}\sum_{i=1}^{N} \frac{f_j(\overline{\mathbf{X}}_i)}{p(\overline{\mathbf{X}}_i)}, \tag{9.112}$$

where $\overline{\mathbf{X}}_i$ are paths from $\mathbf{P}^\infty$ sampled according to the PDF p.

ESTIMATING PIXEL VALUES IN MLT. Let us assume, we have a sufficient large set of random paths $\overline{\mathbf{X}}_0, \overline{\mathbf{X}}_1, \ldots, \overline{\mathbf{X}}_N$, instead of sampled from a given PDF such as in bidirectional path tracing, all paths are samples from an ergodic MCMC process. Then, due to the Ergodic Theorem, a secondary Monte Carlo estimator $F_N^{\mathcal{M}_j,\mathrm{MCMC}}$ for measuring the flux through the pixel $\square_j$ has the form

Ergodic MCMC Process (546)
Ergodic Theorem (546)

$$F_N^{\mathcal{M}_j,\mathrm{MCMC}} \stackrel{\mathrm{def}}{=} E\left(\frac{1}{N-M}\sum_{i=M}^{N}\frac{f_j(\overline{\mathbf{X}}_i)}{p(\overline{\mathbf{X}}_i)}\right). \tag{9.113}$$

If h_j represents the filter function for pixel $\square_j$ and f represents all other factors of the measurement contribution function—the BSDFs, geometry terms, and the emitted radiance from a light source that has been hit—then the measurement equation can be estimated by:

$$\mathcal{M}_j = E\left(\frac{1}{N-M}\sum_{i=M}^{N}\frac{h_j(\overline{\mathbf{X}}_i)\, f(\overline{\mathbf{X}}_i)}{p(\overline{\mathbf{X}}_i)}\right). \tag{9.114}$$

Assuming the PDF where we sample from is chosen, as:

$$p(\overline{\mathbf{x}}) \stackrel{\mathrm{def}}{=} \frac{f(\overline{\mathbf{x}})}{\int_{\mathbf{P}^\infty} f(\overline{\mathbf{x}})\, d\mu_\infty(\overline{\mathbf{x}})}, \tag{9.115}$$

then the representation of the above estimator simplifies to

$$F_N^{\mathcal{M}_j,\mathrm{MCMC}} = E\left(\frac{1}{N-M}\sum_{i=M}^{N} b\, h_j(\overline{\mathbf{X}}_i)\right), \tag{9.116}$$

where $b \stackrel{\mathrm{def}}{=} \int_{\mathbf{P}^\infty} f(\overline{\mathbf{x}})\, d\mu_\infty(\overline{\mathbf{x}})$.

REMARK 9.18 (Start-up Bias) *As we known from our discussion of* $M(RT)^2$ *the samples* $\overline{\mathbf{X}}_i$ *will be distributed according to* f *only in the limit as* $i \to \infty$. *This is also the reason why we discard the first* M *samples in the estimator* $F_N^{\mathcal{M}_j,\mathrm{MCMC}}$ *under the assumption that the random walk has approximately converged to the equilibrium distribution. Statements about the choice of* M *can only be made in rare cases. If we choose* M *to small, then the samples will be strongly correlated to the starting sample* $\overline{\mathbf{X}}_0$, *which will results in the so-called* start-up bias, *which is not only unsatisfactorily since it will bias the result, but also due to the unnecessarily high cost for the discarded samples.*

In [223, Veach & Guibas 1997] and [221, Veach 1998] an unbiased approach is proposed. Here, it is suggested to sample an initial sample $\overline{\mathbf{X}}_0$ *from some convenient*

```
METROPOLIS LIGHT TRANSPORT {
    X̄ ← INITIAL-PATH()
    image ← {0, ..., 0}
    ∀ i ∈ {1, 2, ..., N} do {
        Y ← MUTATE(X)
        α ← ACCEPT-PROB(X̄ → Ȳ)
        if RANDOM() < α {
            X̄ ← Ȳ
        }
        RECORD-SAMPLE(image, X̄)
}
```

FIGURE 9.27: PSEUDOCODE FOR THE METROPOLIS LIGHT TRANSPORT ALGORITHM.

PDF p_0. *If the density* p_0 *is not the desired equilibrium distribution* $\pi^* = \frac{1}{b} f$, *then the sample* $\overline{\mathbf{X}}_0$ *is assigned a weight*

$$W_0 \stackrel{\text{def}}{=} \frac{f(\overline{\mathbf{X}}_0)}{p_0(\overline{\mathbf{X}}_0)}. \tag{9.117}$$

According to the $M(RT)^2$ *algorithm, then new samples* $\overline{\mathbf{X}}_1, \overline{\mathbf{X}}_2, \ldots, \overline{\mathbf{X}}_N$ *are generated via an ergodic MCMC process starting from* $\overline{\mathbf{X}}_0$. *Also these samples are then weighted with* $W_i = W_0$. *As shown in [223, Veach & Guibas 1997] and [221, Veach 1998], the resulting secondary Monte Carlo estimator*

$$F_N^{\mathcal{M}_j,\mathrm{MCMC}} = E\left(\frac{1}{N}\sum_{i=1}^{N} b\, h_j(\overline{\mathbf{X}}_i)\right) \tag{9.118}$$

is unbiased.

THE STRUCTURE OF THE MLT ALGORITHM. Let us now consider the basic structure of the Metropolis light transport algorithm summarized in Figure 9.27. The algorithm begins with the search for a suitable initial path $\overline{\mathbf{X}}_0$. This path should connect a light source with a pixel. In the basic MLT algorithm, such a path is constructed by bidirectional path tracing via connecting an eye and a light path.

Section 9.3

REMARK 9.19 *Choosing the right initial path is crucial for the further course. In principle, each eye path through any point on the image plane can be used. So, only a single random path* $\overline{\mathbf{X}}_0$ *has to be generated via BDPT. But for different reasons,*

for details see [223, Veach & Guibas 1997] and [221, Veach 1998], Veach recommend to generate more than a single initial path $\overline{\mathbf{X}}_0$ and to run copies of the algorithm in parallel for each initial path, where all samples are accumulated into one image.

Then, the algorithm chooses one of the mutation strategies, which modifies the given path according to certain rules—we will discuss the mutation strategies in more detail in the following section. Depending on the contribution it makes to the ideal image the new mutation is accepted or rejected with a carefully chosen probability. Based on these mutation strategies, the Metropolis light transport algorithm generates a sequence of random paths $\overline{\mathbf{X}}_0, \overline{\mathbf{X}}_1, \ldots, \overline{\mathbf{X}}_N$, where each $\overline{\mathbf{X}}_i, i > 0$ is obtained by a random mutation of the path $\overline{\mathbf{X}}_{i-1}$. If a path is accepted, then MLT updates the current image which is stored in memory as a 2-dimensional array of pixel values. For this, the algorithm has to find the point where the path sample $\overline{\mathbf{X}}_i$ intersect the image plane, and it has to update all those pixels whose filter support contains the hit point of $\overline{\mathbf{X}}_i$ with the image plane.

REMARK 9.20 *Due to [220, Veach & Guibas 1997], the MLT algorithm is unbiased, handles general geometric and scattering model and uses little storage. It performs especially well on problems that are usually considered difficult, e.g. those involving bright indirect light, small geometric holes, or glossy surfaces.*

9.4.2 MUTATION STRATEGIES

As we have shown in the preceding section, the idea behind the Metropolis light transport algorithm is quite different from Monte Carlo path tracing and Bidirectional path tracing. Instead to sample from a PDF and to evaluate a function at the generated samples, MLT generates samples proportional to the unknown function, which, in case of the light transport equation, corresponds to the unknown radiance distribution in the scene. So the algorithm explores, if an energetic, hard to find, path is found, the neighborhood of this path, where probably more good paths can be found. Obviously, this strategy is a clever idea, since MLT concentrates work in the bright regions of a scene. Here, MLT counts on a series of so-called *mutation strategies*, resulting in a change of the original path by slight shifting single vertices or edges, respectively, by adding to or deleting edges from the original path. There are many illumination situations in a scene where such a local exploration of the path space can lead to faster convergence as compared to the other ray based rendering methods.

Section 9.1
Section 9.3
PDF (176)

REQUIREMENTS TO MUTATION STRATEGIES. As the Metropolis light transport algorithm is based on the $M(RT)^2$ algorithm, consecutive paths generated via MLT are correlated, which, compared with a Monte Carlo sampling strategy, can lead to higher variance. To minimize the error in the final image, in [220, Veach 1997] some properties are required to a good mutation strategy:

Section 6.5.3.2

- *High Acceptance Probability*: Obviously, small acceptance probability leads to a long series of mutations which are rejected. This then means that the current image is updated with the same path for a long time, which appears as noise.
- *Large Changes to a Path*: Even if the acceptance probability is high, mutations with small changes only implies a bad cover of the image plane.
- *Ergodicity*: Mutations should not be limited in its average too much. So, the whole path space can be explored independent on the initial path, and paths are prevented to stuck in some subregions of the path space.
- *Changes to the Image Location*: To guarantee a good cover of the image plane, a mutation strategy should be applied to the first edge of an eye path.
- *Stratification*: The contributions of all paths to shading a pixel should be the same for all pixels.
- *Low Cost*: Mutations should involve as little as possible computational effort and usage of resources.

In the following, we will describe three different mutation strategies that are implemented in Eric Veach's MLT algorithm: *bidirectional mutations*, *perturbations*, and *lens subpath mutations*.

BIDIRECTIONAL MUTATIONS. The heart of the MLT algorithm are *bidirectional mutations*. Let us consider a path $\overline{\mathbf{X}}_j = \mathbf{X}_{j0}\mathbf{X}_{j1}\dots\mathbf{X}_{jk}$ generated via the Metropolis light transport algorithm. The idea behind a bidirectional mutation is to choose a subpath $\mathbf{X}_{jl}\mathbf{X}_{jl+1}\dots\mathbf{X}_{jm}, 0 < l < m < k$ with probability

$$p_d[l,m] = p_1[l,m] \cdot p_2[l,m], \tag{9.119}$$

where $p_1[l,m]$ depends only on the number of the edges of the subpath and the purpose of $p_2[l,m]$ is to avoid mutations with low acceptance. If the algorithm has chosen such a subpath, it will be deleted from $\overline{\mathbf{X}}_j$, which leads to two, detached from each other, subpaths $\mathbf{X}_{j0}\dots\mathbf{X}_{j1}$ and $\mathbf{X}_{jm}\dots\mathbf{X}_{jk}$ with one or more vertices.

REMARK 9.21 *It should be clear, that we are interested in the deletion of short subpaths, which are cheap to replace, and whose mutations results from small changes to the current path.*

After deleting the chosen subpath from $\overline{\mathbf{X}}_j$, the algorithm generates a new subpath. Since it is desirable that the new subpath is similar to the deleted subpath, which increases the acceptance probability, MLT choses the new subpath length with high probability $p_{a,1}$ similar to the length of the deleted subpath, where the number of vertices by which the light and the eye path of the new subpath must be extended are determined via a probability

distribution $p_{a,2}$. Even in this step it should be ensured, that the changes with respect to deleted subpath are not so large, which leads to high acceptance probability of a mutation.

Afterwards, the light and the eye path of the original path are extended by the number of vertices determined in the above step. This requires the generation of a ray via sampling the BSDF and casting the ray into the scene to find the first surface intersected. If any of the remaining subpaths was empty, then MLT initially must sample a point on a light source, respectively, the lens of the involved camera. If the endnotes of the new eye and light path are visible, the two subpaths are joined together, resulting in a new mutation $\overline{\mathbf{X}}_{j+1}$.

The acceptance probability of $\overline{\mathbf{X}}_{j+1}$

$$a(\mathbf{X}_j \to \mathbf{X}_{j+1}) \overset{(6.357)}{=} \min\left(1, \frac{f(\mathbf{X}_{j+1})T(\mathbf{X}_{j+1} \to \mathbf{X}_j)}{f(\mathbf{X}_j)T(\mathbf{X}_j \to \mathbf{X}_{j+1})}\right) \tag{9.120}$$

$$= \min\left(1, \frac{Q(\mathbf{X}_{j+1} \to \mathbf{X}_j)}{Q(\mathbf{X}_j \to \mathbf{X}_{j+1})}\right) \tag{9.121}$$

where the term $Q(\mathbf{X}_j \to \mathbf{X}_{j+1})$ is defined as:

$$Q(\mathbf{X}_j \to \mathbf{X}_{j+1}) \overset{def}{=} p_{d[l,m]} \sum_{j=1}^{k_a} \frac{p_a[i-1, k_a - i]}{C_i^{bd}}, \tag{9.122}$$

and C_i^{bd} corresponds to the term $C_{s,t}^*$, that is, the unweighted contribution of the path, given via the i^{th} connecting edge between the light and the eye path.

PERTURBATIONS. Let us consider a caustic, as visualized in Figure 9.28. Now, caustics are generated via small subsets of paths from $\mathbf{P}^\infty$, where paths contribute much more than average to the illuminated region. As bidirectional mutations result in relatively large changes of a path, such a strategy, applied to a *caustic path*, attempts to mutate the path outside the high-contribution region. A solution for preventing this problem, are so-called *perturbations*, i.e. slight shifts of one or more vertices or small changes of directions of a path, while leaving most of the path the same. Eric Veach defined three types of perturbations: *lens perturbations*, *caustic perturbations*, and *multi-chain perturbations*.

$\mathbf{P}^\infty$ (461)

LENS PERTURBATIONS. Let us consider a path $\overline{\mathbf{X}}_j$ generated via the Metropolis light transport algorithm. Then, MLT deletes a subpath $\mathbf{x}_{jm}, \ldots, \mathbf{x}_{jk}$ of the form $\overrightarrow{ES^*D(D|L)}$ of $\overline{\mathbf{X}}_j$—where we are mainly interested in perturbing the lens edge $\mathbf{x}_{k-1}\mathbf{x}_k$—and replaces this subpath by a new mutation. The new subpath origins from a perturbation of the old image location by shifting it a random distance R in a uniformly chosen random direction ϕ and tracing a ray, starting at the new image location, through the scene, where it holds:

$$R = r_2 \exp\left(-\ln\left(\frac{r_2}{r_1}\right) U\right) \tag{9.123}$$

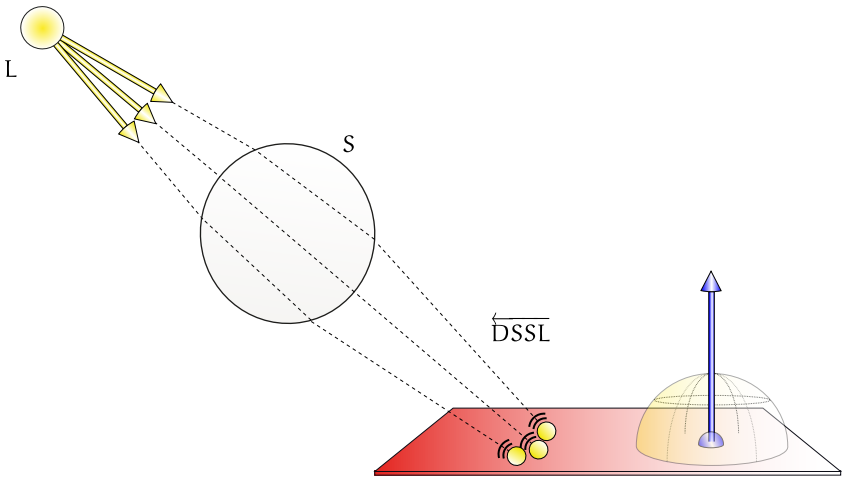


FIGURE 9.28: A CAUSTIC GENERATED VIA LIGHT TRANSPORT PATHS FROM A SMALL SUBSET OF $\mathbf{P}^\infty$. A small subset of light transport paths of $\mathbf{P}^\infty$ are responsible for the caustic seen on a diffuse surface.

with U is uniformly distributed on $[0,1]$ and r_1, r_2 are two values. The path mutation is accepted, if it reaches the same length as the original path, the specular behavior of the new mutation has not been changed with respect to the old subpath, and if the new subpath could be connected successfully with the old part of the transport path, see Figure 9.29.

CAUSTIC PERTURBATIONS. Apart from perturbations of the eye path, also perturbations of the light path are often useful, such as for example when emphasizing the appearance of caustics. Here, MLT perturbates a light subpath with suffix $\mathbf{x}_{jm}, \ldots, \mathbf{x}_{jk}$ of path $\overline{\mathbf{x}}_j$ of type $\overleftarrow{\mathrm{EDS}^*(\mathrm{D}|\mathrm{L})}$. MLT generates a new subpath starting at the vertex $\mathbf{x}_{jm}$ and the edge connecting $\mathbf{x}_{jm}$ and $\mathbf{x}_{jm+1}$ is perturbed by a random amount (θ, ϕ),

$$\theta = \theta_2 \exp\left(-\ln\left(\frac{\theta_2}{\theta_1}\right)U\right), \tag{9.124}$$

where U is uniformly distributed on the unit interval $[0,1]$, see Figure 9.30. Similar to lens perturbations, the new mutation is accepted, if it reaches the same length as the original path, the specular behavior of the new mutation has not been changed with respect to the old subpath, and if the new subpath could be connected successfully with the old part of the transport path.

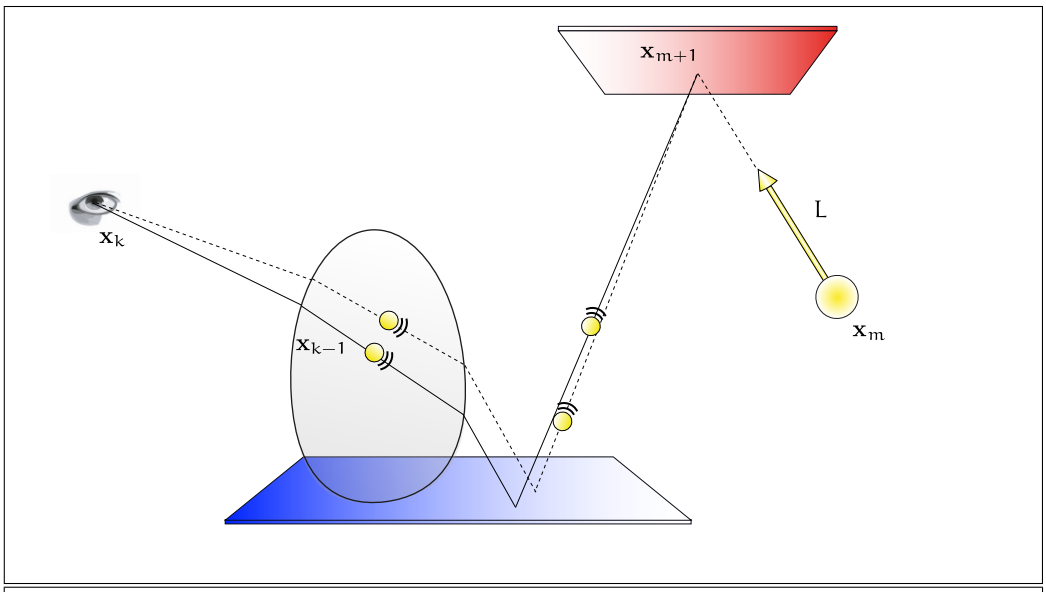


FIGURE 9.29: LENS PERTURBATION. The image location is perturbed by a moving it a random distance in a random direction on the image plane. Reprinted, by permission, from [220, Veach 1997].

REMARK 9.22 (Multi-chain Perturbations) *Neither lens nor caustic perturbations can handle caustics seen through a specular surface, that is, paths of characteristic* $\overleftarrow{\text{ES}^+\text{DS}^+\text{D}(\text{D}|\text{L})}$. *This can be handled by a combination of a lens and a caustic perturbation, also denoted as a* multi-chain perturbation. *In this case, the starting point of a transport path at the image plane is slightly shifted for perturbing the subpath* ES^+D, *and the first edge of the subpath* DS^+D *is changed by a caustic perturbation.*

Finally, let us describe lens subpath mutations, whose goal is to stratify the samples over the image plane, and also to reduce the cost of sampling by re-using subpaths. This strategy should be stratify lens subpaths across the image plane, such that every pixel receives the same number of proposed lens subpath mutations.

LENS SUBPATHS PERTURBATIONS. Also generating a large number of bidirectional samples in the initialization phase of the MLT algorithms does not ensures, that very pixel of the image plane is connected to a light source by a transport path $\overline{\mathbf{X}}_0$, that is, during the rendering step, many pixel may remain black. In order to reduce the existing noise in the image, another mutation strategy is needed, that distributes the mutated paths uniformly over the image plane. The idea is, to apply a specific mutation to the first node of an eye-subpath. For that purpose, MLT deletes a subpath $\mathbf{x}_{j_m}, \ldots, \mathbf{x}_{j_k}$ of the form $\overrightarrow{\text{ES}^*\text{D}(\text{D}|\text{L})}$

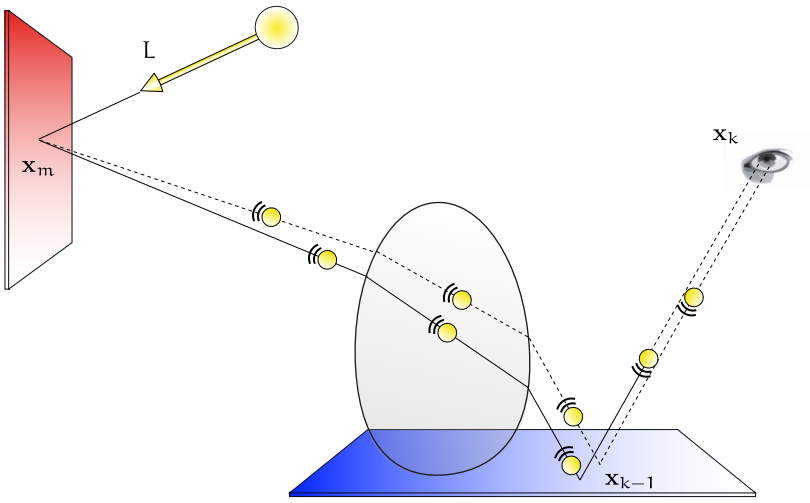


FIGURE 9.30: CAUSTIC PERTURBATION. A new path is generated by perturbing the direction of the ray from the light source by a small amount, and then tracing the perturbed ray through the same sequence of specular reflections and refractions as the original path. Reprinted, by permission, from [220, Veach 1997].

of a given path $\overline{\mathbf{X}}_j$, and replaces this subpath by a new mutation. This new eye-subpath starts at the lens and goes through a new, randomly chosen pixel. The path mutation is accepted, if the required length of the subpath is reached, the specular behavior of the new mutation corresponds to that of the original path, and if the new eye path can be connected successfully with the remaining light subpath of $\mathbf{X}_j$.

REMARK 9.23 *In practice, it has been shown, that the Metropolis algorithm is a very efficient algorithm for rendering images of scenes that include hard-to-find light transport paths. Due to [220, Veach 1997], the MLT algorithm handles general geometric and scattering models, uses little storage, and can be orders of magnitude more efficient than previous unbiased approaches. It performs especially well on problems that are usually considered difficult, e.g. those involving bright indirect light, small geometric holes, or glossy surfaces. Furthermore, it is competitive with previous unbiased algorithms even for relatively simple scenes. But as our derivation has also shown, the implementation of MLT is—due to the complex structure of the mutation strategies—quite complicated.*

9.5 THE PHOTON-MAPPING CONCEPT

With Monte Carlo path tracing, Monte Carlo light tracing, and bidirectional path tracing we have presented three global illumination algorithms that are capable to simulate all light effects within a scene to be rendered. As implementations of Markov processes, all three rendering methods are unbiased algorithms, that can handle arbitrary geometry without meshing and with low memory consumption. Even if MCPT and MCLT have problems with the generation of eye respectively light paths for simulating the one or the other light effect, with bidirectional path tracing we have a rendering technique that unifies the capability to simulate all those effects in a more than a satisfactory manner. But also BDPT has a non-negligible disadvantage: its efficiency, i.e. the run-time of the algorithm.

Section 9.1
Section 9.2
Section 9.3

We will now present another global illumination algorithm called *Photon Mapping*. The *photon-mapping concept* was introduced in [96, Jensen & Christensen 1995]. The algorithm is able to compute some global illumination effects—such as, simulating caustics, diffuse intereflections, as well as subsurface scattering of light in translucent materials, and some other light effects, such as smoke or water vapor in participating media—in scenes containing many complex objects of general form and material properties in a more efficient manner as done via MCPT, MCLT, and BDPT. Due to [97, Jensen & Christensen 2000] a photon-mapping algorithm is significantly faster, and the result looks better since the error in the photon-mapping method is of low frequency which is less noticeable than the high frequency noise of general Monte Carlo methods. But the photon-mapping concept also has a significant disadvantage: It is a biased algorithm. Since it is a consistent method, we can theoretically achieve a correct solution by increasing the number of photons. The prize we pay for this, is memory. As the method is easy to implement, any ray tracer could be extended by an efficient implementation of the photon-mapping concept.

Bias (507)
Consistent (507)

The idea behind the photon-mapping concept is to decouple the geometry of the scene from the illumination of the scene. By storing the illumination in a global data structure, the so-called *photon map*, a renderer—in the classic photon-mapping algorithm a distribution ray tracer—can use these additional data to handle arbitrarily complex geometric scene models, where the photon map works like a cache containing special light paths existing in the scene. Therefore, global illumination algorithms, based on the concept of photon-mapping, are so-called *two-pass algorithms*, where

Photon Map (751)
Distribution Ray Tracing (672)

i) in the 1^{st} pass, the *photon tracing pass*, the photon map data structures are built by tracing photons from the light sources through the scene. The hit points of a photon—together with additional information—at diffuse or slightly glossy surfaces are stored in the photon map, and

Section 9.5.1
Section 9.5.2

ii) in the 2^{nd} pass, *the rendering pass*, the scene is then rendered with the help of a Monte Carlo ray tracer, where at diffuse or slightly glossy locations, instead of

Section 9.5.3

Monte Carlo sampling, the information that is stored in the photon map is used for rendering.

Obvioulsy, the image generation via the photon-mapping concept is done from two direction: From the eye via a Monte Carlo ray tracing strategy and from the light sources in a way like Monte Carlo light tracing works.

9.5.1 PHOTON TRACING

Photon tracing can be interpreted as an essential pre-processing step of any rendering algorithm using the photon-mapping concept. It is the process in which virtual photons are emitted from light sources into the scene. These photons are then traced on their travel over the objects of the scene like in Monte Carlo light tracing. If a photon hits a diffuse or slightly glossy surface on its paths through the scene, the location where the photon has hit the object is stored in a global data structure together with a few additional information. Thus, the photon tracing pass can coarsely be partitioned in three steps, the emission, scattering, and storing of photons.

Photon (246)

PHOTON EMISSION. During the photon tracing pass, any virtual light source within a scene should—similar to a real light source—also emits a large number of photons into the scene to be rendered. Thus, we endow any photon with a certain amount of power, where this power depends on the number of emitted photons and the power of the light source.

DEFINITION 9.1 (Power of a Photon) *Let ☼ be a set of light sources emitting n photons into a scene. Then the power of a photon $\gamma_{h\nu}$, denoted as $\Phi_{\gamma_{h\nu}}$, is computed by dividing the power of all light sources $\Phi_{\text{☼}}$ by the number n of all emitted photons, that is,*

Radiant Power (249)

$$\Phi_{\gamma_{h\nu}} \stackrel{\text{def}}{=} \frac{\Phi_{\text{☼}}}{n} \qquad [W]. \tag{9.125}$$

REMARK 9.24 *To ensure, that all photons emitted into a scene are roughly provided with the same power, more photons should be emitted from brighter light sources than from dim lights.*

Due to Definition 9.1, the power of a photon only depends on the power of the light sources and the number of emitted photons. The type and form of a light source, whether diffuse point light source, directional light, spherical or squared area light source, or a light source with any arbitrary shape and emission profile, plays no role. These properties of a light source have only influence on the way like photons are sent into the scene. While point light sources and spherical light sources emit photons in all directions, directional and area light sources are limited in its possibility to emit photons in all directions.

Section 4.3

The process of emitting photons from any of these light sources can then be simulated via Monte Carlo sampling strategies. Thus for example, we can use uniform sampling over the unit sphere, or use a rejection sampling strategy for emitting photons from diffuse point light sources. By first sampling a point on the unit sphere or a squared area followed by sampling a direction from the hemisphere above this point, we can sent photons from spherical or area lights into a scene. The emission of photons from complex lights with arbitrary shape and emission profile can be simulated via densities proportional to the distribution of the light source. For a more detailed discussion on photon emission, see [95, Jensen 2001].

Uniform Sampling S^2 (193)
Section 6.5.2

PHOTON SCATTERING. After a photon is emitted by a light source, it is traced through the scene using photon tracing. In principle, photon tracing works in exactly the same way as pure-Monte Carlo light tracing.

Section 9.2

When a photon hits an object, it can either be reflected, transmitted, or absorbed. In the first two cases, the power of the photon should be scaled by the reflectivity, respectively, the transmissivity of the involved surfaces. Obviously, this can lead to photons with great discrepancy in their power. Additionally we store a photon within a global data structure, if it hits a non-specular, not too glossy surface. For reasons, which will be clear in the following, photon-mapping algorithms requires that the stored photons have approximately the same power, that is, the technique of attenuation the power of a photon due to scattering at surfaces is not a good choice. A more efficient strategy here is to use a form of Russian roulette.

Russian Roulette (200)

Via Russian roulette, a photon-mapping algorithm decides if a photon, that hits an object surface, is scattered or if it is absorbed. The absorption of a photon via Russian roulette then solves the problem of handling photons with different power. Thus, it makes no different whether we scatter n photons with the half of the power $\frac{\Phi_{\gamma_{h\nu}}}{2}$ of the incoming photons at a surface, or whether we scatter the half of photons with the power $\Phi_{\gamma_{h\nu}}$. The energy in the system remains the same.

For reducing the computational requirements for the photon tracing pass, a photon-mapping algorithm works as follows: If a photon hits an ideal specular surface, the photon tracing algorithm decides via a random variable, U, uniformly distributed on $[0, 1]$, whether the photon is reflected in the mirrored direction, $U < \rho_{dd}$, or if it should be absorbed, $U \geq \rho_{dd}$. The same holds for an ideal diffuse surface, where U is related to ρ_{dh}, except that the outgoing direction of the photon is randomly chosen over the hemisphere above the hit point and the photon is additionally stored within a global data structure, see Figure 9.31.

Uniform Distribution (180)
ρ_{dd} (338)
ρ_{dh} (338)

In the case where a photon hits a surface whose reflectivity, respectively, refractivity is described by a BSDF composed of a diffuse and specular component, a photon-mapping algorithm combines Russian roulette with an importance sampling strategy as follows:

BSDF (375)

$$\begin{array}{lcl} U \in [0, \rho_{dh}] & \Rightarrow & \text{diffuse reflection} \\ U \in (\rho_{dh}, \rho_{dd} + \rho_{dh}] & \Rightarrow & \text{specular reflection} \\ U \in (\rho_{dd} + \rho_{dh}, 1] & \Rightarrow & \text{absorption,} \end{array} \tag{9.126}$$

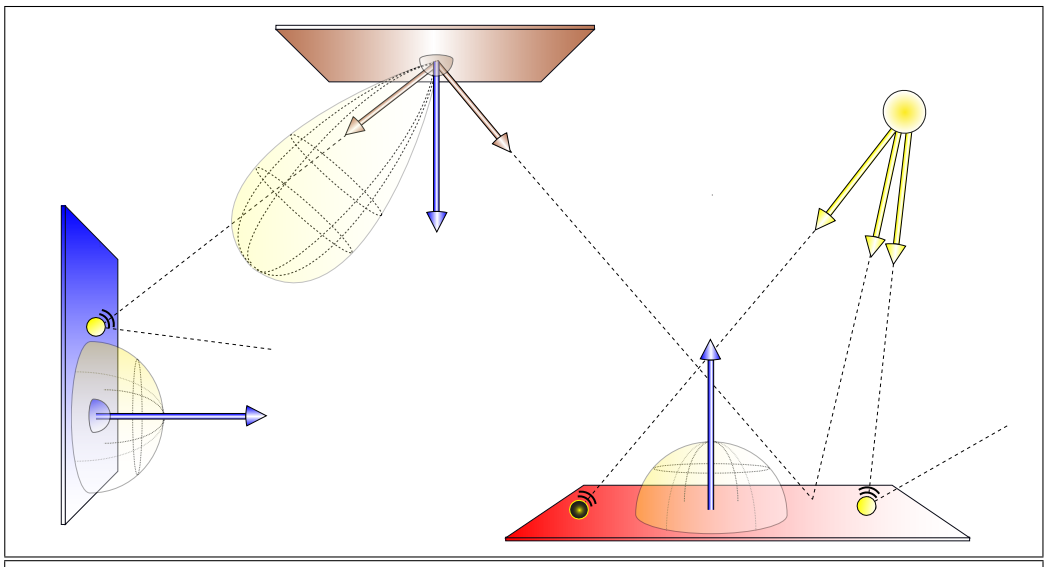


FIGURE 9.31: VISUALIZATION OF PHOTON SCATTERING. The point light source emits 3 photons with power $\frac{1}{3}\Phi_{\odot}$. The yellow colored photons are stored in the photon map with power $\frac{1}{3}\Phi_{\odot}$. The absorbed photon is black-colored, it is also stored in the photon map with power $\frac{1}{3}\Phi_{\odot}$. The upper surface is highly glossy, that is, the incident photon is not stored in the photon map. Recursion depth of the generated photon paths is ≤ 3.

where we assume that it holds: $\rho_{dd} + \rho_{dh} \leq 1$. That is, Russian roulette is used to decide whether the photon is scattered or absorbed, and via importance sampling we choose the type of reflection.

EXAMPLE 9.5 *Let us assume n photons of power $\Phi_{\gamma_{h\nu}} = \frac{\Phi_{\odot}}{n}$ hit a surface, whose BSDF f_s is composed of a diffuse and a specular component, thus: $f_s = f_s^o + f_s^\vee$. With the diffuse reflectance $\rho_{dh} = 0.5$ and the specular reflectance $\rho_{dd} = 0.25$, which implies that the surface is also 25% absorbing, then we get: The half of the photons are diffusely scattered and are also stored in the photon map, 25% of photons are scattered in the mirrored directions, and the rest of incoming photons is absorbed and stored but not propagated.*

REMARK 9.25 (Why Russian Roulette?) *As already mentioned above, the use of Russian roulette in a photon-mapping algorithm is very important. Thus, on the one hand, the unbiased property of Russian roulette guarantees that eliminating work still leads to the correct result. On the other hand, we also circumvent the problem of exponentially increasing number of photons in the scene, since the interaction of a photon at a surface could lead to generating new photons in diffuse and specular directions with*

Russian Roulette (200)

corresponding less power. This leads to less computation time and to less storage requirements of the algorithm.

REMARK 9.26 *The above selection strategy, discussed at the example of monochromatic photons, can also easily be extended to handle photons consisting of more than a single color band, and to BSDFs composed of more than two components. For a more detailed discussion, see [95, Jensen 2001].*

REMARK 9.27 *In [96, Jensen & Christensen 1995], the photon-mapping concept is almost exclusively described via the BRDF. Refraction at surfaces is not explicitly discussed. We use in our discussion of the photon-mapping the concept of the BSDF instead of the BRDF, consequently we speak often also of scattering meaning reflection or refraction at a surface.*

PHOTON STORING. As known from above, photons are only stored in a global data structure if they hit diffuse, or at least non-specular surfaces that are not too glossy. Since the probability, that a photon arrives from the mirrored direction of the incoming ray of a Monte Carlo ray tracer during the rendering pass, is zero, it makes no sense to record the intersection of a photon with a specular surface. So, no useful new information is available. Specular reflection or refraction can best be done by the involved renderer in the rendering pass, commonly a ray tracer or, as used in the classic algorithm, a distribution ray tracer. As we know from previous sections, these renderers are ideal for simulating specular reflections or specular refractions. Therefore, we only store the non-specular photon-surface interaction in a global data structure, the so-called *global photon map*.

Ray Tracing (664)
Distribution Ray Tracing (672)

DEFINITION 9.2 (Photon Map) *A* photon map *is a global data structure that stores the position* $\mathbf{s} \in \partial\mathcal{V}$*, the incoming power* $\Phi_{h\nu}$*, and the incident direction* ω_i *of the interaction of a photon at a diffuse or slightly glossy surface.*

Obviously, the photon map informs about light paths of characteristic $\overleftarrow{D_G(D_G|S_G)^*L}$, where D_G stands for diffuse/slightly glossy reflection and S_G means reflection at a specular/highly glossy surface. The photon map can easily be implemented by an array using the *photon-structure* shown in Figure 9.32.

REMARK 9.28 *Since a photon can bounce back and forth many times between surfaces before it is absorbed, it can also be stored several times in the photon map. Due to the fact that a photon is always stored in the photon map with its hit point, incident direction, and incoming power, we can use a photon to approximate the reflected illumination at several neighboring points on a surface. This is an important observation, which we will exploit in the following section where we approximate the reflected radiance at diffuse surfaces via the information stored in the photon map.*

Section 9.5.2

Due to efficiency reasons, apart from the *global photon map*, that delivers informa-

```
struct photon_hit {
// position of a photon hit
    float x, y, z;
//power packed as 4 chars
    char p[4];
//compressed incident direction
    char phi, theta
//flag used in kd-tree
    short flag;
}
```

FIGURE 9.32: DATA STRUCTURE USED IN THE PHOTON MAP. For each photon hit with a diffuse or slightly glossy surface, the x, y and z coordinates of the hit point, as well as the incoming power and the incident direction of the photon is stored. Additionally a flag is also stored. It plays only a role in the representation of the photon map by a kd-tree used in the rendering step. For representing the power of the photon, Ward's shared-exponent RGB-format is used [230, Ward 1991]. If memory is not of concern, the power of the photon can also be stored using three floats for the red, green, and blue color band.

tion about photon paths of characteristic $\overleftarrow{D_G(D_G|S_G)^*L}$, we also use a so-called *caustics photon map*, containing photons, that have been gone through at least one specular reflection before hitting a diffuse or slightly glossy surface. This corresponds to paths of characteristic $\overleftarrow{D_G S^+ L}$. Unlike all other photons, caustics photons are absorbed after entered into the caustics photon map. While the photons stored in the caustics photon map are specifically traced in direction to specular objects using maps of geometry as seen from the light sources—to reflect the effect of light bundling as precisely as possible—the paths cached via the global photon map implies that the corresponding photons were completely randomly scattered through the scene. The construction of these two photon maps is most easily achieved using two separate photon tracing steps, see Figure 9.33.

REMARK 9.29 *Note: As caustic photons are stored in the caustics photon map as well as in the global photon map, since they last bounce is at a diffuse surface, a photon-mapping algorithm must ensure not to count caustic photons twice. That is, the algorithm can not simply add the information at diffuse surfaces contained in the photon maps.*

Furthermore, a photon can also be reflected at diffuse surfaces, but this is not valid for a caustics photon. A caustics photon is absorbed if it hits a diffuse surface.

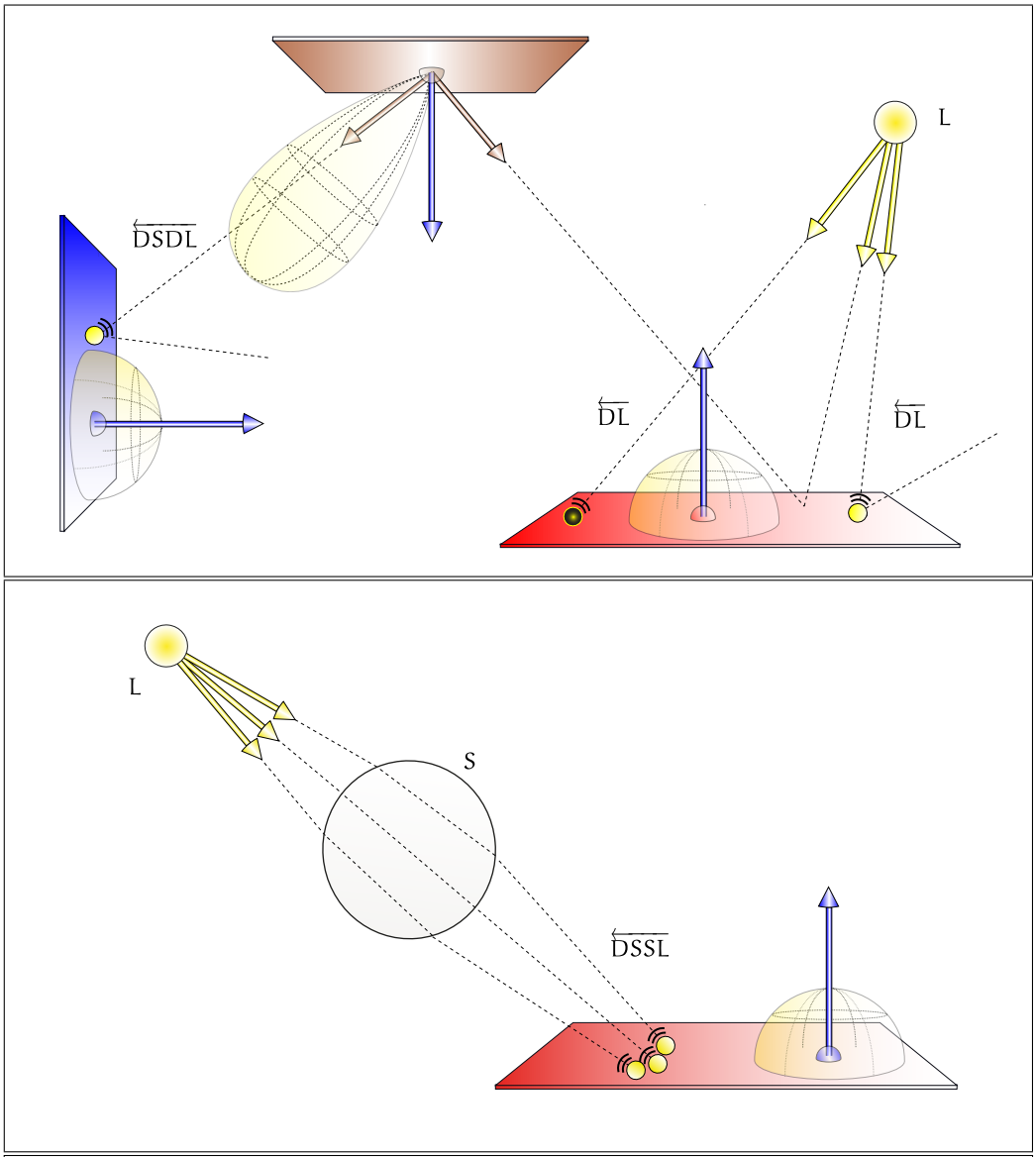


FIGURE 9.33: BUILDING THE GLOBAL AND THE CAUSTICS PHOTON MAP. Photon paths of characteristic $\overleftarrow{D(D|G|S)^*L}$ are stored in the global photon map and paths of characteristic $\overleftarrow{DS^+L}$ are stored in the caustics photon map. Thus, the yellow-colored photons and the black-colored, absorbed photon are stored in the global data structure. Note the difference in the distribution of photons in the global and the caustics photon map. While photons in the global photon map are more coarse distributed, the caustics photon map reveals a dense distribution of caustics photons.

9.5.2 RADIANCE ESTIMATE AND PREPARING THE PHOTON MAP FOR RENDERING

As we have seen, the photon map stores the incoming flux of photons, but the final rendering pass, where the image is generated with the help of a distribution ray tracer, solves the SLTEV (398) SLTEV, which describes the light transport expressed in terms of radiance. Obviously, we have here a discrepancy between the radiometric quantities radiance and flux. As we have seen in Equation (4.127), this discrepancy is particularly noticeable during refraction at interfaces between two media with different refraction indices. While radiance can change at an interface when refracted between two different media, flux, respectively power, remains unchanged. How we can combine these two different radiometric quantities to compute the reflected radiance $L_o(s, \omega_o)$ at a diffuse surface?

Reflectance Equation (321) RADIANCE ESTIMATE. Now, due to the reflectance equation, the exitant radiance at point s onto an opaque surface in direction ω_o is given by:

$$L_o(s, \omega_o) = \int_{\mathcal{H}^2_i(s)} f_r(s, \omega_i \to \omega_o)\, L_i(s, \omega_i)\, d\sigma^{\perp}_s(\omega_i). \tag{9.127}$$

Radiance (250) Interpreting the incident radiance in the scattering equation as the area and solid angle density of radiant power then we can write:

$$\begin{aligned} L_o(s, \omega_o) &\overset{(3.15)}{=} \int_{\mathcal{H}^2_i(s)} f_r(s, \omega_i \to \omega_o)\, \frac{d^2\Phi_i(s, \omega_i)}{d\mu^2(s) d\sigma^{\perp}_s(\omega_i)}\, d\sigma^{\perp}_s(\omega_i) && (9.128)\\ &= \int_{\mathcal{H}^2_i(s)} f_r(s, \omega_i \to \omega_o)\, \frac{d^2\Phi_i(s, \omega_i)}{d\mu^2(s)} && (9.129)\\ &\overset{(3.45)}{=} \int_{\mathcal{H}^2_i(s)} f_r(s, \omega_i \to \omega_o)\, dE(s, \omega_i). && (9.130) \end{aligned}$$

That is, the radiance exitant at point s in direction ω_o can be computed via the differential irradiance $dE(s, \omega_i)$ at point s multiplied with the BRDF. An estimator for $L_o(s, \omega_o)$ then has the form

$$\begin{aligned} F_N^{L_o(s,\omega_o)} &= \sum_{j=1}^{N} f_r(s, \omega_i^j \to \omega_o)\, \Delta E(s, \omega_i^j) && (9.131)\\ &= \sum_{j=1}^{N} f_r(s, \omega_i^j \to \omega_o)\, \frac{\Delta\Phi_{\gamma_{h\nu j}}(s, \omega_i^j)}{\Delta A}, && (9.132) \end{aligned}$$

where ΔA is a small area around point s on a non-specular surface, ω_i^j is the incident direction of photon $\gamma_{h\nu j}$, and $\frac{\Delta\Phi_{\gamma_{h\nu j}}(s,\omega_i^j)}{\Delta A}$ corresponds to the differential irradiance at

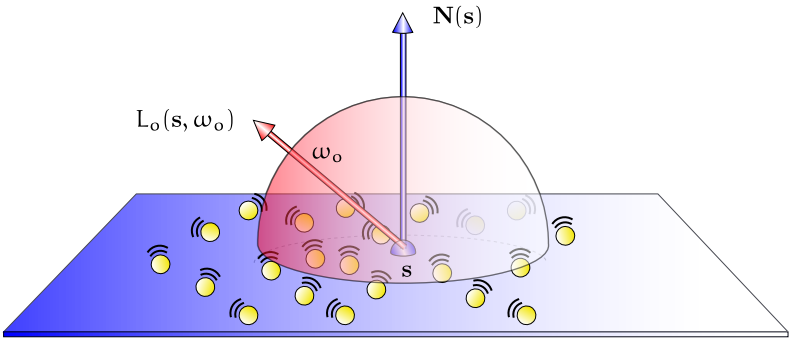


FIGURE 9.34: RADIANCE ESTIMATE. The radiance reflected at point s in direction ω_o is approximated via the local photon density around point s. Therefore, we expand a sphere with center s until it contains enough photons. The radiance $L_o(s, \omega_o)$ is then estimated via the number of photons inside the sphere, their incoming power, and the surface BSDF.

point s. To compute the differential irradiance at s, the algorithm has to find the N closest photons to point s within the photon map and has to compute the area density with respect s. Now, the N closest photons to point s are all contained in a sphere around s with radius r. That is, the algorithm expands a sphere around s until it contains N photons. By assuming that the underlying surface is locally flat around s, we can project the sphere onto the surface and approximate the photon flux density around s via the area of projection of the sphere, that is,

$$\frac{\Delta\Phi_{\gamma_{h\nu_j}}(s, \omega_i^j)}{\Delta A} = \frac{\Delta\Phi_{\gamma_{h\nu_j}}(s, \omega_i^j)}{\pi r^2}, \tag{9.133}$$

see Figure 9.34. Using this relation in Equation (9.132), then the estimator $F_N^{L_o(s,\omega_o)}$ can be written as:

$$F_N^{L_o(s,\omega_o)} = \frac{1}{\pi r^2} \sum_{j=1}^{N} f_r(s, \omega_i^j \to \omega_o)\, \Delta\Phi_{\gamma_{h\nu_j}}(s, \omega_i^j). \tag{9.134}$$

REMARK 9.30 *If the choice of radius r in the radiance estimate is too small there will not be enough photons in the considered sphere which leads to much noise in the resulting image. If r is chosen too large then many photons are contained in the sphere, which will blur the features in the lighting.*

REMARK 9.31 *Assuming that ideal diffuse, or at least diffuse surfaces exist in the scene, and all photons $\gamma_{h\nu_j}$ have the same power—due to Russian roulette—the*

above formula can furthermore be simplified, namely as,

$$F_N^{L_o(s,\omega_o)} = \frac{1}{\pi r^2} f_r(s, \omega_i^j \to \omega_o) \sum_{j=1}^{N} \Delta\Phi_{\gamma_{hvj}} \tag{9.135}$$

$$\overset{(4.161)}{=} \frac{1}{\pi r^2} \rho_{dh}(s) \sum_{j=1}^{N} \Delta\Phi_{\gamma_{hvj}} \tag{9.136}$$

$$\overset{\Phi_{\gamma_{hvj}}=\Phi_{\gamma_{hv}}}{=} \frac{\rho_{dh}(s)N}{\pi r^2} \Phi_{\gamma_{hv}}. \tag{9.137}$$

REMARK 9.32 (Sources of Bias) *As every density estimation technique results in a systematic error, the radiance estimate is the source of bias in a photon-mapping algorithm. Due to [77, Havran & al. 2005] the bias in photon maps can be classified as follows:*

i) Proximity bias, *due to a number of observations close to the photon hit point. Proximity bias causes blurring of edges. This effect can be corrected by increasing the number of photons, by using better density estimation techniques, or by filtering with higher weights for closer photons.*

ii) Boundary bias *is caused by a visible underestimation of illumination on the boundary of objects due to the overestimation of the surface area. Boundary bias results in darkening edges. The darkening on the visible surfaces is well visible.*

iii) Topological bias *is the error due to the assumption that the surface in the neighborhood of the estimated illumination is planar. The underestimation of the area for the curved surface leads to an overestimated result from the density estimation.*

REMARK 9.33 *Indeed, the radiance estimate is the source of bias in a photon-mapping algorithm, but increasing the photon density obviously results in that the radiance estimate will converge to the correct solution. This makes photon mapping to a consistent global illumination algorithm. To ensure the convergence of the radiance estimate to the correct solution it is necessary to use an infinite number of photons in the photon map as well as in the radiance estimate. Additionally, the radius has to converge to zero. These requirements can theoretically be satisfied by using* N *photons in the photon map, but only* N^β *with* $\beta \in]0,1[$ *photons in the radiance estimate. As* N *becomes infinite, both* N *and* N^β*, will become infinite, but* N^β *will be infinitely smaller than* N*, which ensures that* r *will converge to zero, [71, Hachisuka*

& al. 2008]. Then it holds:

$$\lim_{N^\beta \to \infty} F_{N^\beta}^{L_o(s,\omega_o)} = \lim_{N^\beta \to \infty} \frac{1}{\pi r^2} \sum_{j=1}^{N^\beta} f_r(s, \omega_i^j \to \omega_o)\, \Delta\Phi_{\gamma_{hvj}}(s, \omega_i) \tag{9.138}$$

$$= L_o(s, \omega_o). \tag{9.139}$$

Since in the classic photon mapping algorithm all photons are stored in memory, it is impossible to obtain a solution with arbitrary precision. In [71, Hachisuka & al. 2008] a new radiance estimate is described that fulfills the requirements of the above formula without having to store all the photons in memory.

PREPARING THE PHOTON MAP FOR RENDERING. Since the computation of the radiance estimate at a surface point s has to find the N closest photons to s in the photon map, a photon-mapping algorithm needs information about neighborhood relations between photons in the photon map. Now, the photon map is originally constructed in the photon tracing pass as a flat array. Since photons are randomly inserted in this array, we have no chance for efficiently locating the nearest photons to the considered surface point. So that the photon-mapping concept becomes the desirable fast global illumination algorithm, we need a fast data structure for implementing the photon map.

As photons are commonly not uniformly distributed on surfaces, a data structure like an octree is not suitable. A good data structure that satisfies these requirements is a k*d-tree*, where k stands for the dimension of the tree. In principle, a kd-tree is a binary search tree in k dimension, that is, in the case of the photon map, a 3-dimensional binary tree. Each node in a kd-tree contains one photon and has pointers to its left and right subtree. Additionally, we associate with all inner nodes of a kd-tree one axis-orthogonal plane that contains the photon stored in the node, and that intersects one of the spatial dimensions into two half spaces. For more detailed discussion about kd-trees, see [42, de Berg & al. 1997].

The starting point for building a kd-tree for the photon map is the smallest axis-aligned bounding box that contains all photons within the scene. In a first step, we determine the dimension with the largest extent. The photons are projected along this dimension, and the median of points in that dimension is chosen as the root node of the tree representing the photon set. All photons with corresponding smaller coordinates than the median are landing in the left subtree, and the photons with larger coordinates than the median build the right subtree. Based on the bounding boxes for the photons on the left and the right subtree, this procedure is then recursively repeated until all photons are stored in the kd-tree. Figure 9.35 shows the procedure for a kd-tree, with $k = 2$. This balancing strategy has cost $O(n \log n)$, and traversing the tree to find the closest photons costs $O(\log n)$, where n is the number of photons contained in the photon map. Look also at Figure 9.36, where we shown the distribution of photons within a 3d-tree representing the photon map used for rendering the Cornell Box.

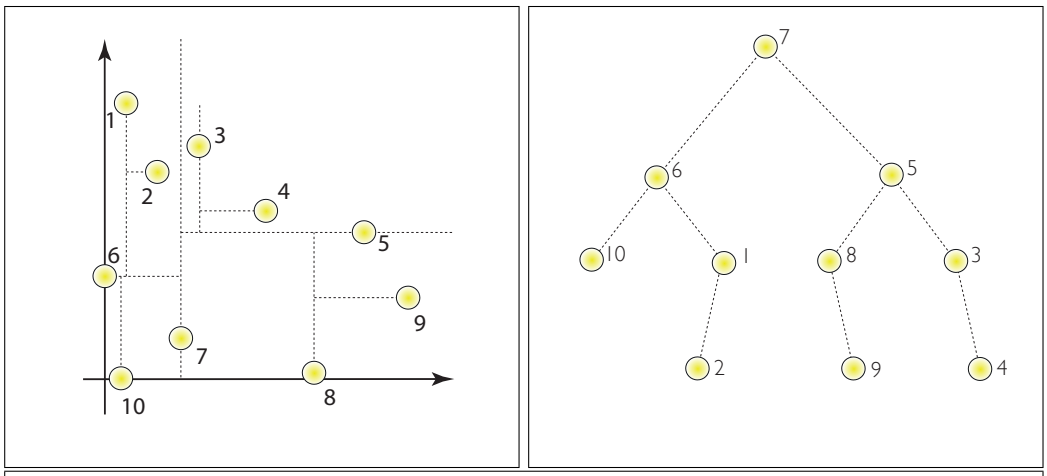


FIGURE 9.35: BALANCING STRATEGY FOR A 2D-TREE. The photon $\gamma_{hv,7}$ is the median in x-direction, the dimension of largest expansion, makes it to the root of the associated 2d-tree. The left subtree then contains all photons with smaller and the right subtree contains all photons with larger x-coordinate than $\gamma_{hv,7}$. Now, $\gamma_{hv,6}$ and $\gamma_{hv,5}$ are the medians in y-direction, the dimension of largest expansion. They are made to childrens of $\gamma_{hv,7}$, etc.

9.5.3 THE RENDERING PASS

Radiance Estimate (759)

Based on the concept of radiance estimate and supported by a simple ray tracer, we can now start to render our first images by visualizing the photon map.

Section 9.5.1
Photon Map (751)
Section 8.3

VISUALIZING THE PHOTON MAP. During the photon tracing step, the photon map is filled with information about the illumination of the scene at non-specular surfaces. Since the photon map is decoupled from the geometry of the model, this additional information about the illumination of a scene can be used by a classic Whitted-style ray tracer for rendering. Thus, the ray tracer handles specular or highly glossy object surfaces as usual, and at diffuse respectively slightly glossy surfaces, instead of tracing shadow rays, the ray tracer uses the photon map to estimate the irradiance at the shading location. That is, the reflected radiance at diffuse materials is approximated via the concept of radiance estimate.

Evidently, it is possible via this method to construct all paths between a sensor and the existing light sources in the scene—classic Whitted-style ray tracing simulates paths of characteristic $\overrightarrow{ES^*_G[D_G]L}$ and the photon map simulates paths of type $\overleftarrow{D_G(D_G|S_G)^*L}$, see Figure 9.37. That is, our simple algorithm for visualizing the photon map based on classic Whitted-style ray tracing combined with the information contained in the photon maps is a real global illumination algorithm. But due to the use of radiance estimate, the algorithm is not unbiased, it is consistent, as shown in the previous section. As you can

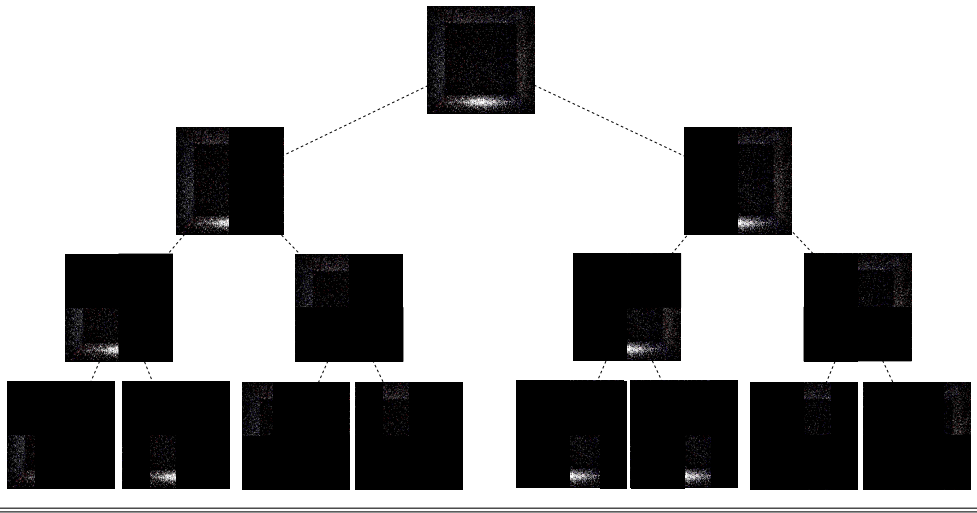


FIGURE 9.36: VISUALIZING THE SEARCH FOR A PHOTON IN THE PHOTON MAP USED FOR RENDERING THE CORNELL BOX. If the photon is not found in the root, the search continues in the left or right subtree that contain all photons with smaller or larger x-coordinate than the photon stored within the root. At nodes of depth 1 photons are distributed to their y-coordinate, and at nodes of depth 2 they are distributed with respect to the z-coordinate, and so on. So the left-most subtree—thus the left-most node of the above tree—contains the fraction of $\frac{n}{8}$ of all photons stored in the photon map, namely the photons that are located within in the lower-left front box of the Cornell Box.

see from Figure 9.38, the resulting images are covered with spotted, noisy regions. The reason for that is the number of photons stored in the photon map and the number of photons used in the radiance estimate. Therefore, direct visualizing the photon map is only recommended when blurry results in the images can be tolerated, such as in the case of fast rendering a scene for previewing.

From Figure 9.38 we also conclude, that more accurate images firstly require very large photon maps and a large number of photons in the radiance estimate. This then leads to bigger kd-trees, which requires not only more storage but also lead to more nearest-neighbor queries, that is, longer run-times.

Radiance Estimate (759)
kd-tree (757)

THE CLASSIC PHOTON-MAPPING ALGORITHM. The classic photon-mapping algorithm uses a distribution ray tracer instead of the simple Whitted-style ray tracer used in the direct visualizing algorithm of the photon map. As any other global illumination algorithm, the goal of a photon-mapping algorithms is to find a solution of the stationary light transport equation in a vacuum, that is, to solve the Fredholm type integral equation

Section 8.4
SLTEV (398)

$$L_o(s, \omega_o) = L_e(s, \omega_o) + \int_{S^2(s)} f_s(s, \omega_i \to \omega_o)\, L_i(s, \omega_i)\, d\sigma_s^{\perp}(\omega_i). \qquad (9.140)$$

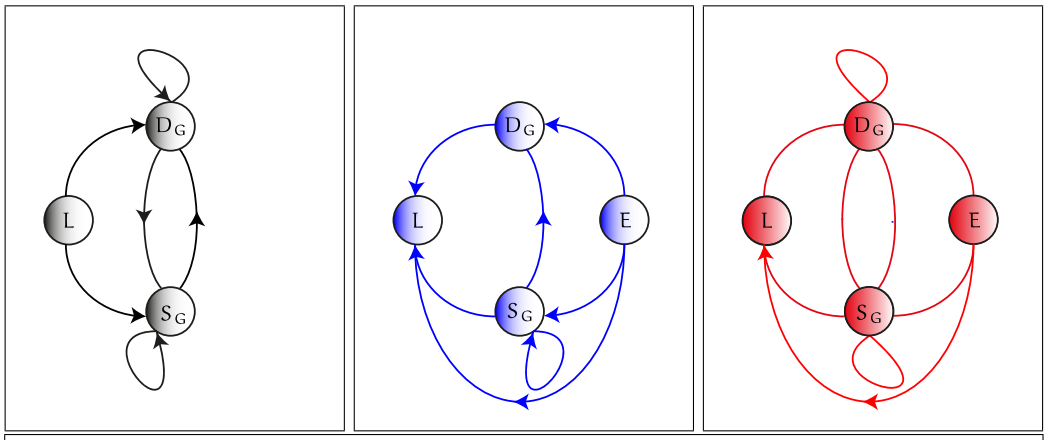


FIGURE 9.37: PATH BETWEEN THE EYE AND LIGHT SOURCES SIMULATED BY A PHOTON-MAPPING ALGORITHM. The graph in the left image visualizes the path characteristic of the global photon map. The image in the center shows the set of eye paths that can be generated by Whitted-style ray tracing. Combining these two methods leads to a global illumination algorithms, which can simulate all light effects more or less well.

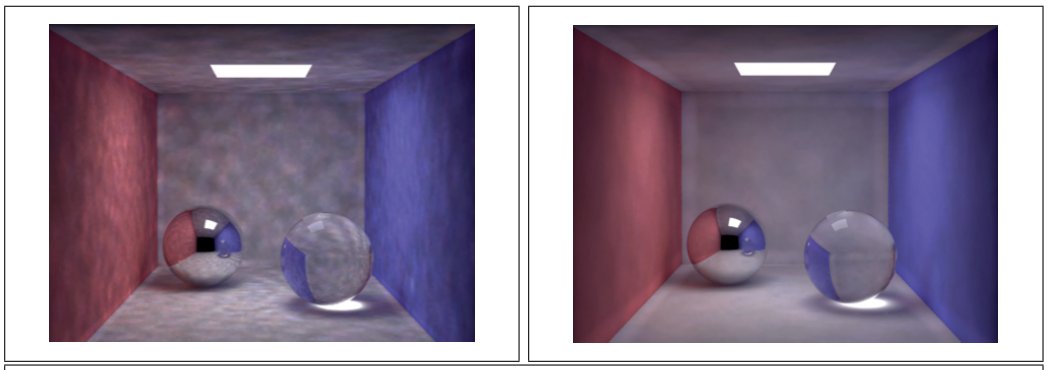


FIGURE 9.38: VISUALIZING THE PHOTON MAP. Cornell box with chrome and glass spheres. Left, direct visualization of the photon map using 10,000 photons where 100 photons are used in the radiance estimate; right, 500,000 photons are stored in the photon map and 500 are used for the radiance estimate. You can recognize spotted, noisy regions in both images, these are stronger in the left than in the image on the right side, where more photons are involved. Image courtesy of Henrik Wann Jensen, USCD.

Compared with the other global illumination algorithms, the photon-mapping concept now uses, to evaluate the above integral efficiently, apart from the decomposition of the BSDF into a diffuse and a specular component, also a decomposition of the incoming radiance. Thus, the algorithm assumes that L_i is composed of three incoming radiance types, namely,

Composition of BSDF (375)

$$L_i(\mathbf{s},\omega_i) \stackrel{\text{def}}{=} L_{i,\text{☼}}(\mathbf{s},\omega_i) + L_{i,c}(\mathbf{s},\omega_i) + L_{i,d}(\mathbf{s},\omega_i), \tag{9.141}$$

where

- $L_{i,\text{☼}}(\mathbf{s},\omega_i)$ is the incident radiance arriving at surface point $\mathbf{s}$ directly from light sources
- $L_{i,c}(\mathbf{s},\omega_i)$ is the incident caustics radiance arriving at $\mathbf{s}$ as a result of one or more specular reflections or transmissions, and
- $L_{i,d}(\mathbf{s},\omega_i)$ is the incident radiance at $\mathbf{s}$ resulting from one or more diffuse inter-object reflections or transmissions,

for an illustration see Figure 9.39.

Using the decomposition of a BSDF in a diffuse and a specular component and replacing the incoming radiance $L_i(\mathbf{s},\omega_i)$ in Equation (9.140) by Equation (9.141), then the scattering equation looks like this:

Composition of BSDF (375)

$$\begin{aligned}
L_r(\mathbf{s},\omega_o) &= \int_{S^2(\mathbf{s})} f_s(\mathbf{s},\omega_i \to \omega_o)\, L_i(\mathbf{s},\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) && (9.142)\\
&= \underbrace{\int_{S^2(\mathbf{s})} f_s(\mathbf{s},\omega_i \to \omega_o)\, L_{i,\text{☼}}(\mathbf{s},\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) +}_{L_{o,\text{☼}}(\mathbf{s},\omega_o)} \\
&\quad \underbrace{\int_{S^2(\mathbf{s})} f^{\vee}_s(\mathbf{s},\omega_i \to \omega_o)\, (L_{i,c}(\mathbf{s},\omega_i) + L_{i,d}(\mathbf{s},\omega_i))\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) +}_{L_{o,\vee}(\mathbf{s},\omega_o)} && (9.143)\\
&\quad \underbrace{\int_{S^2(\mathbf{s})} f^{o}_s(\mathbf{s},\omega_i \to \omega_o)\, L_{i,c}(\mathbf{s},\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i) +}_{L_{o,c}(\mathbf{s},\omega_o)} \\
&\quad \underbrace{\int_{S^2(\mathbf{s})} f^{o}_s(\mathbf{s},\omega_i \to \omega_o)\, L_{i,d}(\mathbf{s},\omega_i)\, d\sigma^{\perp}_{\mathbf{s}}(\omega_i)\,.}_{L_{o,d}(\mathbf{s},\omega_o)}
\end{aligned}$$

Here, we have split the scattering equation in four different scattering terms: A scattering term for computing direct illumination, a scattering term for computing indirect

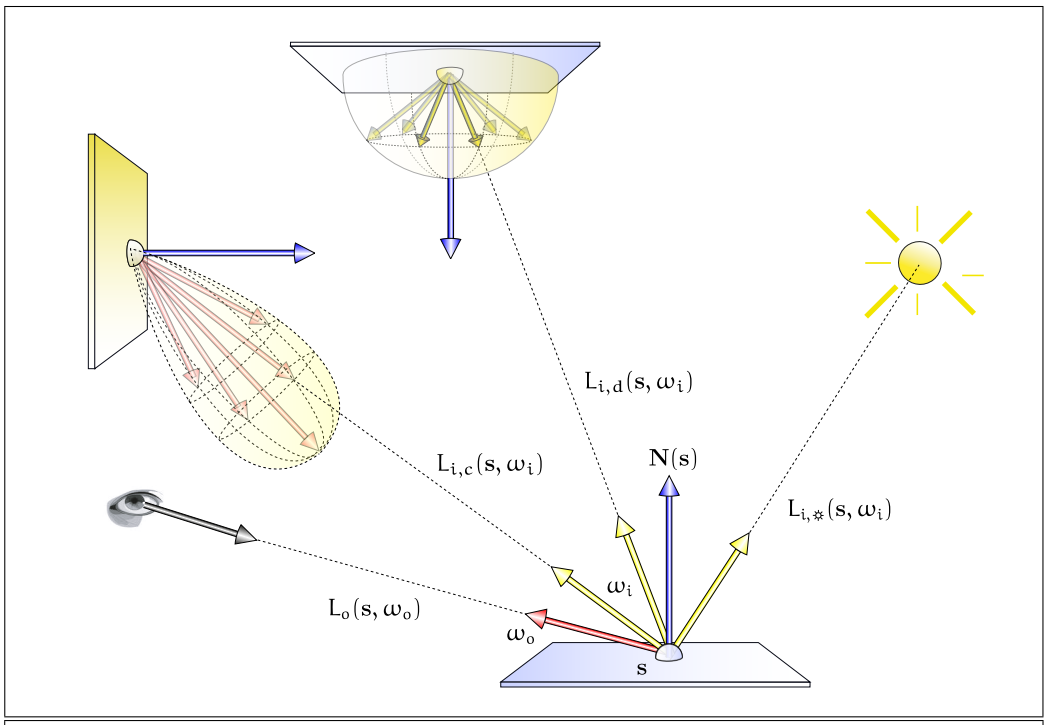


FIGURE 9.39: EVALUATING THE SCATTERING TERM. The decomposition of the incident radiance L_i in the SLTEV into an incoming direct, an incident caustic, and an incoming indirect diffuse radiance component—$L_{i,☼}(s, \omega_i)$, $L_{i,c}(s, \omega_i)$ and $L_{i,d}(s, \omega_i)$—also implies the computation of the scattering part of the SLTEV depending on these incident radiance types.

specular illumination, a scattering term, that specifies the computation of caustics, and a scattering term that simulates multiple diffuse scattering.

The scattering equation in this form is to evaluated via the involved distribution ray tracer using information stored in the photon map. But as we have seen in Section 8.4, distribution ray tracing suffers from the exponentially increase in the number of rays. Thus, the photon-mapping concept uses a few simplifications.

First, the algorithm distinguish between an accurate and an approximate computation. An accurate computation of Equation (9.143) is only used on the first bounce of the primary ray with an object surface, if the surface is seen via a few specular reflections, or if the distance between the ray-surface interaction is closer to the ray origin than a given threshold. Equation (9.143) is only approximately evaluated if the ray contributes little to the pixel radiance or if the ray intersecting a surface has been reflected diffusely. It is clear: the goal of the algorithms is to evaluate the accurate computation of Equation (9.143) as infrequently as possible.

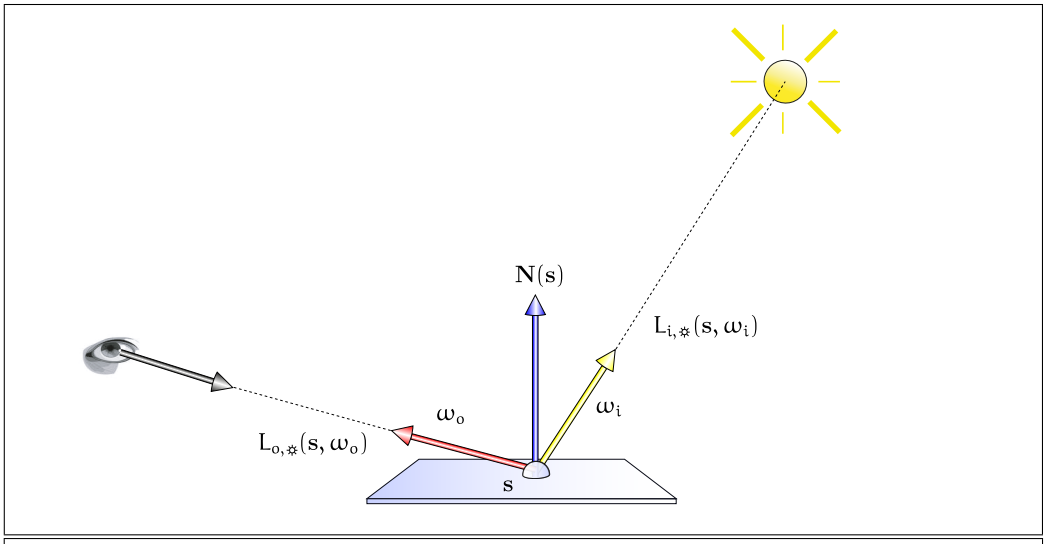


FIGURE 9.40: EVALUATING THE SCATTERING TERM FOR COMPUTING DIRECT ILLUMINATION. The most largest part to the illumination at a surface point comes frequently from light sources. This is also the reason why the direct illumination at point s has to be computed exactly via generating shadow rays.

Let us now discuss the evaluation of each of these scattering terms in more detail.

9.5.3.1 EVALUATING THE SCATTERING TERM FOR COMPUTING DIRECT ILLUMINATION

The scattering term for direct illumination, thus the scattered radiance due to direct illumination, is given by:

$$L_{o,\sun}(s, \omega_o) \stackrel{def}{=} \int_{S^2(s)} f_s(s, \omega_i \rightarrow \omega_o)\, L_{i,\sun}(s, \omega_i)\, d\sigma_s^{\perp}(\omega_i). \tag{9.144}$$

This term is frequently the most important part of the scattered radiance, it can simply be computed via distribution ray tracing by generating shadow rays in direction to the light sources, see Figure 9.40. As the point of interest is seen directly by the eye, $L_{o,\sun}(s, \omega_o)$ has to be evaluated accurately. The approximate evaluation of the direct illumination results from a radiance estimate based on the photon map, where no shadows rays are generated.

In [95, Jensen 2001], also an approach is discussed that uses so-called *shadow photons*. It can lead to considerable speedups in scenes with large area light sources.

9.5.3.2 EVALUATING THE SCATTERING TERM FOR COMPUTING INDIRECT SPECULAR AND GLOSSY ILLUMINATION

The scattering term for indirect specular illumination, thus the scattered radiance at specular surfaces due to indirect illumination, is given by

$$L_{o,\vee}(\mathbf{s},\omega_o) \stackrel{def}{=} \int_{S^2(\mathbf{s})} f_s^{\vee}(\mathbf{s},\omega_i \rightarrow \omega_o)\,(L_{i,c}(\mathbf{s},\omega_i) + L_{i,d}(\mathbf{s},\omega_i))\,d\sigma_{\mathbf{s}}^{\perp}(\omega_i), \qquad (9.145)$$

see Figure 9.41.

This integral is evaluated using distribution ray tracing, where from reasons of efficiency an importance sampling strategy with respect to the involved BSDF can be used for sampling an incoming direction over the unit sphere. Here, the photon map is not used, since the integral is strongly dominated by the specular component of the BSDF.

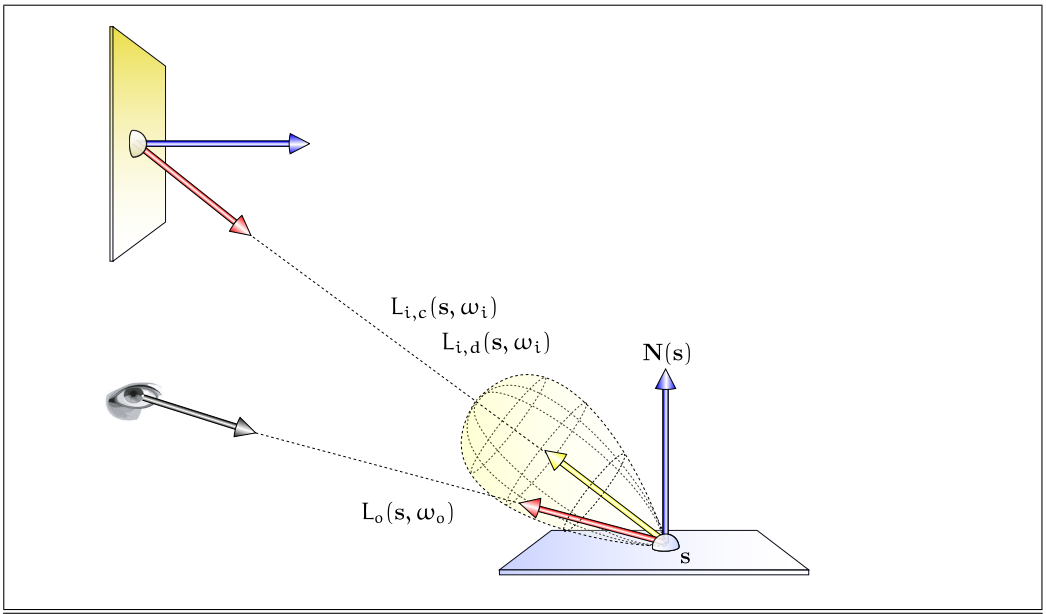


FIGURE 9.41: EVALUATING THE SCATTERING TERM FOR COMPUTING INDIRECT SPECULAR AND GLOSSY ILLUMINATION. If the primary ray hits a point $\mathbf{s}$ at a specular or a gloss surface, then the exitant radiance in direction ω_o is evaluated via Monte Carlo ray tracing. Due to reasons of efficiency an importance sampling strategy with respect to the involved BSDF can be used for sampling an incoming direction over the unit sphere. Here, the photon map is not used, since the integral is strongly dominated by the specular and gloss component of the BSDF.

9.5.3.3 EVALUATING THE SCATTERING TERM FOR COMPUTING CAUSTICS

The scattering term for computing caustics, thus the scattered radiance at diffuse surfaces due to scattering at one or more specular surfaces, is given by

$$L_{o,c}(\mathbf{s},\omega_o) \stackrel{def}{=} \int_{S^2(\mathbf{s})} f_s^o(\mathbf{s},\omega_i \rightarrow \omega_o)\, L_{i,c}(\mathbf{s},\omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i), \tag{9.146}$$

see Figure 9.42.

When an accurate value of $L_{o,c}(\mathbf{s},\omega_o)$ is required, the integral is evaluated via radiance estimate using the information from the caustics photon map. An approximate solution is given by radiance estimate of the global photon map.

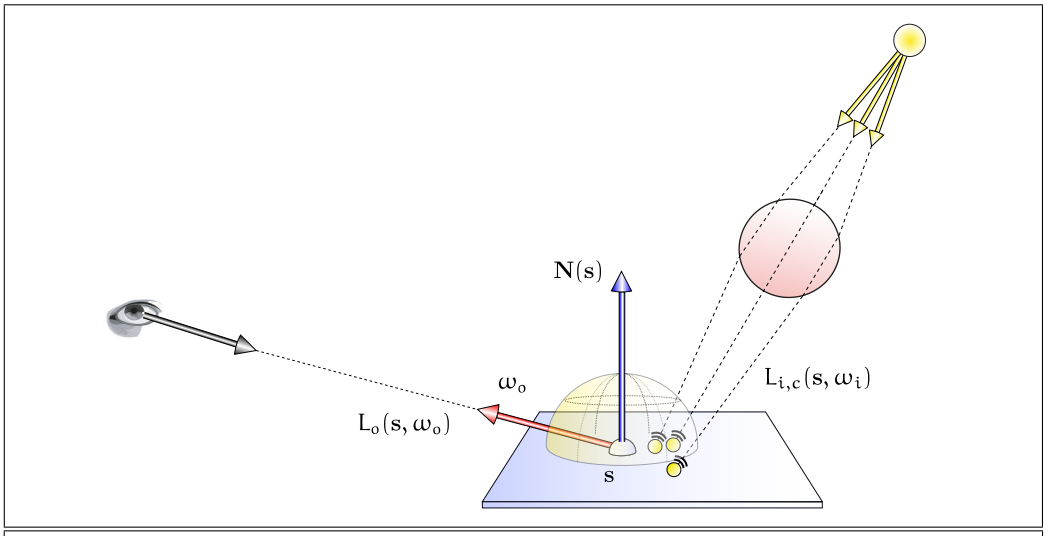


FIGURE 9.42: EVALUATING THE SCATTERING TERM FOR COMPUTING CAUSTICS.
Caustics that are seen directly by the eye are rendered using the caustics photon map via a radiance estimate around surface point s that was hit by the primary ray starting at the eye.

9.5.3.4 EVALUATING THE SCATTERING TERM FOR MULTIPLE DIFFUSE SCATTERING

The scattering term for multiple diffuse scattering, thus the scattered radiance at diffuse surfaces due to scattering at one or more diffuse surfaces, is given by

$$L_{o,d}(\mathbf{s},\omega_o) \stackrel{def}{=} \int_{S^2(\mathbf{s})} f_s^o(\mathbf{s},\omega_i \rightarrow \omega_o)\, L_{i,d}(\mathbf{s},\omega_i)\, d\sigma_{\mathbf{s}}^{\perp}(\omega_i), \tag{9.147}$$

see Figure 9.43.

An accurately value of the integral can be computed via distribution ray tracing, while an approximate value is given via the radiance estimate from the global photon map, which contains the direct, indirect, and caustic illumination contributions.

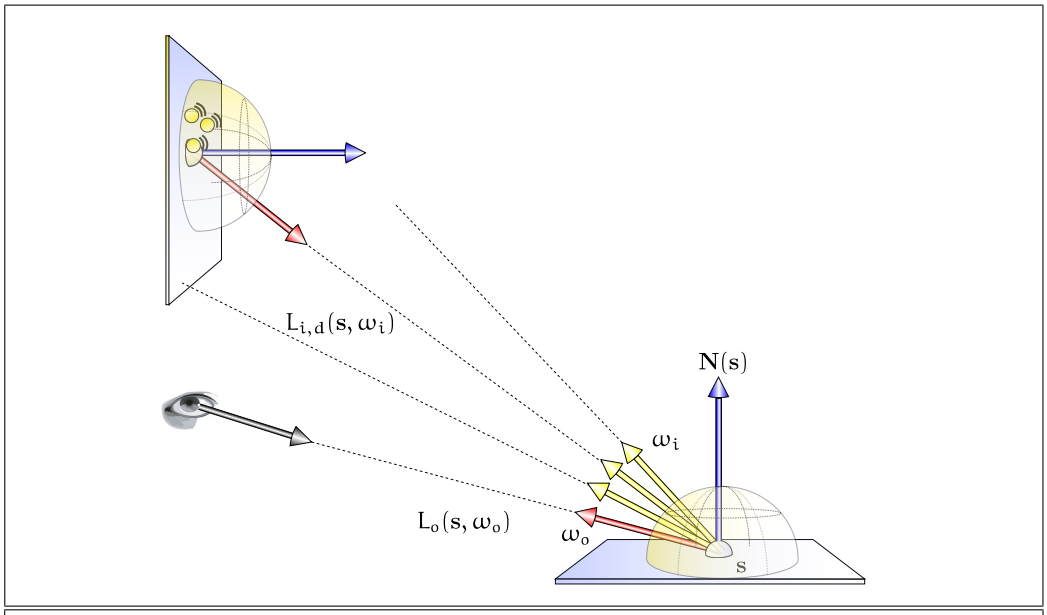


FIGURE 9.43: EVALUATING THE SCATTERING TERM FOR MULTIPLE DIFFUSE SCATTERING. The scattered radiance $L_o(s, \omega_o)$ at diffuse surfaces due to scattering at one or more diffuse surfaces, is exactly evaluated via distribution ray tracing where a number of samples are used to estimate the incident radiance. For an approximate value a radiance estimate from the global photon map can be used.

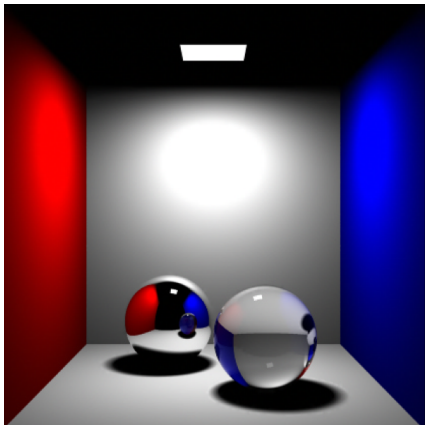
With the use of a distribution ray tracer for computing the direct as well as the indirect illumination, except for caustics, we circumvent the problem of generating a very large number of photons which are needed for rendering accurate images.

REMARK 9.34 (Final Gathering) *An extension of the classic photon mapping algorithm is* final gathering. *Final gathering is a technique from computer graphics that is often used in connection with a global illumination algorithm to enhance the quality of an image, in particular with respect to indirect diffuse illumination. Based on a coarse precomputed solution of the light distribution in a scene, final gathering computes a more accurate per-pixel illumination value via a Monte Carlo ray tracing strategy.*

Chapter 8

Starting at the hit point s *of a primary ray with an object in the scene, the algorithm distributes a large number,* $200 - 5,000$*, of so-called* final gathering rays, FGRs, *over the stratified hemisphere around point* s *into the scene. At the hit point*

Section 6.6.4



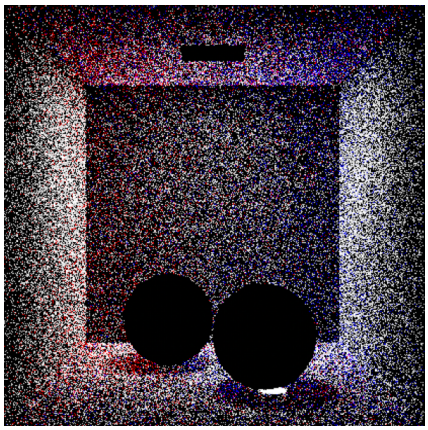


FIGURE 9.44: VISUALIZATION OF THE PHOTON MAP. The Cornell box with glass and chrome spheres. Left, the ray traced image with direct illumination as well as specular reflection and transmission. In the right image, the photons are visualized in the associated photon map. You can recognize the caustic under the glass sphere as a high density of photons induced by the refraction at the sphere. You can also see that photons at the floor and the ceiling as well as at the walls are shaded with the corresponding colors. Note: the reflective and refractive objects are painted black, since they correspond to location in the scene, for which no photons are stored in the photon map. Image courtesy of Henrik Wann Jensen, USCD.

of an FGR with a diffuse or not too glossy object surface the algorithm then computes a radiance estimate via the global photon map and uses this indirect diffuse contribution as the incident indirect light, $\mathrm{L}_{\mathrm{i,d}}(\mathbf{s}, \omega_{\mathrm{i}})$*, for the evaluation of the exitant radiance,* $\mathrm{L}_{\mathrm{o,d}}(\mathbf{s}, \omega_{\mathrm{o}})$*, at point* $\mathbf{s}$ *in direction* ω_{o}*, see Figure 9.45. Usually, final gathering is limited to a single bounce only, but multiple bounces of a final gathering ray is possible, with the consequence of longer run-times. Note: The case where final gathering is limited to paths of length* 1 *corresponds to the classic algorithm, where the ray tracer distributes a large number of rays at the hit point of the primary ray into the scene, and approximates the indirect incident radiance via a radiance estimate of the secondary rays at diffuse hit points.* Section 9.5.2

A better sampling strategy for distributing the final gathering rays such as a stratification of the upper hemisphere around point $\mathbf{s}$*, indicated above, could be to sample the incident directions* ω_{i} *according to a distribution that matches the shape of the integrand of Equation (9.147), that is, to sample according the involved BSDF or the incident radiance* $\mathrm{L}_{\mathrm{i,i}}$ *or to apply a multiple importance sampling strategy.* Section 6.6.9

For diffuse scenes where the indirect illumination various slowly, final gathering often improves the quality of the global illumination solution. Here, the algorithm eliminates photon map artifacts such as low frequency noise and dark corners.

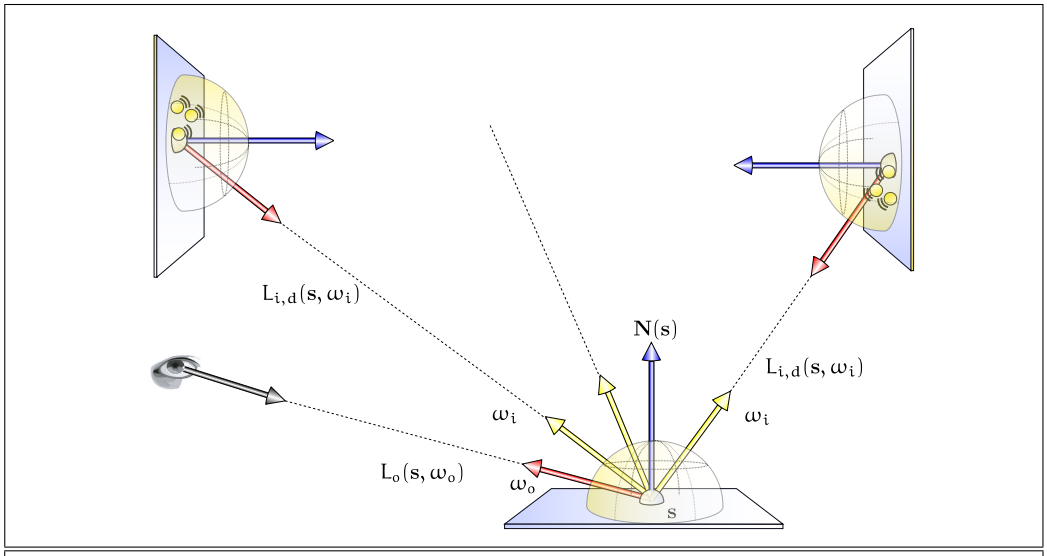


FIGURE 9.45: FINAL GATHERING IN THE PHOTON-MAPPING ALGORITHM. For evaluating the exitant radiance $L_o(s, \omega_o)$ at point s, the final gathering step traces a large set of rays to sample the incident radiance at s. Note, the final gathering rays are not recursively traced if they hit a surface object. Light contribution from hit points of FGRs with specular surfaces are not accounted for and contributions from diffuse surfaces are computed via radiance estimates from the photon map.

REMARK 9.35 (Photon Mapping vs Monte Carlo Path Tracing) *The main benefit of the photon-mapping concept compared with Monte Carlo path tracing is efficiency. The price we pay for that is the extra memory used to store the photons. For most scenes photon-mapping algorithms are significantly faster, and the results look better since the error in the methods are of low frequency which is less noticeable than the high frequency noise of general Monte Carlo methods.*

Section 9.1

Unlike path tracing, bidirectional path tracing and Metropolis light transport, photon mapping is not an unbiased *rendering algorithm. The photon-mapping algorithm is biased. If we use to few photons, it will create artifacts, no matter how long the render time is for the second pass. As already mentioned the algorithm can be made consistent, that is, it can converges to a correct solution to the rendering equation under the assumption, that infinitely many photons are used in the photon map.*

Consistent 507

REMARK 9.36 (Photon Mapping vs Finite Element Radiosity Methods) *Compared with finite element radiosity, photon maps have the advantage that no meshing is required. The radiosity algorithm is faster for simple diffuse scenes but as the complexity of the scene increases, photon-mapping tends to scale better. Additionally, photon-mapping*

Chapter 10

methods also handle non-diffuse surfaces and caustics.

9.6 INSTANT GLOBAL ILLUMINATION

Recall from algorithms like Monte Carlo path tracing or bidirectional path tracing, they all suffer from the fact that for producing noise free or at least noise reduced images an enormous number of primary rays have to be traced through a pixel. Now tracing these rays on its travel through a scene leads to long rendering times. This is perhaps acceptable for applications such as high-quality, offline computer graphics, and physically-correct rendering but it is surly not acceptable for interactive graphical applications. The traditional ray tracing algorithms are completely unsuitable for interactive ray tracing, even if the illumination of the scenes which we will render is not very complex. Therefore, we now describe a global illumination algorithm that allows to simulate the most important global illumination effects at realtime rates: *Instant global illumination*.

Instant global illumination, also abbreviated *IGI*, [228, Wald & al. 2002], [226, Wald & al. 2003] and [225, Wald 2004], was developed by the graphics group around Philipp Slusallek at Saarland University in 2002 for the availability of realtime ray tracing for achieving interactive global illumination.

Based on the idea of *virtual point lights* from [103, Keller 1997], instant global illumination generates a small number of light-carrying paths at the light sources and traces these paths on their travel through the scene. At the hit points of such a light path with an object, IGI places a so-called *virtual point light, VPL*, that can now illuminate the entire scene and not just a single pixel. The scene is then rendered with a conventional ray tracer, where shadow rays in direction to the VPLs are used to approximate the indirect radiance distribution in the scene, see Figure 9.46. If a VPL is visible from the point to be shaded, their contribution is accumulated to the indirect illumination component. The direct lighting needs to be done in the usual manner, by recursively tracing rays to account perfect specular reflection and refraction. So, IGI can simulate all of the most important kinds of illumination, thus, hard as well as smooth shadows, direct and indirect illumination, reflections, refraction, and—combined with photon mapping—even some simple forms of caustics, without the noise that is characteristic for MCPT and BDPT, see Figure 9.47.

CREATING THE VIRTUAL POINT LIGHT SOURCES IN INSTANT GLOBAL ILLUMINATION. In the following discussion, we describe instant global illumination as a variant of bidirectional path tracing, where in a first step, starting at the eye, $\mathbf{z}_0$, an eye path

$$\overline{\mathbf{z}} = \mathbf{z}_0\mathbf{z}_1 \tag{9.148}$$

of length one is created. This is done as in BDPT, where we sample $\mathbf{z}_0$ due to the BDPT (717)

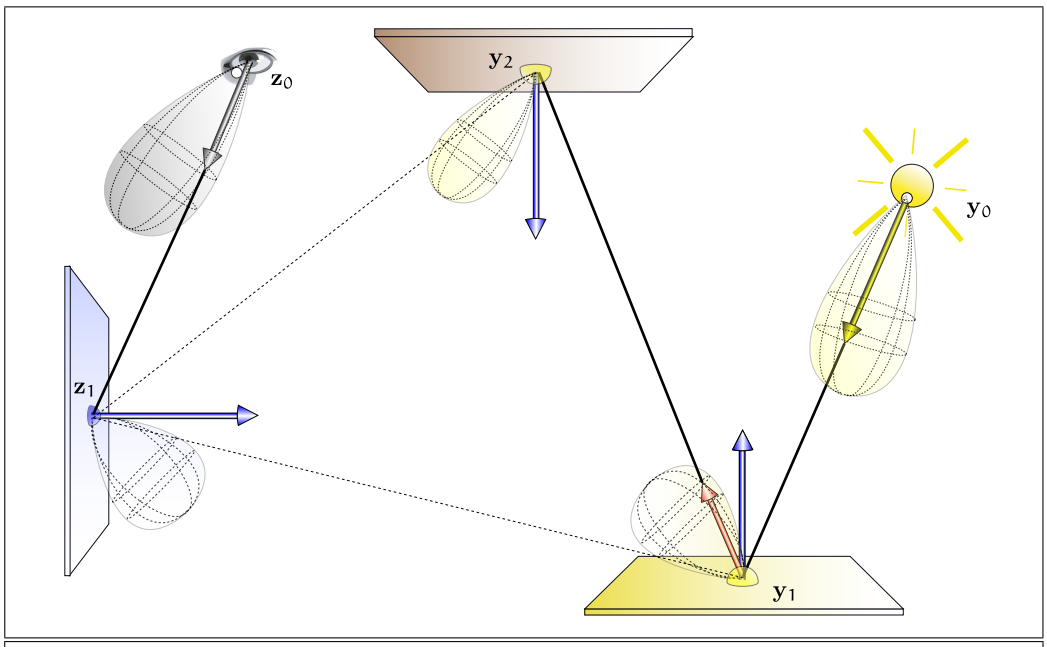


FIGURE 9.46: TRANSPORT PATHS IN INSTANT GLOBAL ILLUMINATION. An eye path $\overline{\mathbf{z}} = \mathbf{z}_0\mathbf{z}_1$ and an light path $\overline{\mathbf{y}} = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2$. At all nodes of the light path, $\overline{\mathbf{y}}$, instant global illumination deposits a virtual point light. Via shadow rays to the virtual point lights, IGI gathers the contributions from the virtual point lights for shading the surface point $\mathbf{z}_1$.

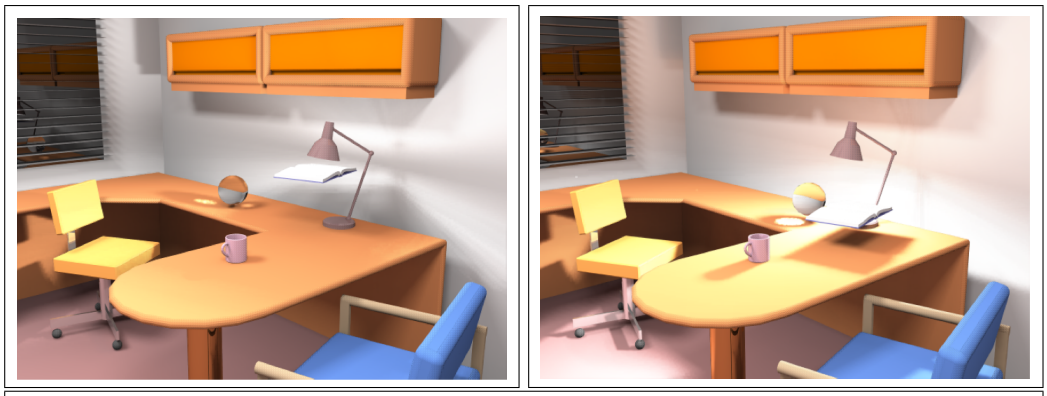


FIGURE 9.47: SCENES RENDERED WITH INSTANT GLOBAL ILLUMINATION. An animated office scene, where a glass ball is rolling over the table, and a book is moved towards the light source. Note the caustic due to the glass ball, and the indirect illumination from the book producing smooth shadows on the wall. All affects are even correctly reflected in the window. All scenes run interactively at several frames per second. Image courtesy by Ingo Wald.

probability density function, p_{μ^2}, uniformly on the pixel area $\square_j$ with respect to the Lebesgue area measure μ^2 and generate a ray according to the PDF, $p_{\sigma^\perp}$, with respect to the projected solid angle measure $\sigma^\perp$. Due to Equation (9.49), the density for sampling the eye path $\bar{z} = z_0 z_1$ is then given by: PDF (176) $p_{\sigma^\perp}$ (88)

$$p(\bar{z}) = p_{\mu^2}(z_0)\, p_{\sigma^\perp}(z_0 \to z_1 | z_0)\, G(z_0 \leftrightarrow z_1). \tag{9.149}$$

Slightly modified to bidirectional path tracing, IGI then generates a small number m of light paths,

$$\bar{y}_i = y_{i_0} \dots y_{i_n}, \qquad 1 \le i \le m, \tag{9.150}$$

where y_{i_0} is sampled via the PDF p_{μ^2} at one of the light sources in the scene, and the successor y_{i_j} of $y_{i_{j-1}}$ is sampled by choosing and casting a ray from the current subpath to the new sampled vertex with respect to the projected solid angle measure $\sigma^\perp$. The density for sampling the $(j+1)^{th}$ vertex of one of the subpaths $\bar{y}_i$ is then given by the conditional density that y_{i_j} is chosen given $y_{i_{j-1}}$ multiplied with the PDF for generating the associated subpath $y_{i_0} \dots y_{i_{j-1}}$, that is,

$$p(\bar{y}_i) \overset{(9.47)}{=} p_{\mu^2}(y_{i_0}) \prod_{j=1}^{n} \left(p_{\sigma^\perp}(y_{i_{j-1}} \to y_{j_i} | y_{i_{j-2}} \to y_{i_{j-1}})\, G(y_{i_{j-1}} \leftrightarrow y_{j_i}) \right). \tag{9.151}$$

REMARK 9.37 *Instead to sample the starting point y_{i_0} uniformly on the area of a light source, y_{i_0} should be sampled due to the amount of power emitted by a light source compared with the existing other light sources. This can improve the efficiency of IGI.*

Except of generating light paths $\bar{y}_i$, instant global illumination also places at the surface location of each vertex $y_{i_j}, 1 \le j \le n, 1 \le i \le m$ a so-called *virtual point light*. Connecting the i^{th} light path $\bar{y}_i$ with the eye path $\bar{z}$ via the edge $y_{i_n} z_1$ then results in a transport path

$$\begin{aligned} \bar{x}_i &= \bar{y}_i \bar{z} && (9.152)\\ &= y_{i_0} \dots y_{i_n} z_1 z_0. && (9.153) \end{aligned}$$

Obviously, the PDF from which $\bar{x}_i$ can be sampled from is given by the product of $p(\bar{y}_i)$ and $p(\bar{z})$, that is,

$$\begin{aligned} p(\bar{x}_i) &= p(\bar{y}_i) \cdot p(\bar{z}) && (9.154)\\ &= p_{\mu^2}(y_{i_0}) \prod_{j=1}^{n_i} \left(p_{\sigma^\perp}(y_{i_{j-1}} \to y_{j_i} | y_{i_{j-2}} \to y_{i_{j-1}})\, G(y_{i_{j-1}} \leftrightarrow y_{j_i}) \right) \cdot \\ &\quad p_{\mu^2}(z_0)\, p_{\sigma^\perp}(z_0 \to z_1 | z_0)\, G(z_0 \leftrightarrow z_1), && (9.155) \end{aligned}$$

see Figure 9.48, and the associated measurement contribution function $f_j(\overline{\mathbf{x}}_i)$ has the form

$$\begin{aligned} f_j(\overline{\mathbf{x}}_i) \quad = \quad & L_e(\mathbf{y}_{i_0}) \cdot \mathcal{G}(\mathbf{y}_{i_0} \leftrightarrow \mathbf{y}_{1_i}) \cdot \prod_{j=1}^{i_n-1} \left(f_s(\mathbf{y}_{i_{j-1}} \rightarrow \mathbf{y}_{i_j} \rightarrow \mathbf{y}_{i_{j+1}}) \cdot \mathcal{G}(\mathbf{y}_{i_j} \leftrightarrow \mathbf{y}_{i_{j+1}}) \right) \cdot \\ & f_s(\mathbf{y}_{i_n-1} \rightarrow \mathbf{y}_{i_n} \rightarrow \mathbf{z}_1) \cdot \mathcal{G}(\mathbf{y}_{i_n} \leftrightarrow \mathbf{z}_1) \cdot f_s(\mathbf{z}_0 \rightarrow \mathbf{z}_1 \rightarrow \mathbf{y}_{i_n}) \cdot \\ & \mathcal{G}(\mathbf{z}_0 \leftrightarrow \mathbf{z}_1) \cdot W_e^j(\mathbf{z}_0), \end{aligned} \tag{9.156}$$

see Figure 9.48, and for details check the discussion in Section 9.3.

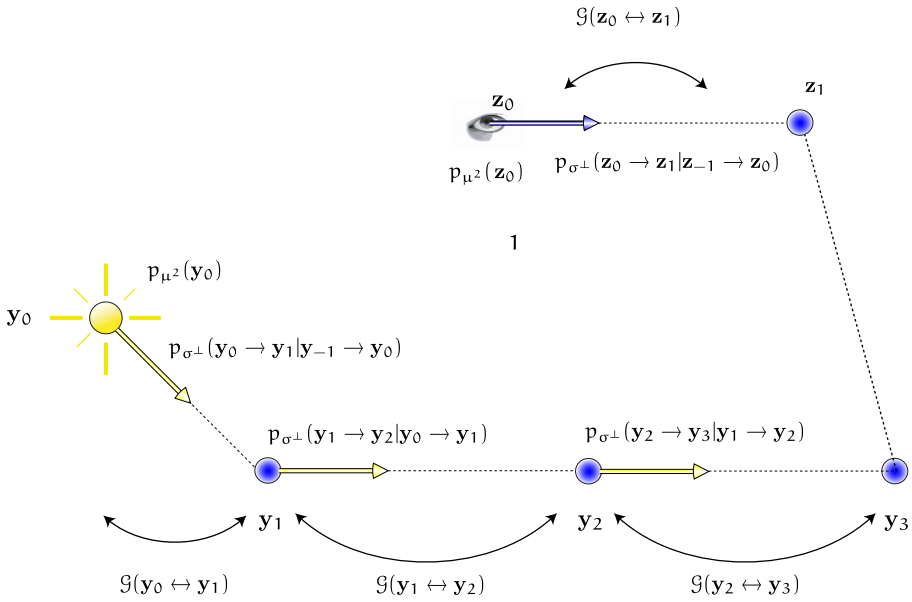


FIGURE 9.48: THE PROBABILITY DENSITY FUNCTION FOR GENERATING A TRANSPORT PATH. Shown is a transport $\overline{\mathbf{x}}_5 = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2\mathbf{y}_3\mathbf{z}_1\mathbf{z}_0$ path of length 5. The probability of computing the light path $\overline{\mathbf{y}} = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2\mathbf{y}_3$ is given by the product of the probabilities for sampling the point $\mathbf{y}_0$ and the direction ω_i, ω_i^1 and ω_i^2. The same holds for the eye path $\overline{\mathbf{x}} = \mathbf{x}_0\mathbf{x}_1$. The probability for computing the connecting edge $\mathbf{y}_3 \leftrightarrow \mathbf{z}_1$ is one.

Using these two relations, then the contribution of a transport path $\overline{\mathbf{x}}_i$ to shading

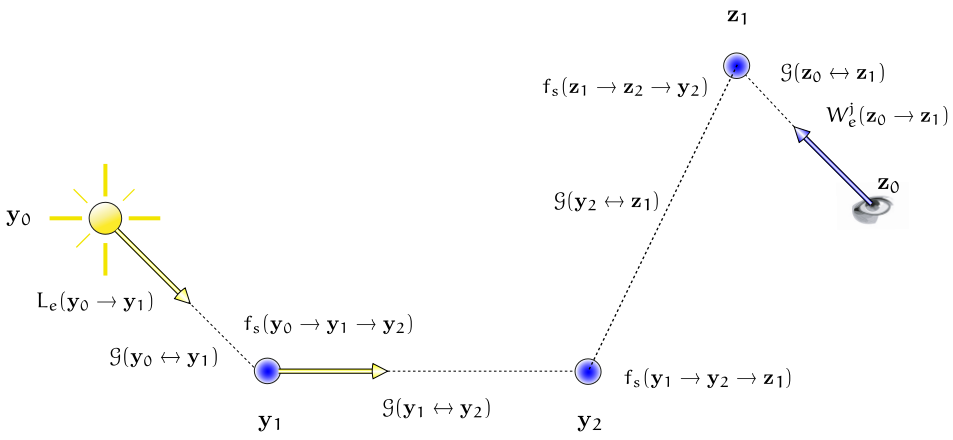


FIGURE 9.49: COMPUTING THE MEASUREMENT CONTRIBUTION FUNCTION FOR A TRANSPORT PATH. The transport path $\overline{\mathbf{x}}_4 = \mathbf{y}_0\mathbf{y}_1\mathbf{y}_2\mathbf{z}_1\mathbf{z}_0$ is composed of a light path of length 2 and an eye path of length 1. The measurement contribution function for the path $\overline{\mathbf{x}}_4$ can be computed via the product of the geometry terms, the BSDFs as well as the emitted radiance and importance along the transport path.

the pixel $\square_j$ is given by

$$\frac{f_j(\overline{\mathbf{x}}_i)}{p(\overline{\mathbf{x}}_i)} = \left(\frac{L_e(\mathbf{y}_{i_0})}{p_{\mu^2}(\mathbf{y}_{i_0})} \cdot \prod_{j=1}^{i_n-1} \frac{f_s(\mathbf{y}_{i_{j-1}} \to \mathbf{y}_{i_j} \to \mathbf{y}_{i_{j+1}})}{p_{\sigma^\perp}(\mathbf{y}_{i_{j-1}} \to \mathbf{y}_{j_i} | \mathbf{y}_{i_{j-2}} \to \mathbf{y}_{i_{j-1}})} \right) \cdot$$

$$f_s(\mathbf{y}_{i_n-1} \to \mathbf{y}_{i_n} \to \mathbf{z}_1) \cdot \mathcal{G}(\mathbf{y}_{i_n} \leftrightarrow \mathbf{z}_1) \cdot f_s(\mathbf{z}_0 \to \mathbf{z}_1 \to \mathbf{y}_{i_n}) \cdot \quad (9.157)$$

$$\frac{W_e^j(\mathbf{z}_0)}{p_{\mu^2}(\mathbf{z}_0)\, p_{\sigma^\perp}(\mathbf{z}_0 \to \mathbf{z}_1 | \mathbf{z}_0)}, \quad (9.158)$$

where we assume: $p_{\sigma^\perp}(\mathbf{y}_{0_i} \to \mathbf{y}_{1_i} | \mathbf{y}_{-1_i} \to \mathbf{y}_{0_i}) = p_{\sigma^\perp}(\mathbf{y}_{0_i} \to \mathbf{y}_{1_i})$.

Obviously, the first row of $\frac{f_j(\overline{\mathbf{x}}_i)}{p(\overline{\mathbf{x}}_i)}$ describes the contributions of all subpaths of $\overline{\mathbf{y}}_i$ for shading the pixel $\square_j$. So, a good strategy is, if instant global illumination stores with the VPL, $\mathbf{y}_{i_k}$, also the contribution of the subpath $\overline{\mathbf{y}}_{i_k} = \mathbf{y}_{i_0} \ldots \mathbf{y}_{i_k}$, that is, the weight of the path implied by the emitted radiance, the product of the BSDF terms, and the sampling densities. Approximating the BSDF at the last vertex of a light path by a constant Lambertian term furthermore simplifies the formula for the contributions from a virtual point light. Now, only the BSDF at $\mathbf{z}_1$ determines whether the light path $\overline{\mathbf{y}}$ and the eye path $\overline{\mathbf{x}}$ can be connected.

Lambertian BRDF (349)

THE RENDERING STEP IN INSTANT GLOBAL ILLUMINATION. After generating the VPLs, we are now ready to begin rendering. For that, IGI computes the direct illumination

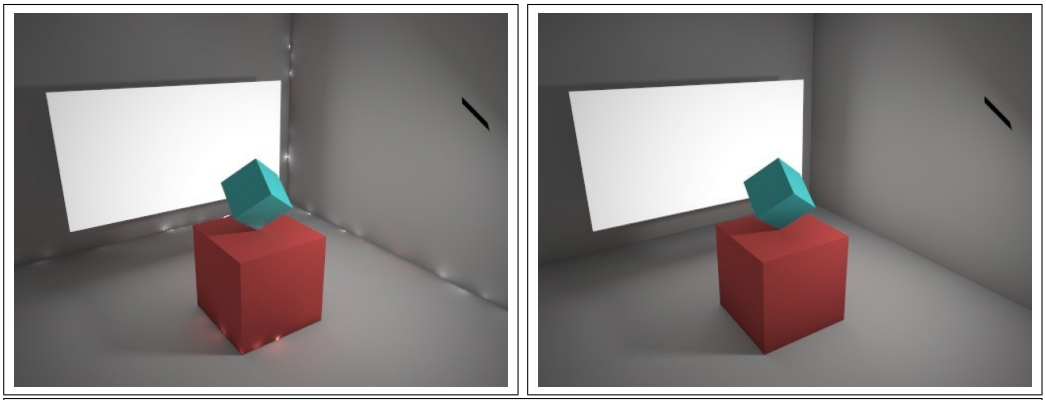


FIGURE 9.50: EFFECTS OF THE WEAK SINGULARITY IN INSTANT GLOBAL ILLUMINATION. The weak singularity linked with the geometry term—caused if a VPL and the point being shaded are close together—can lead to bright splotches in the rendered images, especially noticeable at the corners of a scene. Bounding the geometry term to be no longer larger then a fixed value or adding a constant to the denominator, eliminates the splotches but makes the algorithm biased. The image on the right is slightly darker than that on the left-hand side. Image courtesy by Simon Brown.

contribution to $\mathbf{z}_1$ as usual by recursively tracing rays through the scene to account perfect specular reflection and refraction. Afterwards, the indirect lighting component has to be computed. This is done via shadow rays in direction to the VPLs of the current involved light path $\overline{\mathbf{y}}_i$.

One problem of instant global illumination is the weak singularity linked with the geometry term if a virtual point light source, $*$, and the point being shaded are close together. Due to the denominator $\frac{1}{\|*-\mathbf{s}\|_2^2}$, the geometry term can become very large leading to bright splotches in the rendered images, especially noticeable at the corners of a scene, see Figure 9.50. Indeed this is mathematically correct, but without any practical use. To circumvent this problem a series of different possibilities are available: a very simple method is to add a small constant to the denominator, another possibility could be to clamp the geometry term so that it is no larger than a given upper limit $\overline{G}$. Indeed, this eliminates the artifacts caused, but also introduces bias, since it makes the images slightly darker than they are really.

Geometry Term (129)

REMARK 9.38 *Since the algorithm uses the same set of light paths for coloring all pixels of the image plane, images, rendered with instant global illumination look smoother and lack the typical noisy character of images produced with bidirectional path tracing or Monte Carlo path tracing. As shadow rays in IGI can be traced significantly faster than with BDPT or MCPT, high-quality results can be achieved in seconds even on single processor machines, see [225, Wald 2004].*

REMARK 9.39 *In the mean time, the group around Phillip Slussalek, has extended the method even to handle massively complex scenes with millions of polygons and thousands of light sources, [226, Wald & al. 2003] and [225, Wald & al. 2004]*

9.7 REFERENCE LITERATURE AND FURTHER READING

Monte Carlo path tracing was introduced in [98, Kajiya 1986] as a stochastic based rendering algorithm for solving the rendering equation. In his seminal paper, James Kajiya, derived the rendering equation in terms of intensity instead of radiance, see Remark 4.47.

Commonly, Monte Carlo path tracing is presented in connection with the recursive ray tracing algorithm, namely, as a stochastic variant of ray tracing where at each hit point of a ray with an object only a single reflected or refracted ray is randomly generated. The most textbooks about global illumination such as, [95, Jensen 2001], [50, Dutré & al. 2003], [187, Shirley & Morley 2003] and [51, Dutré & al. 2006] as well as [205, Suffern 2007] follow also this approach when presenting Monte Carlo path tracing. In [158, Pharr & Humphreys 2004] and [159, Pharr & Humphreys 2010], Monte Carlo path tracing is discussed based on the path integral formulation of the stationary light transport.

But we have decided to go another way. Our approach for describing Monte Carlo path tracing is based on the method of successive integral substitution for solving Fredholm integral equations of the 2^{nd} kind, introduced in Section 6.7. Applied to the light transport equation within a vacuum, this method led us in Section 8.4 to the mathematical basis of the distribution ray tracing algorithm, firstly introduced in [40, Cook 1984]. Instead to evaluate the kernel—that is, the BRDF, BTDF, or the BSDF—of the light transport equation via a large number of rays at the hit point of a ray with any surface object within a scene, we decided to generate only a single ray depending on the properties of the surface that has been hit. The resulting algorithm corresponds to a discrete-time Markov process for solving the light transport equation in free space, also called pure-Monte Carlo path tracing,

A similar idea to our approach for introducing path tracing is used in [68, Glassner 1995]. Glassner generates a random walk for solving the light transport equation within a vacuum. In [220, Veach 1997], path tracing is then addressed as a special variant of the bidirectional path tracing algorithm, where transport paths only start at the eye of the observer. Other good, brief, and easily understandable references for path tracing are the master thesis [90, Hutchinson 1993] and the PhD thesises [183, Shirley 1991], [191, Slusallek 1995] and [116, Lafortune 1996]. In particular Lafortune and [50, Dutré & al. 2003], [51, Dutré & al. 2006] discusses a series of variance reduction techniques with respect to their applicability to path tracing for solving the light transport equation, resulting in various optimizations of the original algorithm.

Monte Carlo light tracing is discussed in more detail in [52, Dutré] and [49, Dutré 1994]. The algorithm is based on the potential transport equation, from [150, Pattanaik & Mudur 1993], [151, Pattanaik & Mudur 1995], as the dual algorithm of Monte Carlo path tracing. For obtaining a solution to the adjoint equation, they proposed a Monte Carlo quadrature and random-walk techniques that simulates light propagation starting from the light sources. An extensive comparison of Monte Carlo path tracing and Monte Carlo light tracing is done in [116, Lafortune 1996], where the most interesting variance reduction techniques are applied to both algorithms resulting in a a series of Monte Carlo estimators for both approaches. We present Monte Carlo light tracing as the dual algorithm to path tracing based on the method of successive integral substitution for solving Fredholm integral equations of the 2^{nd} kind, introduced in Section 6.7.

Bidirectional path tracing was at first introduced in [116, Lafortune 1996] and a year later independently developed in [221, Veach 1998]. Although both variants of bidirectional path tracing lead to similar results, they are based on two different mathematical frameworks. While Eric Lafortune's variant of bidirectional path tracing starts from the formulation of the global reflectance distribution function, Eric Veach's variant of bidirectional path tracing is based on the path integral formulation of the light transport problem. Both algorithms have its origin in [8, Arvo 1993], where the rendering of caustics was described by means of a ray tracing algorithm, that takes its starting point in one of the light sources. We have decided to present bidirectional path tracing as a Markov process based rendering method applied to the path integral formulation, that is, Section 9.3 is exclusively built on [221, Veach 1998]. A very short overview of Veach's BDPT algorithms can also be read in [95, Jensen 2001], [158, Pharr & Humphreys 2004] and [159, Pharr & Humphreys 2010], the version of Eric Lafortune is shortly discussed in [50, Dutré & al. 2003], [51, Dutré & al. 2006].

The photon-mapping concept was introduced in [96, Jensen & Christensen 1995], which, together with [95, Jensen 2001] can be considered as our main resources for our presentation of photon mapping in Section 9.5. In a series of papers, Jensen discussed the applicability of photon mapping in many fields of computer graphics. For an overview of the photon mapping algorithm see also [50, Dutré & al. 2003], [51, Dutré & al. 2006], and [158, Pharr & Humphreys 2004], and [159, Pharr & Humphreys 2010].

Instant global illumination, [226, Wald & al. 2003] and [225, Wald 2004] was developed by the graphics group around Philipp Slusallek at Saarland University in 2002 for the availability of realtime ray tracing for achieving interactive global illumination. The algorithm is based on the idea of the virtual point light from instant radiosity [103, Keller 1997]. In the mean time, the group around Phillip Slussalek, has extended the method even to handle massively complex scenes with millions of polygons and thousands of light sources, [226, Wald & al. 2003] and [227, Wald 2004]. Instant global illumination is also discussed in [159, Pharr & Humphreys 2010].

CHAPTER TEN

FINITE ELEMENT METHODS BASED RENDERING ALGORITHMS

Apart from Monte Carlo rendering algorithms, discussed in detail in the last two chapters, there exists another approach for solving the global illumination problem: the *radiosity method*. While Monte Carlo algorithms are based on stochastic principles from probability theory, radiosity methods are based on a *finite element approach*. The idea behind finite element methods is to approximate a complex, infinite-dimensional problem by a simpler, finite dimensional problem, for which a solution can easily be found. With respect to the stationary light transport equations, which are all infinite-dimensional integral equations, this means, that we transform an integral equation into a system of linear equations over a finite dimensional function space, such as the linear normed space $(\mathbb{R}^n, \|\cdot\|)$. Then, fast iterative solvers can be used to solve the corresponding linear system. Since the solution to a linear system of equations only exists in a finite-dimensional function space, such a solution also represents only an approximate to the real solution of the underlying integral equation.

Section 2.3.3.2.3
Linear Integral Equation (127)
$(\mathbb{R}^n, \|\cdot\|)$ (861)
Function Space (28)

Interpreted as 3D rendering algorithms, radiosity methods are global illumination algorithms based on the principle of energy conservation. In its classical variants they work only on purely diffuse surfaces, see Figure 10.1. Unlike Monte Carlo rendering algorithms, which trace an image pixel by pixel, the radiosity approach is view independent. Instead of evaluating the SLTEV for directions and locations determined by the position of the camera and the pixels of the image plane, a radiosity algorithm solves the SLTEV at single locations distributed over the whole surfaces of the environment. As all surfaces in a scene are assumed to be diffuse, the only information we need to create an image is, how much light is being reradiated by each surface, and which objects are visible through a pixel of the image plane. This then allows to efficiently perform interactive walkthroughs of simulated environments since the distribution of light on each surface is unaffected by the movement of the camera, where we only have to recompute which objects are visible through the pixel under consideration. Compared to ray tracing based methods, this then takes less work when generating new views from a scene model.

SLTEV (398)

FIGURE 10.1: THE INTERIOR OF LE CORBUSIER'S CHAPEL AT RONCHAMP. The illumination was computed using radiosity, with the sunbeams added by stochastic ray tracing during rendering. The model was created by Paul Boudreau, Keith Howie, and Eric Haines at 3D/EYE, Inc. with Hewlett-Packards ARTCore Radiosity and Ray Tracing library. Image courtesy of Eric Haines.

OVERVIEW OF THIS CHAPTER. The present chapter is only a high-level introduction to finite element methods based rendering algorithms, that is, we mainly discuss the classical radiosity method and its finite element approach in more detail. Useful variations of the basic algorithm for reducing the size and complexity of the problem, such as the *progressive refinement* technique or *hierarchical* methods, will not be discussed. We also will not talk about meshing strategies. For all these topics we refer the reader to the books by [36, Cohen & Wallace 1993], [190, Sillion & Puech 1994], and [68, Glassner 1995].

Section 10.1 The chapter is structured as follows: First, we present the classical radiosity formulation based on the assumption that light transport is considered under vacuum conditions in a scene consisting of purely diffuse, opaque surfaces and show how the classical discrete Section 10.2 radiosity equation can be solved via methods from numerical analysis. Afterwards, we discuss in more detail finite element approaches—such as the *collocation* and the *Galerkin method*—to solve the global illumination problem under more weaker restrictions as those Section 10.3 given above. We also present the structure of a typical radiosity algorithm for image synthesis, and finally, we will shortly talk about the advantages and disadvantages of ray Section 10.4 tracing and finite element based algorithms.

10.1 THE CLASSICAL RADIOSITY FORMULATION

The radiosity method has its origin in the 1950s as a method for computing radiant heat exchange between surfaces [189, Siegel & Howell 1992]. In 1984, then it was specifically

adapted to solve the global illumination problem by researchers at Fukuyama and Hiroshima Universities [139, Nishita & Nakamae 1985] and at the Program of Computer Graphics at Cornell University [70, Goral & al. 1984]. To simplify the method, it was assumed that all scattering within a scene is perfectly diffuse. In contrast to Monte Carlo rendering algorithms, this assumption then implies, that radiosity algorithms only account for light transport paths of characteristic $\overleftrightarrow{\mathrm{ED}^*\mathrm{L}}$.

The classical radiosity formulation can be considered as a four step procedure: partitioning the scene, discretizing the continuous SLTEV, computing the form factor matrix, and solving the resulting linear system. In the first step of the procedure, the scene to be rendered has to be partioned into a mesh of disjoint surface patches, typically quadrilateral or triangular elements that all have constant radiosity. In the next step, the continuous stationary light transport equation is then transformed into a discrete radiosity equation, that is, a system of linear equations, where the coefficients of the associated matrix, the so-called *form factors*, represent the energy transfer between the patches. They can be computed via the reflectivity of the patches and the scene geometry. As, the heart of a radiosity algorithm is the computation of the form factor matrix we discuss the fundamental concept of the *form factor* in more detail, derive properties of form factors, and show how the computation of the form factors can be done more efficiently. In the last step of the classical radiosity formulation, then the discrete linear system is solved by an iterative solver resulting in a finite dimensional approximate to real solution of the continuous SLTEV.

Section 10.1.1
Section 10.1.2
Section 10.1.3
Section 10.1.4

10.1.1 FROM THE SLTEV TO THE CLASSICAL RADIOSITY INTEGRAL EQUATION

Recall, the hemispherical form of the stationary light transport equation in vacuum, expressed in terms of incident and exitant radiance, has the form

SLTEV (399)

$$\mathrm{L_o}(\mathbf{s}_i, \omega_o^i) = \mathrm{L_e}(\mathbf{s}_i, \omega_o^i) + \int_{\mathcal{H}^2_+(\mathbf{s}_i)} \mathrm{f_r}(\mathbf{s}_i, \omega_i^i \to \omega_o^i)\, \mathrm{L_i}(\mathbf{s}_i, \omega_i^i)\, \mathrm{d}\sigma^\perp_{\mathbf{s}_i}(\omega_i^i), \qquad (10.1)$$

where $\mathbf{s}_i$ is a point at an object surface, and ω_i^i and ω_o^i correspond to incident and exitant directions over the upper hemisphere about point $\mathbf{s}_i$, see Figure 10.2.

Based on the *radiosity assumption* that all surfaces in the scene are Lambertian diffuse reflectors, the exitant radiance at surface point $\mathbf{s}_i$ does not depend on the outgoing direction ω_o^i, that is, the exitant radiance $\mathrm{L_o}$ is a function of position $\mathbf{s}_i$ only. So, we can write:

Lambertian Reflector (349)

$$\mathrm{L}(\mathbf{s}_i \to \mathbf{s}_{i-1}) = \mathrm{L_o}(\mathbf{s}_i, \omega_o^i), \qquad (10.2)$$

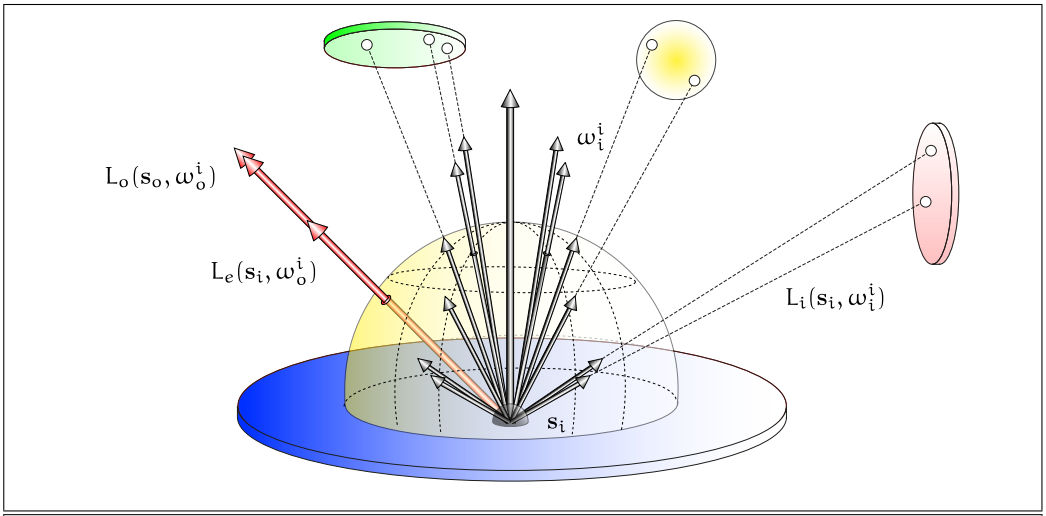


FIGURE 10.2: THE HEMISPHERICAL FORM OF THE STATIONARY LIGHT TRANSPORT EQUATION IN A VACCUM. The radiance exitant at point s_i in direction ω_o^i is composed of the emitted radiance at s_i in direction ω_o^i and the incident radiance at point s_i from direction ω_i^i integrated over the hemisphere about s_i.

where $s_{i-1} = \gamma(s_i, \omega_o^i)$ is a point on any surface within the scene visible from s_i. The same argument obviously holds for the emitted radiance L_e, that is, we get:

$$L_e(s_i \to s_{i-1}) = L_e(s_i, \omega_o^i) \tag{10.3}$$

for a surface point s_i.

Now, apart from the outgoing radiance, also the BRDF is independent of directions in the case of ideal diffuse reflectors. Due to Lemma 4.4 and Remark 4.18 the BRDF is
ρ_{dh} (338) then coupeled to the diffuse reflectance, that is, the directional-hemispherical reflectance, ρ_{dh}, via:

$$f_r \equiv f_r^o = \frac{\rho_{dh}(s_i)}{\pi}. \tag{10.4}$$

Using these results, then the SLTEV can be reformulated in the following mixed hemispherical-3-point form:

$$L(s_i \to s_{i-1}) = L_e(s_i \to s_{i-1}) + \frac{\rho_{dh}(s_i)}{\pi} \int_{\mathcal{H}^2_+(s_i)} L_i(s_i, \omega_i^i)\, d\sigma^\perp_{s_i}(\omega_i^i), \tag{10.5}$$

where the incident radiance L_i still depends on the incident direction ω_i^i. Obviously, the integral in Equation (10.5) corresponds to the incident flux density, which depends on the

exitances of all other surfaces. To compute the exitance leaving a surface, it is just the incident flux density that has to be evaluated.

Now, the principle of radiance invariance in a vacuum says that radiance incident at point $\mathbf{s}_i$ from direction ω_i^i can be expressed in terms of radiance exitant from point $\mathbf{s}_j$ in direction $\omega_o^j = -\omega_i^i$, that is, Radiance Invariance (253)

$$\begin{aligned} L_i(\mathbf{s}_i, \omega_i^i) &= L_o(\mathbf{s}_j, \omega_o^j) && (10.6)\\ &= L(\mathbf{s}_j \to \mathbf{s}_i), && (10.7)\end{aligned}$$

where it holds: $\mathbf{s}_j = \gamma(\mathbf{s}_i, \omega_i^i)$.

Integrating Equation (10.5) over all surfaces patches $\partial\mathcal{V}$ of the scene instead over the hemisphere about point $\mathbf{s}_i$ then requires to transform the integration measure $\sigma^\perp$ into the 2-dimensional Lebesgue area measure, μ^2. As shown in Equation (2.199), this measure transform leads to an integral equation where no directional variable appear anymore, namely, $\sigma^\perp$ (88)

$$L(\mathbf{s}_i \to \mathbf{s}_{i-1}) = L_e(\mathbf{s}_i \to \mathbf{s}_{i-1}) + \frac{\rho_{dh}(\mathbf{s}_i)}{\pi} \int_{\partial\mathcal{V}} L(\mathbf{s}_j \to \mathbf{s}_i)\, \mathcal{G}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i) d\mu^2(\mathbf{s}_j). \qquad (10.8)$$

In this equation, the term $\mathcal{G}$ corresponds to the geometry term from Equation (2.353), thus:

$$\mathcal{G}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i) \overset{def}{=} \frac{\left|\cos\theta_o^j \cos\theta_i^i\right|}{\|\mathbf{s}_j - \mathbf{s}_i\|_2^2} \mathcal{V}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i), \qquad (10.9)$$

with the visibility function $\mathcal{V}$ introduced in Box 2.1.

As radiance on purely diffuse surfaces does not depend on the outgoing direction, see Example 3.4, radiosity and radiance can be used interchangeably to characterize the light leaving such surfaces. So, we can express the outgoing as well as the emitted radiances in the SLTEV by the identities,

$$L(\mathbf{s}_i \to \mathbf{s}_{i-1}) \overset{(3.85)}{=} \frac{B(\mathbf{s}_i)}{\pi} \quad \text{and} \quad L(\mathbf{s}_j \to \mathbf{s}_i) \overset{(3.85)}{=} \frac{B(\mathbf{s}_j)}{\pi} \qquad (10.10)$$

as well as

$$L_e(\mathbf{s}_i \to \mathbf{s}_{i-1}) \overset{(3.85)}{=} \frac{B_e(\mathbf{s}_i)}{\pi}. \qquad (10.11)$$

This then leads to the following form of the SLTEV, expressed in terms of radiosities, namely,

$$\frac{B(\mathbf{s}_i)}{\pi} = \frac{B_e(\mathbf{s}_i)}{\pi} + \frac{\rho_{dh}(\mathbf{s}_i)}{\pi} \int_{\partial\mathcal{V}} \frac{B(\mathbf{s}_j)}{\pi}\, \mathcal{G}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\mu^2(\mathbf{s}_j). \qquad (10.12)$$

Multiplying both sides of this equation by π results in the so-called *radiosity integral equation*:

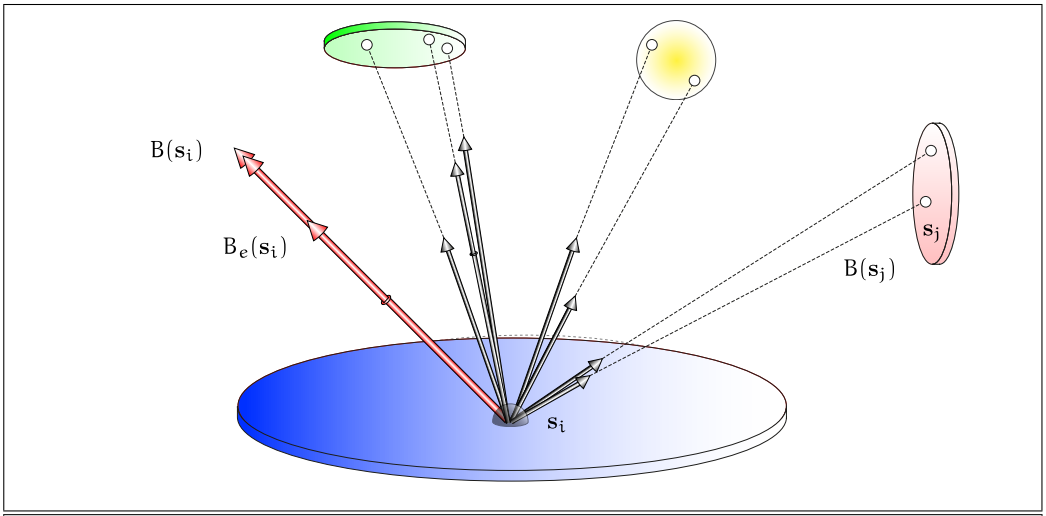


FIGURE 10.3: THE CLASSICAL RADIOSITY INTEGRAL EQUATION. The radiosity at point s_i is composed of the emittance at point s_i and the radiosity incident at point s_i that comes from all points s_j at surfaces visible from point s_i.

DEFINITION 10.1 (The Classical Radiosity Integral Equation) *Let* $B(s_i)$ *be the radiosity leaving surface point* s_i, $B_e(s_i)$ *the corresponding radiosity emitted from light sources, and* $B(s_j)$ *the radiosity at point* s_j. *Then, the equation*

$$B(s_i) = B_e(s_i) + \rho_{dh}(s_i) \int_{\partial \mathcal{V}} B(s_j)\, \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j), \tag{10.13}$$

which describes the scattering behavior of light at diffuse object surfaces in a vacuum, is called the classical radiosity integral equation, *see Figure 10.3.*

$\mathcal{G}$ (129) *The term* $\mathcal{G}'(s_j \leftrightarrow s_i)$ *is called the* radiosity geometry term, *defined by:*

$$\mathcal{G}'(s_j \leftrightarrow s_i) \stackrel{def}{=} \frac{\mathcal{G}(s_j \leftrightarrow s_i)}{\pi}. \tag{10.14}$$

Obviously, the radiosity integral equation describes an arbitrary scalar function across the surfaces. It is a composition of the exitance B_e that describes the emission of light sources and the directional-hemispherical reflectance multiplied by the amount of power received from the environment.

Exitance (264)

As in the case of the SLTEV, also the radiosity integral equation is an integral equation of type which very rarely has a closed-form analytic solution, that is, to find a solution of the radiosity integral equation, we have to use methods from numerical mathematics.

Integral Equation (127)

10.1.2 DISCRETIZING THE CLASSICAL RADIOSITY INTEGRAL EQUATION

As we will see in Section 10.2, the radiosity method is based on a finite element approach. Now, in Section 2.3.3.2.3 we discussed finite element strategies for solving linear integral operator equations. The idea underlying these methods was to approximate an infinite dimensional function space—which contains the solution of a Fredholm integral equation of the 2^{nd} kind—by a finite-dimensional subspace. In this subspace, we then have to find a function that is in some sense a *good* approximation to the true solution. From the multitude of finite element strategies for solving Fredholm type integral equation, we presented two different methods:

Integral Operator Equation (131)

Section 2.3.3.2.2

i) the collocation method, as a finite basis approach, and

ii) the Galerkin method, a so-called projection method.

In Section 10.2 then we will show, that the classical radiosity equation, which we derive in this section, can be considered as a simplification of the more general Galerkin formulation to the SLTEV.

Considered as a finite element technique, the idea behind the classical radiosity method is to generate in a first step a partition of the surfaces of the scene $\partial\mathcal{V}$ into a collection of n disjoint, so-called *surface patches* P_j, thus,

$$\partial\mathcal{V} = \bigcup_{j=1}^{n} P_j, \tag{10.15}$$

where each patch P_j has the Lebesgue measure $\mu^2(P_j) = A_j$.

Based on this partition, the classical radiosity integral equation can then be written as the sum of n integrals over the patches P_j, namely,

$$\begin{aligned} B(\mathbf{s}_i) &= B_e(\mathbf{s}_i) + \rho_{dh}(\mathbf{s}_i) \int_{\bigcup_{j=1}^{n} P_j} B(\mathbf{s}_j)\, \mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\mu^2(\mathbf{s}_j) && (10.16)\\ &= B_e(\mathbf{s}_i) + \rho_{dh}(\mathbf{s}_i) \sum_{j=1}^{n} \int_{P_j} B(\mathbf{s}_j)\, \mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\mu^2(\mathbf{s}_j), && (10.17) \end{aligned}$$

where $\mathbf{s}_i$ is a point on a fix patch P_i and $\mathbf{s}_j$ are different points on all other patches $P_j \neq P_i$.

With the additional assumption that the radiosity is constant over each patch P_j, that is, it holds $B(\mathbf{s}_j) = B_j$, the classical radiosity integral equation can further be simplified by moving the radiosity outside the integral, so, we can write:

$$B(\mathbf{s}_i) \overset{B(\mathbf{s}_j)=B_j}{=} B_e(\mathbf{s}_i) + \rho_{dh}(\mathbf{s}_i) \sum_{j=1}^{n} B_j \int_{P_j} \mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\mu^2(\mathbf{s}_j). \tag{10.18}$$

With $s_i \in P_i$, the constant radiosity value B_i can then be computed via an area-weighted average of the point-radiosities $B(s_i)$, namely,

$$
\begin{aligned}
B_i &\overset{def}{=} \frac{1}{A_i}\int_{P_i} B(s_i)\, d\mu^2(s_i) && (10.19)\\
&\overset{(10.18)}{=} \frac{1}{A_i}\int_{P_i} B_e(s_i)\, d\mu^2(s_i) + && (10.20)\\
&\quad \frac{1}{A_i}\int_{P_i} \rho_{dh}(s_i)\left(\sum_{j=1}^{n} B_j \int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j)\right) d\mu^2(s_i)\\
&= \frac{1}{A_i}\int_{P_i} B_e(s_i)\, d\mu^2(s_i) + && (10.21)\\
&\quad \sum_{j=1}^{n} \frac{1}{A_i} B_j \int_{P_i} \rho_{dh}(s_i) \int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) d\mu^2(s_i).
\end{aligned}
$$

Even this equation can furthermore simplified by assuming that the reflectance is also constant across each patch P_i. Setting $\rho_{dh}(s_i) = \rho_i$ for each $s_i \in P_i$ leads to:

$$
\begin{aligned}
\frac{1}{A_i}\int_{P_i} B(s_i)\, d\mu^2(s_i) &= \frac{1}{A_i}\int_{P_i} B_e(s_i)\, d\mu^2(s_i) + && (10.22)\\
&\quad \rho_i \sum_{j=1}^{n} \frac{1}{A_i} B_j \int_{P_i}\int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) d\mu^2(s_i).
\end{aligned}
$$

Using the indentities

$$B_i = \frac{1}{A_i}\int_{P_i} B(s_i)\, d\mu^2(s_i) \qquad (10.23)$$

as well as

$$B_{ei} = \frac{1}{A_i}\int_{P_i} B_e(s_i)\, d\mu^2(s_i), \qquad (10.24)$$

then the above equation leads to the so-called *classical radiosity equation*. It is defined as follows:

DEFINITION 10.2 (The Classical Discrete Radiosity Equation) *Let B_i be the radiosity leaving surface patch P_i, B_j the corresponding radiosities leaving surface patches P_j, furthermore, let B_{e_i} the exitance from patch P_i into the scene. Then, the equation*

$$B_i = B_{ei} + \rho_i \sum_{j=1}^{n} F_{ij} B_j, \qquad (10.25)$$

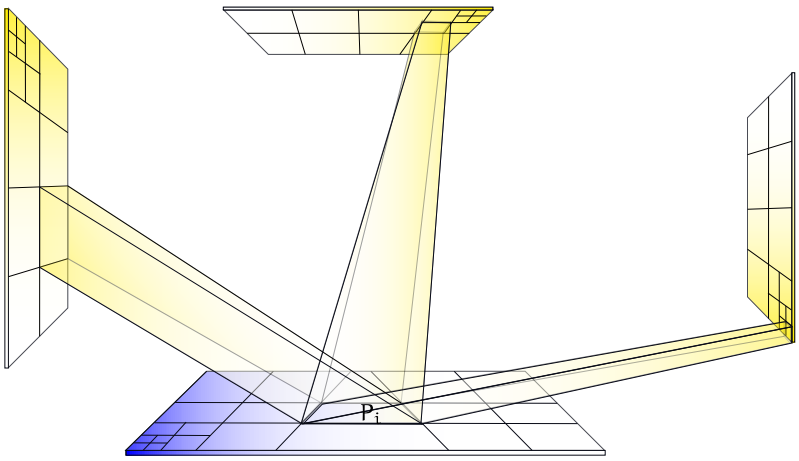


FIGURE 10.4: THE CLASSICAL RADIOSITY EQUATION. The radiosity of patch P_i corresponds to the emittance of the patch and the proportion of power leaving patch P_i that is received by all other patches visible from P_i.

with

$$F_{ij} \stackrel{def}{=} \frac{1}{A_i} \int_{P_i} \int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) d\mu^2(s_i) \tag{10.26}$$

is called the classical discrete radiosity equation *with the* classical form factors F_{ij} *between the patches* P_i *and* P_j*, see Figure 10.4., where the form factor* F_{ij} *describes the fraction of energy leaving patch* P_i *that reaches patch* P_j*.*

REMARK 10.1 *If we try to interprete the classical discrete radiosity equation in a physical way at this time, we have a problem. Obviously, the first term on the right-hand side of (10.25) is the emitted radiosity of patch* P_i*. But the second term makes us problems if we want to give them a physically meaning. Here we have to combine the reflectance of patch* P_i *with the radiosity from patch patch* P_j *and the fraction of light which leaves patch* P_i *and arrives at patch* P_j*. Obviously, the classical discrete radiosity equation given in the above form can physically not really be interpreted.*

REMARK 10.2 *As they are defined as the value of a double integral over the radiosity geometry term* $\mathcal{G}'$ *the form factors* F_{ij} *are only depending on the geometry between the patches* P_i *and* P_j*. That is, only the shape, distance, and orientation as well as the visibility of the involved surface patches are relevant for computing the form factors, the energy flowing between the patches does not play any role.* $\mathcal{G}'$ (782)

Now, Equation (10.25) is only valid for patch P_i. To find an approximate solution to the stationary light transport problem under the assumption made in the above discussion, we also have to compute the radiosities of the other $n-1$ patches. Thus, the stationary light transport problem, expressed in terms of radiosities, corresponds to a linear system of n radiosity equations that couples B_i to B_j, one for each patch $P_i, 1 \leq i \leq n$. The unknowns of this system of equations, which is given by:

$$B_i = B_{e_i} + \rho_i \sum_{j=1}^{n} F_{ij} B_j, \qquad 1 \leq i \leq n, \tag{10.27}$$

are the n patch-radiosities $B_1, \ldots, B_n$ of the partition $\bigcup_{j=1}^{n} P_j$.

Linear Operator Equation (61)

The above linear system of equations can also be expressed in form of a linear operator equation, namely as:

$$\mathbf{B} = \mathbf{B}_e + (\boldsymbol{\rho}\mathbf{F})\mathbf{B}, \tag{10.28}$$

where $\mathbf{B} = (B_1, \ldots, B_n)^T$ is an unknown n-dimensional radiosity vector, $\mathbf{B}_e = (B_{e_1}, \ldots, B_{e_n})^T$ are the given exitances, $\boldsymbol{\rho}$ is the $n \times n$ diagonal matrix of reflectances:

$$\boldsymbol{\rho} \stackrel{def}{=} \begin{pmatrix} \rho_1 & 0 & \ldots & 0 \\ 0 & \rho_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \rho_n \end{pmatrix} \tag{10.29}$$

and $\mathbf{F}$ is the quadratic $n \times n$ matrix of the classical form factors, thus,

$$\mathbf{F} \stackrel{def}{=} \begin{pmatrix} F_{11} & F_{12} & \ldots & F_{1n} \\ F_{21} & F_{22} & \ldots & F_{2n} \\ \vdots & & \ddots & \vdots \\ F_{n1} & F_{n2} & \ldots & F_{nn} \end{pmatrix}, \tag{10.30}$$

which plays the role of a transport operator.

In matrix-vector notation, the classical radiosity system can then be written as:

$$\begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_n \end{pmatrix} = \begin{pmatrix} B_{e_1} \\ B_{e_2} \\ \vdots \\ B_{e_n} \end{pmatrix} + \begin{pmatrix} \rho_1 F_{11} & \rho_1 F_{12} & \ldots & \rho_1 F_{1n} \\ \rho_2 F_{21} & \rho_2 F_{22} & \ldots & \rho_2 F_{2n} \\ \vdots & & \ddots & \vdots \\ \rho_n F_{n1} & \rho_n F_{n2} & \ldots & \rho_n F_{nn} \end{pmatrix} \cdot \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_n \end{pmatrix}. \tag{10.31}$$

As the operator equation from (10.28) shows, we can reformulate this system of equations as:

$$\begin{aligned} \mathbf{B}_e &= \mathbf{B} - (\boldsymbol{\rho}\mathbf{F})\mathbf{B} && (10.32) \\ &= \underbrace{(\mathbf{I} - \boldsymbol{\rho}\mathbf{F})}_{\mathbf{M}}\mathbf{B} && (10.33) \\ &= \mathbf{M}\mathbf{B}, && (10.34) \end{aligned}$$

where $\mathbf{M}$ is denoted as the *classical radiosity matrix* given by

$$\mathbf{M} \stackrel{\text{def}}{=} \begin{pmatrix} 1-\rho_1 F_{11} & -\rho_1 F_{12} & \dots & -\rho_1 F_{1n} \\ -\rho_2 F_{21} & 1-\rho_2 F_{22} & \dots & -\rho_2 F_{2n} \\ \vdots & & \ddots & \vdots \\ -\rho_n F_{n1} & -\rho_n F_{n2} & \dots & 1-\rho_n F_{nn} \end{pmatrix}. \tag{10.35}$$

In matrix-vector notation, the classical radiosity equation can then be written in the form:

$$\begin{pmatrix} 1-\rho_1 F_{11} & -\rho_1 F_{12} & \dots & -\rho_1 F_{1n} \\ -\rho_2 F_{21} & 1-\rho_2 F_{22} & \dots & -\rho_2 F_{2n} \\ \vdots & & \ddots & \vdots \\ -\rho_n F_{n1} & -\rho_n F_{n2} & \dots & 1-\rho_n F_{nn} \end{pmatrix} \cdot \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_n \end{pmatrix} = \begin{pmatrix} B_{e1} \\ B_{e2} \\ \vdots \\ B_{en} \end{pmatrix}. \tag{10.36}$$

EXAMPLE 10.1 (Space Considerations for the Classical Radiosity Method) *From computer science it is known that the efficiency of an algorithm with respect to its run time and space behavior is measured as a function of the size of the input to the problem. At this location, let us shortly talk about the space complexity of a radiosity algorithm. The run time behavior of a radiosity algorithm will then be discussed later, when we present techniques for solving the radiosity system of equations.*

Let us consider a reasonable complex scene consisting of 10^5 *surface patches. The construction of the radiosity matrix* $\mathbf{M}$ *via a classic radiosity algorithm then requires the computation of* 10^{10} *form factors, namely, the coefficients of* $\mathbf{M}$*. Allowing* 4 *bytes per floating-point number for each form factor means that we need* 40 *GB of RAM. Obviously, this is not well scalable, but as we will see, there are methods that work without computing all these form factors.*

10.1.3 THE CLASSICAL FORM FACTORS

Let us now discuss the concept of the classical form factor a little bit more closely. Since it is the most time-consuming part of any radiosity algorithm, we are interested in techniques that makes it possible to compute the form factors in a simple and efficient way. For that purpose, let us first derive some properties of the form factors that are useful for constructing the classical radiosity matrix $\mathbf{M}$. M (787)

The form factor matrix $\mathbf{F}$ represents the most important component of the radiosity matrix $\mathbf{M}$. Their coefficients, thus the entries F_{ij}, are central to the radiosity method and to understanding the propagation of light within a scene. Defined via the four-dimensional F (786) M (787)

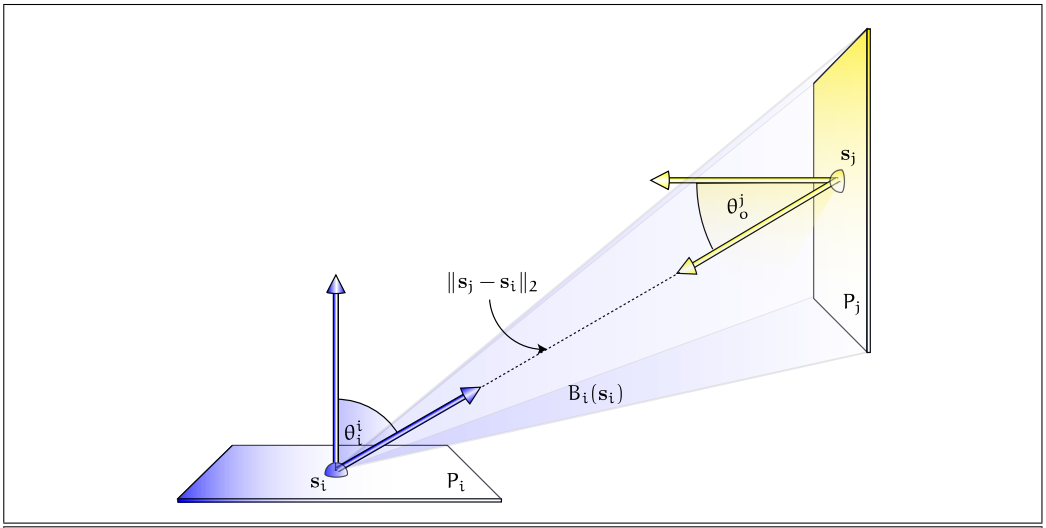


FIGURE 10.5: THE GEOMETRY FOR DEFINING THE CONCEPT OF THE FORM FACTOR. The form factor describes the fraction of energy which leaves element P_i and arrives at element P_j. It is a dimensionless quantity, that depends only on the scene geometry.

integral

$$F_{ij} \stackrel{def}{=} \frac{1}{A_i} \int_{P_i} \left(\int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i) \, d\mu^2(s_j) \right) d\mu^2(s_i) \tag{10.37}$$

$$= \frac{1}{A_i} \int_{P_i} \left(\int_{P_j} \frac{\left|\cos\theta_o^j \cos\theta_i^i\right|}{\pi \|s_j - s_i\|_2^2} \mathcal{V}(s_j \leftrightarrow s_i) \, d\mu^2(s_j) \right) d\mu^2(s_i), \tag{10.38}$$

see Figure 10.5, where s_i and s_j are points on the surface patches P_i respectively P_j, form factors are dimensionless constants, which, due to its definition, have a very simple physical interpretation: F_{ij} *represents the fraction of energy leaving surface patch* P_i *that arrives directly at patch* P_j. Here, we have assumed that the set of surfaces $\partial\mathcal{V}$ in the environment was covered with a collection of patches $\bigcup_{i=1}^{n} P_i$.

EXAMPLE 10.2 (Three Different Types of Form Factors) *In Section 10.1.3.2 we will show, that the form factor from Equation (10.38) can only be solved analytically for the simplest geometries, such as opposing and perpendicular rectangles, circles or polygons. A significant simplifications for calculating* F_{ij} *results from the fact if one of the patches* P_i *or* P_j *is small compared with the distance between* P_i *and* P_j. *Under these conditions, the cosines occurring in the form factor integral as well as the distance* $\|s_j - s_i\|_2^2$ *are nearly constant. As the integration over the differential patch* dP_i *reduces to a multiplication with the Lebesgue area measure of* dP_i *then the*

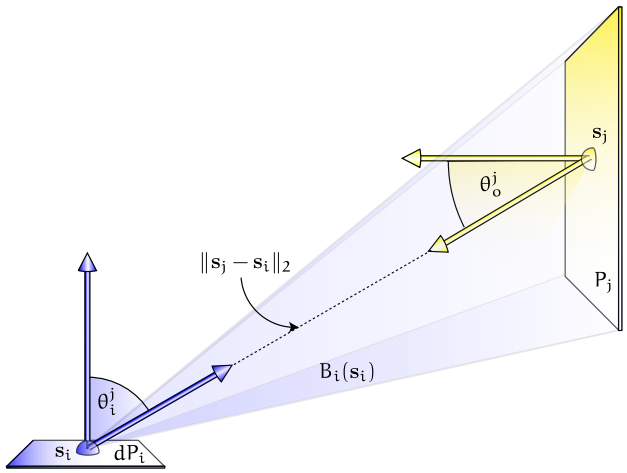


FIGURE 10.6: THE GEOMETRY FOR DEFINING THE DIFFERENTIAL-TO-FINITE ELEMENT FORM FACTOR.

differential-to-finite element form factor $F_{dP_i P_j}$ *can be written as:*

$$F_{dP_i P_j} \overset{(2.207)}{=} \int_{P_j} \frac{\left|\cos\theta_o^j \cos\theta_i^j\right|}{\pi \|s_j - s_i\|_2^2} \mathcal{V}(s_j \leftrightarrow s_i)\, d\mu^2(s_j), \tag{10.39}$$

$$= \int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) \tag{10.40}$$

see Figure 10.6.

Similar to this derivation, we can also compute the form factor between two differential patches dP_i *and* dP_j. *For the* differential-to-differential area form factor $F_{dP_i dP_j}$ *it must hold:*

$$F_{dP_i dP_j} \overset{(2.200)}{=} \mathcal{G}'(s_j \leftrightarrow s_i) \tag{10.41}$$

$$= \frac{\left|\cos\theta_o^j \cos\theta_i^j\right|}{\pi \|s_j - s_i\|_2^2} \mathcal{V}(s_j \leftrightarrow s_i)\, d\mu^2(s_j). \tag{10.42}$$

REMARK 10.3 *Usually, the form factor between two surface patches* P_i *and* P_j *should be denoted by* $F_{P_i P_j}$. *We use this notation only in the case, where differential patches are involved, thus* $F_{dP_i P_j}, F_{P_i dP_j}$ *or* $F_{dP_i dP_j}$, *sometimes we also write* $F_{s_i P_j}, F_{P_i s_j}$ *or* $F_{s_i s_j}$ *for* $s_i \in dP_i$ *and* $s_j \in dP_j$. *In all other cases, we use the simplified notation,* F_{ij}, *to represent the portion of total power leaving patch* P_i *that is received by patch* P_j.

$\mathcal{G}'$ (782) As already mentioned above, due to the definition of the radiosity geometry term $\mathcal{G}'$, a form factor is solely a function of geometry of the scene to be rendered. That is, it does not depend on the reflective or emissive properties of the surfaces, but solely on the shape, distance, and orientation as well as the visibility of the involved surface patches. This fact then allows, that surface properties can be changed without repeating the expensive form factor computation.

REMARK 10.4 *In the case where a surface patch is a light source, the form factor itself represents the direct illumination of the other patch per unit area of emissive power from the source.*

10.1.3.1 PROPERTIES OF THE CLASSICAL FORM FACTORS

As we have seen in Example 10.1, not seldom a complex scene consist of more than 10^5 surface patches, resulting in a radiosity matrix with several billions of coefficients. Now, for each of these coefficients we have to compute the four-dimensional integral from Relation (10.26). Integrals of this type can only be solved in a closed form for special geometric arrangements of the elements, that is, we must often use numerical methods for solving the form factor integral. So, the construction of the radiosity matrix $\mathbf{M}$ can be seen as the most-time consuming part of any radiosity algorithms. If we can improve this process, mainly with respect to the number of form factors, then we are well on the way to derive acceptable algorithms for solving the radiosity equation. For that purpose, we will now work out a series of useful properties of form factors—mainly based on the fact that a form factor is defined as a Lebesgue integral—that can be used to eliminate unnecessary work when computing the coefficients of the radiosity matrix $\mathbf{M}$.

NON-NEGATIVITY OF THE CLASSICAL FORM FACTORS. Due to its definition, the integrand $\mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)$, defined via,

$$\mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i) \overset{(10.26)}{=} \frac{\left|\cos\theta_o^j \cos\theta_i^i\right|}{\pi\|\mathbf{s}_j - \mathbf{s}_i\|_2^2}\mathcal{V}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i) \tag{10.43}$$

is positive or zero. The non-negativity of the integrand then implies that even the Lebesgue integral over this term, i.e. the associated form factor F_{ij}, is not negative. Furthermore it holds, only in the case where the patches P_i and P_j are mutually invisible:

$$F_{ij} = 0, \tag{10.44}$$

for $i \neq j$. Obviously, the non-negativity of the classical form factor for the case $i = j$, thus $F_{ii} = 0$, requires that the patch P_i is planar or convex.

RECIPROCITY OF THE CLASSICAL FORM FACTORS. Another useful property of the classical form factor is the principle of reciprocity. It is also a consequence from the definition of the form factor as a Lebesgue integral.

LEMMA 10.1 (Reciprocity Relation of the Classical Form Factors) *Let F_{ij} be the classical form factor between the surface patches P_i and P_j for $1 \leq i,j \leq n$ as introduced in Definition 10.2, then F_{ij} satisfy the reciprocity relation:*

$$A_i F_{ij} = A_j F_{ji}. \tag{10.45}$$

PROOF 10.1 *Applying the Theorem of Fubini-Tonelli to the definition of the form factor leads to:*

$$\begin{aligned}
A_i F_{ij} &\overset{(10.26)}{=} A_i \frac{1}{A_i} \int_{P_i} \left(\int_{P_j} \mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\mu^2(\mathbf{s}_j) \right) d\mu^2(\mathbf{s}_i) && (10.46)\\
&\overset{\text{Theorem 2.6}}{=} A_j \frac{1}{A_j} \int_{P_j} \left(\int_{P_i} \mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\mu^2(\mathbf{s}_i) \right) d\mu^2(\mathbf{s}_j) && (10.47)\\
&\overset{\mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i) = \mathcal{G}'(\mathbf{s}_i \leftrightarrow \mathbf{s}_j)}{=} A_j \frac{1}{A_j} \int_{P_j} \left(\int_{P_i} \mathcal{G}'(\mathbf{s}_i \leftrightarrow \mathbf{s}_j)\, d\mu^2(\mathbf{s}_i) \right) d\mu^2(\mathbf{s}_j) && (10.48)\\
&\overset{(10.26)}{=} A_j F_{ji}. && (10.49)
\end{aligned}$$

The reciprocity relation of the classical form factor has the beautiful property, that with the existence of the form factor F_{ij} the form factor F_{ji} is also given, namely via the relation

$$F_{ji} = \frac{A_i}{A_j} F_{ij}. \tag{10.50}$$

This means, utilizing of the reciprocity property of the classical form factors reduces the number of evaluating the form factor integral from (10.26) to the halve.

EXAMPLE 10.3 *Let us consider such a complex scene as introduced in Example 10.1, consisting of 10^5 surface patches. The associated radiosity matrix $\mathbf{M}$ then contains 10^{10} entries. Assuming that all patches are planar, then it holds: $F_{ii} = 0$ for $1 \leq i \leq 10^5$. Using the property from Equation (10.50), namely $F_{ji} = \frac{A_i}{A_j} F_{ij}$, then the number of from factors which must be evaluated reduces to $10^5(10^5 - 1)/2$. Even if this is much smaller then the number of form factors in Example 10.1, a radiosity algorithms has to evaluate almost 5 billions of form factor integrals.*

REMARK 10.5 (A More Intuitive Formulation of the Classical Radiosity Equation) *Let us consider once more the classical radiosity equation from Relation (10.25), namely,*

$$B_i = B_{e_i} + \rho_i \sum_{j=1}^{n} F_{ij} B_j. \tag{10.51}$$

Multiplying the classical radiosity equation by the area of the surface patch A_i

and using the reciprocity relation of the classical form factor leads to:

$$\underbrace{A_i B_i}_{\Phi_i} = A_i B_{e_i} + \rho_i \sum_{j=1}^{n} F_{ij} A_i B_j \tag{10.52}$$

$$\overset{(10.50)}{=} \underbrace{A_i B_{e_i}}_{\Phi_{e_i}} + \rho_i \underbrace{\sum_{j=1}^{n} F_{ji} \underbrace{A_j B_j}_{\Phi_j}}_{\Phi_i}, \tag{10.53}$$

Φ (249) *where Φ_i, Φ_{e_i} is the power leaving respectively emitted by the patch P_i.*

Obviously, the classical radiosity equation can now be interpreted as follows: The Radiant Power (249) *power leaving patch P_i is the sum of two terms, namely, the power emitted directly by patch P_i, and the power reflected at P_i after propagated from all patches P_j visible to P_i.*

The second term then tells us to look around at every patch P_j in the environment to determine the area power density of that patch. The form factor F_{ij} then makes a statement about, how much of the power density of P_j reaches patch P_i. The area power densities from all patches that contribute to patch P_i are then accumulated and scaled by the reflectivity ρ_i of patch P_i. Adding this amount to the emitted power per unit area on patch P_i then results in the outgoing radiosity of P_i.

Law of Energy Conservation (332) ENERGY CONSERVATION OF THE CLASSICAL FORM FACTORS. Due to the principle of energy conservation, the following lemma holds for any surface patch P_i.

LEMMA 10.2 (Energy Conservation of the Classical Form Factors) *Let F_{ij} be the classical form factor between the surface patches P_i and P_j for $1 \leq i, j \leq n$ in a closed scene. Then the form factor satisfy the principle of energy conservation from 4.2.2.3, that is, for all i, with $1 \leq i \leq n$, it holds:*

$$\sum_{j=1}^{n} F_{ij} = 1. \tag{10.54}$$

Note: If the scene is not closed, energy can be lost, that is, the sum of the form factors is less than 1.

PROOF 10.2 *Using the partition $\partial\mathcal{V} = \bigcup_{j=1}^{n} P_j$ then it holds for the form factor F_{ij}*

between the patch P_i *and all other patches* P_j*:*

$$\sum_{j=1}^{n} F_{ij} \overset{(10.26)}{=} \frac{1}{A_i} \sum_{j=1}^{n} \int_{P_i} \left(\int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) \right) d\mu^2(s_i) \tag{10.55}$$

$$\overset{(2.248)}{=} \frac{1}{A_i} \int_{P_i} \left(\sum_{j=1}^{n} \int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) \right) d\mu^2(s_i) \tag{10.56}$$

$$\overset{(10.15)}{=} \frac{1}{A_i} \int_{P_i} \left(\int_{\partial \mathcal{V}} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) \right) d\mu^2(s_i). \tag{10.57}$$

Changing the integration measure within the inner integral—from Lebesgue area measure to the projected solid angle measure—the inner integral can be expressed as an integral over the upper hemisphere $\mathcal{H}^2_+$ *about* s_i*, i.e. we can write:*

$$\sum_{j=1}^{n} F_{ij} = \frac{1}{A_i} \sum_{j=1}^{n} \int_{P_i} \left(\int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) \right) d\mu^2(s_i) \tag{10.58}$$

$$\overset{(2.199)}{=} \frac{1}{A_i} \int_{P_i} \left(\frac{1}{\pi} \int_{\mathcal{H}^2_+(s_i)} d\sigma^{\perp}_{s_i}(\omega_i) \right) d\mu^2(s_i) \tag{10.59}$$

$$\overset{(2.300)}{=} \frac{1}{A_i} \int_{P_i} \frac{\pi}{\pi} d\mu^2(s_i) = 1, \tag{10.60}$$

where the factor $\frac{1}{\pi}$ *in the second line comes from the definition of the classical radiosity geometry factor,* $\mathcal{G}'$*, for details see Equation (10.14).*

ADDITIVITY OF THE CLASSICAL FORM FACTORS. Let us consider three disjoint patches P_i, P_j and P_k. Then it should be clear that the power emitted from patch P_i after received from the patches $P_j \cup P_k$ is equal to the sum of the power emitted from patch P_i after received by each of the patches, that is, it holds the following lemma:

LEMMA 10.3 (Additivity of the Classical Form Factors) *Let* $\bigcup_{j=1}^{n} P_j$ *be a disjoint partition of a closed scene. Then the classical form factors are* additive*, that is,*

$$F_{i(\bigcup_{j=1}^{n} j)} = \sum_{j=1}^{n} F_{ij}. \tag{10.61}$$

PROOF 10.3 *The additivity of the form factor is a direct consequence of the countably*

additivity of the Lebesgue integral from Lemma 2.2, since it holds:

$$F_{i(\bigcup_{j=1}^{n} j)} \overset{(10.26)}{=} \frac{1}{A_i} \int_{P_i} \left(\int_{\bigcup_{j=1}^{n} P_j} G'(s_j \leftrightarrow s_i) \, d\mu^2(s_j) \right) d\mu^2(s_i) \tag{10.62}$$

$$\overset{(2.248)}{=} \frac{1}{A_i} \int_{P_i} \left(\sum_{j=1}^{n} \int_{P_j} G'(s_j \leftrightarrow s_i) \, d\mu^2(s_j) \right) d\mu^2(s_i) \tag{10.63}$$

$$\overset{(2.248)}{=} \sum_{j=1}^{n} \frac{1}{A_i} \int_{P_i} \left(\int_{P_j} G'(s_j \leftrightarrow s_i) \, d\mu^2(s_j) \right) d\mu^2(s_i) \tag{10.64}$$

$$\overset{(10.26)}{=} \sum_{j=1}^{n} F_{ij}. \tag{10.65}$$

The additivity property of the classical form factor is a useful tool to determine the full form factor by decomposing surface patches into simpler shapes or sub-elements.

EXAMPLE 10.4 *Let us show the additivity of the classical form factors at the example of three disjoint patches P_i, P_j and P_k as visualized in Figure 10.7. Due to the above lemma, we get:*

$$F_{i(j \cup k)} \overset{(10.26)}{=} \frac{1}{A_i} \int_{P_i} \left(\int_{P_j \cup P_k} G'(s \leftrightarrow s_i) \, d\mu^2(s) \right) d\mu^2(s_i) \tag{10.66}$$

$$\overset{(2.248)}{=} \frac{1}{A_i} \int_{P_i} \left(\int_{P_j} G'(s_j \leftrightarrow s_i) \, d\mu^2(s_j) \right) d\mu^2(s_i) + \tag{10.67}$$

$$\frac{1}{A_i} \int_{P_i} \left(\int_{P_k} G'(s_k \leftrightarrow s_i) \, d\mu^2(s_k) \right) d\mu^2(s_i) \tag{10.68}$$

$$\overset{(10.26)}{=} F_{ij} + F_{ik}. \tag{10.69}$$

REMARK 10.6 *For a partition $\bigcup_{j=1}^{n} P_j$ of a closed scene, where the patches are not required to be disjoint, a slightly modified formula exists:*

$$F_{i(\bigcup_{j=1}^{n} j)} = \sum_{j=1}^{n} F_{ij} - \sum_{1 \le j < k \le n} F_{i(j \cap k)} \tag{10.70}$$

where $P_j \cap P_k \neq \emptyset$.

REMARK 10.7 *Note: The reverse statement of the above lemma is not true, that is,*

$$F_{(\bigcup_{j=1}^{n} j)i} \neq \sum_{j=1}^{n} F_{ji}. \tag{10.71}$$

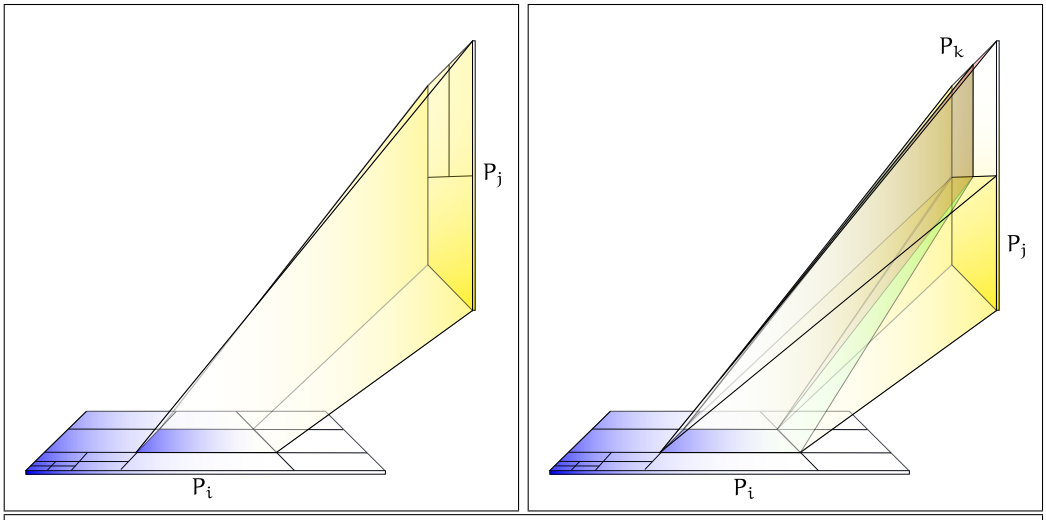


FIGURE 10.7: ADDITIVITY OF THE CLASSICAL FORM FACTORS. The form factor $F_{i(j\cup k)}$ between three patches P_i, P_j and P_k can easily be computed by addition of the form factor, F_{ij}, between P_i and P_j, and the form factor F_{ik}. Note: The additivity of form factors is only valid in one direction. Due to Remark 10.7 it does not hold: $F_{(j\cup k)i} = F_{ji} + F_{ki}$.

This can easily be seen by:

$$F_{(\bigcup_{j=1}^{n} j)i} \overset{(10.26)}{=} \frac{1}{\sum_{j=1}^{n} A_j} \int_{\bigcup_{j=1}^{n} P_j} \left(\int_{P_i} \mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\mu^2(\mathbf{s}_i) \right) d\mu^2(\mathbf{s}_j) \quad (10.72)$$

$$\overset{(2.248)}{=} \frac{1}{\sum_{j=1}^{n} A_j} \sum_{j=1}^{n} \underbrace{\int_{P_j} \left(\int_{P_i} \mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\, d\mu^2(\mathbf{s}_i) \right) d\mu^2(\mathbf{s}_j)}_{A_j F_{ji}} \quad (10.73)$$

$$\overset{(10.26)}{=} \frac{1}{\sum_{j=1}^{n} A_j} \sum_{j=1}^{n} A_j F_{ji} \quad (10.74)$$

$$= \sum_{j=1}^{n} \frac{A_j}{\sum_{j=1}^{n} A_j} F_{ji} \quad (10.75)$$

$$\neq \sum_{j=1}^{n} F_{ji} \quad (10.76)$$

for $\frac{A_j}{\sum_{j=1}^{n} A_j} \neq 1$. *Thus, the form factor from the union of* $\bigcup_{j=1}^{n} P_j$ *to element* P_i *is the area average of all individual patches* P_j.

10.1.3.2 CHARACTERIZING THE CLASSICAL FORM FACTOR SOLUTIONS

Let us consider the taxonomy of form factor algorithms, visualized in Figure 10.8. As we can see from the diagram, there exists two main branches in the tree of solving algorithms for the form factor integral:

- analytic approaches, and
- numeric methods.

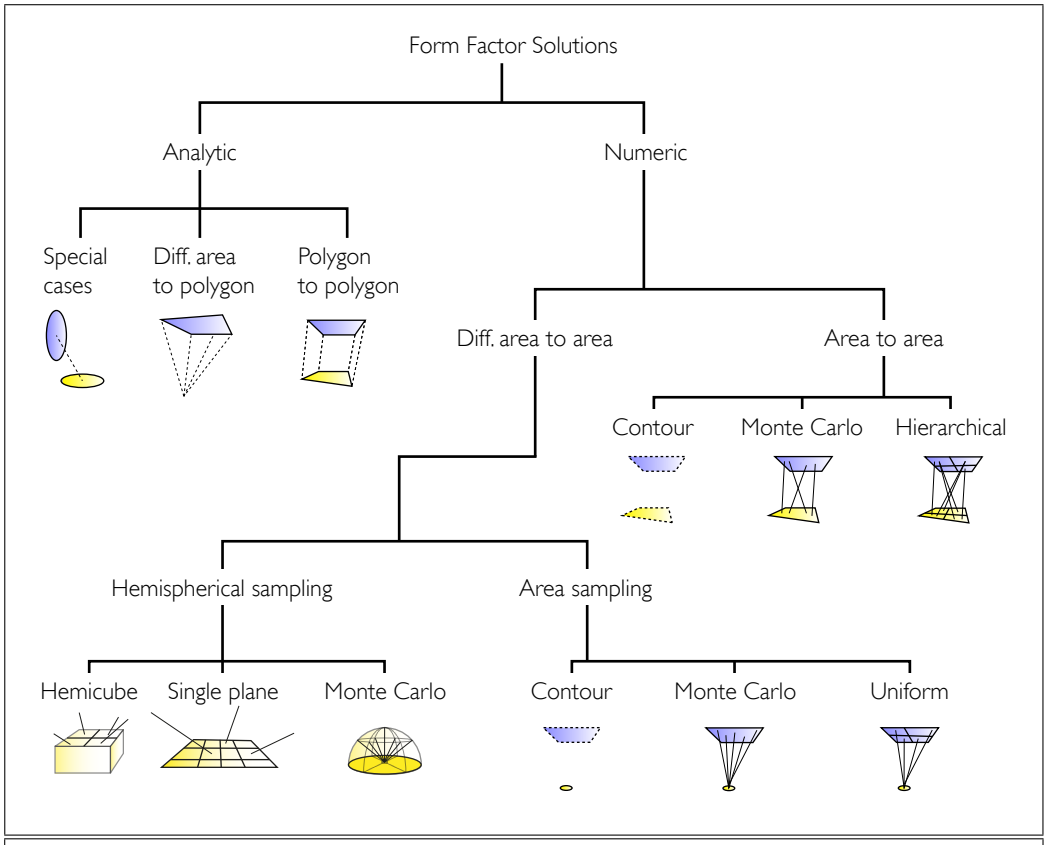


FIGURE 10.8: A TAXONOMY OF FORM FACTOR ALGORITHMS. There are two main branches in the tree of algorithms for computing the form factor integral: analytic and numerical methods. Closed form factor formulae are only available for various differential and finite geometries. There is no closed form solution, neither for the general form factor integral nor for the form factor integral associated with complex shapes. In all these cases, numerical approaches are required to approximate the resulting form factor integral. Reprinted, by permission, from [36, Cohen & Wallace 1993].

As the diagram shows, closed form factor formulae are only available for various differential and finite geometries, such as opposing and perpendicular rectangles, circles or polygons. Even, via form factor algebra, form factors for the union or difference of simple areas can be computed from the form factors to these individual cases. But there is no closed form solution, neither for the general form factor integral nor for the form factor integral associated with complex shapes. In all these cases, numerical approaches are required to approximate the resulting form factor integral.

Section 10.1.3.2.1

Section 10.1.3.2.2

10.1.3.2.1 CLOSED FORM SOLUTIONS FOR FORM FACTORS

Successfully solving the radiosity equation requires the accurate computation of the radiosity matrix $\mathbf{M}$, whose entries contain the classical form factor integral from Relation (10.26). Thus, the first idea for solving the form factor integral was to find an analytical solution.

M (787)

Since the visibility function $\mathcal{V}$ in the form factor integral can usually not be captured analytically, all algorithms, that tries to solve the form factor integral analytically, assume that the patches within a scene are all fully visible to each other. This means, that the visibility term in the form factor integral is equal to one, and the form factor integral can simplified be expressed as:

$$F_{ij} \stackrel{def}{=} \frac{1}{A_i} \int_{P_i} \left(\int_{P_j} \mathcal{G}'(\mathbf{s}_j \leftrightarrow \mathbf{s}_i) \, d\mu^2(\mathbf{s}_j) \right) d\mu^2(\mathbf{s}_i) \tag{10.77}$$

$$= \frac{1}{A_i} \int_{P_i} \left(\int_{P_j} \frac{\left| \cos\theta_o^j \cos\theta_i^i \right|}{\pi \|\mathbf{s}_j - \mathbf{s}_i\|_2^2} \, d\mu^2(\mathbf{s}_j) \right) d\mu^2(\mathbf{s}_i). \tag{10.78}$$

DIRECT INTEGRATION. Now, the integral from Equation (10.78) can be integrated directly, but, due to its complexity, unfortunately only for very simple arrangements, such as parallel and perpendicular rectangles, circles, and hollow tubes [88, Howel 1982]. Although the patches are assumed to be simple and unoccluded, the resulting formulae are anything but simple.

As we have seen in Section 10.1.3, a significant simplification for the computation of form factors results if the area of patch P_i is small compared with the distance to patch P_j. In this case, the luminous surface emitter P_i can be modeled as a point light source at point $\mathbf{s}_i \in P_i$, and the associated form factor is given via the unocluded differential-to-finite element form factor $F_{\mathbf{s}_i P_j}$ from Equation (10.39), namely,

$$F_{\mathbf{s}_i P_j} \stackrel{def}{=} \int_{P_j} \frac{\left| \cos\theta_o^j \cos\theta_i^i \right|}{\pi \|\mathbf{s}_j - \mathbf{s}_i\|_2^2} \, d\mu^2(\mathbf{s}_j). \tag{10.79}$$

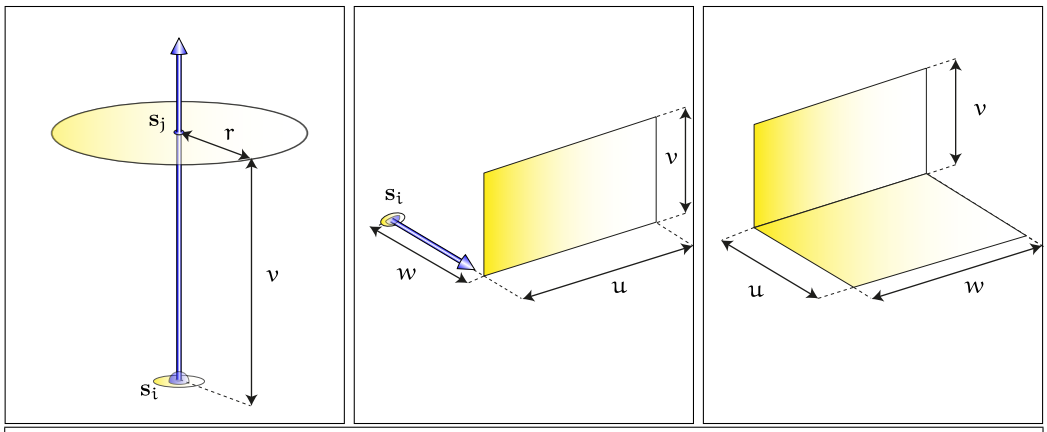


FIGURE 10.9: SIMPLE ARRANGEMENTS FOR COMPUTING CLOSED SOLUTIONS FOR FORM FACTORS. Left, the geometry for computing the form factor between point s_i and a disc with radius r perpendicular to the direction joining s_i and center s_j of the disc. In the center, the geometry for the form factor between a surface point s_i and a parallel rectangle perpendicular to the direction joining s_i to the lower left corner, and on the right, the geometry for computing the form factor between two perpendicular rectangles having a common edge. Reprinted, by permission, from [190, Sillion & Puech 1994].

EXAMPLE 10.5 (Simple Examples for Computing Closed Solutions for Form Factors) *Let us consider the simple arrangements shown in Figure 10.9. The classical form factor between the surface point s_i and the disc perpendicular to the direction joining s_i and the center s_j of the disc P_{disc} is given by:*

$$F_{s_i P_{disc}} \overset{(10.39)}{=} \int_{P_{disc}} \frac{\left|\cos\theta_o^j \cos\theta_i^i\right|}{\pi \left\|s_j - s_i\right\|_2^2} \, d\mu^2(s_j) \tag{10.80}$$

$$= \frac{r^2}{r^2 + v^2}. \tag{10.81}$$

The form factor between point s_i and a parallel rectangle perpendicular to the direction joining s_i to the lower left corner corresponds to:

$$F_{s_i P_{rect}} \overset{(2.207)}{=} \int_{P_{rect}} \frac{\left|\cos\theta_o^j \cos\theta_i^i\right|}{\pi \left\|s_j - s_i\right\|_2^2} \, d\mu^2(s_j). \tag{10.82}$$

Due to [190, Sillion & Puech 1994] a closed solution for the form factor $F_{s_i P_{rect}}$ can be computed via:

$$F_{s_i P_{rect}} = \frac{1}{2\pi} \left\{ \frac{X}{\sqrt{1+X^2}} \tan^{-1}\left(\frac{Y}{\sqrt{1+Y^2}}\right) + \frac{X}{\sqrt{1+X^2}} \tan^{-1}\left(\frac{Y}{\sqrt{1+Y^2}}\right) \right\} \tag{10.83}$$

with $X = \frac{u}{w}$ *and* $Y = \frac{v}{w}$.

For the sake of completeness, we give also the closed form solution of the form factor F_{ij} *between two perpendicular rectangles having a common edge, as seen in the right image of Figure 10.9. Due to [190, Sillion & Puech 1994] it holds:*

$$\begin{aligned} F_{ij} = & \frac{1}{\pi X}\left\{X\tan^{-1}\left(\frac{1}{X}\right) + X\tan^{-1}\left(\frac{1}{Y}\right) - \sqrt{X^2+Y^2}\tan^{-1}\left(\frac{1}{\sqrt{X^2+Y^2}}\right)\right\} + \\ & \frac{1}{4\pi X}\left\{\ln\left(\frac{(1+X^2)(1+Y^2)}{1+X^2+Y^2}\right) + X^2\ln\left(\frac{X^2(1+X^2+Y^2)(1+X^2)}{X^2+Y^2}\right) + \right. \\ & \left. Y^2\ln\left(\frac{Y^2(1+X^2+Y^2)(1+Y^2)}{X^2+Y^2}\right)\right\} \end{aligned} \tag{10.84}$$

with $X = \frac{u}{w}$ *and* $Y = \frac{v}{w}$.

REMARK 10.8 *In [180, Schröder & Hanrahan 1993] a closed form solution for general polygon-to-polygon form factors is derived. The resulting formula consist of a long series of complex terms and is based on the contour integral.*

Contour Integral (803)

Apart from direct integration, another technique that can help us to derive analytical formulas for form factors is *form factor algebra*.

FORM FACTOR ALGEBRA. Under the notion of the *form factor algebra* we understand a system of rules that can be used to compute new form factors for more complex geometries from already given formulas for simple form factors. Typical operations of a form factor algebra are the additivity and the reciprocity relation of the classical form factors, as introduced in Section 10.1.3.1, thus,

$$F_{i(\bigcup_{j=1}^{n} j)} = \sum_{j=1}^{n} F_{ij} \tag{10.85}$$

$$F_{i(\bigcup_{j=1}^{n} j)} = \sum_{j=1}^{n} F_{ij} - \sum_{1\le j<k\le n} F_{i(j\cap k)} \tag{10.86}$$

$$F_{ij} = \frac{A_j}{A_i} F_{ji}, \tag{10.87}$$

where the first formula only holds for a partition of the scene consisting of disjoint surface patches $\bigcup_{j=1}^{n} P_j$.

Due to [190, Sillion & Puech 1994], the form factor algebra is an interesting alternative to the brute force computation of form factors, but it is difficult to apply them automatically.

REMARK 10.9 *Let us finally summarize:*

i) When it is known that no occluders are located between two interacting surfaces then the analytic form factor calculation offer the best accuracy at reasonable cost for planar surfaces.

ii) For the partial occlusion case, the analytical formula usually cannot be derived.

iii) Computing visibility in the form factor integral is like solving a hidden surface problem from the point of view of each patch in the scene, usually the most costly part of the radiosity computation. It can easily but costly be solved by ray tracing.

Section (664)

10.1.3.2.2 NUMERICAL SOLUTIONS FOR FORM FACTORS

As we have seen in the last section closed form analytical solutions to the form factor integral, are only be available for various simple differential and finite geometries. Form factor algebra can also be helpful to produce a closed solution by combining already known form factor formulae to these individual cases. But there is no closed form solution, neither for the general form factor integral nor for the form factor integral associated with complex shapes. In all these cases, numerical approaches are required to approximate the resulting form factor integral.

Section ?? NUMERICAL INTEGRATION. In Chapter 6 we have shortly talked about numerical integration and we have presented with the *Newton-Cotes formulas* and the *Gauss rules* the two most popular numerical procedure for integrating functions. Both procedures are suitable for an approximative evaluation of the form factor integral. As we have seen, the more samples are selected to evaluate the kernel, the more accurate is the approximation, where the cost of the approximation are directly related to the number of kernel evaluations.

MONTE CARLO INTEGRATION. Often, the kernel in the light transport equation is discontinuous and of high dimension. So, we have seen in Section ?? that the convergence of the Gauss rules is of order $O(N^{\frac{-r}{s}})$. This entails, for large $s, s > 5$, highly complex and time-consuming procedures, inappropriate for calculating the integral. As an alternative, we should use *Monte Carlo integration* in connection with variance reduction techniques detailed discussed in Chapter 6.

EXAMPLE 10.6 (A Trivial Monte Carlo Strategy for Form Factor Calculation) *The form factor calculation between two patches P_i and P_j requires to solve a double integral over the surfaces of the involved patches. This can easily be done by sampling pairs of uniformly distributed random variables $(\mathbf{X}_i, \mathbf{X}_j)$ with $\mathbf{X}_k = (X_{k_1}, X_{k_2}), k = i, j$ from the probability space $(P_i \times P_j, \mathfrak{B}(P_i \times P_j), \mathbb{P}_{\mathbf{X}})$. The associated PDF, $p_{\mathbf{X}_i, \mathbf{X}_j}$, is then*

Uniform Distributed RV (180)
Probability Space (163)

given by:

$$p_{\mathbf{X}_i,\mathbf{X}_j}(\mathbf{x}_i,\mathbf{x}_j) = \frac{1}{\mu^2(P_i)\mu^2(P_j)} = \frac{1}{A_i A_j}, \tag{10.88}$$

where we assume that it holds: $\mu^2(P_k) = A_k$ *for* $k \in \{i,j\}$.

A secondary Monte Carlo estimator $F_N^{F_{ij}}$ *then has the form:*

$$F_N^{F_{ij}}(\mathbf{X}_i,\mathbf{X}_j) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{A_i}\frac{\mathcal{G}'(\mathbf{X}_j \leftrightarrow \mathbf{X}_i)}{p_{\mathbf{X}_i,\mathbf{X}_j}(\mathbf{X}_i,\mathbf{X}_j)} \tag{10.89}$$

$$= \frac{A_j}{N}\sum_{i=1}^{N}\mathcal{G}'(\mathbf{X}_j \leftrightarrow \mathbf{X}_i). \tag{10.90}$$

10.1.3.2.2.1 HEMISPHERE SAMPLING FOR DIFFERENTIAL-TO-FINITE-AREA FORM FACTORS

In Definition 10.2, we introduced the classical form factor as a double surface integral, where the integration domain of the inner integral corresponds to the surface patch P_j and integration with respect to the outer integral works over the considered patch P_i. Using the measure transformation from Equation (2.196), then the inner integration within the form factor integral can be replaced by an integration over the hemisphere, resulting in:

$$F_{ij} \overset{(2.196)}{=} \frac{1}{A_i}\int_{P_i}\int_{\mathcal{H}^2_+(\mathbf{s}_i)}\frac{\left|\cos\theta_i^i\right|}{\pi}\mathcal{V}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\,d\sigma_{\mathbf{s}_i}(\omega_i)d\mu^2(\mathbf{s}_i). \tag{10.91}$$

The form factor F_{ij} can now easily be evaluated by a Monte Carlo scheme using uniformly distributed random numbers from probability space $(P_i \times \mathcal{H}^s_+, \mathfrak{B}(P_i \times \mathcal{H}^s_+), \mathbb{P})$. This form of the form factor integral is often used in algorithms to compute the form factors from patch P_i to all elements at once.

Often the inner integral in Formula (10.91) is only evaluated for a single point on patch P_i. An interesting example for this is the Nusselt analog, introduced in Example 2.34. Based on Equation (10.91), the form factor $F_{\mathbf{s}P_j}$ was derived via:

$$F_{\mathbf{s}P_j} \overset{def}{=} \int_{\mathcal{H}^2_+(\mathbf{s}_i)}\frac{\left|\cos\theta_i^i\right|}{\pi}\mathcal{V}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\,d\sigma_{\mathbf{s}_i}(\omega_i) \tag{10.92}$$

$$\overset{(2.207)}{=} \int_{P_j}\frac{\left|\cos\theta_i^i\cos\theta_o^j\right|}{\pi\|\mathbf{s}_j-\mathbf{s}_i\|_2^2}\mathcal{V}(\mathbf{s}_j \leftrightarrow \mathbf{s}_i)\,d\mu^2(\mathbf{s}_j). \tag{10.93}$$

In Example 2.34, we have derived this formula, by projecting the area of the patch P_j, visible from the center of the upper hemisphere about point $\mathbf{s}_i$, onto $\mathcal{H}^2_+(\mathbf{s}_i)$. Dividing the orthogonal projection of the radial projected patch by the area of base of the hemisphere then results in the Nusselt analog.

Let us now consider two algorithms that are based on the Nusselt analog:

EXAMPLE 10.7 *Let us discretize the upper hemisphere $\mathcal{H}^2_+$ about point s_i into a finite set of* cells, *corresponding to small area patches onto $\mathcal{H}^2_+(s_i)$. We can then compute so-called* Δ-form factors, $F_{s_i\Delta_j}$, *where Δ_j denotes the j^{th} cell on $\mathcal{H}^2_+(s_i)$. Obviously, $F_{s_i\Delta_j}$ can be computed by multiplying its area $\mu^2(\Delta_j)$ by the $\cos\theta^j_o$ divided by π, thus,*

$$F_{s_i\Delta_j} = \frac{\mu^2(\Delta_j)\,\cos\theta^j_o}{\pi}. \tag{10.94}$$

By projecting the patch P_j onto $\mathcal{H}^2_+(s_i)$ and registering, which of the cells are covered by the projection of P_j, the differential-to-finite-aera form factor $F_{s_iP_j}$ can then easily be approximated by the sum of the Δ-form factors of the covered cells Δ_j.

In practice, the above algorithm is not really usable, as there are problems with the partition of the hemisphere into a finite set of equal sized patches. But, the Nusselt analog is the base of an other algorithm, which was the first algorithm who has made the radiosity method applicable: the hemicube method.

EXAMPLE 10.8 (The Hemicube Method) *The idea behind the hemicube method is, to project all surfaces of the scene onto the surface area of a hemicube about the center s_i of a differential patch, where the surface of the hemicube himself is partitioned into a finite net of small rectangular or squared patches, the so-called hemicube-pixels. As in the previous example, we can then compute the differential-to-finite-area Δ-form factors. The form factor F_{sP_j} can then be approximated by summing up all Δ-form factors of the hemicube-pixels, that was covered by the projection of the patch P_j onto the surface area of the hemicube.*

10.1.3.2.2.2 AREA SAMPLING FOR DIFFERENTIAL-TO-FINITE-AREA AND FINITE-TO-FINITE-AREA FORM FACTORS

For computing the differential-to-finite-area form factor, in the previous section we have transformed the inner surface integral of the form factor integral into a hemispherical integral. But we can also determine the differential-to-finite-area form factor based on the form factor integral as a double surface integral. Then, the differential-to-finite-area form factor has the form:

$$F_{s_iP_j} = \int_{P_j} \frac{\left|\cos\theta^i_i \cos\theta^j_o\right|}{\pi\|s_j - s_i\|^2_2}(s_j \leftrightarrow s_i)\,d\mu^2(s_j) \tag{10.95}$$

Integrals of this type can easily be evaluated via Monte Carlo methods. Monte carlo methods are also the preferred numerical procedures for computing the finite-to-finite-area form factor, thus, the double surface integral

$$F_{ij} = \frac{1}{A_i}\int_{P_i}\int_{P_j} \frac{\left|\cos\theta^i_i \cos\theta^j_o\right|}{\pi\|s_j - s_i\|^2_2}(s_j \leftrightarrow s_i)\,d\mu^2(s_j)\,d\mu^2(s_i) \tag{10.96}$$

from Definition 10.2. A naive Monte Carlo methods could be to sample points $\mathbf{X}_i$, which are uniformly distributed on patch P_i, according a probability density and evaluating the differential-to-finite-area form factor $F_{\mathbf{X}_i P_j}$ as the average sum of the function evaluated at the chosen samples.

CONTOUR INTEGRAL. Another strategy that can be useful to find solutions for the form factor integral comes from vector analysis and is based on *Stoke's theorem*. Using Stokes' Theorem, the classical form factor integral from Equation (10.78) can be reduced to a so-called *contour integral* over the boundaries of the two involved patches P_i and P_j. Due to [200, Sparrow & Cess 1978] the form factor integral from Equation (10.78) can then be written as

$$F_{ij} \stackrel{def}{=} \frac{1}{2\pi A_i} \oint_{C_i} \oint_{C_j} \ln(r) \langle d\mu(\mathbf{x}_i), d\mu(\mathbf{x}_j) \rangle, \tag{10.97}$$

where C_i and C_j are the boundaries of the elements P_i and P_j, $r = \|\mathbf{x}_j - \mathbf{x}_i\|_2$ is the distance between the points $\mathbf{x}_i = (x_{i_1}, x_{i_2}, x_{i_3}) \in C_i$ and $\mathbf{x}_j = (x_{j_1}, x_{j_2}, x_{j_3})$ on each boundary, and $d\mu(x_i)$ and $d\mu(x_j)$ are the differential vectors along the contours of the patches. In the case of polygon contours, the integral from Relation (10.97) can be evaluated relatively easily by means of numerical methods, such as quadrature methods, as introduced in Section **??**.

$\langle \cdot, \cdot \rangle$ (859)

EXAMPLE 10.9 (Unoccluded Form Factor between a Point and a Convex Planar Polygon) *In [139, Nishita & Nakame 1985] the contour integral was applied to the geometry of a polygon resulting in a surprisingly simple analytic solution for the form factor between a point and a convex planar Polygon with n vertices, namely,*

$$F_{\mathbf{s}P_j} = \frac{1}{2\pi} \sum_{k=0}^{n-1} \langle \mathbf{N}(\mathbf{s}), \Gamma_k \rangle. \tag{10.98}$$

Here, $\mathbf{N}(\mathbf{s})$ corresponds to the normal of surface at point $\mathbf{s}$ and Γ_k is a vector oriented in the direction of the cross product of the two vectors starting at $\mathbf{s}$ and ending in polygon vertices k and $k+1, k \geq 0$.

REMARK 10.10 *Form factor computation via the contour integral can be done for many polygons and shapes [189, Siegel & Howel 1992], but it is quite impracticable for our purposes. The contour integral approach is also used in the paper by [70, Goral & al. 1984] to simulate the light transport only in simple unoccluded environments. It cannot be extended to handle complex scenes with occluded polygons.*

10.1.4 SOLVING THE CLASSICAL DISCRETE RADIOSITY EQUATION

Solving the discrete formulation of the radiosity integral equation means solving a linear system of equations. In Section 2.3.3.2.4, we have already presented direct and iterative

methods for solving linear systems, such as the Gaussian elimination, or the iteration methods by Jacobi and Gauss-Seidel. In this section, we will pick up these techniques,- discuss their interpretation when applied to the classical discrete radiosity equation, but we will also introduce with the *Shouthwell relaxation* a new method that generates intermediate solutions that let the user monitor the process of rendering an image.

Section 10.1.4.1
Section 10.1.4.2

10.1.4.1 DIRECT METHODS

The radiosity matrix given in the form of Equation (10.35) is an $n \times n$ matrix without any useful properties. But making use of the reciprocity property of the form factor from Equation (10.45), namely,

$$F_{ji} = \frac{A_i}{A_j} F_{ji}, \tag{10.99}$$

then the radiosity matrix $\mathbf{M}$ can be made symmetric. If we also make use of the energy conservation of the classical form factors, derived in Section 10.1.3.1, then we ensure that with $\rho_i \leq 1$ the following inequality holds for all diagonal entries of $\mathbf{M}$:

$$\begin{aligned}
|m_{ii}| &= |1 - \rho_i F_{ii}| && (10.100)\\
&\overset{(10.54)}{=} \left| \sum_{j=1}^{n} F_{ij} - \rho_i F_{ii} \right| && (10.101)\\
&\overset{\rho_i < 1}{>} \left| \sum_{j=1}^{n} \rho_i F_{ij} - \rho_i F_{ii} \right| && (10.102)\\
&\overset{\Delta\text{-inequality}}{\geq} \sum_{\substack{j=1 \\ j \neq i}}^{n} |\rho_i F_{ij}| && (10.103)\\
&= \sum_{\substack{j=1 \\ j \neq i}}^{n} |m_{ij}|, && (10.104)
\end{aligned}$$

where we have used that the reflectivity ρ_i at the surfaces and the property that form factors are all non-negative real numbers.

Strong Row Sum Criterion (160)

Now, this inequality corresponds just to the strong row sum criterion from Lemma 2.62, which implies, that the radiosity matrix $\mathbf{M}$ of the classical radiosity equation system

$$\begin{aligned}
\mathbf{B}_e &= \mathbf{M}\mathbf{B} && (10.105)\\
&= (\mathbf{I} - \boldsymbol{\rho}\mathbf{F})\mathbf{B} && (10.106)
\end{aligned}$$

is invertible. So, a first idea for solving this system is to find the analytic solution

$$\mathbf{M}^{-1}\mathbf{B}_e = \mathbf{B}. \tag{10.107}$$

Replacing $\mathbf{M}$ by $\mathbf{I} - \boldsymbol{\rho}\mathbf{F}$, then Equation (10.107) can also be expressed as:

$$\mathbf{B} = \mathbf{M}^{-1}\mathbf{B}_e \tag{10.108}$$

$$\overset{(10.34)}{=} (\mathbf{I} - \boldsymbol{\rho}\mathbf{F})^{-1}\mathbf{B}_e. \tag{10.109}$$

But, this form of the solution of the classical radiosity system is known to us from our discussions in Section 2.3.3.1.1. So, the inverse matrix $(\mathbf{I} - \boldsymbol{\rho}\mathbf{F})^{-1}$ can be expressed via the Neumann series, that is, $(\mathbf{I} - \boldsymbol{\rho}\mathbf{F})^{-1}$ can be written as an infinite series of powers of the matrix $\boldsymbol{\rho}\mathbf{F}$, namely as,

Neumann Series (135)

$$(\mathbf{I} - \boldsymbol{\rho}\mathbf{F})^{-1} = \sum_{i=0}^{\infty} (\boldsymbol{\rho}\mathbf{F})^i \tag{10.110}$$

$$= \mathbf{I} + \boldsymbol{\rho}\mathbf{F} + (\boldsymbol{\rho}\mathbf{F})^2 + (\boldsymbol{\rho}\mathbf{F})^3 + \dots. \tag{10.111}$$

Applied to the radiosity system this leads to:

$$\mathbf{B} = \sum_{i=0}^{\infty} (\boldsymbol{\rho}\mathbf{F})^i \mathbf{B}_e \tag{10.112}$$

$$= \mathbf{B}_e + \boldsymbol{\rho}\mathbf{F}\mathbf{B}_e + (\boldsymbol{\rho}\mathbf{F})^2\mathbf{B}_e + (\boldsymbol{\rho}\mathbf{F})^3\mathbf{B}_e + \dots, \tag{10.113}$$

where each term $(\boldsymbol{\rho}\mathbf{F})^i\mathbf{B}_e$ represents the i^{th} bounce of the initially emitted light. Obviously, $\mathbf{B}_e$ corresponds to the direct illumination, $\boldsymbol{\rho}\mathbf{F}\mathbf{B}_e$ represents the illumination after one bounce, $(\boldsymbol{\rho}\mathbf{F})^2\mathbf{B}_e$ represents the illumination after two bounces, and so on. Under the assumption that the operator $\boldsymbol{\rho}\mathbf{F}$ is contracting—in the case where an operator is a matrix, we also speak of the *spectral radius*, that is, the absolute value of its largest eigenvalue has to be smaller than one—a solution to the classical radiosity system can then easily be determined by adding up the powers of the product of the matrices $\boldsymbol{\rho}$ and $\mathbf{F}$.

As already shown in Example 10.1, systems of equations underlying radiosity methods are quite large and relatively full which means that the associated matrices require, depending on the mesh discretization of the the scene, an enormous amount of storage. Coupled with this, direct solution methods, such as the Gaussian elimination, require $O(n^3)$ operations to solve a linear system of equations or to compute the inverse of the associated matrix, where n is the number of unknowns in the system. So, these methods are not suitable for solving a classical radiosity system. Here, other methods, as detailed discussed in Section 2.3.3.2.4.2, have been proven to be more efficient solvers.

10.1.4.2 RELAXATION METHODS

Let us now discuss the application of iteration methods, presented in Section 2.3.3.2.4.2, to the discrete radiosity equation from Definition 10.2. In particular, we will physically interprete the Jacobi and the Gauss-Seidel iteration when applied to the radiosity equation

system. Furthermore, we will present with the *Southwell relaxation*, a rather less known iteration method for solving linear systems of equations that has a significant advantage in contrast to the two other relaxation methods. It can produce an approximate solution of the radiosity problem without computing the entire radiosity matrix. Since computing the form factor matrix is the most costly part in a radiosity algorithms, Southwell relaxation scheme promises a boost in computation time for solving the classical radiosity equation.

SOLVING THE CLASSICAL RADIOSITY EQUATION VIA JACOBI AND GAUSS-SEIDEL ITERATION. In the previous section, we have shown, that the classical radiosity matrix satisfies the strong row sum criterion, that is, the classical radiosity problem can be solved via the classical iteration methods. As we have seen, the Jacobi iteration is the simplest iterative technique for solving systems of linear equations. Unfortunately, in practice it does not play any role as a solver for the radiosity system of equations, but it provides us with instructive insights into the physics behind the radiosity equation.

Jacobi Iteration (155)

Assuming all patches within a scene are planar or convex, then the form factors $F_{ii}, 1 \leq i \leq n$ are zero, which implies, that the diagonal elements of $\boldsymbol{\rho}\mathbf{F}$ are zero. This then allows, that the radiosity matrix $\mathbf{M}$ can be decomposed into a diagonal matrix, $\mathbf{I}$, a strictly lower diagonal matrix, $\mathbf{L}$, and a strictly upper diagonal matrix, $\mathbf{U}$, resulting in:

$\boldsymbol{\rho}\mathbf{F}$ (787)
$\mathbf{M}$ (787)

$$\mathbf{M} = \mathbf{I} - \mathbf{L} - \mathbf{U}, \tag{10.114}$$

where $(l_{ij})_{1 \leq i < j \leq n} = \rho_i F_{ij}$, and $(u_{ij})_{1 \leq j < i \leq n} = \rho_i F_{ij}$.

As the form factor matrix $\mathbf{M}$ can be decomposed in such a way, the radiosity equation system fulfills the prerequisites that it can be solved via one of the classical iteration methods. The i^{th} component of the new iterate $\mathbf{B}^{(k+1)}$ for the radiosity vector $\mathbf{B}$, computed via the Jacobi iteration, then looks like this:

$$B_i^{(k+1)} \overset{(2.502)}{=} B_{ei} - \rho_i \sum_{\substack{j=1 \\ j \neq i}}^{n} F_{ij} B_j^{(k)} \tag{10.115}$$

and computed via the Gauss-Seidel method it holds:

$$B_i^{(k+1)} \overset{(2.514)}{=} B_{ei} - \rho_i \sum_{j=1}^{i-1} F_{ij} B_j^{(k)} - \rho_i \sum_{j=i+1}^{n} F_{ij} B_j^{(k)} \tag{10.116}$$

with $k \geq 0$, $1 \leq i \leq n$ and $\mathbf{B}_e^{(0)} = \mathbf{B}_e$.

PHYSICAL INTERPRETATION OF THE JACOBI AND GAUSS-SEIDEL ITERATION. Let us consider once more the radiosity equation system from Relation (10.105) and its exact solution formulated as a Neumann series, namely,

$$\mathbf{B} \quad = \quad \mathbf{B}_e + \boldsymbol{\rho}\mathbf{F}\mathbf{B}_e + (\boldsymbol{\rho}\mathbf{F})^2\mathbf{B}_e + (\boldsymbol{\rho}\mathbf{F})^3\mathbf{B}_e + \ldots, \tag{10.117}$$

Now, as the Neumann series can be expressed via a recursive sequence, $(\mathbf{B}_n)_{n\in\mathbb{N}_0}$, from $\mathbb{R}^n$, where the vector valued function $\mathbf{B}_e$ is the starting value and the matrix product $\boldsymbol{\rho}\mathbf{F}$ corresponds to the operator defining the sequence, we get:

$$\mathbf{B}_{n+1} \overset{def}{=} \mathbf{B}_e + \sum_{i=1}^{n+1} (\boldsymbol{\rho}\mathbf{F})^i \mathbf{B}_e \tag{10.118}$$

$$= \mathbf{B}_e + (\boldsymbol{\rho}\mathbf{F}) \sum_{i=0}^{n} (\boldsymbol{\rho}\mathbf{F})^i \mathbf{B}_e \tag{10.119}$$

$$\overset{(10.118)}{=} \mathbf{B}_e + (\boldsymbol{\rho}\mathbf{F})\mathbf{B}_n, \qquad n \geq 0. \tag{10.120}$$

Obviously, this sequence converges to the exact solution of the classical discrete radiosity equation from Definition 10.2, that is, the propagation of light through an environment itself can be interpreted as an iterative method. While the Jacobi iteration simulates the propagation of light bouncing from surface to surface within a scene, the Gauss-Seidel iteration tries to anticipate the amount of light that each surface will receive from the next iteration of reflections.

In order to explain these statements more exactly, let us express the computations of $B_i^{(k+1)}$ resulting from the classical iteration procedures in terms of radiant power. For the Jacobi iteration we get, after multiplying the whole equation with the Lebesgue area measure $A_i = \mu^2(P_i)$ of patch P_i:

Radiant Power (249)

$$\underbrace{A_i B_i^{(k+1)}}_{\Phi_i^{(k+1)}} = A_i B_{e_i} - \rho_i \sum_{j=1}^{n} F_{ij} A_i B_j^{(k)} \tag{10.121}$$

$$\overset{(10.45)}{=} \underbrace{A_i B_{e_i}}_{\Phi_{e_i}} - \rho_i \sum_{j=1}^{n} F_{ji} \underbrace{A_j B_j^{(k)}}_{\Phi_j^{(k)}} \tag{10.122}$$

and for the iterate computed via the Gauss-Seidel iteration it holds:

$$\underbrace{A_i B_i^{(k+1)}}_{\Phi_i^{(k+1)}} = \underbrace{A_i B_{e_i}}_{\Phi_{e_i}} - \rho_i \sum_{j=1}^{i-1} F_{ji} \underbrace{A_j B_j^{(k)}}_{\Phi_j^{(k)}} - \rho_i \sum_{j=i+1}^{n} F_{ji} \underbrace{A_j B_j^{(k)}}_{\Phi_j^{(k)}}. \tag{10.123}$$

Based on these representations, the classical iteration methods can be interpreted as follows: In a single step, both procedures select a patch P_i and compute the emitted power Φ_i from this patch into the environment, as the sum of the self emitted power of patch P_i and the reflected fraction of power *gathered* from all other patches P_j in the scene, that are visible from patch P_i, see Figure 10.10. Obviously, both procedures compute the dot product of a vector of radiosities with a column of the radiosity matrix multiplied by the

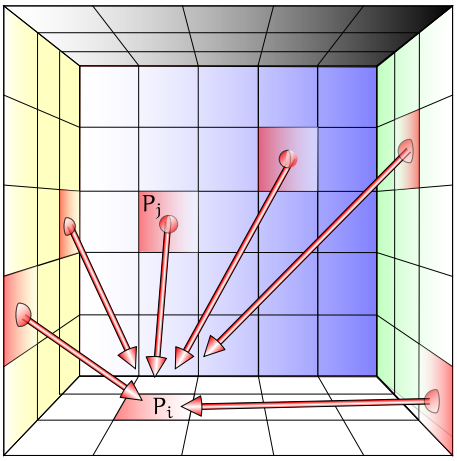


FIGURE 10.10: A GATHERING STEP OF THE CLASSICAL ITERATION METHODS. The radiant power reflected from patch P_i into the environment, is the sum of the self emitted power of patch P_i and the reflected fraction of power *gathered* from all other patches P_j within the scene that are visible from patch P_i.

reflectance ρ_i of patch Pi, see Figure 10.10.

THE SOUTHWELL RELAXATION. In the following, let us consider a linear system of equations of the form

$$\mathbf{MB} = \mathbf{B}_e, \tag{10.124}$$

like the discrete radiosity equation from Definition 10.2, where $\mathbf{M}$ is a $n \times n$ matrix with coefficients from $\mathbb{R}$ and $\mathbf{B_e}$ is a vector from $\mathbb{R}^n$.

As already mentioned above, the idea behind a relaxation method is that at each step of the algorithm one of the components of the residual vector $\mathbf{r}^{(k+1)}$ will be set to zero. The Gauss-Seidel iteration satisfies this idea, as it relaxes the components of the residual vector in turn, that is, at first the component $r_1^{(k+1)}$, then $r_2^{(k+1)}$, and finally $r_n^{(k+1)}$. The *Southwell relaxation*, which we will present in detail below, goes an other way. Instead to relax each component of the residual vector in turn, the algorithms selects the i^{th} equation of the system for which $\mathbf{r}^{(k+1)}$ has the largest residual, where we define

$$\max_{1\le i\le n} \mathbf{r}^{(k+1)} \stackrel{def}{=} \max_{1\le i\le n} \mathbf{r}^{(k+1)} \left(B_{e_i} - \sum_{j=1}^{n} m_{ij} B_j^{(k)} \right). \tag{10.125}$$

Since a large residual component implies that the associated component of the vector $\mathbf{B}$ has to be updated several times successively, it is difficult to predict when a particular

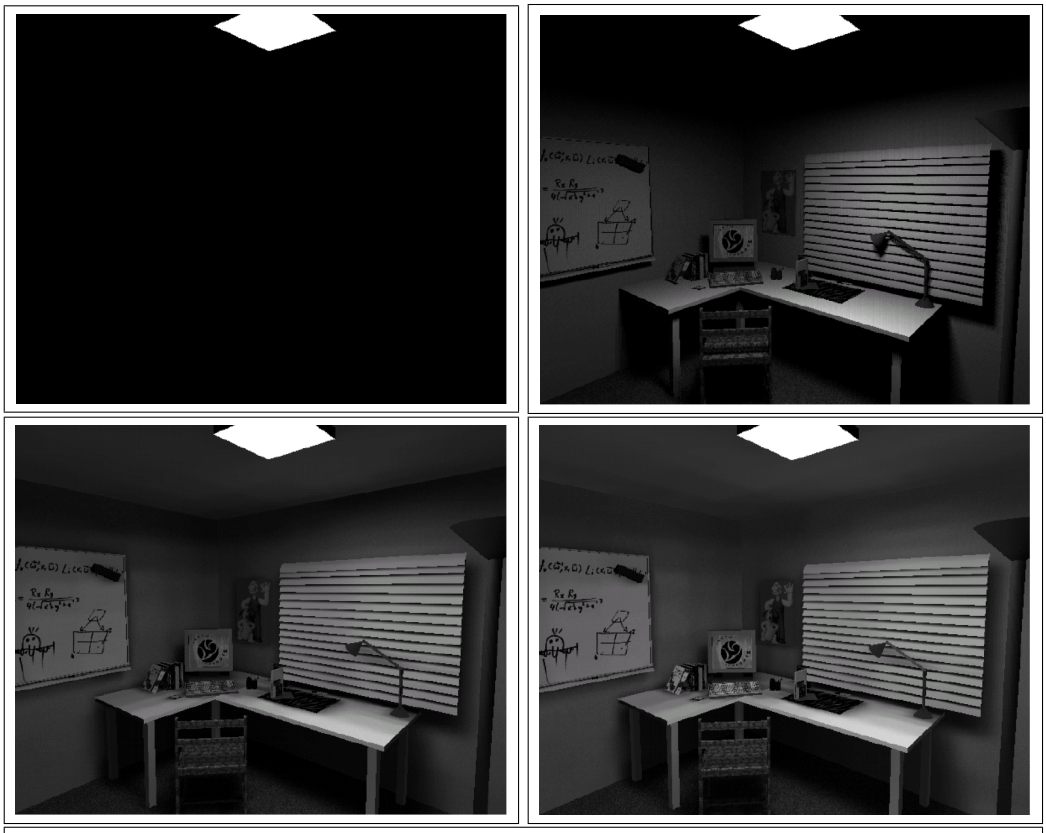


FIGURE 10.11: RADIOSITY SOLUTION VIA THE JACOBI ITERATION. The upper left image shows the radiosity solution after the first Jacobi iteration, so, the radiosity vector represents the emittance of the light source. The other images show the radiosity solution after $2, 3$ and 4 Jacobi iterations.Image courtesy by Karol Myszkowski.

variable will be changed in Southwell relaxation. Therefore we can not transfer the concepts of a step and an iteration cycle as introduced with the classical iteration method to the Southwell algorithm. Here, the superscript k indicates the step number instead as the iteration cycle like in Jacobi and Gauss-Seidel algorithm.

Let us assume, the i^{th} component of the residual has to vanish at step $k+1$, that is, it must hold:

$$0 = r_i^{(k+1)} \tag{10.126}$$

$$= B_{e_i} - \sum_{j=1}^{n} m_{ij} x_j^{(k+1)}. \tag{10.127}$$

Since it is the i^{th} component of the vector $\mathbf{B}$ that has to be modified, it holds

$B_j^{(k+1)} = B_j^{(k)}$ for all $j \neq i$. The new value of $B_i^{(k+1)}$ can then be extracted from Equation (10.127), namely,

$$B_i^{(k+1)} = \frac{1}{m_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} m_{ij} B_j^{(k)} \right). \tag{10.128}$$

Due to Equation (2.505), this equation can also be written in terms of the residual namely as

$$B_i^{(k+1)} = B_i^{(k)} + \frac{r_i^{(k)}}{m_{ii}}, \tag{10.129}$$

that is, this relaxation not only sets $r_i^{(k+1)}$ to zero, but also leads to a modification of all other components of the residual $\mathbf{r}^{(k+1)}$.This can easily be shown as follows: As $\mathbf{B}^{(k+1)}$ and $\mathbf{B}^{(k)}$ are only different in the i^{th} component, we obtain for the residual $\mathbf{r}^{(k+1)}$:

$$\begin{aligned} \mathbf{r}^{(k+1)} &= \mathbf{b} - \mathbf{M}\mathbf{B}^{(k+1)} && (10.130) \\ &= \mathbf{r}^{(k)} - \mathbf{M}(\mathbf{B}^{(k+1)} - \mathbf{B}^{(k)}) && (10.131) \\ &= \mathbf{r}^{(k)} - \mathbf{M} \left(0, \ldots, 0, \frac{r_i^{(k)}}{m_{ii}}, 0, \ldots, 0 \right)^T, && (10.132) \end{aligned}$$

that is , only one column of matrix $\mathbf{M}$ is needed to update all residual values.

In summary, we can say: Using the starting vector $\mathbf{B}_e = \mathbf{0}$ and based on an iterate $\mathbf{B}^{(k)}$ as well its associated residual $\mathbf{r}^{(k)}$, the Southwell relaxation identifies the index i of the residual component with the greatest absolute value. The new iterate $B_i^{(k+1)}$ is then computed due to Equation (10.128), and all residual components are updated using Equation (10.132). As during each iteration, only a single column of the matrix $\mathbf{M}$ is needed even only one row of $\mathbf{M}$ has to be computed., see Figure 10.12.

Let us now show why the Southwell relaxation scheme converges for all possible initial vectors.

LEMMA 10.4 *Given be the classical discrete radiosity from Definition 10.2, where* $\mathbf{M}$ *is an invertible* $n \times n$*-matrix, with the property of diagonal dominance. Then, the Southwell relaxation converge for every starting vector* $\mathbf{B}_e$ *towards the exact solution of Equation (10.25).*

Diagonal Dominance (156)

PROOF 10.4 *In order to prove that Southwell relaxation converges for a diagonal dominant matrix, it suffices to establish*

$$\lim_{k \to \infty} \|\mathbf{r}^{(k)}\| = 0, \tag{10.133}$$

where $\|\cdot\|$ *is the* 1*-norm from Definition A.20.*

SOUTHWELL RELAXATION {

$\forall\ B_i^{(0)} \in \mathbf{B}^{(0)}, r_i^{(0)} \in \mathbf{r}^{(0)}$ do {

$B_i^{(0)} = 0$

$r_i^{(0)} = B_{e_i}$

}

while (not converged) {

pick i, with $\max_{1 \le i \le n} \left| r_i^{(k)} \right|$

$B_i^{(k+1)} = B_i^{(k)} + \frac{r_i^{(k)}}{m_{ii}}$

$\forall\ r_j^{(k+1)} \in \mathbf{r}^{(k+1)}$ do {

$r_j^{(k+1)} = r_j^{(k)} - \frac{m_{ji} r_i^{(k)}}{m_{ii}}$

}

}

FIGURE 10.12: SOUTHWELL RELAXATION.

Expressing Equation (10.132) in terms of the j^{th} *component of the residual and using the fact, that for the relaxed component of the residual it must hold* $r_i^{(k+1)} = 0$*, then it holds for the* 1*-norm of the residual:*

$$\left\| \mathbf{r}^{(k+1)} \right\| = \sum_{\substack{j=1 \\ i \neq j}}^{n} \left| r_i^{(k)} - \frac{m_{ij}}{m_{jj}} r_j^{(k)} \right| \tag{10.134}$$

Therefore

$$\begin{aligned} \|\mathbf{r}^{(k+1)}\| &\le \sum_{\substack{j=1 \\ i \neq j}}^{n} \left| r_i^{(k)} \right| + \sum_{\substack{j=1 \\ i \neq j}}^{n} \left| \frac{m_{ij}}{m_{jj}} r_j^{(k)} \right| && (10.135) \\ &\le \|\mathbf{r}^{(k)}\| - \left| r_j^{(k)} \right| + \left| r_j^{(k)} \right| \sum_{\substack{j=1 \\ i \neq j}}^{n} \left| \frac{m_{ij}}{m_{jj}} \right|. && (10.136) \end{aligned}$$

Now, due to the strong column sum criterion there exists a scalar value t *exists such that it holds:* Column sum Criterion (160)

$$0 < t < 1 \qquad \textit{and} \qquad \sum_{\substack{i=1 \\ i \neq j}}^{n} \left| \frac{m_{ij}}{m_{jj}} \right| < t \qquad \textit{for } 1 \le j \le n. \tag{10.137}$$

Using this result in Equation (10.136) leads to

$$\|\mathbf{r}^{(k+1)}\| \leq \|\mathbf{r}^{(k)}\| - (1-t)\left|r_j^{(k)}\right|. \tag{10.138}$$

If we assume that the j^{th} component $r_j^{(k)}$ is the largest component of the residual vector, then it holds

$$\|\mathbf{r}^{(k)}\| \leq n \cdot \left|r_j^{(k)}\right|. \tag{10.139}$$

Multiplying both sides by $(1-t)$ and dividing by n yields

$$(1-t)\left|r_j^{(k)}\right| \geq \frac{1-t}{n}\|\mathbf{r}^{(k)}\|. \tag{10.140}$$

We can now substitute Equation (10.140) into Equation (10.136), which leads to

$$\begin{aligned}
\|\mathbf{r}^{(k+1)}\| &\leq \|\mathbf{r}^{(k)}\| - (1-t)\left|r_j^{(k)}\right| && (10.141)\\
&\leq \|\mathbf{r}^{(k)}\| - \frac{1-t}{n}\|\mathbf{r}^{(k)}\| && (10.142)\\
&\leq \left(1 - \frac{1-t}{n}\right)\|\mathbf{r}^{(k)}\|. && (10.143)
\end{aligned}$$

Choosing

$$T = 1 - \frac{1-t}{n}, \tag{10.144}$$

then we get with obviously

$$\begin{aligned}
\|\mathbf{r}^{(k+1)}\| &\leq T\|\mathbf{r}^{(k)}\| && (10.145)\\
&\leq T^{k+1}\|\mathbf{r}^{(0)}\|. && (10.146)
\end{aligned}$$

Since $T < 1$ the above inequality implies $\lim_{k\to\infty} T^{k+1}\|\mathbf{r}^{(0)}\| = 0$, which is equivalent to $\lim_{k\to\infty}\|\mathbf{r}^{(0)}\| = 0$. But this means, that the sequence $(\mathbf{B}^{(k)})_{k\in\mathbb{N}_0}$ converges to a solution of the system $\mathbf{MB} = \mathbf{B}_e$. As the proof is independent from the choice of the starting vector $\mathbf{B}_e$, we conclude that the Southwell relaxation converges for all starting vectors.

PHYSICAL INTERPRETATION OF THE SOUTHWELL RELAXATION. From our discussion of the form factor, we know that the coefficient F_{ij} of the radiosity matrix $\mathbf{M}$ represents the proportion of the total power leaving patch P_i that is received by patch P_j. Furthermore, we know from our derivation of the Southwell relaxation that the patch emittances can physically be interpreted as the first guess for the patch radiosities. Due to its Definition,

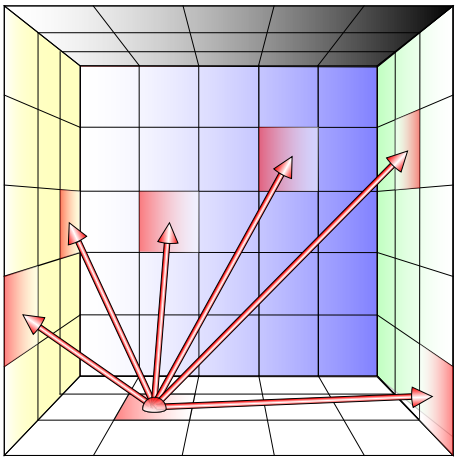


FIGURE 10.13: A SHOOTING STEP OF THE SOUTHWELL RELAXATION. ...

the residual can then be considered as a measure for the difference between the emittance of a patch and its reflected radiosity. This implies the idea, that the energy of a patch can be partitioned into an *amount of shot*, or *already distributed* and an *amount of unshot or undistributed energy*. The shot radiosity is the power per unit area we can see if we look at a patch, and the unshot radiosity is the radiosity of the surface patch, that this element has already received by other patches, but that it has not yet forwarded. In a Southwell relaxation step we then look for the patch with the largest undistributed radiosity and sends this radiosity into the scene until the solution is not convered. Compared with Gauss-Seidel item, where the radiosity of a surface patch is determined by gathering radiosities from other patches, a Southwell relaxation step can be interpreted as a *shooting step*, since energy from a patch is shot towards all other patches, see Figure 10.13.

10.2 THE FINITE ELEMENT RADIOSITY APPROACH

Recall, the classical discrete radiosity equation, as an approximate of the radiosity integral equation within a vacuum, is based on strict assumptions. Apart from the assumption, that all existing surfaces in the environment are Lamberitan, in particular the radiosity value was assumed to take a constant value across each patch of the scene. Since solutions to linear Fredholm integral equations of the 2^{nd} kind live in infinite-dimensional function spaces, a solution of the discrete radiosity system—as an element of an n-dimensional function space spanned by a set of constant basis functions—corresponds only to a coarse solution of the radiosity integral equation over the object surfaces, $\partial\mathcal{V}$, within a scene.

Classical Discrete REQ (784)
Radiosity Integral Equation (782)
Lambertian (349)
Fredholm Integral Equation (127)
Function Space (28)
$\partial\mathcal{V}$ (41)

Recall furthermore, constant radiosity meant that the continuous radiosity function in Equation (10.1) was replaced by a linear combination of step functions with constant values across the patches. Obviously, due to this trivial replacement, many information about the radiosity across a patch is lost. This means that the error between the exact solution, defined on the whole function space $\mathcal{L}^2(\partial\mathcal{V}, \mu^2)$, and an approximate, only valid across n disjoint patches, can be very high. A better approach would be to approximate the continuous radiosity function via a set of basis functions, $\{\phi_1, \ldots, \phi_n\}$, that are closer to the radiosity distribution across a patch, such as linear, quadratic, or other polynomial functions. The finite element approach, as introduced in Section 2.3.3.2.3, does in this regard useful services.

$\mathcal{L}^2(\cdot,\cdot)$ (107) Basis (857)

For the following discussion let $\mathcal{L}^2(\partial\mathcal{V}, \mu^2)$ be the space of all square Lebesgue-integrable functions defined over the set of all surfaces $\partial\mathcal{V}$ of a given scene. Furthermore, $(\partial\mathcal{V}, \mathcal{P}_m, \mathcal{N}_n)$ be a finite element, where the domain $\partial\mathcal{V}$ is partitioned into a finite mesh $\bigcup_{i=1}^{m} \overline{P}_i$ of disjoint surface patches, $\mathcal{N}_n = \{s_1, \ldots, s_n\}$ is the set of nodal points located at the boundaries of the finite mesh $\bigcup_{i=1}^{m} \overline{P}_i$, and $\mathcal{P}_m = \{N_1, \ldots, N_n\}$ be a finite set of piecewise polynomial basis function of degree m that have small supports, in that they are nonzero only in a small region, for details see the construction of the basis function N_i in Section 2.3.3.2.3.

$\mathcal{L}^2(\cdot,\cdot)$ (107) Finite Element (150) Mesh (147) Nodal Points (147)

The basis functions $N_1, \ldots, N_n$ then span a subspace $\mathcal{U}_n \leq \mathcal{L}^2(\partial\mathcal{V}, \mu^2)$. Equipped with the inner product

Subspace (855) $\langle\cdot,\cdot\rangle$ (860)

$$\langle f, g \rangle \stackrel{def}{=} \int_{\partial\mathcal{V}} f(s_i) g(s_i) \, d\mu^2(s_i), \tag{10.147}$$

we then could measure the quality of an approximate solution, $B_n(s_i)$, with respect to the error

$$\epsilon(s_i) \stackrel{def}{=} \|\mathbf{B}(s_i) - B_n(s_i)\|, \tag{10.148}$$

where the norm $\|\cdot\|$ arises from the inner product defined in Equation (10.147). Here, the concept of the residual, as known from our discussion about the convergence behavior of iterative solvers for linear systems is helpful.

$\|\cdot\|$ (860) Residual (144)

Obviously, the radiosity approximate $B_n \in \mathcal{U}_n$ can be written as a linear combination of the basis functions N_j, that is,

$$B_n(s_i) = \sum_{j=1}^{n} B_j N_j(s_i). \tag{10.149}$$

We then define the residual function r that should us give information about the quality of this approximate, by:

$$r(s_i) \stackrel{def}{=} B_n(s_i) - B_e(s_i) - \int_{\partial\mathcal{V}} B_n(s_j) \, \mathcal{G}'(s_j \to s_i) \, d\mu^2(s_j). \tag{10.150}$$

$\mathcal{L}^2(\cdot,\cdot)$ (107)

As the emitted radiosity function B_e is an element of the function space $\mathcal{L}^2\left(\partial\mathcal{V}, \mu^2\right)$, the residual function r can not be of finite dimension. Due to its definition, r is generally not identically zero. Obviously, the reason for that is, that the approximate B_n, as a finite dimensional function, does not satisfy the radiosity integral equation, thus,

$$B_n(\mathbf{s}_i) \neq B_e(\mathbf{s}_i) - \int_{\partial\mathcal{V}} B_n(\mathbf{s}_j)\, \mathcal{G}'(\mathbf{s}_j \to \mathbf{s}_i)\, d\mu^2(\mathbf{s}_j). \tag{10.151}$$

Only in the case, where $B_n = B$, the residual is zero. That is, to find a good approximation B_n the coefficients of B_n should be chosen, that the residual will be small. This implies that the residual should be *minimized*,

$$\min_{B_n \in \mathcal{U}_n} \left\| B_e(\mathbf{s}_i) - \int_{\partial\mathcal{V}} B_n(\mathbf{s}_j)\, \mathcal{G}'(\mathbf{s}_j \to \mathbf{s}_i)\, d\mu^2(\mathbf{s}_j) \right\| \tag{10.152}$$

or expressed in other words, we search a function B_n in $\mathcal{U}_n$ that is closest to the exact solution.

Now, instead to find an approximate B_n in $\mathcal{U}_n$ that makes the residual small, we can also go in the other direction, namely, projecting the residual from $\mathcal{L}^2(\partial\mathcal{V}, \mu^2)$ into the subspace $\mathcal{U}_n$ spanned by the set of basis function $N_1, \ldots, N_n$. This approach then results in the Galerkin method, introduced in Section 2.3.3.2.2.

A FINITE ELEMENT RADIOSITY APPROACH BASED ON THE GALERKIN METHOD. Recall at Equation (2.471), the Galerkin approach for solving a linear Fredholm integral operator equation results in n equations of the type

$$\sum_{j=1}^{n} B_j \underbrace{\langle (\mathbf{I} - \mathbf{K}) N_j(\mathbf{s}_i), N_i(\mathbf{s}_i) \rangle}_{m_{ij}} = \langle g(\mathbf{s}_i), N_i(\mathbf{s}_i) \rangle, \qquad 1 \leq i \leq n, \tag{10.153}$$

where $\mathbf{I}$ is the identity-operator and $\mathbf{K}$ corresponds to the integral operator in the Fredholm type equation. Now, this equation can also be written in matrix-vector notation as

Integral Operator (130)

$$\mathbf{MB} = \mathbf{b}, \tag{10.154}$$

where $\mathbf{B} = (B_1, \ldots, B_n)^T$ is the vector of unknowns of the system, the right-hand side, $\mathbf{b} = (\langle B_e, N_1 \rangle, \ldots, \langle B_e, N_n \rangle)^T$, is the n-dimensional vector, whose i^{th} component corresponds to the inner product of the emitted radiosity B_e and the i^{th} basis function N_i thus:

$$b_i = \langle B_e(\mathbf{s}_i), N_i(\mathbf{s}_i) \rangle = \int_{\partial\mathcal{V}} B_e(\mathbf{s}_i) N_i(\mathbf{s}_i)\, d\mu^2(\mathbf{s}_i), \tag{10.155}$$

and the coefficients $\left(m_{ij}\right)_{1\leq i,j\leq n}$ are given by:

$$\begin{aligned}
m_{ij} &= \langle (\mathbf{I}-\mathbf{K})N_j(s_i), N_i(s_i)\rangle && (10.156)\\
&= \langle N_i(s_i), (\mathbf{I}-\mathbf{K})N_j(s_i)\rangle && (10.157)\\
&= \langle N_i(s_i), N_j(s_i)\rangle - \langle N_i(s_i), (\mathbf{K}N_j)(s_i)\rangle && (10.158)\\
&= \langle N_i(s_i), N_j(s_i)\rangle - \left\langle N_i(s_i), \rho_{dh}(s_i)\int_{\partial\mathcal{V}} N_j(s_j)\,\mathcal{G}'(s_j\leftrightarrow s_i)d\mu^2(s_j)\right\rangle && (10.159)\\
&\overset{(10.147)}{=} \int_{\partial\mathcal{V}} N_i(s_i), N_j(s_i)d\mu^2(s_i) - \\
&\quad \int_{\partial\mathcal{V}} N_i(s_i)\,\rho_{dh}(s_i)\left(\int_{\partial\mathcal{V}} N_j(s_j)\,\mathcal{G}'(s_j\leftrightarrow s_i)d\mu^2(s_j)\right)d\mu^2(s_i), && (10.160)
\end{aligned}$$

Inner Product (859)

where we have used the symmetry and the linearity property of the inner product in the second and the third step of the derivation.

Obviously, the finite element radiosity approach, based on the Galerkin method, attempts to find an approximate $B_n \in \mathcal{U}_n$ that leads, due to Equation (10.153), to a residual function r that is in average zero over the patches of the domain $\partial\mathcal{V}$.

A FINITE ELEMENT RADIOSITY APPROACH BASED ON THE COLLOCATION METHOD. Apart from the Galerkin method, in Section 2.3.3.2.2 we have also presented the collocation method as an approximative solver for Fredholm integral equations of the 2^{nd} kind. The associated linear system, adapted to the radiosity integral equation, was given by:

$$B_e(s_i) = \sum_{j=1}^{n} B_j \underbrace{(\mathbf{I}-\mathbf{K})N_j(s_i)}_{m_{ij}}, \qquad 1\leq j\leq n, \tag{10.161}$$

Integral Operator (130)

where $\mathbf{I}$ is the identity-operator and $\mathbf{K}$ corresponds to the integral operator in the radiosity equation.

Rewritten in matrix-vector notation, we get

$$\mathbf{AB} = \mathbf{b}, \tag{10.162}$$

where $\mathbf{B} = (B_1,\ldots,B_n)^T$ is the vector of unknowns of the system, $\mathbf{b} = (B_e(s_1),\ldots,B_e(s_n))^T$ is the n-dimensional vector, whose i^{th} component corresponds to the the emitted radiosity B_e at collocation point s_i, and the coefficients $\left(m_{ij}\right)_{1\leq i,j\leq n}$ are given by:

$$\begin{aligned}
m_{ij} &= (\mathbf{I}-\mathbf{K})N_j(s_i) && (10.163)\\
&= N_j(s_i) - (\mathbf{K}N_j)(s_i) && (10.164)\\
&= N_j(s_i) - \rho_{dh}(s_i)\int_{\partial\mathcal{V}} N_j(s_j)\,\mathcal{G}'(s_j\leftrightarrow s_i)d\mu^2(s_j). && (10.165)
\end{aligned}$$

Obviously, the finite element radiosity approach, based on the collocation method, attempts to find an approximate $B_n \in \mathcal{U}_n$ that leads, due to Equation (10.161), to a

residual, function r that is in exactly zero only at the nodal points s_i. Compared with the Galerkin approach, where the residual function is in average zero over all patches, we could expect, that the approximate B_n obtained via the collocation approach is not so closed to the exact solution B like an approximate obtained via the Galerkin approach. We can say, that the Galerkin method *extracts* more information from the kernel, since it approximates the kernel via an integral, where the collocation method evaluates the kernel only at the collocation points. However, we pay for a better approximate from the Galerkin method also with more effort for the needed calculation.

REMARK 10.11 (The Classical Radiosity Approach) *Let us consider once more the finite element approach based on the Galerkin method, where the set of basis functions* $\{N_1, \ldots, N_n\}$ *are box functions, defined by:*

$$N_j(s) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } s_i \in P_i \\ 0 & \text{if } s_i \notin P_i. \end{cases} \tag{10.166}$$

Due to Equation (10.160) the coefficients $(a_{ij})_{1 \le i,j \le n}$ *are given by:*

$$\begin{aligned} m_{ij} &= \int_{\partial \mathcal{V}} N_i(s_i) N_j(s_i) d\mu^2(s_i) - & (10.167) \\ & \int_{\partial \mathcal{V}} N_i(s_i)\, \rho_{dh}(s_i) \left(\int_{\partial \mathcal{V}} N_j(s_j)\, \mathcal{G}'(s_j \leftrightarrow s_i) d\mu^2(s_j) \right) d\mu^2(s_i), & (10.168) \end{aligned}$$

that is,

$$\int_{\partial \mathcal{V}} N_i(s_i) N_j(s_i) d\mu^2(s_i) = \left\{ \begin{array}{ll} \int_{P_i} d\mu^2(s_i) & \text{if } i = j \\ 0 & \text{otherwise} \end{array} \right\} = \delta_{ij} A_i, \tag{10.169}$$

where δ_{ij} *is the Kronecker symbol and* A_i *denotes the Lebesgue area measure of the patch* P_i.

In a similar way we get for the i^{th} *component of the right-hand side of the linear system,* $\mathbf{AB} = \mathbf{b}$*:*

$$\begin{aligned} b_i &= \int_{\partial \mathcal{V}} B_e(s_i) N_i(s_i) d\mu^2(s_i) & (10.170) \\ &= \int_{P_i} B_e(s_i) d\mu^2(s_i) = B_{ei} A_i, & (10.171) \end{aligned}$$

where B_{ei} *is the area average emission for patch* P_i.

Last but not least, for the second term in Equation (10.160) it holds:

$$\begin{aligned} & \int_{\partial \mathcal{V}} N_i(s_i)\, \rho_{dh}(s_i) \left(\int_{\partial \mathcal{V}} N_j(s_j)\, \mathcal{G}'(s_j \leftrightarrow s_i) d\mu^2(s_j) \right) d\mu^2(s_i) & (10.172) \\ &= \rho_i \int_{P_i} \left(\int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i) d\mu^2(s_j) \right) d\mu^2(s_i), & (10.173) \end{aligned}$$

where we have assumed that the reflectivity over the patches is constant.

Putting all these things together, then we get:

$$\sum_{j=1}^{n} B_j \left(\delta_{ij} A_i - \rho_i \int_{P_i} \left(\int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i) d\mu^2(s_j) \right) d\mu^2(s_i) \right) = B_{ei} A_i \tag{10.174}$$

and after dividing both sides by the area of patch P_i:

$$\sum_{j=1}^{n} B_j \left(\delta_{ij} - \rho_i \frac{1}{A_i} \int_{P_i} \left(\int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i) d\mu^2(s_j) \right) d\mu^2(s_i) \right) = B_{ei}. \tag{10.175}$$

Using the formula for the classical form factor from Definition (10.2), then leads to:

$$\sum_{j=1}^{n} B_j \left(\delta_{ij} - \rho_i F_{ij} \right) = B_{ei}, \tag{10.176}$$

which is equivalent to Equation (10.25), namely:

$$B_i = B_{ei} + \rho_i \sum_{j=1}^{n} B_j F_{ij}. \tag{10.177}$$

The above derivation shows, that the classical radiosity approach is a special variant of the finite element approach based on the Galerkin method, where the set of basis function are box functions defined on the patches P_i, all patches are assumed to be Lambertian, and the source term B_e emits a constant radiosity value.

REMARK 10.12 (A Finite Element Radiosity Approach Based on the Method of Weighted Residual) *As we have seen in the previous remark, the classical radiosity approach is a special case of the finite element approach based on the Galerkin method. But also the Galerkin method is a variant of a more general approach: The method of* weighted residuals, *also called the* MWR-method.

The idea behind the method of weighted residual is to force the residual to zero in some average sense over the integration domain using n appropriate weighting functions w_i, that is, solving the equation

$$\langle w_i(s_i), r(s_i) \rangle = \int_{\partial \mathcal{V}} w_i(s_i) r(s_i) d\mu^2(s_i) = 0, \qquad 1 \le i \le n. \tag{10.178}$$

According to the choice of the weighting functions, there are at least four MWR-algorithms:

i) the collocation method,

ii) the least squares method,

iii) the method of moments, and

iv) the Galerkin method.

As a result, we get always a system of n *linear equations. The Galerkin method follows this idea. It selects the basis functions, used to approximate the radiosity function, as weighting functions. With respect to the above formula, the Galerkin method then leads to a system of linear equations of type*

$$\langle N_i(s_i), r(s_i)\rangle = \int_{\partial V} N_i(s_i) r(s_i) d\mu^2(s_i) = 0, \qquad 1 \le i \le n. \tag{10.179}$$

THE FORM FACTOR IN THE FINITE ELEMENT RADIOSITY APPROACH. Before we conclude this section, let us talk about the role of the form factor in the finite element radiosity approach. The form factor in the finite element radiosity approach is defined as:

$$\int_{P_i} N_i(s_i) \left(\int_{P_j} N_j(s_j)\, \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) \right) d\mu^2(s_i). \tag{10.180}$$

In the case of constant basis functions N_i we can assume that N_i is equal to 1 only over the patch P_i and otherwise zero, that is, the general form factor corresponds to:

$$\int_{P_i} \left(\int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) \right) d\mu^2(s_i) \tag{10.181}$$

$$= \quad A_i \frac{1}{A_i} \int_{P_i} \left(\int_{P_j} \mathcal{G}'(s_j \leftrightarrow s_i)\, d\mu^2(s_j) \right) d\mu^2(s_i) \tag{10.182}$$

$$\overset{(10.26)}{=} \quad A_i F_{ij}, \tag{10.183}$$

which can also be written as $A_i F_{ij}$.

As we have see in Section 10.1.3, the classical form factor F_{ij} represents the fraction of energy leaving patch P_i that directly arrives at patch P_j. In the generalized case, where the basis functions can also be chosen as linear, quadratic, or polynomial functions, the form factor F_{ij} represents the weighted effect of energy leaving patch P_i under the support of one basis function on the energy of patch P_j of another basis function. That is, the form factor makes a statement on the strength of the coupling between the two associated basis function N_i and N_j. It should be clear, that the general form factor—due to the presence of the basis functions within the integrand—is more difficult to evaluate than its classical analogue. Note: The general form factor can not be divided by the area term, as we did it with the classical form factor.

10.3 THE RADIOSITY PIPELINE

In the previous sections, we have presented the radiosity method as it applies to computer graphics. Thus, we have derived the discrete radiosity equation from the continuous stationary light transport equation in a vacuum, we have talked about methods for computing the coefficients of the radiosity matrix, and we have shown how the radiosity equation system can easily and efficiently be solved via procedures from numerical mathematics. As result, we have get an n-dimensional vector of patch radiosities, but for image synthesis applications we are interested in generating displayable images. For that purpose, let us now shortly talk, how this radiosity approach can be used to produce simulations and images.

A typical simple radiosity algorithm for image synthesis can be modeled via the three-stage *radiosity pipeline* from Figure 10.14, that is,

I) *building the radiosity matrix*

II) *solving the radiosity matrix*, and

III) *visualization*.

The radiosity pipeline excepts as input the geometry and the physics of the scene to be rendered, that is, geometric information about the objects within a scene and the physical properties of the materials used.

INPUT FOR THE RADIOSITY PIPELINE. As shown in Section 10.1.3, form factor calculation is build on the discretization of the scene to be rendered. That is, in a preprocessing step the scene must be modeled—via a finite element mesh consisting of a disjoint partition of the object surfaces—into small patches. Then a nodal point $\mathbf{s}_i$ has to be selected on each patch, commonly the midpoint of the patch, since scenes are mostly discretized in rectangular meshes. Later, these will become the unknowns B_i of the radiosity equation system, $\mathbf{MB} = \mathbf{B}_e$, where discrete radiosity values are stored. Furthermore, we have to decide how radiosity changes across a surface patch, that is, we have to choose the basis functions $N_i, 1 \leq i \leq n$ representing the radiosity.

BUILDING THE FORM FACTOR MATRIX. The representation of the coefficients of the radiosity by $m_{ij} = I - \rho_i F_{ij}, 1 \leq i, j \leq n$ suggests firstly to compute the form factors F_{ij} based on the geometrical information supplied in the scene description. This can be done using the methods introduced in Section 10.1.3.

BUILDING THE RADIOSITY MATRIX. Based on the form factor matrix $\mathbf{F}$ and the identity matrix, $\mathbf{I}$, the radiosity matrix $\mathbf{M}$,

$$\mathbf{M} = \mathbf{I} - \boldsymbol{\rho}\mathbf{F}, \tag{10.184}$$

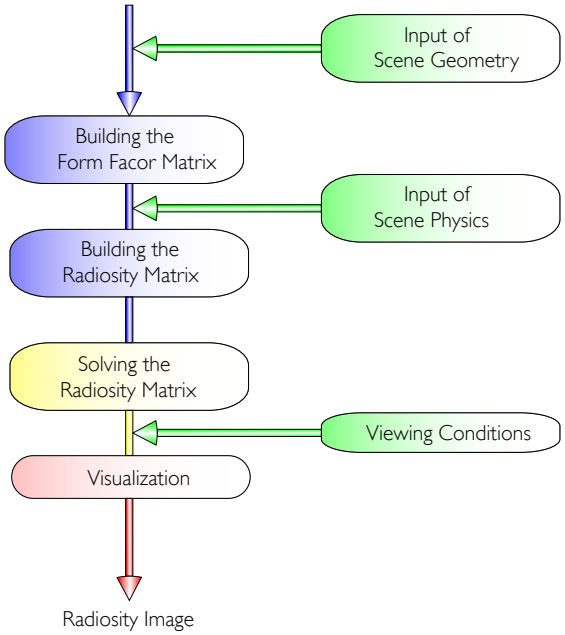


FIGURE 10.14: THE RADIOSITY PIPELINE. From the model of the radiosity pipeline, it is easily recognizable, that after the forma factor matrix is build, the physic of the scene—the reflectivity's of the surface patches, the existence of the light sources, and the characteristic of a patch, light source of not light source—can be changed, without to recompute the form factor matrix. The same holds for the rendering step, if the radiosity matrix is solved, the viewing conditions of the scene can be easily changed without to redo all previous computations.

has to built with the help of the physical properties of the surfaces, given by the diagonal matrix $\boldsymbol{\rho}$.

SOLVING THE RADIOSITY MATRIX. After the radiosity matrix $\mathbf{M}$ is build, we can set up the linear system

$$\mathbf{MB} = \mathbf{B}_e \tag{10.185}$$

with the help of the exitances $\mathbf{B}_e$ from the input of the radiosity pipeline. Then, the system is solved via one of the iteration methods from Section 10.1.4. As result, we obtain the radiosity vector $\mathbf{B}$ of unknown nodal radiosities. Via these nodal radiosities and the chosen basis functions, $N_i, 1 \leq i \leq n$, then a functional form for the variation of radiosity across a patch is derived.

VISUALIZATION. In the final stage of the radiosity pipeline, then a rendering step has to be performed using the functional form for the variation of radiosity across a patch. Another technique could be to use a variety of shaders, such as a flat shader, which shade each patch with the nodal radiosity associated to the patch. As, rendering via flat shading results in low quality images, we can also use continuous shaders across the patches for rendering. Thus, simple bilinear interpolation such as *Gouraud shading* delivers more better results. For this, the nodal radiosity of neighboring patches has to be distributed on the vertices of the mesh, which can then be interpolated across the area of the patches.

REMARK 10.13 *It should be clear, that the radiosity pipeline, as introduced above, serves only as a model for the structure of a radiosity algorithm. Thus, the basic structure of a radiosity algorithm can be diversely extended, see [36, Cohen & Wallace 1993].*

10.4 RAY TRACING VS RADIOSITY

As we have seen in the last three chapters, there are two different approaches for solving the global illumination problem: Approaches that are based on

- Markov processes, and
- finite element based algorithms.

Both approaches solve the global illumination problem, but each of these methods itself has problems that make it unsuitable for rendering all kinds of illumination effects. Let us shortly summarize the most important properties of these both rendering techniques:

MARKOV PROCESS BASED RENDERING ALGORITHMS. As global illumination algorithms, Markov process based rendering algorithms can simulate all possible light effects within a scene more or less well. Thus, they can handle arbitrary geometries and BSDFs, but often provide still noisy images even if they use many thousands of primary ray, where this noise is perceived by the human eye as very disturbing. In particular the simulation of indirect illumination and the illumination from multiple diffuse reflections within a Monte Carlo rendering algorithm is very expensive and prohibitively costly. As, Markov process based rendering algorithms solve the stationary light transport equation only for those points that are visible in an image, changing the view position or orientation by more than a small amount usually requires repeating the entire ray-tracing process from scratch.

FINITE ELEMENT BASED RENDERING ALGORITHMS. The radiosity method is based on a finite element approach for solving the global illumination problem. In its classical version,

it simulates only the diffuse propagation of light through a scene. As the light-transfer calculations are based solely on the geometry of the environment, radiosity procedures are view-independent algorithms, which makes them suitable for simulating walk-throughs through complex scene models. This method requires a meshing of the scene, where the resolution of the mesh determines the precision of the incident illumination and propagated light. Radiosity algorithms produce noise free images. For the accurate treatment of complex geometries, or even complex BSDFs these methods, however, bring a very high computational and storage effort. They also tend to visible artifacts in the display.

REMARK 10.14 *As mentioned above, each of the previously presented methods has problems that make it unsuitable for practical use. Therefore, in practice combinations of these methods was used for a time, with the goal to exploit their strengths and to skip their weaknesses. As the problems of radiosity algorithms—run time and storage requirements—can not be avoided, today almost combinations of Monte Carlo ray tracing methods and photon mapping approaches are used. These hybrid methods offer all the advantages of Markov process based ray tracing methods and photon mapping. At the same time, they avoid the artifacts of the photon mapping, and the noise of the Monte Carlo ray tracing method.*

10.5 REFERENCE LITERATURE AND FURTHER READING

Compared with ray tracing, the number of literature sources that deals with radiosity is comprehensible. There are three excellent books on radiosity [36, Cohen & Wallace 1993], [13, Ashdown 1994], and [190, Sillion & Puech 1994]. Starting with radiometric quantities, over the derivation of the radiosity equation until to algorithms and improvement strategies for solving the radiositiy equation, they leave no question on radiosity unanswered. While the books by Cohen and Wallace as well as Sillion and Puech are rather mathematical organized, Ashdown's book is a step-by-step guidance for the development of the fully functional, radiosity renderer HELIOS for Microsoft WINDOWS. He shows how it is possible to understand the basic, classical radiosity method, only with knowledge of vectors and matrices from a basic course on linear algebra. Hence, we recommend [13, Ashdown 1994] to the reader that is interested in a coarse overview of the radiosity method. For people who want to delve deeper into the topic, [36, Cohen & Wallace 1993] and [190, Sillion & Puech 1994] are the right references.

Chapter 11 of our book is mainly built on these three sources, where we have attempt to emphasize the mathematical character of the radiosity method. Thus, when deriving the finite element approach of the radiosity method, we have exactly formulated the underlying mathematical framework, such as the concept of the finite element, the required function

spaces, and the numerical approaches for solving Fredholm integral equations of the 2^{nd} kind.

An excellent modern survey of radiosity is offered in [68, Glassner 1995]. For a short introduction into the topic, we recommend also [55, Encarnacao & al. 1997]. The tutorial [211, Talbot 1999] is a short review of [36, Cohen & Wallace 1993]. A very nice introduction to the radiosity methods may also be found in [192, Slusallek & al. 1993]. [193, Slusallek & al. 1993] also discusses radiosity and relaxation methods and places them in the context of the literature on solving linear systems of equations. There is also a series of beautiful textbooks on computer graphics, which discuss the classical radiosity algorithm shortly and on a very heigh level. Here we allude [62, Foley & al. 1987], [233, Watt & Watt 1992], [78, Hearn & Baker 1994] and [232, Watt 1999].

Under the aspect that form factors can be interpreted as probabilities, in [50, Dutré & al. 2003], the radiosity problem is discussed with the goal to achieve algorithms that solve the radiosity equation using stochastic sampling. In the same direction goes [21, Bekaert & al. 1998], [20, Bekaert 1999], where bridges are build between hierarchical radiosity and Monte Carlo radiosity.

From the multiplicity of Master and PhD thesises that deals with radiosity methods let us mentioned only a few of them. First, [82, Heckbert 1991] and [34, Christensen 1995]. Heckbert shows that radiosity is a finite element method for solving Fredholm integral equations and Christensen deals with hierarchical techniques for efficient solution of the glossy global illumination problem using radiosity methods. For an alternative Galerkin radiosity formulation based on piecewise smooth illumination functions that incorporates curved surfaces directly, see [240, Zatz 1992]. In [64, Gibson 1995] is discussed how very complex radiosity solutions can be computed quickly and efficiently. Radiosity methods was also extended to the light transport in participating media. So, in [177, Schirmacher 1996], the radiosity method was applied to the volume radiosity equation, and in [89, Hubeli & al. 1999], the concept of the global cube, as a hardware-accelerated hierarchical volume radiosity technique is presented. The concept of instant radiosity was firstly introduced in [103, Keller 1997] and in [207, Suykens 2002], radiosity is combined with bidirectional path tracing.

CHAPTER ELEVEN

APPENDIX

A SIMPLE USEFUL MATHEMATICAL CONCEPTS FROM LINEAR ALGEBRA AND CALCULUS

The present section can be interpreted as a refresher to useful concepts from linear algebra and calculus which are necessarily needed to understand the mathematical foundations of realistic rendering. Students familiar with basic calculus and linear algebra can skip this section since it can be considered as an introductory preparation for the subsequent sections.

We will start with basics from calculus, that is, sets, relations, and operators and introduce with the Euclidean space $\mathbb{R}^3$ a first simple example of a linear space of finite dimension, well-known form linear algebra. This space is ideal for a descriptive explanation of the most important concepts and constructs from linear algebra. Based on the knowledges about sets and linear algebra, we then penetrate a little deeper into the theory of abstract linear spaces and afterwards we address ourselves to the field of differential calculus. We will finish this section with first, few introductory insights into integration theory and the stochastically method of Monte Carlo integration, which we will discuss in more detail in Chapter 2 and Chapter 6.

Section A.1
Section A.2
Section A.3
Section A.4
Section A.5

A.1 SETS AND FUNCTIONS

Commonly, *Mathematics* takes as its starting point the idea of the existence of collections of mathematical objects, such as, *numbers*, *vectors*, and *functions*, also known as *sets*. Such collections are typically endowed with additional algebraic structures. When this is done, it becomes possible to elaborate with their properties and to build up a coherent theory.

As *sets* are clearly basic to a proper study of mathematics, we start our study of mathematical concepts from linear algebra and calculus with some introductory aspects of set theory. After introducing the different sets of *numbers*, we discuss the mathematical concepts of the *algebra* and the σ-*algebra*, which serve as the fundamental set theoretical constructs for generating abstract measures in *integration* and *probability theory*. Then, we speak about *Cartesian products* of sets and introduce the Euclidean spaces $\mathbb{R}, \mathbb{R}^2$ and $\mathbb{R}^3$ as first simple examples of Cartesian products. Since the domains of integral equations in global illumination are generally the unit sphere, the hemisphere or subsets of these, we introduce *polar* and *spherical coordinates* as alternative representations of points in $\mathbb{R}^2$ and $\mathbb{R}^3$. After this, we discuss the concept of the *relation*, repeat the well-known definition of *real* and *complex valued operators*, often also better known as mappings or functions, and present some examples of useful functions which play an important role in analyzing realistic rendering methods.

Section 2.2

Section 2.3

SETS. A *set* is any well-defined collection of objects. These objects—in our context mainly numbers, vectors, and functions—are called *elements* or *members* of a set. Usually, a set is denoted by a capital letter, for example A. If the object x is a member of A, we write $x \in A$ and read as x is an element of A or x belongs to A. Likewise, the expression $x \notin A$ means, that x is not an element of set A. Furthermore, we call a set A the *empty set*, if A contains no elements.

Let us assume A and B are two sets, then we say A is a *subset* of B, if each element of A is a member of B, this is denoted as $A \subset B$. According to this definition every set is, of course, a subset of itself. We call A a *proper subset* of B, if A is indeed a subset of B and if furthermore B also contains elements, that do not belong to A. If it is desirable to indicate that a non-empty set A is a subset of B, which is possibly the set A itself, we write $A \subseteq B$. As a consequence we write $A \not\subset B$, if A is not a subset of B. Two sets A and B are *equal*, if they contain exactly the same elements, in this case we write $A = B$. According to this definition, it is clear that two sets A and B are equal, if and only if, $A \subset B$ and $B \subset A$.

REMARK A.1 *From now on, we make the assumption, that all sets under discussion are subsets of a single fixed set called the universe. The universe is denoted by $\mathcal{U}$, whereas the definition of $\mathcal{U}$ varies from one context to another.*

UNION, INTERSECTION, DIFFERENCE, AND COMPLEMENT OF SETS. In mathematics there exist operations, so-called *set operation*, that allow to construct new sets from given sets: the *union*, the *intersection*, the *difference*, and the *complement* of a set.

The *union* of two sets A and B, written as $A \cup B$, is the set consisting of all elements that are in A *or* in B, thus $A \cup B \stackrel{\text{def}}{=} \{x \,|\, x \in A \text{ or } x \in B\}$. Under the *intersection* of the sets A and B, written as $A \cap B$, we understand the set of all elements that belong to both A *and* B, that is: $A \cap B \stackrel{\text{def}}{=} \{x \,|\, x \in A \text{ and } x \in B\}$. In addition to the union and intersection

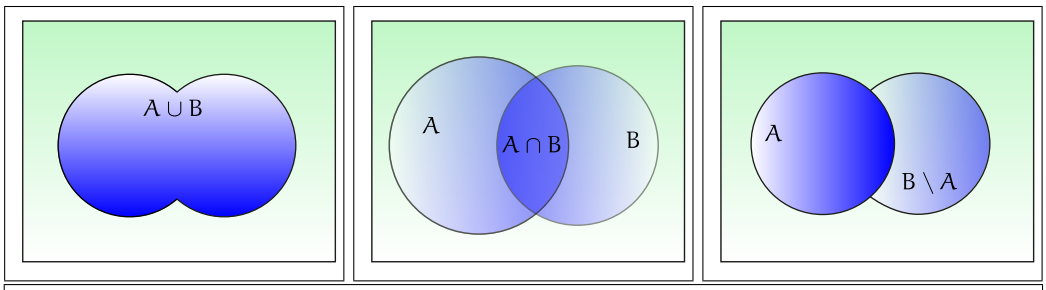


FIGURE A.1: OPERATIONS ON SETS. Visualization of union, $A \cup B$, intersection, $A \cap B$, and difference, $B \setminus A$, of two non empty sets A and B with $A, B \subset \mathcal{U}$. The universe is formed by a rectangle surrounding the sets to be represented whereas its subsets are points or areas within this rectangle.

of sets, we declare the *difference* of two sets A and B, written as $A \setminus B$, as the set of all elements of A that do not belong to B, that is: $A \setminus B \stackrel{\text{def}}{=} \{x \,|\, x \in A \text{ and } x \notin B\}$. Based on the operation of intersection, we call two non-empty sets A, B *disjoint*, if it holds: $A \cap B = \emptyset$. Finally the *complement* of a set A written as $\overline{A}$, is the set of all elements not in A, thus $\overline{A} \stackrel{\text{def}}{=} \{x \in \mathcal{U} \,|\, x \notin A\}$. All these operations on sets can visualized graphically in form of so-called *Venn diagrams*, see Figure A.1.

SETS OF NUMBERS. Now, mathematics deals with numbers, and thus, numbers are the most important types of sets. In our following discussion we make again and again use of the sets of numbers $\mathbb{N} = \{1, 2, 3, \ldots\}, \mathbb{Z}, \mathbb{Q}$, and $\mathbb{R}$, i.e. the set of *natural* or *positive integers*, *integers*, *rational*, and *real numbers*.

It is known that a rational number is a number, that can be expressed as the ratio of two integers p, q, i.e. the set of rational numbers $\mathbb{Q}$ can be written as $\left\{x \,|\, x = \frac{p}{q}, p, q \in \mathbb{Z}\right\}$ with $q \neq 0$. An important property of $\mathbb{Q}$ is, that the set of rational numbers is *countably infinite*, i.e. the positive natural numbers can be evenly matched with the rational numbers or in other words any rational numbers can be indexed by a positive integer. This implies, that $\mathbb{N}$ and $\mathbb{Q}$ have the same cardinality. The difference $\mathbb{R} \setminus \mathbb{Q}$ is called the set of *irrational numbers*. A very important irrational number is $\sqrt{2}$, which cannot be written as fraction of two integers. It is convenient to think of $\mathbb{R}$ as being represented by an infinitely long line, called the *real axis* or the *real line*, where every real number corresponds to a point on this line. It should also be known, that $\mathbb{R}$ is an *uncountable set*, that is, the real numbers can not be labeled by numbers of $\mathbb{N}$.

Section 2.2.1

Section 2.2.1

In addition to the the well-known sets of numbers introduced above, the *set of complex numbers* $\mathbb{C}$ plays an import role in many application areas of mathematics and physics. In mathematics, the set of complex numbers $\mathbb{C}$ is defined to be the set of numbers of form $z = a + ib$, with $i = \sqrt{-1}$ and $a, b \in \mathbb{R}$. Given a complex number $z = a + ib$,

Section 2.1.1

the number a is called the *real part* of z, in sign $\Re(z)$, and b is denoted as the *imaginary part* of z, written as $\Im(z)$. We define the *complex conjugate* $\bar{z}$ of z as the complex number $a - ib$. With the complex conjugate, $\bar{z}$, then it holds for a complex number z: $|z| \stackrel{def}{=} \sqrt{z \cdot \bar{z}} = \sqrt{a^2 + b^2}$.

Inner Product (859)

σ-ALGEBRAS. In probability theory, we are often interested in the outcome of a random experiment, such as flipping a coin or throwing a die. Thus, for computing the probability that a die shows a prime number, we have to determine the probability with which the subset $\{2, 3, 5\}$ of $\{1, 2, 3, 4, 5, 6\}$ occurs, that is, we have to deal with collections of subsets of a given set. With the mathematical concept of the σ-*algebra*, as an example of a collection of sets, we now introduce the perhaps most important set theoretical construct in measure theory. Without the concept of the σ-algebra, it would not be possible to construct the *measure spaces* in *integration* and *probability theory* on which we are interested to find solutions to the global illumination problem.

Section 2.2
Section 2.2.4
Section 2.4

DEFINITION A.1 (Algebra and σ-algebra) *Let $\mathcal{U}$ be the universe, a non-empty collection $\mathfrak{U}$ of subsets of $\mathcal{U}$ is called an* algebra, *if in addition to $\mathcal{U} \in \mathfrak{U}$, the relations $A, B \in \mathfrak{U}$ imply that $A \cup B \in \mathfrak{U}$ and $A \setminus B \in \mathfrak{U}$. An algebra of sets is called a* σ- algebra, *if together with an arbitrary sequence of sets $A_1, A_2, \ldots, A_n, \ldots$ it contains the union $\bigcup_{i=1}^{\infty} A_i$.*

In other words, a σ-algebra is a non-empty set, which is closed under the usual set theoretic operations of countable unions, countable intersections, and countable complements.

The most simplest σ-algebra over a given set A is the *power set*, $\mathfrak{P}(A)$. It is defined as the collection of all subset of A. Evidently, the power set of a finite set A with n elements is a σ-algebra consisting of 2^n elements.

Section 2.4

EXAMPLE A.1 *Let Ω be the set consisting of the elements 0 and 1. Evidently, it holds: $\mathfrak{P}(\Omega) = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$. It should be clear that $\mathfrak{P}(\Omega)$ is a σ-algebra, consisting of 2^2 elements. $\mathfrak{P}(\Omega)$ plays an important role in probability theory. There, it can be interpreted as the events of an experiment with the two outcomes 0 and 1, whereas $\emptyset$ is the impossible event, $\{0\}$ or $\{1\}$ are the elementary events and $\{0, 1\}$ is the certain event, see Figure A.2.*

Section 2.4

EXAMPLE A.2 *Contrary to the previous example, let us now assume that the base set Ω is countably infinite, that is, $\Omega = \{\omega_1, \omega_2, \ldots\}$. Due to Definition A.1, the power set $\mathfrak{P}(\Omega)$ is a σ-algebra consisting of countably infinite sets. That is, $\mathfrak{P}(\Omega)$ contains the empty set $\emptyset$, the set Ω itself, as well as the one-element sets $\{\omega_i\}, i \geq 1$, all sets of size 2, thus $\{\omega_i, \omega_j\}, i < j, i \geq 1$, the sets $\{\omega_i, \omega_j, \omega_k\}, i < j < k, i \geq 1$, containing three elements each other, etc..*

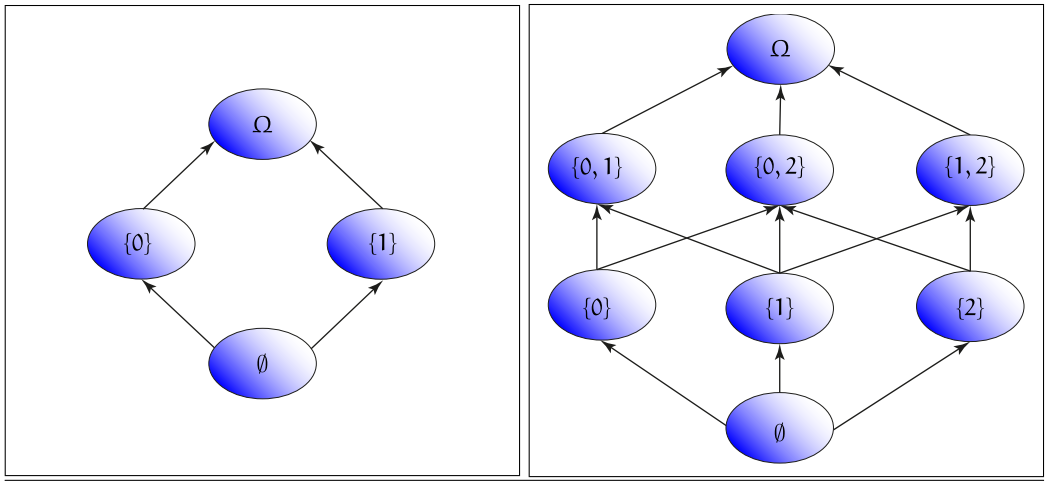


FIGURE A.2: THE POWER SET OF A FINITE SET. Let $\Omega = \{0, 1\}$. Obviously it holds: $\mathfrak{P}(\Omega) = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$, i.e. the power set $\mathfrak{P}(\Omega)$ contains the empty set, the base set Ω, and all 1-element, and 2-element subsets of Ω. The right image shows the graph of $\mathfrak{P}(\Omega)$ for the three-element set $\Omega = \{0, 1, 2\}$.

SUBSETS OF $\mathbb{R}$. Above we introduced the set of numbers $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. From all of these sets, the perhaps almost important set of numbers in mathematics is the set of the real numbers, $\mathbb{R}$. Now, very often our interest is focused not on the whole real line but only on a portion of it, called an *interval*. If $a, b \in \mathbb{R}$, such that $a \leq b$, then we define:

i) the *open interval* $(a, b) \stackrel{\text{def}}{=} \{x | x \in \mathbb{R} \quad a < x < b\}$,

ii) the *closed interval* $[a, b] \stackrel{\text{def}}{=} \{x | x \in \mathbb{R} \quad a \leq x \leq b\}$,

iii) the *half-open interval* $(a, b] \stackrel{\text{def}}{=} \{x | x \in \mathbb{R} \quad a < x \leq b\}$, and

iv) the *half-open interval* $[a, b) \stackrel{\text{def}}{=} \{x | x \in \mathbb{R} \quad a \leq x < b\}$,

see Figure A.3. In this context, *open* and *closed* means, that the endpoints a and b of an interval are included in or excluded from the set.

EXAMPLE A.3 *Let $I = [a, b), a, b \in \mathbb{R}$ be a fixed half-interval and $\mathfrak{I}$ be the collection of all half-intervals $[\alpha, \beta) \subseteq [a, b)$. Clearly, $\mathfrak{I}$ is not an σ-algebra because, generally speaking, neither the union nor the difference of two half-intervals is a half-interval.*

CARTESIAN PRODUCTS. Apart of the set operation such as the union, the intersection, the difference, and the complement of a set, which are usually applied to two ore more Section 2.1.3

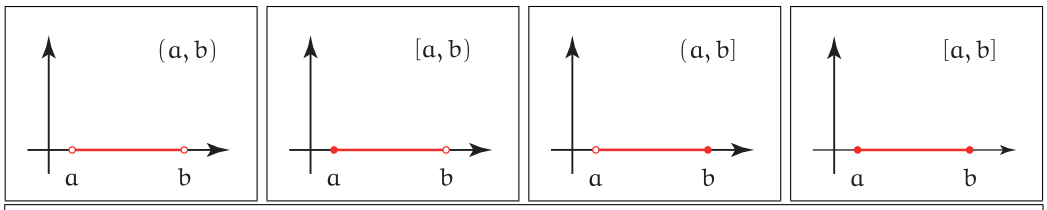


FIGURE A.3: INTERVALS. Four different types of intervals on the real line. Left, an open interval, the next two images show half-open intervals, and the image on the right-hand side visualizes a closed interval.

different sets, new sets can also be constructed via Cartesian products over the same or different sets.

Let A, B be two arbitrary, non-empty sets, then their *Cartesian product*, $A \times B$, is defined to be the set of *all ordered pairs* (a, b), with $a \in A$ and $b \in B$, that is,

$$A \times B \stackrel{\text{def}}{=} \{(a, b) \,|\, a \in A, b \in B\}. \tag{A.1}$$

The idea of the Cartesian product may be extended to products of more than two sets. For example the Cartesian product $A_1 \times A_2 \times \ldots \times A_n$ is defined to be the set of *all ordered n-tuples* $(a_1, a_2, \ldots, a_n)$, where $a_1 \in A_1, a_2 \in A_2, \ldots, a_n \in A_n$.

Section 2.4.3

EXAMPLE A.4 (The Euclidean Plane $\mathbb{R}^2$ and the Euclidean Space $\mathbb{R}^3$) *Obviously the well-known 2-dimensional Euclidean plane can be represented by*

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}, \tag{A.2}$$

that is, the Cartesian product of the real line in 2 dimensions. Then any point in the plane can be written as an ordered pair (x, y) *of real numbers.*

As one can easily see, the situation just described can generalized to higher dimensions. So, the set $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ *is the set of all ordered triples* (x, y, z) *of real numbers, i.e.*

$$\mathbb{R}^3 \stackrel{\text{def}}{=} \{\mathbf{x} = (x, y, z) \,|\, x, y, z \in \mathbb{R}\}. \tag{A.3}$$

Geometrically, $\mathbb{R}^3$ *can be interpreted as the Euclidean space, where any element* $\mathbf{x} \in \mathbb{R}^3$ *can be interpreted as a point in this space and is identified with coordinates* $x, y,$ *and* z.

Generally, we can define $\mathbb{R}^n$ *to be the set of all ordered* n*-tupels of real numbers, namely by:*

$$\mathbb{R}^n \stackrel{\text{def}}{=} \underbrace{\mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}}_{n \times \text{times}} = \{\mathbf{x} = (x_1, x_2, \ldots, x_n) \,|\, x_i \in \mathbb{R}, 1 \leq i \leq n\}. \tag{A.4}$$

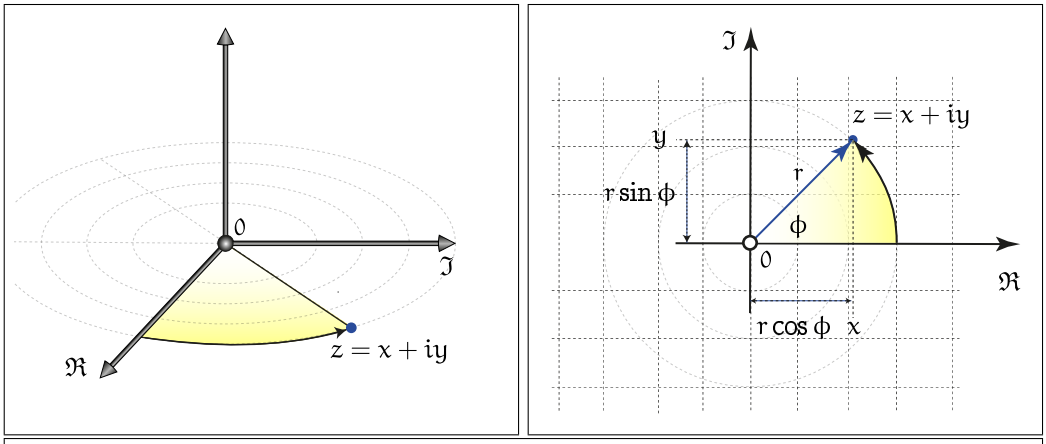


FIGURE A.4: COMPLEX PLANE $\mathbb{C} = \mathbb{R}^2$. A point z of the complex plane can be represented by a pair of (x, y), where x is the real part of z and y is the imaginary part.

EXAMPLE A.5 (Complex Plane $\mathbb{C} = \mathbb{R}^2$) *Another example, based on the Cartesian product $\mathbb{R}^2$, is the complex plane $\mathbb{C}$. The complex plane $\mathbb{C}$ can be formulated as the Cartesian product of the real and imaginary part of complex numbers. So, any complex number $z = a + ib$ can be represented graphically as a point $(a, b) = (\Re(z), \Im(z))$ of real numbers, in the complex plane, see Figure A.4. According to Euler's formula, which shows a deep relationship between the trigonometric functions and the complex exponential function, every complex number z can also be written in the form $z = re^{i\theta} = r(\cos\theta + i\sin\theta)$, where $\cos\theta$ and $\sin\theta$ are the polar coordinates of z in the complex plane.*

Polar Coordinates (832)

We close our discussions about Cartesian products with an important example from probability theory. It combines the concept of the Cartesian product with the power set, and the σ-algebra: the stochastic experiment of flipping a coin n-times.

EXAMPLE A.6 (Stochastic Experiment of Flipping a Coin n-times) *The outcome of this experiment is the n-times Cartesian product of $\{0, 1\}$, i.e. the set $\Omega = \{0, 1\}^n$ of all ordered n-tuples $(\omega_1, \ldots, \omega_n)$ of length n with $\omega_i \in \{0, 1\}$. Again, the power set $\mathfrak{P}(\Omega)$ is the collection of all subsets of Ω, consisting of the empty set, the set Ω, as well as all 1-element, 2-element, ..., and $n-1$-element sets of elements of Ω, resulting in 2^{2^n} sets. Obviously, $\mathfrak{P}(\Omega)$ is a σ-algebra. We leave the proof to the interested reader.*

REMARK A.2 (Flipping a Coin Infinitely Times) *From our discussion above, we conclude that the outcome of flipping a coin infinitely times is the set of all sequences ω_n of infinite length with $\omega_n \in \{0, 1\}, n \in \mathbb{N}$. We also conclude that the power set $\mathfrak{P}(\Omega)$ is*

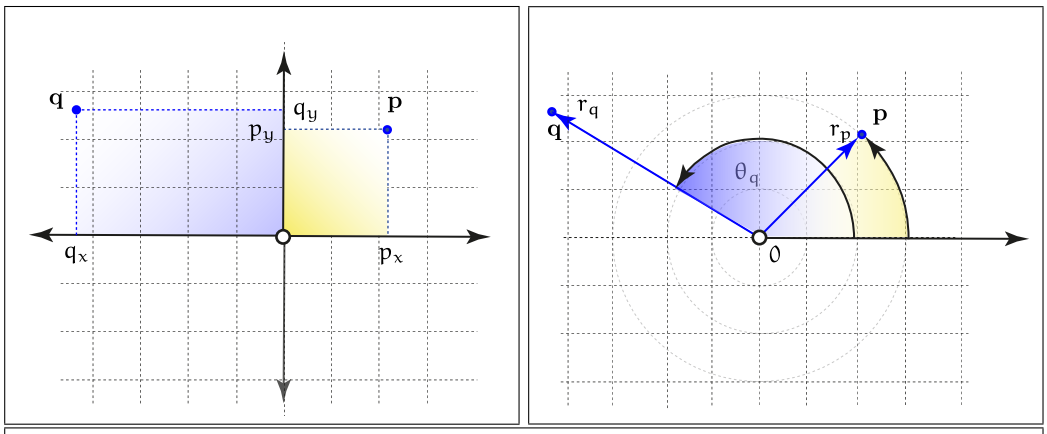


FIGURE A.5: FROM CARTESIAN TO POLAR COORDINATES. In the left image, two points p and q are given via its coordinates (x, y) in a 2-dimensional Cartesian coordinate system. In the right image, the same points are described in polar coordinates, (r, θ). Points in polar coordinates can be specified by the pole 0 and a ray starting at the pole in direction to the point, called the polar axis. The distance form the pole is called the *radial coordinate*, the *radius* or the *length* and the angle is the *angular coordinate*, also called the *polar angle*.

the collection of all subsets of Ω, consisting of the empty set, the set Ω, as well as all 1-element, 2-element, ... sets of elements of Ω. Obviously this set has infinitely many elements. It is left to the reader to show, that $\mathfrak{P}(\Omega)$ is a σ-algebra.

POLAR AND SPHERICAL COORDINATES. Instead of describing a point $\mathbf{p}$ in the Euclidean plane by its coordinates (x, y) with respect to two perpendicular axes, we can also represent it as a pair of a *distance* to the origin of the given coordinate system and an *angle* between the horizontal axis and a ray through the point. Thus the point is described by $(r, \theta)^T$ with $r \geq 0$ and $\theta \in [0, 2\pi)$.

Using the usual axes, where x, y are the ordinary coordinates of our point, then it is easy to see:

$$x = r\cos\theta \quad \text{and} \quad y = r\sin\theta \tag{A.5}$$

with $r = \sqrt{x^2 + y^2}$. Now, a point $(x, y)^T$ on the plane can be represented by its polar coordinates $(r\cos\theta, r\sin\theta)^T \in [0, \infty) \times [0, 2\pi)$, see Figure A.5.

Section 2.3

Because the integration domains of the integral equations in global illumination are generally the unit sphere, the hemisphere, or subsets of these, we present in addition to the above polar coordinates also *spherical coordinates* as alternative representations of points in $\mathbb{R}^3$.

From Figure A.6, we conclude, that a 3D point $\mathbf{p} = (x, y, z)^T$ can be represented by the direction of a line from the origin trough the point $\mathbf{p}$ and a distance from origin to $\mathbf{p}$,

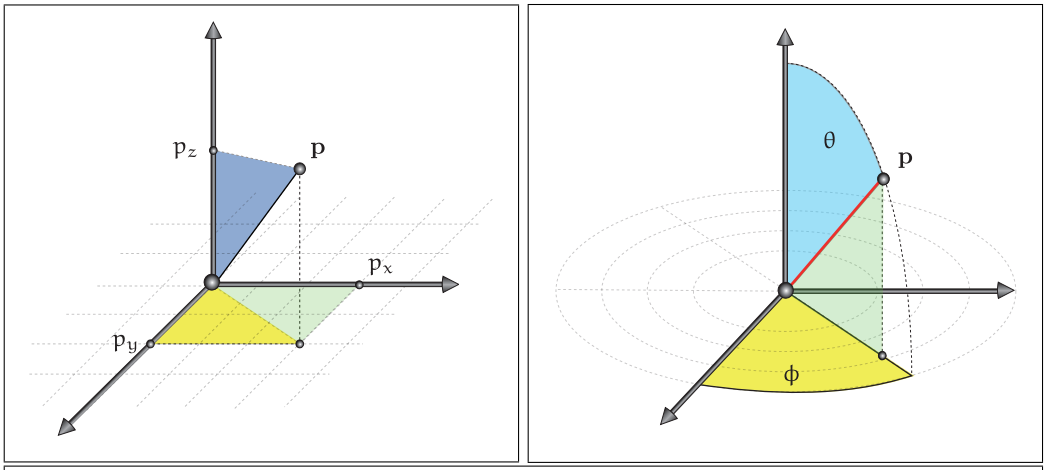


FIGURE A.6: SPHERICAL COORDINATES. In the left image, a point p is given via its coordinates (x, y, z) in a 3-dimensional Cartesian coordinate system. In the right image, the same point is described in spherical coordinates, (r, θ, ϕ). Points in polar coordinates can be specified by the pole 0 and a ray starting at pole in direction to the point, called the polar axis. The distance form the pole is called the *radial coordinate*, the *radius* or the *length*, the angle θ is the *inclination angle* or *polar angle*, and the angle ϕ is called the *azimuthal angle*.

i.e. by a triple of coordinates $(r, \theta, \phi)^T$ with $r \geq 0$, $0 \leq \theta \leq \pi$ and $0 \leq \phi < 2\pi$, where θ and ϕ indicates the direction and r informs about the distance of $\mathbf{p}$ from origin. Here $\theta \in [0, \pi]$ describes the angle made by a line from the origin to $\mathbf{p}$ with the z axis of a left-handed Cartesian coordinate system, and $\phi \in [0, 2\pi)$ describes the angle made by the projection of this line onto the xy plane with the x axis. Obviously it holds:

$$(x, y, z)^T = (r \sin\theta \cos\phi, r \sin\theta \sin\phi, r \cos\theta)^T \tag{A.6}$$

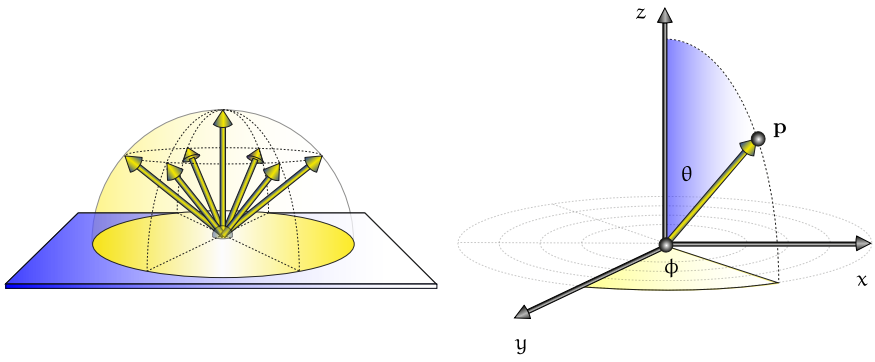
with $r = \sqrt{x^2 + y^2 + z^2}$. A point (x, y, z) in Euclidean space can now be represented by its spherical coordinates $(r \sin\theta \cos\phi, r \sin\theta \sin\phi, r \cos\theta)^T \in [0, \infty) \times [0, \pi] \times [0, 2\pi)$.

BOX A.1 (Directions)

Considering the unit sphere centered around the origin of a 3-dimensional Cartesian coordinate system, then all points on the sphere have the same distance $r = 1$. In this case any point $\mathbf{p}$ on the sphere can be defined by only two of the three spherical coordinates (r, θ, ϕ), namely (θ, ϕ), dropping the coordinate r. Here $\theta \in [0, \pi]$ describes the angle made by a line from the origin to $\mathbf{p}$ with the z axis of a left-handed Cartesian coordinate system, and $\phi \in [0, 2\pi)$ describes the angle made by the projection of this line onto the xy plane with the x axis, see the Figure below.

As many functions from field of realistic rendering need to be integrated over directions incident at a given surface point—in other words over points on the unit sphere or the unit hemisphere—rather than the whole space, we introduce a shortened notation for directions. Therefore we define a direction $\omega \in [0, \pi] \times [0, 2\pi)$ by:

$$\omega \stackrel{\text{def}}{=} (\theta, \phi) \stackrel{\text{def}}{=} (\theta, \phi, 1). \tag{A.7}$$



RELATIONS. Before we introduce the concept of the operator as a mapping between sets, let us shortly speak about relations, the mathematical concept, that surrounds the concept of the operator as a special case.

Let A, B be two non-empty sets, then a *relation* R is a subset of the Cartesian product of A and B, that is: $R \subseteq A \times B$. According to this definition, the points $(x, x) \in \mathbb{R}^2$ lying on a line through the origin of the Euclidean plane forms a relation.

Now, there are a types of relations that are useful for our further consideration on measure theory, particularly relations, in which the ordered pairs come from only a single set, that is: Given a set A, we consider relations on A. Such a relation, in sign $\sim$, is *reflexive*, if $a \sim a$, it is *symmetric*, if $a \sim b \Rightarrow b \sim a$, and it is *transitive*, if $a \sim b$ and $b \sim c \Rightarrow a \sim c$ holds, $\forall a, b, c \in A$. A relation that is reflexive, symmetric, and transitive is called an *equivalence relation*. Equivalence relations have the beautiful property that they specify how to partition the set $A \times A$ into subsets such that every element of the

Section 2.2.3

larger set is exactly in one of the subsets. Then, we say two elements of the larger set are *equivalent* with respect to a equivalence relation if and only if they are also elements of the same set. To show that an equivalence relation forms a partition of $A \times A$, we suppose, that $\sim$ defines an equivalence relation on the set A. Then for each $a \in A$, we can define the set $[a] \stackrel{def}{=} \{x \in A | x \sim a\}$, that is, the set of all elements of A, which are equivalent to a. Because it is easy to see, that $[a] \cap [b] = \emptyset$, an equivalence relation $\sim$ on A partitions A into disjoint *equivalence classes* $[a]$, resulting in $A = \bigcup_{a \in A} [a]$. If we would suppose that $[a] \cap [b] \neq \emptyset$ holds, then there must be an element $c \in A$ with $a \sim c$ and $c \sim b$, hence $a \sim b$, so that $[a] = [b]$. Section 2.2.3

EXAMPLE A.7 *As a simple example, we can construct an equivalence relations $\sim$ on $\mathbb{R} \times \mathbb{R}$, where the equivalence classes are defined by $[a] \stackrel{def}{=} \{(x, a) | x \in \mathbb{R}, a \text{ fix}\}$, that is, the set of all points lying on the horizontal line $y = a$. As one can easily see, the set of all equivalence classes $\{[a] \,|\, a \in \mathbb{R}\}$ defines a partition of the Euclidean plane in an uncountable set of lines parallel to the x-axis.*

FUNCTIONS. The perhaps most important concept in mathematics is that of the function, a special kind of an *operator*, that describes a rule how elements are mapped from a set to another set. As the concept of the operator is fundamental in any field of mathematics, we will refresh useful properties of operators and introduce some functions, which play an important role in our following discussions. For that purpose, let $\mathcal{S}$ and $\mathcal{T}$ be arbitrary nonempty sets.

DEFINITION A.2 (Operators) *An operator f from the subset $\mathrm{Dom}(f) \subseteq \mathcal{S}$ into the set $\mathcal{T}$, written as*

$$f : \mathrm{Dom}(f) \to \mathcal{T}, \tag{A.8}$$

is a mapping that assigns each $x \in \mathrm{Dom}(f)$ a unique element $f(x) \in \mathcal{T}$, thus

$$x \mapsto f(x), \tag{A.9}$$

where $f(x)$ is the image of x under the mapping f. We refer the set $\mathcal{T}$ as the image of the operator *and call the set $\mathrm{Dom}(f)$, the* domain *of f. The set $\mathrm{Im}(f)$, defined by:*

$$\mathrm{Im}(f) \stackrel{def}{=} \{f(x) | x \in \mathrm{Dom}(f)\}, \tag{A.10}$$

is called the range *of f.*

We denote an operator f a real *or* complex valued function, *if $\mathrm{Im}(f) \subset \mathbb{K}$ with $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$.*

Obviously, all real functions of a single or several variables are operators between the set $\mathbb{R}$, respectively $\mathbb{R}^s$, and the image area $\mathbb{R}$, with $s \geq 2$.

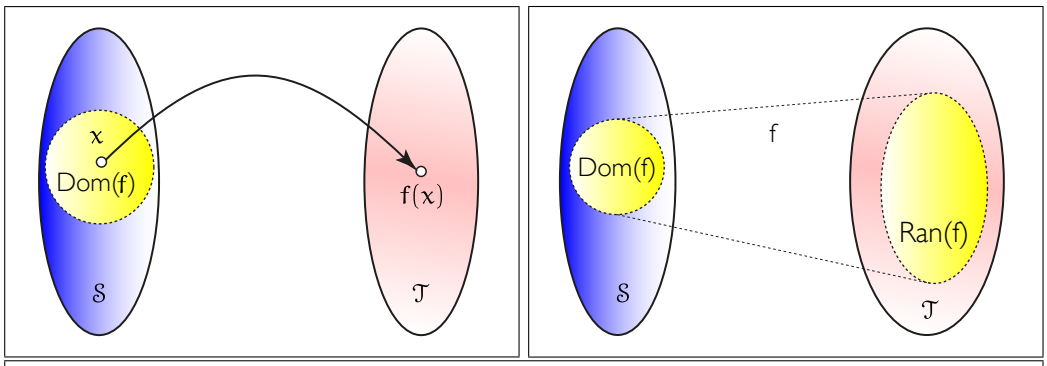


FIGURE A.7: OPERATORS. The mathematical concept of an operator as a mapping between two sets $\mathcal{S}$ and $\mathcal{T}$. The set $\mathrm{Dom}(f) \subseteq \mathcal{S}$ is the domain of the operator, $\mathrm{Dom}(f)$ is called the image, and the set $\{f(x)|x \in \mathrm{Dom}(f)\}$ is the range of the operator.

EXAMPLE A.8 *i) Let S be the real line and* $T = \{0, 1\}$*, then the function* D

$$D : \mathbb{R} \longrightarrow T \tag{A.11}$$

given by:

$$x \longmapsto f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases} \tag{A.12}$$

is called the Dirichlet function*. This function, that maps a rational number to one and an irrational number to zero, plays an important role in measure theory. The Dirichlet function is a real valued function defined on the real line, which can not be graphically visualized exactly by a corresponding graph. How we will see shortly, the Dirichlet function is the most famous example of a function, which is not Riemann- but Lebesgue-integrable.*

Section 2.2.3 Dirichlet Funktion (106) Lebesgue-integrable (105)

ii) Now let S be the Euclidean space $\mathbb{R}^3$ *and* $T = \mathbb{R}$*, then the function*

$$\|\cdot\|_2 : \mathbb{R}^3 \longrightarrow \mathbb{R}, \quad \mathbf{x} \mapsto \sqrt{x_1^2 + x_2^2 + x_3^2} \tag{A.13}$$

is a real valued function defined on $\mathbb{R}^3$*, which returns the distance of point* $\mathbf{x}$ *from origin of a three dimensional Cartesian coordinate system.*

Section A.3 $\|\cdot\|_2$ (861) Section 2.1.3 Section 2.1.1

In one of the following sections about the Euclidean space $\mathbb{R}^3$*, the function* $\|\cdot\|_2$ *is introduced as the so-called* Euclidean norm*, a mathematical concept, which allows us, to measure the length of vectors or the distance between points, vectors, and functions. Thus, the Euclidean norm is used in all ray tracing algorithms to determine, which hit point of a ray with objects in the scene is closest to the camera. The construct of the norm also allows to introduce the concept of the* limit of sequences,

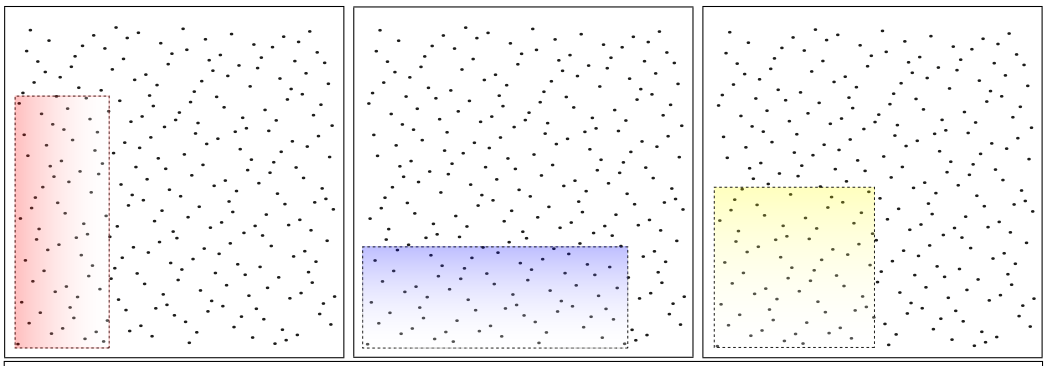


FIGURE A.8: A FIRST ENCOUNTER WITH THE CONCEPT OF DISCREPANCY. A 256-element point set within $\mathbf{I}^2 = [0,1]^2$. The three colored rectangles have the same Lebesgue area measure, that is, the term $\left|\frac{\#(\mathbf{P}\cap\mathbf{B})}{N} - \text{area}(\mathbf{B})\right|$ for the two rectangles is equal to $\left|\frac{59}{256} - \text{area}(\mathbf{B})\right|$. For the square we have $\left|\frac{65}{256} - \text{area}(\mathbf{B})\right|$.

fundamental in linear normed spaces.

iii) A set function *is a function defined on a system $\mathfrak{A}$ of sets. It maps any subset $A \in \mathfrak{A}$ to a real or complex number.*

Let $\mathfrak{A}$ be the collection of all subsets of $[a,b]$. Then, we can define a set function l by:

$$l : \mathfrak{A} \to \mathbb{R}^{\geq 0} \tag{A.14}$$

with

$$[\alpha,\beta] \mapsto l([\alpha,\beta]) = \beta - \alpha \tag{A.15}$$

for $[\alpha,\beta] \subseteq [a,b]$. It is obviously, that the function l assigns an interval $[\alpha,\beta] \subseteq [a,b]$ its length. As we shall see, the concept of the set function plays the central role when defining a measure. Measure (79)

iv) Another interesting example of a set function could be the function $D_N(\mathbf{B})$ defined by:

$$D_N(\mathbf{B}) \quad \stackrel{\text{def}}{=} \quad \max_{\mathbf{B}\subseteq[0,1]^2} \left|\frac{\#(\mathbf{P}\cap\mathbf{B})}{N} - \textit{area}(\mathbf{B})\right|, \tag{A.16}$$

whereas $\mathbf{P}$ is a fixed N-element point set form $[0,1]^2$, $\mathbf{B}$ is an axis-aligned rectangle with lower-left corner at the origin, and $\#$ is the so-called counting measure, *which returns the number of points contained in the intersection of $\mathbf{P}$ and $\mathbf{B}$, see Figure A.8.* # (81)

The function $D_N(\mathbf{B})$ *is a slightly modified, 2-dimensional version of the* discrepancy*, a mathematical concept for measuring the deviation of a point set from its ideal distribution, used in quasi-Monte Carlo integration. It returns the maximum difference between the fraction of points of a* N*-element point set* $\mathbf{P}$ *inside one of the subrectangles* $\mathbf{B}$ *of the set of all axis-aligned rectangles* $[0,1]^2$*.*

Section 7.1

v) Let us suppose $S = [0,\pi] \times [0,2\pi)$ *and* $T = \mathbb{R}^3$*, then the function*

$$f_\omega : [0,\pi] \times [0,2\pi) \longrightarrow \mathbb{R}^3, \quad \omega = (\theta,\phi) \mapsto (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta) \tag{A.17}$$

assigns every line through the origin its intersection point with the unit sphere around the origin. Here we have a function, which is defined on directions around a point.

vi) Now let $S = \left[0,\frac{\pi}{2}\right] \times [0,2\pi)$ *and* $T = \mathbb{R}$*, then the function*

$$I : \left[0,\frac{\pi}{2}\right] \times [0,2\pi) \longrightarrow \mathbb{R}^3, \quad \omega = (\theta,\phi) \mapsto \frac{C}{2\pi} \tag{A.18}$$

assigns every direction around any point the constant $\frac{C}{2\pi}$ *with* $C > 0$*. In computer graphics, such a function can be used to describe the physical process of diffuse reflection on any surface point, where the reflected amount of light in every direction is* $\frac{C}{2\pi}$*. Functions of these kind play an important role in radiometry and the field of realistic rendering.*

Chapter 3

vii) Last but not least, let us consider a mapping $\mathbf{T}$ *between the three and the two-dimensional Cartesian product constructed over the real number set* $\mathbb{R}$*, that is,*

$$\mathbf{T} : \mathbb{R}^3 \rightarrow \mathbb{R}^2 \tag{A.19}$$

$$\mathbf{x} = (x_1,x_2,x_3)^T \mapsto \mathbf{T}\mathbf{x} = (x_1,x_2)^T. \tag{A.20}$$

Obviously, this function maps a triple $(x_1,x_2,x_3)^T$ *to the tupel* $(x_1,x_2)^T$*, that is, the operator* $\mathbf{T}$ *discards the* 3^{rd} *component of* $\mathbf{x}$*. Operators of this type, also refers to as* projection operators*, will be discussed in more detail in Section 2.1.*

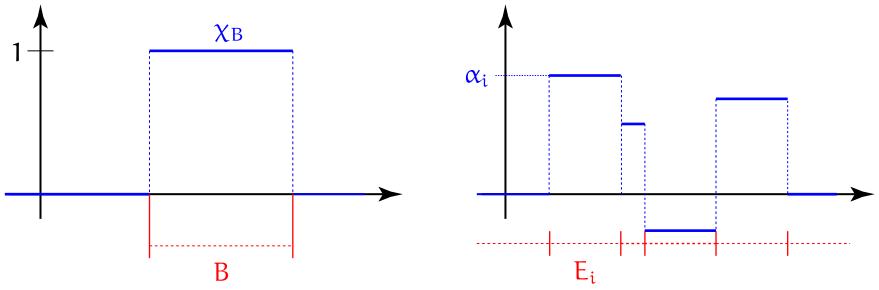
BOX A.2 (The Characteristic Function, χ, and the Concept of Simple Functions)

A very important function in our discussions about measure and integration theory is the *caracteristic function* χ, defined over space S with respect to a subset $B \subset S$ by:

$$\chi_B : S \longrightarrow \{0,1\}, \quad x \mapsto \chi_B(x) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if} \quad x \in B \\ 0 & \text{otherwise,} \end{cases} \tag{A.21}$$

see the left Figure below.

When deriving the Lebesgue integral, we need the concept of *simple functions*, where we call a function *simple*, if it takes only finitely different values. If we additionally assume that all values of a simple function are finite, then the characteristic function from (A.21) is a trivial example of a simple function.



As it is easy seen from the right Figure shown below, every simple function can be written as a sum of characteristic functions of pairwise disjoint sets. Therefore let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be different values of a simple function f. If we set

$$E_i = \{x \in S | f(x) = \alpha_i\} \quad \text{and} \quad S = \bigcup_{i=1}^{n} E_i, \quad \text{with } E_i \cap E_j = \emptyset, \quad i \neq j \tag{A.22}$$

then a simple function can formally be defined by:

$$f(x) = \sum_{i=1}^{n} \alpha_i \chi_{E_i}(x), \tag{A.23}$$

where S is any domain and $\alpha_i \in \mathbb{R}, 1 \leq i \leq n$.

Let us consider the function $f(x) = e^x$, with domain $\mathrm{Dom}(f) = \mathrm{Im}(f) = \mathbb{R}$ and $\mathrm{Im}(f) = \mathbb{R}^{>0}$. Obviously, this function has the property, that each image $y = f(x)$ can be mapped to its preimage x, that is, we get a new function f^{-1} by reversing the rule underlying the mapping $f : \mathbb{R} \to \mathbb{R}^{>0}$, thus,

$$f^{-1} : \mathbb{R}^{>0} \quad \to \quad \mathbb{R} \tag{A.24}$$

$$y \quad \mapsto \quad f^{-1}(y). \tag{A.25}$$

It should be clear, that the characteristic function χ_B from above does not share this property with the exponential function e^x. The reason for this is, that the range of the characteristic function compared with its domain is obviously to small for defining a new mapping. In mathematics, we say also, the characteristic function is not *injective*. Obvioulsy, operators can be characterized with respect to the rule, that they describe, that is, the rule underlying an operator can lead to different types of the operators.

DEFINITION A.3 (Injective, Surjective, and Bijective Operators) *Let* f *be an operator between the sets* $\mathcal{S}$ *and* $\mathcal{T}$. *The mapping* f *is referred to as* injective *or* one-to-one, *if no two distinct elements of* $\mathrm{Dom}(f)$ *are mapped to the same image within* $\mathrm{Im}(f)$, *that is,*

$$x_1 \neq x_2 \Longrightarrow f(x_1) \neq f(x_2). \tag{A.26}$$

If the range of f *is identical to the image of* f, *that is, if it holds:*

$$\mathrm{Ran}(f) = \mathcal{T}, \tag{A.27}$$

then the operator f *is denoted as* surjectve *and we say:* $\mathbf{T}$ *maps* $\mathcal{S}$ *onto* $\mathcal{T}$.

Furthermore, we call an operator bijective, *if it is surjective and injective at the same time.*

REMARK A.3 (Invertible Operators) *Bijective operators, such as the exponential function* $f(x) = e^x$*—we let the proof of the bijectivity of the exponential function as a simple exercise to the interested reader—have a special property: they are* invertible. *Invertibility of an operator* f, *described by,*

$$f : \mathcal{S} \rightarrow \mathcal{T} \tag{A.28}$$

$$x \mapsto f(x) \tag{A.29}$$

means, that the rule underlying the operator can also be reversed, that is, the rule

$$f^{-1} : \mathrm{Ran}(f) \rightarrow \mathcal{S} \tag{A.30}$$

$$y \mapsto f^{-1}(y) = f^{-1}(f(x)) = x, \tag{A.31}$$

is an operator too.

The notion of the invertibility of an operator is central for the mathematical concept of the measurable function *used in measure, integration, and probability theory. Thus for example, the fundamental notion of the* random variable *defined on a corresponding probability space is based on the concept of the measurable function.*

Section 2.2.3
Section 2.4.2

A.2 THE EUCLIDEAN SPACE $\mathbb{R}^3$ AS A FIRST SIMPLE EXAMPLE OF A LINEAR SPACE

The mathematical concept of the *linear function space* is one of the main concepts in our discussions about methods for solving the global illumination problem. For defining a

Global Illumination Problem (6)

linear function space, we need the concept of an *abstract linear space* from linear algebra or calculus. To imagine what mathematical constructs underly the concept of the linear space, it is helpful to introduce the concept of the *Euclidean space* $\mathbb{R}^3$ as a first simple example of a linear space. Because the $\mathbb{R}^3$ provides a good intuitive model for the behavior of elements in more abstract linear spaces— such as functions spaces, that need not to have a geometric interpretation—we now present the most important concepts of linear algebra on the basis of the Euclidean space $\mathbb{R}^3$.

In mathematics, a *vector space*, also called a *linear space*, is defined as a collection of objects, which we call *vectors*. In the Euclidean spaces vectors can be represented by ordered pairs or triples of real numbers, and can be visualized by arrows characterized by a length and a direction. We will show in this section, that, satisfying certain axioms, vectors can be added using the *parallelogram rule* and that they can be multiplied by a real number resulting in changes in *direction* and in *length* of the vector. Additionally, we will present a set of important mathematical constructs, which play a fundamental role in the theory of vector spaces, such as the constructs of the *linear combination, linear dependance*, and that of the *basis*, as well as the concepts of the *norm*, *orthogonality*, the *inner product*, and the *metric*. Finally, we discuss the mathematical construct of a *matrix* as an operator between Euclidean spaces over $\mathbb{R}$.

THE LINEAR SPACE $\mathbb{R}^3$. In Example A.4 of the last section, we introduced the Euclidean plane $\mathbb{R}^2$ and the Euclidean space $\mathbb{R}^3$ as simple examples of Cartesian products over the field $\mathbb{R}$. Now any triple $\mathbf{x} \in \mathbb{R}^3$ with

Cartesian Product (829)

$$\mathbf{x} \stackrel{\text{def}}{=} (x_1, x_2, x_3)^T \equiv \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \in \mathbb{R}^3 \tag{A.32}$$

is called a *point* or a *vector*, whereas the real numbers $x_i, 1 \leq i \leq 3$, are denoted as the *components* or the *coordinates* of vector $\mathbf{x}$.

As is easily seen from Figure A.9, any vector $\mathbf{x} \in \mathbb{R}^3$ can be identified with the directed line segment that has its initial point at the origin and its end point with the Cartesian coordinates given by the components of $\mathbf{x}$. By identifying vectors with directed line segments, we shall follow the convention that any line segment with the same direction and the same length may be used to present the same vector $\mathbf{x}$.

We say two vectors $\mathbf{x}$ and $\mathbf{y}$ are *equal*, written $\mathbf{x} = \mathbf{y}$, if they have the same number of components and if corresponding components are equal. The *sum* of two vectors $\mathbf{x}$ and $\mathbf{y}$, written as $\mathbf{x}+\mathbf{y}$, is the vector, which we obtain by adding the corresponding components, that is, the addition of two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined via:

$$\mathbf{x}+\mathbf{y} \stackrel{\text{def}}{=} (x_1+y_1, x_2+y_2, x_3+y_3)^T \equiv \begin{pmatrix} x_1+y_1 \\ x_2+y_2 \\ x_3+y_3 \end{pmatrix}. \tag{A.33}$$

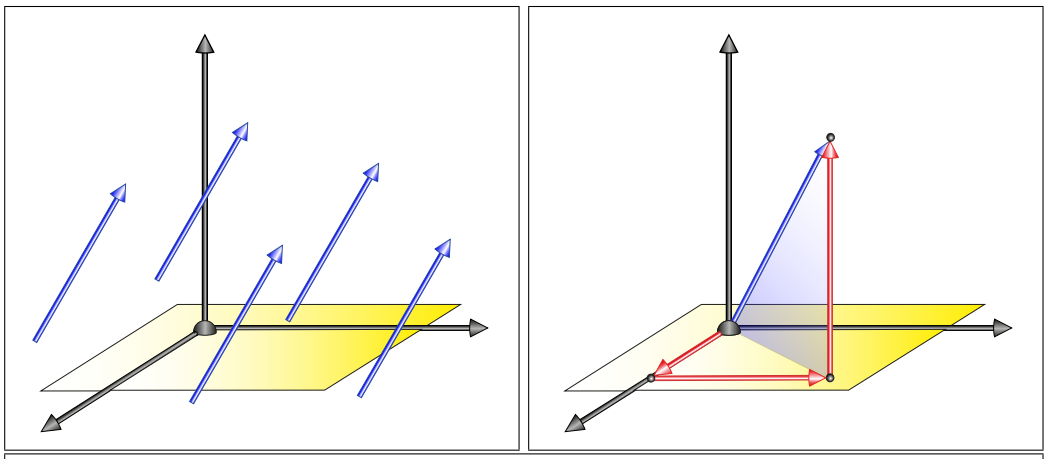


FIGURE A.9: VECTORS. The left image shows the representation of a vector in $\mathbb{R}^3$ as a representant of the set to all arrows with the same length and the same direction. On the right-hand side a vector is shown as the result of the addition of thee other vectors.

Obviously, in the above equation, $\mathbf{x}+\mathbf{y}$ is the diagonal of a parallelogram, which has $\mathbf{x}$ and $\mathbf{y}$ as two adjacent sides. This is illustrated in Figure A.10. The vector $\mathbf{x}+\mathbf{y}$ can be drawn by placing the initial point of $\mathbf{y}$ at the terminal point of $\mathbf{x}$ and then drawing the directed line segment from the initial point of $\mathbf{x}$ to the end point of $\mathbf{y}$. This *heads to tails* construction, shown in Figure A.10, is called the *parallelogram rule* for adding vectors.

We can also multiply a vector $\mathbf{x} \in \mathbb{R}^3$ with a real number α, a so-called *scalar*, by multiplying the components of $\mathbf{x}$ with α, defined by:

$$\alpha \cdot \mathbf{x} \stackrel{\text{def}}{=} (\alpha x_1, \alpha x_2, \alpha x_3)^T \equiv \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \alpha x_3 \end{pmatrix}, \tag{A.34}$$

Section A.3 where we call the multiplication defined by Equation (A.34) the *s-multiplication*, or the *scalar multiplication*. From our interpretation above it is clear that all these vectors lie on a line passing through the origin, for this see Figure A.10.

Multiplying a vector $\mathbf{x}$ with -1 results in the vector $-\mathbf{x}$. By defining $\mathbf{x}-\mathbf{y} \stackrel{\text{def}}{=} \mathbf{x}+(-\mathbf{y})$, we can also subtract two vectors, and conclude that $(\mathbb{R}^3,+)$ satisfies the axioms of an Abelean group. That is, the operation of addition is commutative and associative, and there exist only a single element, the zero element $\mathbf{0}$, for which it holds: $\mathbf{x}+\mathbf{0}=\mathbf{0}+\mathbf{x}=\mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^3$. Additionally, for any vector $\mathbf{x} \in \mathbb{R}^3$ there exist only one vector $-\mathbf{x}$, which is inverse to $\mathbf{x}$. Equipped with the definition of the s-multiplication we call $\mathbb{R}^3$, as well as
Section A.3 $\mathbb{R}^2$, a *linear space* or a *vector space*, if the *distributive law*, $\alpha(\mathbf{x}+\mathbf{y}) = \alpha\mathbf{x}+\alpha\mathbf{y}$, and the rule $\alpha(\beta\,\mathbf{x}) = (\alpha\,\beta)\mathbf{x}$ even holds.

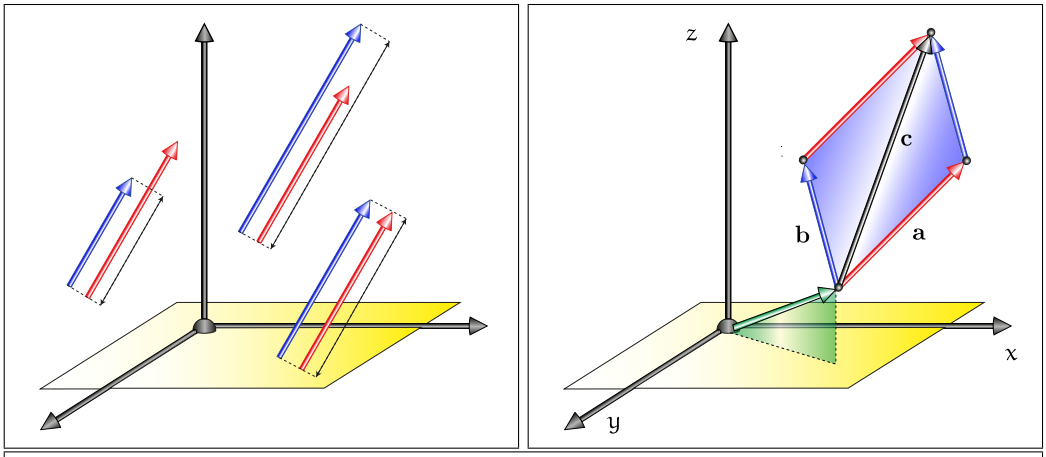


FIGURE A.10: VECTOR OPERATIONS. Left, the visualization of the s-multiplication applied to the blue vector in $\mathbb{R}^3$. Multiplication with a scalar $\alpha < 1$ shortens the vector, multiplication with $\alpha > 1$ enlarges the vector. On the right-hand side the visualization of the parallelogram rule for adding two vectors $\mathbf{a}, \mathbf{b}$ of $\mathbb{R}^3$ resulting in a new vector $\mathbf{c} \in \mathbb{R}^3$.

LINEAR COMBINATION AND LINEAR DEPENDANCE. As is easily seen from Figure A.10, any vector $\mathbf{x} = (x_1, x_2, x_3)^T$ can be represented as sum of multiples of three vectors $\mathbf{e}_1 = (1,0,0)^T, \mathbf{e}_2 = (0,1,0)^T$, and $\mathbf{e}_3 = (0,0,1)^T$, if we write:

$$\begin{aligned} \mathbf{x} &= x_1\,\mathbf{e}_1 + x_2\,\mathbf{e}_2 + x_3\,\mathbf{e}_3 && \text{(A.35)}\\ &= x_1(1,0,0)^T + x_2(0,1,0)^T + x_3(0,0,1)^T && \text{(A.36)}\\ &= (x_1,0,0)^T + (0,x_2,0)^T + (0,0,x_3)^T && \text{(A.37)}\\ &= (x_1,x_2,x_3)^T. && \text{(A.38)} \end{aligned}$$

In linear algebra the Relation (A.35) is denoted as a *linear combination*. As is easily seen, any vector $\mathbf{x} \in \mathbb{R}^3$ corresponds to the diagonal of a parallelepiped with adjacent edges $(x_1,0,0)^T, (0,x_2,0)^T$, and $(0,0,x_3)^T$. The parallelogram law along the edges of this parallelepiped then shows that it holds: $\mathbf{x} = (x_1,x_2,x_3)^T$. The observation then implies the following definition of a linear combination of vectors in $\mathbb{R}^3$:

DEFINITION A.4 (Linear Combination in $\mathbb{R}^3$) *A vector $\mathbf{y} \in \mathbb{R}^3$ is a* linear combination *of vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathbb{R}^3$, if there exist scalars $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$, such that it holds:* Section A.3

$$\mathbf{y} = \alpha_1\,\mathbf{x}_1 + \alpha_2\,\mathbf{x}_2 + \alpha_3\,\mathbf{x}_3. \qquad \text{(A.39)}$$

Closely related to the notion of linear combination is the concept of linear dependance.

DEFINITION A.5 (Linear Dependance in $\mathbb{R}^3$) *We say the vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathbb{R}^3$ are* linearly dependent *if there exists scalars $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$, not all zero, such that it*

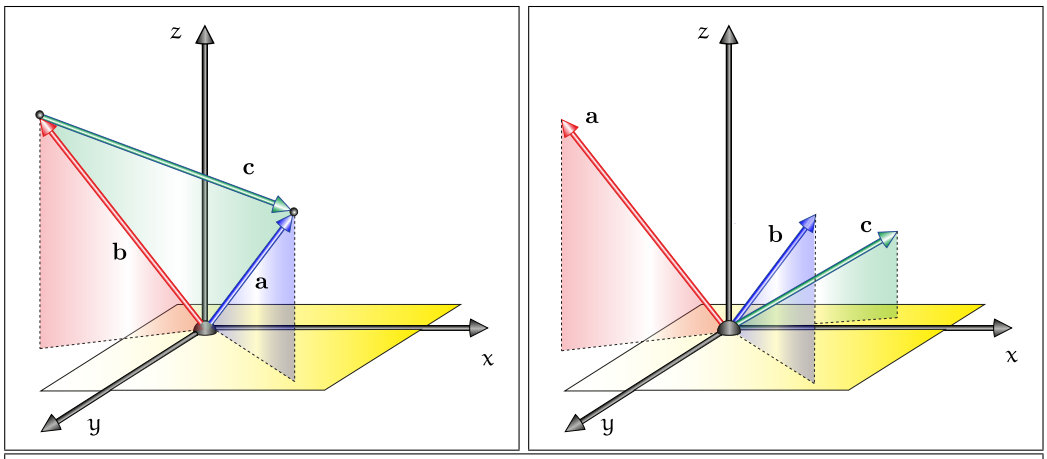


FIGURE A.11: LINEAR DEPENDENT AND LINEAR INDEPENDENT VECTORS. Three linear dependent vectors in the left image in contrast to three linear independent vectors in the right image. Linear dependence in $\mathbb{R}^3$ means that a vector can be described as the addition of two other vectors. Thus, the vector $\mathbf{a}$ in the left image corresponds to the sum of $\mathbf{b}$ and $\mathbf{c}$, but no vector in the right image can be represented as a sum of the two remaining vectors.

holds:

$$\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \alpha_3 \mathbf{x}_3 = \mathbf{0}, \tag{A.40}$$

where $\mathbf{0} = (0,0,0)^T$*, otherwise we say the vectors are linearly independent, see Figure* Section A.3 *A.11.*

Due to this definition the set of vectors $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ are obviously linear independent. Additionally we remark, that more than three vectors are always linear dependent, since any of these vectors can be represented as a linear combination of the three others. We omit the simple proofs of these statements to the interested reader.

BASIS AND DIMENSION. Taking our focus to the vectors $\mathbf{e}_1, \mathbf{e}_2$, and $\mathbf{e}_3 \in \mathbb{R}^3$. As these vectors are linearly independent, we can conclude, that any vector $\mathbf{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ can be represented as a linear combination of $\mathbf{e}_1, \mathbf{e}_2$, and $\mathbf{e}_3$, with $\alpha_i = x_i, 1 \leq i \leq 3$. A set of vectors with these properties will be denoted as a *basis*.

Section A.3 DEFINITION A.6 (Basis in $\mathbb{R}^3$) *A finite set* $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ *of elements of* $\mathbb{R}^3$ *is said to* span $\mathbb{R}^3$*, if every* $\mathbf{x} \in \mathbb{R}^3$ *can be written in the form*

$$\mathbf{x} = \sum_{i=1}^{3} \alpha_i \mathbf{b}_i \tag{A.41}$$

for some real numbers $\alpha_i, 1 \leq i \leq 3$. *The set* $\mathbf{B}$ *is denoted as a* basis *of* $\mathbb{R}^3$, *if and only if the following holds:*

i) $\mathbf{B}$ *spans* $\mathbb{R}^3$, *and*

ii) $\mathbf{B}$ *is a set of linearly independent vectors of* $\mathbb{R}^3$.

REMARK A.4 (Dimension of $\mathbb{R}^3$) *The number of elements that forms a basis of a linear space* $\mathbb{R}^3$ *is called the dimension of* $\mathbb{R}^3$. *Because the set* $\mathbf{E} \stackrel{\text{def}}{=} \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ *builds a basis in the Euclidean space* $\mathbb{R}^3$, *we say* $\mathbb{R}^3$ *is a 3-dimensional linear space, and we write:* $\dim \mathbb{R}^3 = 3$.

INNER PRODUCT. Let us consider the basis $\mathbf{E}$ of $\mathbb{R}^3$ once more, but now, a little more in detail. From Figure A.10, we can see, that these vectors are *perpendicular* to each other. In mathematics one also says: the vectors are *orthogonal*. Multiplying, for example, the components of any two vectors from $\mathbf{E}$ and adding the resulting products, then we yield for $\mathbf{e}_i = (e_{i_1}, e_{i_2}, e_{i_3})^T$ and $\mathbf{e}_j = (e_{j_1}, e_{j_2}, e_{j_3})^T$:

$$e_{i_1} e_{j_1} + e_{i_2} e_{j_2} + e_{i_3} e_{j_3} = 0, \tag{A.42}$$

since all products $e_{i_k} e_{j_k} = 0$ for $i \neq j$ and $1 \leq k \leq 3$.

The product of two vectors from $\mathbb{R}^3$, defined by multiplying corresponding components and adding the resulting products, is called an *inner product*.

DEFINITION A.7 (Inner Product $\langle \cdot, \cdot \rangle_{\mathbb{R}^3}$ and Inner Product Space $(\mathbb{R}^3, \langle \cdot, \cdot \rangle_{\mathbb{R}^3})$) *Let* $\langle \cdot, \cdot \rangle_{\mathbb{R}^3}$ Section A.3 *be a mapping from* $\mathbb{R}^3 \times \mathbb{R}^3$ *to* $\mathbb{R}$, *defined by:*

$$(\mathbf{x}, \mathbf{y}) \mapsto \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^3} \stackrel{\text{def}}{=} \sum_{i=1}^{3} x_i \cdot y_i, \tag{A.43}$$

then $\langle \cdot, \cdot \rangle_{\mathbb{R}^3}$ *is called an* inner product *in* $\mathbb{R}^3$, *if* $\langle \cdot, \cdot \rangle_{\mathbb{R}^3}$ *satisfies the following axioms:*

i) $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{R}^3} \geq 0$ *and* $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{R}^3} = 0$ *iff* $\mathbf{x} = \mathbf{0}$ *(positive-definiteness)*

ii) $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^3} = \langle \mathbf{y}, \mathbf{x} \rangle_{\mathbb{R}^3}$ *(symmetry)*

ii) $\langle \alpha \mathbf{x} + \beta \mathbf{y}, \mathbf{z} \rangle_{\mathbb{R}^3} = \alpha \langle \mathbf{x}, \mathbf{z} \rangle_{\mathbb{R}^3} + \beta \langle \mathbf{y}, \mathbf{z} \rangle_{\mathbb{R}^3}$ *with* $\alpha, \beta \in \mathbb{R}$ *(bilinearity).*

With the inner product $\langle \cdot, \cdot \rangle_{\mathbb{R}^3}$ *the Eucledian space* $\mathbb{R}^3$ *is an* inner product space, *denoted by* $(\mathbb{R}^3, \langle \cdot, \cdot \rangle_{\mathbb{R}^3})$.

The positive definitness, the symmetry and the bilinearity of the inner product are the central formal properties of an inner product. As we will see in the next section, all following statements are consequences of these properties.

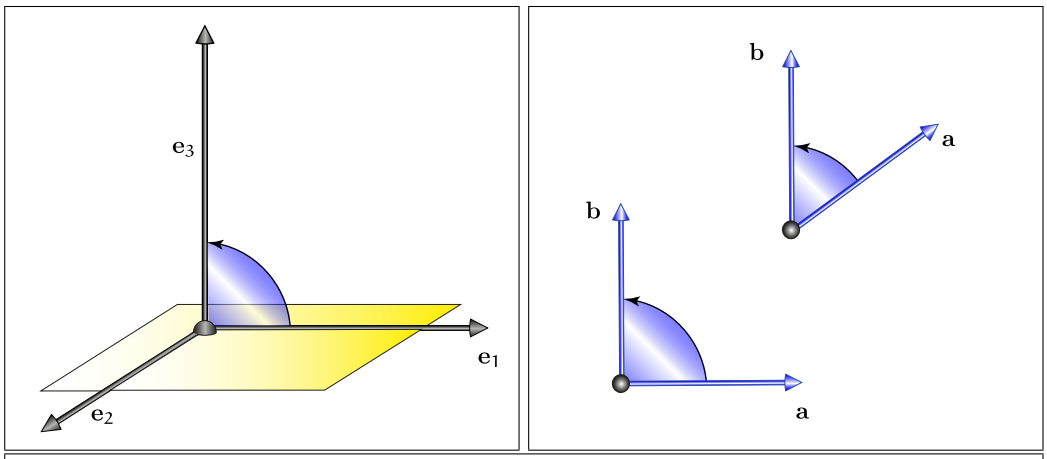


FIGURE A.12: ORTHOGONALITY. The image on the left-hand side shoes the standard basis $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ of $\mathbb{R}^3$, a linear independent set of vectors, where each vector is orthogonal to all other vectors of $\mathbf{E}$. The image on the right shows a pairs of orthogonal and a pair of not orthogonal vectors in $\mathbb{R}^3$.

REMARK A.5 *If it is clear from context, we often not take into account the index in the notation of an inner product.*

From the observation above, we conclude that two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ are *orthogonal*, if there inner product yields zero. This then implies the following definition of the concept of orthogonality in $\mathbb{R}^3$.

Section A.3 DEFINITION A.8 (Orthogonality in $\mathbb{R}^3$) *Based on the definition of an inner product, two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ are denoted as orthogonal, if it holds:*

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^3} = 0. \tag{A.44}$$

NORM. It is well known, that many physical quantities, such as velocity or force, can be represented by vectors. Because vectors are characterized not only by its direction but also by its length we need the concept of a *norm*, which provides information about the length of a vector.

Section A.3 DEFINITION A.9 (The Euclidean Norm $\|\cdot\|_2$ and the Linear Normed Space $(\mathbb{R}^3, \|\cdot\|_2)$) *Let $\|\cdot\|_2$ be a mapping from $\mathbb{R}^3$ to $\mathbb{R}$, defined by:*

$$\mathbf{x} \mapsto \|\mathbf{x}\|_2 \stackrel{\text{def}}{=} \sqrt{\sum_{i=1}^{3} x_i^2}, \tag{A.45}$$

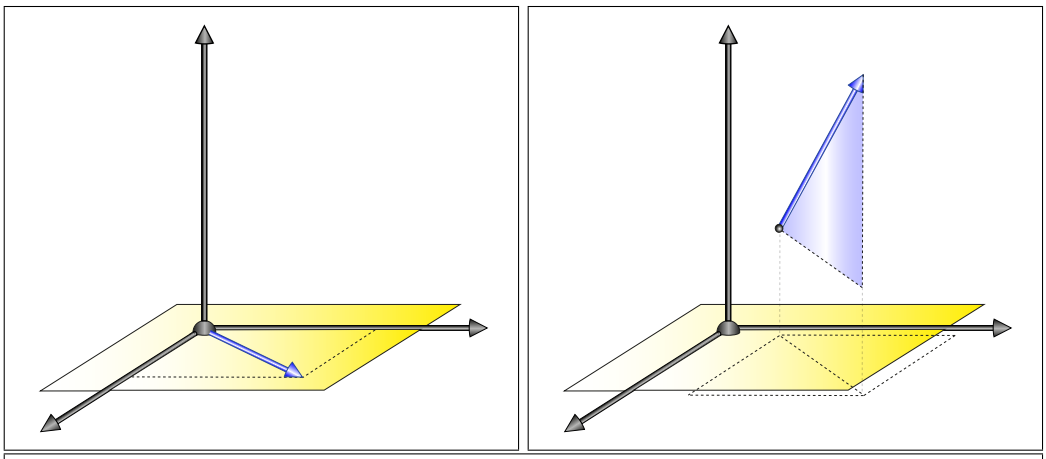


FIGURE A.13: NORM OR LENGTH OF A VECTOR. The images show a two and a three-dimension vector. The norm of the two-dimensional vector can be computed via the x and the y-coordinates of the vector. For computing the norm of a three-dimensional vector, also the z-coordinate of the vector is required.

then $\|\cdot\|_2$ *is called the* length *or the* Euclidean norm *of vector* $\mathbf{x} \in \mathbb{R}^3$, *if* $\|\cdot\|_2$ *satisfies the following axioms:*

i) $\|\mathbf{x}\|_2 \geq 0$ *and* $\|\mathbf{x}\|_2 = 0$ *iff* $\mathbf{x} = \mathbf{0}$ *(positive-definiteness)*

ii) $\|\alpha\,\mathbf{x}\|_2 = \alpha\,\|\mathbf{x}\|_2$ *with* $\alpha \in \mathbb{R}$ *(homogenity)*

ii) $\|\mathbf{x}+\mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2$ *(triangle inequality).*

As the space $\mathbb{R}^3$ *is endowed with the norm* $\|\cdot\|_2$, *we call* $\mathbb{R}^3$ *a* linear normed space *denoted by the tuple* $(\mathbb{R}^3, \|\cdot\|_2)$.

EXAMPLE A.9 *It can easily be shown, that all vectors from* $\mathbf{E}$ *have length* 1, *i.e.* $\|\mathbf{e}_i\|_2 = 1$ *for* $1 \leq i \leq 3$. *As any vector with length one is called a* unit vector, *we can say, that the basis* $\mathbf{E}$ *of* $\mathbb{R}^3$ *consists of three orthogonal unit vectors. In mathematics orthogonal unit vectors are also denoted as* orthonormal vectors. *The basis* $\mathbf{E}$ *is then also called an* orthonormal basis *of the Euclidean space* $\mathbb{R}^3$. *Obviously, any vector* $\mathbf{x} \in \mathbb{R}^3$ *can be normalized via* $\frac{1}{\|\mathbf{x}\|_2}\mathbf{x}$.

EXAMPLE A.10 (Normal of a Plane in $\mathbb{R}^3$ and the Tangent Space $T(\mathbf{x})$) *Let* $\mathbf{a}$ *and* $\mathbf{s}$ *be two points and* $\mathbf{N}(\mathbf{s})$ *be a unit vector in* $\mathbb{R}^3$ *starting at* $\mathbf{s}$. *Then we can define a plane* M *passing through* $\mathbf{a}$ *and perpendicular to* $\mathbf{N}(\mathbf{s})$ *as the collection of all points* $\mathbf{x}$, *such that the vector* $\mathbf{x} - \mathbf{a}$ *is orthogonal to* $\mathbf{N}(\mathbf{s})$, *see Figure A.14. According to*

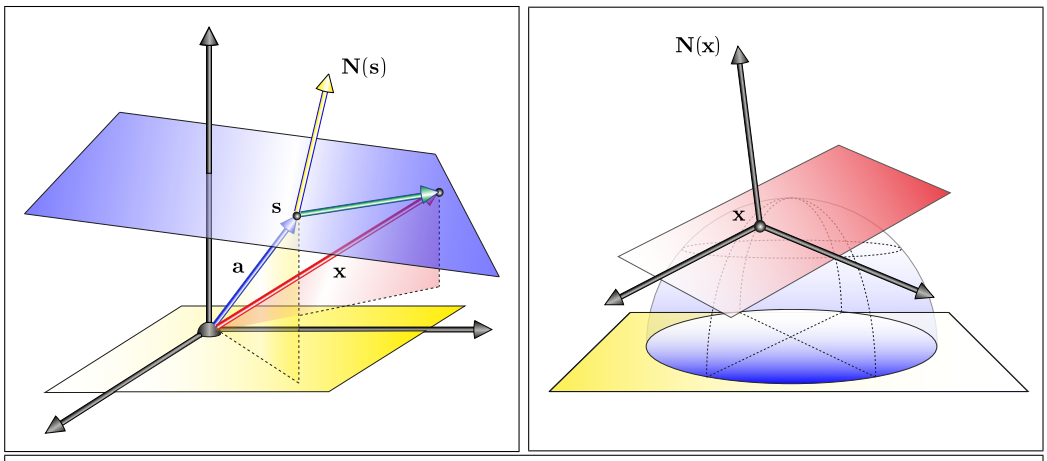


FIGURE A.14: PLANE AND TANGENT SPACE IN $\mathbb{R}^3$. Left, the construction of plane via the vector $\mathbf{a}$ and the normal vector $\mathbf{N}(\mathbf{s})$ at point $\mathbf{s}$. The tangent space at point $\mathbf{x}$ corresponds to all vectors $\mathbf{y}$ that are orthogonal to the normal vector at $\mathbf{x}$.

our definition from above, this corresponds to the condition

$$\langle \mathbf{x} - \mathbf{a}, \mathbf{N}(\mathbf{s}) \rangle_{\mathbb{R}^3} = 0. \tag{A.46}$$

In mathematics, Equation (A.46) is also called the Hesse normal equation *of a plane. Instead of saying that* $\mathbf{N}(\mathbf{s})$ *is orthogonal to* M*, one also says that* $\mathbf{N}(\mathbf{s})$ *is the* normal *of the plane* M *at point* $\mathbf{s}$.

If we now define $T_M(\mathbf{x}) \stackrel{\text{def}}{=} \{\mathbf{y} \in \mathbb{R}^3 | \langle \mathbf{y}, \mathbf{N}(\mathbf{x}) \rangle_{\mathbb{R}^3} = 0\}$*, then* $T_M(\mathbf{x})$ *satisfies the axioms required to a linear space—we omit the simple proof and leave it to the interested reader. It is called the* tangent space *at point* $\mathbf{x}$*, i.e. the space of all vectors in* $\mathbb{R}^3$ *that are perpendicular to the surface normal* $\mathbf{N}(\mathbf{x})$ *at* $\mathbf{x}$*, see Figure A.14.*

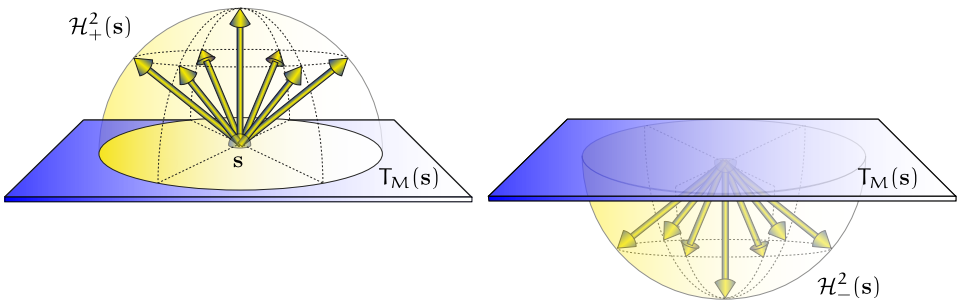
BOX A.3 (Lower and Upper Hemisphere as well as the Unit Sphere)

Suppose a surface point $\mathbf{s}$ is given on a surface $\mathcal{M}$ in $\mathbb{R}^3$ and let S^2 be the unit sphere centered around $\mathbf{s}$. Obviously the tangent space $T_{\mathcal{M}}(\mathbf{s})$ imposes a partitioning of S^2 into two *hemispheres*. Let $\mathbf{N}(\mathbf{s})$ denote the surface normal at $\mathbf{s} \in \mathcal{M}$, and ω be a direction. Then we designate the hemisphere where $\langle \mathbf{N}(\mathbf{s}), \omega \rangle_{\mathbb{R}^3} > 0$ the *positive* or *upper hemisphere* and define it by the set:

$$\mathcal{H}^2_+(\mathbf{s}) \stackrel{\text{def}}{=} \{\omega \in \mathbb{R}^3 | \langle \omega, \mathbf{N}(\mathbf{s}) \rangle_{\mathbb{R}^3} > 0\}. \tag{A.47}$$

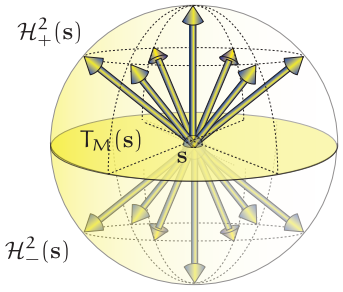
Similarly, we define the *lower hemisphere* by:

$$\mathcal{H}^2_-(\mathbf{s}) \stackrel{\text{def}}{=} \{\omega \in \mathbb{R}^3 | \langle \omega, \mathbf{N}(\mathbf{s}) \rangle_{\mathbb{R}^3} < 0\}. \tag{A.48}$$



Via $\mathcal{H}^2_+(\mathbf{s})$, $\mathcal{H}^2_-(\mathbf{s})$, and the tangent space $T_{\mathcal{M}}(\mathbf{x})$ we can now construct the unit sphere around the surface point $\mathbf{s}$ by:

$$S^2 \stackrel{\text{def}}{=} \mathcal{H}^2_+(\mathbf{s}) \cup T_{\mathcal{M}}(\mathbf{s}) \cup \mathcal{H}^2_-(\mathbf{s}). \tag{A.49}$$



REMARK A.6 (Angle between two Vectors) *In the Euclidean space* $\mathbb{R}^3$ *the inner product* $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^3}$ *is not only related to the length of a vector,* $\|\cdot\|_2$. *Applied to two vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *it can also be used to compute the angle* θ *between these two vectors, since it holds:*

$\|\cdot\|_2$ (846)

$$\theta = \cos^{-1}\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^3}}{\|\mathbf{x}\|_2 \, \|\mathbf{y}\|_2}\right). \tag{A.50}$$

If the vectors $\mathbf{x}$ *and* $\mathbf{y}$ *are normalized, then the dot product gives the cosine of the angle between them, thus:*

$$\cos\theta = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{R}^3}. \tag{A.51}$$

Keep your eyes open, we will use this identity in the following over and over again.

CROSS PRODUCT. Apart from the inner product, there exists another type of a product in and only in the vector space $\mathbb{R}^3$, the so-called *cross product.*

DEFINITION A.10 (Cross Product on $\mathbb{R}^3$) *Let* $\mathbf{x}, \mathbf{y}$ *be two vectors of* $\mathbb{R}^3$. *The* cross product *is a mapping from* $\mathbb{R}^3 \times \mathbb{R}^3$ *to* $\mathbb{R}^3$, *for which applies:*

$$\begin{aligned} \mathbf{x} \times \mathbf{y} &\stackrel{\text{def}}{=} (x_2 y_3 - y_3 x_2, x_3 y_1 - x_1 y_3, x_1 y_2 - x_2 y_1)^T && \text{(A.52)} \\ &= \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \sin\theta \, \mathbf{N}(\mathbf{s}), && \text{(A.53)} \end{aligned}$$

where θ *is the angle between the two vectors* $\mathbf{x}$ *and* $\mathbf{y}$, *and* $\mathbf{N}(\mathbf{s})$ *is a unit vector perpendicular to* $\mathbf{x}$ *and* $\mathbf{y}$ *starting at point* $\mathbf{s}$.

Contrary to the inner product, which assigns two given vectors a real number, the cross product results in a vector, which is orthogonal to the arguments of the mapping. As easily seen, the vector product of two vectors of $\mathbf{E}$ delivers the remaining vector of $\mathbf{E}$, i.e it holds for example: $\mathbf{e}_1 \times \mathbf{e}_2 = \mathbf{e}_3$. In contrast to the inner product, the algebra defined by the cross product is neither associative nor commutative.

REMARK A.7 (Interpretation of the Cross Product as the Area of a Parallelogram) *The cross product has an interesting practical use: It can be interpreted as the positive area of a parallelogram lying on a plane spanned by two vectors* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$.

For that, let us consider Figure A.15, where two vectors $\mathbf{a}$ *and* $\mathbf{b}$, *attached at point* $\mathbf{s}$ *span a plane. Obviously, the parallelogram with the sides* $\|\mathbf{a}\|_2$ *and* $\|\mathbf{b}\|_2$ *has area*

$$A = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \sin\theta \stackrel{(A.53)}{=} \|\mathbf{a} \times \mathbf{b}\|_2 \tag{A.54}$$

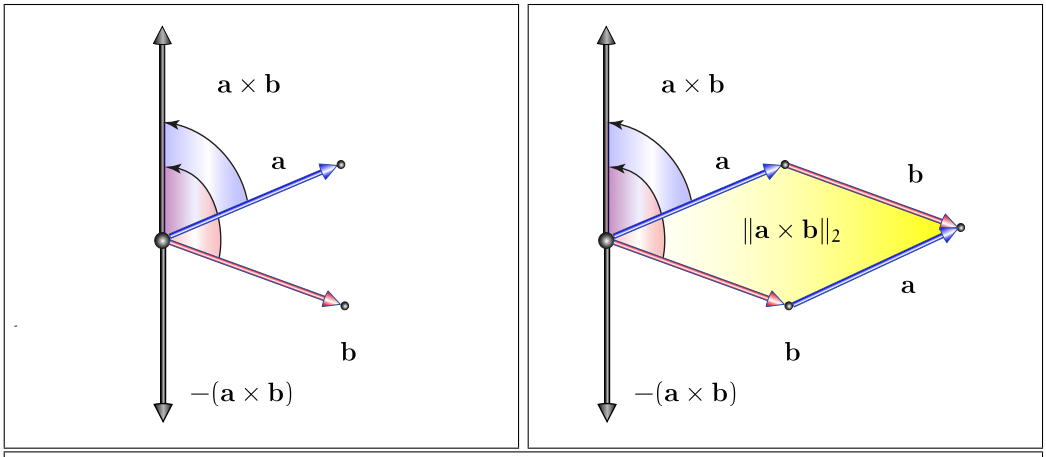


FIGURE A.15: CROSS PRODUCT AND THE AREA OF A PARALLELLOGRAM. The cross product of two vectors $\mathbf{a}, \mathbf{b}$ results in a new vector $\mathbf{a} \times \mathbf{b}$, which is orthogonal to $\mathbf{a}$ and $\mathbf{b}$. The norm of the vector $\mathbf{a} \times \mathbf{b}$ corresponds to the area of a parallelogram spanned by $\mathbf{a}$ and $\mathbf{b}$.

where θ is the angle between the two vectors $\mathbf{a}$ and $\mathbf{b}$. We leave the proof of this simple statement to the interested reader as an exercise.

We use this property of the cross product when deriving the solid angle measures. Thus, we will cover the unit sphere with a net of infinitesimal small parallelograms and will use this parallelograms—analog to the derivation of the area Lebesgue measure via rectangles—to define the solid angle and the projected solid angle measure. Section 2.2.2

METRIC. The final geometric property of $\mathbb{R}^3$ that we wish to introduce, is the concept of the *metric*. In mathematics, a metric is a function, which can be used to determine the distance between elements of a set. Equipped with a metric, we are able to study the convergence behavior of sequences of vectors and to decide, if a given sequence of vectors converge towards an element of $\mathbb{R}^3$. Section A.3

DEFINITION A.11 (The Metric Δ on $\mathbb{R}^3$) *Let $\mathbf{x}, \mathbf{y}, \mathbf{z}$ be three vectors of $\mathbb{R}^3$. A* metric *on $\mathbb{R}^3$ is a mapping Δ from $\mathbb{R}^3 \times \mathbb{R}^3$ to $\mathbb{R}$, which satisfies the following axioms:*

i) $\Delta(\mathbf{x}, \mathbf{x}) \geq 0$ *and* $\Delta(\mathbf{x}, \mathbf{x}) = 0$ *iff* $\mathbf{x} = \mathbf{0}$ *(positive-definiteness)*

ii) $\Delta(\alpha \mathbf{x}, \mathbf{y}) = \alpha \Delta(\mathbf{x}, \mathbf{y})$ *with* $\alpha \in \mathbb{R}$ *(homogenity)*

ii) $\Delta(\mathbf{x} + \mathbf{y}) \leq \Delta(\mathbf{x}, \mathbf{y}) + \Delta(\mathbf{y}, \mathbf{z})$ *(triangle inequality).*

A set with a metric Δ defined on it is called a metric space.

Contrary to the norm or an inner product, a metric requires less structure on the underlying set for its definition. Rather than generating a metric from scratch, we will use the concept of a norm in order to define a corresponding metric by

$$\Delta(\mathbf{x},\mathbf{y}) \stackrel{\text{def}}{=} \|\mathbf{x}-\mathbf{y}\|_2. \tag{A.55}$$

When a metric Δ is generated via a norm, we say that Δ is *generated by* the norm $\|\cdot\|_2$. Obviously, if we construct a metric Δ on $\mathbb{R}^3$ in accordance with Equation (A.55), then $\left(\mathbb{R}^3,\Delta\right)$ will be a metric space, where Δ is generated by the norm $\|\cdot\|_2$ on $\mathbb{R}^3$.

Obviously, the metric from Equation (A.55) provides information about the distance between two points in $\mathbb{R}^3$, that is, in the metric space $\mathbb{R}^3$ it is now possible to introduce the mathematical concept of *convergence of sequences* with members from $\mathbb{R}^3$, which will be discussed explicitly in one of the following section.

Section A.3
Section 2.1.1

Operator (835)

LINEAR MAPPINGS. Let us recall the concept of the operator introduced in the last section. In Example A.8 we constructed an operator $\mathbf{T}$ between the Euclidean spaces $\mathbb{R}^3$ and $\mathbb{R}^2$, who mapped a triple $\mathbf{x} = (x_1,x_2,x_3)^T$ onto the tuple $\mathbf{Tx} = (x_1,x_2)^T$. Evidently, this mapping drops the 3^{rd} component of the given vector $\mathbf{x} = (x_1,x_2,x_3)^T$, resulting in a 2-dimensional vector.

With our knowledge about vector spaces from this section, we can now identify the operator from above as the 2-dimensional projection of a 3-dimensional vector, for an illustration see Figure A.16. The projection from $\mathbb{R}^3$ to $\mathbb{R}^2$ is then described by the following construct, a so-called *matrix*

$$\mathbf{T} \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \tag{A.56}$$

with

$$\mathbf{x} \mapsto \mathbf{Tx}, \tag{A.57}$$

defined via the so-called *matrix-vector product*,

$$\mathbf{Tx} \stackrel{\text{def}}{=} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \tag{A.58}$$

$$= \begin{pmatrix} 1\cdot x_1 + 0\cdot x_2 + 0\cdot x_3 \\ 0\cdot x_1 + 1\cdot x_2 + 0\cdot x_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2. \tag{A.59}$$

Chapter 10

REMARK A.8 *Finite element methods for solving the radiosity equations are based on such a type of projection methods described by operators between special linear spaces.*

Due to its definition, our projection operator $\mathbf{T}$ satisfies the property,

$$\mathbf{T}(\mathbf{x}+\mathbf{y}) = \mathbf{Tx}+\mathbf{Ty}, \tag{A.60}$$

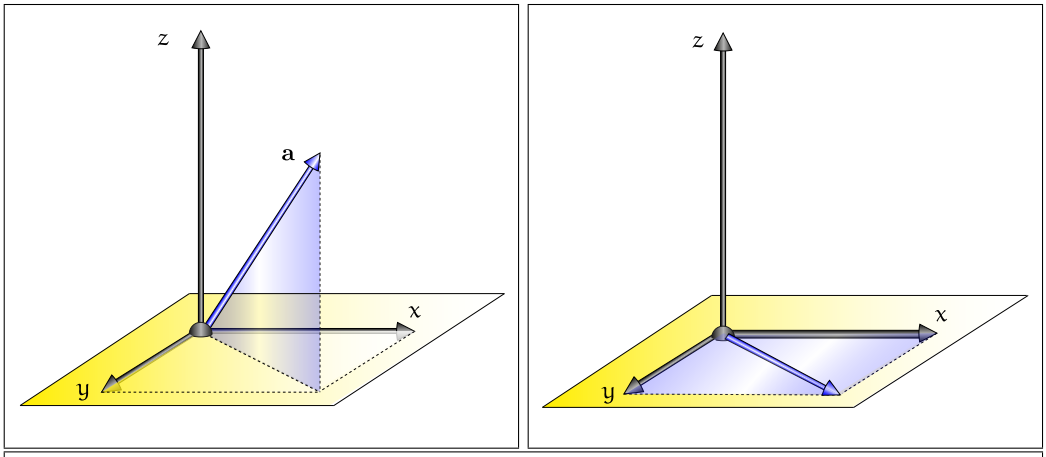


FIGURE A.16: A PROJECTION OPERATOR. A three-dimensional vector and its projection onto the xy-plane. One can easily, that the x and y-coordinates of $\mathbf{a}$ remains unchanged, while the z-coordinate of $\mathbf{a}$ is dropped.

what can easily be proofed via the definition of the matrix-vector product. Operators satisfying this property are called *linear operators*—the proof is easy and is left to the reader as an exercise. As the example of the exponential function $f(x) = e^x$ shows—$e^{x+y} = e^x\, e^y \neq e^x + e^y$—not all operators are linear operators.

DEFINITION A.12 (Linear Mappings on $\mathbb{R}^2$ resepctively $\mathbb{R}^3$) *In the following, $\mathbb{S}$ denotes the linear spaces $\mathbb{R}^2$ and $\mathbb{R}^3$ resepctively. Let $\mathbf{T}$ be a rule, that maps an element $\mathbf{x}$ of $\mathbb{S}$ to an element $\mathbf{Tx} \in \mathbb{S}$, then this rule is called a* linear mapping *between the linear space $\mathbb{R}^2$ respectively $\mathbb{R}^3$, if $\forall\, \mathbf{x}, \mathbf{y} \in \mathbb{S}$ and $\forall\, \alpha \in \mathbb{R}$ it holds:* Section 2.1.4

$$\mathbf{T}(\alpha\,\mathbf{x} + \beta\,\mathbf{y}) \quad = \quad \alpha\,\mathbf{Tx} + \beta\,\mathbf{Ty} \quad (\textit{linearity}). \tag{A.61}$$

In the above definition, the operator $\mathbf{T}$ is described by a so-called $n \times m$ *matrix*, i.e., a scheme of numbers of the form:

$$\begin{pmatrix} s_{11} & \dots & s_{1m} \\ \vdots & \dots & \vdots \\ s_{n1} & \dots & s_{nm} \end{pmatrix} \tag{A.62}$$

with $s_{ij} \in \mathbb{R}, 1 \leq i \leq n, 1 \leq j \leq m$ and $n, m \in \{2, 3\}$.

As already mentioned above, the linearity property of an operator from Equation (A.61) preserves the operations of vector addition and s-multiplication. Operators, that satisfies this property, are build the core of functional analysis, since they can be used to transform very complicated equation resulting from real-world processes, such as differential or integral equations, into very simple and easily solvable linear equations. Section 2.1 Section 2.1.4 Section 2.1.5

EXAMPLE A.11 *As shown above, the projection operator* $\mathbf{T}$ *from Example A.8 is a linear operator. But* $\mathbf{T}$ *is not an invertible linear operator, since it is not injective. So,* $\mathbf{T}$ *maps two different vectors* $\mathbf{x} = (x_1, x_2, x_3)^T$ *and* $\mathbf{y} = (x_1, x_2, y)^T$ *with* $y \neq x_3$ *onto the same vector* $\mathbf{Tx} = (x_1, x_2)^T$. *Since* $\mathbf{T}$ *is indeed a surjective but not an injective mapping,* $\mathbf{T}$ *is even not a bijective operator.*

A.3 ABSTRACT LINEAR SPACES

Vector Space (842) In order to make use of the concept of the vector as a directed ray starting at a point in a specified direction, and apply it to concrete applications, it was necessary to equip the Norm (846) set of vectors with additional structures. This results most importantly in the definition Cross Product (850) of the vector space, the length, and distance of vectors as well as of the cross product and Inner Product (845) the inner product.

However, if we extend our attention from vector sets to include generally arbitrary sets, then it is likely, that without additional structures the analyst will find these very sterile and rather inappropriate as bases of concrete analytical problems. The question then follows is: Which properties must be added to these sets in order to enable them to provide us with a sufficiently interesting and useful theory for the analysis of concrete applications? The key to this problem is to be found in the concept of the *abstract linear space*.

We start in this section by defining the mathematical construct of the *abstract linear space* and show, that a linear space can represented as a sum of *linear subspaces*. Based on the notions of the *linear combination, basis*, and *inner product* we introduce *orthonormal sets* and *bases* in *finite-dimensional* and *infinite-dimensional spaces*. After that, we present—via the *norm* in a linear space—a few useful properties of sets in linear spaces, which we need for the development of the Lebesgue integral in Section 2.2.4, these are: *bounded, open*, and *closed sets, covers* as well as the *supremum* and *infimum* of a set, and the concept of the *boundary-* and *accumulation point*. Because the mathematical construct of the sequence is fundamental in functional analysis, we also define *metric spaces*, and speak about the *limit of sequences* in metric spaces.

Vector Space (842) ABSTRACT LINEAR SPACES. The normed Euclidean vector space $\mathbb{R}^3$ has been an intuitive guide in our development thus far. Now, it is not required, that the elements of a linear $\mathbb{R}, \mathbb{C}$ (827) space are tuples or triples of numbers from $\mathbb{R}$ or $\mathbb{C}$ as in the case of $\mathbb{R}^3$. Elements of linear spaces can also be polynomials, solutions of differential or integral equations, or general mappings between given sets. To construct and analyze such linear spaces it is required to introduce the general concept of the *abstract linear space*.

DEFINITION A.13 (Abstract Linear Space) *Let* S *be a set, and* $\mathbb{K}$ *be either the set of real* $\mathbb{R}, \mathbb{C}$ (827) *numbers or the set* $\mathbb{C}$ *of complex numbers, which both are being referred to here as*

scalars. Then $\mathcal{S}$ *is called an* abstract linear space*, or briefly a* linear space*, if it can be combined with an operation* $+$ *called* addition *and an operation of* multiplication by a scalar*, the* s-multiplication *for short, and satisfies the following axioms:*

i) $(\mathcal{S},+)$ *is an* Abelian group*, i.e. the operation* $+$ *is commutative, associative and there exists exactly one zero element* 0*, and for every* x *of* $\mathcal{S}$*, there exists exactly one inverse element* $-x$*, with* $0+x=x+0=x$ *and* $x+(-x)=(-x)+x=0$

ii) $(\alpha\cdot\beta)x=\alpha(\beta x),\ \forall\, x\in\mathcal{S},\alpha,\beta\in\mathbb{K}$

iii) $(\alpha+\beta)x=\alpha x+\beta x,\ \forall\, x\in\mathcal{S},\alpha,\beta\in\mathbb{K}$

iii) $\alpha(x+y)=\alpha x+\alpha y,\ \forall\, x,y\in\mathcal{S},\alpha\in\mathbb{K}$

iv) $(1\cdot x)=x,\ \forall\, x\in\mathcal{S}$.

When $\mathbb{K}$ *is chosen to be the set of real numbers, then we call* $\mathcal{S}$ *a* real linear space*, while it is referred to as a* complex linear space*, if* $\mathbb{K}=\mathbb{C}$.

EXAMPLE A.12 (The Linear Space $\mathbb{R}^n$) *The set* $\mathbb{R}^n$ *is given by* $\{\mathbf{x}\,|\,\mathbf{x}=(x_1,x_2,\ldots,x_n)\in\mathbb{R}^n,\ x_i\in\mathbb{R},1\le i\le n\}$*. Obviously* $(\mathbb{R}^n,+,\cdot)$ *is a real linear space because* $(\mathbb{R}^n,+,\cdot)$ *satisfies the axioms of Definition A.13, with* $\mathbf{0}=(0,0,\ldots,0)$ *as the zero element,* $-\mathbf{x}=(-x_1,-x_2,\ldots,-x_n)$ *as the inverse to any* $\mathbf{x}\in\mathbb{R}^n$ *as well as* $+$ *as the component-wise addition and* $\cdot$ *as the component-wise s-multiplication. We leave the proof as an exercise to the interested reader.*

EXAMPLE A.13 (The Linear Space of Polynomials of Degree $n-1$ on the Interval $[0,1]$) *Considering the coordinates of a vector* $\mathbf{a}\in\mathbb{R}^n$ *as the coefficients of a polynom of degree* $n-1$ *defined on* $\mathbb{R}$*, then the vector* $\mathbf{a}$ *can be interpreted as a polynomial of the form* $\sum_{i=1}^{n}a_ix^{i-1}$ *in the variable* $x\in\mathbb{R}$*; thus, the vector* $\mathbf{a}=(1,2,\ldots,n)^T$ *represents the polynomial* $\sum_{i=1}^{n}ix^{i-1}$*. Denoting the set of all polynomials of degree* $n-1$ *by* $\mathcal{P}_{n-1}$*, then with* $(\mathbb{R}^n,+,\cdot)$ *also* $(\mathcal{P}_{n-1},+,\cdot)$ *satisfies the axioms of a real linear space, where* 0 *is the zero polynomial and* $\sum_{i=1}^{n}-a_ix^{i-1}$ *is the inverse polynomial to the polynomial* $\sum_{i=1}^{n}a_ix^{i-1}$*, see Figure A.17,*

Generally, arbitrary subsets of linear spaces are not linear spaces, but if a subset of a linear space is closed with respect to additon and s-multiplication, then such a subset is also a *linear subspace*.

DEFINITION A.14 (Linear Subspace) *A subset* $\mathcal{S}'$ *of a linear space* $\mathcal{S}$*, which is also a linear space, is called a* linear subspace *and we write* $\mathcal{S}'\le\mathcal{S}$.

If $\mathcal{S}'$ and $\mathcal{S}''$ are subspaces of a linear space $\mathcal{S}$, then we can construct the *sum* $\mathcal{S}'+\mathcal{S}''$ of these subspaces, to be the set of all elements of $\mathcal{S}$ of the form $x'+x''$ with $x'\in\mathcal{S}'$ and $x''\in\mathcal{S}''$. According to this construction, we call the linear space $\mathcal{S}$ the *direct sum* of $\mathcal{S}'$ and $\mathcal{S}''$, denoted by $\mathcal{S}'\oplus\mathcal{S}''$, if it holds $\mathcal{S}=\mathcal{S}'+\mathcal{S}''$ and $\mathcal{S}'\cap\mathcal{S}''=\mathbf{0}$.

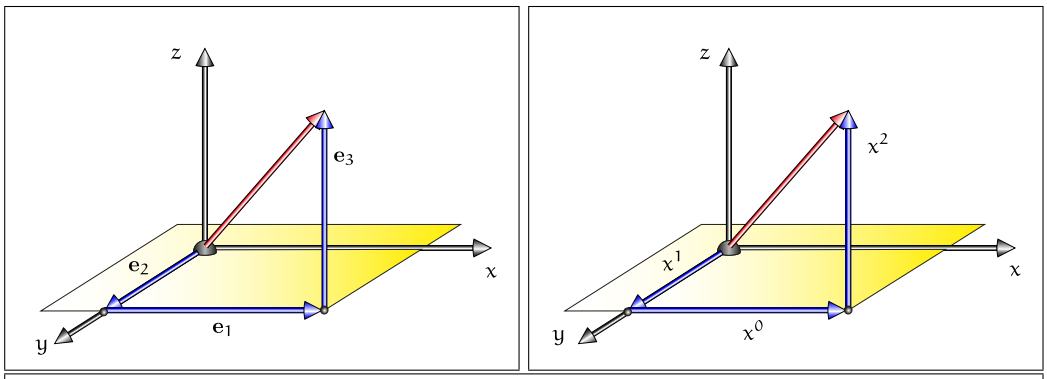


FIGURE A.17: VECTOR SPACE ISOMORPHISM BETWEEN $\mathbb{R}^3$ AND $\mathcal{P}_3$. A vector representing an element of space $\mathbb{R}^3$. It can be used to represent a polynomial of space $\mathcal{P}_3$, where the basis $E = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is replaced by the powers of x, namely: x^0, x^1, x^2.

EXAMPLE A.14 *Since $\mathbb{R}^2$ is not a subset of $\mathbb{R}^3$—the set of all tuples over $\mathbb{R}$ is not contained in the set of all triples over $\mathbb{R}$—it does not hold: $\mathbb{R}^2 \leq \mathbb{R}^3$. Considering the set $\mathbb{R}'^2 \stackrel{\text{def}}{=} \{(x_1, x_2, 0) \,|\, x_i \in \mathbb{R}, 1 \leq i \leq 2\}$, thus the set of all points from $\mathbb{R}^3$ lying on the plane $x_3 = 0$ through the origin, see Figure A.18, then it is easily to see, that $\mathbb{R}'^2$ satisfies the axioms of a real linear space. Since $\mathbb{R}'^2 \subset \mathbb{R}^3$ we get: $\mathbb{R}'^2 \leq \mathbb{R}^3$. Obviously every set $\mathbb{R}'^m \stackrel{\text{def}}{=} \{(x_1, x_2, \ldots, x_m, \underbrace{0, \ldots, 0}_{(n-m)-\text{times}}) \,|\, x_i \in \mathbb{R}, 1 \leq i \leq m\}$, is a linear subspace of $\mathbb{R}^n$, i.e., $\mathbb{R}'^2 \leq \mathbb{R}'^3 \leq \mathbb{R}^n$.*

$\mathbb{R}^3$ (841)

Furthermore, any element $x \in \mathbb{R}^n$ can be written as $x'^m + x'^{n-m}$, with $x'^m \in \mathbb{R}'^m$ and $x'^{n-m} \in \mathbb{R}'^{n-m} = \{(\underbrace{0, \ldots, 0}_{(m-\text{times})}, x_{m+1}, \ldots, x_n) \,|\, x_i \in \mathbb{R}, m+1 \leq i \leq n\}$. Because it holds $\mathbb{R}'^m \cap \mathbb{R}'^{n-m} = \mathbf{0}$, we can conclude: $\mathbb{R}^n = \mathbb{R}'^m \oplus \mathbb{R}'^{n-m}$.

This technique of generating a new linear space from two given linear spaces will be used in Chapter 5. There we are interested in the development of mathematical models of light and importance transport over a linear normed function space.

Function Space (28)

Now it is possible to generalize many of the fundamental concepts of the Euclidean space $\mathbb{R}^3$ from the preceeding section to abstract linear spaces. For that, we will start first with the concepts of the *linear combination*, the *basis*, the *inner product*, and the *orthogonality*, and introduce then, in analogy to the length of a vector, the concept of a *norm*, and a *metric*.

LINEAR COMBINATION AND BASIS. To describe a linear space S economically, we are interested in the construction of a subset of elements of S, which can be used to specify

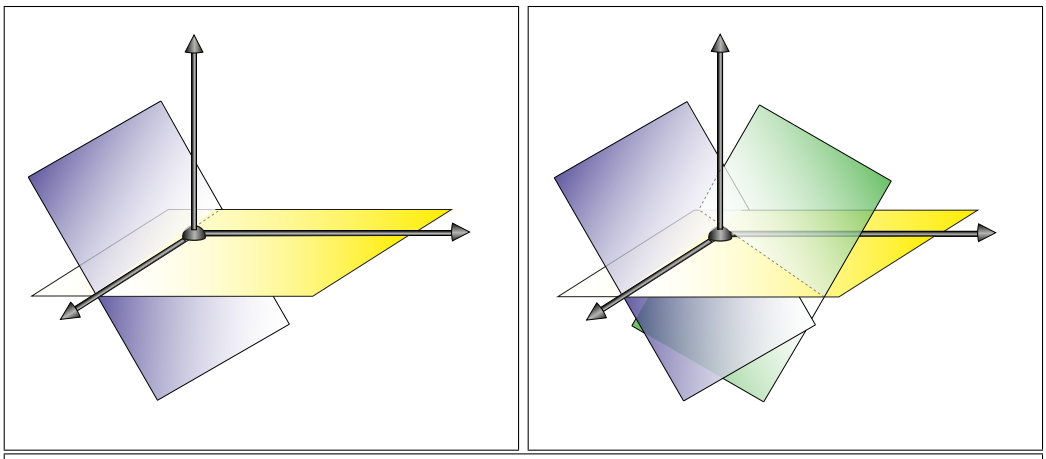


FIGURE A.18: LINEAR SUBSPACES. Planes and lines passing through the origin are subspaces of $\mathbb{R}^n$, since sums of multiples of points of such a plane or a line also lie on the plane or the line.

every member of $\mathcal{S}$ as a combination of these elements. Such a procedure is based on the concepts of the *linear combination* and the *basis*.

DEFINITION A.15 (Linear Combination) *Let* $B = \{b_1, b_2, \ldots, b_n\}$ *be a finite set of elements of a linear space* $\mathcal{S}$ *and* x *be an element of* $\mathcal{S}$. *We call* x *a* linear combination *of elements of* B, *if it may be represented as a weighted sum of the elements* $b_i, 1 \leq i \leq n$, *that is,*

$$x = \sum_{i=1}^{n} \alpha_i b_i, \tag{A.63}$$

with $\alpha_i \in \mathbb{K}$ *and* $\mathbb{K} = \mathbb{R}$ *or* $\mathbb{C}$.

DEFINITION A.16 (Basis) *The set* B *from above is called a* basis *of* $\mathcal{S}$, *if every element* $x \in \mathcal{S}$ *may be represented as a linear combination of elements of* B *and none of the elements* $b_i, 1 \leq i \leq n$ *may be formulated as a linear combination of* $b_j, 1 \leq j \leq n, i \neq j$.

If it holds $n < \infty$, *the linear space* $\mathcal{S}$ *is referred to as a* finite-dimensional *linear space, whereas we write:* $\dim \mathcal{S} = n$, *otherwise it is called an* infinite-dimensional *linear space.* Section 2.1.1

EXAMPLE A.15 (The Linear Space $\mathbb{R}^n$) *It is easily seen, that the set* $\mathbf{E} = \{\mathbf{e}_1 = (1, 0, \ldots, 0), \mathbf{e}_2 = (0, 1, \ldots, 0), \ldots, \mathbf{e}_n = (0, 0, \ldots, 1)\}$ *forms a basis of* $\mathbb{R}^n$, *since* $\mathbf{E}$ *is a linearly independent set of vectors, i.e.* $\mathbf{0} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \ldots + \alpha_n\mathbf{e}_n \Leftrightarrow \alpha_i = 0, 1 \leq i \leq n$ Example A.12

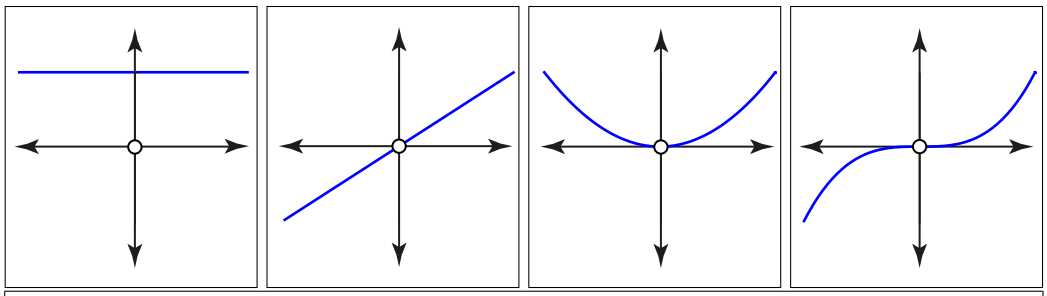


FIGURE A.19: A BASIS FOR THE LINEAR SPACE OF POLYNOMIALS OF DEGREE 3. The four functions 1, x, x^2, and x^3 build a basis of the linear space $\mathcal{P}_3$ of polynomials of degree 3.

and any element $\mathbf{x} \in \mathbb{R}^n$ *can be written as a linear combination of elements of* $\mathbf{E}$. *Applying component-wise addition and the multiplication with scalars leads to:* $\mathbf{x} = (x_1, x_2, \ldots, x_n) = x_1(1, 0, \ldots, 0) + x_2(0, 1, \ldots, 0) + \ldots + x_n(0, 0, \ldots, 1)$. *As this basis consists of* n *vectors one says the linear space* $\mathbb{R}^n$ *has dimension* n, *thus* $\dim \mathbb{R}^n = n$.

Example A.13 EXAMPLE A.16 (The Linear Space of Polynomials of Degree $n-1$, $\mathcal{P}_{n-1}$) *Suppose* $p(x) = \sum_{i=1}^{n} \alpha_i x^{i-1} \in \mathcal{P}_{n-1}$, *then* p *can be identified with the vector* $(\alpha_1, \alpha_2, \ldots, \alpha_n) = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \ldots + \alpha_n \mathbf{e}_n$ *where* $\mathbf{e}_i \in \mathbf{E} = \{1, x, x^2, \ldots, x^{n-1}\}, 1 \leq i \leq n$. *From this, we conclude, that any polynomial of degree* $n-1$ *can be represented by a linear combination of elements of set* $\mathbf{E}$. *Since the set* $\mathbf{E}$ *is linearly independent,* $\mathbf{E}$ *forms a basis of the linear space* P_{n-1} *and it holds:* $\dim \mathcal{P}_{n-1} = n$. *For a visualization of the basis functions of* $\mathcal{P}_3$, *see Figure A.19.*

REMARK A.9 *If we have a basis of a linear space* S, *then we can control the linear space, since the basis already contains all essential information about* S. *This implies, that the basis can not be more compressed.*

REMARK A.10 *As we will see later, the spaces of greatest interest underlying realistic rendering procedures are infinite-dimensional function spaces. Further below, we will present with* $C([0,1])$ *a first simple example of an infinite-dimensional linear function space: the space of all continuous functions over a closed interval. Nevertheless, it will often be of use to consider finite-dimensional spaces.*

Section 2.1.1
Section 2.1.1
Section 2.4

Orthogonality in $\mathbb{R}^3$ (846)
$\mathbb{R}^3$ (841)
Inner Product in $\mathbb{R}^3$ (845)

INNER PRODUCT. Now, we turn to the question: Is it possible to adapt the concept of *orthogonality*—introduced in the last section with respect to the Euclidean space $\mathbb{R}^3$—to general linear spaces, where we cannot argue with its geometrical interpretation? Inspired by the inner product of the Euclidean space, now we define an inner product valid in general abstract linear spaces.

DEFINITION A.17 (The Inner Product $\langle\cdot,\cdot\rangle_S$ and the Inner Product Space $(S,\langle\cdot,\cdot\rangle_S)$) *Let S be a complex or real linear space, the* inner product $\langle\cdot,\cdot\rangle_S$ *on the linear space S is a mapping from $S\times S$ to $\mathbb{K}=\mathbb{R}$ or $\mathbb{C}$, which satisfies the following axioms for any member $x,y,z\in S$ and $\alpha,\beta\in\mathbb{K}$:*

i) $\langle x,y\rangle_S \geq 0$ *and* $\langle x,x\rangle_S = 0$ *iff* $x=0$ *(positive-definiteness)*

ii) $\begin{cases} \langle x,y\rangle_S = \overline{\langle y,x\rangle}_S & \text{if} \quad \mathbb{K}=\mathbb{C} \\ \langle x,y\rangle_S = \langle y,x\rangle_S & \text{otherwise} \end{cases}$ *(symmetry)*

iii) $\langle \alpha x+\beta y,z\rangle_S = \alpha\langle x,z\rangle_S + \beta\langle y,z\rangle_S$ *(linearity),*

where $\overline{\langle\cdot,\cdot\rangle_S}$ *is the* conjugate complex. Conjugate Complex (828)

A linear space S, endowed with an inner product $\langle\cdot,\cdot\rangle_S$, is called an inner product space. *We denote an inner product space S by $(S,\langle\cdot,\cdot\rangle_S)$.* Section 2.1.1

Let $(S,\langle\cdot,\cdot\rangle_S)$ be an inner product space, then a number of important concepts from vector algebra and calculus may be transferred onto it. Thus, in an inner product space $(S,\langle\cdot,\cdot\rangle_S)$, not only the fundamental *Cauchy-Schwartz inequality*[1] holds but also *orthogonality*, the well-known and important concept from Euclidean space $\mathbb{R}^3$, may also be adapted. Thus, two elements $x,y\in S$ are termed *orthogonal*, if it holds: Orthogonality in $\mathbb{R}^3$ (846)

$$\langle x,y\rangle_S = 0, \tag{A.65}$$

in this case we write also $x\perp y$.

EXAMPLE A.17 (The Inner Product $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ and the Inner Product Space $(\mathbb{R}^n,\langle\cdot,\cdot\rangle_{\mathbb{R}^n})$) Example (A.12)
The inner product $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ *of the linear space $\mathbb{R}^n$ is given by:*

$$\langle \mathbf{x},\mathbf{y}\rangle_{\mathbb{R}^n} = (x_1,x_2,\ldots,x_n)\cdot\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \overset{\text{def}}{=} \sum_{i=1}^{n} x_i\cdot y_i. \tag{A.66}$$

If we endow $\mathbb{R}^n$ with the inner product $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ it is called an inner product space, *shortly also $(\mathbb{R}^n,\langle\cdot,\cdot\rangle_{\mathbb{R}^n})$. Due to Equation (A.66) one can easily see, that the vectors $\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n\in\mathbb{R}^n$ are orthogonal, as it holds:*

$$\langle e_i,e_j\rangle_{\mathbb{R}^n} = 0 \tag{A.67}$$

for $i\neq j, 1\leq i,j\leq n$.

[1] For each two elements x,y of the inner product space $(S,\langle\cdot,\cdot\rangle_S)$ the following applies:

$$|\langle x,y\rangle_S| \leq \langle x,x\rangle_S^{\frac{1}{2}}\langle y,y\rangle_S^{\frac{1}{2}} \tag{A.64}$$

Based on the notion of the inner product, now we can define another important concept: the *orthogonal complement* of a set. Together with the construct of the *orthonormal set* as well as the *orthonormal basis*, which we will also introduce further below, it plays a central role in the development of numerical solutions methods for Fredholm integral equations of the 2^{nd} based on projection and finite basis methods.

Section 2.3.3.2.2

DEFINITION A.18 (Orthogonal Complement) *Let $(\mathcal{S}, \langle\cdot,\cdot\rangle_{\mathcal{S}})$ be an inner product space and $\mathcal{S}'$ any subspace of $\mathcal{S}$, then the* orthogonal complement *of $\mathcal{S}'$ is defined to be the set*

$$\mathcal{S}'^{\perp} \stackrel{\text{def}}{=} \{x \in \mathcal{S} \mid \langle x, x'\rangle_{\mathcal{S}} = 0,\ \forall x' \in \mathcal{S}'\}. \tag{A.68}$$

EXAMPLE A.18 *As we have seen in Example A.14, $\mathbb{R}^n$ can be written as the direct sum of the spaces $\mathbb{R}'^m$ and $\mathbb{R}'^{n-m}$, i.e., $\mathbb{R}^n = \mathbb{R}'^m \oplus \mathbb{R}'^{n-m}$. Considering now any members $x'^m = (x_1, \ldots, x_m, 0, \ldots 0)$ and $x'^{n-m} = (0, \ldots 0, x_{m+1}, \ldots, x_n)$, then obviously it holds for the inner product of two vectors: $\langle x'^m, x'^{n-m}\rangle_{\mathbb{R}^n} = 0$. That is, the space $\mathbb{R}'^m$ is the orthogonal complement of $\mathbb{R}'^{n-m}$ and vice versa. Hence we conclude, that the linear space $\mathbb{R}^n$ can be decomposed in two subspaces $\mathbb{R}'^m$ and $\mathbb{R}'^{n-m}$, which are orthogonal to each other, that is, we can write $\mathbb{R}^n = \mathbb{R}'^m \oplus \mathbb{R}'^{n-m}$. The proof of this statement is very easy, hence we leave it as an exercise to the interested reader.*

As we will see a little bit later, it is just this property—namely, the decomposition of a linear space in two mutually orthogonal subspaces—of an inner product space, which we will extend to arbitrary linear spaces, more precisely speaking to Hilbert spaces. *It provides the mathematical basis for the so-called* finite element methods *underlying* radiosity algorithms.

Section 2.1.1
Section 10

NORM. To measure the elements of an abstract linear space, now we will introduce in analogy to the concept of the length of a vector, a real valued non-negative function $\|\cdot\|$ over a linear space $\mathcal{S}$, the so-called *norm*.

Lenght of a Vector in $\mathbb{R}^3$ (846)

As with the definition of an inner product, this notation may be abstracted in a natural way, if we start from scratch with an arbitrary linear space $\mathcal{S}$.

DEFINITION A.19 (The Norm $\|\cdot\|$ and the Linear Normed Space $(\mathcal{S}, \|\cdot\|)$) *A* norm $\|\cdot\|$ *on the linear space $\mathcal{S}$ is a mapping from $\mathcal{S}$ to $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, which satisfies the following properties for any member $x, y \in \mathcal{S}, \alpha \in \mathbb{K}$:*

i) $\|x\| \geq 0$ *and* $\|x\| = 0$ *iff* $x = 0$ *(positive-definiteness)*

ii) $\|\alpha x\| = \alpha\|x\|$ *(homogenity)*

iii) $\|x + y\| \leq \|x\| + \|y\|$ *(triangle inequality)*

Endowed with the norm $\|\cdot\|$ the linear space $\mathcal{S}$ is called a linear normed space, *also written as $(\mathcal{S}, \|\cdot\|)$.*

Obviously, apart positive-definiteness and homogenity, a norm satisfies the triangle equation, which abstracts the situation, that results from the parallelogram law for addition of vectors.

Parallelogram Law (842)

EXAMPLE A.19 (The Linear Normed Space $(\mathbb{R}^n, \|\cdot\|_2)$) *Let us consider the inner product space* $(\mathbb{R}^n, \langle\cdot,\cdot\rangle_{\mathbb{R}^n})$ *from Example (A.17), then a norm on* $\mathbb{R}^n$ *can be defined according to:*

Example A.17

$$\|\mathbf{x}\|_2 \stackrel{def}{=} \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle} = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{A.69}$$

$\|\cdot\|_2$ *is called the* Euclidean norm *on* $\mathbb{R}^n$. *It serves to measure the length of a vector. With this norm, the linear space* $\mathbb{R}^n$ *will become a linear normed space, which we will denote as* $(\mathbb{R}^n, \|\cdot\|_2)$.

EXAMPLE A.20 (The Family of Norms $\|\cdot\|_p$ in $\mathbb{R}^n$) *Apart from the Euclidean norm* $\|\cdot\|_2$, *we can define a whole family of norms in the space* $\mathbb{R}^n$ *by:*

$$\|\mathbf{x}\|_p \stackrel{def}{=} \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, \tag{A.70}$$

where $1 \leq p < \infty$.

Obviously, the case $p = 2$ *corresponds to the Euclidean norm from Example A.19. Another important norm is the case* $p = 1$, *thus,*

$$\|\mathbf{x}\|_1 \stackrel{def}{=} \sum_{i=1}^{n} |x_i|^p. \tag{A.71}$$

We leave the proof, that $\|\cdot\|_p$ *fulfills the requirements to a norm, to the interested reader.*

The concept of the norm can now be used to define an other useful mathematical construct: the notion of the *orthonormal set* and the *orthonormal basis*. In Section 2.3.3.2.2, we will use these concepts to construct functions—defined in finite-dimensional linear function spaces—that are in some sense good approximations to solutions of Fredholm integral equations of the 2^{nd} kind.

DEFINITION A.20 (Orthonormal Set and Orthonormal Basis) *Let* $(\mathcal{S}, \langle\cdot,\cdot\rangle_{\mathcal{S}})$ *be an inner product space. We call the set* $\mathcal{B}_\phi = \{\phi_1, \phi_2, \ldots, \phi_n, \ldots\}$ *of members of* $\mathcal{S}$ *an* orthonormal set, *if it holds:*

$$\langle \phi_i, \phi_j\rangle_{\mathcal{S}} = 0 \tag{A.72}$$

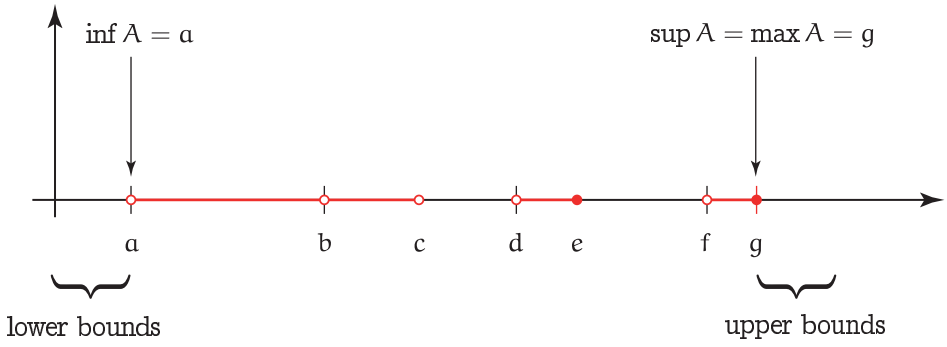


FIGURE A.20: A BOUNDED SET. The set A as the union of the open intervals $(a, b), (b, c)$ and the half-open intervals $(d, e]$ and $(f, g]$. Obviously, A is a bounded set with $\inf A = a$ and $\sup A = \max A = g$. While the set of lower bounds is given by all reals equals or smaller than a, the upper bounds are all numbers equals or greater than the maximum of A, namely g.

for $i \neq j$ *with* $\|\phi_i\| = 1$ *for all* $i, j \geq 1$.

If $\mathcal{S}$ *is of finite dimension, thus* $\dim \mathcal{S} = n$, *then we call the orthonormal set* $\mathcal{B}_\phi = \{\phi_1, \ldots, \phi_n\}$ *an orthonormal basis of* $\mathcal{S}$. *As a basis, the elements of* $\mathcal{B}_\phi$ *span the space* $\mathcal{S}$, *such that any element of* $\mathcal{S}$ *can be written as a linear combination of orthonormal members of* $\mathcal{B}_\phi$.

Section 2.1.1

Example A.15 EXAMPLE A.21 (Orthonormal Basis of $\mathbb{R}^n$) *A trivial orthonormal basis of the inner product space* $(\mathbb{R}^n, \langle \cdot, \cdot \rangle_{\mathbb{R}^n})$ *is the basis* $\mathbf{E}$ *introduced in Example A.15. We have seen in Example A.17, that in each case, two members of* $\mathbf{E}$ *are orthogonal. Since it holds* $\|e_i\|_2 = 1$, $1 \leq i \leq n$, *the set* $\mathbf{E}$ *is also a collection of orthonormal vectors, which spans* $\mathbb{R}^n$.

SOME USEFUL PROPERTIES OF SETS BASED ON THE NORM. Many properties of sets, which are relevant in our discussions, are based on the norm. In the following we will introduce the most important for our interest, where we always assume, that A is any arbitrary non-empty subset of a linear normed space $(\mathcal{S}, \|\cdot\|)$.

Section 2.2 DEFINITION A.21 (Bounded Set, Supremum and Infimum of a Set) *We say* A *is a* bounded set, *if for all elements* x *of* A *there exists a positive real number* S, *such that* $\|x\| \leq S$. *In that case,* S *is called an* upper bound, *and* $-S$ *is referred to as a* lower bound *of* A, *see Figure A.20. For the* least upper *and the* greatest lower bound *of a set* A, *we will use the well-known notions of the* supremum, sup A, *and the* infimum, inf A.

It should be known, that the supremum and the infimum of a bounded set A, must

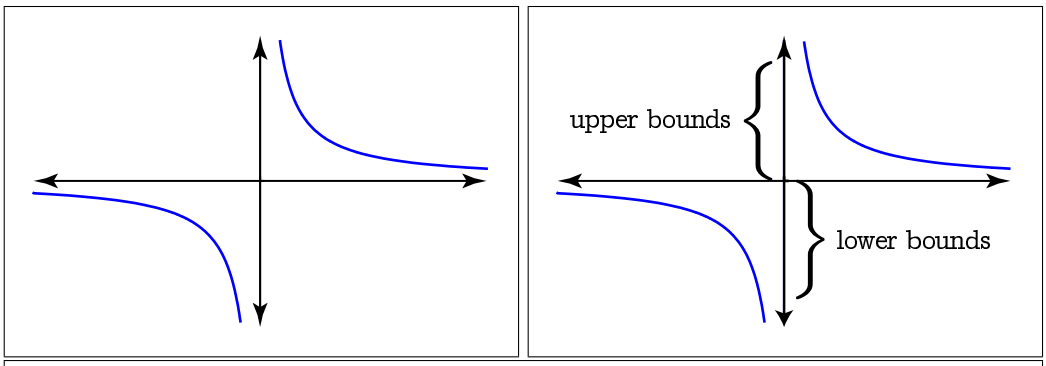


FIGURE A.21: BOUNDED FUNCTIONS. The real-valued function $f(x) = \frac{1}{x}$ in the left image is unbounded on its domain $\mathbb{R} \setminus \{0\}$. The restrictions $f|_{\mathbb{R}^+}$ and $f|_{\mathbb{R}^-}$, visualized on the right are bounded above respectively bounded below with $\inf_{\mathbb{R}^+} f = \sup_{\mathbb{R}^-} f = 0$.

not be necessarily elements of A. But if it holds, that these quantities are contained in the set A, then we call $\sup A$, the *maximum* of A, i.e. $\max A = \sup A$, and $\inf A$, the *minimum* of A, thus $\min A = \sup A$.

EXAMPLE A.22 (Bounded Intervals in $\mathbb{R}$) *Let us consider the open interval* (a, b), $a, b \in \mathbb{R}$. *Obviously* (a, b) *is a bounded set, with* $\frac{a}{2}$ *as a lower and* $2b$ *as an upper bound. The infimum of* (a, b) *is the element* $a \notin (a, b)$ *and for the supremum it holds:* $\sup(a, b) = b$, *that is,* $\inf(a, b) \notin (a, b)$ *as well as* $\sup(a, b) \notin (a, b)$. *Contrary to this, it holds for the closed interval* $[a, b]$*:* $\inf[a, b] = \min[a, b] = a$ *and* $\sup[a, b] = \max[a, b] = b$. Intervals in $\mathbb{R}$ (829)

DEFINITION A.22 (Bounded Function) *In the special case, that the set* A *is the range of a function* f*, we say* f *is a* bounded function*, if and only if* $A = \mathrm{Ran}(f)$ *is a bounded set.* Section 2.1.4

EXAMPLE A.23 (Bounded Functions) *i) Obviously, the characteristic function* χ_B *is a bounded function, as it holds:* $\|\chi_B(x)\| \leq 1 \Leftrightarrow -1 \leq \chi_B(x) \leq 1$ *with* $\sup \chi_B = 1$ *as well as* $\inf \chi_B = 0$. χ_B (839)

ii) On the other side, it's also quite plain, that the function $f : \mathbb{R} \setminus \{0\} \to \mathbb{R} \setminus \{0\}$ *which maps* x *to* $\frac{1}{x}$ *has neither a minimum nor a maximum on its domain, see Figure A.21. If we restrict our considerations only to the negative real line, then* f *is bounded above by all non-negative real numbers, that is:* $\sup_{\mathbb{R}^-} f = 0$. *Something near it occurs, if we restrict our considerations to the positive real line. Here, the function* f *is bounded below by all non-positive real numbers, and it holds:* $\inf_{\mathbb{R}^+} f = 0$. *Even though* $0 \notin \mathrm{Ran}(f)$*, we have* $\inf_{\mathbb{R}^+} f = \sup_{\mathbb{R}^-} f = 0$.

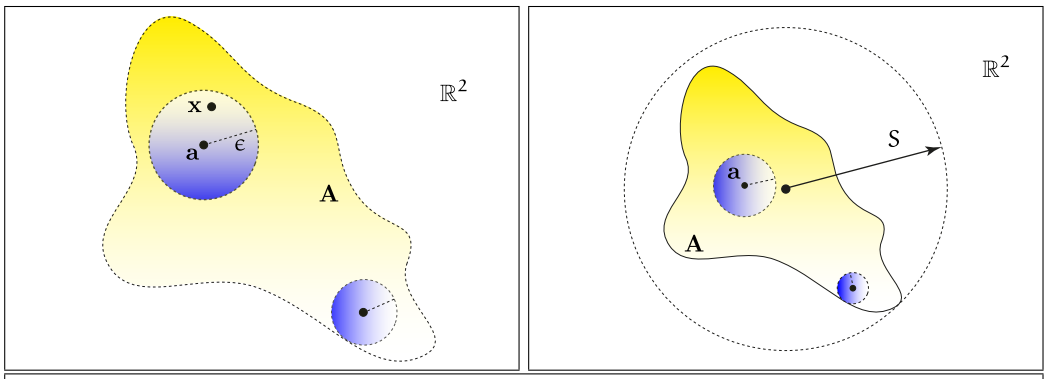


FIGURE A.22: OPEN, CLOSED AND BOUNDED SETS. The left image shows an open set $A \subset \mathbb{R}^2$ with two labeled inner points. The image on the right-hand side represents a closed and bounded set B with labeled inner and boundary points. Obviously, the set A is bounded, because there exists a bound S—select S to be the radius of a circle, which contains A—such that $\forall \mathbf{x} \in A$ it holds: $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2} < S$.

Section 2.1.2

DEFINITION A.23 (Interior Points and Boundary Points of a Set) *A point* $\mathbf{x}$ *of* A *is referred to as an* interior point *of* A*, if there exists a ball around* $\mathbf{x}$ *with radius* $\epsilon > 0$*, which lies completly in* A*. That is, in addition to* $\mathbf{x}$ *the ball contains only points of* A*, for which the following holds:*

$$\{\mathbf{a} \in A \,|\, \|\mathbf{x} - \mathbf{a}\| < \epsilon\} \subset A. \tag{A.73}$$

The point $\mathbf{x}$ *is denoted as a* boundary point *of* A*, if every ball around* $\mathbf{x}$ *contains at least one point out of the complement of* A.

Complement of a Set (827)

Based on the definition of the interior point, we are now ready, to address ourselves to the topological concept of open and closed sets. They are the basis for the definition of the concept of the Borel[2] σ-algebra, which is fundamental for measure, integration, and probability theory.

Borel σ-algebra (865)

Section 2.2.2

DEFINITION A.24 (Open and Closed Set) *A non-empty set* A *is denoted as* open*, if the set contains only interior points. The set* A *is referred to as* closed*, if its complement is open, see Figure A.22.*

[2]Named after *Émile Borel*, who implicitly introduced 1898 in his *Leçon sur la théorie des fonctions* the Borel subsets of the unit interval and denoted, that it is possible to define the notion of length for these sets, which possesses the fundamental property of a σ-algebra.

REMARK A.11 *An equivalent definition for closed sets is based on the concept of the point of accumulation. Therefore, we say a point* $\mathfrak{a}$ *of a non-empty set* A *is denoted as a* point of accumulation, *if every open ball around* $\mathfrak{a}$, *no matter how small, contains at least one point of* A, *which is different from* $\mathfrak{a}$. *If a non-empty set* A *contains all of its points of accumulation, then we call* A closed.

DEFINITION A.25 (Borel σ-Algebra) *Let* $\mathfrak{R}$ *be the collection of all* σ*-algebras containing all open intervals of* $\mathbb{R}$, *then the intersection of these* σ*-algebras, written as:* Section 2.2.1 σ-algebra (828)

$$\mathfrak{B}(\mathbb{R}) \stackrel{\text{def}}{=} \bigcap \mathfrak{R}, \tag{A.74}$$

is denoted as the Borel σ-algebra *generated by all open intervals of* $\mathbb{R}$. *The elements of* $\mathfrak{B}(\mathbb{R})$ *are called the Borel sets of real numbers.*

As one can see by means of the construction of $\mathfrak{B}(\mathbb{R})$, all open intervals belong to $\mathfrak{B}(\mathbb{R})$, and since $\mathfrak{B}(\mathbb{R})$ is a σ-algebra, all open sets—as countable union of open intervals—are Borel sets. With open sets, closed sets as complements of open sets are also Borel sets, etc. Since each countable set is a countable union of closed intervals of the form $[a, a], a \in \mathbb{R}$, countable sets are as well Borel sets. Similarly we can argue for half-open and closed intervals, they are all Borel sets since it holds: $[a, b) = \{a\} \cup (a, b)$ and $[a, b] = \{a\} \cup (a, b) \cup \{b\}$. In particular $\mathbb{N}, \mathbb{Z}$ and $\mathbb{Q}$ are the Borel sets of natural, integer and rational numbers. It is easily seen, that the set of irrational numbers—as the complement of $\mathbb{Q}$—is also a Borel set. Countability (827)

EXAMPLE A.24 (The Borel σ-algebras: $\mathfrak{B}([0,1])$, $\mathfrak{B}([a,b])$, $\mathfrak{B}([a,b]^n)$, and $\mathfrak{B}(\mathbb{R}^n)$) *Similar to the Definition of the Borel* σ*-algebra* $\mathfrak{B}(\mathbb{R})$, *we define the Borel* σ*-algebras* $\mathfrak{B}([0,1])$, $\mathfrak{B}([a,b])$, $\mathfrak{B}([a,b]^n)$,*and* $\mathfrak{B}(\mathbb{R}^n)$, *as the* σ*-algebras generated by all open intervals of* $[0,1]$, $[a,b]$, $[a,b]^n$ *as well as* $\mathbb{R}^n$. *They play, in particular, a central role in our further considerations about random variables and random vectors.* Section 2.4.2 Section 2.4.3

REMARK A.12 *The definition of the Borel* σ*-algebra is very flexible; as long as we start with all intervals of a particular type, these collections generate the same Borel* σ*-algebra.*

As σ*-algebras may normally not be indicated by directly writing down their elements, they are often defined by indicating a so-called* generator. *In the case of the Borel* σ*-algebra the set of open subsets* $\mathcal{O} \subset \mathbb{R}$ *has pointed out as the generator, which means that* $\mathfrak{B}(\mathbb{R})$ *may be regarded as the smallest* σ*-algebra generated by open subsets* $O \in \mathcal{O}$ *using* $\cup, \cap$ *and complement operations.*

Another important notion useful for understanding the concept of a *measure*, is the mathematical construct of an *open cover*. Measure (79)

DEFINITION A.26 (Open Cover) *In mathematics, an* open cover *of a set* A *is a collection* Section 2.2.1

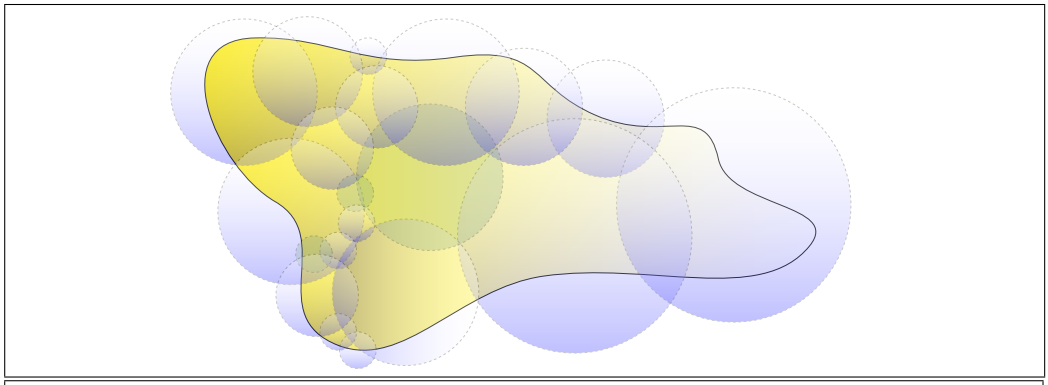


FIGURE A.23: OPEN COVER. An open cover of a closed 2-dimensional set by a finite set of open circles with different radii.

of open sets $\{O_i | i \in I\}$, $I \subseteq \mathbb{N}$, *such that* A *is a subset of the union of* O_i, *that is:*

$$A \subseteq \bigcup_{i \in I} O_i. \tag{A.75}$$

For an illustration of an open cover, see Figure A.23.

Section 7.2 EXAMPLE A.25 (Open and Closed Sets, as well as Open Covers) *Let us consider the set of primes, a subset of the real numbers. It is easily seen, that this set is not open in* $\mathbb{R}$, *because all balls around a prime, with radius smaller then* 1, *contains no other prime. On the other side, the set of primes is closed, because the set has no points of accumulation, which implies, that it contains all of its points of accumulation. According to the definition above, the set of primes is closed. We can cover the set of primes by open intervals of the form* $(p - \epsilon, p + \epsilon)$, *where* p *is a prime and* ϵ *is a positive real number.*

Section 7.2 *An other interesting example is the set* $U = \bigcup_{n \in \mathbb{N}} \left(n, n + \frac{1}{n}\right)$, *i.e. a countable union of open intervals of the real line. As countable union of open sets,* U *is open.* U *is not closed, because* U *does not contains any of its points of accumulation, which are given by the set* $\left\{n, n + \frac{1}{n} \middle| n \in \mathbb{N}\right\}$. *Obviously a cover for* U *can be constructed via the countable union of open intervals of the form* $\left(n - \frac{1}{n}, n + \frac{2}{n}\right), n \in \mathbb{N}$, *see Figure A.24.*

SEQUENCES IN METRIC SPACES. If we now declare, as consequence of the norm on the linear normed space $(S, \|\cdot\|)$, a real valued function Δ_S on $S \times S$ with

$$\Delta_S(x, y) \stackrel{def}{=} \|x - y\| \text{ for all } x, y \in S, \tag{A.76}$$

Δ (851) then Δ_S clearly satisfies the characteristics of a *metric*, where a *metric* is defined as a

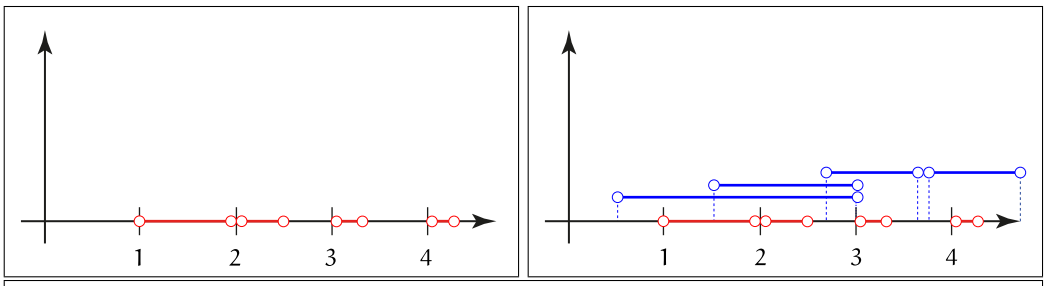


FIGURE A.24: AN OPEN SET WITH ASSOCIATED OPEN COVER. The set $U = \bigcup_{n\in\mathbb{N}} \left(n, n + \frac{1}{n}\right)$ from Example A.25 with an associated open cover.

distance function Δ_S defined on the space S, that is:

$$\Delta_S : S \times S \to \mathbb{R}^{\geq 0} \tag{A.77}$$

with

$$\Delta_S(x, y) \geq 0, \quad x, y \in S. \tag{A.78}$$

In addition to non-negativity and symmetry, that is, $\Delta_S(x, y) \geq 0$ and $\Delta_S(x, y) = \Delta_S(y, x)$ a metric also satisfies the triangle inequality $\Delta_S(x, z) \leq \Delta_S(x, y) + \Delta_S(y, z)$, and the uniqueness $\Delta_S(x, y) = 0 \Leftrightarrow x = y$, where $x, y, z \in S$.

SEQUENCES. Obviously, $\|x - y\|$ describes the distance between two elements of S. Thus, a linear normed space equipped with the metric Δ_S as its distance function, $(S, \|\cdot\|)$ permits the definition of the *limit* of a sequence $(x_n)_{n\in\mathbb{N}}$ of elements of S. This in turn implies, that the sequence $(x_n)_{n\in\mathbb{N}}$ converges towards x, which must not be an element of S. This behavior of a sequence is symbolized by $x_n \to x$ and must be read as follows: $\forall \epsilon > 0, \epsilon \in \mathbb{R}$, there exists an index $N(\epsilon)$, such that $\|x_n - x\| < \epsilon, \forall n \geq N(\epsilon)$. Another, more informal way of declaring this behavior could be: Pick any positive number ϵ, then the sequence $(x_n)_{n\in\mathbb{N}}$ is said to converge towards x, if it is always possible to make the difference between x_n and x smaller than ϵ by choosing n large enough, larger than some number N. Section 2.1.1 Section 2.1.5

REMARK A.13 *Clearly it is possible to define different norms over a linear space, each of which also induces a different topology, thus providing us with various different definitions of the size and distance of elements or of the convergence of sequences of elements out of S.*

EXAMPLE A.26 (Convergence Behavior of Monte Carlo Methods) *As we will see in more detail in a later chapter the convergence behavior of Monte Carlo methods for evalu-* Chapter 6

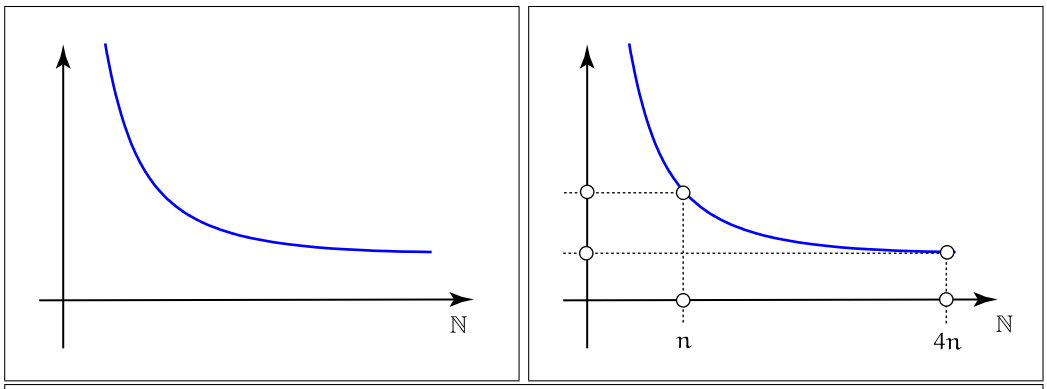


FIGURE A.25: CONVERGENCE BEHAVIOR OF MONTE CARLO METHODS. The more work is put in a Monte Carlo algorithm, the smaller is the resulting error. Although we will always be able to improve the result, we newer exactly reach it in general. Due to the convergence behavior of $O\left(\frac{1}{\sqrt{n}}\right)$ quadrupling the work of the algorithm will only halve the error.

ating integrals is of order $O\left(\frac{1}{\sqrt{n}}\right)$*, for this see Figure A.25. This statement means, that quadrupling the work of the algorithm will only halve the error. As easily seen,* Section 6.4 *the sequence* $x_n = \frac{1}{\sqrt{n}}$ *converges towards* 0*, since* $\left|\frac{1}{\sqrt{n}}\right| < \epsilon, \forall n \geq N(\epsilon)$*, if* $N(\epsilon) = \frac{1}{\epsilon^2}$ *holds and* $|\cdot|$*, the absolute value of a real number, represents the norm on* $\mathbb{R}$*. As a result, we conclude: The more work is put in a Monte Carlo algorithm, the smaller is the resulting error. Although we will always be able to improve the result, we newer exactly reach it in general.*

C(Λ) (28) CONTINUITY. In physics, many natural phenomena will be modeled by quantities, which may be represented by continuous functions. Informally, a real valued function can be declared to be *continuous*, if it is possible to draw the graph of the function without lifting one's pen. With this intuitive definition, a function is *discontinuous*, if its graph has a break or if it is an interrupted curve, perhaps unbounded at some points of its domain.

Now, for our subsequent work, this intuitive definition of continuity is insufficient. Therefore, we need a definition of continuity, that agrees with our intuition and which is also robust enough to be used in all mathematical situation.

DEFINITION A.27 (Continuity of a Function) *Let* f *be a function defined on a metric space* $(\mathcal{S}, \Delta_{\mathcal{S}})$ *with values in another metric space* $(\mathcal{T}, \Delta_{\mathcal{T}})$*. The function* f *is called* continuous at a point $x_0 \in \mathcal{S}$*, if for any positive real number* ϵ*, there exists a positive real number* δ*, such that* $\forall x \in \mathcal{S}$ *holds:*

$$\Delta_{\mathcal{S}}(x, x_0) < \delta \quad \Rightarrow \quad \Delta_{\mathcal{T}}(f(x), f(x_0)) < \epsilon, \tag{A.79}$$

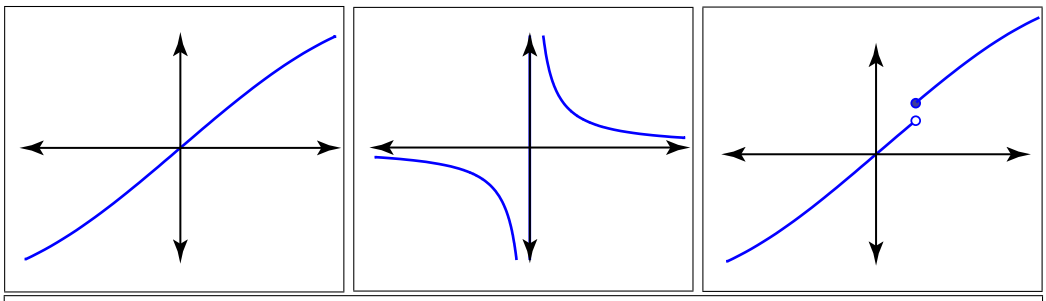


FIGURE A.26: CONTINUOUS AND DISCONTINUOUS REAL VALUED FUNCTIONS. Two countinuous and a discontinuous functions. The discontinuous functions has a break in its graph. Note, the function in the center image is continuous, although it has a singularity, i.e., an infinite value at the point $x_0 = 0$.

that is,

$$\|x - x_0\|_S < \delta \quad \Rightarrow \quad \|f(x) - f(x_0)\|_T < \epsilon. \tag{A.80}$$

If the function f *is continuous at every point* x_0 *of a set* $A \subset S$*, then* f *is said to be* continuous on A.

EXAMPLE A.27 *The Dirichlet function from Example A.8 is one of the most famous discontinuous functions in mathematics. To show this, we assume* x_0 *be an irrational number. Independent of a horizontal* ϵ*-strip around* 0*, there is no choice of* δ*, such that in the vertical* δ*-strip around* x_0 *there are only irrational numbers. Obviously, the same argument holds for the continuity at a point* $x_0 \in \mathbb{Q}$ *and a horizontal* δ*-strip around* 1.

Dirichlet Function (836)

A.4 A BIT OF DIFFERENTIAL CALCULUS

Calculus is the field in mathematics that is focused on the discussion of limits, infinite series, functions, derivatives, and integrals. It can be partitioned into two major branches, *differential calculus* and *integral calculus*.

Since it plays a central role in our further discussions, we repeat in this section the well known concept of the *derivative* from differential calculus and discuss classical integral calculus shortly in the following section.

DERIVATIVES. In particular in Chapter 3, we will often encounter mathematical notations of the form

$$\frac{d}{d\mu(x)}\Phi, \qquad \frac{d}{dx}\Phi, \qquad \frac{d\,\Phi}{d\mu(x)}, \qquad \text{or} \qquad \frac{d\,\Phi}{dx}, \tag{A.81}$$

the so-called *Leibniz notation* of differential calculus. They all symbolize the mathematical construct of a *derivative*.

The derivative of a function at a point is a way of interpreting the small-scale behavior of this function near that point. It provides information about the change of a quantity in response to changes in some other quantity. With respect to the notations from Equation (A.81), a quotient like $\frac{d\,\Phi}{dx}$ means, that a finite, measurable difference $\Delta\Phi = \Phi(x + \Delta x) - \Phi(x)$ of a quantity Φ is to be divided by a difference Δx in some other quantity.

The *derivative* of the function Φ with respect to x is then described by the process of continuously shrinking down the quantities $\Delta\Phi$ and Δx until they are immeasurably small, mathematically this can be expressed by:

$$\frac{dQ(x)}{dx} \stackrel{def}{=} \lim_{\Delta x \to 0} \frac{\Delta Q}{\Delta x}. \tag{A.82}$$

Even if the denominator on the right in the above equation goes to zero, the quotient will not be increased but will be remain finite, as the numerator also shrinks down.

Since derivatives are infinitesimally small quantities, they don't play an important role in practice. In practice one measures finite amount of or change in some physical quantity, such as $\Delta\Phi$ and divides by a finite amount of or change of the other quantity Δx. The smaller the quantity in the denominator can be made the better will be the quality of the measurement [127, McCluney 1994].

Open Interval (829)

DEFINITION A.28 (Derivative) *Let f be a function defined on an open interval $]a, b[\subset \mathbb{R}$ with values in $\mathbb{R}$ and $h \in \mathbb{R}$. Provided, that the limit*

$$\left(\frac{d}{dx} f\right)(x_0) \stackrel{def}{=} \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0))}{h} \tag{A.83}$$

exists for any $x_0 \in]a, b[$, then we call it the derivative *of f at point x_0.*

Obviously, the derivative $\frac{d}{dx} f$ is a function whose domain is the set of all points $x \in]a, b[$ where the limit from Equation (A.83) exists.

PARTIAL DERIVATIVE. Now, the most functions that we are encounter in our discussions on light transport problems, such as the radiometric quantities, are not functions of a single variable, but functions of several variables. With respect to the derivative of these functions, we are mainly interested in their *partial derivatives*.

Chapter 3

[174, Rudin 1998]

A partial derivative of a function of several variables is its derivative with respect to one of its variables, where all other variables as treated as constant.

DEFINITION A.29 (Partial Derivative) *Let* f *be a function defined on an open set* $E \subset \mathbb{R}^n$ *with values in* $\mathbb{R}$ *and* $h \in \mathbb{R}$. *Furthermore let* $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ *be the basis of* $\mathbb{R}^n$. *Provided, that the limit* Basis (857) Open Set (864)

$$\left(\frac{\partial}{\partial x_i} f\right)(\mathbf{x}_0) \stackrel{def}{=} \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0))}{h} \tag{A.84}$$

$$= \lim_{h \to 0} \frac{f(x_{0_1}, \ldots, x_{0_i} + h, \ldots x_{0_n}) - f(x_{0_1}, \ldots, x_{0_n})}{h}. \tag{A.85}$$

exists, then $\left(\frac{\partial}{\partial x_i} f\right)(\mathbf{x}_0)$ *is called the* i^{th} partial derivative *of* f *at point* $\mathbf{x_0}$.

REMARK A.14 *If we choose* $s = 1$ *in the above definition, then Definition A.29 also covers the derivative of a real-valued function of a single variable, thus*

$$\left(\frac{d}{dx} f\right)(x) \stackrel{def}{=} \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{A.86}$$

Partial derivatives play a fundamental role in vector calculus and differential geometry.

DIRECTIONAL DERIVATIVE. An other type of a derivative of functions of several variables is the *directional derivative*. It represents the instantaneous rate of change of the function with respect to the direction of a given vector. Obviously, the directional derivative generalizes the concept of the partial derivative, which can be interpreted as a direction derivative in one of the coordinate axes.

DEFINITION A.30 (Directional Derivative) *Let* $\alpha \in \mathbb{R}$, ω *be a vector and* $\mathbf{x}$ *a point from* $\mathbb{R}^n$. *Let furthermore* f *be a function defined on an open set* $E \subset \mathbb{R}^n$ *with values in* $\mathbb{R}$, *then the* directional derivative *of* f *in direction* ω *is defined as:* Open Set (864)

$$\frac{\partial}{\partial \alpha} f(\mathbf{x} + \alpha\omega)\Big|_{\alpha=0} \stackrel{def}{=} \left\langle \left(\frac{\partial}{\partial x_1} f, \ldots, \frac{\partial}{\partial x_n} f\right), (\omega_1, \ldots, \omega_n) \right\rangle, \tag{A.87}$$

where $\langle \cdot, \cdot \rangle$ *is the standard inner product of* $\mathbb{R}^n$. Inner Product (859)

When deriving the integral form of the stationary particle transport equation, we need the concept of the directional derivative in the special case $n = 3$. Here the directional derivative is given by Section 4.1.3

$$\frac{\partial}{\partial \alpha} f(\mathbf{x} + \alpha\omega)\Big|_{\alpha=0} \stackrel{def}{=} \left\langle \underbrace{\left(\frac{\partial}{\partial x_1} f, \frac{\partial}{\partial x_2} f, \frac{\partial}{\partial x_3} f\right)}_{\nabla f}, (\omega_1, \omega_2, \omega_3) \right\rangle \tag{A.88}$$

$$= \langle \nabla f, \omega \rangle, \tag{A.89}$$

whereas ∇f is called the *gradient* of f. Gradient (53)

THE JACOBIAN MATRIX. All functions considered until now in this section were mappings into the real numbers. Let us now study the local change of real, vector-valued functions with images in the Euclidean space $\mathbb{R}^r$. The derivative of such functions, including also the one-dimensional case, are described by the *Jacobian matrix*, often also shortly called the *Jacobian*. The Jacobian is a matrix, whose coefficients are given by the partial derivatives of all component functions of a vector-valued function.

$\mathbb{R}^r$ (830)
Matrix (853)

DEFINITION A.31 (The Jacobian Matrix) *Let $\mathbf{f}$ be a vector-valued function defined on an open set $E \subset \mathbb{R}^n$ with values in $\mathbb{R}^r$. Furthermore let $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ and $(\mathbf{e}'_1, \ldots, \mathbf{e}'_r)$ be the bases of $\mathbb{R}^n$ and $\mathbb{R}^r$, the components of the r-dimensional function $\mathbf{f}$ are given by the real-valued functions $f_1, \ldots, f_r$, whereas it holds:*

Open Set (864)
Basis (857)

$$\mathbf{f}(\mathbf{x}) = \sum_{i=1}^{r} f_i(\mathbf{x})\, \mathbf{e}'_i. \tag{A.90}$$

Provided, that the limits

$$\left(\frac{\partial}{\partial x_j} f_i\right)_{1 \le i \le r, 1 \le j \le n}, \tag{A.91}$$

exists, then we define the Jacobian *of $\mathbf{f}$ by:*

$$J_f(\mathbf{x}_0) \overset{def}{=} \left(\frac{\partial}{\partial x_j} f_i\right)_{1 \le i \le r, 1 \le j \le n} \tag{A.92}$$

$$= \begin{pmatrix} \frac{\partial}{\partial x_1} f_1(\mathbf{x}_0) & \frac{\partial}{\partial x_2} f_1(\mathbf{x}_0) & \cdots & \frac{\partial}{\partial x_n} f_1(\mathbf{x}_0) \\ \frac{\partial}{\partial x_1} f_2(\mathbf{x}_0) & \frac{\partial}{\partial x_2} f_2(\mathbf{x}_0) & \cdots & \frac{\partial}{\partial x_n} f_2(\mathbf{x}_0) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial}{\partial x_1} f_r(\mathbf{x}_0) & \frac{\partial}{\partial x_2} f_r(\mathbf{x}_0) & \cdots & \frac{\partial}{\partial x_n} f_r(\mathbf{x}_0) \end{pmatrix}, \tag{A.93}$$

that is, we define the Jacobian as the matrix of all partial derivatives.

In Monte Carlo integration we encounter again and again the problem of sampling random variables distributed according to density functions that can be computed via the transformation of known probability distributions. A typical example in this context is the transformation of uniformly on $[0,1]^2$ distributed random variables to random variables which are uniformly distributed on the unit circle or the unit sphere. Such a process needs the computation of the Jacobian for the polar coordinate transformation.

Random Variable (168)
Probability Distribution (80)
Uniform Distribution (180)

EXAMPLE A.28 (The Polar Coordinate Transformation) *Let us consider the function f that maps the polar coordinates (r, θ) to the Cartesian coordinates (x, y), it is given by:*

$$f : [0, \infty) \times [0, 2\pi) \to \mathbb{R}^2 \tag{A.94}$$

with

$$(r,\theta) \mapsto f(r,\theta) \stackrel{def}{=} (x,y) \tag{A.95}$$
$$= (r\cos\theta, r\sin\theta). \tag{A.96}$$

Due to Definition A.31 the Jacobian J_f is given by:

$$J_f = \begin{pmatrix} \frac{d}{dr}x & \frac{d}{d\theta}x \\ \frac{d}{dr}y & \frac{d}{d\theta}y \end{pmatrix} \tag{A.97}$$
$$= \begin{pmatrix} \frac{d}{dr}r\cos\theta & \frac{d}{d\theta}r\cos\theta \\ \frac{d}{dr}r\sin\theta & \frac{d}{d\theta}r\sin\theta \end{pmatrix} \tag{A.98}$$
$$= \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}. \tag{A.99}$$

Another interesting example is the spherical coordinate transformation, it is not only important for the representation of directions (r,θ,ϕ) as points (x,y,z), but it plays also an important role in our sampling theory.

EXAMPLE A.29 (The Spherical Coordinate Transformation) *Let us consider the function f that maps the spherical coordinates (r,θ,ϕ) to the Cartesian coordinates (x,y,z), it is given by:*

$$f : [0,\infty) \times [0,\pi] \times [0,2\pi) \to \mathbb{R}^3 \tag{A.100}$$

with

$$(r,\theta,\phi) \mapsto f(r,\theta,\phi) \stackrel{def}{=} (x,y,z) \tag{A.101}$$
$$= (r\sin\theta\cos\phi, r\sin\theta\sin\phi, r\cos\theta). \tag{A.102}$$

Due to Definition A.31 the Jacobian J_f is given by:

$$J_f = \begin{pmatrix} \frac{d}{dr}x & \frac{d}{d\theta}x & \frac{d}{d\phi}x \\ \frac{d}{dr}y & \frac{d}{d\theta}y & \frac{d}{d\phi}y \\ \frac{d}{dr}z & \frac{d}{d\theta}z & \frac{d}{d\phi}z \end{pmatrix} \tag{A.103}$$
$$= \begin{pmatrix} \frac{d}{dr}r\sin\theta\cos\phi & \frac{d}{d\theta}r\sin\theta\cos\phi & \frac{d}{d\phi}r\sin\theta\cos\phi \\ \frac{d}{dr}r\sin\theta\sin\phi & \frac{d}{d\theta}r\sin\theta\sin\phi & \frac{d}{d\phi}r\sin\theta\sin\phi \\ \frac{d}{dr}r\cos\theta & \frac{d}{d\theta}r\cos\theta & \frac{d}{d\phi}r\cos\theta \end{pmatrix} \tag{A.104}$$
$$= \begin{pmatrix} \sin\theta\cos\phi & r\cos\theta\cos\phi & -r\sin\theta\sin\phi \\ \sin\theta\sin\phi & r\cos\theta\sin\phi & r\sin\theta\cos\phi \\ \cos\theta & -r\sin\theta & 0 \end{pmatrix}. \tag{A.105}$$

THE JACOBIAN DETERMINANT. Considering real, vector-valued function from $\mathbb{R}^n$ to $\mathbb{R}^n$, then the Jacobian matrix is a square matrix, that means, we can build its determinant, the so-called *Jacobian determinant*, often also simply denoted as the *Jacobian*. The Jacobian determinant plays a central role when we transform a multidimensional integral from one integration domain into another.

DEFINITION A.32 (The Jacobian Determinant) *Let* $\mathbf{f}$ *be a vector-valued function defined on an open set* $E \subset \mathbb{R}^n$ *with values in* $\mathbb{R}^n$. *Provided, that the limits*

Open Set (864)

$$\left(\frac{\partial}{\partial x_j} f_i\right)_{1 \le i \le n, 1 \le j \le n}, \tag{A.106}$$

exist, then we define the Jacobian determinat *of* $\mathbf{f}$ *by:*

$$\det\left(J_f(\mathbf{x})\right) \stackrel{def}{=} \det\left(\frac{\partial}{\partial x_j} f_i\right)_{1 \le i \le n, 1 \le j \le n}. \tag{A.107}$$

EXAMPLE A.30 *Obviously, for the Jacobian determinant of the Jacobian matrix* J_f *from Example A.28 we have*

$$\det\left(J_f\right) = r\cos^2\theta + r\sin^2\theta = r \tag{A.108}$$

and for the Jacobian determinant of the Jacobian matrix J_f *from Example A.29 we get:*

$$\det\left(J_f\right) = r^2 \sin\theta. \tag{A.109}$$

A.5 A FIRST ENCOUNTER WITH THE LEBESGUE INTEGRAL AND MONTE CARLO INTEGRATION

Integrals are not only of importance for evaluating the length of curves, the area of surfaces, and the volume of n-dimensional abstract mathematical constructs or for solving differential equation, but they also provide the basis for describing many laws of nature by means of integral equations.

Now, in mathematics there exists many different types of integrals. The two most relevant are the *Riemann integral*, known from school, and the *Lebesgue integral*, which is the notion of integral in functional analysis and modern mathematics. Applied to computation of lengths, areas, and volumes, we will detect nearly no difference between the two types of integrals, except in advanced applications particularly with regard to functional analysis. Because the Lebesgue integral is more general and powerful as the ordinary Riemann integral, it is used as the fundamental basis of the theory of integral equations, which we need for a deeper understanding of the equations describing the global illumination problem.

This section serves as a review of the well-known concept of the *Riemann integral* as well as a first short encounter with the *Lebesgue integral*. We start in this section by motivating the integral as the area of a region between the graph of a function over a given integration domain. We will speak about integration domains—in particular those used in the integral equations of global illumination theory—and we will make some remarks about the historical development of the integral. That is, we present the way from Cauchy's definition of the integral to the integral notion by Lebesgue, which is of particular interest for us. Afterwards, we demonstrate on low-level the differences between the two widely-used types of integrals: the Riemann and the Lebesgue integral and we will conclude the section with a short insight into the world of Monte Carlo integration, where probabilistic approaches were used for the numerical solution of integrals.

Section 2.3
Section 2.2.4
Chapter 6

AREAS AND INTEGRATION DOMAINS. In mathematics the integral of a real-valued, continuous function f over the integration domain $[a, b]$ is denoted by

Continuous Function (869)

$$\int_a^b f(x)\,dx. \tag{A.110}$$

This integral has an elementary geometrical meaning, in that it describes the area of the region in the xy-plane bounded by the function f, the x-axis and the vertical lines $x = a$ and $x = b$. In multivariable calculus, it is shown, that the choice of intervals as domains of integrals is not mandatory. It is also possible to integrate functions of several variables over regions other than intervals on the real line, for example over bounded or unbounded areas in the xy-plane, volumes in $\mathbb{R}^3$, or higher dimensions.

Bounded Set (862)

In the following, we are interested in domains of integrals, which occur in so-called *integral equations*, as we will introduced them in the first chapter in form of the *rendering* and the *radiosity equation*. The integration domains of these integrals are always of a special kind. Instead off to integrate over a simple interval or an area, often we have to evaluate an integral over Cartesian products of subsets of $\mathbb{R}^2$ and the unit sphere as well as the upper or lower hemisphere.

Rendering Equation (400)
Radiosity Equation (782)

BOX A.4 (Integration Domains)

A common problem in computer graphics is the computation of the flux of any kind of stuff through a surface patch coming from or going in a given range of directions. As we will see in Chapter 3, we can solve this problem by computing the particles, which go out or come from any direction of the lower or upper hemisphere over any point on the patch.

Mathematically, this can be formulated as an integral over an integration domain, which can be constructed via the Cartesian product of the surface patch A and $\left[0, \frac{\pi}{2}\right] \times [0, 2\pi)$ or $\left[\frac{\pi}{2}, \pi\right] \times [0, 2\pi)$, that is, $A \times \left[0, \frac{\pi}{2}\right] \times [0, 2\pi)$ and $A \times \left[\frac{\pi}{2}, \pi\right] \times [0, 2\pi)$ respectively.

Due to our notation from Box (A.3) we can formulate the integration domains over a surface patch A and the upper or lower hemisphere, centered around a surface point s, in the future as:

$$A \times \mathcal{H}^2_+(s) \equiv A \times \left[0, \frac{\pi}{2}\right] \times [0, 2\pi) \tag{A.111}$$

and

$$A \times \mathcal{H}^2_-(s) \equiv A \times \left[\frac{\pi}{2}, \pi\right] \times [0, 2\pi). \tag{A.112}$$

SOME HISTORICAL REMARKS TO INTEGRATION THEORY. *Augustin-Louis Cauchy* can be considered as the intrinsic founder of the concept of the integral. Based on Cavalieri's and Fermat's approaches, Cauchy formulated a constructive definition of the integral of any arbitrary function, which is continuous on a closed interval $[a, b] \subset \mathbb{R}$. Forming a partition $a = x_0 < x_1 < \ldots < x_n = b$ of the integration domain with equidistant subintervals $[x_{i-1}, x_i], i = 1, \ldots, n$, Cauchy's idea was to take the left endpoints of the subinterval $[x_{i-1}, x_i]$ and considering the limit of the sum

Continuous Function (869)

$$\lim_{n\to\infty} \sum_{i=0}^{n} f(x_{i-1})(x_i - x_{i-1}). \tag{A.113}$$

If the limit of these sum exists, he called it the *integral* of f over the integration domain $[a, b]$.

Based on Cauchy's procedure, *Bernhard Riemann* asked the question: "In which cases is a function integrable and in which it is not integrable?" resulting in a weaker requirement than Cauchy's requirement of continuity. Riemann showed, that Cauchy's integral exists, if the function f is bounded on $[a, b]$, for this see Figure A.27. Informally spoken, a function f is *Riemann-integrable* on $[a, b]$, if and only if f does oscillate large only on very small sets. Until today, the Riemann integral, defined as the limit of the *upper* and *lower Riemann-Darboux integral*

Bounded Function (863)

$$\int_a^b f(x)\, dx \stackrel{def}{=} \overline{\int_a^b} f(x)\, dx = \underline{\int_a^b} f(x)\, dx, \tag{A.114}$$

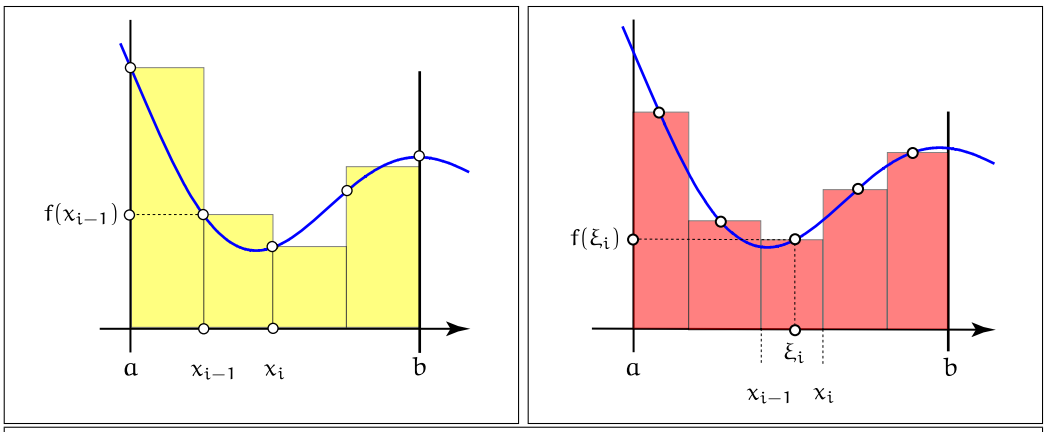


FIGURE A.27: CAUCHY'S AND RIEMANN'S CONSTRUCTION OF THE INTEGRAL. Left Cauchy's definition of the integral restricted to continuous functions. On the right, Riemann's definition of the integral—as the limit of the upper and lower Rieman-Darboux integral—based on a weaker requirement than Cauchy's requirement of continuity.

where the corresponding Riemann-Darboux integrals are given by:

$$\overline{\int_a^b} f(x)\,dx \stackrel{def}{=} \lim_{n\to\infty} \sum_{i=1}^{n} \sup_{x\in[x_{i-1},x_i]} f(x)\,(x_i - x_{i-1}) \tag{A.115}$$

$$\underline{\int_a^b} f(x)\,dx \stackrel{def}{=} \lim_{n\to\infty} \sum_{i=1}^{n} \inf_{x\in[x_{i-1},x_i]} f(x)\,(x_i - x_{i-1}), \tag{A.116}$$

is the most widely-used integral notion in mathematics.

It were *Camille Jordan* and *Emile Borel*, which introduced as the first, the concept of the *measure* in the theory of integration. Contrary to the previous investigations in integration theory, they partitioned an interval $[a, b]$ in so-called *measurable sets* instead of subintervals. With its strict description of a measure, finally Emile Borel clears the way to the introduction of a new type of integral, which is more universally valid as the common Riemann integral: the *Lebesgue integral*.

Measure (79)
Measurable Set (80)
Lebesgue Integral (105)

Resulting from a constructive process based on countably infinite covers, the *Lebesgue integral*, named after the french mathematician *Henri Lebesgue*, can be seen as the result of the fusion of the mathematical concepts of measure and integration.

Countable Inifinite Set (827)
Cover (865)

BOX A.5 (The Flux Through a Surface Patch Formulated as Lebesgue Integral)

As we will see later, the flux through any patch M can be computed via:

$$\int_{M\times\mathcal{H}^2_+(\mathbf{s})} L(\mathbf{x},\omega)\,d\mu(\mathbf{x},\omega) = \int_M \left(\int_{\mathcal{H}^2_+(\mathbf{s})} L(\mathbf{x},\omega)\,d\sigma^\perp(\omega)\right) d\mu(\mathbf{x}) \tag{A.117}$$

and in case of integration over the lower hemisphere, it holds:

$$\int_{M\times\mathcal{H}^2_-(\mathbf{s})} L(\mathbf{x},\omega)\,d\mu(\mathbf{x},\omega) = \int_M \left(\int_{\mathcal{H}^2_-(\mathbf{s})} L(\mathbf{x},\omega)\,d\sigma^\perp(\omega)\right) d\mu(\mathbf{x}). \tag{A.118}$$

In the equations above, already we use the notation of the Lebesgue integral. Here $d\mu$ and $d\sigma^\perp$ denote so-called measures. As we will see in Section 2.2, the notion of the measure will be fundamental in derivating the Lebesgue integral, which we will present in Section 2.2.4.

Lebesgue Integral (105) Measure Space (80)

SHORT REVIEW: RIEMANN VS LEBESGUES INTEGRAL. The *Lebesgue integral* is the type of integral used in modern mathematics, which makes it possible to define lengths, areas, and volumes in arbitrary abstract measure spaces. In the special case of real numbers, the Lebesgue integral, based on the Lebesgue measure, represents a real generalization of the ordinary Riemann integral.

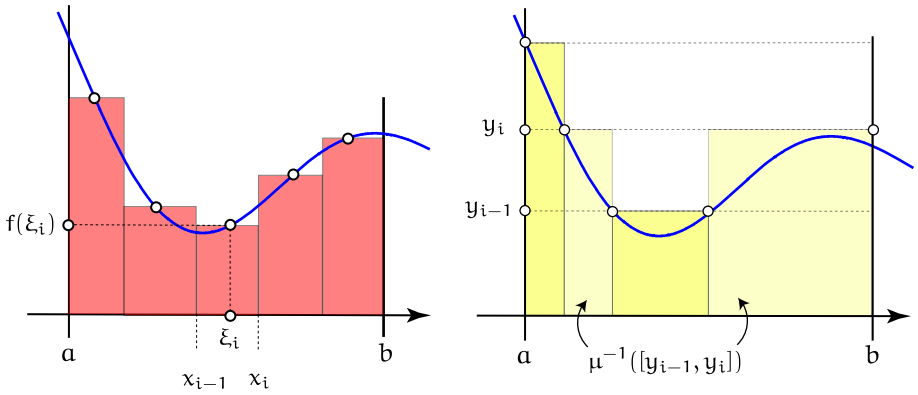
Sequence of Functions (30) Simple Function (839)

Contrary to the Riemann integral, which is defined by the limit of the area of a sequence of step functions, the Lebesgue integral is based on the limit of the area of a sequence of simple functions. Graphically this means: the Riemann integral can be visualized by vertical strips of the area under the graph of a function, the Lebesgue integral by horizontal strips of the area under the graph of a function, for this, see the figures in Box (A.6).

BOX A.6 (Riemann Integral vs Lebesgues Integral)

Let $(\mathcal{S}, \|\cdot\|)$ be a linear, normed space. The *Riemann integral* of a real-valued function f is based on the idea of splitting the integration domain $S \subset \mathcal{S}$ into a finite number n of subdomains, where the k-th subdomain has the area measure $\Delta x_i = \|x_i - x_{i-1}\|$, and then considering so-called *Riemann-sums* of the form $\lim_{n\to\infty} \sum_{i=1}^{n} f(\xi_i)\ \Delta x_i$ with $\xi_i \in [x_{i-1}, x_i]$

$$\int_S f(x)dx \stackrel{def}{=} \lim_{n\to\infty} \sum_{i=1}^{n} f(\xi_i)\ \Delta x_i. \tag{A.119}$$



In contrast to the Riemann integral, the *Lebesgues integral* of a function f is based not on the idea by further subdivisons of the integration domain, but by refining the approximation of f by very simple functions, i.e. by functions that take on a finite number of values. Provided, that we have no problems with the subsets M_i on which the functions take their constant values, the integral of f can be approximated by a sum of the form $\lim_{n\to\infty} \sum_{i=1}^{n} y_i\ d\mu_i$, where $d\mu_i \stackrel{def}{=} \mu(f^{-1}(M_i))$ denotes the measure of the subdomain M_i:

$$\int_S f(x)d\mu(x) \stackrel{def}{=} \lim_{n\to\infty} \sum_{i=1}^{n} y_i\ d\mu_i. \tag{A.120}$$

A SHORT PREVIEW TO MONTE CARLO INTEGRATION. Monte Carlo integration is a technique for the approximate evaluation of definite, multidimensional integrals. A Monte Carlo algorithm evaluates the integrand at randomly chosen points of the integration domain. Summing up this values provides us—in dependance of the number of randomly points—a good approximation of the proper value of the integral. Chapter 6

BOX A.7 (Monte Carlo Integration)

Let $f(x)$ be a 1D function, which we wish to integrate over a 1-dimensional domain $[a, b]$, i.e.,

$$\int_{[a,b]} f(x)\,dx. \tag{A.121}$$

The basic idea behind *Monte Carlo integration*, detailed discussed in Chapter 6, is to evaluate this integral by computing the mean value of $f(x)$ over the intervall $[a, b]$, and then multiply this mean by the area of the intervall $(b-a)$. For this purpose we generate N independent and uniformly distributed random variables $X_1, X_2, \ldots, X_N$ in $[a, b]$ and average the values of $f(x)$ at this N locations. This gives:

$$F_N \stackrel{def}{=} (b-a)\,\frac{1}{N}\sum_{i=1}^{N} f(X_i), \tag{A.122}$$

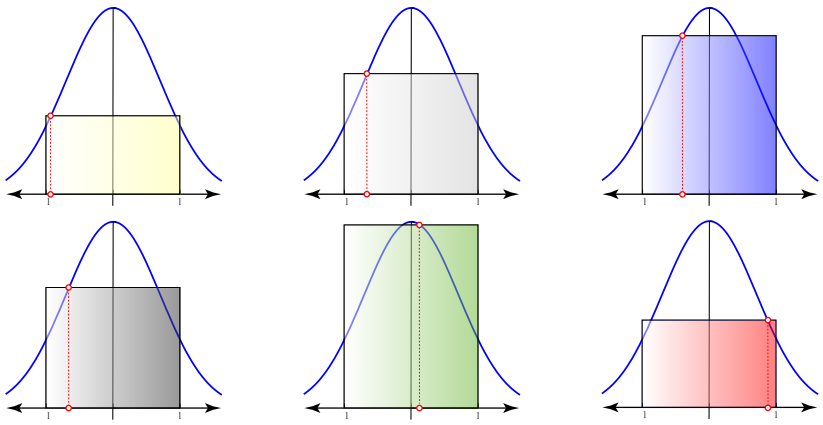
where F_N is called the *Monte Carlo estimator* of the intergral. Inceasing the number of samples, N, this estimator becomes more accurate and in the limit we will find that it holds:

$$\lim_{N\to\infty} F_N = \int_{[a,b]} f(x)\,dx. \tag{A.123}$$

The drawback of Monte Carlo integration is its slow convergence. As we will see in Section 6.4, the convergence rate of Monte Carlo integration ist $\frac{1}{\sqrt{N}}$, i.e., to halve the error we must quadruple the number of samples.

Let us show with the help of a simple example how Monte Carlo integration works. For that purpose, we compute the following 1D integral via 6 independent, uniformly distributed random variables X_i, drawn from $[-1, 1]$, we get:

$$\int_{[-1,1]} e^{-x^2}\,dx \quad\Longrightarrow\quad F_6 \stackrel{def}{=} \frac{1}{2}\sum_{i=1}^{6} e^{-X_i^2}. \tag{A.124}$$



B LIST OF SYMBOLS

For radiosity methods we suggest [36, Cohen & Wallace 1993], [13, Ashdown 1994], and [190, Sillion & Puech 1994], they can be classified as the standard works about radiosity. While [13, Ashdown 1994] qualifies for beginners, which are not familiar with the mathematics behind the global illumination equations, the other two require a deeper mathematical background for their comprehension.

C REFERENCE LITERATURE AND FURTHER READING

The appendix can be considered as a preparatory section for Chapter 2. It was written for the reader unfamiliar with calculus or basic linear algebra. Here, we present, in a short manner, the most important concepts from these areas. For the reader who is interested in a low-level introduction into these fields, we also recommend the undergraduate textbooks [120, Lang 1968], [121, Lang 1987] and Schaum's outline series [201, Spiegel 1995]. There is an endless list of literature which deals with calculus and linear algebra, such as: [18, Barner & Flohr 19989], [17, Barner & Flohr 1991], [110, König 1984], [173, Rudin 1976], [174, Rudin 1998], [2, Amann & Escher 1998], [3, Amann & Escher 1999], [4, Amann & Escher 2001] and [112, Kowalsky 1979].

GLOSSARY

L, radiance... 25

σ, solid angle measure... 25

name beschreibung... 25

BIBLIOGRAPHY

[1] AKENINE-MÖLLER TOMAS, HAINES ERIC AND HOFFMAN NATY. REAL-TIME RENDERING. A K PETERS, LTD. 888 WORCESTER STREET, SUITE 230, WELLESLEY, MA 02482, THIRD EDITION, 2008.

[2] AMANN HERBERT UND ESCHER JOACHIM. ANALYSIS I. BIRKHÄUSER VERLAG, BASEL, 1998.

[3] AMANN HERBERT UND ESCHER JOACHIM. ANALYSIS II. BIRKHÄUSER VERLAG, BASEL, 1999.

[4] AMANN HERBERT UND ESCHER JOACHIM. ANALYSIS III. BIRKHÄUSER VERLAG, BASEL, 2001.

[5] APPLE ARTHUR. SOME TECHNIQUES FOR SHADING MACHINE RENDERINGS OF SOLIDS. *In Proceedings of the Spring Joint Computer Conference*, 32:37–45, 1968.

[6] ARVO JAMES. BACKWARD RAY TRACING. *In ACM SIGGRAPH '86 Course Notes - Developments in Ray Tracing*, PAGES 259–263, 1986.

[7] ARVO JAMES. THE ROLE OF FUNCTIONAL ANALYSIS IN GLOBAL ILLUMINATION. *Rendering Techniques, edited by P.M. Hanrahan and W. Purgarthofer*, 1991.

[8] ARVO JAMES. LINEAR OPERATORS AND INTEGRAL EQUATIONS IN GLOBAL ILLUMINATION. *ACM SIGGRAPH, Global Illumination course notes*, CHAPTER 2, AUGUST 1993.

[9] ARVO JAMES. TRANSFER EQUATIONS IN GLOBAL ILLUMINATION. *In Global Illumination, SIGGRAPH '93 Course Notes*, VOL. 42, AUGUST 1993.

[10] ARVO JAMES. *Analytic Methods for Simulated Light Transport*. PHD THESIS, YALE UNIVERSITY, 1995.

[11] ARVO JAMES AND KIRK DAVID. PARTICLE TRANSPORT AND IMAGE SYNTHESIS. *SIGGRAPH '90 Proceedings of the 17th annual conference on Computer graphics and interactive techniques*, PAGES 63–66, 1990.

[12] ARVO JAMES, TORRANCE KENNETH AND SMITS BRIAN. A FRAMEWORK FOR THE ANALYSIS OF ERROR IN GLOBAL ILLUMINATION ALGORITHMS. *Computer Graphics Proceedings, Annual Conference Series, ACM SIGGRAPH*, PAGES 75–84, 1994.

[13] ASHDOWN IAN. RADIOSITY A PROGRAMMER'S PERSPECTIVE. JOHN WILEY & SONS, NEW YORK, FIRST EDITION, 1994.

[14] ASHIKHMIN MICHAEL AND SHIRLEY PETER. AN ANISOTROPIC PHONG BRDF MODEL. *Journal of Graphics Tools*, 5:6–74, 2000.

[15] ASH ROBERT B. AND DOLÉANS-DADE CATHERINE A. PROBABILITY & MEASURE THEORY. ACADEMIC PRESS, SAN DIEGO, SECOND EDITION, 2000.

[16] ATKINSON KENDALL AND HAN WEIMIN. THEORETICAL NUMERICAL ANALYSIS. SPRINGER, SECOND EDITION, 2007.

[17] BARNER MARTIN UND FLOHR FRIEDRICH. ANALYSIS I. WALTER DE GRUYTER, NEW YORK, FOURTH EDITION, 1989.

[18] BARNER MARTIN UND FLOHR FRIEDRICH. ANALYSIS II. WALTER DE GRUYTER, NEW YORK, SECOND EDITION, 1991.

[19] BECKMANN PETR AND SPIZZICHINO ANDRE. THE SCATTERING OF ELECTROMAGNETIC WAVES FROM ROUGH SURFACES. ARTECH HOUSE INC., FIRST EDITION, 1987.

[20] BEKAERT PHILIPPE. *Hierarchical and Stochastic Algorithms for Radiosity*. PHD THESIS, KATHOLIKE UNIVERISTAIT LEUVEN, 1999.

[21] BEKAERT PHILIPPE, NEUMANN LÁSZLÓ, NEUMANN ATTILA, SBERT MATEU AND WLLEMS YVES D. HIERARCHICAL MONTE CARLO RADIOSITY. 9TH EUROGRAPHICS WORKSHOP ON RENDERING, VIENNA, AUSTRIA, JUNE 1998.

[22] BEREZANSKY YURIJ. M., SHEFTEL ZINOVIJ G. AND GEORGIJ F. US. FUNCTIONAL ANALYSIS VOL. I. OPERATOR THEORY ADVANCES AND APPLICATIONS VOL. 85. BIRKHÄUSER VERLAG, BASEL, FIRST EDITION, 1996.

[23] BIERBAUER JÜRGEN AND EDEL YVES. SOME GOOD TERNARY (t, m, s)-NETS. DEPARTMENT OF MATHEMATICAL SCIENCES, MICHIGAN TECHNOLOGICAL UNIVERSITY, HOUGHTON, MICHIGAN, USA, FEBRUARY 1997.

[24] BLASI PHILIPPE, LE SAEC BERTRAND AND SCHLICK CHRISTOPHE. A RENDERING ALGORITHM FOR DISCRETE VOLUME DENSITY OBJECTS. *In Computer Graphics Forum (Eurographics '93)*, 12(3):201–210, 1993.

[25] BLINN JAMES F. MODELS OF LIGHT REFLECTION FOR COMPUTER SYTHESIZED PICTURES. *Computer Graphics (Proc. Siggraph '77), 11(2): 192-198*, JULY 1977.

[26] BOHREN CRAIG F. AND HUFFMAN DONALD R. ABSORPTION AND SCATTERING OF LIGHT BY SMALL PARTICLES. WILEY-VCH, WEINHEIM, FIRST EDITION, 2004.

[27] BORN MAX AND WOLF EMIL. PRINCIPLES OF OPTICS. CAMBRIDGE, UNIVERSITY PRESS, CAMBRIDGE, SEVENTH EDITION, 1999.

[28] BOSS EMMANUEL, MOBLEY CURTIS AND ROESLER COLLIN. OCEAN OPTICS BOOK. OCEAN OPTICS WEB BOOK, 2011 CREATIVE COMMONS ATTRIBUTION LICENSE, 2011.

[29] BRENNER SUSANNE C. AND SCOTT L. RIDGWAY. THE MATHEMATICAL THEORY OF FINITE ELEMENT METHODS. SPRINGER-VERLAG, NEW YORK, 1994.

[30] BURK FRANK. LEBESGUES MEASURE AND INTEGRATION. PURE AND APPLIED MATHEMATICS: A WILEY - INTERSIENCE SERIES OF TEXTS, MONOGRAPHS, AND TRACTS. JOHN WILEY & SONS, NEW YORK, FIRST EDITION, 1998.

[31] CAPIŃSKI MAREK AND KOPP EKKEHARD. MEASURE, INTEGRAL AND PROBABILITY. SPRINGER UNDERGRADUTE MATHEMATICS SERIES. SPRINGER-VERLAG, LONDON, THIRD EDITION, 2000.

[32] CHAI SOO BONG. LEBESGUES INTEGRATION. SPRINGER-VERLAG, NEW YORK BERLIN HEIDELBERG, SECOND EDITION, 1995.

[33] CHANDRASEKHAR SUBRAHMANYAN. RADIATIVE TRANSFER. DOVER PUBLICATIONS INC., OXFORD, FIRST EDITION, 1960.

[34] CHRISTENSEN PER H. *Hierarchical Techniques for Glossy Global Illumination*. PHD THESIS, UNVERSITY OF WASHINGTHON, 1995.

[35] CHRISTENSEN PER H. ADJOINTS AND IMPORTANCE IN RENDERING: AN OVERVIEW. *IEEE Transactions on Visualization and Computer Graphics*, 9(3), JULY-SEPTEMBER 2003.

[36] COHEN MICHAEL F. AND WALLACE JOHN R. RADIOSITY AND REALISTIC IMAGE SYNTHESIS. ACADEMIC PRESS PROFESSIONAL, CAMBRIDGE, MASSACHUSETTS, FIRST EDITION, 1993.

[37] COMNINOS PETER. MATHEMATICAL AND COMPUTER PROGRAMMING TECHNIQUES FOR COMPUTER GRAPHICS. SPRINGER-VERLAG, LONDON, FIRST EDITION, 2006.

[38] COOK ROBERT. STOCHASTIC SAMPLING IN COMPUTER GRAPHICS. *ACM Transaction on Graphics*, 5(1):51–72, 1986.

[39] COOK ROBERT L., TORRANCE KENETH E. A REFLECTANCE MODEL FOR COMPUTER GRAPHICS. *ACM Transactions on Graphics*, 1:7 – 24, JANUAR 1982.

[40] COOK ROBERT, PORTER THOMAS AND CARPENTER LAUREN. DISTRIBUTED RAY TRACING. *Computer Graphics*, 18(3):137–145, JULY 1984.

[41] DAUCET ARNAUD, JOHANSEN ADAM M. AND TADIĆ VLADISLAV B. ON SOLVING INTEGRAL EQUATIONS USING MARKOV CHAIN MONTE CARLO METHODS. *ACM Applied Mathematics and Computation*, 1(216):2869–2880, 2010.

[42] DE BERG MARC, VAN KREVELD MARC, OVERMARS MARC AND SCHWAZKOPF OTFRIED. COMPUTATIONAL GEOMETRY. SPRINGER-VERLAG, BERLIN HEIDELBERG, 1997.

[43] KELLY DEMPSKI AND EMMANUEL VIALE. ADVANCED LIGHTING AND MATERIALS WITH SHADERS. WORDWARE PUBLISHING, INC., SAN DIEGO, SEVENTH EDITION, 2005.

[44] DITCHBURN R. W. LIGHT. DOVER PUBLICATIONS INC., 31 EAST 2ND STREET, MINEOLA, NEW YORK, 11501, 1991.

[45] DOPKIN DAVID P., EPPSTEIN DAVID AND MITCHELL DON P. COMPUTING THE DISCREPANCY WITH APPLICATIONS TO SUPERSAMPLING PATTERNS. *In Proceeding SCG' 93, Proceedings of the ninth annual symposium on Computational geometry, pp. 47-52*, 1993.

[46] DRABEK PAVEL UND KUFNER ALOIS. INTEGRALGLEICHUNGEN. MATHEMATIK FÜR INGENIEURE UND NATURWISSENSCHAFTLER. B. G. TEUBNER VERLAGSGESELLSCHAFT, 1996.

[47] DUTRÉ PHILIP. *Mathematical Framework and Monte Carlo Algorithms for Global Illumination in Computer Graphics*. PHD THESIS, KATHOLIKE UNIVERISTAIT LEUVEN, 1996.

[48] DUTRÉ PHILIP. GLOBAL ILLUMINATION COMPENDIUM, 2003.

[49] DUTRÉ PHILIP AND WILLEMS YVES D. IMPORTANCE-DRIVEN LIGHT TRACING. *Proceedings of the Fifth Eurographics Worksshop on Rendering, pp. 185-194, Darmstadt, Germany, June*, 1994.

[50] DUTRÉ PHILIP, BEKAERT PHILIPPE AND BALA KAVITA. ADVANCED GLOBAL ILLUMINATION. NATICK, MASSACHUSETTS. A K PETERS, 2003.

[51] DUTRÉ PHILIP, BEKAERT PHILIPPE AND BALA KAVITA. ADVANCED GLOBAL ILLUMINATION. AK PETERS, NATICK, MASSACHUSETTS, SECOND EDITION, 2006.

[52] DUTRÉ PHILIP, LAFORTUNE ERIC P. AND WILLEMS YVES D. MONTE CARLO LIGHT TRACING WITH DIRECT COMPUTATION OF PIXEL INTENSITIES. *Proceedings of CompuGraphics, pp. 128-137, Alvor, Portugal, December*, 1993.

[53] EDWARDS R.E. FUNCTIONAL ANALYSIS. HOLT, RINEHART AND WINSTON, NEW YORK SAN FRANCISCO TORONTO LONDON, 1965.

[54] ELSTRODT JÜRGEN. MASS- UND INTEGRATIONSTHEORIE. SPRINGER, BERLIN HEIDELBERG NEW YORK, ERSTE EDITION, 1996.

[55] ENCARNACAO JOSÉ, STRASSER WOLFGANG AND KLEIN REINHARD. GRAPHISCHE DATENVERARBEITUNG 2. OLDENBURG, MÜNCHEN, VIERTE EDITION, 1997.

[56] ENGL HEINZ W. INTEGRALGLEICHUNGEN. SPRINGER LEHRBUCH MATHEMATIK. SPRINGER, 1997.

[57] EVANS MICHAEL AND SWARTZ TIM. APPROXIMATING INTEGRALS VIA MONTE CARLO AND DETERMINISTIC METHODS. OXFORD-UNIVERSITY PRESS, OXFORD, FIRST EDITION, 2000.

[58] FAURE HENRY. MONTE-CARLO AND QUASI-MONTE-CARLO METHODS FOR NUMERICAL INTEGRATION. TECHNICAL REPORT, INSTITUT DE MATHÉMATIQUES DE LUMINY, U.P.R. 9016 CNRS 163 AVENUE DE LUMINY, CASE 907, F-13288 MARSEILLE CEDEX 09 FRANCE, 2009.

[59] FEYNMAN RICHARD P. QED: THE STRANGE THEORY OF LIGHT AND MATTER. PRINCETON UNIVERSITY PRESS, PRINCETON, FIRST EDITION, 1985.

[60] FISHMAN GEORGE S. MONTE CARLO, VOLUME CONCEPTS, ALGORITHMS, AND APLICATIONS. SPRINGER-VERLAG, LONDON, FIRST EDITION, 1996.

[61] FLEET DAVID AND HERTZMANN AARON. CS418 RADIOMETRY AND REFLECTION. AVAILABLE ONLINE.

[62] FOLEY JAMES D., VAN DAM ANDRIES, FEINER STEVEN K. AND HUGHES JOHN F. COMPUTER GRAPHICS, PRINCIPLES AND PRACTICE. ACADEMIC PRESS, LONDON, FIRST EDITION, 1987.

[63] GENTLE JAMES E. RANDOM NUMBER GENERATION AND MONTE CARLO METHODS. SPRINGER-VERLAG, NEW YORK BERLIN HEIDELBERG, FIRST EDITION, 1998.

[64] GIBSON SIMON. EFFICIENT RADIOSITY FOR COMPLEX ENVIRONMENTS. MASTER'S THESIS, UNIVERSITY OF MANCHESTER, 1995.

[65] GILKS W. R., RICHARDSON S. AND SPIEGELHALTER D. J. MARKOV CHAIN MONTE CARLO IN PRACTIVE. CHAMANN & HALL, LONDON, FIRST EDITION, 1996.

[66] GLASSNER ANDREW S. AN INTRODUCTION TO RAY TRACING. ACADEMIC PRESS, LONDON, FIRST EDITION, 1989.

[67] GLASSNER ANDREW S. PRINCIPLES OF DIGITAL IMAGE SYNTHESIS VOLUME 1. MORGAN KAUFMANN, SAN FRANCISCO, FIRST EDITION, 1995.

[68] GLASSNER ANDREW S. PRINCIPLES OF DIGITAL IMAGE SYNTHESIS VOLUME 2. MORGAN KAUFMANN, SAN FRANCISCO, FIRST EDITION, 1995.

[69] GOESELE MICHAEL. *New Acquisition Techniques for Real Objects and Light Sources*. PHD THESIS, UNIVERSITÄT DES SAARLANDES, 2004.

[70] GORAL C.M., TORRANCE K.E., GREENBERG D.P., AND BATTAILE B. MODELLING THE INTERACTION OF LIGHT BETWEEN DIFFUSE SURFACES. *Computer Graphics (SIGGRAPH '84 Proceedings)*, PAGES 212–222, JULY 1984.

[71] HACHISUKA TOSHIYA, OGAKI SHINJI AND JENSEN HENRIK WANN. PROGRESSIVE PHOTON MAPPING. *ACM Transactions on Graphics, (Proceedings of SIGGRAPH Asia 1008) 27(5), 130:1-130:8*, 2008.

[72] HACKBUSCH WOLFGANG. INTEGRAL EQUATIONS. BIRKHÄUSER VERLAG, BASEL, FIRST EDITION, 1995.

[73] HAMMERSLEY J. AND HANDSCOMP D. C. MONTE CARLO METHODS. JOHN WILEY & SONS, NEW YORK, FIRST EDITION, 1964.

[74] HANRAHAN PAT. MONTE CARLO PATH TRACING. *State of the Art in Monte Carlo Ray Tracing for Realistic Image Synthesis, SIGGRAPH 2001, Course 29, pp. 71-89*, 2001.

[75] HANRAHAN PAT AND KRUEGER WOLFGANG. REFLECTION FROM LAYERED SURFACES DUE TO SUBSURFACE SCATTERING. *Proceedings of SIGGRAPH' 93, Computer Graphics Proceedings , Annual Conference Series) edited by James T.Kajiya, pp. 165-174, New York: ACM Press*, 1993.

[76] HANRAHAN PAT, RUSINKEWICZ SZYMON AND MARSCHNER STEVE. CS448C: TOPICS IN COMPUTER GRAPHICS, APPEARANCE MODELS FOR COMPUTER GRAPHICS AND VISION, STANFORD UNIVERSITY, 2000.

[77] HAVRAN VLASTIMIL, BITTNER JIRI, HERZOG ROBERT AND SEIDEL HANS-PETER. RAY MAPS FOR GLOBAL ILLUMINATION. *Eurographics Symposium on Rendering*, 2005.

[78] HEARN DONALD AND BAKER M. PAULINE. COMPUTER GRAHICS. PRENTICE HALL, INC., ENGLEWOOD CLIFFS, NEW JERSEY, SECOND EDITION, 1994.

[79] HECHT EUGENE. THEORY AND PROBLEMS OF OPTICS. MCGRAW-HILL INC., NEW YORK, FIRST EDITION, 1975.

[80] HECHT EUGENE. OPTICS. ADDISON WESLEY PUBLISHING COMPANY, NEW YORK, FOURTH EDITION, 2001.

[81] HECKBERT PAUL S. ADAPTIVE RADIOSITY TEXTURES FOR BIDIRECTIONAL RAY TRACING. *Computer Graphics (SIGGRAPH '90 Conference Proceedings)*, 24(4):145–154, AUGUST 1990.

[82] HECKBERT PAUL S. *Simulating Global Illumination using Adaptive Meshing*. PHD THESIS, UNIVERSITY OF CALIFORNIA, BERKLEY, CA 94720, 1991.

[83] HENYEY L.G. AND GREENSTEIN J.L. DIFFUSE RADIATION IN THE GALAXY. *Astrophysics Journal*, 93:70–83, 1941.

[84] HESSE CHRISTIAN. ANGWEWANDTE WAHRSCHEINLICHKEITSTHEORIE. VIEWEG & SOHN VERLAGSGESELLSCHAFT MBH, BRAUNSCHWEIG WIESBADEN, ERSTE EDITION, 2003.

[85] HE XIAO D. AND TORRANCE KENNNETH E. AND SILLION FRANCOIS X. AND GREENBERG DONALD P. A COMPREHENSIVE PHYSICAL MODEL FOR LIGHT REFLECTION. *Computer Graphics (Proc. Siggraph '91)*, 25(4):175–186, JULY 1991.

[86] HOFFMAN-JORGENSEN J. PROBABILITY THEORY WITH A VIEW TOWARD STATISTICS, VOLUME I OF *Chapman & Hall Probability Series*. CHAPMAN HALL, LONDON, FIRST EDITION, 1994.

[87] HOPCROFT JOHN E. AND ULLMAN JEFFREY D. INTRODUCTION TO AUTOMATA THEORY, LANGUAGES AND COMPUTATION. ADDISON-WESLEY PUBLISHING COMPANY, INC., READING, MASSACHUSETTS, FIRST EDITION, 1979.

[88] HOWELL JOHN R. A CATALOG OF RADIATION CONFIGURATION FACTORS. MCGRAW-HILL INC., NEW YORK, USA, 1982.

[89] HUBELI ALEXANDER, LIPPERT LARS AND GROSS MARKUS. THE GLOBAL CUBE A HARDWARE-ACCELERATED HIERARCHICAL VOLUME RADIOSITY TECHNIQUE. CS TECHNICAL REPORT 331, ETH, EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE ZÜRICH, OCTOBER 1999.

[90] HUTCHINSON DAVID J. MONTE CARLO PATH TRACING FOR PHOTOREALISM. MASTER'S THESIS, UNIVERSITY OF MANCHESTER, 1993.

[91] ISHIMARU AKIRA. WAVE PROPAGATION AND SCATTERING IN RANDOM MEDIA. OXFORD UNIVERSITY PRESS, OXFORD, ENGLAND, FIRST EDITION, 1997.

[92] JAN JUKKA KAINULAINEN. REFLECTION MODELS IN REAL-TIME GRAPHICS. MASTER'S THESIS, HELSINKI UNIVERSITY OF TECHNOLOGY, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, 2003.

[93] JENSEN HENRIK AND BUHLER JUAN. A RAPID HIERARCHICAL RENDERING TECHNIQUE FOR TRANSLUCENT MATERIALS. *ACM Transactions on Graphics*, PAGES 576–581, 2002.

[94] JENSEN HENRIK, MARSCHNER STEVE, LEVOY MARC AND HANRAHAN PAT. A PRACTICAL MODEL FOR SUBSURFACE LIGHT TRANSPORT. *Proceedings of SIGGRAPH 2001, Computer Graphics Proceedings , Annual Conference Series edited by E. Fiume, Reading, MA: Addison Wesley*, PAGES 511–518, 2001.

[95] JENSEN HENRIK WANN. REALISTIC IMAGE SYNTHESIS USING PHOTON MAPPING. A K PETERS, NATICK, MASSACHUSETTS, FIRST EDITION, 2001.

[96] JENSEN HENRIK WANN AND CHRISTENSEN NIELS JØRGEN. PHOTON MAPS IN BIDIRECTIONAL MONTE CARLO RAY TRACING OF COMPLEX OBJECTS. *Computers and Graphics*, 19(2):215–224, MARCH 1995.

[97] JENSEN HENRIK WANN AND CHRISTENSEN NIELS JØRGEN. A PRACTICAL GUIDE TO GLOBAL ILLUMINATION USING PHOTON MAPS. SIGGRAPH 2000 COURSE 8, JULY, 23 2000.

[98] KAJIYA JAMES T. THE RENDERING EQUATION. *Computer Graphics (SIGGRAPH '86 Conference Proceedings)*, 20(4):143–150, AUGUST 1986.

[99] KALOS MALVIN H. AND WHITLOCK PAULA. A. MONTE CARLO METHODS. JOHN WILEY & SONS, NEW YORK, FIRST EDITION, 1986.

[100] KELLER ALEXANDER. MONTE CARLO AND BEYOND. VORLESUNG COMPUTER GRAPHIK II, UNIVERSITÄT DES SAARLANDES, 2002.

[101] KELLER ALEXANDER. THE FAST CALCULATION OF FORM FACTORS USING LOW DISCREPANCY SEQUENCES. *In Proc. Spring Conference on Computer Graphics (SCCG '96)*, PAGES 195–2004, 1996.

[102] KELLER ALEXANDER. THE QUASI-RANDOM WALK. *In Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Lecture Notes in Statistics, Springer*, 127:277–291, 1996.

[103] KELLER ALEXANDER. INSTANT RADIOSITY. *In Proceedings of SIGGRAPH '97, Computer Graphics Proccedings, Annual Conference Series, Los Angeles*, PAGES 49–56, 1997.

[104] KELLER ALEXANDER. *Quasi-Monte Methods for Photorealistic Image Synthesis*. PHD THESIS, UNIVERSITÄT KAISERSLAUTERN, 1998.

[105] KELLER ALEXANDER AND HEINRICH STEFAN. QUASI-MONTE CARLO METHODS IN COMPUTER GRAPHICS: THE RADIANCE EQUATION. *Technicl Report 243/94*, 1994.

[106] KELLER ALEXANDER AND HEINRICH STEFAN. QUASI-MONTE CARLO METHODS IN COMPUTER GRAPHICS: THE GLOBAL ILLUMINATION PROBLEM. *Lecute Notes in App. Math.*, 32:455–469, 1996.

[107] KOLB CRAIG, MITCHELL DON AND HANRAHAN PAT. A REALISTIC CAMERA MODEL FOR COMPUTER GRAPHICS. *Proceedings of SIGGRAPH 95*, 1995.

[108] KOLLIG THOMAS AND KELLER ALEXANDER. EFFICIENT BIDIRECTIONAL PATH TRACING BY RANDOMIZED QUASI-MONTE CARLO INTEGRATION. *In Monte Carlo and Quasi-Monte Carlo Methods 2000,*, PAGES 290–305, 2002.

[109] KOLLIG THOMAS AND KELLER ALEXANDER. EFFICIENT MULTIDIMENSIONAL SAMPLING. *EUROGRAPHICS 2002*, 21(3), 2002.

[110] KÖNIG HEINZ. ANALYSIS I. BIRKHÄUSER VERLAG, BASEL, 1984.

[111] KÖNIG HEINZ. MEASURE AND INTEGRATION. SPRINGER, BERLIN HEIDELBERG NEW YORK, FIRST EDITION, 2000.

[112] KOWALSKY HANS-JOACHIM. LINEARE ALGEBRA. WALTER DE GRUYTER, NINTH EDITION, 1979.

[113] KRESS RAINER. LINEAR INTEGRAL EQUATIONS. SPRINGER, BERLIN HEIDELBERG NEW YORK, SECOND EDITION, 1999.

[114] KREYSZIG ERWIN. INTRODUCTORY FUNCTIONAL ANALYSIS WITH APPLICATION. JOHN WILEY & SONS, NEW YORK SANTA BARBARA LONDON SYDNEY TORONTO, 1978.

[115] LAFORTUNE E., FOO S.C., TORRANCE K. AND GREENBERG D. NON-LINEAR APPROXIMATION OF REFLECTANCE FUNCTIONS. *In Proceedings of SIGGRAPH '97, Computer Graphics Proccedings, Annual Conference Series, Los Angeles,* PAGES 117–126, 1997.

[116] LAFORTUNE ERIC P. *Mathematical Models and Monte Carlo Algorithms for Physically Based Rendering.* PHD THESIS, KATHOLIKE UNIVERSITAIT LEUVEN, BELGIUM, 1996.

[117] LAFORTUNE ERIC P. AND WILLEMS YVES D. USING THE MODIFIED PHONG REFLECTANCE MODEL FOR PHYSICALLY BASED RENDERING. TECHNICAL REPORT CW 197, DEPARTMENT OF COMPUTING SCIENCE, K.U. LEUVEN, NOVEMBER 1994.

[118] LAFORTUNE ERIC P. AND WILLEMS YVES D. THE AMBIENT TERM AS A VARIANCE REDUCTION TECHNIQUE FOR MONTE CARLO RAY TRACING. *Proceedings of the Fufth Eurographics Workshop on Rendering,* PAGES 163–171, JUNE 1994.

[119] LAFORTUNE ERIC P. AND WILLEMS YVES D. BI-DIRECTIONAL PATH TRACING. *Proceedings of SIGGRAPH 97, Computer Graphics Proceedings, Annual Conference Series, edited by Turner Whitte, Reading, MA: Addison Wesley,* PAGES 117–126, 1997.

[120] LANG SERGE. ANALYSIS I. ADDISON WESLEY PUBLISHING COMPANY, INC, 1968.

[121] LANG SERGE. CALCULUS OF SEVERAL VARIABLES. SPRINGER-VERLAG, THIRD EDITION, 1987.

[122] LAUSCHKE STEPHEN MARIUS. PRECOMPUTED RADIANCE TRANSFER (PRT). TECHNISCHE UNIVERSITÄT MÜNCHEN, FAKULTÄT FÜR INFORMATIK, LEHRSTUHL COMPUTER GRAFIK; HS: ASPECTS OF GAME ENGINE DESIGN, 2006.

[123] LEBEDEV L. P., VOROVICH I.I. AND GLADWELL G.M.L. FUNCTIONAL ANALYSIS. KLUWER ACADEMIC PUBLISHERS, NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW, SECOND EDITION, 2003.

[124] LENSCH HENDRIK P.A. REALISTIC MATERIALS IN COMPUTER GRAPHICS, INTRODUCTION, COURSE 10, STANFORD UNIVERSITY, 2005, 2005.

[125] MARSCHNER STEVE. CS667 LECTURE NOTES: RADIOMETRY. AVAILABLE ONLINE, SEPTEMBER 2009.

[126] MATHAR RUDOLF AND PFEIFER DIETHMAR. STOCHASTIK FÜR INFORMATIKER. LEITFÄDEN UND MONOGRAPHIEN DER INFORMATIK. B. G. TEUBNER VERLAGSGESELLSCHAFT, STUTTGART, ERSTE EDITION, 1990.

[127] MCCLUNEY WILLIAM ROSS. INTRODUCTION TO RADIOMETRY AND PHOTOMETRY. ARTECH HOUSE, BOSTON, LONDON, FIRST EDITION, 1994.

[128] MERTENS TOM, KAUTZ JAN, BEKAERT PHILIPPE, VAN REETH FRANK AND SEIDEL HANS-PETER. EFFICIENT RENDERING OF LOCAL SUBSURFACE SCATTERING. *In Proceedings of the 11th Pacific Conference on Computer Graphics and Applications*, PAGE 51, 2003.

[129] METROPOLIS N, ROSENBLUTH A. W., ROSENBLUTH M. N., TELLER A. H. AND TELLER E. EQUATIONS OF STATE CALCULATIONS BY FAST COMPUTING MACHINES. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[130] MEYN S. P. AND TWEEDIE R. L. MARKOV CHAINS AND STOCHASTIC STABILITY. SPRINGER-VERLAG, LONDON, FIRST EDITION, 1993.

[131] MIAOHUA JIANG AND ELIZABETH MCNAMARA. EVALUATING THE QUASI-MONTE CARLO METHOD FOR DISCONTINUOUS INTEGRANDS. PUBLICATIONS 10, DEPARTMENT OF MATHEMATICS, WAKE FOREST UNIVERSITY, WINSTON SALEM NC 27109, 2002.

[132] MILLER IRWIN AND MILLER MARYLEES. JOHN E. FREUND'S MATHEMATICAL STAISTICS. PRENTICE HALL, INC., SIXTH EDITION, 1999.

[133] MITCHELL DON P. AND NETREVALI ARUN N. . RECONSTRUCTION FILTERS IN COMPUTER GRAPHICS. *Computers and Graphics (Proc. Siggraph '88)*, 22(4):221–228, JULY 1988.

[134] MOROKOFF WILLIAM J. AND CALFISH RUSSEL E. QUASI-MONTE CARLO INTEGRATION. TECHNICAL REPORT, MATHEMATICAL DEPARTMENT, UCLA, 1995.

[135] NICODEMUS F. E. AND RICHMOND J.C. AND HSIA J.J. AND GINSBERG I.W. AND LIMPERIS T. GEOMETRICAL CONSIDERATIONS AND NOMENCLATURE FOR REFLECTANCE. *National Bureau of Standards (US)*, 1977.

[136] NICODEMUS FRED E. RADIANCE. *American Journal of Physics*, 31(5):368–377, 1963.

[137] NICODEMUS FRED E. SELF-STUDY MANUAL ON OPTICAL RADIATION MEASUREMENTS: PART I—CONCEPTS, CHAPTERS 4 AND 5. TECHNICAL NOTE NO. 910-2, NATIONAL BURO OF STANDARDS (US), FEBRUAR 1978.

[138] NIEDEREITER HARALD. RANDOM NUMBER GENERATION AND QUASI-MONTE CARLO METHODS. SIAM, PHILADELPHIA, FIRST EDITION, 1992.

[139] NISHITA T. AND NAKAMAE E. CONTINUOUS TONE REPRESENATTION OF 3-D OBJECTS TAKING ACCOUNT OF SHADOWS AND INTERREFELCTIONS. *Computer Graphics (Proc. Siggraph '85)*, PAGES 213–222, JULY 1984.

[140] NISHITA TOMOYUKI, MIYAWAKI YASUHIRO AND NAKAMAE EIHACHIRO. A SHADING MODEL FOR ATMOSPHERIC SCATTERING CONSIDERING LUMINOUS INTENSITY DISTRIBUTION OF LIGHT SOURCES. *Computer Graphics (Proceedings of SIGGRAPH 87)*, 21(4):303–310, 1987.

[141] OLANO MARC, HART JOHN C., HEIDRICH WOLFGANG AND MCCOOL MICHAEL. REAL-TIME SHADING. A K PETERS, NATICK, MASSACHUSETTS, FIRST EDITION, 2002.

[142] OREN MICHAEL AND NAYAR SHREE K. GENERALIZATION OF LAMBERT'S REFLECTANCE MODEL. *Computer Graphics (SIGGRAPH '94 Conference Proceedings)*, PAGES 239–246, JULY 1994.

[143] OWEN, A. B. ORTHOGONAL ARRAYS FOR COMPUTER EXPERIMENTS, INTEGRATION AND VISUALIZATION. IN *Statistica Sinica*, VOLUME 2, PAGES 439–452, 1992.

[144] OWEN ART B. LATTICE SAMPLING REVISITED: MONTE CARLO VARIANCE OF MEANS OVER RANDOMIZED ORTHOGONAL ARRAYS. *The Annals of Statistics*, 22:930 – 945, 1994.

[145] OWEN ART B. MONTE CARLO EXTENSION OF QUASI-MONTE CARLO. *Proceedings of the 1998 Winter Simulation Conference*, 1998.

[146] OWEN ART B. QUASI-MONTE CARLO SAMPLING. *In Monte Carlo Ray Tracing, SIGGRAPH 2003 Course 44*, PAGES 69–88, JULY 2003.

[147] PALMER JAMES H. AND GRANT BARABARA G. THE ART OF RADIOMETRY. SPIE PRESS BOOK, 2009.

[148] PALMER JAMES M. RADIOMETRY AND PHOTOMETRY FAQ.

[149] PATTANAIK S. N. AND MUDUR S.P. . EFFICIENT POTENTIAL EQUATION SOLUTIONS FOR GLOBAL ILLUMINATION COPUTATION. *Computers and Graphics*, 17(4):387–396, 1993.

[150] PATTANAIK S. N. AND MUDUR S.P. . THE POTENTIAL EQUATION AND IMPORTANCE IN ILLUMINATION COMPUTATIONS. *Computer Graphics Forum*, 12(2):131–136, 1993.

[151] PATTANAIK S. N. AND MUDUR S.P. . ADJOINT EQUATIONS AND RANDOM WALKS FOR GLOBAL ILLUMINATION COMPUTATION. *ACM Transactions on Graphics*, 14(1):77–102, 1995.

[152] PAULY MARK. ROBUST MONTE CARLO METHODS FOR PHOTOREALISTIC RENDERING OF VOLUMETRIC EFFECTS. MASTER'S THESIS, UNIVERSITÄT KAISERSLAUTERN, FACHBEREICH INFORMATIK, 1999.

[153] PEERCY M. S. LINEAR COLOR REPRESENTATIONS FOR FULL SPECTRAL RENDERING. *ACM SIGGRAPH '93 Proceedings*, PAGES 191–198, 1993.

[154] PEGORARO VINCENT. PARTICIPATING MEDIA, REALISTIC IMAGE SYNTHESIS, UNIVERSITÄT DES SAARLANDES, 2010.

[155] PÉREZ JOSÉ-PHILIPPE. OPTIK. SPEKTRUM AKADEMISCHER VERLAG GMBH, HEIDELBERG, ERSTE EDITION, 1996.

[156] PETTY GRANT W. A FIRST COURSE IN ATMOSPHERIC RADIATION. SUNDOG PUBLISHING, UNIVERSITY OF WISCONSIN-MADISON, MADISON, WISCONSIN, SECOND EDITION, 2006.

[157] PFANZAGL JOHANN. ELEMENTARE WAHRSCHEINLICHKEITSTHEORIE. WALTER DE GRUYTER, BERLIN, ZWEITE EDITION, 1991.

[158] PHARR MATT AND HUMPHREYS JEFFREY. PHYSICALLY BASED RENDERING. MORGAN KAUFMANN PUBLISHERS, INC., FIRST EDITION, 2004.

[159] PHARR MATT AND HUMPHREYS JEFFREY. PHYSICALLY BASED RENDERING. MORGAN KAUFMANN PUBLISHERS, INC., SECOND EDITION, 2010.

[160] PHONG BUI-TONG. ILLUMINATION FOR COMPUTER GENERATED PICTURES. *Communications of the ACM*, 18(6):311–317, JUNE 1975.

[161] PIPKIN ALLEN C. A COURSE ON INTEGRAL EQUATIONS. TEXTS IN APPLIED MATHEMATICS 9. SPRINGER-VERLAG, NEW YORK, FIRST EDITION, 1991.

[162] PITTMAN JIM. PROBABILITY. SPRINGER, NEW YORK, FIRST EDITION, 1999.

[163] POMRANING G. C. THE EQUATIONS OF RADIATION HYDRODYNAMICS. PERGAMON PRESS, NEW YORK, 1973.

[164] PRAHL SCOTT ALAN. *Light Transport in Tisuue*. PHD THESIS, UNIVERSITY OF TEXAS AT AUSTIN, 1988.

[165] PREISENDORFER RUDOPH W. RADIATIVE TRANSFER IN DISCRETE SPACES. PERGAMON PRESS INC., 1965.

[166] PREISENDORFER RUDOPH W. HYDROLOGIC OPTICS, VOLUME I. INTRODUCTION. U.S: DEPARTMENT OF COMMERCE, NATIONAL OCEANIC & ATMOSPHERIC ADMINISTRATION ENVIRONEMENTAL RESEARCH LABORATORIES, HONOLULU, HAWAII, 1976.

[167] PREISENDORFER RUDOPH W. AND TYLER JOHN E. RADIANCE. TECHNICAL REPORT, VISIBILITY LABORATORY, UNIVERSITY OF CALIFORNIA, SAN DIEGO, 1958.

[168] PREISENDORFER RUDOPH W. AND TYLER JOHN E. THE MEASUREMENT OF LIGHT IN NATURAL WATERS. GRIPPS INSTITUTION OF OCEANOGRAPHY, NOVEMBER 1958.

[169] REDDY B. DAYA. INTRODUCTORY FUNCTIONAL ANALYSIS. TEXTS IN APPLIED MATHEMATICS 27. SPRINGER-VERLAG, 1991.

[170] ROBERT CHRISTIAN P. AND CASELLA GEORGE. MONTE CARLO STATISTCAL METHODS. SPRINGER, 1999.

[171] ROSS SHELDON M. INTRODUCTION TO PROBABILITY MODELS. ACADEMIC PRESS, SAN DIEGO, SEVENTH EDITION, 2000.

[172] RUBINSTEIN REUVEN Y. SIMULATION AND THE MONTE CARLO METHODS. JOHN WILEY & SONS, NEW YORK, FIRST EDITION, 1981.

[173] RUDIN WALTER. PRINCIPLES OF MATHEMATICAL ANALYSIS, VOLUME THIRD. MCGRAW-HILL INC., NEW YORK SAN FRANCISCO LONDON SYDNEY TOKYO TORONTO, 1976.

[174] RUDIN WALTER. ANALYSIS. OLDENBURG, ROSENHEIMER STRASSE 145, D-81671 MÜNCHEN, 1998.

[175] RUSINKIEWICZ SZYMON. A SURVEY OF BRDF REPRESENTATION FOR COMPUTER GRAPHICS, WINTER 1997.

[176] RYNNE BRYAN P. AND YOUNGSON MARTIN A. LINEAR FUNCTIONAL ANALYSIS, VOLUME SECOND. SPRINGER, 2008.

[177] SCHIRMACHER HARTMUT. HIERARCHISCHE VOLUMEN-RADIOSITY. TECHNICAL REPORT, FRIEDRICH-ALEXANDER-UNIVERSITÄT ERLANGEN-NÜRNBERG, 1996.

[178] SCHLICK CHRISTOPHE. A CUSTOMIZABLE REFLECTANCE MODEL FOR EVERYDAY RENDERING. *In Fourth Eurographics Workshop on Rendering*, *EUROGRAPHICS*, PAGES 73–84, JUNE 1993.

[179] SCHMEISSER GERHARD AND SCHIRMEIER HORST. PRAKTISCHE MATHEMATIK. WALTER DE GRUYTER, NEW YORK, 1976.

[180] SCHRÖDER P. AND HANRAHAN P. A CLOSED FORM EXPRESSION FOR THE FORM FACTOR BETWEEN TWO POLYGONS. TECH. REP. CS-404–93, DEPARTMENT OF COMPUTER SCIENCE, PRINCETON UNIVERSITY, 1993.

[181] SHIRLEY PETER S. *Physically Based Lighting Calculations for Computer Graphics*. PHD THESIS, UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN, 1990.

[182] SHIRLEY PETER S. DISCREPANCY AS A QUALITY MEASURE FOR SAMPLE DISTRIBUTIONS. *Proceedings of EUROGRAPHICS*, PAGES 183–199, 1991.

[183] SHIRLEY PETER S. *Physically Based Lighting Calculations for Computer Graphics*. PHD THESIS, UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN, 1991.

[184] SHIRLEY PETER S. HYBRID RADIOSITY AND MONTE CARLO METHODS. *ACM SIGGRAPH '94 Advanced Topics in Radiosity Course Notes, Chapter 11*, PAGES 1–24, 1994.

[185] SHIRLEY PETER S. REALISTIC RAY TRACING. A K PETERS, NATIK, MASSACHUSETTS, FIRST EDITION, 2000.

[186] SHIRLEY PETER S. FUNDAMENTALS OF COMPUTER GRAPHICS. A K PETERS, NATIK, MASSACHUSETTS, FIRST EDITION, 2002.

[187] SHIRLEY PETER S. AND MORLEY R. KEITH. REALISTIC RAY TRACING. A K PETERS, SECOND EDITION, 2003.

[188] SHIRLEY PETER S., WANG CHANGYAW AND ZIMMERMANN KURT. MONTE CARLO TECHNIQUES FOR DIRECT LIGHTING COMPUTATIONS. *ACM Transactions on Graphics*, 15(1):1–36, JANUARY 1996.

[189] SIEGEL R., AND HOWELL J. R. THERMAL RADIATION HEAT TRANSFER. HEMISPHERE PUBLISHING CORPORATION, NEW YORK, THIRD EDITION, 1992.

[190] SILLION FRANCOIS X. AND PUECH CLAUDE. RADIOSITY AND GLOBAL ILLUMINATION. MORGAN KAUFMANN PUBLISHERS, INC., SAN FRANCISCO, CALIFORNIA, FIRST EDITION, 1994.

[191] SLUSALLEK PHILIPP. *Vision - An Architecture for Physically-Based Rendering*. PHD THESIS, TECHNISCHE FAKULTÄT DER UNIVERSITÄT ERLANGEN-NÜRNBERG, 1995.

[192] SLUSALLEK PHILIPP, COHEN MICHAEL F. AND GORTLER STEVEN. *Radiosity and Relaxation Methods, Progressive Refinement is Southwell Relaxation*. UNIVERSITÁT TÚBINGEN, PRINCETON UNIVERSITY, 1993.

[193] SLUSALLEK PHILIPP, PAUL JEAN-CLAUDE AND SILLION FRANCOIS X. *Using Realistic Lighting in Modern Graphics Applications*. EUROGRAPHICS, POITIER, 1996.

[194] SLUSALLEK PHILIPP UND STAMMINGER MARK, 2000. PHOTOREALISTISCHE BILDSYNTHESE, VORLESUNG COMPUTER GRAPHIK, UNIVERSITÄT DES SAARLANDES.

[195] SNYDER WILLIAM C. AND WAN ZHENGMING. BRDF MODELS TO PREDICT SPECTRAL REFLECTANCE AND EMISSIVITY IN THE THERMAL INFRARED. *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, 36(1), JANUARY 1998.

[196] SNYDER WILLIAM C. AND WAN ZHENGMING. RECIPROCITY OF THE BIDIRECTIONAL REFLECTANCE DISTRIBUTION FUNCTION (BRDF) IN MEASUREMENTS AND MODELS OF STRUCTURED SURFACES. *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, 36(2), MARCH 1998.

[197] SOBOL I. M. DIE MONTE-CARLO-METHODE. DEUTSCHER VERLAG DER WISSENSCHAFTEN, BERLIN, VIERTE EDITION, 1985.

[198] SPANIER JEROME AND GELBARD E.M. MONTE CARLO PRINCIPLES AND NEUTRON TRANSPORT PROBLEMS. ADDISON WESLEY, READING, MASSACHUSETTS, 1969.

[199] SPANIER JEROME AND LI LIMING. QUASI-MONTE CARLO METHODS FOR INTEGRAL EQUATIONS. *In Lecture Notes in Computer Statistics, Springer, 127*, PAGES 382–397, 1998.

[200] SPARROW, E., AND CESS, R. RADIATION HEAT TRANSFER. HEMISPHERE PUBLISHING CORPORATION, WASHINGTON, 1978.

[201] SPIEGEL MURRAY R.. VECTOR ANALYSIS AND AN INTRODUCTION TO TENSOR ANALYSIS. SCHAUM'S OUTLINE SERIES, MCGRAW-HILL, INC., NEW YORK, 1995.

[202] STOER JOSEF. EINFÜHRUNG IN DIE NUMERISCHE MATHEMATIK, VOLUME FIRST. SPRINGER-VERLAG, BERLIN HEIDELBERG NEW YORK, 1979.

[203] STOER JOSEF AND BULIRSCH ROLAND. EINFÜHRUNG IN DIE NUMERISCHE MATHEMATIK, VOLUME SECOND. SPRINGER-VERLAG, BERLIN HEIDELBERG NEW YORK, 1978.

[204] STROOCK DANIEL W. AN ITRODUCTION TO MARKOV PROCESSES. SPRINGER, BERLIN, FIRST EDITION, 2005.

[205] SUFFERN KEVIN. RAY TRACING FROM THE GROUND UP. A K PETERS, A K PETERS, LTD. 888 WORCESTER STREET, SUITE 230, WELLESLEY, MA 02482, FIRST EDITION, 2007.

[206] SUN YINLONG, DREW MARK S. AND FRACCHIA F. DAVID. REPRESENTING SPECTRAL FUNCTIONS BY A COMPOSITE MODEL OF SMOOTH AND SPIKY COMPONENTS FOR EFFICIENT FULL-SPECTRUM PHOTOREALISM. *IEEE Workshop on Photometric Modeling of Computer Vision and Graphics*, PAGES 4–11, JUNE 1999.

[207] SUYKENS- DE LAET FRANK. *On Robust Monte Carlo Algorithms for Multi-pass Global Illumination*. PHD THESIS, KATHOLIKE UNIVERISTAIT LEUVEN, 2002.

[208] SZIRMAY-KALOS LÁSZLÓ. MONTE CARLO METHODS IN GLOBAL ILLUMINATION. SCRIPT, INSTITUTE OF COMPUTER GRAPHICS, VIENNA UNIVERSITY OF TECHNOLOGY, 1999.

[209] SZIRMAY-KALOS LÁSZLÓ. *Photorealistic Image Synthesis using Ray-Bundles*. PHD THESIS, DEPARTMENT OF CONTROL ENGINEERING AND INFORMATION TECHNOLOGY TECHNICAL UNIVERSITY OF BUDAPEST, 2000.

[210] SZIRMAY-KALOS LÁSZLÓ AND PURGATHOFER WERNER. ANALYSIS OF THE QUASI-MONTE CARLO INTEGRATION OF THE RENERING EQUATION. *In Winter School of Computer Graphics '99*, PAGES 281–288, 1999.

[211] TALBOT LEO. *Radiosity*. DEPARTMENT OF COMPUTER SCIENCE, TRINITY COLLEGE, DUBLIN, 1999.

[212] TAN KENS SENG AND BOYLE PHELIM P. APPLICATIONS OF RANDOMIZED LOW DISCREPANCY SEQUENCES TO THE VALUATION OF COMPLEX SECURITIES. *Journal of Economic Dynamics & Control*, 24:1747–1782, 2000.

[213] TAYLOR ANGUS E. AND LAY DAVID C. INTRODUCTION TO FUNCTIONAL ANALYSIS. KRIEGER PUBLISHING, MALABAR, FLORIDA, SECOND EDITION, 1986.

[214] TAYLOR J. E. AN INTRODUCTION TO MEASURE AND PROBABILITY. SPRINGER-VERLAG, NEW YORK BERLIN HEIDELBERG, SECOND EDITION, 1997.

[215] TORRENCE K. E. AND SPARROW E.M. THEORY FOR OFF-SPECULAR RELFECTION FROM ROUGHENED SURFACES. *Journal of the Optical Society of Amerika*, 1(57(9)):1104–1114, 1967.

[216] TSANG LEUNG AND KONG JIN AU. SCATTERING OF ELECTROMAGNETIC WAVES, VOLUME ADVANCED TOPICS. JOHN WILEY & SONS, INC, FIRST EDITION, 2001.

[217] TSANG LEUNG, KONG JIN AU AND DING KUNG-HAU. SCATTERING OF ELECTROMAGNETIC WAVES. JOHN WILEY & SONS, INC, NEW YORK, NEW YORK, USA, FIRST EDITION, 2000.

[218] VAN DE HULST H.C. LIGHT SCATTERING BY SMALL PARTICLES. DOVER PUBLICATIONS, INC., NEW YORK, SECOND EDITION, 1981.

[219] VEACH ERIC. NON-SYMMETRIC SCATTERING IN LIGHT TRANSPORT ALGORITHMS. *Eurographics Eurographics RenderingWorkshop 1996 Proceedings. Also in Rendering Techniques '96, Springer-Verlag, New York*, 1996.

[220] VEACH ERIC. CS448: TOPICS IN COMPUTER GRAPHICS, MATHEMATICAL MODELS FOR COMPUTER GRAPHICS. STANFORD UNIVERSITY, 1997.

[221] VEACH ERIC. *Robust Monte Carlo Methods for Light Transport Simulation*. PHD THESIS, STANFORD UNIVERSITY, 1998.

[222] VEACH ERIC AND GUIBAS LEONIDAS. OPTIMALLY COMBINIG SAMPLING TECHNIQUES FOR MONTE CARLO RENDERING. *SIGGRAPH' 95 Proceedings, Addison-Wesley*, PAGES 419–428, 1995.

[223] VEACH ERIC AND GUIBAS LEONIDAS. METROPOLIS LIGHT TRANSPORT. *In SIGGRAPH 97 Proceedings, Addison-Wesley*, PAGES 65–76, 1997.

[224] VON NEUMANN JOHN. VARIOUS TECHNIQUES USED IN CONNECTION WITH RANDOM DIGITS. *In U.S. Nat Bur. Stand. Applied. Math. Ser.*, (12):36–38, 1951.

[225] WALD INGO. *Real Time Raytracing and Interactive Global Illumination*. PHD THESIS, *Computer Graphics Group, Saarland University, Saarbrücken, Germany*, 2004.

[226] WALD INGO, BENTHIN CARSTEN, AND SLUSALLEK PHILIPP. INTERACTIVE GLOBAL ILLLUMINATION IN COMPLEX AND HIGHLY OCCLUDED ENVIRONEMENTS. *In Eurographics Symposium on Rendering: 14th Eurographics Workshop on Rendering*, PAGES 74–81, 2003.

[227] WALD INGO, DIETRICH ANDREAS AND SLUSALLEK PHILIPP. AN INTERACTIVE OUT-OF-CORE RENDERING FRAMEWORK FOR VISUALIZING MASSIVELY COMPLEX MODELS. *Rendering Techniques 2004: Eurographics Symposium on Rendering, Norrkoeping, Sweden, 2004.*, PAGES 81–92, 2004.

[228] WALD INGO, KOLLIG THOMAS, BENTHIN CARSTEN, KELLER ALEXANDER AND SLUSALLEK PHILIPP. INTERACTIVE GLOBAL ILLLUMINATION USING FAST RAY TRACING. *In Rendering Techniques 2002: 13th Eurographics Workshop on Rendering*, PAGES 15–24, 2002.

[229] WALTER BRUCE. NOTES ON THE WARD BRDF. TECHNICAL REPORT PCG-05-06, CORNELL PROGRAM OF COMPUTER GRAPHICS, 2005.

[230] WARD GREGORY J. REAL PIXELS. *Graphics Gems II, edited by James Arvo*, PAGES 80–83, 1991.

[231] WARD GREGORY J. MEASURING AND MODELING ANISOTROPIC REFLECTION. *Computer Graphics*, 26(2), JULY 1992.

[232] WATT ALAN. 3D-COMPUTER GRAPHICS. ADDISON WESLEY, NEW YORK, THIRD EDITION, 1999.

[233] WATT ALAN AND WATT MARK. ADVANCED ANIMATION AND RENDERING TECHNIQUES, THEORY AND PRACTICE. ADDISON WESLEY, NEW YORK, FIRST EDITION, 1992.

[234] WAZWAZ ABDUL-MAJID. A FIRST COURSE IN INTEGRAL EQUATIONS. WORLD SCIENTIFIC PUBLISHING CO. PTE. LTD., PO BOX 128, FARRER ROAD, SINGAPORE 912805, FIRST EDITION, 1997.

[235] WEISSTEIN ERIC W. CRC CONCIS ENCYCLOPEDIA OF MATHEMATICS. CHAPMAN & HALL/CRC, BOCA RATON, FLORIDA, SECOND EDITION, 2003.

[236] WHITTED TURNER. AN IMPROVED ILLUMINATION MODEL FOR SHADED DISPLAY. *Communications of the ACM*, 32(6):343–349, 1992.

[237] WONG TIEN-TSIN, LUK WAI-SHING AND HENG PHENG-ANN. SAMPLING WITH HAMMERSLEY AND HALTON POINTS. *Journal of Graphics Tools*, 2:9–24, 1997.

[238] WYNN CHRIS. AN INTRODUCTION TO BRDF-BASED LIGHTING. *NVIDIA Corporation*, 2006.

[239] YOSHIDA KÔSAKU. FUNCTIONAL ANALYSIS. SPRINGER-VERLAG, BERLIN HEIDELBERG NEW YORK, 1980.

[240] ZATZ HAROLD R. *Galerkin Radiosity: A Higher Order Solution Method for GLobal Illumination*. PHD THESIS, CORNELL UNIVERSITY, 1992.

[241] ZEIDLER EBERHARD. APPLIED FUNCTIONAL ANALYSIS. SPRINGER, TEXTS IN APPLIED MATHEMATICAL SCIENCES 108, NEW YORK, FIRST EDITION, 1995.

[242] ZSOLT SÁNDOR AND TRAIN KENNETH. QUASI-RANDOM SIMULATION OF DISCRETE CHOICE MODELS. TECHNICAL REPORT, ERASMUS UNIVERSITY ROTTERDAM AND UNIVERSITY OF CALIFORNIA, BERKELEY, 2004.

INDEX