Clustering with Neighborhood Graphs

Dissertation

zur Erlangung des Grades des

Doktors der Naturwissenschaften

der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes.

Eingereicht in Saarbrücken im Jahre 2009 von

Dipl.-Inform. Markus Martin Maier

Tag des Kolloquiums: 22. Februar 2010

Dekan der Naturwisseschaftlich-Technischen Fakultät I: Prof. Dr. Joachim Weickert

Mitglieder des Prüfungsausschusses:

- Prof. Dr. Markus Bläser (Vorsitzender)
- Prof. Dr. Matthias Hein (Berichterstatter)
- Dr. Ulrike von Luxburg (Berichterstatterin)
- Dr. Matthias Seeger (akademischer Mitarbeiter)

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbständig 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.

Tübingen, den 26. Februar 2010

Markus Maier

Inhaltsverzeichnis

Thanks

First of all I would like to thank Bernhard Schölkopf for giving me the opportunity to do the research for this thesis in an open and stimulating environment. I appreciate that he gave me the freedom to follow my research interests in theoretical machine learning and that he advised and supported me in many respects.
I am deeply indebted to my two supervisors Matthias Hein and Ulrike von Luxburg. Matthias Hein introduced me to learning theory while he was still in Tübingen. He kindly agreed to continue to supervise me and collaborate with me after becoming a professor in Saarbrücken. Furthermore, he gave me the opportunity to graduate at Saarland University and supported me very much during the last steps towards the completion of the thesis. Ulrike von Luxburg has been supervising me for more than three years, since she began to work as a postdoc at the MPI for Biological Cybernetics. She was always ready to give valuable advice on research topics and beyond, to discuss openly any question and to support me whenever necessary. Both Matthias Hein and Ulrike von Luxburg are coauthors of my publications. I am grateful to them for the fruitful collaboration and for allowing me to use parts of our papers as a basis for this thesis.
I am also thankful to all current and former members of AGBS. I enjoyed the open atmosphere in the group and that people were always ready to give advice and share their knowledge. In particular I would like to thank Agnes Radl, Jacquelyn Shelton and Michael Hirsch who have read parts of this thesis and my office mates over the years, Florian Steinke, Jan Eichhorn, Kenji Fukumizu, Suvrit Sra and Matthias Hofmann for many interesting conversations, valuable discussions and also many cheerful hours outside the office.
Finally, I would like to thank my family for their love and support during my studies: my parents, Annemarie and Willi Maier, and my girlfriend Mariya.

Summary in English

Graph clustering methods are defined for general weighted graphs. If data is given in the form of points and distances between them, a neighborhood graph, such as the r-graph or kNN-graphs, is constructed and graph clustering is applied to this graph. We investigate the influence of the type and parameter of the neighborhood graph on the clustering results, when n sample points are drawn independently from a density in Euclidean space.

In Chapter 2 we study "cluster identification": the true clusters are the connected components of density level sets and a cluster is identified if its points are a connected component of the graph. We compare (modifications of) the mutual and the symmetric kNN-graph. They behave differently if the goal is to identify the "most significant" clusters, whereas there is no difference if the goal is to identify all clusters. We give the range of k for which the clusters are identified in the graphs and derive the optimal choice of k, which, surprisingly, is linear in n.

In Chapter 3 we study the convergence of the normalized cut (Ncut) and the ratio cut as $n \to \infty$ for cuts in the kNN- and the r-graph induced by a hyperplane. The limits differ; consequently Ncut on a kNN-graph does something systematically different than Ncut on an r-graph! This can be experimentally observed on toy and real data sets. Therefore, graph clustering criteria cannot be studied independently of the type of graph to which they are applied.

Zusammenfassung in deutscher Sprache

Graphclustering ist für gewichtete Graphen definiert. Liegen Daten jedoch in Form von Punkten und Abständen zwischen ihnen vor, wird zuerst ein Nachbarschaftsgraph wie der r-Graph oder kNN-Graphen konstruiert, auf den dann Graphclustering angewandt wird. In dieser Arbeit wird der Einfluss des Nachbarschaftsgraphen auf die Clusteringergebnisse untersucht, wenn n Punkte unabhängig voneinander von einer Dichte im euklidischen Raum gezogen werden.
In Kapitel 2 wird das Problem der "Clusteridentifizierung" betrachtet: die Cluster sind die Zusammenhangskomponenten einer Dichteniveaumenge. Ein Cluster wird identifiziert, wenn seine Punkte eine Zusammenhangskomponente des Graphen bilden. Modifikationen verschiedener kNN-Graphen werden verglichen. Sollen nur die "signifikantesten" Cluster gefunden werden, unterscheidet sich ihr Verhalten, nicht jedoch für die Identifizierung aller Cluster. Es wird gezeigt, für welche k die Cluster identifiziert werden und dass die optimale Wahl von k linear in n ist.
In Kapitel 3 wird die Konvergenz der Kriterien "normalized cut" (Ncut) und "ratio cut" für Schnitte im kNN- und r-Graphen, die von einer Hyperebene induziert werden, gezeigt. Die Grenzwerte unterscheiden sich. Folglich bewirkt Ncut auf einem kNN-Graphen etwas anderes als Ncut auf einem r-Graphen. Dieser Effekt kann experimentell beobachtet werden. Daraus folgt, dass Graphclusteringkriterien nicht getrennt vom Graphtyp betrachtet werden können.

1 Introduction

Some of the most successful clustering algorithms described in the literature are graph clustering algorithms, which find meaningful subgraphs of a given graph. However, we often deal with data that is given in the form of sample points and distances or similarities between them. This is the case, for example, for points sampled from some probability distribution in an underlying metric space. In order to apply graph clustering algorithms in such a setting we first have to construct a graph on the points that reflects the similarity or dissimilarity between them, for example a neighborhood graph. We refer to clustering methods that consist of the construction of a neighborhood graph and the subsequent application of a graph clustering algorithm as graph-based clustering methods.

In machine learning different types of neighborhood graphs (for example, the r-neighborhood graph or the k-nearest neighbor graph) are used, and in the construction of each graph a neighborhood parameter must be chosen (r or k, respectively). Therefore, it is vital to investigate how the choices of the graph type and of the parameter influence the overall results of a method based on neighborhood graphs. In fact, understanding this relationship is even more important in the domain of unsupervised learning, including clustering, than in the domain of supervised learning, since in the former domain we cannot use cross-validation to select the graph type and the parameter. However, for clustering neither empirical studies have been conducted (for example: how sensitive are the results to the graph parameters?), nor do theoretical results exist which lead to well-justified heuristics — rather different researchers use their "gut feeling" to set these parameters. Therefore it is an essential line of research to analyze the interplay between the choice of neighborhood graph type and its parameter and the performance of graph-based clustering algorithms. The final goal of this research is to develop a theoretically sound procedure to choose the graph type and the parameter according to the properties of the desired clustering and to the properties of the distribution of the sample points.

In this thesis we analyze the behavior of different neighborhood graphs and choices of parameter in two graph-based clustering settings. In the following section we give an introduction to graph clustering in general as well as to spectral clustering, one of the most important graph clustering methods. We present the setting of the clustering of points from a density in Euclidean space in which we will apply graph-based clustering methods. In this setting it is possible to study the consistency of a clustering method. We shortly introduce different notions of clustering consistency, since these give hints on which quantities to consider in order to estimate how well an algorithms works. In Section 1.2 we give an overview over different types of neighborhood graphs and present some literature that is interesting for the study of neighborhood graphs in

clustering. In Section 1.3 we provide an overview of the results in this thesis. Finally, in Section 1.4 we introduce some formal definitions and notations that will be important for the rest of this thesis.

1.1 Clustering

Clustering is one of the most important techniques for exploratory data analysis and is used in a wide range of disciplines, such as biology, medicine, psychology, marketing and the social sciences. Interestingly, the study of modern clustering methods began with work in biological taxonomy (see Jardine and Sibson [50]). Hartigan [41] defines clustering as "the grouping of similar objects"; according to Jain and Dubes [49] clustering is "the unsupervised classification of patterns into groups".
Although clustering is an important problem in many areas of application a general theory of clustering has not yet been developed. Indeed, little is known about the theoretical properties of clustering (von Luxburg and Ben-David [86]; refer to this work also for some of the challenges in clustering theory). In Kleinberg [51] a set of axioms for clustering functions that are "independent of any particular algorithm, objective function, or generative data model" is suggested. However, it is shown that, although each of the axioms sounds plausible, there cannot exist a clustering function that satisfies all of them. From this result often the conclusion is drawn that it is impossible to develop a general theory of clustering, without referring to a "particular algorithm, objective function, or generative data model", although not all authors share this opinion (see Ackerman and Ben-David [1]).
The problem of the ill-definedness of clustering can be circumvented by formalizing clustering as the optimization of a clustering quality measure. A clustering quality measure is "a function that, given a data set and its partition into clusters, returns a non-negative real number representing how strong or conclusive the clustering is" (Ackerman and Ben-David [1]). The goal of a clustering algorithm is to find the clustering that optimizes the clustering quality measure. However, for most clustering quality measures used in practice the optimal clustering cannot be efficiently computed. A discussion of the desirable properties of clustering quality measures, an axiomatization and several examples can be found in Ackerman and Ben-David [1].
In this thesis two clustering scenarios will be of special importance: graph clustering, that is, the clustering of graphs into subgraphs, and the clustering of points drawn from a density in Euclidean space. In particular, we study the application of graph clustering algorithms to neighborhood graphs that are constructed on points from a density in Euclidean space.

Graph clustering In graph clustering we assume that we are given a graph consisting of a set of vertices and a set of (possibly weighted) edges between them. The goal is to find a clustering of the nodes, that is, a partition of the nodes into meaningful subgraphs. Note that graph clustering in this sense should not be confused with the

clustering of sets of graphs. A review of the principles and methods of graph clustering is given in Schaeffer [77].
Graph clustering is often based on the fundamental paradigm of intra-cluster density versus inter-cluster sparsity (see Gaertler [36]). The ideal case of a clustering would be disjoint cliques with only few connections between each other. This trade-off between intra-cluster density and inter-cluster sparsity is reflected in many graph clustering quality measures, such as Coverage, Conductance and Performance. An overview of common graph clustering quality measures is given in Gaertler [36], many of which rely on the cut in order to quantify "inter-cluster sparsity". The cut of a partition into two subgraphs is defined as the sum of the weights of the edges from one subgraph to the other. That is, the lower the cut of a cluster, the better it is in terms of "inter-cluster sparsity". However, the cut does not measure "intra-cluster density" and it is typically minimized by very unbalanced partitions, for example cutting off just one vertex. Therefore, a suitable graph clustering quality measure has to compensate for this effect by penalizing partitions with "low intra-cluster density" or by penalizing unbalanced partitions. A cluster has a low "intra-cluster density" if, for example, the ratio of the number of edges in the cluster and the maximally possible number of edges between the points in the cluster (that is, the number of edges in the complete graph with the same number of points) is low. A partition is unbalanced if, for example, the distribution of vertices between the partitions is uneven, or if the total weight of edges originating in a certain partition differs much between partitions. Combining the cut with slight variations of these measures of "balancedness" leads to the graph clustering quality measures Conductance, Ncut and RatioCut. Having such a graph clustering quality measure we would ideally find a clustering by optimizing the quality measure over all possible clusterings. However, most of the relevant clustering quality measures, such as the Conductance, Ncut and RatioCut cannot be efficiently optimized (see Šíma and Schaeffer [89] and Wagner and Wagner [90]). For some graph clustering quality measures relying on a variant of the cut, there are efficient approximation algorithms (see, for example, Arora et al. [3], Arora et al. [4], and Fernandez de la Vega et al. [33]). However, the most popular graph clustering method based on graph clustering quality measures is spectral clustering, which itself is based on Ncut and RatioCut. Here the minimization problem is formulated as a discrete optimization problem, the relaxation of this problem is efficiently solved and the solution is re-transformed to a discrete indicator vector.
Due to the popularity of spectral clustering we only deal with Ncut and RatioCut in Chapter 3, although the Conductance has some theoretically appealing properties. However, the results there can be easily carried over to the Conductance and similar quality measures. We now give a more detailed review of spectral clustering and its relation to Ncut and RatioCut.

A Review of Spectral Clustering One of the most popular graph clustering algorithms is spectral clustering, which is based on the graph clustering quality measures Ncut and RatioCut. It is simple to implement and can be solved efficiently by standard linear

algebra methods. Here we give a short review of spectral clustering for the case of two clusters, which follows von Luxburg [85]. However, spectral clustering can be generalized to arbitrary numbers of clusters m.
As mentioned above, the cut can be used in graph clustering quality measures to measure inter-cluster sparsity. In order to introduce a balancedness constraint we can use the volume vol. These quantities are defined as follows: Given an undirected graph $G = (V, E)$ with nodes $V = \{v_1, \dots, v_n\}$, edges E and weights $w : E \to \mathbb{R}$, and a partition of the nodes V into $(C, V \setminus C)$ we define

$$\operatorname{cut}(C, V \setminus C) = \sum_{u \in C, v \in V \setminus C} \big(w(u,v) + w(v,u)\big)$$

and $\operatorname{vol}(C) = \sum_{u \in C, v \in V} w(u,v)$.
Based on these two elements we define the graph clustering quality measures

$$\operatorname{Ncut}(C, V \setminus C) = \operatorname{cut}(C, V \setminus C)\left(\frac{1}{\operatorname{vol}(C)} + \frac{1}{\operatorname{vol}(V \setminus C)}\right)$$

or "normalized cut" and

$$\operatorname{RatioCut}(C, V \setminus C) = \operatorname{cut}(C, V \setminus C)\left(\frac{1}{|C|} + \frac{1}{|V \setminus C|}\right).$$

Unfortunately, contrary to the computation of the minimum cut, the optimization of Ncut and RatioCut cannot be computed efficiently (see Wagner and Wagner [90]). However, the problem $\min_{C \subseteq V} \operatorname{Ncut}(C, V \setminus C)$ can be shown to be equivalent, up to a constant factor of two, to the discrete optimization problem

$$\begin{aligned} &\min_C f' L f && (1.1)\\ &\text{subject to} \quad f = (f_1, \dots, f_n) \text{ with } f_i = \begin{cases} \sqrt{\frac{\operatorname{vol}(V \setminus C)}{\operatorname{vol}(C)}} & \text{if } v_i \in C \\ -\sqrt{\frac{\operatorname{vol}(V \setminus C)}{\operatorname{vol}(C)}} & \text{if } v_i \notin C \end{cases} \\ &\qquad\qquad Df \perp \mathbf{1} \\ &\qquad\qquad f' D f = \operatorname{vol}(V), \end{aligned}$$

where $D = \operatorname{diag}(d_1, \dots, d_n)$ with $d_i = \sum_{j=1}^n w_{ij}$, $L = D - W$ with $W = ((w_{ij}))_{i,j}$ the weight matrix, and $\mathbf{1} \in \mathbb{R}^n$ the vector with all entries 1. The matrix L is called the unnormalized graph Laplacian. Similarly, the problem $\min_{C \subseteq V} \operatorname{RatioCut}(C, V \setminus C)$ can be formulated as

$$\begin{aligned} &\min_C f' L f && (1.2)\\ &\text{subject to} \quad f = (f_1, \dots, f_n) \text{ with } f_i = \begin{cases} \sqrt{\frac{|V \setminus C|}{|C|}} & \text{if } v_i \in C \\ -\sqrt{\frac{|V \setminus C|}{|C|}} & \text{if } v_i \notin C \end{cases} \\ &\qquad\qquad f \perp \mathbf{1} \\ &\qquad\qquad \|f\| = \sqrt{n}. \end{aligned}$$

Of course, these discrete optimization problems still cannot be solved efficiently. The idea in spectral clustering is to relax the integrality condition and solve for the optimal vector $f \in \mathbb{R}^n$, which can be computed efficiently by linear algebra methods, namely by solving a (generalized) eigenvalue problem involving the graph Laplacian matrix L. In a final step, the real-valued solution vector f of the relaxed problem has to be re-transformed into a discrete indicator vector. In the case of two clusters the simplest way is to use the sign of f as the indicator function, that is, to put the nodes v_i with $f_i < 0$ into one cluster and to put the nodes v_i with $f_i \geq 0$ into the other cluster. A heuristic improvement is the "best threshold cut algorithm": The vertices are sorted according to their f-values. For each index $j = 1, \ldots, n-1$ we compute the Ncut- or RatioCut-value of the cut given by a splitting of the vertices into those with sorted index $\leq j$ and those with sorted index $> j$. The split that provides the best Ncut- or RatioCut-value is chosen. Another variant to obtain a discrete indicator vector considers the coordinates of f_i as points in $\mathbb{R}$ and clusters them into two groups using the k-means algorithm. This variant can be adapted to the case of more than 2 clusters and is used in the experiments in Chapter 3.

Spectral clustering based on the normalized cut objective, that is, the optimization problem in Equation (1.1), is called "normalized spectral clustering" algorithm, whereas the spectral algorithm based on the RatioCut objective, that is, the optimization problem in Equation (1.2), is called "unnormalized spectral clustering" algorithm. The consistency of normalized spectral clustering was shown in von Luxburg et al. [87].

Although spectral clustering works well in practice there are no constant-factor approximation guarantees: Guattery and Miller [40] introduce the so-called "roach graphs" as an example where spectral clustering using the sign of the elements of f to assign the nodes to clusters performs badly: For a roach graph with n nodes the ratio of the best cut found by spectral clustering and the optimal cut is (up to a constant) lower bounded by n. If we use spectral clustering with the best threshold cut, we achieve a constant factor approximation on the roach graphs. However, on the so-called tree-cross-path graphs, which were introduced in Guattery and Miller [40] as well, the ratio of the best cut found by the best threshold cut algorithm and the optimal cut is (up to a constant) lower bounded by $\sqrt[3]{n}$, where n again denotes the number of vertices. Furthermore, it is shown in the same paper that any spectral algorithm that chooses a threshold t and puts the nodes with $f_i < t$ into one cluster and the nodes with $f_i \geq t$ into the other cluster cannot achieve a better quotient than $\sqrt[3]{n}$ in the worst case, no matter how the threshold t is computed.

Bühler and Hein [17] present a generalized version of spectral clustering that uses the graph p-Laplacian instead of the standard graph Laplacian. It is shown that for $p \to 1$ this algorithm converges to the cut with minimum Conductance, which is bounded by twice the minimum of the normalized cut. In experiments the results of this algorithm are often better than that of standard spectral clustering, but it is numerically much more involved.

Recently, there has been some interest in spectral properties of directed graphs (see, for example, Chung [21]) and spectral clustering methods for these (see Meila and Pentney [62]). However, in this thesis we apply spectral clustering only to undirected graphs.

When does graph clustering work? One way to study how well a given clustering method works is to define a clustering quality measure and show that the clusterings found by the method are sufficiently close to the optimal value of the quality measure achievable on the given data. If the graph clustering algorithm itself is based on a different quality measure, this approach seems inconsistent. However, there is some influential work where this approach is used.
In Vempala et al. [84] the clustering quality measure that is proposed compares the minimal conductance within a cluster with the ratio of the weight of inter-cluster and intra-cluster edges. They present worst case guarantees for the clusterings found by a variant of spectral clustering, and show that if there is a good clustering with respect to the proposed measure, this algorithm will find a close approximation. Spielman and Teng [80] consider the ratio of vertices removed to edges cut as the clustering quality measure. They show that this measure is bounded for the clusterings found by some variant of spectral clustering on bounded-degree planar graphs and finite element meshes. In Bilu and Linial [11] the notion of stability of a clustering instance is introduced: an instance is stable if the optimal clustering does not change when the instance is slightly perturbed. It is shown that a spectral heuristic can approximate the maximum cut partition well if the instance is sufficiently stable. It is presumably possible that this kind of analysis could be carried over to other clustering quality measures that use some variant of the cut.
Of course, if the clustering algorithm relies on the approximate minimization of a clustering quality measure and the approximation factor is known, this gives a trivial bound on the worst-case difference between the optimal clustering and the clustering found by the approximation algorithm (in terms of the clustering quality measure). However, it is usually difficult to interpret what such an approximation factor means for the computed clustering. For spectral clustering it is known that it does not provide a constant-factor approximation to the original problem of optimizing Ncut or RatioCut. We refer the reader to the discussion in the review of spectral clustering above.

Graph-based clustering In machine learning, graph clustering algorithms are often applied to data which does not possess an inherent graph structure, but rather consists of points together with corresponding pairwise distance or similarity values. In this case, we perform two steps: First, the similarity information is used to construct a neighborhood graph on the data points, and then a graph clustering algorithm is applied to this neighborhood graph. We call this procedure *graph-based clustering*. In the following paragraph we introduce the clustering scenario of the clustering of points from a probability distribution in Euclidean space. In the rest of this thesis we will study graph-based clustering in this setting.

Clustering of points from a probability distribution in Euclidean space One of the most important clustering scenarios considered in machine learning is the clustering of points drawn from some probability distribution in a metric space. In this thesis we only consider the Euclidean space and probability distributions that have a density

with respect to the Lebesgue measure. In this scenario we assume that there is a "true clustering" of the underlying space that depends on the probability distribution. For example, if the support of the density consists of two balls with unit radius and positive distance from each other, we would intuitively claim that the "true clustering" would be to put the points of one ball into one cluster and the points of the other ball into the other cluster respectively. Having such a ground truth for clustering opens the door to the study of questions of consistency for clustering algorithms. However, it is not easy to define formally what true clusterings are.

In this thesis we consider two approaches of how to define a "true clustering" if the probability distribution is given by a density: In the first approach we consider the connected components of the set of all points where the density exceeds a certain threshold level as the "true clusters", and call these *high-density clusters* (also denoted in the literature as density-contour clusters or population clusters). In this approach only the level set parameter has to be chosen to define the clusters; however, it is not clear how to set this parameter. Note that in general in this clustering model not all the sample points belong to a high-density cluster. Rather there is a distinction between "background noise" and "interesting data", which offers the possibility to account for a foreground-background structure in the data. On the other hand, in many applications it is not desirable to have points which are not assigned to a cluster.

The second approach is to define a true clustering as a partition of the whole space, namely each point of the space belongs to exactly one part of the partition. Here, each sample point lies in exactly one true cluster. However, the partition of the space given the density is not obvious, although there are some intuitive notions of a good boundary between clusters, for example that it cuts through regions of low density rather than through regions of high density. This can be formalized by introducing a quality measure on a subset of the partitions of the space and define the true clustering as the partition out of this set that minimizes the quality measure.

Another approach to the clustering of points from a density, which will not be used in this thesis, is model-based clustering. In the easiest case, it is assumed that the true density can be represented as a mixture of Gaussians, where each Gaussian defines a cluster. Clustering then is the estimation of the mixture parameters. The number of clusters may be estimated by comparing different mixtures using a criterion such as the Bayes information criterion (BIC) for model selection. For an overview of model-based clustering and extensions to this framework, see, for example, Banfield and Raftery [7] and Fraley and Raftery [35].

Clustering consistency in this setting In this thesis we want to investigate how well graph-based clustering algorithms work when different types of neighborhood graphs or parameter choices are used. Having a notion of a "true clustering" makes it possible to evaluate the performance of clustering algorithms by studying, for example, questions of consistency.

Intuitively, consistency means that for more and more sample points the empirical clusters become infinitesimally close to the true clusters. However, the exact definition

of clustering consistency depends on the definition of true clustering we use and the meaning of the term consistency is surprisingly inconsistent in the literature. In the following we first study notions of consistency for the high-density cluster definition and then for the definition of clusters as partitions of the space.
Formally, a "true clustering" of d-dimensional Euclidean space $\mathbb{R}^d$ with probability measure μ defines the

- true clusters $C^{(1)}, \ldots, C^{(m)} \subseteq \mathbb{R}^d$ as subsets of $\mathbb{R}^d$ with $\mu(C^{(i)} \cap C^{(j)}) = 0$ for all $i, j = 1, \ldots, m$; $i \neq j$. Using the high-density cluster model the true clusters are the connected components of the level set of the density and the union of the true clusters in general does not cover the support of μ. In the definition of clustering as a partition of the space the true clustering is often given as the partition that minimizes a clustering quality measure on partitions of the space.

Applying a clustering algorithm on a finite set of sample points $x_1, \ldots, x_n$ sampled from the distribution μ we obtain either empirical or sample clusters:

- Empirical clusters $\hat{C}_n^{(1)}, \ldots, \hat{C}_n^{(m')} \subseteq \mathbb{R}^d$ are subsets of $\mathbb{R}^d$ whose pairwise intersection is a set of measure zero. If the true clustering is given as a partition of the space, the empirical clusters form a partition of $\mathbb{R}^d$ as well.

- Sample clusters $\tilde{C}_n^{(1)}, \ldots, \tilde{C}_n^{(m')}$ are pairwise disjoint subsets of the sample points. In the high-density cluster model not all sample points belong to a sample cluster, but there are also background points, whereas all sample points belong to exactly one sample cluster in the partition model.

If the algorithm computes empirical clusters we assume that the sample clusters are the intersection of the empirical clusters and the sample points: $\tilde{C}_n^{(i)} = \hat{C}_n^{(i)} \cap \{x_1, \ldots, x_n\}$ for $i = 1, \ldots, m'$. Conversely, it is not as simple to obtain reasonable empirical clusters from sample clusters (see discussion below).
In the high-density cluster model a definition of consistency must take into account that there could be sample points that do not belong to any of the true clusters. In the influential paper by Hartigan [42] on the consistency of single-linkage clustering, these points are ignored in the sense that it does not matter with respect to consistency whether a sample cluster contains such points and how many of them.
Hartigan [42] defines "full consistency" to be the property that with probability 1 as the sample size tends to infinity, for each "true" high-density cluster there is a sample cluster containing all of its sample points, and the sample clusters are mutually disjoint. The weaker property of "fractional consistency" is defined to be the property, that asymptotically, there will be two disjoint sample clusters that include a positive fraction of the sample points in the "true" high-density clusters. Wong and Lane [95] define "strong set consistency" for a hierarchical clustering: If, for each sample size, we choose the smallest sample clusters that contain all the sample points of the true high-density clusters, then almost surely for all but finitely many sample sizes the sample clusters are disjoint. Accordingly, we have "weak set consistency" if we replace "almost surely" by "with probability approaching one". Basically, set consistency is the

full consistency of Hartigan [42], where the choice of the hierarchy level is built in into the definition of consistency. Note that all the notions of consistency defined here only compare the true clusters and the sample clusters — there is no need to estimate the empirical clusters from the sample.
A special case of high-density clusters is the situation, where the support of the density consists of finitely many connected subsets having positive distance from each other and where we assume that the density is bounded away from zero on its support. Then the connected components of the support correspond to the high-density clusters if we set the density level to the infimum of the density on its support. In Chapter 2 we call this setting the "noise-free case" because the assumptions cannot hold if we disturb even a very well-behaved density (such as the uniform density on a ball) by Gaussian noise. If we can show "full consistency" in this setting for the appropriate density level parameter, namely the infimum of the density on its support, then the points of each high-density cluster form exactly one empirical cluster and there does not exist any sample point not belonging to a high-density cluster.
When we use the definition of the true clustering as a partition of the space the situation becomes even more diverse than in the high-density cluster model. In the following analysis if the number m' of empirical or sample clusters is smaller than the number m of true clusters, we set the empirical clusters $\hat{C}_n^{(i)} = \emptyset$ and the sample clusters $\tilde{C}_n^{(i)} = \emptyset$ for $i = m' + 1, \ldots, m$.
If the clustering algorithm under consideration estimates the empirical clusters we can use a distance measure between partitions of the space which can account for the probability distribution to compare clusterings. One possibility for such a distance measure would be the probability mass in the symmetric difference of the true and the empirical cluster, where the symmetric difference of two sets A and B is defined as $A \Delta B = (A \setminus B) \cup (B \setminus A)$. That is, a clustering algorithm is said to be consistent if we can order the empirical clusters for each sample size in such a way, that for all $i = 1, \ldots, m$ we have

$$\mu\left(C^{(i)} \Delta \hat{C}_n^{(i)}\right) \to 0 \tag{1.3}$$

for $n \to \infty$. Note here, that it is not necessary that $m = m'$, since the probability in all empirical clusters $\hat{C}_n^{(i)}$ that do not correspond to a true cluster clearly has to approach 0, since the probability mass in $\hat{C}_n^{(1)}, \ldots, \hat{C}_n^{(m)}$ has to approach one if the condition holds.
Many clustering algorithms do not compute the empirical clusters, but rather the sample clusters. In order to use the definition of consistency above, we have to extend the sample clustering to the whole space and thus — at least implicitly — compute empirical clusters. The extension of the clustering is basically a classification task and we can use a classifier such as a suitable k-nearest neighbor classifier.
Another way to define clustering consistency is to use only the sample clusters and the restriction of the true clusters to the sample points. A clustering algorithm is defined to be consistent if for all $i = 1, \ldots, m$ and $n \to \infty$

$$\frac{1}{n} \left| \left(C^{(i)} \cap \{x_1, \ldots, x_n\} \right) \Delta \tilde{C}_n^{(i)} \right| \to 0. \tag{1.4}$$

This definition of consistency seems to be in accordance with the definitions of consistency in the high-density cluster case, where the consistency of an algorithm depended only on the sample clusters. However, it would be interesting to investigate the relationship between this definition and the definition above: Suppose, for example, that the empirical clusters are computed by extending the sample clusters to the whole space using a consistent classifier. Under which conditions does consistency in the sense of Equation (1.4) imply consistency in the sense of Equation (1.3) then?
For clustering algorithms based on the optimization of a clustering quality measures that is defined on partitions of the space it is possible to define consistency through the convergence of the clustering quality measure as in von Luxburg et al. [88] and Bubeck and von Luxburg [16]. If Q denotes the quality measure, that is $Q(\{C^{(1)}, \ldots, C^{(m)}\})$ denotes the value of the quality measure for the partition of the space into $C^{(1)}, \ldots, C^{(m)}$, then a clustering algorithm that computes empirical clusters is consistent, if

$$Q\left(\{\hat{C}_n^{(1)}, \ldots, \hat{C}_n^{(m')}\}\right) \;\to\; Q\left(\{C^{(1)}, \ldots, C^{(m)}\}\right)$$

for $n \to \infty$, where $C^{(1)}, \ldots, C^{(m)}$ are the clusters of the optimal partition according to Q. A similar idea is used in Pollard [72] to show the consistency of k-means clustering. The k-means algorithm is special in the sense, that it automatically computes both, sample clusters and empirical clusters. Here, the possible partitions of the space are restricted to the set of all Voronoi partitions of at most k points. A quality measure that takes into account the density is defined on the partitions (that is, on the k center points). For each finite sample size an estimator of the quality measure is optimized over all partitions. It is shown that the Voronoi centers of the optimal partition of the estimator converge almost surely to the centers of the optimal Voronoi partition of the density, and that the minimum value of the estimator of the quality measure converges almost surely to the actual minimal value of the quality measure on the density.
As demonstrated, it is interesting to study the relationship between clustering quality measures defined on the density and the actual clusterings computed by a clustering algorithm or (ideally) by the minimization of a clustering quality measure on a finite set of sample points. In Chapter 3 of this thesis we fix a partition of the space and build certain types of neighborhood graphs on our sample points. We investigate the limits of Ncut and RatioCut for the cuts in the neighborhood graphs that are induced by the fixed partition of the space. The neighborhood graphs we use are presented in the following section.

1.2 Neighborhood Graphs

Using graphs to model real world problems is one of the most widely used techniques in computer science. This approach usually involves two major steps: building an appropriate graph which represents the problem in a convenient way, and then constructing an algorithm which provides a solution to the problem on the given type of graph. While in some cases there exists an obvious natural graph structure to model

the problem, in other cases one has much more choice when constructing the graph. In this thesis we consider neighborhood graphs, which are a popular choice of graph type in many applications of computer science.

Neighborhood graphs are random geometric graphs that use concepts of "nearness" or neighborhood of points to define a graph. Random geometric graphs are graphs whose vertices are points that are randomly distributed in some metric space. The existence and the weight of an edge between two points depends only on the position of all the points — in other words, it is not random given the points. This distinguishes random geometric graphs from classical random graphs such as Erdős-Rényi random graphs, wherein the edges are inserted randomly (see Bollobas [12] for an overview).

Neighborhood graphs include k-nearest neighbor graphs, r-neighborhood graphs, relative neighborhood graphs, Gabriel graphs, sphere-of-influence graphs and sphere-of-attraction graphs. An overview of these types of neighborhood graphs and their use in statistical pattern recognition is presented in Marchette [59]. In this thesis we only treat the two most common types of neighborhood graphs, the r-neighborhood graph and the k-nearest neighbor graph (often abbreviated to kNN-graph). In the r-neighborhood graph we choose a radius r and connect two points if their distance is less than or equal to r. Since the metric is symmetric there is a canonical way to convert a directed r-neighborhood graph into an undirected one and vice versa. The idea behind the k-nearest neighbor graphs is to connect each point to the k points closest to it. This yields a directed graph and, unlike the metric, the k-nearest neighbor relation is not symmetric. In order to construct an undirected k-nearest neighbor graph we have to decide whether to insert an edge between two points where one point is among the k-nearest neighbors of the other point but not vice versa. In the mutual k-nearest neighbor graph this edge is not inserted — there is only an edge between two points if both points are among the k nearest neighbors of the other point. In the symmetric k-nearest neighbor graph this edge is inserted, there is already an edge between two points if one of them is among the k nearest neighbors of the other one. Formal definitions of these two types of neighborhood graphs can be found in Section 1.4.

Neighborhood graphs are used in a wide range of scientific disciplines: from modeling sensor networks and ad-hoc networks in computer science, to modeling the spread of diseases in medicine and the connections in the brain in neurology. Due to their ubiquitous use in many diverse application areas, neighborhood graphs have recently received a lot of attention in mathematics. It is reasonable to assume that the properties of neighborhood graphs will depend on the distribution of the points on which they are constructed. Thus, in order to study the properties of neighborhood graphs we make some assumptions on the distribution of the points — for example, that they are drawn independently from a probability density.

As previously mentioned, there exists a lot of results in the literature on the properties of different types of neighborhood graphs in various settings. However, there are two drawbacks to most of the published results: First, they usually treat the case of samples from a uniform density on the cube or torus or from a homogeneous Poisson process, and it is not clear how to generalize these results for more general distributions. In clustering, however, we consider the opposite of this scenario: a uniform density does not

contain any clusters at all. Second, the results are on asymptotic convergence properties when the number of points tends to infinity. The question of whether these results also provide clues on the properties of random graphs for finite samples is hardly discussed. However, the ideal results in the context of clustering would be *non-asymptotic* results on the properties of different types of neighborhood graphs on a *finite point set* which has been drawn from a *highly clustered distribution*.
A monograph on many properties of neighborhood graphs is Penrose [67]. The properties that are particularly interesting for clustering are the connectivity of the graph, the size of its connected components, the relationship to the minimum spanning tree (of neighborhood graphs and of the complete graph where an edge between two points is weighted by their distance), and the longest edge of the kNN-graph.
Connectivity, especially of the r-neighborhood graphs, has been studied recently in the context of ad-hoc and sensor networks (see Estrin et al. [31] and Pottie and Kaiser [73]). Most of the results in this area are concerned with random graphs in the two-dimensional plane, since this setting is motivated by the application. Avin and Ercal [5] study the limits of certain properties of random walks on r-neighborhood graphs in the plane. Santi and Blough [76], Bettstetter [9] and Kunniyur and Venkatesh [53] study the connectivity of ad-hoc and sensor networks that are modeled by r-neighborhood graphs. Connectivity results for neighborhood graphs are also closely related to the study of percolation, see for example, Stauffer and Aharony [82], Grimmett [38] and Bollobas and Riordan [13]. In Brito et al. [15] the authors study the connectivity of the random mutual k-nearest neighbor graphs and suggest a test for the presence of outliers in the data based on these connectivity properties. The size of connected components in the 1-nearest neighbor graph on points from a homogeneous Poisson process in $\mathbb{R}^d$ is studied in Kozakova et al. [52]. A central limit theorem for the total edge length and the number of components of the k-nearest neighbor graph on points from a homogeneous Poisson process or from a uniform distribution is shown in Penrose and Yukich [70]. In Penrose and Yukich [71] a weak law of large numbers is shown for the total edge length and is also shown for the number of connected components of the k-nearest neighbor graph on points drawn independently from a density in d-dimensional space. The connection between the minimal spanning tree and the k-nearest neighbor graph is studied in Gonzàlez-Barrios and Quiroz [37]. They show that $k \sim \log n$ is the smallest k such that the k-nearest neighbor graph contains the minimal spanning tree of the complete graph if the weight of the edge between two points is their Euclidean distance. The limit distribution of the length of the longest edge of the 1-nearest neighbor graph on points uniformly distributed in the unit d-cube is shown in Dette and Henze [26]. For points generated by a Poisson process in the unit square a central limit theorem for the total edge length of the k-nearest neighbor graph is shown in Avram and Bertsimas [6]. Penrose [68] shows the weak convergence of the length of the longest edge of the minimal spanning tree and for the 1-nearest neighbor graph for uniform distributions in the unit square. A strong law for the length of the longest edge of the minimal spanning tree for points sampled independently from a density in $\mathbb{R}^d$ is shown in Penrose [69]. However, Caroni and Prescott [18] claim that this result is inapplicable on small data sets that are not uniformly distributed.

1.3 Overview of the results

We wish to study the influence of the construction of the neighborhood graph on the graph-based clustering of points sampled from a density in Euclidean space. Ideally, we would like to be able to decide which neighborhood graph and which neighborhood parameter to choose in order to achieve good results in graph-based clustering. As a step towards this goal we study the interrelation between "true clusters" of densities in Euclidean space, namely either high-density clusters or partitions of the density, on the one hand, and "sample clusters" (found by graph-based clustering methods) and graph clustering quality measures on the other hand.

Chapter 2 In Chapter 2 we use the high-density cluster definition and study the problem of cluster identification: Clusters are said to be roughly identified in a graph, if the connected components of the graph correspond to the true underlying clusters. The connected components may still contain points which do not belong to any high-density cluster. We say that a cluster is exactly identified if the proportion of these points to the cluster points asymptotically approaches zero. The graphs we consider are modifications of the mutual and the symmetric k-nearest neighbor graph: the points at which the estimated density is below a certain threshold as well as too small connected components are removed. We prove bounds on the probability that clusters are identified successfully in these graphs, both in the special case of the noise-free setting, where the density is bounded away from zero on its support, and in the general noisy setting, where the density does not have to be bounded from below. Using these bounds we compare the (modified) mutual and symmetric k-nearest neighbor graphs with respect to cluster identification and determine the optimal choice of the parameter k to maximize the probability of rough cluster identification. It turns out that the optimal choice is surprisingly high, rather of the order n than of the order $\log n$. Our second conclusion is that the major difference between the mutual and the symmetric k-nearest neighbor graph occurs when one attempts to detect the most significant cluster only, that is the cluster that is most easily identified. Furthermore, we show that for growth rates of the parameter k between logarithmic and linear (with suitable constants) the clusters are exactly identified asymptotically almost surely in the constructed graphs. Chapter 2 is based on Maier et al. [56] and Maier et al. [57].

Chapter 3 In Chapter 3 we consider as "true clusters" partitions of $\mathbb{R}^d$ by hyperplanes and clusterings of finite sets of sample points induced by these partitions. On a finite sample we construct the directed r-neighborhood and k-nearest neighbor graphs and investigate the limit for the graph clustering quality measures Ncut and RatioCut for the given clusterings as the sample size tends to infinity. We find that the limit expressions are different for the two different types of neighborhood graph under consideration. Furthermore, we can find simple examples of densities, for which the optimal partition out of a set of reasonable partitions is substantially different. In other words: Ncut on a kNN graph does something systematically different than Ncut on an

r-neighborhood graph. This finding shows that graph clustering quality measures cannot be studied independently of the type of neighborhood graph when they are applied in the graph-based clustering of points from a density in Euclidean space. We also provide examples which demonstrate that these differences can be observed when spectral clustering is applied to toy and real data sets of rather small sample sizes. Chapter 3 is based on Maier et al. [58].

1.4 General definitions and notations

In this section we introduce definitions and notations that will be used throughout the rest of this thesis. For an overview of the notation see also the table of notations on page 127 and the following pages.

Setting: space and sample We always work on the space $\mathbb{R}^d$ endowed with the Euclidean metric dist. The distance is extended to sets $A, B \subseteq \mathbb{R}^d$ via $\text{dist}(A, B) = \inf\{\text{dist}(x, y) \mid x \in A, y \in B\}$, and similarly $\text{dist}(x, A) = \inf\{\text{dist}(x, y) \mid y \in A\}$ for $x \in \mathbb{R}^d$. The Euclidean norm of $x \in \mathbb{R}^d$ is denoted by $\|x\|$, the standard dot product between $x, y \in \mathbb{R}^d$ by $\langle x, y \rangle$.
The data points $x_1, ..., x_n$ are sampled independently from some probability measure μ which has a density p with respect to the Lebesgue measure in $\mathbb{R}^d$. That is, for a measurable set $A \subseteq \mathbb{R}^d$ we set $\mu(A) = \int_A p(x)\, \mathrm{d}x$.

Neighborhood Graphs For a point x_i $(i = 1, \ldots, n)$ let π_i be a bijective mapping of $\{1, \ldots, n-1\}$ to the indices $1, ..., i-1, i+1, ..., n$ such that $\text{dist}(x_i, x_{\pi_i(1)}) \leq \ldots \leq \text{dist}(x_i, x_{\pi_i(n-1)})$. Then for $k \in \{1, \ldots, n-1\}$ the k-nearest neighbor radius (kNN radius) $R^k(x_i)$ of a point x_i is defined as $R^k(x_i) = \text{dist}(x_i, x_{\pi_i(k)})$. For $i = 1, \ldots, n$ set $\text{kNN}(x_i) = \{x_{\pi_i(1)}, \ldots, x_{\pi_i(k)}\}$, that is the k nearest neighbors of point x_i. Note that the mapping π_i and therefore the set $\text{kNN}(x_i)$ is not unique, whereas the kNN radius $R^k(x_i)$ is unique. However, in the rest of this thesis we always construct neighborhood graphs on points randomly sampled from a density in $\mathbb{R}^d$ endowed with the Euclidean metric. In this case there almost surely exists a unique mapping π_i for all $i \in \{1, \ldots, n\}$. Clearly, the k-nearest neighbor relationship is not symmetric, that is $x_j \in \text{kNN}(x_i)$ does not necessarily imply $x_i \in \text{kNN}(x_j)$. That is why we first define the directed neighborhood graphs. All of them have vertex set $x_1, \ldots, x_n$ and we denote the edge set by E. The edge set of the r-neighborhood graph $G_{\text{r}}(n, r)$ is $E = \{(x_i, x_j) \mid i, j = 1, \ldots, n; i \neq j; \text{dist}(x_i, x_j) \leq r\}$, whereas the edge set of the directed k-nearest neighbor graph $G_{\text{kNN}}(n, k)$ is $E = \{(x_i, x_j) \mid i, j = 1, \ldots, n; i \neq j; x_j \in \text{kNN}(x_i)\}$. That is, in the

- *r-neighborhood graph* $G_r(n, r)$ there is an edge from x_i to x_j if $\text{dist}(x_i, x_j) \leq r$, whereas in the
- *k-nearest-neighbor graph* $G_{kNN}(n, k)$ there is an edge from x_i to x_j if $x_j \in \text{kNN}(x_i)$.

Since many algorithms rely on undirected graphs we have to construct undirected graphs from these neighborhood graphs. For the r-neighborhood graph there is a canonical way to do this, since, due to the symmetry of the distance function, if there is an edge from x_i to x_j then there is also an edge from x_j to x_i. Therefore we can just set $E = \{\{x_i, x_j\} \mid i, j = 1, \dots, n; i \neq j; \text{dist}(x_i, x_j) \leq r\}$ for the undirected r-neighborhood graph $G_{\text{r}}^u(n, r)$. As mentioned above the k-nearest neighbor relationship is not symmetric and thus there is not one natural way to construct an undirected graph. However, there are two variants that are widely used: the symmetric k-nearest-neighbor graph $G_{\text{sym}}(n, k)$, in which $E = \{\{x_i, x_j\} \mid i, j = 1, \dots, n; i \neq j; x_i \in \text{kNN}(x_j) \text{ or } x_j \in \text{kNN}(x_i)\}$ and the mutual k-nearest-neighbor graph $G_{\text{mut}}(n, k)$, in which $E = \{\{x_i, x_j\} \mid i, j = 1, \dots, n; i \neq j; x_i \in \text{kNN}(x_j) \text{ and } x_j \in \text{kNN}(x_i)\}$. To summarize, for $i, j = 1, \dots, n; i \neq j$ in the

- *undirected r-neighborhood graph* $G_r^u(n, r)$*:* the points x_i and x_j are connected if $\text{dist}(x_i, x_j) \leq r$,
- *symmetric k-nearest-neighbor graph* $G_{sym}(n, k)$*:* the points x_i and x_j are connected if $x_i \in \text{kNN}(x_j)$ or $x_j \in \text{kNN}(x_i)$,
- *mutual k-nearest-neighbor graph* $G_{mut}(n, k)$*:* the points x_i and x_j are connected if $x_i \in \text{kNN}(x_j)$ and $x_j \in \text{kNN}(x_i)$.

Note that the literature does not agree on the names for the different kNN graphs. In particular, the graph we call "symmetric" usually does not have a special name.

Minimal curvature radius For a smooth hypersurface M in Euclidean space we use the *minimal curvature radius* κ in order to state assumptions on the maximal curvature: At any point $p \in M$ the shape operator s_p is a self-adjoint linear transformation on the tangent space T_pM to M at p. The eigenvalues $r_1^{(p)}, \dots, r_{d-1}^{(p)}$ of s_p are called the principal curvatures of M at p. Define $\kappa^{(p)} = 1/(\max\{|r_i^{(p)}| \mid i = 1, \dots, d-1\})$ if at least one of the principal curvatures is different from 0 and $\kappa^{(p)} = \infty$ otherwise. Then the minimal curvature radius κ is defined as $\kappa = \inf_{p \in M} \kappa^{(p)}$.
The geometric meaning of the minimal curvature radius is the following: For any point p and any unit tangent vector v at p let E_v denote the plane through p spanned by v and the normal n_p in p. The intersection of M and E_v is locally a curve. The Euclidean curvature of this curve in the point p is the dot product of the tangent vector v and the shape operator s_p applied to v. Then $1/\kappa^{(p)}$ (where we set $1/\infty = 0$) is an upper bound on the curvature of such a curve in p. Taking the infimum we obtain a bound over all points p in M.

Further notation For a measurable set $A \in \mathbb{R}^d$ we denote the Lebesgue measure by $\mathcal{L}_d(A)$. Similarly, for a subset B of a $(d-1)$-dimensional affine subspace $\mathcal{L}_{d-1}(B)$ denotes the $(d-1)$-dimensional Lebesgue measure in that subspace. Furthermore, we set $\mathcal{L}_0(A) = |A|$. When applied to a finite union of smooth $(d-1)$-dimensional surfaces

$S \subseteq \mathbb{R}^d$ without boundary or with smooth boundaries, the symbol $\mathcal{L}_{d-1}(S)$ denotes the $(d-1)$-dimensional area and similarly for lower-dimensional entities. Technically, this is an overloading of notation. However, in the context where both definitions are applicable they coincide. Furthermore, it reflects the intuitive notion of the "content" of a set irrespective of the strict mathematical details.
For $x \in \mathbb{R}^d$ and $r \in \mathbb{R}_{\geq 0}$ we set $B(x,r)$ to be the closed ball of Euclidean radius r around x, that is $B(x,r) = \{y \in \mathbb{R}^d \mid \operatorname{dist}(x,y) \leq r\}$. The volume of the Euclidean unit ball in $\mathbb{R}^d$ is denoted by η_d, that is $\eta_d = \mathcal{L}_d(B(0,1))$. We set $\eta_0 = 1$.
$\operatorname{Bin}(n,p)$ denotes the discrete probability density of the binomial distribution with parameters n and p.
In order to state an asymptotic upper bound on the growth and decay rates for sequences $(f_n)_{n\in\mathbb{N}}$, $(g_n)_{n\in\mathbb{N}}$ in $\mathbb{R}_{\geq 0}$ we use the Landau O notation: $f_n = O(g_n)$ if there exist positive constants c and n_0 such that $0 \leq f_n \leq cg_n$ for all $n \geq n_0$.

2 Cluster Identification

2.1 Introduction

In graph-based clustering we use graph clustering algorithms on neighborhood graphs in order to cluster points from a density in Euclidean space. In general, we are interested in the question of how the choices of neighborhood graph and its parameter influence the clustering result. This is related to the question of which graph to choose and how to set the parameter in order to obtain the best clustering results. In this chapter we wish to answer these questions for a very simple graph clustering algorithm.
As described in Section 1.1 graph clustering algorithms are often based on graph clustering quality measures that reflect the trade-off between intra-cluster density and inter-cluster sparsity. A minimal requirement for intra-cluster density would be the connectedness of each cluster in the graph. An extreme case of inter-cluster sparsity would be to require different clusters to be disconnected in the graph. Combining these requirements we come to the simple definition of graph clusters as the connected components of the graph. That is, a simple graph clustering algorithm would identify the connected components of a graph as the graph clusters.
Suppose that we apply this simple algorithm for the graph-based clustering of points sampled from a density in Euclidean space: A neighborhood graph is constructed on the points and the connected components of this graph are taken to be the sample clusters. In this setting we can define a true clustering of the density via the high-density cluster model. Having defined the true clusters we can study "cluster identification": Do the sample clusters found by the graph-based clustering algorithm, that is, the connected components of the neighborhood graph, correspond to the true high-density clusters? Specifically, in this chapter we study the question of how to construct an undirected k-nearest neighbor graph in order to maximize the probability of cluster identification: Is the mutual or the symmetric k-nearest neighbor graph better suited for the task? How should we choose its connectivity parameter k, which determines the size of the neighborhood?
Our results on the choice of the graph type and the parameter k for cluster identification can be summarized as follows. Concerning the question of the choice of k, we obtain the result that k should be chosen surprisingly high, namely of the order of $O(n)$ instead of $O(\log n)$ (the latter would be the rate one would "guess" from results in standard random geometric graphs). Concerning the types of graphs, it turns out that different graphs have their advantages in different situations: if one is only interested in identifying the "most significant" cluster (while some clusters might still not be correctly identified), then the mutual kNN graph should be chosen. If one wants to identify many clusters simultaneously, the bounds show no substantial difference between the

mutual and the symmetric kNN graph.
This chapter is based on Maier et al. [56], which received the E. M. Gold Award at the Conference on Algorithmic Learning Theory in 2007, and Maier et al. [57].

2.2 Main constructions and results

In this section we provide a brief overview of the setup and techniques we use in the following. Mathematically exact statements will follow in the next sections.

Sample and Neighborhood graphs. We are in the setting described in Section 1.4. The neighborhood graphs we consider in this chapter are the undirected k-nearest neighbor graphs $G_{\text{sym}}(n,k)$, and $G_{\text{mut}}(n,k)$ defined there. Most of the questions we study are much easier to solve for the undirected r-neighborhood graph $G_{\text{r}}^{u}(n,r)$, than for kNN graphs. The reason is that whether two points x_i and x_j are connected in the r-graph only depends on $\text{dist}(x_i,x_j)$, while in the kNN graph the existence of an edge between x_i and x_j also depends on the distances of x_i and x_j to all other data points. However, the kNN graph is the one which is mostly used in practice. Hence we focus on kNN graphs. Most of the proofs can easily be adapted for the r-neighborhood graph.

The cluster model. We use the high-density cluster model, that is, we define clusters as the level sets of the density. Given the underlying density p of the data space and a parameter $t > 0$, we define the t-level set $L(t)$ as the closure of the set of all points $x \in \mathbb{R}^d$ with $p(x) \geq t$. Clusters are then defined as the connected components of the t-level set (where the term "connected component" is used in its topological sense and not in its graph-theoretic sense).

The cluster identification problem. Given a finite sample from the underlying distribution, our goal is to identify the sets of points which originate from different connected components of the t-level set, that is from different high-density clusters. In the rest of this chapter we will write shortly "cluster" for the true high-density clusters. We study this problem in two settings:
The noise-free case. Here we assume that the support of the density consists of several connected components which have a positive distance from each other. Between these components, there is only "empty space" (density zero). Each of the connected components is called a cluster. Given a finite sample $x_1,...,x_n$ from such a density, we construct a neighborhood graph G based on this sample. We say that **a cluster is identified in the graph** if there is exactly one corresponding connected component in the neighborhood graph. That is, all of the points originating in the same underlying cluster are connected in the graph, and they are not connected to points from any other cluster. Examples illustrating the influence of the graph parameter k of the symmetric kNN graph on cluster identification in the noise-free case in a simple example can be found in Figure 2.1. We call this setting the noise-free case, because the assumption that

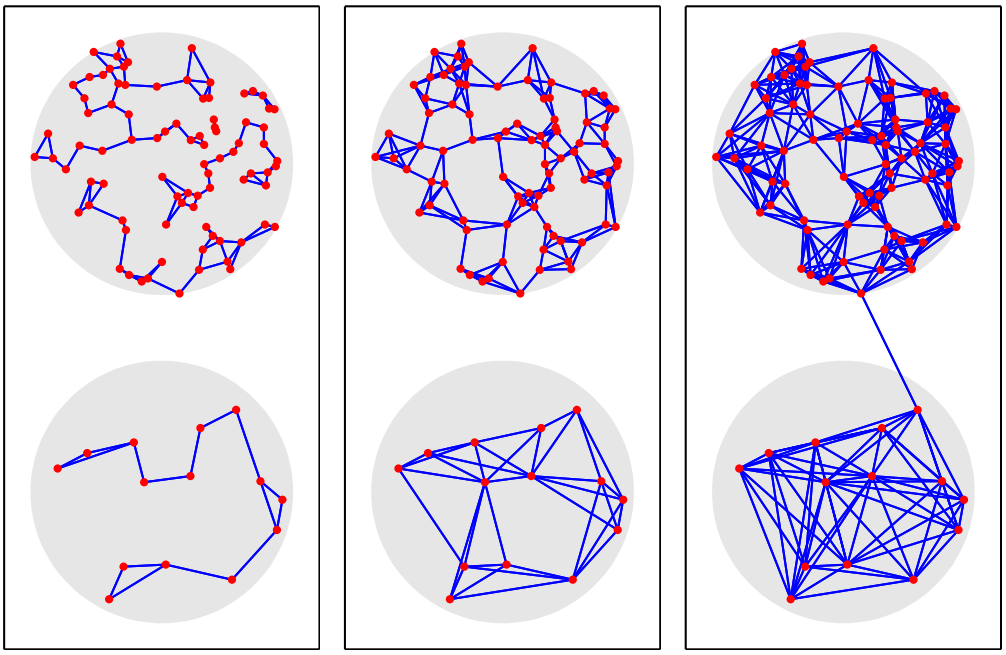

Figure 2.1: The influence of the graph parameter k on cluster identification in the noise-free case: We plot the symmetric k-nearest neighbor graph for $k = 2$ (left), $k = 4$ (middle) and $k = 8$ (right) constructed on a sample of 100 points from two clusters with uniform density (the grey discs). The subgraph of the upper cluster is disconnected for $k = 2$. For $k = 4$ both clusters are identified, whereas for $k = 8$ there are connections between the clusters.

the density is 0 between the clusters cannot hold any more if we add Gaussian noise to all the points, even if the original density was very well-behaved.

The noisy case. Here we no longer assume that the clusters are separated by "empty space", but allow the underlying density to be supported everywhere. Clusters are defined as the connected components of the t-level set $L(t)$ of the density (for a fixed parameter t chosen by the user), and points not contained in this level set are considered as background noise. A point $x \in \mathbb{R}^d$ is called a *cluster point* if $x \in L(t)$ and *background point* otherwise. As in the previous case we will construct a neighborhood graph G on the given sample. However, we will remove points from this graph which we consider as noise. The remaining graph $\tilde{G}$ will be a subgraph of the graph G, containing fewer vertices and fewer edges. As opposed to the noise-free case, we now define two slightly different cluster identification problems. They differ in the way background points are treated. The reason for this more involved construction is that in the noisy case, one cannot guarantee that no additional background points from the neighborhood of the cluster will belong to the graph.

We say that **a cluster is roughly identified** in the remaining graph $\tilde{G}$ if the following

properties hold:

- all sample points from a cluster are contained as vertices in the graph, that is, only background points are dropped,
- the vertices belonging to the same cluster are connected in the graph; that is, there exists a path between every pair of two distinct vertices, and
- every connected component of the graph contains only points of exactly one cluster (and maybe some additional noise points, but no points of a different cluster).

We say that **a cluster is exactly identified in** $\tilde{G}$ if

- it is roughly identified for all but finitely many n almost surely, and
- the ratio of the number of background points and the number of cluster points in the graph $\tilde{G}$ converges almost surely to zero as the sample size approaches infinity.

If all the clusters have been roughly identified, the number of connected components of the graph $\tilde{G}$ is equal to the number of connected components of the level set $L(t)$. However, $\tilde{G}$ might still contain a significant number of background points. In this sense exact cluster identification is a much stronger problem, as we require the fraction of background points in the graph to approach zero. Exact cluster identification is an asymptotic statement, whereas rough cluster identification can be verified on each finite sample. Due to the statement about the background points exact cluster identification is stronger than full consistency in Hartigan [42] and strong set consistency in Wong and Lane [95], because there the background points are ignored. Finally, note that in the noise-free case rough and exact cluster identification coincide.

The clustering algorithms. To determine the clusters in the finite sample, we proceed as follows. Initially, we construct a neighborhood graph on the sample. This graph looks different, depending on whether we allow noise or not:
Noise-free case. Given the data, we simply construct the mutual or the symmetric k-nearest neighbor graph ($G_{\text{mut}}(n,k)$ respectively $G_{\text{sym}}(n,k)$) on the data points, for a certain parameter k, based on the Euclidean distance. Clusters are then the connected components of this graph.
Noisy case. Here we use a more complex procedure:

- As in the noise-free case, construct the mutual (symmetric) kNN graph $G_{\text{mut}}(n,k)$ (resp. $G_{\text{sym}}(n,k)$) on the samples.
- Estimate the density $\hat{p}_n(x_i)$ at every sample point x_i (for example, by kernel density estimation).
- If $\hat{p}_n(x_i) < t'$, remove the point x_i and its adjacent edges from the graph (where t' is a parameter determined later). The resulting graph is denoted by $G'_{\text{mut}}(n,k,t')$ (resp. $G'_{\text{sym}}(n,k,t')$).

- Determine the connected components of $G'_{\text{mut}}(n,k,t')$ (resp. $G'_{\text{sym}}(n,k,t')$), for example by a simple depth-first search.
- Remove the connected components of the graph that are "too small", that is, which contain less than δn points (where δ is a small parameter determined later).
- The resulting graph is denoted by $\tilde{G}_{\text{mut}}(n,k,t',\delta)$ (resp. $\tilde{G}_{\text{sym}}(n,k,t',\delta)$); its connected components are the clusters of the sample.

Note that by removing the small components in the graph the method becomes very robust against outliers and "fake" clusters (small connected components just arising by random fluctuations). In the rest of this chapter the modified graphs $\tilde{G}_{\text{mut}}$ and $\tilde{G}_{\text{sym}}$ will be called mutual and symmetric k-nearest neighbor graph, since the modifications are rather minor with respect to the points in reasonably large high-density regions.

Main results, intuitively. We would like to outline our results briefly in an intuitive fashion. Exact statements can be found in the following sections.

Result 1 (Range of k for successful cluster identification) *Under mild assumptions, and for n large enough, there exists constants $c_1, c_2 > 0$ such that for any $k \in [c_1 \log n, c_2 n]$, all clusters are identified with high probability in both the mutual and symmetric* kNN *graph. This result holds for cluster identification in the noise-free case as well as for the rough and the exact cluster identification problem (the latter seen as an asymptotic statement) in the noisy case (with different constants c_1, c_2).*

For the noise-free case, the lower bound on k has already been proven in Brito et al. [15], for the noisy case it is new. Importantly, in the exact statement of the result all constants have been worked out more carefully than in Brito et al. [15], which is very important for proving the following statements.

Result 2 (Optimal k for cluster identification) *Under mild assumptions, and for n large enough, the parameter k which maximizes the probability of successful identification of one cluster in the noise-free case has the form $k = c_1 n + c_2$, where c_1, c_2 are constants which depend on the geometry of the cluster. This result holds for both the mutual and the symmetric* kNN *graph, but the convergence rates are different (see Result 3). A similar result holds as well for rough cluster identification in the noisy case, with different constants.*

This result is completely new, both in the noise-free and in the noisy case. In the light of the existing literature, it is rather surprising. So far it has been well known that in many different settings the lower bound for obtaining connected components in a random kNN graph is of the order $k \sim \log n$. However, we now can see that *maximizing the probability* of obtaining connected components on a finite sample leads to a dramatic change: k has to be chosen much higher than $\log n$, namely of the order n itself. Moreover, we were surprised ourselves that this result does not only hold in the noise-free case, but can also be carried over to rough cluster identification in the noisy setting.

For exact cluster identification we did not manage to determine an optimal choice of k due to the very difficult setting. For large values of k, small components which can be discarded will no longer exist. This implies that a lot of background points are attached to the real clusters. On the other hand, for small values of k there will exist several small components around the cluster which are discarded, so that there are less background points attached to the final cluster. However, this trade-off is very hard to grasp in technical terms. We therefore leave the determination of an optimal value of k for exact cluster identification as an open problem. Moreover, as exact cluster identification concerns the asymptotic case of $n \to \infty$ only, and rough cluster identification is all one can achieve on a finite sample anyway, it is more than acceptable to be able to prove the optimal rate in this case.

Result 3 (Identification of the most significant cluster) *For the optimal k as stated in Result 2, the convergence rate (with respect to n) for the identification of one fixed cluster $C^{(i)}$ is different for the mutual and the symmetric* kNN *graph. It depends*

- *only on the properties of the cluster $C^{(i)}$ itself in the mutual* kNN *graph*
- *on the properties of the "least significant" (the "worst" out of all) clusters in the symmetric* kNN *graph.*

This result shows that if one is interested in identifying the "most significant" clusters only, it is better to use the mutual kNN graph. When the goal is to identify all clusters, then there is not much difference between the two graphs, because both of them have to deal with the "worst" cluster anyway. Note that this result is mainly due to the different between-cluster connectivity properties of the graphs, the within-cluster connectivity results are not so different (using our proof techniques at least).

Proof techniques, intuitively. Given a neighborhood graph on the sample, cluster identification always consists of two main steps: ensuring that points of the same cluster are connected and that the points of different clusters are not connected to each other. We call these two events "within-cluster connectedness" and "between-cluster disconnectedness" (or "cluster isolation").
In order to treat within-cluster connectedness we work with a covering of the true cluster. We cover the whole cluster by balls of a certain radius $z/4$. Then we wish to ensure that, first, each of the balls contains at least one of the sample points, and second, that points in neighboring balls are always connected in the kNN graph. These are two contradicting goals. The larger the $z/4$ is, the easier it is to ensure that each ball contains a sample point. The smaller the radius $z/4$ is, the easier it is to ensure that points in neighboring balls will be connected in the graph for a fixed number of neighbors k. So the first part of the proof consists in computing the probability that for a given radius $z/4$ both events occur at the same time and finding the optimal radius $z/4$. An example of a $z/4$-covering can be seen in Figure 2.2.
Between-cluster connectivity is easier to treat. Given a lower bound on the distance u between two clusters, all all that is required is to make sure that edges in the kNN

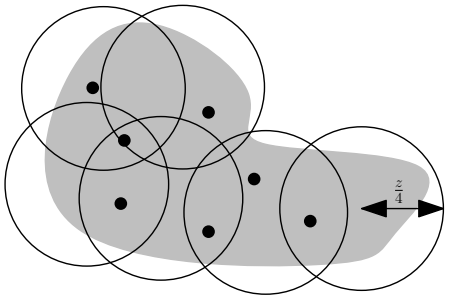

Figure 2.2: A covering of a cluster with balls of radius $z/4$, where each one contains at least one sample point. If the minimal k-nearest-neighbor radius of the sample points is at least z, then points in neighboring balls are connected in the kNN graph.

graph never become longer than u, that is we have to prove bounds on the maximal kNN distance in the sample.
In general, those techniques can be applied with small modifications both in the noise-free and in the noisy case, provided we construct our graphs in the way described above. The complication in the noisy case is that if we just used the standard kNN graph as in the noise-free case, then naturally the whole space would be considered as one connected component, and this would also show up in the neighborhood graphs. Thus, one must artificially reduce the neighborhood graph in order to remove the background component. Only then can one hope to obtain a graph with different connected components corresponding to the different clusters. The way we construct the graph $\tilde{G}$ ensures this. First, under the assumption that the error of the density estimator is bounded by ε, we consider the $(t-\varepsilon)$-level set instead of the t-level set we are interested in. This ensures that we do not remove "true cluster points" in our procedure. A second challenging complication in the noisy case is that with a naive approach, the radius z of the covering and the accuracy ε of the density estimator would be coupled to each other. We would need to ensure that the parameter ε decreases with a certain rate depending on z. This would lead to complications in the proof as well as very slow convergence rates. The trick by which we can avoid this is to introduce the parameter δ and throw away all connected components which are smaller than δn. Thus, we ensure that no small connected components are left over in the boundary of the $(t-\varepsilon)$-level set of a cluster, and all remaining points which are in this boundary strip will be connected to the main cluster represented by the t-level set. Note, that this construction allows us to estimate the number of clusters even without exact estimation of the density.

Building blocks from the literature. To a certain extent, our proofs follow and combine some of the techniques presented in Brito et al. [15], Cuevas et al. [23] and Biau et al. [10].
In Brito et al. [15] the authors study the connectivity of random mutual k-nearest neigh-

bor graphs. They are, however, mainly interested in asymptotic results; they only consider the noise-free case and do not attempt to make statements about the optimal choice of k. Their main result is that in the noise-free case, choosing k at least of the order $O(\log n)$ ensures that in the limit where n tends to infinity connected components of the mutual k-nearest neighbor graph correspond to true underlying clusters.
In Cuevas et al. [23] and Biau et al. [10], the authors study the noisy case and define clusters as connected components of the t-level set of the density. As in our case, the authors use density estimation to remove background points from the sample, but then work with an r-neighborhood graph instead of a k-nearest neighbor graph on the remaining sample. Connectivity of this kind of graph is much easier to treat than the one of k-nearest neighbor graphs, as the connectivity of two points in the r-neighborhood graph does not depend on any other points in the sample (this is not the case in the k-nearest neighbor graph). Cuevas et al. [23] prove results on the estimation of the number of clusters, whereas Biau et al. [10] prove asymptotic results for the estimation of the connected components of the level set $L(t)$. Both papers furthermore do not investigate the optimal choice of their graph parameter r. Moreover, due to our additional step where we remove small components of the graph, we have a much weaker coupling of the density estimator and the clustering algorithm. Therefore, we can provide much faster rates for the estimation of the components.

2.3 General assumptions and notation

Density and clusters. As in Section 1.4, let p be a bounded probability density with respect to the Lebesgue measure on $\mathbb{R}^d$ and μ be the measure on $\mathbb{R}^d$ induced by p. Given a fixed level parameter $t > 0$, the t-level set of the density p is defined as

$$L(t) = \overline{\{x \in \mathbb{R}^d \mid p(x) \geq t\}}.$$

where the bar denotes the topological closure (note that level sets are closed by assumptions in the noisy case, but this is not necessarily the case in the noise-free setting).
Geometry of the clusters. We define clusters as the connected components of $L(t)$ (where the term "connected component" is used in its topological sense). The number of clusters is denoted by m, and the clusters themselves by $C^{(1)}, \dots, C^{(m)}$. We set $\beta_{(i)} = \mu(C^{(i)})$, that means, $\beta_{(i)}$ denotes the probability mass in cluster $C^{(i)}$.
We assume that each cluster $C^{(i)}$ $(i = 1, \dots, m)$ is a compact and connected subset of $\mathbb{R}^d$, whose boundary $\partial C^{(i)}$ is a smooth $(d-1)$-dimensional submanifold in $\mathbb{R}^d$ with minimal curvature radius $\kappa^{(i)} > 0$. For $\nu \leq \kappa^{(i)}$, we define the collar set $Col^{(i)}(\nu) = \{x \in C^{(i)} \mid \operatorname{dist}(x, \partial C^{(i)}) \leq \nu\}$ and the maximal covering radius $\nu^{(i)}_{\max} = \max_{\nu \leq \kappa^{(i)}} \{\nu \mid C^{(i)} \setminus Col^{(i)}(\nu) \text{ connected }\}$. These quantities will be needed for the following reasons: It will be necessary to cover the inner part of each cluster by balls of a certain fixed radius z, and these balls are not supposed to extrude. Such a construction is only possible under assumptions on the maximal curvature of the boundary of the cluster. This will be particularly important in the noisy case, where all statements about the density estimator only hold in the inner part of the cluster.

For an arbitrary $\varepsilon > 0$, the connected component of $L(t-\varepsilon)$ which contains the cluster $C^{(i)}$ is denoted by $C_-^{(i)}(\varepsilon)$. Points in the set $C_-^{(i)}(\varepsilon) \setminus C^{(i)}$ will sometimes be referred to as boundary points. To express distances between the clusters, we assume that there exists some $\tilde{\varepsilon} > 0$ such that $\operatorname{dist}(C_-^{(i)}(2\tilde{\varepsilon}), C_-^{(j)}(2\tilde{\varepsilon})) \geq u^{(i)} > 0$ for all $i, j \in \{1, \ldots, m\}$. The numbers $u^{(i)}$ will represent lower bounds on the distances between cluster $C^{(i)}$ and the remaining clusters. Note that the existence of the $u^{(i)} > 0$ ensures that $C_-^{(i)}(2\varepsilon)$ does not contain any other clusters apart from $C^{(i)}$ for $\varepsilon < \tilde{\varepsilon}$. Analogously to the definition of $\beta_{(i)}$ above, we set $\tilde{\beta}_{(i)} = \mu(C_-^{(i)}(2\tilde{\varepsilon}))$. That means, $\tilde{\beta}_{(i)}$ denotes the mass of the enlarged set $C_-^{(i)}(2\tilde{\varepsilon})$. These definitions are illustrated in Figure 2.3. Furthermore, we introduce a

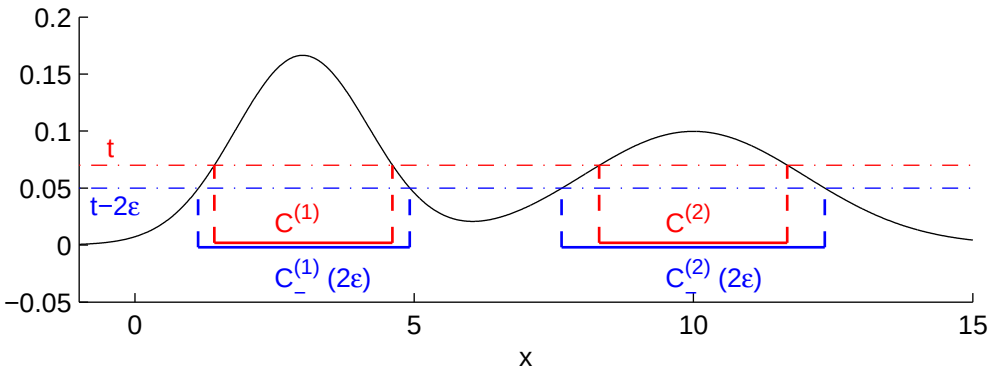

Figure 2.3: An example of our cluster definition. The clusters $C^{(1)}, C^{(2)}$ are defined as the connected components of the t-level set of the density (here $t = 0.07$). The clusters are subsets of the sets $C_-^{(1)}(2\varepsilon), C_-^{(2)}(2\varepsilon)$ (here for $\varepsilon = 0.01$).

lower bound on the probability mass in balls of radius $u^{(i)}$ around points in $C_-^{(i)}(2\tilde{\varepsilon})$

$$\rho^{(i)} \leq \inf_{x \in C_-^{(i)}(2\tilde{\varepsilon})} \mu\left(B(x, u^{(i)})\right).$$

In particular, under our assumptions on the smoothness of the cluster boundary we can set $\rho^{(i)} = O^{(i)} t \eta_d (u^{(i)})^d$ for an *overlap constant*

$$O^{(i)} = \inf_{x \in C_-^{(i)}(2\tilde{\varepsilon})} \left(\frac{\mathcal{L}_d(B(x, u^{(i)}) \cap C_-^{(i)}(2\tilde{\varepsilon}))}{\mathcal{L}_d(B(x, u^{(i)}))} \right) > 0.$$

The way it is constructed, $\rho^{(i)}$ becomes larger the larger the distance of $C^{(i)}$ to all the other clusters and is upper bounded by $\tilde{\beta}_{(i)}$, the probability mass of the extended cluster $C_-^{(i)}(2\tilde{\varepsilon})$.

Example in the noisy case. All assumptions on the density and the clusters are satisfied if we assume that the density p is twice continuously differentiable on a neighborhood

of $\{p = t\}$, for each $x \in \{p = t\}$ the gradient of p at x is non-zero and $\text{dist}(C^{(i)}, C^{(j)}) = u' > u^{(i)}$.

Example in the noise-free case. Here we assume that the support of the density p consists of m connected components $C^{(1)}, \dots, C^{(m)}$ which satisfy the smoothness assumptions above, and such that the densities on the connected components are lower bounded by a positive constant t. Then the noise-free case is a special case of the noisy case.

Sampling. As always in this thesis we assume that our n sample points $x_1, ..., x_n$ are sampled independently from the underlying probability distribution given by the density p.

Density estimation in the noisy case. In the noisy case we will estimate the density at each data point x_j by some estimate $\hat{p}_n(x_j)$. For convenience we state some of our results using a standard kernel density estimator, see Section A.3 for some background on kernel density estimation. However, our results can be easily rewritten with any other density estimate.

Further notation. Let the kNN radius $R^k(x_j)$ of a point x_j be defined as in Section 1.2. $R^{(i)}_{\min}$ denotes the minimal kNN radius of the sample points in cluster $C^{(i)}$, whereas $\tilde{R}^{(i)}_{\max}$ denotes the maximal kNN radius of the sample points in $C^{(i)}_-(2\tilde{\varepsilon})$. Note here the difference in the point sets that are considered.

An overview of the most important notations that are used in this chapter can be found in Table 2.1 and in the list of notations starting on page 127.

Table 2.1: Table of notations

$p(x)$	Density
$\hat{p}_n(x)$	Density estimate in point x
t	Density level set parameter
$L(t)$	t-level set of p
$C^{(1)}, \dots, C^{(m)}$	Clusters, i.e. connected components of $L(t)$
$C^{(i)}_-(\varepsilon)$	Connected component of $L(t-\varepsilon)$ containing $C^{(i)}$
$\beta_{(i)}, \tilde{\beta}_{(i)}$	Probability mass of $C^{(i)}$ and $C^{(i)}_-(2\tilde{\varepsilon})$ Respectively
$p^{(i)}_{\max}$	Maximal density in cluster $C^{(i)}$
$\rho^{(i)}$	Lower bound on probability of balls of radius $u^{(i)}$ around points in $C^{(i)}_-(2\tilde{\varepsilon})$
$\kappa^{(i)}$	Minimal curvature radius of the boundary $\partial C^{(i)}$
$\nu^{(i)}_{\max}$	Maximal covering radius of cluster $C^{(i)}$
$Col^{(i)}(\nu)$	Collar set for radius ν
$u^{(i)}$	Lower bound on the distances between $C^{(i)}$ and other clusters
$\tilde{\varepsilon}$	Parameter such that $\text{dist}(C^{(i)}_-(2\varepsilon), C^{(j)}_-(2\varepsilon)) \geq u^{(i)}$ for all $\varepsilon \leq \tilde{\varepsilon}$
η_d	Volume of the d-dimensional unit ball
k	Number of neighbors in the construction of the graph

2.4 Exact statements of the main results

In this section we are going to state all of our main results in a formal way. In the statement of the theorems we need the following conditions:

- *Condition 1:* Lower and upper bounds on the number of neighbors k,

$$k \geq 4^{d+1} \frac{p_{\max}^{(i)}}{t} \log\left(2\, 8^d\, p_{\max}^{(i)}\, \mathcal{L}_d(C^{(i)})\, n\right),$$
$$k \leq (n-1) \min\left\{ \frac{\rho^{(i)}}{2} - \frac{2\log(\tilde{\beta}_{(i)} n)}{(n-1)}, 2\, 4^d\, \eta_d\, p_{\max}^{(i)} \min\left\{(u^{(i)})^d, (\nu_{\max}^{(i)})^d\right\}\right\}.$$

- *Condition 2:* The density p is twice continuously differentiable with uniformly bounded derivatives, $\beta_{(i)} > 2\delta$, and ε_n sufficiently small such that $\mu\big(\bigcup_i (C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)})\big) \leq \delta/2$.

Condition 1 is necessary for both, the noise-free and the noisy case, whereas Condition 2 is only needed for the noisy case. Note that in Theorems 1 to 3 ε_n is considered small but constant and thus we drop the index n there.
In our first theorem, we present the optimal choice of the parameter k in the mutual kNN graph for the identification of one cluster. This theorem treats both the noise-free and the noisy case.

Theorem 2.1 (Optimal k for identification of one cluster in the mutual kNN graph) *The optimal choice of k for identification of cluster $C^{(i)}$ in $G_{mut}(n,k)$ (noise-free case) respectively rough identification in $\tilde{G}_{mut}(n,k,t-\varepsilon,\delta)$ (noisy case) is*

$$k = (n-1)\Gamma^{(i)} + 1, \quad \textit{with} \quad \Gamma^{(i)} = \frac{\rho^{(i)}}{2 + \frac{1}{4^d}\frac{t}{p_{\max}^{(i)}}},$$

provided this choice of k fulfills Condition 1.
In the noise-free case we obtain with $\Omega_{\textit{noise-free}}^{(i)} = \frac{\rho^{(i)}}{2\, 4^{d+1}\, \frac{p_{\max}^{(i)}}{t} + 4}$ and for sufficiently large n

$$\Pr\left(\textit{Cluster } C^{(i)} \textit{ is identified in } G_{mut}(n,k)\right) \geq 1 - 3\exp\left(-(n-1)\Omega_{\textit{noise-free}}^{(i)}\right).$$

For the noisy case, assume that additionally Condition 2 holds and let $\hat{p}_n$ be a kernel density estimator with bandwidth h. Then there exist constants C_1, C_2 such that if $h^2 \leq C_1\varepsilon$ we get with

$$\Omega_{\textit{noisy}}^{(i)} = \min\left\{ \frac{\rho^{(i)}}{2\, 4^{d+1}\, \frac{p_{\max}^{(i)}}{t} + 4},\ \frac{n}{n-1}\frac{\delta}{6},\ \frac{n}{n-1} C_2\, h^d\, \varepsilon^2 \right\}$$

and for sufficiently large n

$$\Pr\left(\textit{Cluster } C^{(i)} \textit{ roughly identified in } \tilde{G}_{mut}(n,k,t-\varepsilon,\delta)\right) \geq 1 - 8\exp\left(-(n-1)\Omega_{\textit{noisy}}^{(i)}\right).$$

This theorem has several remarkable features. First, we can see that both in the noise-free and in the noisy case, the optimal choice of k is roughly linear in n. A surprising result, given that the lower bound for cluster connectivity in the kNN graphs is $k \sim \log n$. We will discuss the important consequences of this result in the last section.
Second, we can see that for the mutual kNN graph the identification of one cluster $C^{(i)}$ only depends on the properties of the cluster $C^{(i)}$, but not on those of any other cluster. This is a unique feature of the mutual kNN graph which comes from the fact that if cluster $C^{(i)}$ is very "dense", then the neighborhood relationship of points in $C^{(i)}$ never links outside of cluster $C^{(i)}$. In the mutual kNN graph this implies that any connections of $C^{(i)}$ to other clusters are prevented. Note that this is not true for the symmetric kNN graph, where another cluster can simply link into $C^{(i)}$, irrespective of the internal properties of $C^{(i)}$.
For the mutual graph, it therefore makes sense to define the *most significant* cluster as the one with the largest coefficient $\Omega^{(i)}$, since this can be identified with the fastest rate. In the noise-free case one can observe that the coefficient $\Omega^{(i)}$ of cluster $C^{(i)}$ is large given that

- $\rho^{(i)}$ is large, which effectively means a large distance $u^{(i)}$ of $C^{(i)}$ to the closest other cluster,
- $p^{(i)}_{\max}/t$ is small, so that the density is rather uniform inside the cluster $C^{(i)}$.

Note that those properties are the most simple properties one would think of when imagining an "easily detectable" cluster. For the noisy case, a similar analysis still holds as long as one can choose the constants δ, h and ε small enough.
Formally, the result for the identification of single clusters in the symmetric kNN graph looks very similar to the one above.

Theorem 2.2 (Optimal k for identification of one cluster in symmetric kNN graph) *We use the same notation as in Theorem 2.1 and define $\rho_{\min} = \min_{i=1,\dots,m} \rho^{(i)}$. Then all statements about the optimal rates for k in Theorem 2.1 can be carried over to the symmetric kNN graph, provided one replaces $\rho^{(i)}$ with $\rho_{\min}$ in the definitions of $\Gamma^{(i)}$, $\Omega^{(i)}_{\text{noise-free}}$ and $\Omega^{(i)}_{\text{noisy}}$. If Condition 1 holds and the condition $k \leq (n-1)\rho_{\min}/2 - 2\log(n)$ replaces the corresponding one in Condition 1, we have in the noise-free case for a sufficiently large n*

$$\Pr\left(C^{(i)} \text{ is identified in } G_{sym}(n,k)\right) \geq 1 - (m+2)\exp\left(-(n-1)\Omega^{(i)}_{\text{noise-free}}\right).$$

If additionally Condition 2 holds, we have in the noisy case for sufficiently large n

$$\Pr\left(C^{(i)} \text{ roughly identified in } \tilde{G}_{sym}\left(n,k,t-\varepsilon,\delta\right)\right) \geq 1 - (m+7)\exp\left(-(n-1)\Omega^{(i)}_{\text{noisy}}\right).$$

The constant $\rho^{(i)}$ has now been replaced by the minimal $\rho^{(j)}$ among all clusters $C^{(j)}$, that means that the rate of convergence for the symmetric kNN graph is governed by the constant $\rho^{(j)}$ of the "worst" cluster, the one which is most difficult to identify. Intuitively, this "worst" cluster is the one which has the smallest distance to its neighboring

clusters. In contrast, for the mutual kNN graph the rate for identification of $C^{(i)}$ is governed by the cluster $C^{(i)}$ itself. This is a big disadvantage of the symmetric kNN graph if the goal is to only identify the "most significant" clusters. For this purpose the mutual graph has a clear advantage.
On the other hand, as we will see in the next theorem, the difference in the behavior between the mutual and symmetric graphs vanishes as soon as we attempt to identify *all* clusters.

Theorem 2.3 (Optimal k for identification of all clusters in the mutual kNN graph) *We use the same notation as in Theorem 2.1 and define $\rho_{\min} = \min_{i=1,\dots,m} \rho^{(i)}$, $p_{\max} = \max_{i=1,\dots,m} p^{(i)}_{\max}$. The optimal choice of k for the identification of all clusters in the mutual kNN graph in $G_{mut}(n,k)$ (noise-free case) respectively rough identification of all clusters in $\tilde{G}_{mut}(n,k,t-\varepsilon,\delta)$ (noisy case) is given by*

$$k = (n-1)\Gamma^{all} + 1, \quad \textit{with} \quad \Gamma^{all} = \frac{\rho_{\min}}{2 + \frac{1}{4^d}\frac{t}{p_{\max}}},$$

provided this choice of k fulfills Condition 1 for all clusters $C^{(i)}$. In the noise-free case we get the rate

$$\Omega_{\textit{noise-free}} = \frac{\rho_{\min}}{2\,4^{d+1}\,\frac{p_{\max}}{t} + 4},$$

such that for sufficiently large n

$$\Pr\left(\textit{All clusters exactly identified in } G_{mut}(n,k)\right) \geq 1 - 3m\,\exp\left(-(n-1)\Omega_{\textit{noise-free}}\right).$$

For the noisy case, assume that additionally Condition 2 holds for all clusters and let $\hat{p}_n$ be a kernel density estimator with bandwidth h. Then there exist constants C_1, C_2 such that if $h^2 \leq C_1\varepsilon$ we get with

$$\Omega_{\textit{noisy}} = \min\left\{\frac{\rho_{\min}}{2\,4^{d+1}\,\frac{p_{\max}}{t} + 4},\ \frac{n}{n-1}\frac{\delta}{6},\ \frac{n}{n-1}C_2\,h^d\,\varepsilon^2\right\}$$

and for sufficiently large n

$$\Pr\left(\textit{All clusters roughly identified in } \tilde{G}_{mut}(n,k,t-\varepsilon,\delta)\right) \geq 1 - (3m+5)\,\exp\left(-(n-1)\Omega_{\textit{noisy}}\right).$$

We can see that as in the previous theorem, the constant which now governs the speed of convergence is the worst case constant among all the $\rho^{(j)}$. In the setting where we wish to identify all clusters this is unavoidable. Of course the identification of "insignificant" clusters will be difficult, and the overall behavior will be determined by the most difficult case. This is reflected in the above theorem. The corresponding theorem for identification of all clusters in the symmetric kNN graph looks very similar, and we omit it.
So far for the noisy case we have mainly considered the case of rough cluster identification. As we have seen, in this setting the results of the noise-free case are very similar to the ones in the noisy case. Now we would like to conclude with a theorem for exact cluster identification in the noisy case.

Theorem 2.4 (Exact identification of clusters in the noisy case) *Let p be twice continuously differentiable with uniformly bounded derivatives and let the gradient of p be non-zero in a neighborhood of $\{p = t\}$. Let $\hat{p}_n$ be a kernel density estimator with bandwidth $h_n = h_0(\log n/n)^{1/(d+4)}$ for some $h_0 > 0$. For a suitable constant $\varepsilon_0 > 0$ set $\varepsilon_n = \varepsilon_0(\log n/n)^{2/(d+4)}$. Then there exists constants c_1, c_2 such that for $n \to \infty$ and $c_1 \log n \leq k \leq c_2 n$ we obtain*

$$\textit{Cluster } C^{(i)} \textit{ is exactly identified in } \tilde{G}_{mut}(n, k, t - \varepsilon_n, \delta).$$

Note that as opposed to rough cluster identification, which is a statement about a given finite nearest neighbor graph, exact cluster identification is an inherently asymptotic property. The complication in this asymptotic setting is that one has to balance the speed of convergence of the density estimator with the one of the "convergence of the graph". The exact form of the density estimation is not important. Every other density estimator with the same convergence rate would yield the same result, possibly even under weaker assumptions. Finally, note that since it is technically difficult to grasp the graph after the small components have been discarded, we could not prove what the optimal k in this setting should be.

2.5 Proofs

The propositions and lemmas containing the major proof steps are presented in Section 2.5.1. The proofs of the theorems themselves can be found in Section 2.5.2. An overview of the proof structure can be seen in Figure 2.4.

2.5.1 Main propositions for cluster identification

In Proposition 2.5 we identify some events whose combination guarantees the connectedness of a cluster in the graph and at the same time ensures that there is not a connected component of the graph that consists of background points only. The probabilities of the events appearing in the proposition are then bounded in Lemmas 2.6 – 2.9. In Proposition 2.10 and Lemma 2.11 we examine the probability of connections between clusters. The section concludes with Proposition 2.12 and Lemma 2.13, which are used in the exact cluster identification in Theorem 2.4, and some remarks about the differences between the noise-free and the noisy case. In the proofs we make frequent use of Theorem A.1 stating tail bounds for the binomial distribution, which have been introduced by Hoeffding.

Proposition 2.5 (Connectedness of one cluster $C^{(i)}$ in the noisy case) *Let $C_n^{(i)}$ denote the event that in $\tilde{G}_{mut}(n, k, t - \varepsilon_n, \delta)$ (respectively $\tilde{G}_{sym}(n, k, t - \varepsilon_n, \delta)$)*

- *all the sample points from $C^{(i)}$ are contained in the graph,*
- *the sample points from $C^{(i)}$ are connected in the graph,*

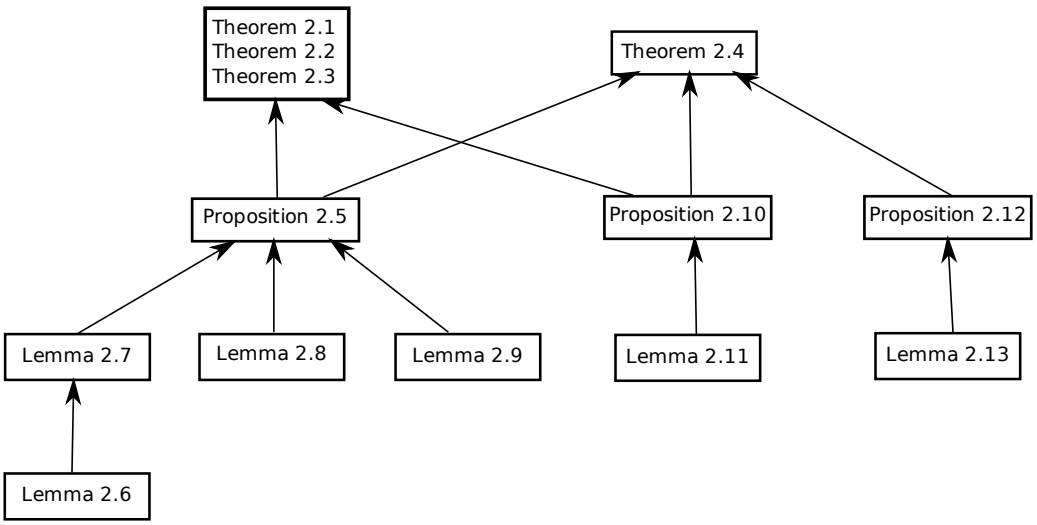

Figure 2.4: The structure of our proofs. Proposition 2.5 deals with within-cluster connectedness and Proposition 2.10 with between-cluster disconnectedness. Proposition 2.12 bounds the ratio of background and cluster points for the asymptotic analysis of exact cluster identification.

- *there exists no component of the graph which consists only of sample points from outside* $L(t)$.

Then under the conditions

1. $\beta_{(i)} > 2\delta$,
2. ε_n *sufficiently small such that* $\mu\big(\bigcup_i (C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)})\big) \leq \delta/2$,
3. $k \geq 4^{d+1} \frac{p_{\max}^{(i)}}{t} \log\left(2\, 8^d\, p_{\max}^{(i)}\, \mathcal{L}_d(C^{(i)})\, n\right)$,
   $k \leq (n-1) 2\, 4^d\, \eta_d\, p_{\max}^{(i)} \min\left\{(u^{(i)})^d, (v_{\max}^{(i)})^d\right\}$,

and for sufficiently large n, *we obtain*

$$\begin{aligned}\Pr\left((\mathcal{C}_n^{(i)})^c\right) &\leq \Pr\left((\mathcal{A}_n^{(i)})^c\right) + \Pr\left((\mathcal{B}_n^{(i)})^c\right) + \Pr(\mathcal{E}_n^c) + \Pr(\mathcal{D}_n^c) \\ &\leq 2\, \exp\left(-\frac{k-1}{4^{d+1}} \frac{t}{p_{\max}^{(i)}}\right) + 2\exp\left(-n\frac{\delta}{6}\right) + 2\Pr(\mathcal{D}_n^c),\end{aligned}$$

where the events are defined as follows:

- $\mathcal{A}_n^{(i)}$: *the subgraph consisting of points from* $C^{(i)}$ *is connected in* $G'_{mut}(n, k, t - \varepsilon_n)$ *(resp.* $G'_{sym}(n, k, t - \varepsilon_n)$*),*

- $\mathcal{B}_n^{(i)}$: *there are more than* δn *sample points from cluster* $C^{(i)}$,
- $\mathcal{E}_n$: *there are less than* δn *sample points in the set* $\bigcup_i \left(C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)}\right)$, *and*
- $\mathcal{D}_n$: $|\hat{p}_n(x_i) - p(x_i)| \leq \varepsilon_n$ *for all sample points* x_i, $i = 1, \ldots, n$.

Proof. We bound the probability of $\mathcal{C}_n^{(i)}$ using the observation that $\mathcal{A}_n^{(i)} \cap \mathcal{B}_n^{(i)} \cap \mathcal{E}_n \cap \mathcal{D}_n \subseteq \mathcal{C}_n^{(i)}$ implies

$$\Pr\left((\mathcal{C}_n^{(i)})^c\right) \leq \Pr\left((\mathcal{A}_n^{(i)})^c\right) + \Pr\left((\mathcal{B}_n^{(i)})^c\right) + \Pr(\mathcal{E}_n^c) + \Pr(\mathcal{D}_n^c). \tag{2.1}$$

This results from the following chain of observations. If the event $\mathcal{D}_n$ holds, no point with $p(x_i) \geq t$ is removed, since on this event $p(x_i) - \hat{p}_n(x_i) \leq \varepsilon_n$ and thus $\hat{p}_n(x_i) \geq p(x_i) - \varepsilon_n \geq t - \varepsilon_n$, which is the threshold in the graph $G'(n, k, t - \varepsilon_n)$.
If the samples in cluster $C^{(i)}$ are connected in $G'(n, k, t - \varepsilon_n)$ ($\mathcal{A}_n^{(i)}$), and there are more than δn samples in the cluster $C^{(i)}$ ($\mathcal{B}_n^{(i)}$), then the resulting component of the graph $G'(n, k, t - \varepsilon_n)$ is not removed in the algorithm and is thus contained in $\tilde{G}(n, k, t - \varepsilon_n, \delta)$.
Conditional on $\mathcal{D}_n$ all remaining samples are contained in $\bigcup_i C_-^{(i)}(2\varepsilon_n)$. Thus all non-cluster samples lie in $\bigcup_i (C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)})$. Given that this set contains less than δn samples, there exists no connected component only consisting of non-cluster points, which implies that all remaining non-cluster points are connected to one of the clusters.
The probabilities for the complements of the events $\mathcal{A}_n^{(i)}$, $\mathcal{B}_n^{(i)}$ and $\mathcal{E}_n$ are bounded in Lemmas 2.7 – 2.9 below. Plugging in those bounds into Equation (2.1) leads to the desired result. □

In the following lemmas we derive bounds for the probabilities of the events introduced in the proposition above.

Lemma 2.6 (Within-cluster connectedness $(\mathcal{A}_n^{(i)})$**)** *As in Proposition 2.5 let* $\mathcal{A}_n^{(i)}$ *denote the event that the points of cluster* $C^{(i)}$ *are connected in* $G'_{mut}(n, k, \varepsilon_n)$ *(respectively* $G'_{sym}(n, k, \varepsilon_n)$*). For* $z \in (0, 4 \min\{u^{(i)}, \nu_{\max}^{(i)}\})$,

$$\Pr\left((\mathcal{A}_n^{(i)})^c\right) \leq n\, \beta_{(i)}\, \Pr(M \geq k) + N \left(1 - t\, \eta_d \frac{z^d}{4^d}\right)^n + \Pr\left(\mathcal{D}_n^c\right),$$

where M *is a* $\mathrm{Bin}(n-1, p_{\max}^{(i)} \eta_d z^d)$*-distributed random variable and we have the upper bound* $N \leq 8^d \mathcal{L}_d(C^{(i)})/(z^d \eta_d)$.

Proof. Given that $\mathcal{D}_n$ holds, all samples lying in cluster $C^{(i)}$ are contained in the graph $G'(n, k, \varepsilon_n)$. Suppose that we have a covering of $C^{(i)} \setminus Col^{(i)}(z/4)$ with balls of radius $z/4$. By construction every ball of the covering lies entirely in $C^{(i)}$, so that t is a lower bound for the minimal density in each ball. If every ball of the covering contains at least one sample point and the minimal kNN radius of samples in $C^{(i)}$ is larger than

or equal to z, then all samples of $C^{(i)} \setminus Col^{(i)}(z/4)$ are connected in $G'(n,k,\varepsilon_n)$ given that $z \leq 4\nu_{\max}^{(i)}$. Moreover, one can easily check that all samples lying in the collar set $Col^{(i)}(z/4)$ are connected to $C^{(i)} \setminus Col^{(i)}(z/4)$. In total, all sample points lying in $C^{(i)}$ are connected. Figure 2.5 illustrates the covering of the collar set and the implications for the connectivity of the kNN graph.

Denote by $\mathcal{F}_z^{(i)}$ the event that one ball in the covering with balls of radius $z/4$ contains no sample point. Formally, $\{R_{\min}^{(i)} > z\} \cap (\mathcal{F}_z^{(i)})^c$ implies connectedness of the samples lying in $C^{(i)}$ in the graph $G'(n,k,\varepsilon_n)$.

Define $N_s = |\{j \neq s \mid x_j \in B(x_s,z)\}|$ for $1 \leq s \leq n$. Then $\{R_{\min}^{(i)} \leq z\} = \cup_{s=1}^n \{\{N_s \geq k\} \cap \{x_s \in C^{(i)}\}\}$. We have

$$\Pr\left(R_{\min}^{(i)} \leq z\right) \leq \sum_{s=1}^{n} \Pr\left(N_s \geq k \,|\, x_s \in C^{(i)}\right) \Pr\left(x_s \in C^{(i)}\right) \leq n\beta_{(i)} \Pr(U \geq k),$$

where $U \sim \mathrm{Bin}(n-1, \sup_{x \in C^{(i)}} \mu(B(x,z))$. The final result is obtained using the upper bound $\sup_{x \in C^{(i)}} \mu(B(x,z)) \leq p_{\max}^{(i)} \eta_d z^d$.

A standard construction using a $z/4$-packing provides us with the covering. Since $z/4 \leq \nu_{\max}^{(i)}$ we know that balls of radius $z/8$ around the packing centers are subsets of $C^{(i)}$ and are disjoint by construction. Thus, the total volume of the N balls is bounded by the volume of $C^{(i)}$ and we get $N(z/8)^d \eta_d \leq \mathcal{L}_d(C^{(i)})$. Since we assume that $\mathcal{D}_n$ holds, no sample lying in $C^{(i)}$ has been discarded. Thus the probability for one ball of the covering being empty can be upper bounded by $(1 - t\,\eta_d\, z^d/4^d)^n$, where we have used the fact that the balls of the covering are entirely contained in $C^{(i)}$ and thus the density is lower bounded by t. In total, a union bound over all balls in the covering yields,

$$\Pr\left(\mathcal{F}_z^{(i)}\right) \leq N\,(1 - t\,\eta_d\, z^d/4^d)^n + \Pr\left(\mathcal{D}_n^c\right).$$

Utilizing both results together yields the final result. □

In Lemma 2.6 we provided a bound on the probability which includes two competing terms for the choice of z. One favors small z whereas the other favors large z. The next lemma will show how to choose the radius z in terms of k.

Lemma 2.7 (Choice of k for within-cluster connectedness $(\mathcal{A}_n^{(i)})$) *If the parameter k fulfils Condition (3) of Proposition 2.5, we have for sufficiently large n*

$$\Pr\left((\mathcal{A}_n^{(i)})^c\right) \leq 2\,\exp\left(-\frac{k-1}{4^{d+1}} \frac{t}{p_{\max}^{(i)}}\right) + \Pr\left(\mathcal{D}_n^c\right).$$

Proof. The upper bound on the probability of $(\mathcal{A}_n^{(i)})^c$ given in Lemma 2.6 has two terms dependent on z. The tail bound for the binomial distribution is small if z is chosen to be small, whereas the term from the covering is small given that z is large. Here, we find a choice for z which is close to optimal. Define $p = p_{\max}^{(i)} \eta_d z^d$ and $\alpha = k/(n-1)$. Using

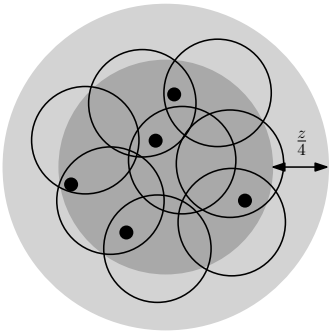

Figure 2.5: A cluster $C^{(i)}$ (in dark grey) with a covering of the set $C^{(i)} \setminus Col^{(i)}(z/4)$ (in light grey), where each ball of the covering contains at least one sample point. Since $z/4 \leq \kappa^{(i)}$ every point in in the collar set $Col^{(i)}(z/4)$ has a distance at most $z/4$ to some point in $C^{(i)} \setminus Col^{(i)}(z/4)$. Thus, if all the balls of the covering contain a sample point, those in the collar set are connected in the k-nearest neighbor graph as well. Due to the condition $z \leq 4\, v_{\max}^{(i)}$ the dark set is connected and thus all the sample points are connected if their minimal k-nearest neighbor radius is greater than z.

Theorem A.1 we obtain for $M \sim \mathrm{Bin}(n-1, p)$ and a choice of z such that $p < \alpha$,

$$\begin{aligned} n\beta_{(i)} \Pr(M \geq k) &\leq n\beta_{(i)} \exp\left(-(n-1)\left(\alpha \log\left(\frac{\alpha}{p}\right) + (1-\alpha)\log\left(\frac{1-\alpha}{1-p}\right)\right)\right) \\ &\leq n\beta_{(i)} \exp\left(-(n-1)\left(\alpha \log\left(\frac{\alpha}{p}\right) + p - \alpha\right)\right), \end{aligned}$$

where we have used $\log(z) \geq (z-1)/z$ for $z > 0$. Now introduce $\theta = \eta_d z^d/\alpha$ such that $p = p_{\max}^{(i)}\, \theta\, \alpha$, where $p \leq \alpha$ implies $0 \leq \theta\, p_{\max}^{(i)} \leq 1$. Then,

$$\begin{aligned} n\beta_{(i)} \Pr(M \geq k) &\leq n\beta_{(i)} \exp\left(-k\left(\log\left(\frac{1}{p_{\max}^{(i)}\theta}\right) + \theta p_{\max}^{(i)} - 1\right)\right) \\ &\leq \exp\left(-\frac{k}{2}\left(\log\left(\frac{1}{p_{\max}^{(i)}\theta}\right) + \theta p_{\max}^{(i)} - 1\right)\right), \end{aligned} \tag{2.2}$$

where we used in the last step an upper bound on the term $n\beta_{(i)}$ which holds given $k \geq \left(2\log(\beta_{(i)} n)\right) / \left(\log(1/(\theta p_{\max}^{(i)})) + \theta p_{\max}^{(i)} - 1\right)$. On the other hand,

$$N\left(1 - \frac{t\,\eta_d\, z^d}{4^d}\right)^n = N \exp\left(n \log\left(1 - \frac{t\,\eta_d\, z^d}{4^d}\right)\right) \leq N \exp\left(-\frac{n\, t\, \eta_d\, z^d}{4^d}\right)$$

where we used $\log(1-x) \geq -x$ for $x \leq 1$. With $\eta_d z^d = \theta\alpha$ and the upper bound on N we get using $n/(n-1) \geq 1$,

$$N \exp\left(-\frac{n\,t\,\eta_d\,z^d}{4^d}\right) \leq \exp\left(-\frac{n\,t\,\theta\,\alpha}{4^d} + \log\left(\frac{\mathcal{L}_d(C^{(i)})8^d}{\theta\,\alpha}\right)\right)$$
$$\leq \exp\left(-k\frac{t\,\theta}{4^d} + \log\left(\frac{\mathcal{L}_d(C^{(i)})8^d}{\theta\,\alpha}\right)\right) \leq \exp\left(-k\frac{t\,\theta}{2\,4^d}\right), \quad (2.3)$$

where the last step holds given $k \geq 2\,4^d \log(n\mathcal{L}_d(C^{(i)})8^d/\theta)/(t\theta)$. Upper bounding the bound in (2.2) with the one in (2.3) requires,

$$\frac{t\theta}{2\,4^d} \leq \frac{1}{2}\left(\log\left(\frac{1}{p_{\max}^{(i)}\theta}\right) + \theta p_{\max}^{(i)} - 1\right).$$

After introducing $\gamma = \theta\, p_{\max}^{(i)}$, the inequality $\gamma t/(4^d p_{\max}^{(i)}) \leq (-\log(\gamma) + \gamma - 1)$ is equivalent to the last one. Note, that $t/(4^d p_{\max}^{(i)}) \leq 1/4$. Thus, the above inequality holds for all $d \geq 1$ given that $-\log(\gamma) \geq 1 - 3\gamma/4$. A simple choice is $\gamma = 1/2$ and thus $\theta = 1/(2p_{\max}^{(i)})$, which fulfills $\theta p_{\max}^{(i)} \leq 1$. In total, we obtain with the result from Lemma 2.6,

$$\Pr\left((\mathcal{A}_n^{(i)})^c\right) \leq 2\,\exp\left(-\frac{k}{4^{d+1}}\frac{t}{p_{\max}^{(i)}}\right) + \Pr\left(\mathcal{D}_n^c\right) \leq 2\,\exp\left(-\frac{k-1}{4^{d+1}}\frac{t}{p_{\max}^{(i)}}\right) + \Pr\left(\mathcal{D}_n^c\right).$$

We insert the choice of θ into the lower bounds on k. One can easily find an upper bound for the maximum of the two lower bounds which gives,

$$k \geq 4^{d+1}\frac{p_{\max}^{(i)}}{t}\log\left(2\,8^d\,p_{\max}^{(i)}\,\mathcal{L}_d(C^{(i)})\,n\right).$$

The upper bound, $z \leq 4\min\{u^{(i)}, v_{\max}^{(i)}\}$, translates into the following upper bound on k, $k \leq (n-1)2\,4^d\,\eta_d\,p_{\max}^{(i)}\min\left\{(u^{(i)})^d, (v_{\max}^{(i)})^d\right\}$. □

The result of this lemma means that if we choose $k \geq c_1 + c_2 \log n$ with two constants c_1, c_2 that depend on the geometry of the cluster and the respective density, then the probability that the cluster is disconnected approaches zero exponentially in k.
Note that due to the constraints on the covering radius, we have to introduce an upper bound on k which depends linearly on n. However, as the probability of connectedness is monotonically increasing in k, the value of the within-connectedness bound for this value of k is a lower bound for all larger k as well. Since the lower bound on k grows with $\log n$ and the upper bound grows with n, there exists a feasible region for k if n is large enough.

Lemma 2.8 (Event $\mathcal{B}_n^{(i)}$) *As in Proposition 2.5 let $\mathcal{B}_n^{(i)}$ denote the event that there are more than δn sample points from cluster $C^{(i)}$. If $\beta_{(i)} > 2\delta$ then*

$$\Pr\left(\left(\mathcal{B}_n^{(i)}\right)^c\right) \leq \exp\left(-\frac{1}{4}n\delta\right).$$

Proof. Let $M^{(i)}$ be the number of samples in cluster $C^{(i)}$. Then $M^{(i)} \sim \text{Bin}(n, \beta_{(i)})$ and we have

$$\begin{aligned}\Pr\left(M^{(i)} < \delta n\right) &\leq \Pr\left(M^{(i)} < \frac{\delta}{\beta_{(i)}}\beta_{(i)}n\right) = \Pr\left(M^{(i)} < \left(1 - \left(1 - \frac{\delta}{\beta_{(i)}}\right)\right)\beta_{(i)}n\right) \\ &\leq \exp\left(-\frac{1}{2}n\beta_{(i)}\left(1 - \frac{\delta}{\beta_{(i)}}\right)^2\right) \leq \exp\left(-\frac{1}{4}n\delta\right),\end{aligned}$$

where we used the tail bound for the binomial distribution in Theorem A.2. □

Lemma 2.9 (Event $\mathcal{E}_n$) *As in Proposition 2.5 let $\mathcal{E}_n$ denote the event that there are less than δn sample points in all the boundary sets $C_-^{(j)}(2\varepsilon_n) \setminus C^{(j)}$ together. If $\sum_{j=1}^m \mu\left(C_-^{(j)}(2\varepsilon_n) \setminus C^{(j)}\right) < \delta/2$, we have $\Pr(\mathcal{E}_n^c) \leq \exp(-\delta n/6)$.*

Proof. By assumption we have $\sum_{j=1}^m \mu(C_-^{(j)}(2\varepsilon_n) \setminus C^{(j)}) < \delta/2$ for the probability mass in the boundary strips. Then the probability that there are at least δn points in the boundary strips can be bounded by the probability that a $\text{Bin}(n, \delta/2)$-distributed random variable V exceeds δn. We obtain

$$\begin{aligned}\Pr(V > \delta n) &= \Pr\left(V > 2\frac{\delta}{2}n\right) = \Pr\left(V > (1+1)\frac{\delta}{2}n\right) \leq \exp\left(-\frac{1}{3}\frac{\delta}{2}n\right) \\ &= \exp\left(-\frac{1}{6}\delta n\right)\end{aligned}$$

using the tail bound from Theorem A.2. □

The proposition and the lemmas above are used in the analysis of within-cluster connectedness. The following proposition deals with between-cluster disconnectedness. We say that a cluster $C^{(i)}$ is *isolated* if the subgraph of $\tilde{G}_{\text{mut}}(n, k, t - \varepsilon_n, \delta)$ (resp. $\tilde{G}_{\text{sym}}(n, k, t - \varepsilon_n, \delta)$) corresponding to cluster $C^{(i)}$ is not connected to another subgraph corresponding to any other cluster $C^{(j)}$ with $j \neq i$. Note, that we assume $\min_{j=1,\ldots,m} \text{dist}(C_-^{(i)}(2\varepsilon_n), C_-^{(j)}(2\varepsilon_n)) \geq u^{(i)}$ for all $\varepsilon_n \leq \tilde{\varepsilon}$. The following proposition bounds the probability for cluster isolation. This bound involves the probability that the maximal k-nearest-neighbor radius is greater than a given threshold. Therefore in Lemma 2.11 we derive a bound for this probability. Note that the paper Maier et al. [56], which this chapter is partly based on, contained an error in the result corresponding to Lemma 2.11, which changed some constants but did not affect the main results.

Proposition 2.10 (Cluster isolation) *Let $\mathcal{I}_n^{(i)}$ denote the event that the subgraph of the samples in $C_-^{(i)}(2\varepsilon_n)$ is isolated in $\tilde{G}_{mut}(n,k,t-\varepsilon_n,\delta)$. Then given that $\varepsilon_n \leq \tilde{\varepsilon}$, $k < \rho^{(i)}n/2 - 2\log(\tilde{\beta}_{(i)}n)$, we obtain*

$$\begin{aligned}\Pr\left((\mathcal{I}_n^{(i)})^c\right) &\leq \Pr\left(\tilde{R}_{\max}^{(i)} \geq u^{(i)}\right) + \Pr\left(\mathcal{D}_n^c\right) \\ &\leq \exp\left(-\frac{n-1}{2}\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)\right) + \Pr\left(\mathcal{D}_n^c\right).\end{aligned}$$

Let $\hat{\mathcal{I}}_n^{(i)}$ be the event that the subgraph of samples in $C_-^{(i)}(2\varepsilon_n)$ is isolated in $\tilde{G}_{sym}(n,k,t-\varepsilon_n,\delta)$. Define $\rho_{\min} = \min_{i=1,\dots,m}\rho^{(i)}$ and $\tilde{\beta}_{\max} = \max_{i=1,\dots,m}\tilde{\beta}_{(i)}$. Then for $\varepsilon_n \leq \tilde{\varepsilon_n}$, $k < \rho_{\min}n/2 - 2\log(\tilde{\beta}_{\max}n)$, we obtain

$$\begin{aligned}\Pr\left((\hat{\mathcal{I}}_n^{(i)})^c\right) &\leq \sum_{j=1}^{m}\Pr\left(\tilde{R}_{\max}^{(j)} \geq u^{(j)}\right) + \Pr\left(\mathcal{D}_n^c\right) \\ &\leq m\,\exp\left(-\frac{n-1}{2}\left(\frac{\rho_{\min}}{2} - \frac{k-1}{n-1}\right)\right) + \Pr\left(\mathcal{D}_n^c\right).\end{aligned}$$

Proof. We have $\Pr((\mathcal{I}_n^{(i)})^c) \leq \Pr((\mathcal{I}_n^{(i)})^c \mid \mathcal{D}_n) + \Pr(\mathcal{D}_n^c)$. Given the event $\mathcal{D}_n$, the remaining points in $\tilde{G}_{\text{mut}}(n,k,t-\varepsilon_n,\delta)$ are samples from $C_-^{(j)}(2\varepsilon_n)$ ($j=1,\dots,m$). By assumption we have for $\varepsilon_n \leq \tilde{\varepsilon}$ that $\min_{j\neq i}\operatorname{dist}(C_-^{(i)}(2\varepsilon_n), C_-^{(j)}(2\varepsilon_n)) \geq u^{(i)}$. In order to have edges from samples in $C_-^{(i)}(2\varepsilon_n)$ to any other part in $\tilde{G}_{\text{mut}}(n,k,t-\varepsilon_n,\delta)$, it is necessary that $\tilde{R}_{\max}^{(i)} \geq u^{(i)}$. Using Lemma 2.11 we can lower bound the probability of this event. For the symmetric kNN graph there can be additional edges from samples in $C_-^{(i)}(2\varepsilon_n)$ to other parts in the graph if those lying in $C_-^{(i)}(2\varepsilon_n)$ are among the k nearest neighbors of samples in $C_-^{(j)}(2\varepsilon_n)$, $j \neq i$. Let u^{ij} be the distance between $C_-^{(i)}(2\tilde{\varepsilon})$ and $C_-^{(j)}(2\tilde{\varepsilon})$. There can be edges from samples in $C_-^{(i)}(2\varepsilon_n)$ to any other part in $\tilde{G}_{\text{sym}}(n,k,\varepsilon_n,\delta)$ if the following event holds: $\{\tilde{R}_{\max}^{(i)} \geq u^{(i)}\} \cup \{\cup_{j\neq i}\{\tilde{R}_{\max}^{(j)} \geq u^{ij}\}\}$. Using a union bound we obtain

$$\Pr\left(\left(\hat{\mathcal{I}}_n^{(i)}\right)^c \mid \mathcal{D}_n\right) \leq \Pr\left(\tilde{R}_{\max}^{(i)} \geq u^{(i)}\right) + \sum_{j\neq i}\Pr\left(\tilde{R}_{\max}^{(j)} \geq u^{ij}\right).$$

With $u^{(j)} \leq u^{ij}$ and Lemma 2.11 we obtain the result for $\tilde{G}_{\text{sym}}(n,k,\varepsilon_n,\delta)$. □

Note that the upper bound on the probability that $C^{(i)}$ is isolated is the same for all clusters in the graph based on the symmetric kNN graph. The upper bound is loose in the sense that it does not respect specific geometric configurations of the clusters where the bound could be smaller. However, it is tight in the sense that the probability that cluster $C^{(i)}$ is isolated in $\tilde{G}_{\text{sym}}(n,k,\varepsilon_n,\delta)$ always depends on the *worst* cluster. This is the main difference to the mutual kNN graph, where the properties of cluster $C^{(i)}$ are

independent of the other clusters. This is illustrated by the following example: Given two clusters $C^{(1)}$ and $C^{(2)}$ in a distance which is larger than their diameter, and let $n^{(1)}$ and $n^{(2)}$ denote the number of points in the clusters. Then a point in $C^{(1)}$ has a k-nearest neighbor in the other cluster if and only if $k > n^{(1)} - 1$ and a similar statement holds for the other cluster. So there are connections between the clusters in the symmetric kNN graph if and only if $k > n^{(1)} - 1$ or $k > n^{(2)} - 1$, whereas there are connections in the mutual kNN graph if and only if $k > n^{(1)} - 1$ and $k > n^{(2)} - 1$. Should the weights $\beta_{(1)}$ and $1 - \beta_{(1)}$ of the clusters be very different, then it is obvious that the condition for the mutual graph is much stronger.
The following lemma states the upper bound for the probability that the maximum k-nearest neighbor radius $\tilde{R}^{(i)}_{\max}$ of samples in $C^{(i)}_{-}(2\varepsilon_n)$ used in the proof of Proposition 2.10.

Lemma 2.11 (Maximal kNN **radius)** *Let $k < \rho^{(i)} n/2 - 2\log(\tilde{\beta}_{(i)} n)$. Then*

$$\Pr\left(\tilde{R}^{(i)}_{\max} \geq u^{(i)}\right) \leq \exp\left(-\frac{n-1}{2}\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)\right).$$

Proof. Define $N_s = |\{j \neq s \,|\, x_j \in B(x_s, u^{(i)})\}|$ for $1 \leq s \leq n$. Then $\{\tilde{R}^{(i)}_{\max} \geq u^{(i)}\} = \bigcup_{s=1}^{n} \{N_s \leq k-1 \cap x_s \in C^{(i)}_{-}(2\tilde{\varepsilon})\}$. Thus,

$$\Pr\left(\tilde{R}^{(i)}_{\max} \geq u^{(i)}\right) \leq \sum_{s=1}^{n} \Pr\left(N_s \leq k-1 \,|\, x_s \in C^{(i)}_{-}(2\tilde{\varepsilon})\right) \Pr\left(x_s \in C^{(i)}_{-}(2\tilde{\varepsilon})\right).$$

Let $M \sim \mathrm{Bin}(n-1, \rho^{(i)})$. Then $\Pr(N_s \leq k-1 \,|\, x_s \in C^{(i)}_{-}(2\tilde{\varepsilon})) \leq \Pr(M \leq k-1)$. Using the tail bound from Theorem A.1 we obtain for $k - 1 < \rho^{(i)}(n-1)$,

$$\begin{aligned}\Pr\left(\tilde{R}^{(i)}_{\max} \geq u^{(i)}\right) &\leq n\,\tilde{\beta}_{(i)}\,\Pr(M \leq k-1) \leq n\tilde{\beta}_{(i)} \exp\left(-(n-1)\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)\right) \\ &\leq \exp\left(-\frac{n-1}{2}\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)\right),\end{aligned}$$

where we use that $\log(x) \geq (x-1)/x$, that $-w/e$ is the minimum of $x\log(x/w)$ attained at $x = w/e$ and $(1 - 1/e) \geq 1/2$. Finally, we use that under the stated condition on k we have $\log(n\tilde{\beta}_{(i)}) \leq [(n-1)\rho^{(i)}/2 - (k-1)]/2$. □

The following proposition quantifies the rate of *exact cluster identification*, that means how fast the fraction of points from outside the level set $L(t)$ approaches zero.

Proposition 2.12 (Ratio of boundary and cluster points) *Let $N_{Cluster}$ and $N_{NoCluster}$ be the number of cluster points and background points in $\tilde{G}_{mut}(n, k, t - \varepsilon_n, \delta)$ (resp.*

$\tilde{G}_{sym}(n, k, t-\varepsilon_n, \delta)$) and let $\mathcal{C}_n^{all}$ denote the event that the points of each cluster form a connected component of the graph. Let $\varepsilon_n \to 0$ for $n \to \infty$ and define $\beta = \sum_{i=1}^m \beta_{(i)}$. Then there exists a constant $\bar{D} > 0$ such that for sufficiently large n,

$$\Pr\left(N_{NoCluster}/N_{Cluster} > 4\frac{\bar{D}}{\beta}\varepsilon_n \,\middle|\, \mathcal{C}_n^{all}\right) \leq \exp\left(-\frac{1}{4}\bar{D}\varepsilon_n n\right) + \exp\left(-n\frac{\beta}{8}\right) + \Pr(\mathcal{D}_n^c).$$

Proof. According to Lemma 2.13 we can find constants $\bar{D}^{(i)} > 0$ such that $\mu(C_-^{(i)}(2\varepsilon_n) \backslash C^{(i)}) \leq \bar{D}^{(i)}\varepsilon_n$ for n sufficiently large, and set $\bar{D} = \sum_{i=1}^m \bar{D}^{(i)}$. Suppose that $\mathcal{D}_n$ holds. Then the only points which do not belong to a cluster lie in the set $\cup_{i=1}^m C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)}$. Some of them might be discarded, but since we are interested in proving an upper bound on $N_{\text{NoCluster}}$ that does not matter. Then with $p = \mathbb{E}\, N_{\text{NoCluster}}/n \leq \bar{D}\varepsilon_n$ and $\alpha = 2\bar{D}\varepsilon_n$ we obtain with Theorem A.1 and for sufficiently small ε_n,

$$\Pr\left(N_{\text{NoCluster}} \geq 2\bar{D}\varepsilon_n n \,\middle|\, \mathcal{C}_n^{\text{all}}, \mathcal{D}_n\right) \leq \exp\left(-nK(\alpha||p)\right) \leq \exp\left(-n\,\varepsilon_n\,\bar{D}\,(2\log(2)-1)\right),$$

where K denotes the Kullback-Leibler divergence. Here we have used that for $p \leq \bar{D}\varepsilon_n$ we have $K(\alpha||p) \geq K(\alpha||\bar{D}\varepsilon_n)$, and with $\log(1+x) \geq x/(1+x)$ for $x > -1$ we have $K(2\bar{D}\varepsilon_n||\bar{D}\varepsilon_n) \geq \bar{D}\varepsilon_n(2\log 2 - 1) \geq \bar{D}\varepsilon_n/4$. Given that $\mathcal{D}_n$ holds and the points of each cluster are a connected component of the graph, we know that all cluster points remain in the graph and we have

$$\Pr\left(N_{\text{Cluster}} \leq \frac{\beta n}{2} \,\middle|\, \mathcal{C}_n^{\text{all}}, \mathcal{D}_n\right) \leq \exp\left(-n\frac{\beta}{8}\right)$$

using Theorem A.1 and similar arguments as above. □

Lemma 2.13 *Assume that $p \in C^2(\mathbb{R}^d)$ with $\|p\|_\infty = p_{\max}$ and that for each x in a neighborhood of $\{p = t\}$ the gradient of p at x is non-zero, then there exists a constant $\bar{D}^{(i)} > 0$ such that for ε_n sufficiently small,*

$$\mu\big(C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)}\big) \leq \bar{D}^{(i)}\varepsilon_n.$$

Proof. Under the conditions on the gradient and ε_n small enough, one has $C_-^{(i)}(2\varepsilon_n) \subseteq C^{(i)} + C_1\varepsilon_n B(0,1)$ for some constant C_1. Here " $+$ " denotes set addition, that is for sets A and B we define $A + B = \{a + b \mid a \in A, b \in B\}$. In the case $d = 1$, due to the connectedness of the clusters $\mathcal{L}_1(C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)}) \leq 2C_1\varepsilon_n$. In the case $d \geq 2$, since the boundary $\partial C^{(i)}$ is a closed smooth $(d-1)$-dimensional submanifold in $\mathbb{R}^d$ with a minimal curvature radius $\kappa^{(i)} > 0$, there exists $\gamma_1 > 0$ and a constant C_2 such that $\mathcal{L}_d(C^{(i)} + \varepsilon_n B(0,1)) \leq \mathcal{L}_d(C^{(i)}) + C_2\varepsilon_n\mathcal{L}_{d-1}(\partial C^{(i)})$ for $\varepsilon_n < \gamma_1$ (see Theorem A.14). Thus, by the additivity of the volume,

$$\begin{aligned}\mathcal{L}_d\big(C_-^{(i)}(2\varepsilon_n) \setminus C^{(i)}\big) &\leq \mathcal{L}_d\big(C^{(i)} + C_1\varepsilon_n B(0,1)\big) - \mathcal{L}_d\big(C^{(i)}\big) \\ &\leq C_1C_2\mathcal{L}_{d-1}\big(\partial C^{(i)}\big)\varepsilon_n.\end{aligned}$$

Since p is bounded, we obtain, $\mu(C_{-}^{(i)}(2\varepsilon_n) \setminus C^{(i)}) \leq C_1\, C_2\, \mathcal{L}_{d-1}(\partial C^{(i)})\, p_{\max}\, \varepsilon_n$, for ε_n small enough. Setting $\bar{D}^{(i)} = C_1\, C_2\, \mathcal{L}_{d-1}(\partial C^{(i)})\, p_{\max}$ the result follows. □

Noise-free case as special case of the noisy one. In the noise-free case, by definition all sample points belong to a cluster. That means

- we can omit the density estimation step, which was used to remove background points from the graph, and drop the event $\mathcal{D}_n$ everywhere,
- we work with $L(t)$ directly instead of $L(t-\varepsilon)$,
- we do not need to remove the small components of size smaller than δn, which was needed to get a grip on the "boundary" of $L(t-\varepsilon) \setminus L(t)$.

In particular, setting $\delta = 0$ we trivially have $\Pr((\mathcal{B}_n^{(i)})^c) = 0$ and $\Pr(\mathcal{E}_n^c) = 0$ for all $i = 1, \ldots, m$ and all $n \in \mathbb{N}$.
As a consequence, we can directly work on the graphs $G_{\text{mut}}(n,k)$ and $G_{\text{sym}}(n,k)$, respectively. Therefore, the bounds we gave in the previous sections also hold in the simpler noise-free case and can be simplified in this setting.

2.5.2 Proofs of the main theorems

Proof of Theorem 2.1. Given that we are working on the complement of the event $\mathcal{I}_n^{(i)}$ of Proposition 2.10, there are no connections in $\tilde{G}_{\text{mut}}(n,k,t-\varepsilon,\delta)$ between the subgraph containing the points of cluster $C^{(i)}$ and points from any other cluster. Moreover, by Proposition 2.5 we know that the event $\mathcal{C}_n^{(i)} = \mathcal{A}_n^{(i)} \cap \mathcal{B}_n^{(i)} \cap \mathcal{E}_n \cap \mathcal{D}_n$ implies that the subgraph of all the sample points lying in cluster $C^{(i)}$ is connected and all other sample points lying not in in the cluster $C^{(i)}$ are either discarded or connected to the subgraph containing all cluster points. That means we have identified cluster $C^{(i)}$. Collecting the bounds from Proposition 2.5 and Proposition 2.10, we obtain

$$
\begin{aligned}
&\Pr\left(\text{Cluster } C^{(i)} \text{ not roughly identified in } \tilde{G}_{\text{mut}}(n,k,t-\varepsilon,\delta)\right) \\
&\quad \leq \Pr\left((\mathcal{I}_n^{(i)})^c\right) + \Pr\left((\mathcal{C}_n^{(i)})^c\right) \\
&\quad \leq \Pr\left((\mathcal{I}_n^{(i)})^c\right) + \Pr\left((\mathcal{A}_n^{(i)})^c\right) + \Pr\left((\mathcal{B}_n^{(i)})^c\right) + \Pr(\mathcal{E}_n^c) + \Pr(\mathcal{D}_n^c) \\
&\quad \leq \exp\left(-\frac{n-1}{2}\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)\right) + 2\,\exp\left(-\frac{k-1}{4^{d+1}}\frac{t}{p_{\max}^{(i)}}\right) \\
&\qquad + 2\exp\left(-n\frac{\delta}{6}\right) + 3\Pr(\mathcal{D}_n^c).
\end{aligned}
$$

In the noise-free case the events $\mathcal{B}_n^{(i)}$, $\mathcal{E}_n$ and $\mathcal{D}_n$ can be ignored. The optimal choice for k follows by equating the exponents of the bounds for $(\mathcal{I}_n^{(i)})^c$ and $(\mathcal{A}_n^{(i)})^c$ and solving

for k. One gets for the optimal k,

$$k = (n-1)\frac{\rho^{(i)}}{2 + \frac{1}{4^d}\frac{t}{p_{\max}^{(i)}}} + 1, \text{ and a rate of } (n-1)\frac{\rho^{(i)}}{2\,4^{d+1}\frac{p_{\max}^{(i)}}{t} + 4}.$$

In the noisy case, we know that if n is sufficiently large we can select a value of ε that is small enough (ε is small and fixed) such that the condition $\sum_{j=1}^{m} \mu(C_{-}^{(j)}(2\varepsilon) \setminus C^{(j)}) < \delta/2$ holds.
According to Corollary A.9 under our conditions on p there exist constants C_1, C_2 such that $\Pr(\mathcal{D}_n^c) \leq \exp(-C_2 n h^d \varepsilon^2)$, if we estimate the density with a kernel density estimator with a bandwidth h that fulfils $h^2 \leq C_1 \varepsilon$. Inserting result into the bounds above the rate of convergence is determined by the worst exponent,

$$\min\left\{\frac{(n-1)\rho^{(i)}}{4} - \frac{k-1}{2}\,, \frac{k-1}{4^{d+1}}\frac{t}{p_{\max}^{(i)}},\ n\frac{\delta}{6},\ C_2 n h^d \varepsilon^2\right\}.$$

However, since the other bounds do not depend on k the optimal choice for k remains the same. □

Proof of Theorem 2.2. Compared to the proof for cluster identification in the mutual kNN graph in Theorem 2.1 the only part which changes is the connectivity event. Here we have to replace the bound on $\Pr((\mathcal{I}_n^{(i)})^c)$ by the bound on $\Pr((\hat{\mathcal{I}}_n^{(i)})^c)$ from Proposition 2.10. With $\rho_{\min} = \min_{i=1,\dots,m} \rho^{(i)}$ we obtain

$$\Pr\left((\hat{\mathcal{I}}_n^{(i)})^c\right) \leq m\,\exp\left(-\frac{n-1}{2}\left(\frac{\rho_{\min}}{2} - \frac{k-1}{n-1}\right)\right) + \Pr(\mathcal{D}_n^c).$$

Following the same procedure as in the proof of Theorem 2.1 provides the result (for both, the noise-free and the noisy case). □

Proof of Theorem 2.3. We set $\mathcal{C}_n^{\text{all}} = \bigcap_{i=1}^{m} \mathcal{C}_n^{(i)}$ and $\mathcal{I}_n^{\text{all}} = \bigcap_{i=1}^{m} \mathcal{I}_n^{(i)}$. By a slight modification of the proof of Proposition 2.5 and $p_{\max} = \max_{i=1,\dots,m} p_{\max}^{(i)}$

$$\begin{aligned}\Pr\left((\mathcal{C}_n^{\text{all}})^c\right) &\leq 2\sum_{i=1}^{m}\exp\left(-\frac{k-1}{4^{d+1}}\frac{t}{p_{\max}^{(i)}}\right) + 2\exp\left(-n\frac{\delta}{6}\right) + 2\Pr(\mathcal{D}_n^c)\\ &\leq 2\,m\,\exp\left(-\frac{k-1}{4^{d+1}}\frac{t}{p_{\max}}\right) + 2\exp\left(-n\frac{\delta}{6}\right) + 2\Pr(\mathcal{D}_n^c).\end{aligned}$$

By a slight modification of the proof of Proposition 2.10 with $\rho_{\min} = \min_{i=1,\dots,m} \rho^{(i)}$,

$$\begin{aligned}\Pr\left((\mathcal{I}_n^{\text{all}})^c\right) &\leq \sum_{i=1}^{m}\exp\left(-\frac{n-1}{2}\left(\frac{\rho^{(i)}}{2} - \frac{k-1}{n-1}\right)\right) + \Pr(\mathcal{D}_n^c)\\ &\leq m\,\exp\left(-\frac{n-1}{2}\left(\frac{\rho_{\min}}{2} - \frac{k-1}{n-1}\right)\right) + \Pr(\mathcal{D}_n^c).\end{aligned}$$

Combining these results we obtain

$$\begin{aligned}
&\Pr\left(\text{Not all Clusters } C^{(i)} \text{ roughly identified in } \tilde{G}_{\text{mut}}\left(n,k,t-\varepsilon,\delta\right)\right) \\
&\quad\leq m\,\exp\left(-\frac{n-1}{2}\left(\frac{\rho_{\min}}{2}-\frac{k-1}{n-1}\right)\right)+3\Pr(\mathcal{D}_n^c) \\
&\qquad+2\,m\,\exp\left(-\frac{k-1}{4^{d+1}}\frac{t}{p_{\max}}\right)+2\exp\left(-n\frac{\delta}{6}\right).
\end{aligned}$$

The result follows with a similar argumentation to the proof of Theorem 2.1. □

Proof of Theorem 2.4. According to Corollary A.9 we have $\sum_{n=1}^{\infty}\Pr(\mathcal{D}_n^c)<\infty$. Moreover, let $\mathcal{C}_n^{\text{all}}$ denote the event that the points of each cluster form a connected component of the graph. Then it can be easily checked with Proposition 2.12 that we have $\sum_{n=1}^{\infty}\Pr(N_{\text{NoCluster}}/N_{\text{Cluster}} > 4\bar{D}\varepsilon_n/\beta \mid \mathcal{C}_n^{\text{all}})<\infty$. Moreover, similar to the proof of Theorem 2.3, one can show that there are constants $c_1, c_2 > 0$ such that for $c_1\log n \leq k \leq c_2 n$ cluster $C^{(i)}$ will be roughly identified almost surely as $n\to\infty$. (Note here that the bounds on k for which our probability bounds hold are also logarithmic and linear, respectively, in n). Thus, the event $\mathcal{C}_n^{\text{all}}$ occurs almost surely and consequently $N_{\text{NoCluster}}/N_{\text{Cluster}}\to 0$ almost surely. □

2.6 Discussion

In this chapter we studied the problem of cluster identification in kNN graphs. As opposed to earlier work (Brito et al. [15], Biau et al. [10]) which was only concerned with establishing connectivity results for a certain choice of k (respectively r in case of an r-neighborhood graph), our goal was to determine for which value of k the probability of cluster identification is maximized. Our work goes considerably beyond Brito et al. [15] and Biau et al. [10], concerning both the results and the proof techniques. In the noise-free case we come to the surprising conclusion that the optimal k is rather linear in n than of the order of $\log n$ as many people had suspected, both for mutual and symmetric kNN graphs. A similar result also holds for rough cluster identification in the noisy case. Both results were quite surprising to us — our first naive expectation based on the standard random geometric graph literature had been that $k\sim\log n$ would be optimal. In hindsight, our results perfectly make sense. The minimal k to achieve within-cluster connectedness is indeed of the order $\log n$. However, clusters can be more easily identified the tighter they are connected. In an extreme case where clusters have a very large distance to each other, increasing k only increases the within-cluster connectedness. Only when the cluster is fully connected (that is, k coincides with the number of points in the cluster, that is k is a positive fraction of n), connections to other clusters start to arise. Then the cluster will not be identified any more. Of course, the standard situation will not be as extreme as this one, but our proofs show that the tendency is the same.

While our results on the optimal choice of k are appealing in theory, in practical application they are often hard to realize. The higher the constant k in the kNN graph is chosen, the less sparse the neighborhood graph becomes, and the more resources we need to compute the kNN graph and to run algorithms on it. This means that one has to face a trade-off: even if in many applications it is impossible to choose k linear in n for computational restrictions, one should attempt to choose k as large as one can afford, in order to obtain the most reliable clustering results.

When comparing the symmetric and the mutual kNN graph, in terms of the within-cluster connectedness both graphs behave similarly. But note that this might be an artifact of our proof techniques: the better connectivity properties of the symmetric kNN graph are not reflected in our within-cluster connectedness results. Concerning the between-cluster disconnectedness, however, the difference between the graph types is reflected in our results: To ensure disconnectedness of one cluster $C^{(i)}$ from the other clusters in the mutual kNN graph, it suffices to make sure that the nearest neighbor of all points of $C^{(i)}$ are again elements of $C^{(i)}$. In this sense, the between-cluster disconnectedness of an individual cluster in the mutual graph can be expressed in terms of properties of this cluster only. In the symmetric kNN graph this is different. Here some other cluster $C^{(j)}$ can link inside $C^{(i)}$, no matter how nicely connected $C^{(i)}$ is. In particular, this affects the setting where the goal is to identify the most significant cluster only. While this is simple in the mutual kNN graph, in the symmetric kNN graph it is not easier than identifying all clusters given that the between-cluster disconnectedness is governed by the worst case.

From a technical point of view there are some aspects about our work which could be improved. First, we believe that the geometry of the clusters does not influence our bounds in a satisfactory manner. The main geometric quantities which enter our bounds are simple things such as the distance of the clusters to each other, the minimal and maximal density on the cluster, and so forth. However, intuitively it seems plausible that cluster identification depends on other quantities as well, such as the shapes of the clusters and the relation of those shapes to each other. For example, we would expect cluster identification to be more difficult if the clusters are in the form of concentric rings than if they are rings with different centers aligned next to each other. Currently we cannot deal with such differences. Second, the covering techniques we use for proving our bounds are not well adapted to small sample sizes. We first cover all clusters completely by small balls, and then require that there is at least one sample point in each of those balls. This leads to the unpleasant side effect that our results are not valid for a very small sample size n. However, we did not find a way to circumvent this construction. The reason is that as soon as one has to prove connectedness of a small sample of cluster points, one would have to explicitly construct a path connecting each two points. While some techniques from percolation theory might be used for this purpose in the two-dimensional setting, we did not see any way to solve this problem in high-dimensional spaces.

In the current chapter, we worked primarily with a cluster definition widely used in the statistics community, namely the high-density cluster definition. In practice, most people try to avoid performing clustering by first applying density estimation – density

estimation is inherently difficult on small samples, in particular in high-dimensional spaces. On the other hand, we have already explained earlier that this inherent complexity of the problem can also pay off. In the end, not only have we detected where the clusters are, but we also know where the data only consists of background noise.
In this chapter we only considered simple "yes/no" events (such as "cluster is connected" or "clusters are not connected to each other") to determine if we have a valid clustering or not. However, graph clustering is often solved via partitioning algorithms that (attempt to) optimize a graph clustering quality measure that favors balanced cuts such as spectral clustering, variants of which are based on the normalized cut Ncut. The next logical step would be to extend our analysis to this partitioning setting, which is technically more challenging: instead of "yes/no" events one has to carefully "count" how many edges one has in different areas of the graph. In Chapter 3 we go one step towards the analysis of graph clustering quality measures in a graph-based clustering setting, although we have not proved results on the optimal choice of k yet.

3 Influence of graph construction on graph-based clustering quality measures

3.1 Introduction

In the graph-based clustering of points from a density in Euclidean space we use graph clustering methods to cluster the neighborhood graph on the points. There are different neighborhood graphs and a neighborhood parameter has to be chosen, so it is an interesting question whether and how the results of graph-based clustering algorithms are affected by the neighborhood graph type and the parameter choice.
One way to study this question is to investigate the behavior of a graph-based clustering algorithms for more and more points: Suppose we apply it on any finite sample of data points from the density. Do the clusterings obtained by this procedure in the limit of infinitely many points correspond to a reasonable partition of the space? If so, to which one? Furthermore, do the partitions differ depending on which neighborhood graph we choose?
Unfortunately, it is hard to make statements about the asymptotic behavior of the solution of most graph-based clustering algorithms. However, many such algorithms are based on a graph clustering quality measure. So we examine a simpler but closely related problem in this chapter: Instead of the asymptotic behavior of clusterings obtained by an algorithm we examine that of graph clustering quality measures: A partition of the space is fixed and on each finite sample from the density a neighborhood graph is constructed. The partition of the space induces a clustering of the neighborhood graph. We examine the limits of the clustering quality measure of this clustering as the sample size tends to infinity and compare the limits for different neighborhood graph types: the directed r-neighborhood and the directed k-nearest neighbor graphs. The graph clustering quality measures we consider are the normalized cut Ncut and the RatioCut, which are used in the derivation of spectral clustering.
To our own surprise, when studying this convergence it turns out that, depending on the type of neighborhood graph, the normalized cut converges to different limit values. That is, the (suitably normalized) values of Ncut tend to different limit functionals, depending on whether we use the r-neighborhood graph or the kNN graph on the finite sample. Intuitively, what happens is as follows: On any given *graph*, the normalized cut is one unique, well-defined mathematical expression. But of course, given a fixed partition of a sample of *points*, this Ncut value is different for different graphs constructed on the sample (different graph constructions put different numbers of edges between points, which leads to different Ncut values). It can now be shown that even after appropriate rescaling, such differences remain visible in the limit for the sample

size tending to infinity. For example, we will see that depending on the type of graph, the limit functional integrates over different powers of the density. This can lead to the effect that the minimizer of Ncut on the kNN graph is different from the minimizer of Ncut on the r-graph.
This means that ultimately, the question about the "best Ncut" clustering, given infinite amount of data, has different answers, depending on which underlying graph we use. This observation opens Pandora's box on clustering criteria: the "meaning" of a clustering criterion does not only depend on the exact definition of the criterion itself, but also on how the graph on the finite sample is constructed. In the case of Ncut this means that Ncut is not just "one well-defined criterion", but it corresponds to a whole bunch of criteria, which differ depending on the underlying graph. More sloppy: Ncut on a kNN graph does something different than Ncut on an r-neighborhood graph. Similar results hold for the RatioCut graph clustering quality measure.
In Section 3.2 we give the formal definitions of the clustering quality measures we analyze and state the assumptions of our setting. Section 3.3 is devoted to the statement (without proofs) of our our results. We investigate how and under which conditions the Ncut criterion converges on the different graphs, and what the corresponding limit expressions are. In Section 3.4 we show experimentally that these findings are not only of theoretical interest, but that they also influence concrete algorithms such as spectral clustering in practice. We give examples of well-clustered distributions (mixtures of Gaussians), where the optimal limit cut on the kNN graph is different from the one on the r-neighborhood graph. Moreover, these results can be reproduced with finite samples. That is, given a finite sample from some well-clustered distribution, normalized spectral clustering on the kNN graph produces systematically different results from spectral clustering on the r-neighborhood graph. In Section 3.6 we give the full technical proofs of our results.
The convergence results of this chapter (without the convergence rates) were published in the paper Maier et al. [58], which received an Outstanding Student Paper Award at the Neural Information Processing Systems (NIPS) Conference in 2008.

3.2 Definitions and assumptions

Given a graph $G = (V, E)$ with weights $w : E \to \mathbb{R}$ and a partition of the nodes V into $(U, V \setminus U)$ we define

$$\mathrm{cut}(U, V \setminus U) = \sum_{u \in U, v \in V \setminus U} \big(w(u,v) + w(v,u)\big),$$

$\mathrm{vol}(U) = \sum_{u \in U, v \in V} w(u,v)$, $\mathrm{card}(U) = |U|$,

$$\mathrm{Ncut}(U, V \setminus U) = \mathrm{cut}(U, V \setminus U)\left(\frac{1}{\mathrm{vol}(U)} + \frac{1}{\mathrm{vol}(V \setminus U)}\right), \tag{3.1}$$

and

$$\text{RatioCut}(U, V \setminus U) = \text{cut}(U, V \setminus U)\left(\frac{1}{\text{card}(U)} + \frac{1}{\text{card}(V \setminus U)}\right). \tag{3.2}$$

We consider the setting described in Section 1.4 and the two types of directed neighborhood graphs defined there: the directed r-neighborhood graph $G_{\text{r}}(n, r)$ and the directed k-nearest neighbor graph $G_{\text{kNN}}(n, k)$. The density p has to fulfil some technical conditions that are defined later.
On the space $\mathbb{R}^d$ we want to study partitions which are induced by some hypersurface S. Given a surface S which separates the space $\mathbb{R}^d$ into two non-empty parts H^+ and H^-, each of which contains positive probability mass, we denote by $\text{cut}_{n,r}(S)$ the number of edges in $G_{\text{r}}(n, r)$ that go from a sample point on one side of the surface to a sample point on the other side of the surface. The corresponding quantity for the directed k-nearest neighbor graph is denoted by $\text{cut}_{n,k}(S)$. For a set $A \subseteq \mathbb{R}^d$ the volume of $\{x_1, \ldots, x_n\} \cap A$ in the graph $G_{\text{r}}(n, r)$ is denoted by $\text{vol}_{n,r}(A)$, and correspondingly $\text{vol}_{n,k}(A)$ in the graph $G_{\text{kNN}}(n, k)$. We further define $\text{card}_n(A)$ to be the number of sample points in A (note that this quantity does not depend on the graph type). Accordingly we define $\text{Ncut}_{n,r}(S)$, $\text{Ncut}_{n,k}(S)$, $\text{RatioCut}_{n,r}(S)$ and $\text{RatioCut}_{n,k}(S)$.
While the setting introduced so far is very general, we make some substantial simplifications in this thesis. First, we consider all graphs as unweighted graphs (the proofs are already technical enough in this setting). We would expect that weights on the edges might lead yet to other limit expressions. Moreover, we consider directed graphs for simplicity. In general, we want to study the setting where one wants to find two clusters which are induced by some hypersurface in $\mathbb{R}^d$. Here we only consider the case where S is a hyperplane. See also the discussion in Section 3.5 on future work that might extend our setting to more general cases.
In the rest of this chapter we assume that the following technical assumptions hold:
General assumptions in the whole chapter:

- *The data points $x_1, ..., x_n$ are drawn independently from some density p on $\mathbb{R}^d$. The measure on $\mathbb{R}^d$ that is induced by p is denoted by μ; that means, for a measurable set $A \subseteq \mathbb{R}^d$ we set $\mu(A) = \int_A p(x)\, \mathrm{d}x$.*
- *The density p is bounded from below and above, that is $0 < p_{\min} \leq p(x) \leq p_{\max}$. In particular, it has compact support C.*
- *In the interior of C, the density p is twice differentiable and $\|\nabla p(x)\| \leq p'_{\max}$ for a $p'_{\max} \in \mathbb{R}$ and all x in the interior of C.*
- *The boundary ∂C of C is a set of Lebesgue measure 0. Furthermore, we can find constants $\gamma > 0$ and $r_\gamma > 0$ such that for all $r \leq r_\gamma$ we have $\mathcal{L}_d(B(x, r) \cap C) \geq \gamma \mathcal{L}_d(B(x, r))$ for all $x \in C$.*
- *The hyperplane S splits $\mathbb{R}^d$ into two halfspaces H^+ and H^- (both including the hyperplane S) with positive probability masses, that is $\mu(H^+) > 0$, $\mu(H^-) > 0$. The normal of S pointing towards H^+ is denoted by n_S. Furthermore, S intersects the interior of C.*

- *The $(d-1)$-dimensional Lebesgue measure of $S \cap \partial C$ is zero, that is $\mathcal{L}_{d-1}(S \cap \partial C) = 0$.*

For the statements giving convergence rates we need the following conditions in dimension $d \geq 2$, which we will refer to later as "rate conditions".
Rate conditions:

- *The boundary ∂C is a compact, smooth $(d-1)$-dimensional surface with minimal curvature radius $\kappa > 0$ and denote by n_x the normal to the surface at the point $x \in \partial C$.*
- *We can find an angle $\alpha \in (0, \pi/2)$ such that $|\langle n_S, n_x \rangle| \leq \cos\alpha$ for all $x \in S \cap \partial C$.*

3.3 Limits of quality measures

In this section we study the asymptotic behavior of the aforementioned quantities for both the unweighted directed kNN graph and the unweighted r-graph. We state simple convergence results under our general assumptions as well as results on the optimal convergence rates under the rate assumptions. Here, "optimal" means the best trade-off between our bounds for different quantities. Note that it might be possible to proof faster convergence rates using a more refined proof method. Detailed proofs can be found in Section 3.6. Let $(k_n)_{n\in\mathbb{N}}$ be an increasing sequence in $\mathbb{N}$. Given a finite sample $x_1, ..., x_n$ from the underlying distribution, we will construct the graph $G_{\mathrm{kNN}}(n, k_n)$ and study the convergence of $\mathrm{Ncut}_{n,k_n}(S)$, the Ncut value induced by S evaluated on the graph $G_{\mathrm{kNN}}(n, k_n)$. Similarly, given a sequence $(r_n)_{n\in\mathbb{N}}$ in $\mathbb{R}$ of radii, we consider the convergence of Ncut_{n,r_n} induced by S on the graph $G_{\mathrm{r}}(n, r_n)$. In the following $\int_S \mathrm{d}s$ denotes the $(d-1)$-dimensional Lebesgue integral in the affine subspace S. Here is our main result for the Ncut clustering quality measure:

Theorem 3.1 (Limit values of Ncut **on different graphs)** *Assume that the general assumptions hold.* **For the** kNN **graph**, *assume that $(k_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ is a sequence in $\mathbb{N}$ with $k_n/n \to 0$. In the case $d = 1$, assume that $k_n/\sqrt{n \log n} \to \infty$, in the case $d \geq 2$ assume $k_n/\log n \to \infty$. Setting*

$$\mathrm{NcutLim}_{kNN} = \frac{2\eta_{d-1}}{(d+1)\eta_d^{1+1/d}} \int_S p^{1-1/d}(s)\,\mathrm{d}s \left(\Big(\int_{H^+} p(x)\,\mathrm{d}x \Big)^{-1} + \Big(\int_{H^-} p(x)\,\mathrm{d}x \Big)^{-1} \right)$$

we have for $n \to \infty$

$$\sqrt[d]{\frac{n}{k_n}}\,\mathrm{Ncut}_{n,k_n}(S) \quad \xrightarrow{a.s.} \quad \mathrm{NcutLim}_{kNN}\,.$$

Let, furthermore, the rate conditions hold. Then the optimal convergence rate is achieved for $k_n = k_0 n^{2/(d+2)} (\log n)^{d/(d+2)}$ if $d \geq 2$ and $k_n = k_0 \sqrt[4]{n^3 \log n}$ if $d = 1$ for suitable constants k_0. For this choice of (k_n) almost surely

$$\left| \sqrt[d]{\frac{n}{k_n}}\,\mathrm{Ncut}_{n,k_n}(S) - \mathrm{NcutLim}_{kNN} \right| = \begin{cases} O\left(\sqrt[d+2]{\frac{\log n}{n}} \right) & \text{if } d \geq 2 \\ O\left(\sqrt[4]{\frac{\log n}{n}} \right) & \text{if } d = 1. \end{cases}$$

For the r-neighborhood graph *let $(r_n)_{n\in\mathbb{N}} \subset \mathbb{R}_{>0}$. Assume $r_n \to 0$ and $nr_n^{d+1}/\log n \to \infty$ for $n \to \infty$. Setting*

$$\mathrm{NcutLim}_r = \frac{2\eta_{d-1}}{(d+1)\eta_d} \int_S p^2(s)\,\mathrm{d}s \left(\Big(\int_{H^+} p^2(x)\,\mathrm{d}x \Big)^{-1} + \Big(\int_{H^-} p^2(x)\,\mathrm{d}x \Big)^{-1} \right)$$

we have for $n \to \infty$

$$\frac{1}{r_n} \mathrm{Ncut}_{n,r_n}(S) \quad \xrightarrow{a.s.} \quad \mathrm{NcutLim}_r\,.$$

If, furthermore, the rate conditions hold, the optimal convergence rate is achieved for $r_n = r_0 \sqrt[d+3]{\log n/n}$ for a suitable constant $r_0 > 0$. For this choice of (r_n) almost surely

$$\left| \frac{1}{r_n} \mathrm{Ncut}_{n,r_n}(S) - \mathrm{NcutLim}_r \right| \;=\; O\left(\sqrt[d+3]{\frac{\log n}{n}} \right).$$

The following theorem is our main result for the RatioCut graph clustering quality measure:

Theorem 3.2 (Limit values of RatioCut **on different graphs)** *Assume the general assumptions hold.* **For the** kNN **graph**, *assume that $(k_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ is a sequence in $\mathbb{N}$ with $k_n/n \to 0$. In the case $d = 1$, assume that $k_n/\sqrt{n \log n} \to \infty$, in the case $d \geq 2$ assume $k_n/\log n \to \infty$. Setting*

$$\mathrm{RatioCutLim}_{kNN} = \frac{2\eta_{d-1}}{(d+1)\eta_d^{1+1/d}} \int_S p^{1-1/d}(s)\,\mathrm{d}s \left(\Big(\int_{H^+} p(x)\,\mathrm{d}x \Big)^{-1} + \Big(\int_{H^-} p(x)\,\mathrm{d}x \Big)^{-1} \right)$$

we have for $n \to \infty$

$$\frac{1}{k_n} \sqrt[d]{\frac{n}{k_n}}\, \mathrm{RatioCut}_{n,k_n}(S) \quad \xrightarrow{a.s.} \quad \mathrm{RatioCutLim}_{kNN}\,.$$

Let, furthermore, the rate conditions hold. Then the optimal convergence rate is achieved for $k_n = k_0 n^{2/(d+2)} (\log n)^{d/(d+2)}$ if $d \geq 2$ and $k_n = k_0 \sqrt[4]{n^3 \log n}$ if $d = 1$ for suitable constants k_0. For this choice of (k_n) almost surely

$$\left| \frac{1}{k_n} \sqrt[d]{\frac{n}{k_n}}\, \mathrm{RatioCut}_{n,k_n}(S) - \mathrm{RatioCutLim}_{kNN} \right| \;=\; \begin{cases} O\left(\sqrt[d+2]{\frac{\log n}{n}} \right) & \text{if } d \geq 2 \\ O\left(\sqrt[4]{\frac{\log n}{n}} \right) & \text{if } d = 1. \end{cases}$$

For the r-neighborhood graph, *assume $r_n > 0$, $r_n \to 0$ and $nr_n^{d+1}/\log n \to \infty$ for $n \to \infty$. Setting*

$$\mathrm{RatioCutLim}_r = \frac{2\eta_{d-1}}{(d+1)\eta_d} \int_S p^2(s)\,\mathrm{d}s \left(\Big(\int_{H^+} p(x)\,\mathrm{d}x \Big)^{-1} + \Big(\int_{H^-} p(x)\,\mathrm{d}x \Big)^{-1} \right)$$

we have for $n \to \infty$

$$\frac{1}{nr_n^{d+1}} \mathrm{RatioCut}_{n,r_n}(S) \quad \xrightarrow{a.s.} \quad \mathrm{RatioCutLim}_r\,.$$

If, furthermore, the rate conditions hold the optimal convergence rate is achieved for $r_n = r_0 \sqrt[d+3]{\log n / n}$ *for a suitable constant* $r_0 > 0$*. For this choice of* (r_n) *almost surely*

$$\left| \frac{1}{n r_n^{d+1}} \operatorname{RatioCut}_{n,r_n}(S) - \operatorname{RatioCutLim}_r \right| \;=\; O\left(\sqrt[d+3]{\frac{\log n}{n}} \right).$$

The proof of these main theorems is based on propositions about the convergence of the suitably normalized random variables $\mathrm{cut}_{n,k}$, $\mathrm{vol}_{n,k}$, card_n and the corresponding random variables for the r-neighborhood graphs.
Since there are a lot of different graph clustering quality measures that are based on these or similar random variables, we state the convergence results explicitly. For the cut and the volume random variables there are two propositions: One stating the (probabilistic) convergence of the random variable to its expectation and the other one stating the (deterministic) convergence of the expectation to some constant. In the case of the cardinality we treat both convergences in one proposition.
In the following the absolute value of the difference between the value of a random variable and its expectation is also called the *variance term*, whereas the difference between the expectation and the limit of the expectation is called the *bias term*.

Proposition 3.3 (Limit values of $\mathbb{E}\,\mathrm{cut}_{n,k_n}$ and $\mathbb{E}\,\mathrm{cut}_{n,r_n}$) *Let the general assumptions hold.*
For the kNN **graph**, *define*

$$\operatorname{CutLim}_{kNN} = \frac{2\eta_{d-1}}{d+1} \eta_d^{-1-1/d} \int_S p^{1-1/d}(s)\, \mathrm{d}s.$$

If $k_n / n \to 0$ *and* $k_n / \log n \to \infty$ *for* $n \to \infty$*, then*

$$\mathbb{E}\left(\frac{1}{n k_n} \sqrt[d]{\frac{n}{k_n}}\, \mathrm{cut}_{n,k_n}(S) \right) \;\to\; \operatorname{CutLim}_{kNN}.$$

Let, furthermore, the rate conditions hold. Then

$$\left| \mathbb{E}\left(\frac{1}{n k_n} \sqrt[d]{\frac{n}{k_n}}\, \mathrm{cut}_{n,k_n}(S) \right) - \operatorname{CutLim}_{kNN} \right| = O\left(\sqrt{\frac{\log n}{k_n}} + \sqrt[d]{\frac{k_n}{n}} \right).$$

Therefore, the optimal convergence rate is

$$\left| \mathbb{E}\left(\frac{1}{n k_n} \sqrt[d]{\frac{n}{k_n}}\, \mathrm{cut}_{n,k_n}(S) \right) - \operatorname{CutLim}_{kNN} \right| = O\left(\sqrt[d+2]{\frac{\log n}{n}} \right).$$

which is achieved for $k_n = k_0 n^{2/(d+2)} (\log n)^{d/(d+2)}$ *and any* $k_0 > 0$*.*
For the r**-neighborhood graph**, *define*

$$\operatorname{CutLim}_r = \frac{2\eta_{d-1}}{d+1} \int_S p^2(s)\, \mathrm{d}s.$$

If $r_n > 0$ and $r_n \to 0$ for $n \to \infty$, then

$$\mathbb{E}\left(\frac{\mathrm{cut}_{n,r_n}(S)}{n^2 r_n^{d+1}}\right) \quad \to \quad \mathrm{CutLim}_r\,.$$

If, furthermore, $nr_n \to \infty$ for $n \to \infty$ and the rate conditions hold, we have

$$\left|\mathbb{E}\left(\frac{\mathrm{cut}_{n,r_n}(S)}{n^2 r_n^{d+1}}\right) - \mathrm{CutLim}_r\right| = O\left(r_n\right).$$

Proposition 3.3 already shows one of the most important differences between the limits of the expected cut for the different graphs: For the r-graph we integrate over p^2, while we integrate over $p^{1-1/d}$ for the kNN graph. This difference comes from the fact that the kNN-radius is a random quantity, which is not the case for the deterministically chosen radius r_n in the r-graph.

Proposition 3.4 (Deviation of cut_{n,k_n} and cut_{n,r_n} from their means) *Let the general assumptions hold.* **For the** kNN **graph**, *define*

$$\mathrm{CutVar}_{kNN}(n,k_n) = \left|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S) - \mathbb{E}\left(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S)\right)\right|.$$

Then

$$\Pr\left(\mathrm{CutVar}_{kNN}(n,k_n) > \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2 n^{1-2/d}k_n^{2/d}}{(3\tau_d)^2}\right),$$

where τ_d denotes the kissing number in dimension d,that is, the number of unit hyperspheres in $\mathbb{R}^d$ which can touch a unit hypersphere without any intersections . In particular, let $k_n/n \to 0$ and assume $k_n/\sqrt{n\log n} \to \infty$ if the dimension $d = 1$ or $k_n/\log n \to \infty$ for $d \geq 2$. Then $\mathrm{CutVar}_{kNN}(n,k_n) \xrightarrow{a.s.} 0$ for $n \to \infty$.
For the r-neighborhood graph, *define*

$$\mathrm{CutVar}_r(n,r_n) = \left|\frac{1}{n^2 r_n^{d+1}}\,\mathrm{cut}_{n,r_n}(S) - \mathbb{E}\left(\frac{1}{n^2 r_n^{d+1}}\,\mathrm{cut}_{n,r_n}(S)\right)\right|.$$

Let $r_n > 0$, $r_n \to 0$ for $n \to \infty$. Then there exists a constant $c > 0$ such that for n sufficiently large and all $\varepsilon > 0$ sufficiently small

$$\Pr\left(\mathrm{CutVar}_r(n,r_n) \geq \varepsilon\right) \leq 2\exp\left(-\frac{nr_n^{d+1}\varepsilon^2}{16c}\right).$$

In particular, if $nr_n^{d+1}/\log n \to \infty$ we have $\mathrm{CutVar}_r(n,r_n) \xrightarrow{a.s.} 0$ for $n \to \infty$.

The following two propositions are used to show the convergence of the suitable normalized volume functionals in both graph types considered. Proposition 3.5 considers the bias term, whereas Proposition 3.6 considers the variance term.

Proposition 3.5 (Limit values of $\mathbb{E}\,\mathrm{vol}_{n,k_n}$ and $\mathbb{E}\,\mathrm{vol}_{n,r_n}$) *Let the general assumptions hold, and let $H = H^+$ or $H = H^-$. Then,* **for the** kNN **graph** *we have*

$$\mathbb{E}\left(\frac{1}{nk_n}\,\mathrm{vol}_{n,k_n}(H)\right) = \mu(H).$$

For the r-neighborhood graph, *we define*

$$\mathrm{VolLim}_r = \eta_d \int_H p^2(x)\,\mathrm{d}x.$$

If $r_n > 0$, $r_n \to 0$ for $n \to \infty$, we have

$$\mathbb{E}\left(\frac{1}{n^2 r_n^d}\,\mathrm{vol}_{n,r_n}(H)\right) \quad \longrightarrow \quad \mathrm{VolLim}_r\,.$$

If, furthermore, the rate conditions hold and $nr_n \to \infty$ for $n \to \infty$,

$$\left|\mathbb{E}\left(\frac{1}{n^2 r_n^d}\,\mathrm{vol}_{n,r_n}(H)\right) - \mathrm{VolLim}_r\right| = O\left(r_n\right).$$

The following proposition states bounds for the variance term of the suitable scaled volume functionals.

Proposition 3.6 (Deviation of vol_{n,k_n} and vol_{n,r_n} from their means) *Let the general assumptions hold and let $H = H^+$ or $H = H^-$.* **For the** kNN **graph** *define*

$$\mathrm{VolVar}_{kNN}(n,k_n) = \left|\frac{1}{nk_n}\,\mathrm{vol}_{n,k_n}(H) - \mathbb{E}\left(\frac{1}{nk_n}\,\mathrm{vol}_{n,k_n}(H)\right)\right|.$$

Then we have $\Pr\left(\mathrm{VolVar}_{kNN}(n,k_n) > \varepsilon\right) \leq 2\exp(-2\varepsilon^2 n)$. *In particular,* $\mathrm{VolVar}_{kNN}(n,k_n) \xrightarrow{a.s.} 0$ *for $n \to \infty$.*
For the r-neighborhood graph, *define*

$$\mathrm{VolVar}_r(n,r_n) = \left|\frac{1}{n^2 r_n^d}\,\mathrm{vol}_{n,r_n} - \mathbb{E}\left(\frac{1}{n^2 r_n^d}\,\mathrm{vol}_{n,r_n}\right)\right|.$$

If $r_n \to 0$ we have for n sufficiently large and all sufficiently small $\varepsilon > 0$

$$\Pr\left(\mathrm{VolVar}_r(n,r_n) \geq \varepsilon\right) \leq 2\exp\left(-\frac{nr_n^d\varepsilon^2}{16 p_{\max}\eta_d \mu(H)}\right).$$

In particular, if $nr_n^d / \log n \to \infty$ we have $\mathrm{VolVar}_r(n,r_n) \xrightarrow{a.s.} 0$ *for $n \to \infty$.*

Finally, we state both the bias term and the variance term of the cardinality, which is used in RatioCut. Note that the bias term is zero and therefore only stated implicitly.

Proposition 3.7 (Limit value of $\mathbb{E}\,\mathrm{card}_n$ and its deviation from the mean) *Let the general assumptions hold and let $H = H^+$ or $H = H^-$. Then we have*

$$\mathbb{E}\left(\frac{1}{n}\,\mathrm{card}_n(H)\right) = \mu(H).$$

Define

$$\mathrm{CardVar}(n) = \left|\frac{1}{n}\,\mathrm{card}_n(H) - \mathbb{E}\left(\frac{1}{n}\,\mathrm{card}_n(H)\right)\right|.$$

Then for every $\varepsilon > 0$ we have $\Pr\left(\mathrm{CardVar}(n) > \varepsilon\right) \leq 2\exp\left(-2\varepsilon^2 n\right)$. In particular, $\mathrm{CardVar}(n) \xrightarrow{a.s.} 0$ for $n \to \infty$.

Other convergence results. In the literature, we know only of one other limit result for graph cuts by Narayanan et al. [66]. The authors study the case of a fully connected graph with Gaussian weights $w_t(x_i, x_j) = 1/(4\pi t)^{d/2}\exp(-dist(x_i, x_j)^2/4t)$. Denoting the corresponding cut value by $\mathrm{cut}_{n,t}$, the authors show that if $t_n \to 0$ such that $t_n > 1/n^{1/(2d+2)}$, then

$$\frac{\sqrt{\pi}}{n\sqrt{t_n}}\,\mathrm{cut}_{n,t_n} \;\to\; \int_S p(s)\,ds \quad \text{a.s.}$$

By comparing this result to ours, we can see that it corroborates our finding: yet another graph leads to yet another limit result (for cut, as the authors did not study the Ncut criterion).

3.4 Examples where different limits of Ncut lead to different optimal cuts

In Theorem 3.1 we have seen that the kNN graph leads to a different limit functional for $\mathrm{Ncut}(S)$ than the r-neighborhood graph. Now we want to show that this difference is not a mathematical subtlety without practical relevance: If we select an optimal cut based on the limit criterion for the kNN graph we can obtain a different result than if we use the limit criterion based on the r-neighborhood graph. Moreover, this finding does not only apply to the limit cuts, but also to cuts constructed on finite samples. This shows that on finite data sets, different constructions of the graph can lead to systematic differences in the clustering results.

Consider a density p in one and two dimensions which is a Gaussian mixture distribution that is set to zero where it is below a threshold value θ and then rescaled properly. That is, if p' denotes the Gaussian mixture distribution of three isotropic multivariate Gaussians whose means only differ in the first coordinate

$$p'(x = (x_1, \ldots, x_d)) = (2\pi)^{-d/2}\sum_{i=1}^{3}\frac{\alpha_i}{\sigma_i^d}\exp\left(-\frac{1}{2\sigma_i^2}\left((x_1 - \mu_i)^2 + \sum_{j=2}^{d} x_j^2\right)\right),$$

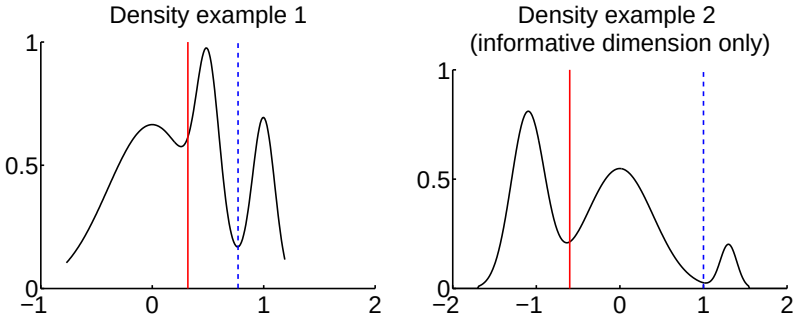

Figure 3.1: Truncated mixtures of three Gaussian densities in the examples in Section 3.4. In the two-dimensional case, we plot the informative dimension (marginal over the other dimension) only. The dashed blue vertical line depicts the optimal limit cut of the r-graph and the solid red vertical line represents the optimal limit cut of the kNN graph.

we set $A = \{x \in \mathbb{R}^d \mid p'(x) \geq \theta\}$, and obtain our thresholded density p by

$$p(x) = \frac{p'(x)\mathbb{1}_A(x)}{\int_A p'(x)\,\mathrm{d}x}.$$

We used the following specific parameters

dim	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3	α_1	α_2	α_3	θ
1	0	0.5	1	0.4	0.1	0.1	0.66	0.17	0.17	0.1
2	−1.1	0	1.3	0.2	0.4	0.1	0.4	0.55	0.05	0.01

Thresholding the density by ignoring the low-density regions is sensible for two reasons: On the one hand we assumed in our theoretical results that the density is bounded away from zero on its support. On the other hand, by sampling from a distribution with areas of very low density we obtain points that are far away from their neighbors. This can lead to isolated points in the r-neighborhood graph and to very large k-nearest neighbor radii for the kNN graph. So we see that the assumptions we made to derive our theoretical results also make sense from an experimental point of view.
In Figure 3.1 we plot the marginal densities for the first dimension. The vertical lines indicate the position of the hyperplanes perpendicular to the x_1-axis that minimize our limit expressions for Ncut (for details see below).
We initially investigate the theoretic limit Ncut values, for hyperplanes which cut perpendicular to the first dimension (which is the "informative" dimension of the data). For the chosen densities, the limit Ncut expressions from Theorem 3.1 can be computed analytically for a given hyperplane. We chose hyperplanes between $x_1 = -2$ and $x_1 = 2$ with distances of 0.1 The plots in Figure 3.2 show the theoretic limits of Ncut as dashed lines, where the horizontal axis indicates the x_1-position of the hyperplane. The vertical lines indicate the minimum of the prediction over all the hyperplanes for the respective

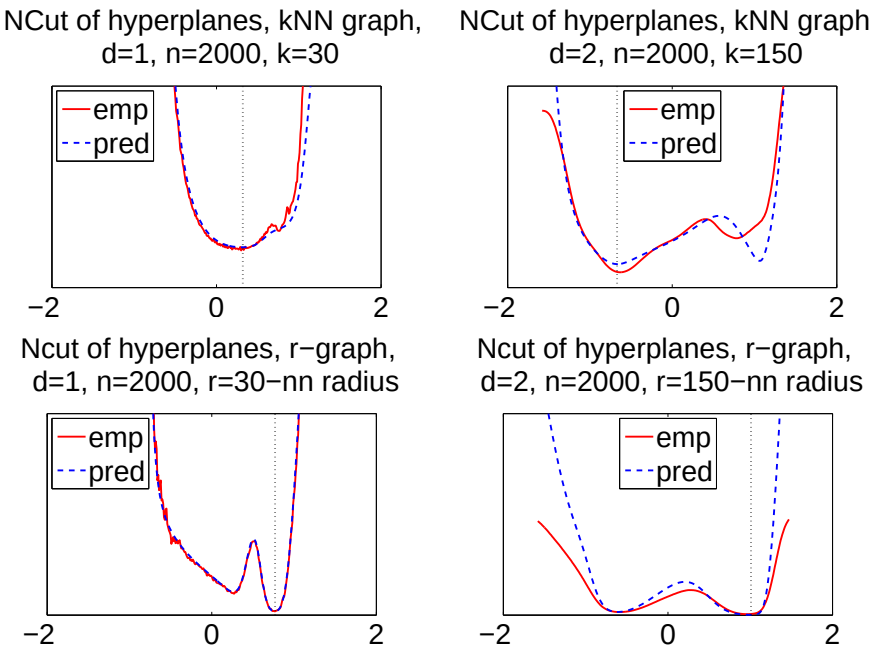

Figure 3.2: Comparison of the theoretical predictions (dashed) and empirical means (solid) for Ncut. The optimal cut is indicated by the dotted line. The top row shows the results for the kNN graph, the bottom row for the r-graph. In the left column the result for dimension 1, in the right column for dimension 2.

graph. In particular, the minimal Ncut value in the kNN case is obtained at a different position than the minimal value in the r-neighborhood case.

This effect can also be observed in a finite sample setting. We sampled $n = 2000$ points from the given distributions and constructed the unweighted symmetric kNN graph. In the results presented here we chose a parameter k for the k-nearest neighbor graph and then set r to the mean k-nearest neighbor radius. This is done to ensure that the graphs are "comparable", namely to make sure that our results are not due to the different neighborhood sizes considered, but rather because of the different graph types. The parameter k was chosen to be small but such that all the graphs we use are connected or have only few isolated points or small components.

We evaluated the empirical Ncut values for all hyperplanes which cut perpendicular to the informative dimension, as detailed in the last paragraph. This experiment was repeated 100 times. Figure 3.2 shows the means of the Ncut values of these hyperplanes, evaluated on the sample graphs. In order to compare the behavior of the empirical functional to the limit we rescaled the functional with the factor that gave the best match between the functions. We can see that the behavior of the empirical mean is very close to the behavior of the limit, that is the two local minima are in approximately the same position and the global minimum of the empirical mean coincides with the global

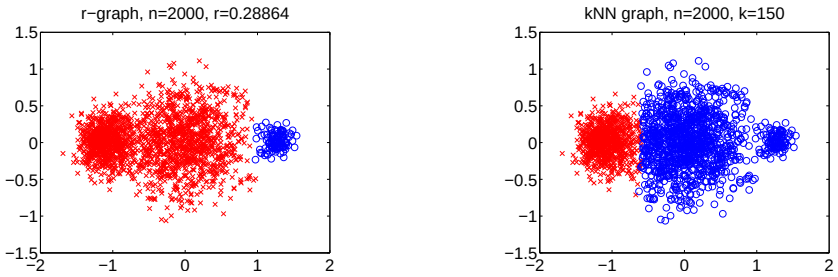

Figure 3.3: Results of spectral clustering in two dimensions, for r-graph (left) and kNN graph (right) with parameters corresponding to each other. We can see that the results differ substantially.

minimum of the limit.
Moreover, we applied normalized spectral clustering (cf. Section 1.1 and von Luxburg [85]) to the mixture data sets. As standard spectral clustering is not defined for directed graphs, we had to use an undirected kNN graph instead of the directed one: We used the symmetric kNN graph, in which two points are connected if one point is among the k nearest neighbors of the other one, or vice versa.
We tried a range of reasonable values for the parameters k and r and the results we obtained were stable over a range of parameters. Here we present the results for the 30- (for $d = 1$) and the 150-nearest neighbor graphs (for $d = 2$) and the r-graphs with corresponding parameter r, that is r was set to be the 30- and 150-nearest neighbor radius (see above).
We compare different clusterings by the minimal matching distance:

$$d_{MM}(\text{Clust}_1, \text{Clust}_2) = \frac{1}{2n} \min_{\pi} \sum_{i=1}^{n} \mathbb{1}_{\text{Clust}_1(x_i) \neq \pi(\text{Clust}_2(x_i))}$$

where the minimum is taken over all permutations π of the labels. In the case of two clusters this distance corresponds to the 0-1-loss as used in classification: a minimal matching distance of 0.39, for instance, means that 39% of the data points lie in different clusters. In our spectral clustering experiment, we could observe that the clusterings obtained by spectral clustering are usually very close to the theoretically optimal hyperplane splits predicted by theory (the minimal matching distances to the optimal hyperplane splits were always in the order of 0.03 or smaller). As predicted by theory, both types of graph give different cuts in the data. An illustration of this phenomenon for the case of dimension 2 can be found in Figure 3.3. To give a quantitative evaluation of this phenomenon, we computed the mean minimal matching distances between clusterings obtained by the same type of graph over the different samples (denoted $\hat{d}_{\text{kNN}-\text{kNN}}$ and $\hat{d}_{r-r}$), and the mean distance $\hat{d}_{\text{kNN}-r}$ between the clusterings obtained by different graph types. In order to compare clusterings of different samples from the same distribution, we have to extend a clustering of one sample to the other sample.

We do this by computing for each point the 11 nearest neighbors from the other clustering and assign the cluster of the majority of the neighbors. Of course, this is another neighborhood size that might influence our results, but we think that this effect can be neglected in the experiments we did with their well-behaved densities.
We obtained the following distances between the clusterings:

Example	$\hat{d}_{\mathrm{kNN}-\mathrm{kNN}}$	$\hat{d}_{r-r}$	$\hat{d}_{\mathrm{kNN}-r}$
1 dim	0.0005 ± 0.0006	0.0003 ± 0.0004	0.346 ± 0.063
2 dim	0.005 ± 0.0023	0.001 ± 0.001	0.49 ± 0.01

We can see that for the same graph, the clustering results are very stable (differences in the order of 10^{-3}), whereas the differences between the kNN graph and the r-neighborhood graph are substantial (0.35 and 0.49, respectively). This difference is exactly the one induced by assigning the middle mode of the density to different clusters, which is the effect predicted by theory.
It is tempting to conjecture that these effects might be due to the fact that the number of Gaussians and the number of clusters we are looking for do not coincide. Yet this is not the case: for a density in one dimension as above but with only two Gaussians with parameters

μ_1	μ_2	σ_1	σ_2	α_1	α_2	θ
0.2	0.4	0.05	0.03	0.8	0.2	0.1

the same effects can be observed. The density is depicted in the left plot of Figure 3.4. In one dimension we can compute the place of the boundary between two clusters, that is the middle between the rightmost point of the left cluster and the leftmost point of the right cluster. We did this for 100 iterations and plotted histograms of the location of the cluster boundary. In the middle and the right plot of Figure 3.4 we see that these coincide with the optimal cut predicted by theory.
Finally, we conducted an experiment similar to the last one on two real data sets (breast cancer and heart from the Data Repository by Gunnar Rätsch [25]). Here we chose the parameters $k = 20$ for both data sets, $r = 3.2$ for breast cancer and $r = 4.3$ for heart (among the parameters we tried, these were the parameters where the results were most stable, that is where $\hat{d}_{\mathrm{kNN}-\mathrm{kNN}}$ and $\hat{d}_{r-r}$ were minimal). Then we ran spectral clustering on different subsamples of the data sets ($n = 200$ for breast cancer, $n = 170$ for heart). To evaluate whether our clusterings were any useful at all, we computed the minimal matching distance between the clusterings and the true class labels and obtained distances of 0.27 for the r-graph and 0.44 for the kNN graph on breast cancer and 0.17 and 0.19 for heart. These results are reasonable (standard classifiers lead to classification errors of 0.27 and 0.17 on these data sets). Moreover, to exclude other artifacts such as different cluster sizes obtained with the kNN or r-graph, we also computed the expected random distances between clusterings, based on the actual cluster sizes we obtained in the experiments. We obtained the following table when we compared clusterings produced with the same graph

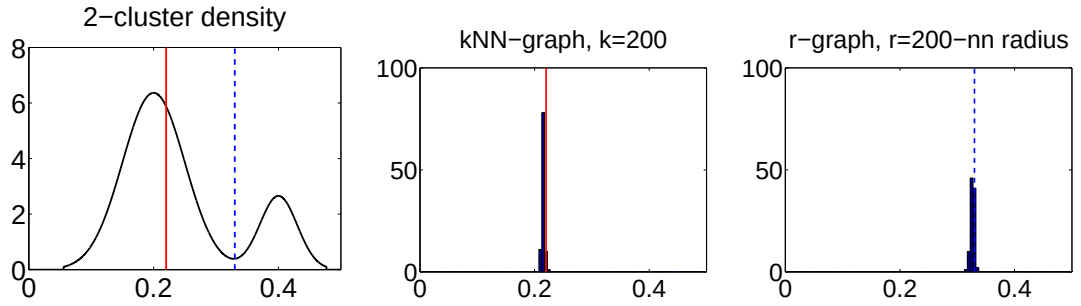

Figure 3.4: The example with the sum of two Gaussians, that is two modes of the density. In the left the figure of the density with the optimal limit cut of the r-graph (dashed blue vertical line) and the optimal limit cut of the kNN graph (the solid red vertical line). The two figures on the right show the histograms of the cluster boundary over 100 iterations for the two graph types.

Example	$\hat{d}_{\mathrm{kNN-kNN}}$	rand. $\hat{d}_{\mathrm{kNN-kNN}}$	$\hat{d}_{r-r}$	rand. $\hat{d}_{r-r}$
breast cancer	0.13 ± 0.15	0.48 ± 0.01	0.14 ± 0.10	0.22 ± 0.01
heart	0.06 ± 0.02	0.47 ± 0.02	0.06 ± 0.02	0.44 ± 0.02

and the following table when we compared clusterings produced with different graph types

Example	$\hat{d}_{\mathrm{kNN}-r}$	rand. $\hat{d}_{\mathrm{kNN}-r}$
breast cancer	0.40 ± 0.10	0.44 ± 0.01
heart	0.07 ± 0.03	0.47 ± 0.02

We can see that in the example of breast cancer, the mean distances $\hat{d}_{\mathrm{kNN-kNN}}$ and $\hat{d}_{r-r}$ are much smaller than the mean distance $\hat{d}_{\mathrm{kNN}-r}$. This shows that the clustering results differ considerably between the two types of graph (and compared to the expected random effects, this difference does not look random at all). For heart, on the other hand, we do not observe significant differences between the two graphs.
This experiment shows that for some data sets a systematic difference between the clusterings based on different graph types exists. But of course, such differences can occur for many reasons. The different limit results might just be one potential reason, and other reasons might exist. However, independent of the reason it is interesting to observe these systematic differences between graph types in real data.

3.5 Discussion

We have investigated the influence of the graph construction on graph-based clustering measures such as the normalized cut and RatioCut. We have seen that depending on the type of graph, the Ncut and RatioCut criteria converge to different limit results. We computed the exact limit expressions for the r-neighborhood graph and the kNN graph.

Moreover, another different limit result for a complete graph using Gaussian weights exists in the literature (see Narayanan et al. [66]). The fact that all these different graphs lead to different clustering criteria shows that these criteria cannot be studied isolated from the graph type they are applied to.
From a theoretical side there are several directions in which our work can be improved. We proved our results for the directed r- and kNN graphs. However, most graph-based clustering methods use the undirected graph. This is not a problem in the case of the r-graph since there the neighborhood relation is symmetric. Yet it is not clear how to prove similar results for the (symmetric or mutual) kNN graph, since in order to decide if there is an edge between two nodes in these graphs, we have to take into account the k-nearest neighbor radii of both points. However, these are stochastically dependent and therefore much harder to estimate. So it would be an interesting line of research to study if and how our results change for undirected k-nearest neighbor graphs.
Another interesting line of research would be to consider weighted graphs. For technical reasons we proved our results in this chapter only for the unweighted graphs. However, in practice weighted graphs are frequently used, since it is reasonable to give edges between points which are far away from each other less weight. This seems to be of particular concern in the case of the k-nearest neighbor graph, since very long edges can occur in regions of low density in this graph. Indeed it would be interesting to examine the influence of different weighting schemes on the limit expressions we have studied so far. The weighting scheme for the k-nearest neighbor graph that seems particularly interesting to us is the following: edges are weighted with Gaussian weights whose variance is set to the mean k-nearest neighbor radius. This graph combines the advantage of the r-neighborhood graph, that only points close to each other are connected with a high weight, with the advantage of the k-nearest neighbor graph that there are no isolated points and under certain conditions the whole graph is connected. Although it would be much more technically involved, it seems possible to adapt our convergence proofs to this weighted graph.
Another valuable extension would be to formally prove our results for other surfaces than hyperplanes. Intuitively, this should not be a problem for a sufficiently smooth surface, since in the limit we only consider local quantities and these surfaces can be approximated in the limit locally by a hyperplane. However, technically the proof would be very involved. Having these results it would not be difficult to prove uniform convergence results over a suitable class of surfaces. Here one just has to take care that a suitably restricted class of candidate surfaces S is used (note that uniform convergence results over the set of all partitions of $\mathbb{R}^d$ are impossible, cf. von Luxburg et al. [88]). Our results on the convergence rate above indicate that for uniform convergence the surface would have to fulfil many different requirements, for example regarding the probability mass on both sides of it, the volume of the intersection of the surface and the support of the density and so on.
For practice it will be important to study how the different limit results influence clustering results. So far, we do not have much intuition about when the different limit expressions lead to different optimal solutions, and when these solutions will show up in practice. The examples we provided above already show that different graphs indeed

can lead to systematically different clusterings in practice. Gaining more understanding of this effect will be an important direction of research if one wants to understand the nature of different graph clustering criteria. Assuming a density with reasonable mild assumptions, our distant goal would be to relate the clustering quality measure to (intuitively understandable) properties of a desired clustering.

3.6 Proofs

This section contains the full technical proofs of the results previously stated in Section 3.3 under the assumptions defined in Section 3.2. This section is divided into five parts: The first two parts are devoted to the proof of convergence for the bias and variance term of the scaled cut-functionals, that is the proofs of Proposition 3.3 and Proposition 3.4. In the third part we proof the corresponding convergences for the volume functionals, that is Proposition 3.5 and Proposition 3.6, and in the fourth part for the cardinality, that is Proposition 3.7. Finally, in the last part we proof the main Theorems 3.1 and 3.2.
In each of the parts we first describe the ideas of the proofs in a more intuitive way, then we state and prove the necessary lemmas, if any. Finally we present the full technical proofs of the propositions or theorems.
An overview of the structure of our proofs can be seen in Figure 3.5 on page 73.

3.6.1 Convergence of the bias term of cut_{n,r_n} and cut_{n,k_n}

In this section we prove Proposition 3.3, that is the convergence of the (suitably scaled) expectation of cut_{n,r_n} and cut_{n,k_n}.
Before we state the formal proofs, we give a proof sketch without technical details in order to identify the different steps of the proof and to motivate the technical lemmas. In the proof the probability mass in the intersection of balls with a certain radius and the other side of the hyperplane will play an important role:

Definition 3.1 *We define* $g : \mathbb{R}^d \times \mathbb{R}_{\geq 0} \to [0,1]$,

$$g(x,r) = \begin{cases} \mu(B(x,r) \cap H^+) & \text{if } x \in H^- \text{ and } x \notin S \\ \mu(B(x,r) \cap H^-) & \text{if } x \in H^+. \end{cases}$$

We also state here the properties of g we will make use of in the proofs:

- For all $x \in \mathbb{R}^d$ and $r > 0$ we clearly have $0 \leq g(x,r) \leq 1$ due to the properties of probability measures.
- $g(x,r)$ is monotonically increasing in the second argument, that is, if $r_2 > r_1$ we have $g(x,r_2) \geq g(x,r_1)$: Without loss of generality we assume $x \in H^+$. Then for $x \in H^-$

$$\begin{aligned} g(x,r_2) - g(x,r_1) &= \mu(B(x,r_2) \cap H^+) - \mu(B(x,r_1) \cap H^+) \\ &= \mu((B(x,r_2) \setminus B(x,r_1)) \cap H^+) \geq 0. \end{aligned}$$

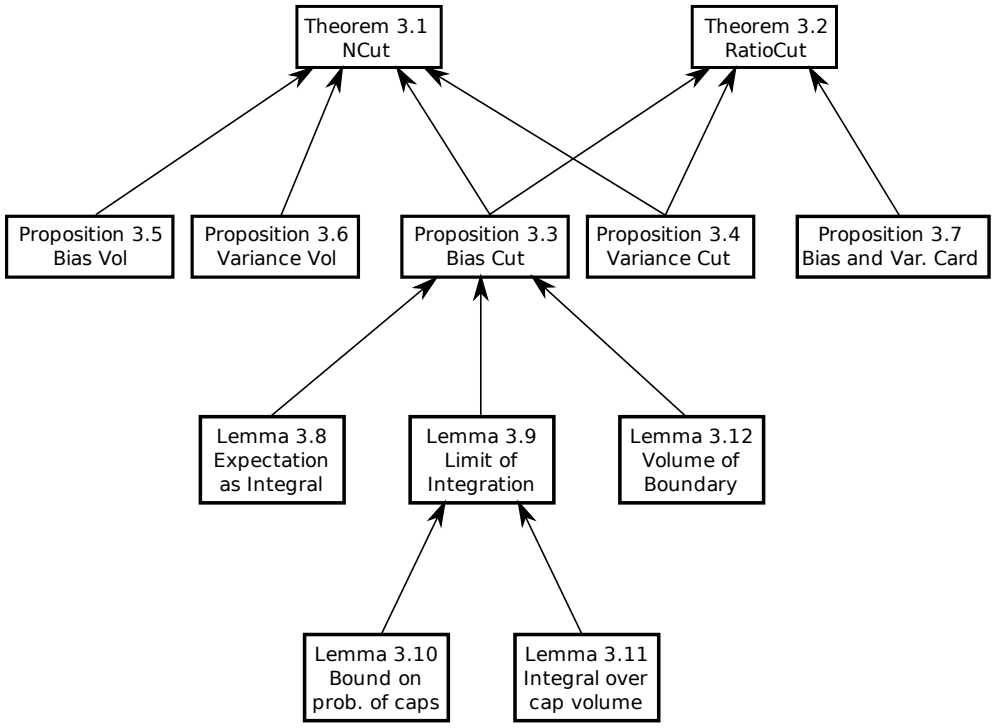

Figure 3.5: The structure of our proofs. The proofs of the two main theorems are in Section 3.6.5. The proof of the proposition concerning the convergence of the bias term for the cut, (Proposition 3.3) and all the results necessary in that proof can be found in Section 3.6.1. The proof for the convergence of the variance term of the cut, (Proposition 3.4) can be found in Section 3.6.2. The proof for the converge of bias and variance term for the volume can be found in Section 3.6.3, whereas the corresponding proof for the cardinality can be found in Section 3.6.4.

The proof sketch mainly deals with the r-neighborhood graph before we show how the proof can be adapted for the k-nearest neighbor graph.
By N_{ij} $(i, j = 1, ..., n; i \neq j)$ we denote the random variable indicating if there is an edge in the r_n-neighborhood graph from point x_i to point x_j and the points are on different sides of the cut surface S. As all points are sampled i.i.d, we have

$$\mathbb{E}\left(\mathrm{cut}_{n,r_n}(S)\right) \;=\; \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\mathbb{E}N_{ij} \;=\; n(n-1)\mathbb{E}N_{12} \;=\; n(n-1)\Pr(N_{12}=1),$$

since N_{12} is an indicator variable. Suppose the position of the first point is x. Then $N_{12} = 1$ if the second point falls in the ball $B(x, r_n)$ (then the points are connected) and it also falls on the other side of the hyperplane, which means that it falls in the intersection

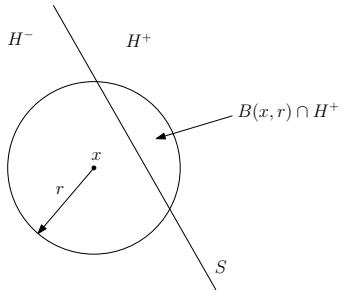

Figure 3.6: For $x \in \mathbb{R}^d$ and $r \in \mathbb{R}$ the function $g(x,r)$ specifies the probability mass in the intersection of the ball $B(x,r)$ with the halfspace that does not contain the point x.

of $B(x, r_n)$ and the other side of the hyperplane. Setting $g(x, r_n)$ as in Definition 3.1 we have $\Pr(N_{12} = 1 \mid x_1 = x) = g(x, r_n)$. Integrating this conditional expectation over all positions of the point x in $\mathbb{R}^d$ gives

$$\mathbb{E}\big(\operatorname{cut}_{n,r_n}(S)\big) \;=\; n(n-1)\int_{\mathbb{R}^d} g(x,r_n)p(x)\,\mathrm{d}x.$$

For the k-nearest neighbor graph the connectedness is harder to treat since it depends on the k-nearest neighbor radius of a point, that is, the distance of a data point to its k-th nearest neighbor, which is itself a random variable. However, we can give upper and lower bounds on the integral that use the expected k-nearest neighbor radius, rescaled by a factor close to 1. The lemma which is used in this part of the proof is Lemma 3.8. The second important idea is that instead of integrating over $\mathbb{R}^d$, we initially integrate over the hyperplane S and then, at each point $s \in S$, along the normal line through s, that is the line $s + tn_S$ for all $t \in \mathbb{R}$. This leads to

$$n(n-1)\int_{\mathbb{R}^d} g(x,r_n)p(x)\,\mathrm{d}x \;=\; n(n-1)\int_S\int_{-\infty}^{\infty} g(s+tn_S,r_n)p(s+tn_S)\,\mathrm{d}t\,\mathrm{d}s.$$

This has two advantages. First, if x is far enough from S (that is, $\operatorname{dist}(x,s) > r_n$ for all $s \in S$), then $g(x, r_n) = 0$ and the corresponding terms in the integral vanish. Second, if x is close to $s \in S$ and the radius r_n is small, then the density on the ball $B(x, r_n)$ can be considered approximately homogeneous, that is $p(y) \approx p(s)$ for all $y \in B(x, r_n)$. Thus,

$$\begin{aligned}\int_{-\infty}^{\infty} g(s+tn_S,r_n)p(s+tn_S)\,\mathrm{d}t &= \int_{-r_n}^{r_n} g(s+tn_S,r_n)p(s+tn_S)\,\mathrm{d}t\\ &\approx 2\int_0^{r_n} p(s)\mathcal{L}_d\big(B(s+tn_S,r_n)\cap H^-\big)p(s)\,\mathrm{d}t.\end{aligned}$$

It is not difficult to see that $\mathcal{L}_d\big(B(s+tn_S,r_n)\cap H^-\big) = r_n^d A(t/r_n)$, where $A(t/r_n)$ denotes the volume of the cap of the unit ball capped at distance t/r_n. Solving the integrals leads to

$$\int_0^{r_n} \mathcal{L}_d\big(B(s+tn_S,r_n)\cap H^-\big)\,\mathrm{d}t \;=\; r_n^{d+1}\int_0^1 A(t)\,\mathrm{d}t = r_n^{d+1}\frac{\eta_{d-1}}{d+1}.$$

This integration is performed in Lemma 3.9. For the r-neighborhood graph this lemma is easy to apply, since we know the radii of the balls we have to integrate over. For the k-nearest neighbor graph we use that for $k_n/n \to 0$ the expected kNN radius converges to zero. Consequently, for large n we only have to integrate over balls of approximately homogeneous density. In a region of homogeneous density $\tilde{p}$, the expected kNN radius is given as $(k_n/((n-1)\eta_d \tilde{p}))^{1/d}$ and we can easily find suitable bounds for the radius. As previously mentioned, in the proof for the k_n-nearest neighbor graph the expected k_n-nearest neighbor radius plays an important role. This is why we define a function $\tilde{r}(x,q)$, which gives us for a point $x \in \mathbb{R}^d$ the expected k_n-nearest neighbor radius if we set $q = k_n/(n-1)$.

Definition 3.2 *Under our general assumptions we define for $x \in \mathbb{R}^d$ and $q \in [0,1]$*

$$\tilde{r}(x,q) = \min\{r \in \mathbb{R}_{\geq 0} \mid \mu(B(x,r)) = q\}.$$

Remark 1 *We must show that $\tilde{r}$ is well-defined: For a fixed $x \in \mathbb{R}^d$ we have $\mu(B(x,0)) = 0$ and due to the compactness of the support of p we can find $\bar{r} \in \mathbb{R}$ such that $\mu(B(x,\bar{r})) = 1$. Now we show that the function $r \to \mu(B(x,r))$ is continuous on the interval $[0,\bar{r}]$. Let $r_1, r_2 \in [0,\bar{r}]$ and $r_1 \leq r_2$. Then*

$$\begin{aligned}\mu(B(x,r_2)) - \mu(B(x,r_1)) &= \mu(B(x,r_2) \setminus B(x,r_1)) \\ &\leq p_{\max} \mathcal{L}_d(B(x,r_2) \setminus B(x,r_1)) = p_{\max}\eta_d(r_2^d - r_1^d).\end{aligned}$$

If $r_2 \to r_1$ we have $\mu(B(x,r_2)) \to \mu(B(x,r_1))$ and therefore the function $r \to \mu(B(x,r))$ is continuous. Consequently, $\{r \mid r \in \mathbb{R}, r \geq 0, \mu(B(x,r)) = q\} \neq \emptyset$. Since the level sets of continuous functions are closed, the minimum exists.

Lemma 3.8 (Expectation of $\mathrm{cut}_{n,r}$ and $\mathrm{cut}_{n,k}$ as integral over $\mathbb{R}^d$) *Let the general assumptions hold, $g : \mathbb{R}^d \times \mathbb{R}_{\geq 0} \to [0,1]$ as in Definition 3.1 and $(r_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}_{>0}$. For the r_n-neighborhood graph we have*

$$\mathbb{E}(\mathrm{cut}_{n,r_n}(S)) = n(n-1)\int_{\mathbb{R}^d} g(x,r_n)p(x)\,\mathrm{d}x.$$

Let $\tilde{r} : \mathbb{R}^d \times [0,1] \to \mathbb{R}_{\geq 0}$ be as in Definition 3.2. Let $(k_n)_{n\in\mathbb{N}} \subseteq \mathbb{N}$ with $k_n < (n-1)/2$ and $(\delta_n)_{n\in\mathbb{N}} \subseteq (0,1/2)$ and $\delta_n k_n > 1$. Then for the k_n-nearest neighbor graph we have

$$\begin{aligned}\mathbb{E}(\mathrm{cut}_{n,k_n}(S)) &\leq n(n-1)\int_{\mathbb{R}^d} g(x,\tilde{r}(x,(1+\delta_n)\alpha_n))p(x)\,\mathrm{d}x + 2\exp(2\log n - \delta_n^2 k_n/4) \\ \mathbb{E}(\mathrm{cut}_{n,k_n}(S)) &\geq n(n-1)\int_{\mathbb{R}^d} g(x,\tilde{r}(x,(1-\delta_n)\alpha_n))p(x)\,\mathrm{d}x - 2\exp(2\log n - \delta_n^2 k_n/4),\end{aligned}$$

where $\alpha_n = k_n/(n-1)$.

Proof. First we consider properties of the expectation of the cut which are independent of the graph type. So we denote the cut by $\mathrm{cut}_n(S)$ and the neighborhood graph by

$G(n)$, such that $G(n) = G_{\mathrm{r}}(n, r_n)$ or $G(n) = G_{\mathrm{kNN}}(n, k_n)$. For $i, j \in \{1, \ldots, n\}$, $i \neq j$ we define a random variable N_{ij} with

$$N_{ij} = \begin{cases} 1 & \text{if } x_i \in H^+, x_j \in H^- \text{ or vice versa and } (x_i, x_j) \text{ edge in } G(n) \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$\mathrm{cut}_n(S) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} N_{ij},$$

and, since the points are independent and identically distributed,

$$\mathbb{E}(\mathrm{cut}_n(S)) = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbb{E}(N_{ij}) = n(n-1)\mathbb{E}(N_{12}).$$

Conditioning on the location of the points and using that $N_{ij} = 0$ if $x_1, x_2 \in H^+$ or $x_1, x_2 \in H^-$, we obtain

$$\begin{aligned} \mathbb{E}(N_{12}) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbb{E}(N_{12} \mid x_1 = x, x_2 = y) p(y) \, \mathrm{d}y \; p(x) \, \mathrm{d}x \\ &= \int_{H^+} \int_{H^-} \mathbb{E}(N_{12} \mid x_1 = x, x_2 = y) p(y) \, \mathrm{d}y \; p(x) \, \mathrm{d}x \\ &\quad + \int_{H^-} \int_{H^+} \mathbb{E}(N_{12} \mid x_1 = x, x_2 = y) p(y) \, \mathrm{d}y \; p(x) \, \mathrm{d}x, \end{aligned} \tag{3.3}$$

where we have used that the joint density of x_1 and x_2 is the product of the marginal densities since the two variables are assumed to be independent. Note that technically the conditional expectation does not exist if $x \notin C$ or $y \notin C$. However, this does not pose a problem since in this case $p(x) = 0$ or $p(y) = 0$, so we can set the conditional expectation to an arbitrary fixed number without changing the integral. In the following we set $\mathbb{E}(N_{12} \mid x_1 = x, x_2 = y) = 0$ if $x \notin C$ or $y \notin C$.
Now we will examine the inner integrals for $x \in H^+$, $y \in H^-$. The other integral can be dealt with similarly.
For the r_n-neighborhood graph we have $N_{12} = 1$ if $\mathrm{dist}(x, y) \leq r_n$ and x and y are on different sides of S, and 0 otherwise. So, for given $x \in H^+ \cap C$ and $y \in H^- \cap C$, we have

$$\mathbb{E}(N_{12} \mid x_1 = x, x_2 = y) = \begin{cases} 1 & \text{if } \mathrm{dist}(x, y) \leq r_n \\ 0 & \text{otherwise,} \end{cases}$$

and thus for a given $x \in H^+ \cap C$

$$\int_{H^-} \mathbb{E}(N_{12} \mid x_1 = x, x_2 = y) p(y) \, \mathrm{d}y = \int_{H^- \cap B(x, r_n)} p(y) \, \mathrm{d}y = g(x, r_n).$$

Applying the same procedure for the other integral, that is for $x \in H^- \cap C$, we obtain for the r_n-neighborhood graph

$$\begin{aligned}\mathbb{E}(N_{12}) &= \int_{H^+} g(x, r_n) p(x)\, \mathrm{d}x + \int_{H^-} g(x, r_n) p(x)\, \mathrm{d}x \\ &= \int_{\mathbb{R}^d} g(x, r_n) p(x)\, \mathrm{d}x. \end{aligned} \tag{3.4}$$

For the k-nearest neighbor graph, we have for $x \in H^+ \cap C$, $y \in H^- \cap C$

$$\mathbb{E}(N_{12} \mid x_1 = x, x_2 = y) = \Pr(C_{12} = 1 \mid x_1 = x, x_2 = y),$$

where C_{12} is the indicator variable of the event that there is an edge from x_1 to x_2. We have

$$\Pr(C_{12} = 1 \mid x_1 = x, x_2 = y) = \Pr(U < k_n),$$

where $U \sim \mathrm{Bin}(n-2, \mu(B(x, \mathrm{dist}(x, y))))$.
Since $0 < \delta_n < 1/2$ and $0 < \alpha_n < 1/2$ we have $(1+\delta_n)\alpha_n \in (0,1)$ and therefore $\tilde{r}(x, (1+\delta_n)\alpha_n)$ and $\tilde{r}(x, (1-\delta_n)\alpha_n)$ exists for every $x \in \mathbb{R}^d$. Now suppose $\mathrm{dist}(x, y) \leq \tilde{r}(x, (1-\delta_n)\alpha_n)$ and let $U' \sim \mathrm{Bin}(n-2, \mu(B(x, \tilde{r}(x, (1-\delta_n)\alpha_n))))$, so $U' \sim \mathrm{Bin}(n-2, (1-\delta_n)\alpha_n)$. Then, under the conditions on δ_n and k_n we can use the tail bound for the binomial distribution in Corollary A.3 and obtain

$$\begin{aligned}\Pr(U < k_n) &\geq \Pr(U' < k_n) = 1 - \Pr(U' > k_n - 1) \\ &\geq 1 - \exp\left(-\frac{1}{3}\frac{(k_n - 1 - (n-2)(1-\delta_n)\alpha_n)^2}{(n-2)(1-\delta_n)\alpha_n}\right).\end{aligned}$$

Now we clearly have (since $0 < \delta_n < 1/2$ and $n > 2$, and using $\alpha_n = k_n/(n-1)$)

$$\begin{aligned}\frac{(k_n - 1 - (n-2)(1-\delta_n)\alpha_n)^2}{(n-2)(1-\delta_n)\alpha_n} &= \frac{\left(k_n - 1 - (n-2)(1-\delta_n)\frac{k_n}{n-1}\right)^2}{(n-2)(1-\delta_n)\frac{k_n}{n-1}} \\ &\geq \frac{\left(k_n - 1 - \left(1 - \frac{1}{n-1}\right)(1-\delta_n)k_n\right)^2}{k_n} = \frac{\left(\left(\delta_n - \frac{\delta_n}{n-1} + \frac{1}{n-1}\right)k_n - 1\right)^2}{k_n} \\ &\geq \frac{\left(\delta_n - \frac{\delta_n}{n-1} + \frac{1}{n-1}\right)^2 k_n^2 - 2\left(\delta_n - \frac{\delta_n}{n-1} + \frac{1}{n-1}\right)k_n}{k_n} \\ &\geq \left(\delta_n + \frac{1-\delta_n}{n-1}\right)^2 k_n - 2\left(\delta_n + \frac{1}{n-1}\right) \geq \delta_n^2 k_n - 2,\end{aligned}$$

where in the last step we use $k_n/(n-1) \leq 1$. That is, in the case of a small distance $\mathrm{dist}(x, y) \leq \tilde{r}(x, (1-\delta_n)\alpha_n)$ we have

$$\Pr(C_{12} = 1 \mid x_1 = x, x_2 = y) \geq 1 - \exp\left(-\frac{\delta_n^2}{3}k_n + \frac{2}{3}\right). \tag{3.5}$$

Suppose $\operatorname{dist}(x,y) \geq \tilde{r}(x,(1+\delta_n)\alpha_n)$ and let $V' \sim \operatorname{Bin}(n-2, \mu(B(x,\tilde{r}(x,(1+\delta_n)\alpha_n))))$, that is $V' \sim \operatorname{Bin}(n-2,(1+\delta_n)\alpha_n)$. Then under our conditions on δ_n and k_n we can use the tail bound for the binomial distribution in Corollary A.3 and obtain

$$\Pr(U < k_n) \leq \Pr(V' < k_n) \leq \exp\Big(-\frac{1}{2}\frac{((n-2)(1+\delta_n)\alpha_n - k_n)^2}{(n-2)(1+\delta_n)\alpha_n}\Big).$$

As $0 < \delta_n < 1/2$ and $n > 2$

$$\begin{aligned}\frac{((n-2)(1+\delta_n)\alpha_n - k_n)^2}{(n-2)(1+\delta_n)\alpha_n} &= \frac{\left(\left(1-\frac{1}{n-1}\right)(1+\delta_n)k_n - k_n\right)^2}{(n-2)(1+\delta_n)\frac{k_n}{n-1}} = \frac{\left(\left(\delta_n - \frac{1+\delta_n}{n-1}\right)k_n\right)^2}{(1+\delta_n)k_n} \\ &\geq \frac{\delta_n^2}{1+\delta_n}k_n - \frac{2\delta_n\frac{1+\delta_n}{n-1}}{1+\delta_n}k_n = \frac{\delta_n^2}{1+\delta_n}k_n - 2\delta_n\frac{k_n}{n-1} \geq \frac{\delta_n^2}{2}k_n - 1.\end{aligned}$$

In the case $\operatorname{dist}(x,y) \geq \tilde{r}(x,(1+\delta_n)\alpha_n)$ we have

$$\Pr(C_{12} = 1 \mid x_1 = x, x_2 = y) \leq \exp\left(-\frac{\delta_n^2}{4}k_n + \frac{1}{2}\right). \tag{3.6}$$

We set $\mathcal{X} = \{x_1 = x, x_2 = y\}$, $B_n^- = B(x_1, \tilde{r}(x,(1-\delta_n)\alpha_n))$ and $B_n^+ = B(x,\tilde{r}(x,(1+\delta_n)\alpha_n))$. Then

$$\begin{aligned}\int_{H^-} \mathbb{E}(N_{12} \mid \mathcal{X})p(y)\,\mathrm{d}y &= \int_{H^-\cap C} \Pr(C_{12}=1 \mid \mathcal{X})p(y)\,\mathrm{d}y \\ &= \int_{H^-\cap B_n^-\cap C} \Pr(C_{12}=1 \mid \mathcal{X})p(y)\,\mathrm{d}y + \int_{H^-\cap (B_n^-)^c\cap C} \Pr(C_{12}=1 \mid \mathcal{X})p(y)\,\mathrm{d}y \\ &\geq \int_{H^-\cap B_n^-\cap C} \Pr(C_{12}=1 \mid \mathcal{X})p(y)\,\mathrm{d}y \geq \min_{y\in H^-\cap B_n^-\cap C} \Pr(C_{12}=1 \mid \mathcal{X}) \int_{H^-\cap B_n^-} p(y)\,\mathrm{d}y \\ &= g(x,\tilde{r}(x,(1-\delta_n)\alpha_n)) \min_{y\in H^-\cap B_n^-\cap C} \Pr(C_{12}=1 \mid \mathcal{X}),\end{aligned}$$

and using the result of Equation (3.5)

$$\begin{aligned}&\geq g(x,\tilde{r}(x,(1-\delta_n)\alpha_n))\left(1 - \exp\left(-\frac{\delta_n^2}{3}k_n + \frac{2}{3}\right)\right) \\ &\geq g(x,\tilde{r}(x,(1-\delta_n)\alpha_n)) - \exp\left(-\frac{\delta_n^2}{3}k_n + \frac{2}{3}\right).\end{aligned}$$

On the other hand, using the result of Equation (3.6)

$$\begin{aligned}
\int_{H^-} \mathbb{E}(N_{12} \mid \mathcal{X}) p(y)\, \mathrm{d}y &= \int_{H^- \cap C} \Pr(C_{12} = 1 \mid \mathcal{X}) p(y)\, \mathrm{d}y \\
&= \int_{H^- \cap B_n^+ \cap C} \Pr(C_{12} = 1 \mid \mathcal{X}) p(y)\, \mathrm{d}y + \int_{H^- \cap (B_n^+)^c \cap C} \Pr(C_{12} = 1 \mid \mathcal{X}) p(y)\, \mathrm{d}y \\
&\leq \int_{H^- \cap B_n^+} p(y)\, \mathrm{d}y + \max_{y \in H^- \cap (B_n^+)^c \cap C} \Pr(C_{12} = 1 \mid \mathcal{X}) \int_{H^- \cap (B_n^+)^c} p(y)\, \mathrm{d}y \\
&\leq g(x, \tilde{r}(x, (1+\delta_n)\alpha_n)) + \max_{y \in H^- \cap (B_n^+)^c \cap C} \Pr(C_{12} = 1 \mid \mathcal{X}) \\
&\leq g(x, \tilde{r}(x, (1+\delta_n)\alpha_n)) + \exp\left(-\frac{\delta_n^2 k_n}{4} + \frac{1}{2}\right).
\end{aligned}$$

The same analysis can be carried out for the other integral, that is for $x \in H^-$. Inserting these bounds for the inner integral into Equation (3.3) we obtain

$$\begin{aligned}
\mathbb{E}(N_{12}) &\geq \int_{H^+} \left(g(x, \tilde{r}(x, (1-\delta_n)\alpha_n)) - \exp\left(-\frac{\delta_n^2}{3} k_n + \frac{2}{3}\right) \right) p(x)\, \mathrm{d}x \\
&\quad + \int_{H^-} \left(g(x, \tilde{r}(x, (1-\delta_n)\alpha_n)) - \exp\left(-\frac{\delta_n^2}{3} k_n + \frac{2}{3}\right) \right) p(x)\, \mathrm{d}x \\
&= \int_{\mathbb{R}^d} g(x, \tilde{r}(x, (1-\delta_n)\alpha_n)) p(x)\, \mathrm{d}x - \exp\left(-\frac{\delta_n^2}{3} k_n + \frac{2}{3}\right) \\
&\geq \int_{\mathbb{R}^d} g(x, \tilde{r}(x, (1-\delta_n)\alpha_n)) p(x)\, \mathrm{d}x - 2\exp\left(-\frac{\delta_n^2}{4} k_n\right),
\end{aligned}$$

where we use $\exp(2/3) < 2$ and

$$\begin{aligned}
\mathbb{E}(N_{12}) &\leq \int_{\mathbb{R}^d} g(x, \tilde{r}(x, (1+\delta_n)\alpha_n))\, \mathrm{d}x + \exp\left(-\frac{\delta_n^2}{4} k_n + \frac{1}{2}\right) \\
&\leq \int_{\mathbb{R}^d} g(x, \tilde{r}(x, (1+\delta_n)\alpha_n))\, \mathrm{d}x + 2\exp\left(-\frac{\delta_n^2}{4} k_n\right).
\end{aligned}$$

□

The following lemma is a technical lemma, which essentially states that the integral over the whole space in Lemma 3.8 can be replaced by an integral over the hyperplane S under certain conditions. In the proof of Proposition 3.3 we show that the conditions in this lemma hold for the r-neighborhood graph and the k-nearest neighbor graph. We give here a more intuitive description of the conditions: The parameter $\tilde{r}_n(x)$ in the original integral must be bounded from above uniformly over all x in C by $r_n^{\max}$. Intuitively, $r_n^{\max}$ denotes an upper bound on the distance of points to which a sample point at position x can be connected, that is on the neighborhood radius. The second condition means that for a point s on the hyperplane and far enough from the boundary of C we can bound the radii $\tilde{r}_n(x)$ of points x close to s in terms of $r_n(s)$. The radius $r_n(s)$

is basically the expected neighborhood radius of a sample point at s. The third condition gives lower and upper bounds on the density close to the hyperplane. The values of ν_n and ξ_n describe how close the estimates of the neighborhood radius and the density are to the actual value. This "closeness" in the second and the third condition depends on the upper bound on the neighborhood radius $r_n^{\max}$. For the graphs under consideration we show that ξ_n can be chosen linear in $r_n^{\max}$ and ν_n is set to zero for the r-graph and linear in $r_n^{\max}$ for the kNN-graph. In fact, making stronger differentiability assumptions on the density we could improve the dependence of ν_n and ξ_n on $r_n^{\max}$. However, this would not improve the convergence rates in the end, since these are determined by ν_n, ξ_n and $\mathcal{L}_{d-1}(S \cap \mathcal{R}_n)$, where the last can be shown to be linear in $r_n^{\max}$ under the rate conditions.

Lemma 3.9 (Integral over $\mathbb{R}^d$ bounded in terms of integral over S) *Let $(r_n)_{n\in\mathbb{N}}$, $(\tilde{r}_n)_{n\in\mathbb{N}}$ be sequences of functions $r_n, \tilde{r}_n : \mathbb{R}^d \to \mathbb{R}$ for all $n \in \mathbb{N}$ and $(r_n^{\max})_{n\in\mathbb{N}}$ a sequence of reals. Suppose the following conditions hold:*

1. *$r_n(x) \leq r_n^{\max}$ and $\tilde{r}_n(x) \leq r_n^{\max}$ for all $x \in C$ and $n \in \mathbb{N}$,*
2. *we can find a sequence $(\nu_n)_{n\in\mathbb{N}} \subseteq [0,1)$ such that if $B(s, 3r_n^{\max}) \subseteq C$ for $s \in S$ then for all $x \in B(s, r_n^{\max})$ we have*
$$\sqrt[d]{1-\nu_n}\, r_n(s) \leq \tilde{r}_n(x) \leq \sqrt[d]{1+\nu_n}\, r_n(s), \tag{3.7}$$
3. *we can find a sequence $(\xi_n)_{n\in\mathbb{N}} \subseteq (0,1)$ such that if $B(s, 3r_n^{\max}) \subseteq C$ for $s \in S$ then for all $y \in B(s, 3r_n^{\max})$ we have*
$$(1-\xi_n)p(s) \leq p(y) \leq (1+\xi_n)p(s). \tag{3.8}$$

Then we have

$$\begin{aligned}\int_{\mathbb{R}^d} p(x)g(x, \tilde{r}_n(x))\,\mathrm{d}x \leq &(1+\xi_n)^2(1+\nu_n)^{1+1/d}\frac{2\eta_{d-1}}{d+1}\int_S p^2(s) r_n^{d+1}(s)\,\mathrm{d}s \\ &+ 2\eta_d p_{\max}^2 (r_n^{\max})^{d+1}\mathcal{L}_{d-1}(S\cap\mathcal{R}_n),\end{aligned}$$

and

$$\begin{aligned}\int_{\mathbb{R}^d} p(x)g(x, \tilde{r}_n(x))\,\mathrm{d}x \geq &(1-\xi_n)^2(1-\nu_n)^{1+1/d}\frac{2\eta_{d-1}}{d+1}\int_S p^2(s) r_n^{d+1}(s)\,\mathrm{d}s \\ &- 2\eta_{d-1} p_{\max}^2 (r_n^{\max})^{d+1}\mathcal{L}_{d-1}(S\cap\mathcal{R}_n),\end{aligned}$$

where we have set $\mathcal{R}_n = \{x \in \mathbb{R}^d \mid \operatorname{dist}(x, \partial C) \leq 3r_n^{\max}\}$.

Proof. We have by a translation and rotation of the coordinate system

$$\begin{aligned}\int_{\mathbb{R}^d} p(x)g(x, \tilde{r}_n(x))\,\mathrm{d}x &= \int_S \int_{-\infty}^{\infty} p(s+tn_S)g(s+tn_S, \tilde{r}_n(s+tn_S))\,\mathrm{d}t\,\mathrm{d}s &(3.9)\\ &= \int_S h_n(s)\,\mathrm{d}s, &(3.10)\end{aligned}$$

where we have set

$$h_n(s) = \int_{-\infty}^{\infty} p(s + tn_S) g(s + tn_S, \tilde{r}_n(s + tn_S)) \, \mathrm{d}t. \tag{3.11}$$

Let $\mathcal{R}_n$ be defined as above, $\mathcal{I}_n = C \setminus \mathcal{R}_n$ and $\mathcal{A}_n = \mathbb{R}^d \setminus (\mathcal{I}_n \cup \mathcal{R}_n)$. Then we can decompose the integral into

$$\int_S h_n(s) \, \mathrm{d}s = \int_{S \cap \mathcal{I}_n} h_n(s) \, \mathrm{d}s + \int_{S \cap \mathcal{R}_n} h_n(s) \, \mathrm{d}s + \int_{S \cap \mathcal{A}_n} h_n(s) \, \mathrm{d}s.$$

Let $s \in S \cap \mathcal{A}_n$. Then $\operatorname{dist}(s, C) \geq r_n^{\max}$ and thus for $|t| \leq r_n^{\max}$ we have $p(s + tn_S) = 0$. If $|t| > r_n^{\max}$ and $s + tn_S \notin C$ then $p(s + tn_S) = 0$ as well. Otherwise if $s + tn_S \in C$ we have $\tilde{r}_n(s + tn_S) \leq r_n^{\max}$ but $d(s + tn_S, S) \geq r_n^{\max}$ and thus $g(s + tn_S, \tilde{r}_n(s + tn_S)) = 0$. Therefore

$$\int_{S \cap \mathcal{A}_n} h_n(s) \, \mathrm{d}s = 0. \tag{3.12}$$

Now let $s \in S \cap \mathcal{R}_n$. We have for any $s \in S$ and $t \in \mathbb{R}$

$$p(s + tn_s) g(s + tn_S, \tilde{r}_n(s + tn_S)) \leq p_{\max} g(s + tn_S, r_n^{\max}),$$

since either $p(s + tn_S) = 0$ (for $s + tn_S \notin C$) or $p(s + tn_S) \leq p_{\max}$ and $\tilde{r}_n(s + tn_S) \leq r_n^{\max}$ (for $s + tn_S \in C$). Therefore we have for $s \in S \cap \mathcal{R}_n$

$$\begin{aligned} h_n(s) &\leq \int_{-r_n^{\max}}^{r_n^{\max}} p_{\max} g(s + tn_S, r_n^{\max}) \, \mathrm{d}t \leq \int_{-r_n^{\max}}^{r_n^{\max}} p_{\max}^2 \eta_d (r_n^{\max})^d \, \mathrm{d}t \\ &= p_{\max}^2 \eta_d (r_n^{\max})^d 2 r_n^{\max} = 2 p_{\max}^2 \eta_d (r_n^{\max})^{d+1}, \end{aligned}$$

and thus

$$\begin{aligned} \int_{S \cap \mathcal{R}_n} h_n(s) \, \mathrm{d}s &\leq 2 p_{\max}^2 \eta_d (r_n^{\max})^{d+1} \int_{S \cap \mathcal{R}_n} 1 \, \mathrm{d}s \\ &= 2 p_{\max}^2 \eta_d (r_n^{\max})^{d+1} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n). \end{aligned} \tag{3.13}$$

Finally, we consider the case $s \in S \cap \mathcal{I}_n$, that means $B(s, 3r_n^{\max}) \subseteq C$.
Since $\tilde{r}_n(x) \leq \sqrt[d]{1 + \nu_n} r(s)$ for $x \in B(s, r_n^{\max})$ and $\nu_n < 1$, we have $\tilde{r}_n(x) \leq 2 r_n^{\max}$ and therefore $B(x, \tilde{r}_n(x)) \subseteq B(s, 3r_n^{\max}) \subseteq C$ by definition. That is, we can use the definition of ξ_n in Assumption 3 and the monotonicity of g to obtain

$$\begin{aligned} h_n(s) &= \int_{-\infty}^{\infty} p(s + tn_S) g(s + tn_S, \tilde{r}_n(s + tn_S)) \, \mathrm{d}t \\ &\leq \int_{-\sqrt[d]{1+\nu_n} r_n(s)}^{\sqrt[d]{1+\nu_n} r_n(s)} (1 + \xi_n) p(s) g(s + tn_S, \sqrt[d]{1 + \nu_n} r_n(s)) \, \mathrm{d}t \\ &= (1 + \xi_n) p(s) \int_{-\sqrt[d]{1+\nu_n} r_n(s)}^{\sqrt[d]{1+\nu_n} r_n(s)} g(s + tn_S, \sqrt[d]{1 + \nu_n} r_n(s)) \, \mathrm{d}t. \end{aligned}$$

Setting $A(t) = \mathcal{L}_d(B(0,1) \cap \{x = (x^{(1)}, \ldots, x^{(d)}) | x^{(1)} \geq t\})$ and applying Lemma 3.10 we obtain

$$\begin{aligned} h_n(s) &\leq (1+\xi_n)p(s) \int_{-\sqrt[d]{1+\nu_n} r_n(s)}^{\sqrt[d]{1+\nu_n} r_n(s)} (1+\xi_n)p(s)(\sqrt[d]{1+\nu_n} r_n(s))^d A\left(\frac{|t|}{\sqrt[d]{1+\nu_n} r_n(s)}\right) \mathrm{d}t \\ &= (1+\xi_n)^2(1+\nu_n)p^2(s)r_n^d(s)2 \int_0^{\sqrt[d]{1+\nu_n} r_n(s)} A\left(\frac{t}{\sqrt[d]{1+\nu_n} r_n(s)}\right) \mathrm{d}t. \end{aligned}$$

Substituting in the integral $u = t/(\sqrt[d]{1+\nu_n} r_n(s))$ we have $\mathrm{d}t = \sqrt[d]{1+\nu_n} r_n(s)\, \mathrm{d}u$ and obtain

$$\begin{aligned} h_n(s) &\leq (1+\xi_n)^2(1+\nu_n)p^2(s)r_n^d(s)2 \int_0^1 A(u) \sqrt[d]{1+\nu_n} r_n(s)\, \mathrm{d}u \\ &= (1+\xi_n)^2(1+\nu_n)^{1+1/d}p^2(s)r_n^{d+1}(s)2 \int_0^1 A(u)\, \mathrm{d}u \\ &= (1+\xi_n)^2(1+\nu_n)^{1+1/d} \frac{2\eta_{d-1}}{d+1} p^2(s)r_n^{d+1}(s), \end{aligned}$$

where we apply Lemma 3.11 in the last step. Hence

$$\int_{S \cap \mathcal{I}_n} h_n(s)\, \mathrm{d}s \leq (1+\xi_n)^2(1+\nu_n)^{1+1/d} \frac{2\eta_{d-1}}{d+1} \int_{S \cap \mathcal{I}_n} p^2(s)r_n^{d+1}(s)\, \mathrm{d}s. \tag{3.14}$$

Similarly, we show

$$\int_{S \cap \mathcal{I}_n} h_n(s)\, \mathrm{d}s \geq (1-\xi_n)^2(1-\nu_n)^{1+1/d} \frac{2\eta_{d-1}}{d+1} \int_{S \cap \mathcal{I}_n} p^2(s)r_n^{d+1}(s)\, \mathrm{d}s.$$

With $\mathcal{R}_n$, $\mathcal{A}_n$, and $\mathcal{I}_n$ as above we certainly have

$$\begin{aligned} &\int_S p^2(s)r_n^{d+1}(s)\, \mathrm{d}s \\ &\quad = \int_{S \cap \mathcal{I}_n} p^2(s)r_n^{d+1}(s)\, \mathrm{d}s + \int_{S \cap \mathcal{R}_n} p^2(s)r_n^{d+1}(s)\, \mathrm{d}s + \int_{S \cap \mathcal{A}_n} p^2(s)r_n^{d+1}(s)\, \mathrm{d}s \\ &\quad = \int_{S \cap \mathcal{I}_n} p^2(s)r_n^{d+1}(s)\, \mathrm{d}s + \int_{S \cap \mathcal{R}_n} p^2(s)r_n^{d+1}(s)\, \mathrm{d}s. \end{aligned}$$

Therefore,

$$\begin{aligned} &\int_{S \cap \mathcal{I}_n} p^2(s)r_n^{d+1}(s)\, \mathrm{d}s \\ &\qquad = \int_S p^2(s)r_n^{d+1}(s)\, \mathrm{d}s - \int_{S \cap \mathcal{R}_n} p^2(s)r_n^{d+1}(s)\, \mathrm{d}s \\ &\qquad \geq \int_S p^2(s)r_n^{d+1}(s)\, \mathrm{d}s - p_{\max}^2 (r_n^{\max})^{d+1} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n), \end{aligned}$$

and finally

$$\begin{aligned}
\int_S h_n(s)\,\mathrm{d}s &= \int_{S\cap\mathcal{I}_n} h_n(s)\,\mathrm{d}s + \int_{S\cap\mathcal{R}_n} h_n(s)\,\mathrm{d}s + \int_{S\cap\mathcal{A}_n} h_n(s)\,\mathrm{d}s \\
&\geq \int_{S\cap\mathcal{I}_n} h_n(s)\,\mathrm{d}s \\
&\geq (1-\xi_n)^2(1-\nu_n)^{1+1/d}\frac{2\eta_{d-1}}{d+1}\int_{S\cap\mathcal{I}_n} p^2(s)r_n^{d+1}(s)\,\mathrm{d}s \\
&\geq (1-\xi_n)^2(1-\nu_n)^{1+1/d}\frac{2\eta_{d-1}}{d+1}\int_{S} p^2(s)r_n^{d+1}(s)\,\mathrm{d}s \\
&\quad - (1-\xi_n)^2(1-\nu_n)^{1+1/d}\frac{2\eta_{d-1}}{d+1}p_{\max}^2(r_n^{\max})^{d+1}\mathcal{L}_{d-1}(S\cap\mathcal{R}_n) \\
&\geq (1-\xi_n)^2(1-\nu_n)^{1+1/d}\frac{2\eta_{d-1}}{d+1}\int_{S} p^2(s)r_n^{d+1}(s)\,\mathrm{d}s \\
&\quad - 2\eta_{d-1}p_{\max}^2(r_n^{\max})^{d+1}\mathcal{L}_{d-1}(S\cap\mathcal{R}_n).
\end{aligned}$$

On the other hand, combining the bounds in (3.14), (3.13) and (3.12), we obtain

$$\begin{aligned}
\int_S h_n(s)\,\mathrm{d}s &= \int_{S\cap\mathcal{I}_n} h_n(s)\,\mathrm{d}s + \int_{S\cap\mathcal{R}_n} h_n(s)\,\mathrm{d}s + \int_{S\cap\mathcal{A}_n} h_n(s)\,\mathrm{d}s \\
&\leq (1+\xi_n)^2(1+\nu_n)^{1+1/d}\frac{2\eta_{d-1}}{d+1}\int_{S} p^2(s)r_n^{d+1}(s)\,\mathrm{d}s \\
&\quad + 2p_{\max}^2\eta_d(r_n^{\max})^{d+1}\mathcal{L}_{d-1}(S\cap\mathcal{R}_n).
\end{aligned}$$

□

Lemma 3.10 (Lower and upper bound on probability mass in caps) *Let the general assumptions and notations hold. Let $s \in S$, $t \in \mathbb{R}$ and $r > 0$. Then $g(s + tn_S, r) = 0$ if $|t| \geq r$. Otherwise, if $\tilde{p}_{\min} \leq p(y) \leq \tilde{p}_{\max}$ for all $y \in B(s + tn_S, r)$ we have*

$$\tilde{p}_{\min} r^d A\left(\frac{|t|}{r}\right) \leq g(s + tn_S, r) \leq \tilde{p}_{\max} r^d A\left(\frac{|t|}{r}\right),$$

where

$$A(t) = \mathcal{L}_d\left(B(0,1) \cap \left\{z = (z_1, \ldots, z_d) \in \mathbb{R}^d \;\middle|\; z_1 \geq t\right\}\right).$$

Proof. Suppose $t \geq 0$ so that $s + tn_S \in H^+$ (the other case can be treated analogously). We use that by a translation and a rotation of the coordinate system (such that the origin is at $s + tn_S$ and $-n_S$ is the direction of the first unit vector) and the invariance of the Lebesgue measure with respect to linear transformations we have

$$\mathcal{L}_d(B(s + tn_S, r) \cap H^-) = \mathcal{L}_d\left(B(0,r) \cap \left\{z = (z_1, \ldots, z_d) \in \mathbb{R}^d \;\middle|\; z_1 \geq t\right\}\right).$$

By scaling we obtain

$$\mathcal{L}_d\left(B(0,r)\cap\left\{z=(z_1,\ldots,z_d)\in\mathbb{R}^d \;\middle|\; z_1\geq t\right\}\right) = r^d\mathcal{L}_d\left(B(0,1)\cap\left\{z=(z_1,\ldots,z_d)\in\mathbb{R}^d \;\middle|\; z_1\geq \frac{t}{r}\right\}\right).$$

Considering that the probability mass can be bounded by the product of the minimal or maximal density and the Lebesgue measure we obtain the statement about the probability mass. □

Lemma 3.11 (Integral over cap volume) *With $A(t)$ defined as in Lemma 3.10 we have*

$$\int_0^1 A(t)\,\mathrm{d}t = \frac{\eta_{d-1}}{d+1}.$$

Proof. We have

$$\begin{aligned}\int_0^1 A(t)\mathrm{d}t &= \int_0^1 \mathcal{L}_d\left(B(0,1)\cap\left\{z=(z_1,\ldots,z_d)\in\mathbb{R}^d \;\middle|\; z_1\geq t\right\}\right)\,\mathrm{d}t\\ &= \int_0^1\int_t^1 \eta_{d-1}\sqrt{1-r^2}^{d-1}\,\mathrm{d}r\,\mathrm{d}t = \int_0^1\int_0^r \eta_{d-1}\sqrt{1-r^2}^{d-1}\,\mathrm{d}t\,\mathrm{d}r\\ &= \int_0^1 \eta_{d-1}\sqrt{1-r^2}^{d-1}\int_0^r \mathrm{d}t\,\mathrm{d}r = \int_0^1 \eta_{d-1} r\sqrt{1-r^2}^{d-1}\,\mathrm{d}r.\end{aligned}$$

Substituting $r=\cos\theta$, we have to integrate from $\theta=\arccos(0)=\pi/2$ to $\theta=\arccos(1)=0$, and have $\mathrm{d}r=-\sin\theta\,\mathrm{d}\theta$. Thus,

$$\int_0^1 A(t)\,\mathrm{d}t = \eta_{d-1}\int_0^{\pi/2}\cos\theta\sin^d\theta\,\mathrm{d}\theta = \eta_{d-1}\left[\frac{1}{d+1}\sin^{d+1}\theta\right]_0^{\pi/2} = \frac{\eta_{d-1}}{d+1}.$$

□

Lemma 3.12 (Convergence $\mathcal{L}_{d-1}(\{s\in S \mid \mathrm{dist}(x,\partial C)\leq\nu_n\})\to 0$ for $\nu_n\to 0$) *Let the general assumptions hold and let $(\nu_n)_{n\in\mathbb{N}}\subseteq\mathbb{R}_{>0}$ be a sequence with $\nu_n\to 0$ for $n\to\infty$. Define*

$$\mathcal{R}_n = \{x\in\mathbb{R}^d \mid \mathrm{dist}(x,\partial C)\leq\nu_n\}.$$

Then we have $\mathcal{L}_{d-1}(S\cap\mathcal{R}_n)\to 0$ for $n\to\infty$. In the case $d=1$, furthermore, $\mathcal{L}_{d-1}(S\cap\mathcal{R}_n)=0$ for all but finitely many n. If $d\geq 2$ and the rate conditions hold then for ν_n sufficiently small

$$\mathcal{L}_{d-1}(S\cap\mathcal{R}_n)\leq\frac{6\nu_n}{\sin(\alpha/2)}\mathcal{L}_{d-2}(S\cap\partial C).$$

Proof. We first show the statements in the more interesting case $d \geq 2$. For the proof of convergence we assume without loss of generality that $(\nu_n)_{n\in\mathbb{N}}$ is monotonically decreasing. If this is not the case then we will show the result for a suitable monotonic subsequence and use the fact that the sequence (ν_n) converges to 0.
We have $\mathcal{R}_1 \supseteq \mathcal{R}_2 \supseteq \ldots$ and $\mathcal{L}_{d-1}(S \cap \mathcal{R}_1) < \infty$. The limit $\lim_{n\to\infty} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n)$ exists because $\mathcal{L}_{d-1}(S \cap \mathcal{R}_1) < \infty$, the sequence $\mathcal{L}_{d-1}(S \cap \mathcal{R}_n)$ is decreasing and bounded from below. By the continuity of the Lebesgue measure

$$\mathcal{L}_{d-1}\left(\bigcap_{n=1}^{\infty}(S \cap \mathcal{R}_n)\right) = \lim_{n\to\infty} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n).$$

On the other hand $S \cap \partial C = \cap_{n=1}^{\infty}(S \cap \mathcal{R}_n)$. Therefore

$$\mathcal{L}_{d-1}(S \cap \partial C) = \lim_{n\to\infty} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n).$$

Suppose $\mathcal{L}_{d-1}(S \cap \mathcal{R}_n) \not\to 0$ for $n \to \infty$ and remember that $\mathcal{L}_{d-1}(S \cap \partial C) = 0$ by assumption. The limit $v = \lim_{n\to\infty} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n)$ exists and by our assumption $v > 0$. Therefore $\mathcal{L}_{d-1}(S \cap \partial C) = v > 0$, which is a contradiction.
In order to prove the convergence rate we first show that for sufficiently small ν_n we have

$$S \cap \mathcal{R}_n = \{s \in S \mid \operatorname{dist}(s, \partial C) \leq \nu_n\} \subseteq \left\{ s \in S \;\middle|\; \operatorname{dist}(s, S \cap \partial C) \leq \frac{2\nu_n}{\sin(\alpha/2)} \right\}.$$

In the following let $\operatorname{dist}_{\partial C}(x, y)$ denote the geodesic distance in ∂C of $x, y \in \partial C$. We have for $x, y \in \partial C$

$$\begin{aligned} \left|\langle n_y, n_S\rangle\right| &= \left|\langle n_y - n_x + n_x, n_S\rangle\right| = \left|\langle n_y - n_x, n_S\rangle + \langle n_x, n_S\rangle\right| \\ &\leq \left|\langle n_y - n_x, n_S\rangle\right| + \left|\langle n_x, n_S\rangle\right|, \end{aligned}$$

and with the Cauchy-Schwartz inequality

$$\begin{aligned} &\leq \|n_y - n_x\|\|n_S\| + |\langle n_x, n_S\rangle| = \|n_y - n_x\| + |\langle n_x, n_S\rangle| \\ &\leq \frac{\operatorname{dist}_{\partial C}(x, y)}{\kappa} + \left|\langle n_x, n_S\rangle\right|, \end{aligned}$$

where in the last step we use that, according to Lemma A.15, $\|n_x - n_y\| \leq \operatorname{dist}_{\partial C}(x, y)/\kappa$. In particular, if $x \in S \cap \partial C$, $y \in \partial C$ and $\operatorname{dist}_{\partial C}(x, y) \leq \kappa(\cos(\alpha/2) - \cos\alpha)$ we have $\left|\langle n_y, n_S\rangle\right| \leq \cos(\alpha/2)$.
Define

$$B_{irr} = \{x \in \partial C \mid \operatorname{dist}_{\partial C}(x, S \cap \partial C) \geq \kappa(\cos(\alpha/2) - \cos\alpha)\},$$

which is a compact set. Therefore the function $\operatorname{dist}(., S)$ must attain its minimum on B_{irr} and we set $b_{irr} = \operatorname{argmin}_{x\in B_{irr}} \operatorname{dist}(x, S)$. Clearly $\operatorname{dist}(b_{irr}, S) > 0$, since otherwise $b_{irr} \in S$ which is a contradiction to the construction of B_{irr}.

In the following suppose that $\nu_n < \text{dist}(B_{irr}, S)$. Let s be a point in S with $\text{dist}(s, \partial C) \leq \nu_n$ and let $s' \in \partial C$ such that $\text{dist}(s, s') = \text{dist}(s, \partial C)$ (such a point must exist due to the compactness of ∂C and the continuity of $\text{dist}(s, .)$). In the following we set $h = \text{dist}(s, s')$. If $h = 0$ we have $s = s'$ and thus $\text{dist}(s, S \cap \partial C) = 0$. Next we will consider the case $h > 0$.
Since $\text{dist}(s', s) = h \leq \nu_n$ but $\text{dist}(B_{irr}, S) > \nu_n$ we have $s' \notin B_{irr}$ and therefore for the normal $n_{s'}$ that $|\langle n_S, n_{s'} \rangle| \leq \cos(\alpha/2)$. Let T denote the tangential hyperplane to ∂C in the point s' and n_T its normal vector, that is $n_T = n_{s'}$. Since ∂C is a smooth $(d-1)$-dimensional surface we can represent it locally as a smooth function over the tangent plane. That is, there is a radius R such that for any unit vector $u \in T$ the intersection of ∂C with the plane spanned by n_T and u can be represented locally by a smooth function $f_u : \mathbb{R} \to \mathbb{R}$, such that the intersection curve is given by

$$s' + \xi u + f_u(\xi) n_T,$$

for $|\xi| \leq R$. Furthermore $f_u(0) = 0$ and $f_u'(0) = 0$, that is with a Taylor-expansion around 0 we obtain

$$f_u(\xi) = f_u(0) + \xi f_u'(0) + \frac{\xi^2}{2} f_u''(\theta\xi) = \xi^2 f_u''(\theta\xi)$$

with $\theta \in (0, 1)$.
Now we show that the vector $s - s'$ is perpendicular to the tangent plane T in s'. Let u be a unit vector in T. For $|\xi| \leq R$ the point

$$s'' = s' + \xi u + f_u(\xi) n_T$$

is on the surface ∂C. Now we have

$$\begin{aligned} \|s - s''\|^2 &= \|s - s' - \xi u - f_u(\xi) n_T\|^2 \\ &= \|s - s'\|^2 + \|\xi u - f_u(\xi) n_T\|^2 - 2\langle s - s', \xi u + f_u(\xi) n_T \rangle \\ &= \|s - s'\|^2 + \xi^2 + f_u^2(\xi) - 2\xi \langle s - s', u \rangle - 2 f_u(\xi) \langle s - s', n_T \rangle \end{aligned}$$

and using the Taylor expansion of f_u with a $\theta \in (0, 1)$

$$= \|s - s'\|^2 + \xi^2 + \frac{\xi^4}{4} f_u''^2(\theta\xi) - 2\xi \langle s - s', u \rangle - \xi^2 f_u''(\theta\xi) \langle s - s', n_T \rangle.$$

We can find $R_2 > 0$ such that $|f_u''(x)| \leq 2|f_u''(0)|$ for $-R_2 \leq x \leq R_2$. Therefore, if $|\xi| \leq R_2$ we have

$$\|s - s''\|^2 \leq \|s - s'\|^2 - 2\xi \langle s - s', u \rangle + \xi^2 + 4\frac{\xi^4}{4} |f_u''(0)|^2 + 2\xi^2 |f_u''(0)| |\langle s - s', n_T \rangle|.$$

Suppose $\langle s - s', u \rangle \neq 0$. Then we can find ξ such that for the corresponding point s'' we have

$$\|s - s''\|^2 < \|s - s'\|^2,$$

That is we have found a point s'' on ∂C that is closer to s than s'. This is a contradiction to $\text{dist}(s, s') = \text{dist}(s, \partial C)$! Therefore $\langle s - s', u \rangle = 0$ and since the tangent vector u was arbitrary, $s - s'$ is perpendicular to the tangent plane T.
Due to the condition on the angle α the intersection $S \cap T$ is a $(d-2)$-dimensional affine subspace of R^d. Now let p be the orthogonal projection of s' onto the subspace $S \cap T$, that is

$$p = \operatorname*{argmin}_{x \in S \cap T} \|s' - x\|^2.$$

Clearly $s' - p$ is a vector in the hyperplane T and therefore perpendicular to $s - s'$, that is by Pythagoras

$$\|s - p\|^2 = \|s - s' + s' - p\|^2 = \|s - s'\|^2 + \|s' - p\|^2,$$

and thus p is also the orthogonal projection of s onto $S \cap T$.
Define $v_1 = (s - p)/\|s - p\|$ and $v_2 = (s' - p)/\|s' - p\|$. Clearly, v_1 is a vector in the hyperplane S and v_2 in T. Due to the choice of p the vectors n_S, n_T and v_1, v_2 span a (2-dimensional) plane and we have $\langle n_S, v_1 \rangle = 0$ and $\langle n_T, v_2 \rangle = 0$. Since we are in a 2-dimensional subspace it is clear that $|\langle v_1, v_2 \rangle| = |\langle n_S, n_T \rangle|$. Now we consider the plane

$$E = \left\{ s' + u_1 \frac{p - s'}{\|p - s'\|} + u_2 \frac{s - s'}{\|s - s'\|} \;\middle|\; u_1, u_2 \in \mathbb{R} \right\}$$

and in this plane we consider the coordinate system with its origin at s' and orthonormal basis $(p - s')/\|p - s'\|$ and $(s - s')/\|s - s'\|$. In this coordinate system $s' = (0, 0)$ and the intersection of T and E is the horizontal line. Since $h = \|s - s'\|$ we have $s = (0, h)$. The cosine of the angle at p in the triangle spanned by s,s' and p is $|\langle v_1, v_2 \rangle|$ and therefore the angle is bounded from below by $\alpha/2$. In the following we assume that this angle equals $\alpha/2$. Then we have $p = (h/\tan(\alpha/2), 0)$. Thus the intersection of S with E is given by the function $g_S : \mathbb{R} \to \mathbb{R}$ with $g_S(x) = h - x\tan(\alpha/2)$. Let f denote the local representation of the surface from above. Clearly we can find a radius $R_3 > 0$ such for all $0 \le x \le R_3$ we have $f(x) \ge f_l(x)$ with $f_l(x) = -(x\tan(\alpha/2))/2$. Since the minimal curvature radius is bounded away from 0 on ∂C we can even find the radius $R_3 > 0$ such that it is independent of the exact s'. The intersection of f_l with g_S is the point $(2h/\tan(\alpha/2), -h)$. Figure 3.7 illustrates these functions in the plane E.
Now suppose that $h \le (R_3 \tan(\alpha/2))/2$. Then $f(0) = 0 < g_S(0)$ but

$$f\left(\frac{2h}{\tan(\alpha/2)}\right) \ge f_l\left(\frac{2h}{\tan(\alpha/2)}\right) = -h = g_S\left(\frac{2h}{\tan(\alpha/2)}\right),$$

and therefore there exists a point $q = (q_1, q_2) \in S \cap \partial C \cap E$ with $0 \le q_1 \le 2h/\tan(\alpha/2)$. Clearly,

$$\begin{aligned}\text{dist}(s, q) &\le \sqrt{4h^2 + \frac{4h^2}{\tan^2(\alpha/2)}} = 2h\sqrt{1 + \frac{\cos^2(\alpha/2)}{\sin^2(\alpha/2)}} = 2h\sqrt{1 + \frac{1 - \sin^2(\alpha/2)}{\sin^2(\alpha/2)}} \\ &= \frac{2h}{\sin(\alpha/2)},\end{aligned}$$

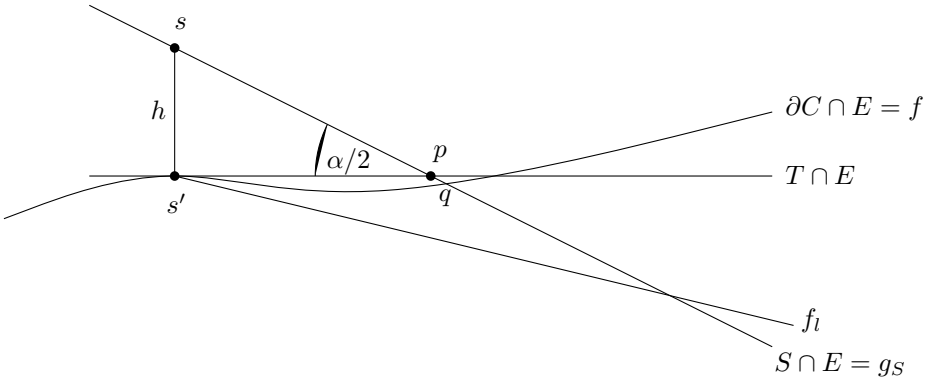

Figure 3.7: An illustration of the points s, s', p and the functions g_S, f in the plane E where the angle between n_T and n_S equals $\alpha/2$. Locally the function f is bounded from below by the linear function g_S. Since g_S and f_l intersect there must also be an intersection point q of f and g_S. The distance of q to s cannot be greater than the distance of the intersection of g_S and f_l to s.

and therefore $\operatorname{dist}(s, S \cap \partial C) \leq 2h/\sin(\alpha/2)$. Since s was arbitrary, we have shown that

$$S \cap \mathcal{R}_n = \{s \in S \mid \operatorname{dist}(s, \partial C) \leq \nu_n\} \subseteq \left\{ s \in S \;\middle|\; \operatorname{dist}(s, S \cap \partial C) \leq \frac{2\nu_n}{\sin(\alpha/2)} \right\}.$$

According to Lemma A.18 $S \cap \partial C$ consists of finitely many connected components in the $(d-1)$-dimensional subspace S and the relative boundary of each one is a closed, smooth $(d-2)$-dimensional surface without a boundary.
Therefore, we have with Lemma A.14, applied to the $(d-1)$-dimensional affine subspace S, that for ν_n sufficiently small

$$\begin{aligned}\mathcal{L}_{d-1}\left(\left\{ s \in S \;\middle|\; \operatorname{dist}(s, S \cap \partial C) \leq \frac{2\nu_n}{\sin(\alpha/2)} \right\}\right) &\leq 3\mathcal{L}_{d-2}(S \cap \partial C)\frac{2\nu_n}{\sin(\alpha/2)} \\ &= \frac{6\nu_n}{\sin(\alpha/2)}\mathcal{L}_{d-2}(S \cap \partial C).\end{aligned}$$

In the case $d = 1$ the hyperplane S is a single point and the condition $\mathcal{L}_{d-1}(S \cap \partial C) = 0$ implies that $S \notin \partial C$. Due to the compactness of C we can find an $\varepsilon > 0$ such that $B(S, \varepsilon) \cap C = \emptyset$. That means, for ν_n sufficiently small $S \cap \mathcal{R}_n = \emptyset$ and therefore $\mathcal{L}_0(S \cap \mathcal{R}_n) = 0$ for all but finitely many n. □

Proof of Proposition 3.3. We first show the statement for the ***r*-neighborhood graph**. According to Lemma 3.8 and the linearity of the expectation we have

$$\mathbb{E}\left(\frac{\operatorname{cut}_{n,r_n}(S)}{n^2 r_n^{d+1}}\right) = \frac{n-1}{n}\frac{1}{r_n^{d+1}}\int_{\mathbb{R}^d} g(x, r_n)p(x)\,\mathrm{d}x.$$

Setting $r_n(x) = \tilde{r}(x) = r_n^{\max} = r_n$ in Lemma 3.9 we can clearly choose $\nu_n = 0$ for all $n \in \mathbb{N}$.
Assume $s \in S$ and $B(s, 3r_n^{\max}) \subseteq C$. For all $y \in B(s, 3r_n^{\max})$ we have

$$p(s) - 3p'_{\max} r_n^{\max} \leq p(y) \leq p(s) + 3p'_{\max} r_n^{\max},$$

or, written differently and using $r_n^{\max} = r_n$,

$$p(s)\left(1 - 3\frac{p'_{\max}}{p(s)} r_n\right) \leq p(y) \leq p(s)\left(1 + 3\frac{p'_{\max}}{p(s)} r_n\right).$$

Setting $\xi_n = 3p'_{\max} r_n / p_{\min}$ Equation (3.8) in Lemma 3.9 holds if $\xi_n < 1$. In the following let n be sufficiently large such that $r_n < p_{\max}/(3p_{\min})$, that is $\xi_n < 1$. Applying Lemma 3.9 we obtain

$$\begin{aligned}
\mathbb{E}\left(\frac{\mathrm{cut}_{n,r_n}(S)}{n^2 r_n^{d+1}}\right) &\leq \frac{n-1}{n}\frac{1}{r_n^{d+1}}\Bigg((1+\xi_n)^2(1+\nu_n)^{1+1/d}\frac{2\eta_{d-1}}{d+1}\int_S p^2(s) r_n^{d+1}(s)\,\mathrm{d}s \\
&\qquad + 2\eta_d p_{\max}^2 (r_n^{\max})^{d+1}\mathcal{L}_{d-1}(S \cap \mathcal{R}_n)\Bigg) \\
&\leq (1+\xi_n)^2 \frac{2\eta_{d-1}}{d+1}\int_S p^2(s)\,\mathrm{d}s + 2\eta_d p_{\max}^2 \mathcal{L}_{d-1}(S \cap \mathcal{R}_n), \qquad (3.15)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\left(\frac{\mathrm{cut}_{n,r_n}(S)}{n^2 r_n^{d+1}}\right) &\geq \frac{n-1}{n}\frac{1}{r_n^{d+1}}\Bigg((1-\xi_n)^2(1-\nu_n)^{1+1/d}\frac{2\eta_{d-1}}{d+1}\int_S p^2(s) r_n^{d+1}(s)\,\mathrm{d}s \\
&\qquad - 2\eta_{d-1} p_{\max}^2 (r_n^{\max})^{d+1}\mathcal{L}_{d-1}(S \cap \mathcal{R}_n)\Bigg) \\
&\geq \frac{n-1}{n}(1-\xi_n)^2 \frac{2\eta_{d-1}}{d+1}\int_S p^2(s)\,\mathrm{d}s - 2\eta_{d-1} p_{\max}^2 \mathcal{L}_{d-1}(S \cap \mathcal{R}_n). \qquad (3.16)
\end{aligned}$$

We simplify the expressions in Equation (3.15) using $\xi_n < 1$ and $\eta_d \leq 6$ for all $d \geq 1$

$$\mathbb{E}\left(\frac{\mathrm{cut}_{n,r_n}(S)}{n^2 r_n^{d+1}}\right) \leq \frac{2\eta_{d-1}}{d+1}\int_S p^2(s)\,\mathrm{d}s + 18\xi_n \int_S p^2(s)\,\mathrm{d}s + 12 p_{\max}^2 \mathcal{L}_{d-1}(S \cap \mathcal{R}_n),$$

and

$$\mathbb{E}\left(\frac{\mathrm{cut}_{n,r_n}(S)}{n^2 r_n^{d+1}}\right) \geq \frac{2\eta_{d-1}}{d+1}\int_S p^2(s)\,\mathrm{d}s - 6\left(2\xi_n + \frac{1}{n}\right)\int_S p^2(s)\,\mathrm{d}s - 12 p_{\max}^2 \mathcal{L}_{d-1}(S \cap \mathcal{R}_n).$$

Clearly $\int_S p^2(s)\,\mathrm{d}s \leq p_{\max}^2 \mathcal{L}_{d-1}(S \cap C)$, and thus

$$\begin{aligned}
&\left|\mathbb{E}\left(\frac{\mathrm{cut}_{n,r_n}(S)}{n^2 r_n^{d+1}}\right) - \frac{2\eta_{d-1}}{d+1}\int_S p^2(s)\,\mathrm{d}s\right| \\
&\quad \leq 6\,(3\xi_n + 1/n)\, p_{\max}^2 \mathcal{L}_{d-1}(S \cap C) + 12 p_{\max}^2 \mathcal{L}_{d-1}(S \cap \mathcal{R}_n) \\
&\quad = 6\left(\frac{9p'_{\max}}{p_{\min}} r_n + \frac{1}{n}\right) p_{\max}^2 \mathcal{L}_{d-1}(S \cap C) + 12 p_{\max}^2 \mathcal{L}_{d-1}(S \cap \mathcal{R}_n).
\end{aligned}$$

Since we assumed $r_n \to 0$ for $n \to \infty$ the convergence towards zero for $n \to \infty$ of the first term is clear. Under the condition $\mathcal{L}_{d-1}(S \cap \partial C) = 0$ we have by Lemma 3.12 that $\mathcal{L}_{d-1}(S \cap \mathcal{R}_n) \to 0$ for $n \to \infty$, and therefore the second term also converges towards zero for $n \to \infty$.
Under the condition $nr_n \to \infty$ we have for n sufficiently large, $nr_n \geq p_{\min}/p'_{\max}$, and thus

$$6\left(\frac{9p'_{\max}}{p_{\min}}r_n + \frac{1}{n}\right)p_{\max}^2\mathcal{L}_{d-1}(S \cap C) \leq \frac{60p'_{\max}}{p_{\min}}p_{\max}^2\mathcal{L}_{d-1}(S \cap C)r_n.$$

For $d = 1$ we have with Lemma 3.12 that $\mathcal{L}_{d-1}(S \cap \mathcal{R}_n) = 0$ for all but finitely many n. For $d \geq 2$ and under the rate conditions we have with the same lemma for n sufficiently large

$$12p_{\max}^2\mathcal{L}_{d-1}(S \cap \mathcal{R}_n) \leq 12p_{\max}^2\frac{18r_n}{\sin(\alpha/2)}\mathcal{L}_{d-2}(S \cap \partial C) = \frac{216\mathcal{L}_{d-2}(S \cap \partial C)}{\sin(\alpha/2)}p_{\max}^2 r_n.$$

In the following we show the statement for the ***k*-nearest neighbor graph**. According to Lemma 3.8 and the linearity of the expectation we have for $\delta_n < 1/2$ and $k_n < (n-1)/2$

$$\begin{aligned}\mathbb{E}\left(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S)\right) &\leq \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{\mathbb{R}^d} g(x, \tilde{r}(x,(1+\delta_n)\alpha_n))p(x)\,\mathrm{d}x \\ &\quad + \frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}2\exp(2\log n - \delta_n^2 k_n/4) \\ &\leq \frac{n-1}{k_n}\sqrt[d]{\frac{n}{k_n}}\int_{\mathbb{R}^d} g(x, \tilde{r}(x,(1+\delta_n)\alpha_n))p(x)\,\mathrm{d}x \\ &\quad + 2\exp(2\log n - \delta_n^2 k_n/4). \end{aligned} \tag{3.17}$$

In the following, we bound the integral

$$\int_{\mathbb{R}^d} g(x, \tilde{r}(x,(1+\delta_n)\alpha_n))p(x)\,\mathrm{d}x$$

using Lemma 3.9.
Set

$$r_n^{\max} = \sqrt[d]{\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}} \qquad \text{and} \qquad r_n(x) = \sqrt[d]{\frac{(1+\delta_n)\alpha_n}{p(x)\eta_d}}.$$

By definition $r_n(x) \leq r_n^{\max}$ for all $x \in C$ since $p_{\min} \leq p(x)$ and $\gamma \leq 1$. We identify $\tilde{r}_n(x) = \tilde{r}(x,(1+\delta_n)\alpha_n))$ and have $\tilde{r}_n(x) \leq r_n^{\max}$ for all $x \in C$ given that $r_n^{\max} < r_\gamma$: Computing the probability mass in balls of radius $r_n^{\max}$ for an arbitrary $x \in C$, we have

$$\begin{aligned}\mu(B(x, r_n^{\max})) &= \mu\left(B\left(x, \sqrt[d]{\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}}\right)\right) \geq p_{\min}\mathcal{L}_d\left(B\left(x, \sqrt[d]{\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}}\right) \cap C\right) \\ &\geq p_{\min}\gamma\mathcal{L}_d\left(B\left(x, \sqrt[d]{\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}}\right)\right) = p_{\min}\gamma\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}\eta_d \\ &= (1+\delta_n)\alpha_n,\end{aligned}$$

and therefore we have $\tilde{r}_n(x) \leq r_n^{\max}$ for all $x \in C$.
Now we consider the second condition in Lemma 3.9. Assume that $s \in S$ and $B(s, 3r_n^{\max}) \subseteq C$. For all $y \in B(s, 3r_n^{\max})$ we have

$$p(s) - 3p'_{\max} r_n^{\max} \leq p(y) \leq p(s) + 3p'_{\max} r_n^{\max},$$

or, written differently,

$$p(s)\left(1 - 3\frac{p'_{\max}}{p(s)} r_n^{\max}\right) \leq p(y) \leq p(s)\left(1 + 3\frac{p'_{\max}}{p(s)} r_n^{\max}\right).$$

Setting $\xi_n = 3p'_{\max} r_n^{\max} / p_{\min}$ we have

$$p(s)(1 - \xi_n) \leq p(y) \leq p(s)(1 + \xi_n). \tag{3.18}$$

Now we show that for $\xi_n < 1/2$ we can set $\nu_n = 2\xi_n$. Under this condition we have $\nu_n < 1$,

$$1 + 2\xi_n = \frac{(1 + 2\xi_n)(1 - \xi_n)}{1 - \xi_n} = \frac{1 + \xi_n - 2\xi_n^2}{1 - \xi_n} > \frac{1}{1 - \xi_n},$$

and

$$1 - 2\xi_n = \frac{(1 - 2\xi_n)(1 + \xi_n)}{1 + \xi_n} = \frac{1 - \xi_n + 2\xi_n^2}{1 + \xi_n} < \frac{1}{1 + \xi_n}.$$

Therefore, for $x \in B(s, r_n^{\max})$ we have

$$\begin{aligned}
\mu(B(x, \sqrt[d]{1 + 2\xi_n} r_n(s)) &> \mu\left(B\left(x, \sqrt[d]{\frac{1}{1 - \xi_n}} r_n(s)\right)\right) \\
&\geq (1 - \xi_n) p(s) \mathcal{L}_d\left(B\left(x, \sqrt[d]{\frac{1}{1 - \xi_n}} r_n(s)\right)\right) \\
&= (1 - \xi_n) p(s) \left(\sqrt[d]{\frac{1}{1 - \xi_n}} r_n(s)\right)^d \eta_d \\
&= p(s) \eta_d r_n(s)^d = (1 + \delta_n)\alpha_n,
\end{aligned}$$

and

$$\begin{aligned}
\mu(B(x, \sqrt[d]{1 - 2\xi_n} r_n(s)) &< \mu\left(B\left(x, \sqrt[d]{\frac{1}{1 + \xi_n}} r_n(s)\right)\right) \\
&\leq (1 + \xi_n) p(s) \mathcal{L}_d\left(B\left(x, \sqrt[d]{\frac{1}{1 + \xi_n}} r_n(s)\right)\right) \\
&= (1 + \xi_n) p(s) \left(\sqrt[d]{\frac{1}{1 + \xi_n}} r_n(s)\right)^d \eta_d \\
&= p(s) \eta_d r_n(s)^d = (1 + \delta_n)\alpha_n.
\end{aligned}$$

The strict inequalities hold, because the balls $B(x, \sqrt[d]{1+2\xi_n} r_n(s))$ and $B(x, \sqrt[d]{1-2\xi_n} r_n(s))$ lie completely within C. Therefore, we have

$$\sqrt[d]{1-2\xi_n} r_n(s) \leq \tilde{r}_n(x) \leq \sqrt[d]{1+2\xi_n} r_n(s).$$

Inserting this into Lemma 3.9 we obtain:

$$\begin{aligned}\int_{\mathbb{R}^d} p(x) g(x, \tilde{r}_n(x)) \, \mathrm{d}x \leq (1+\xi_n)^2 (1+\nu_n)^{1+1/d} \frac{2\eta_{d-1}}{d+1} \int_S p^2(s) r_n^{d+1}(s) \, \mathrm{d}s \\ + 2\eta_d p_{\max}^2 (r_n^{\max})^{d+1} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n),\end{aligned}$$

and since $\nu_n = 2\xi_n > 0$ and $\xi_n < 1/2$

$$\begin{aligned}&\leq (1+16\xi_n) \frac{2\eta_{d-1}}{d+1} \int_S p^2(s) \left(\frac{(1+\delta_n)\alpha_n}{p(s)\eta_d} \right)^{1+1/d} \mathrm{d}s \\ &\quad + 2\eta_d p_{\max}^2 \left(\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min} \eta_d} \right)^{1+1/d} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n) \\ &\leq (\alpha_n(1+\delta_n))^{1+1/d} \left[\frac{2\eta_{d-1}(1+16\xi_n)}{(d+1)\eta_d^{1+1/d}} \int_S p^{1-1/d}(s) \, \mathrm{d}s \right. \\ &\quad \left. + \frac{2p_{\max}^2}{(\gamma p_{\min})^{1+1/d} \eta_d^{1/d}} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n) \right].\end{aligned}$$

Since $\alpha_n = k_n/(n-1)$ we have

$$\frac{n-1}{k_n} \sqrt[d]{\frac{n}{k_n}} \alpha_n^{1+1/d} = \frac{n-1}{k_n} \sqrt[d]{\frac{n}{k_n}} \left(\frac{k_n}{n-1} \right)^{1+1/d} = \sqrt[d]{\frac{n}{n-1}} \leq 1 + \frac{2}{n}.$$

Employing this result in Equation (3.17), using $n \geq 2$ and $\delta < 1$

$$\begin{aligned}\mathbb{E}\left(\frac{1}{nk_n} \sqrt[d]{\frac{n}{k_n}} \operatorname{cut}_{n,k_n}(S) \right) &\leq \left(1+\frac{2}{n}\right)(1+\delta_n)^2 \left[\frac{2\eta_{d-1}(1+16\xi_n)}{(d+1)\eta_d^{1+1/d}} \int_S p^{1-1/d}(s) \, \mathrm{d}s \right. \\ &\quad \left. + \frac{2p_{\max}^2}{(\gamma p_{\min})^{1+1/d} \eta_d^{1/d}} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n) \right] + 2\exp(2\log n - \delta_n^2 k_n/4) \\ &\leq \left(1+\frac{2}{n}\right)(1+3\delta_n)(1+16\xi_n) \frac{2\eta_{d-1}}{(d+1)\eta_d^{1+1/d}} \int_S p^{1-1/d}(s) \, \mathrm{d}s \\ &\quad + \frac{16p_{\max}^2}{(\gamma p_{\min})^{1+1/d} \eta_d^{1/d}} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n) + 2\exp(2\log n - \delta_n^2 k_n/4) \\ &\leq \frac{2\eta_{d-1}}{(d+1)\eta_d^{1+1/d}} \int_S p^{1-1/d}(s) \, \mathrm{d}s \\ &\quad + \left(3\delta_n + 64\xi_n + \frac{72}{n}\right) \frac{2\eta_{d-1}}{\eta_d^{1+1/d}} p_{\max}^{1-1/d} \mathcal{L}_{d-1}(S \cap C) \\ &\quad + \frac{16p_{\max}^2 \eta_d}{(\gamma p_{\min} \eta_d)^{1+1/d}} \mathcal{L}_{d-1}(S \cap \mathcal{R}_n) + 2\exp(2\log n - \delta_n^2 k_n/4).\end{aligned}$$

For the lower inequality we obtain similarly with Lemma 3.9

$$\int_{\mathbb{R}^d} p(x)g(x,\tilde{r}_n(x))\,\mathrm{d}x \geq (\alpha_n(1-\delta_n))^{1+1/d}\left[\frac{2\eta_{d-1}(1-6\xi_n)}{(d+1)\eta_d^{1+1/d}}\int_S p^{1-1/d}(s)\,\mathrm{d}s - \frac{2\eta_{d-1}p_{\max}^2}{(\gamma p_{\min})^{1+1/d}\eta_d^{1+1/d}}\mathcal{L}_{d-1}(S\cap\mathcal{R}_n)\right].$$

Inserting this result back into Equation (3.17) and using $\eta_d \leq 6$ for all $d \geq 1$

$$\begin{aligned}\mathbb{E}\left(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S)\right) &\geq (1-2\delta_n)(1-6\xi_n)\frac{2\eta_{d-1}}{(d+1)\eta_d^{1+1/d}}\int_S p^{1-1/d}(s)\,\mathrm{d}s \\ &\quad - \frac{2\eta_{d-1}p_{\max}^2}{(\gamma p_{\min})^{1+1/d}\eta_d^{1+1/d}}\mathcal{L}_{d-1}(S\cap\mathcal{R}_n) - 2\exp(2\log n - \delta_n^2 k_n/4) \\ &\geq \frac{2\eta_{d-1}}{(d+1)\eta_d^{1+1/d}}\int_S p^{1-1/d}(s)\,\mathrm{d}s - (2\delta_n+6\xi_n)\frac{2\eta_{d-1}}{\eta_d^{1+1/d}}p_{\max}^{1-1/d}\mathcal{L}_{d-1}(S\cap C) \\ &\quad - \frac{12p_{\max}^2}{(\gamma p_{\min}\eta_d)^{1+1/d}}\mathcal{L}_{d-1}(S\cap\mathcal{R}_n) - 2\exp(2\log n - \delta_n^2 k_n/4).\end{aligned}$$

Combining the lower and the upper bound from above and using

$$\xi_n = 3\frac{p'_{\max}}{p_{\min}}r_n^{\max} = 3\frac{p'_{\max}}{p_{\min}}\sqrt[d]{\frac{(1+\delta_n)\alpha_n}{\gamma p_{\min}\eta_d}} \leq \frac{6p'_{\max}}{p_{\min}^{1+1/d}\sqrt[d]{\gamma\eta_d}}\sqrt[d]{\frac{k_n}{n-1}} \leq \frac{12p'_{\max}}{p_{\min}^{1+1/d}(\gamma\eta_d)^{1/d}}\sqrt[d]{\frac{k_n}{n}}$$

we obtain

$$\begin{aligned}&\left|\mathbb{E}\left(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S)\right) - \frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_S p^{1-1/d}(s)\,\mathrm{d}s\right| \\ &\leq \left(3\delta_n + 64\frac{12p'_{\max}}{p_{\min}^{1+1/d}(\gamma\eta_d)^{1/d}}\sqrt[d]{\frac{k_n}{n}} + \frac{72}{n}\right)\frac{2\eta_{d-1}}{\eta_d^{1+1/d}}p_{\max}^{1-1/d}\mathcal{L}_{d-1}(S\cap C) \\ &\quad + \frac{96p_{\max}^2}{(\gamma p_{\min}\eta_d)^{1+1/d}}\mathcal{L}_{d-1}(S\cap\mathcal{R}_n) + 2\exp(2\log n - \delta_n^2 k_n/4).\end{aligned}$$

Since for $n \geq 2$

$$\frac{1}{n} \leq \frac{1}{\sqrt[d]{n}} = \frac{1}{\sqrt[d]{k_n}}\sqrt[d]{\frac{k_n}{n}},$$

we can subsume the term $72/n$ under two times the $\sqrt[d]{k_n/n}$-term for sufficiently large n and obtain

$$\begin{aligned}&\left|\mathbb{E}\left(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S)\right) - \frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_S p^{1-1/d}(s)\,\mathrm{d}s\right| \\ &\leq \left(3\delta_n + \frac{1536p'_{\max}}{p_{\min}^{1+1/d}(\gamma\eta_d)^{1/d}}\sqrt[d]{\frac{k_n}{n}}\right)\frac{2\eta_{d-1}}{\eta_d^{1+1/d}}p_{\max}^{1-1/d}\mathcal{L}_{d-1}(S\cap C) \\ &\quad + \frac{96p_{\max}^2}{(\gamma p_{\min}\eta_d)^{1+1/d}}\mathcal{L}_{d-1}(S\cap\mathcal{R}_n) + 2\exp(2\log n - \delta_n^2 k_n/4).\end{aligned}$$

Choosing $\delta_n = 4\sqrt{(\log n)/k_n}$ and applying Lemma 3.12, the right hand side converges to zero for $n \to \infty$ and thus we obtain the convergence result.
Now we will show the convergence rates. For $d = 1$ we have with Lemma 3.12 that $\mathcal{L}_{d-1}(S \cap \mathcal{R}_n) = 0$ for all but finitely many n. For $d \geq 2$ and under the rate conditions we have with the same lemma for n sufficiently large

$$\begin{aligned}\frac{96p_{\max}^2}{(\gamma p_{\min}\eta_d)^{1+1/d}}\mathcal{L}_{d-1}(S \cap \mathcal{R}_n) &\leq \frac{96p_{\max}^2}{(\gamma p_{\min}\eta_d)^{1+1/d}}\frac{18r_n^{\max}}{\sin(\alpha/2)}\mathcal{L}_{d-2}(S \cap \partial C) \\ &= \frac{96p_{\max}^2}{(\gamma p_{\min}\eta_d)^{1+1/d}}\frac{18}{\sin(\alpha/2)}\sqrt[d]{\frac{(1+\delta_n)k_n}{\gamma p_{\min}\eta_d n}}\mathcal{L}_{d-2}(S \cap \partial C) \\ &\leq \frac{3456p_{\max}^2\mathcal{L}_{d-2}(S \cap \partial C)}{(\gamma p_{\min}\eta_d)^{1+2/d}\sin(\alpha/2)}\sqrt[d]{\frac{k_n}{n}}.\end{aligned}$$

In the following we do not consider the case $d = 1$ separately, since the proof in this case is the same as for the case $d \geq 2$ when we ignore the $\mathcal{L}_{d-2}(S \cap \partial C)$-term.
Plugging in the result for $d \geq 2$ under the rate conditions we obtain

$$\begin{aligned}&\left|\mathbb{E}\left(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\operatorname{cut}_{n,k_n}(S)\right) - \frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_S p^{1-1/d}(s)\,\mathrm{d}s\right| \\ &\quad\leq \left(3\delta_n + \frac{1536p'_{\max}}{p_{\min}^{1+1/d}(\gamma\eta_d)^{1/d}}\sqrt[d]{\frac{k_n}{n}}\right)\frac{2\eta_{d-1}}{\eta_d^{1+1/d}}p_{\max}^{1-1/d}\mathcal{L}_{d-1}(S \cap C) \\ &\qquad+ \frac{3456p_{\max}^2\mathcal{L}_{d-2}(S \cap \partial C)}{(\gamma p_{\min}\eta_d)^{1+2/d}\sin(\alpha/2)}\sqrt[d]{\frac{k_n}{n}} + 2\exp(2\log n - \delta_n^2 k_n/4).\end{aligned}$$

Clearly there is a trade-off in the choice of δ_n: For the convergence of the first term δ_n should go to zero quickly, whereas for the convergence of the third term δ_n should be as large as possible. We have

$$\exp\left(2\log n - \frac{\delta_n^2 k_n}{4}\right) = \exp\left(\log n\left(2 - \frac{\delta_n^2 k_n}{4\log n}\right)\right) = n^{2-\frac{\delta_n^2 k_n}{4\log n}}.$$

Therefore, the lowest growth rate such that this term converges is $\delta_n = a\sqrt{(\log n)/k_n}$ with $a > \sqrt{12}$. In the following we set $\delta_n = 4\sqrt{(\log n)/k_n}$ such that

$$\exp\left(2\log n - \frac{\delta_n^2 k_n}{4}\right) = \frac{1}{n^2}.$$

In fact, this choice of a is arbitrary, since for every $a > \sqrt{12}$ the exponential term converges faster than $1/n$ and therefore faster than $\sqrt[d]{\frac{k_n}{n}}$. If δ_n is set in this way we have $\delta_n \to 0$ since we assumed $k_n/\log n \to \infty$ for $n \to \infty$, and thus for n sufficiently large we have $\delta_n < 1/2$ and $\delta_n k_n > 1$. Clearly, for a value of n that is sufficiently large we can

subsume the last term $\exp(2\log n - \delta_n^2 k_n/4) = 2/n^2$ under two times the $\sqrt[d]{k_n/n}$ term in the bracket and obtain

$$\begin{aligned}\left|\mathbb{E}\left(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S)\right) - \frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_S p^{1-1/d}(s)\,\mathrm{d}s\right| \\ \leq \left(3\delta_n + \frac{3072 p'_{\max}}{p_{\min}^{1+1/d}(\gamma\eta_d)^{1/d}}\sqrt[d]{\frac{k_n}{n}}\right)\frac{2\eta_{d-1}}{\eta_d^{1+1/d}}p_{\max}^{1-1/d}\mathcal{L}_{d-1}(S\cap C) \\ + \frac{3456 p_{\max}^2 \mathcal{L}_{d-2}(S\cap\partial C)}{(\gamma p_{\min}\eta_d)^{1+2/d}\sin(\alpha/2)}\sqrt[d]{\frac{k_n}{n}}.\end{aligned}$$

Plugging in the definition of δ_n and setting

$$c_1 = \frac{24\eta_{d-1}}{\eta_d^{1+1/d}}p_{\max}^{1-1/d}\mathcal{L}_{d-1}(S\cap C)$$

and

$$c_2 = \frac{6144 p'_{\max}\eta_{d-1}p_{\max}^{1-1/d}\mathcal{L}_{d-1}(S\cap C)}{p_{\min}^{1+1/d}(\gamma\eta_d)^{1/d}\eta_d^{1+1/d}} + \frac{3456 p_{\max}^2\mathcal{L}_{d-2}(S\cap\partial C)}{(\gamma p_{\min}\eta_d)^{1+2/d}\sin(\alpha/2)}$$

we obtain

$$\left|\mathbb{E}\left(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S)\right) - \frac{2\eta_{d-1}}{d+1}\eta_d^{-1-1/d}\int_S p^{1-1/d}(s)\,\mathrm{d}s\right| \leq c_1\sqrt{\frac{\log n}{k_n}} + c_2\sqrt[d]{\frac{k_n}{n}}.$$

Setting $k_n = k_0 n^{2/(d+2)}(\log n)^{d/(d+2)}$ for any $k_0 > 0$ we have

$$\sqrt{\frac{\log n}{k_n}} = \sqrt{\frac{1}{k_0}n^{-\frac{2}{d+2}}(\log n)^{1-\frac{d}{d+2}}} = \sqrt{\frac{1}{k_0}}n^{-\frac{1}{d+2}}(\log n)^{\frac{1}{d+2}} = \sqrt{\frac{1}{k_0}}\sqrt[d+2]{\frac{\log n}{n}},$$

and

$$\sqrt[d]{\frac{k_n}{n}} = \sqrt[d]{k_0 n^{\frac{2}{d+2}-1}(\log n)^{\frac{d}{d+2}}} = \sqrt[d]{k_0 n^{-\frac{d}{d+2}}(\log n)^{\frac{d}{d+2}}} = \sqrt[d]{k_0}\sqrt[d+2]{\frac{\log n}{n}}.$$

That is, for this choice of k_n the two terms converge equally fast. Clearly, if the growth rate of k_n is faster, then the convergence of the second term is slower, whereas for a slower growth rate of k_n the convergence of the first term is slower. Therefore, the convergence rate achieved for this choice of k_n is optimal. □

3.6.2 Convergence of the variance term of cut_{n,r_n} and cut_{n,k_n}

This section considers the proof of Proposition 3.4, that is with the convergence of the variance term of cut_{n,r_n} and cut_{n,k_n}, which is the convergence of the suitably scaled random variables cut_{n,r_n} and cut_{n,k_n} to their expectations.
In the case of the kNN graph we use McDiarmid's bounded differences inequality with a kissing number argument to obtain the bounded differences condition to derive exponential decay rates for the deviation probabilities and thus convergence in probability. In the case of the r-neighborhood graph the same is achieved using Theorem A.6, that is a concentration-of-measure inequality for self-bounding functions. Almost sure convergence can be obtained using the Borel-Cantelli lemma.

Proof of Proposition 3.4. We first show the statement for the **k-nearest neighbor graph**. Let $x_1, \ldots, x_n$ be points drawn i.i.d. from our density p and let $\bar{x}_i \in \mathbb{R}^d$. Let $\mathrm{cut}^{(i)}_{n,k_n}(S)$ denote the cut induced by S in the k_n-nearest neighbor graph that is constructed on the points $x_1, \ldots, x_{i-1}, \bar{x}_i, x_{i+1}, \ldots, x_n$. The number of outgoing edges of each point x_i is k_n and according to Miller et al. [63] the number of incoming edges is bounded by $\tau_d k_n$, where τ_d denotes the kissing number in d dimensions, that is the number of unit hyperspheres in $\mathbb{R}^d$ which can touch a unit hypersphere without any intersections. Thus, changing the position of point x_i to $\bar{x}_i$ at most $k_n + 2\tau_d k_n \leq 3\tau_d k_n$ edges across the cut can change. This implies

$$\left|\mathrm{cut}_{n,k_n}(S) - \mathrm{cut}^{(i)}_{n,k_n}(S)\right| \leq 3\tau_d k_n.$$

Hence,

$$\left|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S) - \frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}^{(i)}_{n,k_n}(S)\right| \leq \frac{3\tau_d k_n}{nk_n}\sqrt[d]{\frac{n}{k_n}} = \frac{3\tau_d}{n}\sqrt[d]{\frac{n}{k_n}}.$$

Thus by McDiarmid's inequality,

$$\Pr\left(\left|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S) - \mathbb{E}\left(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S)\right)\right| > \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2}{n(\frac{3\tau_d}{n}\sqrt[d]{\frac{n}{k_n}})^2}\right) = 2\exp\left(-\frac{2\varepsilon^2 n^{1-2/d}k_n^{2/d}}{(3\tau_d)^2}\right).$$

Therefore, for every $\varepsilon > 0$ we have

$$\sum_{n=1}^{\infty}\Pr\left(\left|\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S) - \mathbb{E}\left(\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{cut}_{n,k_n}(S)\right)\right| > \varepsilon\right) < \infty,$$

if $n^{1-2/d}k_n^{2/d}/\log n \to \infty$. That is, under this condition we have almost sure convergence by Borel-Cantelli. In the case $d = 1$, the condition $k_n/\sqrt{n\log n} \to \infty$ implies that $n^{1-2/d}k_n^{2/d}/\log n = k_n^2/(n\log n) \to \infty$. In the case $d = 2$ we have $n^{1-2/d}k_n^{2/d}/\log n =$

$k_n / \log n$ and thus we have almost sure convergence if $k_n / \log n \to \infty$. For $d \geq 3$ we have $n^{1-2/d} k_n^{2/d} / \log n \geq n^{1/3} / \log n$, which certainly diverges to infinity and thus implies almost sure convergence.
Now we proof the statement for the ***r*-neighborhood graph**. For $j, l \in \{1, \ldots, n\}$, $j \neq l$ set

$$N_{j,l} = \begin{cases} 1 & \text{if } (x_j, x_l) \text{ edge in } G_{\mathrm{r}}(n, r_n) \text{ and } x_j \text{ and } x_l \text{ on different sides of } S, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$\mathrm{cut}_{n,r_n}(S) = \sum_{j=1}^{n} \sum_{\substack{l=1 \\ l \neq j}}^{n} N_{j,l}.$$

Define

$$g(x_1, \ldots, x_n) = \frac{1}{2n} \mathrm{cut}_{n,r_n}(S) = \frac{1}{2n} \sum_{j=1}^{n} \sum_{\substack{l=1 \\ l \neq j}}^{n} N_{j,l}.$$

We demonstrate that g fulfills the self-bounding property of Definition A.2 in order to apply Theorem A.6.
We have

$$g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \frac{1}{2n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{l=1 \\ l \neq i,j}}^{n} N_{j,l}.$$

Then, using the symmetry $N_{j,i} = N_{i,j}$,

$$g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \frac{1}{2n} \sum_{\substack{l=1 \\ l \neq i}}^{n} N_{i,l} + \frac{1}{2n} \sum_{\substack{j=1 \\ j \neq i}}^{n} N_{j,i} = \frac{1}{n} \sum_{\substack{l=1 \\ l \neq i}}^{n} N_{i,l}$$

and since $0 \leq N_{i,j} \leq 1$ and $n - 1 < n$,

$$0 \leq g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1.$$

Furthermore

$$\sum_{i=1}^{n} (g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\substack{l=1 \\ l \neq i}}^{n} N_{i,l} = 2g(x_1, \ldots, x_n).$$

Consequently, g is $(2, 0)$-self-bounding and we can apply the concentration-of-measure inequality for self-bounding functions in Theorem A.6. We have

$$\mathbb{E} g(x_1, \ldots, x_n) = \mathbb{E}\left(\frac{1}{2n} \mathrm{cut}_{n,r_n}(S)\right) = \frac{1}{2} n r_n^{d+1} \mathbb{E}\left(\frac{1}{n^2 r_n^{d+1}} \mathrm{cut}_{n,r_n}(S)\right).$$

According to Equation (3.15) in the proof of Proposition 3.3 and by the linearity of expectation we have

$$\mathbb{E}g(x_1,\ldots,x_n) \leq \frac{1}{2} n r_n^{d+1} \left(\left(1 + \frac{3p'_{\max}}{p_{\min}} r_n\right)^2 \frac{2\eta_{d-1}}{d+1} \int_S p^2(s)\,\mathrm{d}s + 2\eta_d p_{\max}^2 \mathcal{L}_{d-1}(S \cap \mathcal{R}_n) \right).$$

We set

$$c = \max_n \left(\left(1 + \frac{3p'_{\max}}{p_{\min}} r_n\right)^2 \frac{2\eta_{d-1}}{d+1} \int_S p^2(s)\,\mathrm{d}s + 2\eta_d p_{\max}^2 \mathcal{L}_{d-1}(S \cap \mathcal{R}_n) \right),$$

which exists since the expression on the right-hand side is monotonically decreasing in r_n and r_n converges to 0 from above. We obtain $\mathbb{E}g(x_1,\ldots,x_n) \leq cnr_n^{d+1}/2$. Applying Theorem A.6 we have for $t \leq \mathbb{E}g(x_1,\ldots,x_n)$

$$\begin{aligned}\Pr\left(|g(x_1,\ldots,x_n) - \mathbb{E}g(x_1,\ldots,x_n)| \geq t\right) &\leq 2\exp\left(-\frac{t^2}{8\mathbb{E}g(x_1,\ldots,x_n)}\right) \\ &\leq 2\exp\left(-\frac{t^2}{4cnr_n^{d+1}}\right).\end{aligned}$$

Now, for $nr_n^{d+1}\varepsilon/2 \leq \mathbb{E}g(x_1,\ldots,x_n)$, that is $\varepsilon \leq \mathbb{E}(n^{-2}r_n^{-d-1}\operatorname{cut}_{n,r_n}(S))$, we have

$$\begin{aligned}&\Pr\left(\left|\frac{1}{n^2 r_n^{d+1}}\operatorname{cut}_{n,r_n}(S) - \mathbb{E}\left(\frac{1}{n^2 r_n^{d+1}}\operatorname{cut}_{n,r_n}(S)\right)\right| \geq \varepsilon\right) \\ &= \Pr\left(\left|\frac{1}{2n}\operatorname{cut}_{n,r_n}(S) - \mathbb{E}\left(\frac{1}{2n}\operatorname{cut}_{n,r_n}(S)\right)\right| \geq nr_n^{d+1}\varepsilon/2\right) \\ &= \Pr\left(|g(x_1,\ldots,x_n) - \mathbb{E}g(x_1,\ldots,x_n)| \geq nr_n^{d+1}\varepsilon/2\right) \\ &\leq 2\exp\left(-\frac{(nr_n^{d+1}\varepsilon/2)^2}{4cnr_n^{d+1}}\right) = 2\exp\left(-\frac{nr_n^{d+1}\varepsilon^2}{16c}\right).\end{aligned}$$

Here we use that the expectation $\mathbb{E}(n^{-2}r_n^{-d-1}\operatorname{cut}_{n,r_n}(S))$ exists and converges to a positive limit for $n \to \infty$. Under the condition $nr_n^{d+1}/\log n \to \infty$ we have for every $\varepsilon > 0$

$$\sum_{n=1}^{\infty} 2\exp\left(-\frac{nr_n^{d+1}\varepsilon^2}{16c}\right) < \infty,$$

and thus we have almost sure convergence by the Borel-Cantelli lemma. □

3.6.3 Convergence of bias and variance terms for vol_{n,r_n} and vol_{n,k_n}

In the following we give the proof of Proposition 3.5 concerning the convergence of the bias term of the volume, and Proposition 3.6 concerning the convergence of the variance term of the volume.

The ideas in the proof of the convergence of the bias term include the following: In the graph $G_{\mathrm{kNN}}(n, k_n)$ there are exactly k_n outgoing edges from each node. Thus the expected number of edges originating in H depends on the number of sample points in H only, which is binomially distributed with parameters n and $\mu(H)$. For the graph $G_{\mathrm{r}}(n, r_n)$ we decompose the volume into the contributions of all the points, and for a single point we condition on its location. The number of outgoing edges, provided the point is at position x, is the number of other points in $B(x, r_n)$, which is binomially distributed with parameters $(n-1)$ and $\mu(B(x, r_n))$. If r_n is sufficiently small we can approximate $\mu(B(x, r_n))$ by $\eta_d r_n^d p(x)$ under our conditions on the density.
In order to show the convergence of the variance term of the volume we use McDiarmid's inequality for the k-nearest neighbor graph and a concentration-of-measure inequality for self-bounding functions for the r-neighborhood graph.

Proof of Proposition 3.5. First we state the proof for the ***k*-nearest neighbor graph**. Here, the expected number of points in H is $n\mu(H)$, each of them has exactly k_n outgoing edges, thus

$$\mathbb{E}(\mathrm{vol}_{n,k_n}(H)) = nk_n\mu(H).$$

Now we state the proof for the ***r*-neighborhood graph**. Let $\mathcal{E}_{n,r_n}$ denote the edges of the graph $G_{\mathrm{r}}(n, r_n)$. With

$$M_i = \begin{cases} |\{(x_i, x_j) \in \mathcal{E}_{n,r_n} \parallel j = 1, \ldots, n\}| & \text{if } x_i \in H \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\mathrm{vol}_{n,r_n}(H)) = M_1 + \ldots + M_n$$

and thus, due to the independent identical distribution of the sample point,

$$\mathbb{E}(\mathrm{vol}_{n,r_n}(H)) = n\mathbb{E}(M_1).$$

Conditioning on the position of x_1, we have

$$\begin{aligned} \mathbb{E}(\mathrm{vol}_{n,r_n}(H)) &= n \int_{\mathbb{R}^d \cap C} \mathbb{E}(M_1 \mid x_1 = x) p(x) \, \mathrm{d}x \\ &= n \int_{H \cap C} (n-1) \mu(B(x, r_n)) p(x) \, \mathrm{d}x \end{aligned}$$

and thus

$$\mathbb{E}\Big(\frac{1}{n^2 r_n^d} \mathrm{vol}_{n,r_n}(H)\Big) = \frac{n-1}{n} \frac{1}{r_n^d} \int_{H \cap C} \mu(B(x, r_n)) p(x) \, \mathrm{d}x. \tag{3.19}$$

Setting $\mathcal{R}_n = \{x \in H \cap C \mid \mathrm{dist}(x, \partial(H \cap C)) \leq r_n\}$ and $\mathcal{I}_n = (H \cap C) \setminus \mathcal{R}_n$, we have

$$\begin{aligned} &\frac{1}{r_n^d} \int_{H \cap C} \mu(B(x, r_n)) p(x) \, \mathrm{d}x \\ &\qquad = \frac{1}{r_n^d} \int_{\mathcal{R}_n} \mu(B(x, r_n)) p(x) \, \mathrm{d}x + \frac{1}{r_n^d} \int_{\mathcal{I}_n} \mu(B(x, r_n)) p(x) \, \mathrm{d}x. \end{aligned}$$

Let $x \in \mathcal{I}_n$. Under our conditions on the differentiability of p we have for all $y \in B(x, r_n)$

$$|p(y) - p(x)| \leq p'_{\max} r_n.$$

Hence, we can approximate the integral

$$\begin{aligned}
\frac{1}{r_n^d} \int_{\mathcal{I}_n} \mu(B(x, r_n)) p(x) \, \mathrm{d}x &\leq \frac{1}{r_n^d} \int_{\mathcal{I}_n} (p(x) + p'_{\max} r_n) r_n^d \eta_d p(x) \, \mathrm{d}x \\
&= \frac{1}{r_n^d} \int_{\mathcal{I}_n} p(x)^2 r_n^d \eta_d \, \mathrm{d}x + \frac{1}{r_n^d} \int_{\mathcal{I}_n} p'_{\max} r_n r_n^d \eta_d p(x) \, \mathrm{d}x \\
&\leq \eta_d \int_{\mathcal{I}_n} p(x)^2 \, \mathrm{d}x + \eta_d p'_{\max} r_n \int_{\mathcal{I}_n} p(x) \, \mathrm{d}x \\
&\leq \eta_d \int_{\mathcal{I}_n} p(x)^2 \, \mathrm{d}x + \eta_d p'_{\max} r_n.
\end{aligned}$$

Similarly we can show the lower bound, and thus

$$\Big| \frac{1}{r_n^d} \int_{\mathcal{I}_n} \mu(B(x, r_n)) p(x) \, \mathrm{d}x - \eta_d \int_{\mathcal{I}_n} p^2(x) \, \mathrm{d}x \Big| \leq \eta_d p'_{\max} r_n.$$

Now we turn to the border strip $\mathcal{R}_n$. We have

$$\begin{aligned}
&\frac{1}{r_n^d} \int_{\mathcal{R}_n} \mu(B(x, r_n)) p(x) \, \mathrm{d}x - \eta_d \int_{\mathcal{R}_n} p^2(x) \, \mathrm{d}x \\
&\quad \leq \frac{1}{r_n^d} \int_{\mathcal{R}_n} p_{\max} \eta_d r_n^d p(x) \, \mathrm{d}x - \eta_d \int_{\mathcal{R}_n} p^2(x) \, \mathrm{d}x \\
&\quad = \eta_d \int_{\mathcal{R}_n} p_{\max} p(x) \, \mathrm{d}x - \eta_d \int_{\mathcal{R}_n} p^2(x) \, \mathrm{d}x = \eta_d \int_{\mathcal{R}_n} (p_{\max} - p(x)) p(x) \, \mathrm{d}x \\
&\quad \leq \eta_d p_{\max} \int_{\mathcal{R}_n} p(x) \, \mathrm{d}x \\
&\quad \leq \eta_d p_{\max}^2 \mathcal{L}_d(\mathcal{R}_n).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{r_n^d} \int_{H \cap C} \mu(B(x, r_n)) p(x) \, \mathrm{d}x &\leq \eta_d p_{\max}^2 \mathcal{L}_d(\mathcal{R}_n) + \eta_d \int_{\mathcal{I}_n} p^2(x) \, \mathrm{d}x + \eta_d p'_{\max} r_n \\
&\leq \eta_d \int_{H \cap C} p^2(x) \, \mathrm{d}x + \eta_d p_{\max}^2 \mathcal{L}_d(\mathcal{R}_n) + \eta_d p'_{\max} r_n.
\end{aligned}$$

Inserting this in into Equation (3.19) we obtain

$$\begin{aligned}
\mathbb{E}\Big(\frac{1}{n^2 r_n^d} \operatorname{vol}_{n, r_n}(H) \Big) &\leq \frac{n-1}{n} \left(\eta_d \int_{H \cap C} p^2(x) \, \mathrm{d}x + \eta_d p_{\max}^2 \mathcal{L}_d(\mathcal{R}_n) + \eta_d p'_{\max} r_n \right) \\
&\leq \eta_d \int_{H \cap C} p^2(x) \, \mathrm{d}x + \eta_d p_{\max}^2 \mathcal{L}_d(\mathcal{R}_n) + \eta_d p'_{\max} r_n.
\end{aligned}$$

On the other hand,

$$\begin{aligned}\frac{1}{r_n^d}\int_{H\cap C}\mu(B(x,r_n))p(x)\,\mathrm{d}x &\geq \frac{1}{r_n^d}\int_{\mathcal{I}_n}\mu(B(x,r_n))p(x)\,\mathrm{d}x\\ &\geq \eta_d\int_{\mathcal{I}_n}p^2(x)\,\mathrm{d}x-\eta_d p'_{\max}r_n\\ &= \eta_d\int_{H\cap C}p^2(x)\,\mathrm{d}x-\eta_d\int_{\mathcal{R}_n}p^2(x)\,\mathrm{d}x-\eta_d p'_{\max}r_n\\ &\geq \eta_d\int_{H\cap C}p^2(x)\,\mathrm{d}x-\eta_d p_{\max}^2\mathcal{L}_d(\mathcal{R}_n)-\eta_d p'_{\max}r_n.\end{aligned}$$

Using this in Equation (3.19) we obtain

$$\begin{aligned}\mathbb{E}\Big(\frac{1}{n^2r_n^d}\operatorname{vol}_{n,r_n}(H)\Big) &\geq \frac{n-1}{n}\left(\eta_d\int_{H\cap C}p^2(x)\,\mathrm{d}x-\eta_d p_{\max}^2\mathcal{L}_d(\mathcal{R}_n)-\eta_d p'_{\max}r_n\right)\\ &\geq \eta_d\int_{H\cap C}p^2(x)\,\mathrm{d}x-\eta_d p_{\max}^2\mathcal{L}_d(\mathcal{R}_n)-\eta_d p'_{\max}r_n-\frac{\eta_d}{n}\int_{H\cap C}p^2(x)\,\mathrm{d}x\\ &\geq \eta_d\int_{H\cap C}p^2(x)\,\mathrm{d}x-\eta_d p_{\max}^2\mathcal{L}_d(\mathcal{R}_n)-\eta_d p'_{\max}r_n-\frac{\eta_d p_{\max}}{n}.\end{aligned}$$

Combining lower and upper bound we obtain

$$\left|\mathbb{E}\Big(\frac{1}{n^2r_n^d}\operatorname{vol}_{n,r_n}(H)\Big)-\eta_d\int_{H\cap C}p^2(x)\,\mathrm{d}x\right|\leq \eta_d p_{\max}^2\mathcal{L}_d(\mathcal{R}_n)+\eta_d p'_{\max}r_n+\frac{\eta_d p_{\max}}{n}.$$

For $d=1$ we clearly have $\mathcal{L}_1(\mathcal{R}_n)\leq 2\mathcal{L}_0(\partial(H\cap C))r_n$, which shows the convergence rate. Now let $d\geq 2$. Without loss of generality we assume that the sequence (r_n) is monotonically decreasing. Otherwise we choose a monotonic subsequence and use the fact that the sequence (r_n) converges to 0. Then $\mathcal{R}_1\supseteq\mathcal{R}_2\supseteq\ldots$ and therefore by the continuity of the Lebesgue measure

$$\lim_{n\to\infty}\mathcal{L}_d(\mathcal{R}_n)=\mathcal{L}_d\left(\cap_{i=1}^{\infty}\mathcal{R}_n\right)=\mathcal{L}_d\left(\partial(H\cap C)\right).$$

We have $\partial(H\cap C)\subseteq\partial H\cup\partial C$. Therefore $\mathcal{L}_d\left(\partial(H\cap C)\right)\leq\mathcal{L}_d(\partial H)+\mathcal{L}_d(\partial C)=0$, since ∂H is the hyperplane S and $\mathcal{L}_d(\partial C)=0$ by assumption. This shows the convergence of the expectation.

Now we prove the convergence rate under the rate conditions for $d\geq 2$. Due to Corollary A.17 the support of the density C can consist of only finitely many connected components. The boundary of the intersection of a connected component of C with H consists of a submanifold of ∂C, possibly with a smooth boundary $S\cap\partial C$, and, possibly, finitely many connected components of $C\cap S$, each with a smooth boundary (see Lemma A.18). Thus the boundary of $H\cap C$ is the union of finitely many connected, closed and smooth $(d-1)$-dimensional surfaces without boundaries or with smooth boundaries. For r_n sufficiently small we have with Lemma A.14

$\mathcal{L}_d(\mathcal{R}_n) \leq 3r_n\mathcal{L}_{d-1}(\partial(H \cap C))$. Consequently,

$$\left|\mathbb{E}\Big(\frac{1}{n^2 r_n^d}\operatorname{vol}_{n,r_n}(H)\Big) - \eta_d \int_{H\cap C} p^2(x)\,\mathrm{d}x\right| \leq 3\eta_d p_{\max}^2 \mathcal{L}_{d-1}(\partial(H\cap C))r_n + \eta_d p'_{\max} r_n + \frac{\eta_d p_{\max}}{n}.$$

Under the condition that $nr_n \to \infty$ for $n \to \infty$ we have

$$\left|\mathbb{E}\Big(\frac{1}{n^2 r_n^d}\operatorname{vol}_{n,r_n}(H)\Big) - \eta_d \int_{H\cap C} p^2(x)\,\mathrm{d}x\right| \leq \left(3\eta_d p_{\max}^2 \mathcal{L}_{d-1}(\partial(H\cap C)) + \eta_d p'_{\max} + \frac{\eta_d p_{\max}}{nr_n}\right) r_n,$$

and for sufficiently large n such that $nr_n \geq p_{\max}/p'_{\max}$

$$\left|\mathbb{E}\Big(\frac{1}{n^2 r_n^d}\operatorname{vol}_{n,r_n}(H)\Big) - \eta_d \int_{H\cap C} p^2(x)\,\mathrm{d}x\right| \leq \left(3p_{\max}^2 \mathcal{L}_{d-1}(\partial(H\cap C)) + 2p'_{\max}\right)\eta_d r_n.$$

Clearly $\int_{H\cap C} p^2(x)\,\mathrm{d}x = \int_H p^2(x)\,\mathrm{d}x$ since $p(x) = 0$ for $x \notin C$. □

In the following we prove Proposition 3.6, the convergence of the variance term of the volume. The statement for the k-nearest neighbor graph is proved using McDiarmid's inequality whereas the statement for the r-neighborhood graph is shown using an inequality for the concentration of self-bounding functions.

Proof of Proposition 3.6. First we state the proof for the **k-nearest neighbor graph**. Changing the position of one point will change the volume by at most k_n, thus with McDiarmid's inequality for every $\varepsilon > 0$

$$\Pr\left(\left|\frac{1}{nk_n}\operatorname{vol}_{n,k_n}(H) - \mathbb{E}\left(\frac{1}{nk_n}\operatorname{vol}_{n,k_n}(H)\right)\right| > \varepsilon\right) < 2\exp\left(-\frac{2\varepsilon^2}{n(\frac{k_n}{nk_n})^2}\right) = 2\exp\left(-2\varepsilon^2 n\right).$$

Clearly, for every $\varepsilon > 0$

$$\sum_{n=1}^{\infty} \Pr\left(\left|\frac{1}{nk_n}\operatorname{vol}_{n,k_n}(H) - \mu(H)\right| > \varepsilon\right) < \infty,$$

and therefore we have almost sure convergence by the Borel-Cantelli lemma.
Now, we state the proof for the **r-neighborhood graph**. For $j, l \in \{1,\ldots,n\}$, $j \neq l$ set

$$N_{j,l} = \begin{cases} 1 & \text{if } (x_j, x_l) \text{ edge in } G_{\mathrm{r}}(n, r_n) \text{ and } x_j \in H \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$\mathrm{vol}_{n,r_n}(H) = \sum_{j=1}^{n} \sum_{\substack{l=1 \\ l \neq j}}^{n} N_{j,l}.$$

Define

$$g(x_1, \ldots, x_n) = \frac{1}{2n} \mathrm{vol}_{n,r_n} = \frac{1}{2n} \sum_{j=1}^{n} \sum_{\substack{l=1 \\ l \neq j}}^{n} N_{j,l}.$$

We show that g fulfills the self-bounding property of Definition A.2 in order to apply Theorem A.6.
We have

$$g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \frac{1}{2n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{l=1 \\ l \neq i,j}}^{n} N_{j,l}.$$

Then, using symmetry,

$$g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \frac{1}{2n} \sum_{\substack{l=1 \\ l \neq i}}^{n} N_{i,l} + \frac{1}{2n} \sum_{\substack{j=1 \\ j \neq i}}^{n} N_{j,i} = \frac{1}{n} \sum_{\substack{l=1 \\ l \neq i}}^{n} N_{i,l},$$

and since $0 \leq N_{i,l} \leq 1$ and $n - 1 < n$,

$$0 \leq g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1.$$

Furthermore

$$\sum_{i=1}^{n} (g(x_1, \ldots, x_n) - g_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\substack{l=1 \\ l \neq i}}^{n} N_{i,l} = 2g(x_1, \ldots, x_n).$$

Consequently, g is $(2, 0)$-self-bounding and we can apply the concentration-of-measure inequality for self-bounding functions in Theorem A.6.
We have

$$\mathbb{E} g(x_1, \ldots, x_n) = \frac{1}{2n} \sum_{j=1}^{n} \sum_{\substack{l=1 \\ l \neq j}}^{n} \mathbb{E} N_{j,l} = \frac{1}{2} (n-1) \, \mathbb{E} N_{1,2} \leq \frac{1}{2} n \Pr(N_{1,2} = 1).$$

Conditioning on the location of point x_1 we have $\Pr(N_{1,2} = 1) = 0$ if $x_1 \notin H$. Otherwise $\Pr(N_{1,2} = 1) \leq p_{\max} r_n^d \eta_d$. Therefore

$$\Pr(N_{1,2} = 1) \leq p_{\max} r_n^d \eta_d \Pr(x_i \in H) = p_{\max} r_n^d \eta_d \mu(H),$$

and thus

$$\mathbb{E}g(x_1,\ldots,x_n) \leq \frac{1}{2}np_{\max}r_n^d\eta_d\mu(H) = \frac{1}{2}p_{\max}\eta_d\mu(H)nr_n^d.$$

Applying Theorem A.6 we have for $t \leq \mathbb{E}g(x_1,\ldots,x_n)$

$$\begin{aligned}\Pr\left(|g(x_1,\ldots,x_n) - \mathbb{E}g(x_1,\ldots,x_n)| \geq t\right) &\leq 2\exp\left(-\frac{t^2}{8\mathbb{E}g(x_1,\ldots,x_n)}\right)\\ &\leq 2\exp\left(-\frac{t^2}{4p_{\max}\eta_d\mu(H)nr_n^d}\right).\end{aligned}$$

Now, for $nr_n^d\varepsilon/2 \leq \mathbb{E}g(x_1,\ldots,x_n)$, that is $\varepsilon \leq \mathbb{E}(n^{-2}r_n^{-d}\operatorname{vol}_{n,r_n})$, we have

$$\begin{aligned}&\Pr\left(\left|\frac{1}{n^2r_n^d}\operatorname{vol}_{n,r_n} - \mathbb{E}\left(\frac{1}{n^2r_n^d}\operatorname{vol}_{n,r_n}\right)\right| \geq \varepsilon\right) = \Pr\left(\left|\frac{1}{2n}\operatorname{vol}_{n,r_n} - \mathbb{E}\left(\frac{1}{2n}\operatorname{vol}_{n,r_n}\right)\right| \geq nr_n^d\varepsilon/2\right)\\ &= \Pr\left(|g(x_1,\ldots,x_n) - \mathbb{E}g(x_1,\ldots,x_n)| \geq nr_n^d\varepsilon/2\right)\\ &\leq 2\exp\left(-\frac{(nr_n^d\varepsilon/2)^2}{4p_{\max}\eta_d\mu(H)nr_n^d}\right) = 2\exp\left(-\frac{nr_n^d\varepsilon^2}{16p_{\max}\eta_d\mu(H)}\right).\end{aligned}$$

Here we use that under our general assumptions the expectation $\mathbb{E}(n^{-2}r_n^{-d}\operatorname{vol}_{n,r_n})$ exists and converges to a positive limit for $n \to \infty$.
Under the condition $nr_n^d/\log n \to \infty$ we have for every $\varepsilon > 0$

$$\sum_{n=1}^{\infty} 2\exp\left(-\frac{nr_n^d\varepsilon^2}{16p_{\max}\eta_d\mu(H)}\right) < \infty,$$

and thus we have almost sure convergence by Borel-Cantelli. □

3.6.4 Convergence of bias and variance terms for card_n

Proof of Proposition 3.7. Clearly the expected number of points in H is $n\mu(H)$, thus

$$\mathbb{E}(\operatorname{card}_n(H)) = n\mu(H).$$

Changing the position of one point will change $\operatorname{card}_n(H)$ by at most 1, thus with McDiarmid's inequality

$$\Pr\left(\left|\frac{1}{n}\operatorname{card}_n(H) - \mu(H)\right| > \varepsilon\right) < 2\exp\left(-\frac{2\varepsilon^2}{n(\frac{1}{n})^2}\right) = 2\exp\left(-2\varepsilon^2 n\right).$$

Clearly, for every $\varepsilon > 0$

$$\sum_{n=1}^{\infty}\Pr\left(\left|\frac{1}{n}\operatorname{card}_n(H) - \mu(H)\right| > \varepsilon\right) < \infty,$$

and therefore we have almost sure convergence. □

3.6.5 Proofs of the main Theorems

We give a detailed proof of Theorem 3.1, which states the convergence and convergence rates for Ncut on both, the k-nearest neighbor graph and the r-neighborhood graph. Theorem 3.2 can be proven analogously and we do not present the details here.
In the proofs we first use the fact that the (properly rescaled) random variables Ncut_{n,r_n} and $\mathrm{RatioCut}_{n,r_n}$ are simple functions of the (properly rescaled) random variables cut_{n,r_n}, vol_{n,r_n} and card_n. Using Corollary A.13 we can decompose the distance of Ncut_{n,r_n} and $\mathrm{RatioCut}_{n,r_n}$ to their suspected limits into a weighted sum of the distances of cut_{n,r_n}, vol_{n,r_n} and card_n from their respective limits, where the weights depend on the limits but not on the sample size n. In a second step we decompose the latter distances into bias and variance term. Then we can apply Propositions 3.3 – 3.7 to bound these terms and find the optimal choice of r_n such that we achieve the highest almost sure convergence rate. Analogously for Ncut_{n,k_n} and $\mathrm{RatioCut}_{n,k_n}$.

Proof of Theorem 3.1. For the r-neighborhood graph we set $C_n = n^{-2} r_n^{-(d+1)} \mathrm{cut}_{n,r_n}(S)$, $C = (2\eta_{d-1}/(d+1)) \int_S p^2(s)\, \mathrm{d}s$, $V_n^+ = n^{-2} r_n^{-d} \mathrm{vol}_{n,r_n}(H^+)$, $V^+ = \eta_d \int_{H^+} p^2(x)\, \mathrm{d}x$, $V_n^- = n^{-2} r_n^{-d} \mathrm{vol}_{n,r_n}(H^-)$, $V^- = \eta_d \int_{H^-} p^2(x)\, \mathrm{d}x$. Note that we overload the notation C here. However, in this proof we do not have to refer to the support of p which was also denoted by C before. Then

$$
\begin{aligned}
&\left| \frac{1}{r_n} \mathrm{Ncut}_{n,r_n}(S) - \frac{2\eta_{d-1}}{(d+1)\eta_d} \int_S p^2(s)\, \mathrm{d}s \left(\left(\int_{H^+} p^2(x)\, \mathrm{d}x \right)^{-1} + \left(\int_{H^-} p^2(x)\, \mathrm{d}x \right)^{-1} \right) \right| \\
&\quad = \left| \frac{1}{r_n} \mathrm{Ncut}_{n,r_n}(S) - \frac{2\eta_{d-1}}{d+1} \int_S p^2(s)\, \mathrm{d}s \left(\left(\eta_d \int_{H^+} p^2(x)\, \mathrm{d}x \right)^{-1} + \left(\eta_d \int_{H^-} p^2(x)\, \mathrm{d}x \right)^{-1} \right) \right| \\
&\quad = \left| \frac{1}{n^2 r_n^{d+1}} \mathrm{cut}_{n,r_n}(S) \left(\frac{1}{\frac{1}{n^2 r_n^d} \mathrm{vol}_{n,r_n}(H^+)} + \frac{1}{\frac{1}{n^2 r_n^d} \mathrm{vol}_{n,r_n}(H^-)} \right) - C \left(\frac{1}{V^+} + \frac{1}{V^-} \right) \right| \\
&\quad = \left| C_n \left(\frac{1}{V_n^+} + \frac{1}{V_n^-} \right) - C \left(\frac{1}{V^+} + \frac{1}{V^-} \right) \right|.
\end{aligned}
$$

For the k-nearest neighbor graph we set $C_n = n^{-1+1/d} k_n^{-1-1/d} \mathrm{cut}_{n,k_n}(S)$, $C = (2\eta_{d-1}/(d+1)) \eta_d^{-1-1/d} \int_S p^{1-1/d}(s)\, \mathrm{d}s$, $V_n^+ = (nk_n)^{-1} \mathrm{vol}_{n,k_n}(H^+)$, $V^+ = \int_{H^+} p(x)\, \mathrm{d}x$, $V_n^- = (nk_n)^{-1} \mathrm{vol}_{n,k_n}(H^-)$, and $V^- = \int_{H^+} p(x)\, \mathrm{d}x$. Then

$$
\begin{aligned}
&\left| \sqrt[d]{\frac{n}{k_n}} \mathrm{Ncut}_{n,k_n}(S) - \frac{2\eta_{d-1}}{(d+1)\eta_d^{1+1/d}} \int_S p^{1-1/d}(s)\, \mathrm{d}s \left(\left(\int_{H^+} p(x)\, \mathrm{d}x \right)^{-1} + \left(\int_{H^-} p(x)\, \mathrm{d}x \right)^{-1} \right) \right| \\
&\quad = \left| \frac{1}{nk_n} \sqrt[d]{\frac{n}{k_n}} \mathrm{cut}_{n,k_n}(S) \left(\frac{1}{\frac{1}{nk_n} \mathrm{vol}_{n,k_n}(H^+)} + \frac{1}{\frac{1}{nk_n} \mathrm{vol}_{n,k_n}(H^-)} \right) - C \left(\frac{1}{V^+} + \frac{1}{V^-} \right) \right| \\
&\quad = \left| C_n \left(\frac{1}{V_n^+} + \frac{1}{V_n^-} \right) - C \left(\frac{1}{V^+} + \frac{1}{V^-} \right) \right|.
\end{aligned}
$$

Under the conditions $|C_n - C| \leq C$, $|V_n^+ - V^+| \leq V^+/2$ and $|V_n^- - V^-| \leq V^-/2$ we

have with Corollary A.13

$$\begin{aligned}
&\left|C_n\left(\frac{1}{V_n^+}+\frac{1}{V_n^-}\right)-C\left(\frac{1}{V^+}+\frac{1}{V^-}\right)\right| \\
&\quad\leq \frac{4C}{(V^+)^2}|V_n^+-V^+|+\frac{4C}{(V^-)^2}|V_n^--V^-|+\frac{V^++V^-}{V^+V^-}|C_n-C| \\
&\quad\leq \frac{4C}{(V^+)^2}\left(|V_n^+-\mathbb{E}V_n^+|+|\mathbb{E}V_n^+-V^+|\right)+\frac{4C}{(V^-)^2}\left(|V_n^--\mathbb{E}V_n^-|+|\mathbb{E}V_n^--V^-|\right) \\
&\qquad+\frac{V^++V^-}{V^+V^-}\left(|C_n-\mathbb{E}C_n|+|\mathbb{E}C_n-C|\right) \\
&\quad\leq \frac{4C}{(V^+)^2}|\mathbb{E}V_n^+-V^+|+\frac{4C}{(V^-)^2}|\mathbb{E}V_n^--V^-|+\frac{V^++V^-}{V^+V^-}|\mathbb{E}C_n-C| \\
&\qquad+\frac{4C}{(V^+)^2}|V_n^+-\mathbb{E}V_n^+|+\frac{4C}{(V^-)^2}|V_n^--\mathbb{E}V_n^-|+\frac{V^++V^-}{V^+V^-}|C_n-\mathbb{E}C_n|,
\end{aligned}$$

and the conditions above hold for $|C_n-\mathbb{E}C_n|\leq C/2$, $|\mathbb{E}C_n-C|\leq C/2$, $|V_n^+-\mathbb{E}V_n^+|\leq V^+/4$, $|\mathbb{E}V_n^+-V^+|\leq V^+/4$, and $|V_n^--\mathbb{E}V_n^-|\leq V^-/4$, $|\mathbb{E}V_n^--V^-|\leq V^-/4$.
Note that the three terms in the second to last line can be seen as bias terms, whereas the terms in the last line can be seen as variance terms, so we have effectively done a standard decomposition into bias and variance terms.
Assuming that the general conditions hold, the non-probabilistic convergence of the bias terms is shown in Propositions 3.3 and 3.5, and almost sure convergence of the variance terms is shown in Propositions 3.3 and 3.5 provided that the respective conditions on n, k_n and r_n hold.
Hence, for sufficiently large n all the terms become sufficiently small for our conditions to hold and we have

$$\left|C_n\left(\frac{1}{V_n^+}+\frac{1}{V_n^-}\right)-C\left(\frac{1}{V^+}+\frac{1}{V^-}\right)\right|\xrightarrow{a.s.}0.$$

In the following paragraphs we show the **convergence rates** for both graph types. Therefore we assume that the rate conditions hold.
First we show the convergence rate for the ***r*-neighborhood graph**. Due to Propositions 3.3 and 3.5 we can find a constant C_{bias} independent of the choice of r_n such that under the condition $nr_n\to\infty$ and for n sufficiently large $|\mathbb{E}V_n^+-V^+|\leq C_{bias}r_n$, $|\mathbb{E}V_n^--V^-|\leq C_{bias}r_n$, and $|\mathbb{E}C_n-C|\leq C_{bias}r_n$, that is all three bias terms can be bounded by $C_{bias}r_n$.
According to Proposition 3.4 we have for a constant $c>0$

$$\Pr\left(|C_n-\mathbb{E}C_n|\geq C_{bias}r_n\right)\leq 2\exp\left(-\frac{nr_n^{d+1}C_{bias}^2r_n^2}{16c}\right),$$

and according to Proposition 3.6

$$\Pr\left(|V_n^+-\mathbb{E}V_n^+|\geq C_{bias}r_n\right)\leq 2\exp\left(-\frac{nr_n^dC_{bias}^2r_n^2}{16p_{\max}\eta_d\mu(H^+)}\right),$$

and a similar term for $\Pr\left(|V_n^- - \mathbb{E}V_n^-| \geq C_{bias} r_n\right)$. Set

$$C_{rate} = 2\left(\frac{4C}{(V^+)^2} + \frac{4C}{(V^-)^2} + \frac{V^+ + V^-}{V^+V^-}\right) C_{bias}.$$

Then for n sufficiently large, and a suitable constant $\tilde{C} > 0$

$$\begin{aligned}
&\Pr\left(\left|C_n\left(\frac{1}{V_n^+} + \frac{1}{V_n^-}\right) - C\left(\frac{1}{V^+} + \frac{1}{V^-}\right)\right| \geq C_{rate} r_n\right) \\
&\quad \leq 2\exp\left(-\frac{n r_n^{d+3} C_{bias}^2}{16c}\right) + 2\exp\left(-\frac{n r_n^{d+2} C_{bias}^2}{16 p_{\max} \eta_d \mu(H^+)}\right) + 2\exp\left(-\frac{n r_n^{d+2} C_{bias}^2}{16 p_{\max} \eta_d \mu(H^-)}\right) \\
&\quad \leq 6\exp\left(-\tilde{C} n r_n^{d+3}\right).
\end{aligned}$$

Setting $r_n = r_0 \sqrt[d+3]{(\log n)/n}$ with $r_0 = \sqrt[d+3]{2/\tilde{C}}$ we have $-\tilde{C} n r_n^{d+3} = -2\log n$ and therefore

$$\sum_{n=1}^{\infty} \Pr\left(\left|C_n\left(\frac{1}{V_n^+} + \frac{1}{V_n^-}\right) - C\left(\frac{1}{V^+} + \frac{1}{V^-}\right)\right| \geq C_{rate} r_n\right) < \infty. \tag{3.20}$$

Application of the Borel-Cantelli-Lemma shows that the event

$$\left|C_n\left(\frac{1}{V_n^+} + \frac{1}{V_n^-}\right) - C\left(\frac{1}{V^+} + \frac{1}{V^-}\right)\right| \geq C_{rate} r_n$$

almost surely can occur for only finitely many $n \in \mathbb{N}$.

Note that this is the optimal convergence rate: For a faster convergence rate we would have to choose a faster convergence of r_n to 0 since the convergence of the bias terms is determined by r_n. However, the sum of probabilities in Equation (3.20) would diverge then.

Now we deal with the ***k*-nearest neighbor graph**. We remark first that we can ignore the bias terms of the volume since $|\mathbb{E}V_n^+ - V^+| = 0$ and $|\mathbb{E}V_n^- - V^-| = 0$.

According to Proposition 3.3 the optimal convergence rate for the bias term of the cut is $|\mathbb{E}C_n - C| = O(\sqrt[d+2]{(\log n)/n})$ which is achieved for $k_n = k_0 n^{2/(d+2)} (\log n)^{d/(d+2)}$ and a constant $k_0 > 0$. That is, for this choice of k_n we can find a constant C_{bias}, which may depend on k_0, such that $|\mathbb{E}C_n - C| \leq C_{bias} \sqrt[d+2]{(\log n)/n}$.

With Proposition 3.4 we have, plugging in the optimal rate for the bias term with a

factor $\tilde{C} > 0$,

$$\begin{aligned}
\Pr\left(|C_n - \mathbb{E}C_n| \geq \tilde{C}\sqrt[d+2]{\frac{\log n}{n}}\right) &\leq 2\exp\left(-\frac{2\tilde{C}^2((\log n)/n)^{2/(d+2)}n^{1-2/d}k_n^{2/d}}{(3\tau_d)^2}\right) \\
&= 2\exp\left(-\frac{2\tilde{C}^2k_0^{2/d}}{(3\tau_d)^2}(\log n)^{2/(d+2)}n^{1-2/d-2/(d+2)}n^{4/(d(d+2))}(\log n)^{2/(d+2)}\right) \\
&= 2\exp\left(-\frac{2\tilde{C}^2k_0^{2/d}}{(3\tau_d)^2}(\log n)^{4/(d+2)}n^{1-4/(d+2)}\right) \\
&\leq 2\exp\left(-\log n\,\frac{2\tilde{C}^2k_0^{2/d}}{(3\tau_d)^2}\left(\frac{n}{\log n}\right)^{1-4/(d+2)}\right).
\end{aligned}$$

For $d = 2$ we have $1 - 4/(d+2) = 0$. Setting $k_0 = ((3\tau_2)^2/\tilde{C}^2)^{d/2}$ we obtain in this case

$$\Pr\left(|C_n - \mathbb{E}C_n| \geq \tilde{C}\sqrt[4]{\frac{\log n}{n}}\right) \leq \frac{2}{n^2}.$$

For $d > 2$ we have $1 - 4/(d+2) > 0$ and thus $(n/(\log n))^{1-4/(d+2)} \to \infty$. Therefore, for $d \geq 2$ and any choice of k_0

$$\sum_{i=1}^{\infty}\Pr\left(|C_n - \mathbb{E}C_n| \geq \tilde{C}\sqrt[d+2]{\frac{\log n}{n}}\right) < \infty.$$

Applying Borel-Cantelli we obtain that the event $|C_n - \mathbb{E}C_n| \geq \tilde{C}\sqrt[d+2]{(\log n)/n}$ almost surely occurs only finitely often.
Clearly, we cannot find a better rate of k_n, since for any other rate of k_n the convergence of the bias term would become slower.
For the variance terms of the volume we have with Proposition 3.6

$$\begin{aligned}
\Pr\left(|V_n^- - \mathbb{E}V_n^-| \geq \tilde{C}\sqrt[d+2]{\frac{\log n}{n}}\right) &\leq 2\exp\left(-2\tilde{C}^2\left(\frac{\log n}{n}\right)^{2/(d+2)}n\right) \\
&= 2\exp\left(-2\tilde{C}^2(\log n)^{2/(d+2)}n^{d/(d+2)}\right),
\end{aligned}$$

which for all $d \geq 2$ implies

$$\sum_{i=1}^{\infty}\Pr\left(|V_n^- - \mathbb{E}V_n^-| \geq \tilde{C}\sqrt[d+2]{\frac{\log n}{n}}\right) < \infty.$$

With Borel-Cantelli we obtain that the event $|V_n^- - \mathbb{E}V_n^-| \geq \tilde{C}\sqrt[d+2]{(\log n)/n}$ almost surely occurs only finitely often, and similarly for the other volume term.

Combining the terms as we do in the convergence proof we obtain that we can find a $k_0 > 0$ and a constant $\hat{C}$ such that almost surely there exists an $n_0 > 0$ with

$$\left| \sqrt[d]{\frac{n}{k_n}} \operatorname{Ncut}_{n,k_n}(S) - \operatorname{NcutLim}_{\mathrm{kNN}} \right| \leq \hat{C} \sqrt[d+2]{\frac{\log n}{n}}$$

for all $n \geq n_0$.
For $d = 1$ we have according to Proposition 3.3 that we can find a constant C_{bias} not depending on k_0 such that $|\mathbb{E}C_n - C| \leq C_{bias}(\sqrt{(\log n)/k_n} + k_n/n)$. Plugging these rates in into Proposition 3.4 we obtain

$$\begin{aligned} \Pr\left(|C_n - \mathbb{E}C_n| \geq C_{bias} \sqrt{\frac{\log n}{k_n}} \right) &\leq 2 \exp\Big(- \frac{2C_{bias}^2((\log n)/k_n)\ n^{-1}k_n^2}{(3\tau_1)^2} \Big) \\ &= 2 \exp\Big(- \log n\ \frac{2C_{bias}^2}{(3\tau_1)^2} \frac{k_n}{n} \Big), \end{aligned}$$

and

$$\begin{aligned} \Pr\left(|C_n - \mathbb{E}C_n| \geq C_{bias} \frac{k_n}{n} \right) &\leq 2 \exp\Big(- \frac{2C_{bias}^2(k_n^2/n^2)\ n^{-1}k_n^2}{(3\tau_1)^2} \Big) \\ &\leq 2 \exp\Big(- \log n\ \frac{2C_{bias}^2}{(3\tau_1)^2} \frac{k_n^4}{n^3 \log n} \Big). \end{aligned}$$

Since we assume $k_n/n \to 0$ for $n \to \infty$ we cannot find a rate for k_n such that $|C_n - \mathbb{E}C_n| \geq C_{bias}\sqrt{\frac{\log n}{k_n}}$ with very low probability. However, choosing $k_n = \sqrt[4]{k_0 n^3 \log n}$ with $k_0 = (3\tau_1)^2/C_{bias}^2$ we obtain

$$\Pr\left(|C_n - \mathbb{E}C_n| \geq C_{bias} \frac{k_n}{n} \right) \leq \frac{2}{n^2}.$$

Furthermore,

$$\sqrt{\frac{\log n}{k_n}} \leq \sqrt{\frac{\log n}{\sqrt[4]{k_0 n^3 \log n}}} = k_0^{-1/8} n^{-3/8} (\log n)^{3/8} = k_0^{-1/8} \left(\frac{\log n}{n} \right)^{3/8}$$

and

$$\sqrt[d]{\frac{k_n}{n}} = \frac{k_n}{n} = \frac{\sqrt[4]{k_0 n^3 \log n}}{n} = k_0^{1/4} n^{-1/4} (\log n)^{1/4} = k_0^{1/4} \left(\frac{\log n}{n} \right)^{1/4}.$$

Therefore, we have

$$\begin{aligned} |\mathbb{E}C_n - C| &\leq C_{bias} \left(\sqrt{\frac{\log n}{k_n}} + \frac{k_n}{n} \right) = C_{bias} \left(k_0^{-1/8} \left(\frac{\log n}{n} \right)^{3/8} + k_0^{1/4} \left(\frac{\log n}{n} \right)^{1/4} \right) \\ &\leq C_{bias} \left(k_0^{-1/8} + k_0^{1/4} \right) \left(\frac{\log n}{n} \right)^{1/4} \end{aligned}$$

and for the variance term

$$\Pr\left(|C_n - \mathbb{E}C_n| \geq C_{bias} k_0^{1/4} \left(\frac{\log n}{n}\right)^{1/4}\right) \leq \frac{2}{n^2},$$

which implies

$$\sum_{n=1}^{\infty} \Pr\left(|C_n - \mathbb{E}C_n| \geq C_{bias} k_0^{1/4} \left(\frac{\log n}{n}\right)^{1/4}\right) < \infty.$$

With Borel-Cantelli this event can occur only finitely often.
This choice of k_n is optimal: If we chose a higher rate for k_n then the convergence of $|\mathbb{E}C_n - C|$ would become slower. On the other hand, the lowest rate of ε_n for which the sum of $\Pr(|C_n - \mathbb{E}C_n| \geq \varepsilon_n)$ over n converges is $\varepsilon_n \sim \sqrt{n \log n}/k_n$. Therefore, if we chose a lower rate of k_n then the rate of the variance term would get worse.
For the variance terms of the volume we have with Proposition 3.6 and for any constant $\tilde{C} > 0$

$$\begin{aligned}\Pr\left(|V_n^- - \mathbb{E}V_n^-| \geq \tilde{C}\sqrt[4]{\frac{\log n}{n}}\right) &\leq 2\exp\left(-2\tilde{C}^2\left(\frac{\log n}{n}\right)^{2/4} n\right) \\ &= 2\exp\left(-2\tilde{C}^2\sqrt{n\log n}\right).\end{aligned}$$

The sum over these events converges and therefore only finitely many of these events can occur by Borel-Cantelli.
Combining the terms for the cut and the volume we obtain the result.

□

Proof of Theorem 3.2. We have for the r-neighborhood graph

$$\frac{1}{n^2 r_n^{d+1}} \operatorname{cut}_{n,r_n}(S)\left(\frac{1}{\frac{1}{n}\operatorname{card}_n(H^+)} + \frac{1}{\frac{1}{n}\operatorname{card}_n(H^-)}\right) = \frac{1}{n r_n^{d+1}} \operatorname{RatioCut}_{n,r_n}(S),$$

and similarly for the k-nearest neighbor graph

$$\frac{1}{nk_n}\sqrt[d]{\frac{n}{k_n}} \operatorname{cut}_{n,k_n}(S)\left(\frac{1}{\frac{1}{n}\operatorname{card}_n(H^+)} + \frac{1}{\frac{1}{n}\operatorname{card}_n(H^-)}\right) = \frac{1}{k_n}\sqrt[d]{\frac{n}{k_n}} \operatorname{RatioCut}_{n,k_n}(S).$$

Analogously to the proof of Theorem 3.1 and with Proposition 3.7 instead of the corresponding statements for the volume one can show:

$$\frac{1}{n r_n^{d+1}} \operatorname{RatioCut}_{n,r_n}(S) \xrightarrow{a.s.} \frac{2\eta_{d-1}}{d+1} \int_S p^2(s)\,\mathrm{d}s \left(\left(\int_{H^+} p(s)\,\mathrm{d}s\right)^{-1} + \left(\int_{H^+} p(s)\,\mathrm{d}s\right)^{-1}\right),$$

and

$$\frac{1}{k_n}\sqrt[d]{\frac{n}{k_n}}\,\mathrm{RatioCut}_{n,k_n}(S) \xrightarrow{a.s.} \frac{2\eta_{d-1}}{(d+1)\eta_d^{1+1/d}}\int_S p^{1-1/d}(s)\,\mathrm{d}s\left(\left(\int_{H^+} p(s)\,\mathrm{d}s\right)^{-1}+\left(\int_{H^+} p(s)\,\mathrm{d}s\right)^{-1}\right),$$

and the corresponding statements about the convergence rates, where we consider that the expressions for the cut do not change and the variance term of card_n is similar to the variance term of the volume for the k-nearest neighbor graph. All the conditions on k_n in the proof of Theorem 3.1 came from the cut. □

A Mathematical Appendix

A.1 Tail bounds for sums of random variables

Basic large deviation inequalities for the binomial distribution are due to Bernstein and Chernoff [19]. These results have been generalized by Hoeffding [48] to the case of the sum of independent random variables that are bounded. However, we only present the Hoeffding bound for the case of binomial variables.

Theorem A.1 (Hoeffding, [48]) *Let $M \sim \mathrm{Bin}(n,p)$ and define $\alpha = k/n$. Then, if $\alpha \geq p$ we have*

$$\Pr(M \geq k) \leq \exp\left(-n\,K(\alpha||p)\right),$$

and for $\alpha \leq p$ we have

$$\Pr(M \leq k) \leq \exp\left(-n\,K(\alpha||p)\right),$$

where $K(\alpha||p)$ is the Kullback-Leibler divergence of $(\alpha, 1-\alpha)$ and $(p, 1-p)$,

$$K(\alpha||p) = \alpha \log\Big(\frac{\alpha}{p}\Big) + (1-\alpha)\log\Big(\frac{1-\alpha}{1-p}\Big).$$

The following tail bound, which according to Srivastav and Stangier [81] is most useful for binomial distributions with small expectations and probabilities respectively, is due to Angluin and Valiant [2].

Theorem A.2 (Angluin and Valiant [2]) *Let $M \sim \mathrm{Bin}(n,p)$ and $0 < \delta \leq 1$. Then*

$$\Pr(M > (1+\delta)np) \leq \exp\Big(-\frac{1}{3}\delta^2 np\Big),$$
$$\Pr(M < (1-\delta)np) \leq \exp\Big(-\frac{1}{2}\delta^2 np\Big).$$

Corollary A.3 *Let $M \sim \mathrm{Bin}(n,p)$ and define $\alpha = k/n$. Then,*

$$\Pr(M > k) \leq \exp\Big(-\frac{1}{3}\frac{(k-np)^2}{np}\Big) \qquad \text{for } p < \alpha \leq 2p,$$
$$\Pr(M < k) \leq \exp\Big(-\frac{1}{2}\frac{(np-k)^2}{np}\Big) \qquad \text{for } 0 \leq \alpha < p.$$

Proof. We show the proof for the first inequality, the proof for the second inequality is similar. We have

$$\Pr(M > k) = \Pr\left(M > \frac{k}{np}np\right) = \Pr\left(M > \left(1 + \left(\frac{k}{np} - 1\right)\right)np\right).$$

Using $\alpha = k/n$ and Theorem A.2 we obtain for $0 < \alpha/p - 1 \leq 1$

$$\Pr(M > k) \leq \exp\left(-\frac{1}{3}\left(\frac{\alpha}{p} - 1\right)^2 np\right).$$

The condition $0 < \alpha/p - 1 \leq 1$ translates into the condition $p < \alpha \leq 2p$. □

The following inequality for the sum of independent random variables is a classical result by Bernstein. It can be derived from Bennett's inequality published in Bennett [8] by applying an elementary inequality (refer to Lugosi [55] for details). We follow the presentation in Rao [75], where it is called "Bennett's inequality".

Theorem A.4 (Bernstein's inequality) *Let Y_i, $1 \leq i \leq n$ be i.i.d. random variables with $\mathbb{E}(Y_i) = 0$ and let $S_n = \sum_{i=1}^n Y_i$. If Y_1 takes values in $[a, b]$ with probability one and if $g = b - a$, $\sigma^2 = \mathbb{E}(Y_i^2) < \infty$, then, for all $\varepsilon > 0$*

$$\Pr\left(\left|\frac{1}{n}S_n\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{n\varepsilon^2}{2\sigma^2 + g\varepsilon}\right).$$

A.2 Concentration-of-measure inequalities

In this section we follow the presentation of the corresponding concentration-of-measure inequalities in Lugosi [55] but quote a stronger result for the concentration of self-bounding functions from McDiarmid and Reed [61]. Let $\mathcal{X}$ be a measurable space.

Definition A.1 (bounded differences property) *A function $g : \mathcal{X}^n \to \mathbb{R}$ has the bounded differences property, if for some non-negative constants $c_1, \ldots, c_n$, we have for all $1 \leq i \leq n$*

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathcal{X}} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i.$$

Theorem A.5 (McDiarmid's bounded differences inequality [60]) *Assume the function g satisfies the bounded differences property with constants $c_1, \ldots, c_n$ and that $x_1, \ldots, x_n$ are independent identically distributed random variables taking values in $\mathcal{X}$. Set $Z = g(x_1, \ldots, x_n)$. Then*

$$\Pr\left(|Z - \mathbb{E}Z| > t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Definition A.2 ((a,b)-self-bounding function) *Let $a, b \geq 0$ and $g : \mathcal{X}^n \to \mathbb{R}$ a non-negative function. Define $g_i : \mathcal{X}^{n-1} \to \mathbb{R}$ as*

$$g_i(x_1, \dots x_{i-1}, x_{i+1}, \dots, x_n) = \inf_{x' \in \mathcal{X}} g(x_1, \dots x_{i-1}, x', x_{i+1}, \dots, x_n).$$

Then the function g is (a,b)-self-bounding if for all $x_1, \dots, x_n \in \mathcal{X}$ and all $i = 1, \dots, n$

$$0 \leq g(x_1, \dots, x_n) - g_i(x_1, \dots x_{i-1}, x_{i+1}, \dots, x_n) \leq 1$$

and

$$\sum_{i=1}^{n} (g(x_1, \dots, x_n) - g_i(x_1, \dots x_{i-1}, x_{i+1}, \dots, x_n)) \leq a g(x_1, \dots, x_n) + b.$$

Theorem A.6 (McDiarmid and Reed [61]) *Assume that $g : \mathcal{X}^n \to \mathbb{R}$ is an (a,b)-self-bounding function and that $x_1, \dots, x_n$ are independent identically distributed random variables taking values in $\mathcal{X}$. Set $Z = g(x_1, \dots, x_n)$. Then for every $t > 0$*

$$\Pr\left(Z - \mathbb{E}Z \geq t\right) \leq \exp\left(-\frac{t^2}{2(a\mathbb{E}Z + b + at)}\right),$$

and

$$\Pr\left(Z - \mathbb{E}Z \leq -t\right) \leq \exp\left(-\frac{t^2}{2(a\mathbb{E}Z + b + t/3)}\right).$$

A.3 Density Estimation

We give a simple review of some facts on kernel density estimation that are used in Chapter 2. We loosely follow the presentation in Rao [75], Theorem A.7 is similar to Theorem 3.1.4 of this book, whereas our Theorem A.8 is similar to Theorem 3.1.5 in Rao [75]. We refer the reader to Silverman [79] and Devroye and Lugosi [27] for background reading on density estimation.

Definition A.3 (Kernel density estimator) *Let p be a density in $\mathbb{R}^d$ that is two times continuously differentiable with bounded derivatives and let $p(u) \leq p_{\max}$ for all $u \in \mathbb{R}^d$. Let $K : \mathbb{R}^d \to \mathbb{R}$ be a kernel function with*

- $\int_{\mathbb{R}^d} K(u)\, \mathrm{d}u = 1$
- $K(u) = K(-u)$ *for all* $u \in \mathbb{R}^d$
- $K(u) \geq 0$ *for all* $u \in \mathbb{R}^d$
- $\sup_{z \in \mathbb{R}^d} K(z) < \infty$
- $\int_{\mathbb{R}^d} \|u\|^2 K(u)\, \mathrm{d}u < \infty$

Let $x_1, \dots, x_n$ be sample points from p and let $h_n > 0$. Then the function

$$\hat{p}_n(x) = \frac{1}{nh_n^d} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h_n}\right)$$

is called the kernel density estimator with kernel K and bandwidth h_n.

All the conditions on the kernel function K hold, for example for the Gaussian kernel

$$K(x) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|x\|^2\right),$$

and the multivariate Epanechnikov kernel

$$K(x) = \begin{cases} \frac{d+2}{2\eta_d}\left(1 - \|x\|^2\right) & \text{for } \|x\| \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

which has some theoretically appealing properties.

Theorem A.7 (Bias of kernel density estimation) *Let the assumptions in Definition A.3 hold. Then there exists a constant C_1 such that for all $x \in \mathbb{R}^d$ and h_n sufficiently small*

$$|\mathbb{E}(\hat{p}_n(x)) - p(x)| \leq C_1 h_n^2.$$

Proof. We have

$$\begin{aligned} \mathbb{E}(\hat{p}_n(x)) &= \frac{1}{nh_n^d} \sum_{j=1}^{n} \int_{\mathbb{R}^d} K\left(\frac{x - y}{h_n}\right) p(y)\, \mathrm{d}y = \int_{\mathbb{R}^d} \frac{1}{h_n^d} K\left(\frac{x - y}{h_n}\right) p(y)\, \mathrm{d}y \\ &= \int_{\mathbb{R}^d} K(u)\, p(x - uh_n)\, \mathrm{d}u. \end{aligned}$$

Now we make a Taylor expansion of p around x and obtain

$$p(x - uh_n) = p(x) - \sum_{i=1}^{d} h_n u_i \frac{\partial p(x)}{\partial x_i} + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} h_n^2 u_i u_j \frac{\partial^2 p(x - \theta_u u h_n)}{\partial x_i \partial x_j}$$

with $\theta_u \in (0, 1)$ for all $u \in \mathbb{R}^d$. Clearly,

$$\int_{\mathbb{R}^d} K(u) p(x)\, \mathrm{d}u = p(x).$$

We have for all $i = 1, \dots, d$ with the substitution $v = -u$

$$\int_{\mathbb{R}^d} u_i K(u)\, \mathrm{d}u = \int_{\mathbb{R}^d} -v_i K(-v)\, \mathrm{d}v = -\int_{\mathbb{R}^d} v_i K(v)\, \mathrm{d}v$$

since $K(v) = K(-v)$ by assumptions. Therefore the integral must be 0 and we have

$$\int_{\mathbb{R}^d} K(u) \sum_{i=1}^{d} h_n u_i \frac{\partial p(x)}{\partial x_i} \, \mathrm{d}u = h_n \sum_{i=1}^{d} \frac{\partial p(x)}{\partial x_i} \int_{\mathbb{R}^d} u_i K(u) \, \mathrm{d}u = 0.$$

Thus,

$$\begin{aligned} |\mathbb{E}(\hat{p}_n(x)) - p(x)| &\leq \left| \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \int_{\mathbb{R}^d} K(u) h_n^2 u_i u_j \frac{\partial^2 p(x - \theta_u u h_n)}{\partial x_i \partial x_j} \, \mathrm{d}u \right| \\ &\leq h_n^2 \frac{d^2}{2} \sup \left\{ \left| \frac{\partial^2 p(z)}{\partial x_i \partial x_j} \right| \;\middle|\; z \in \mathbb{R}^d; i, j, k = 1, \ldots, d \right\} \int_{\mathbb{R}^d} \|u\|^2 K(u) \, \mathrm{d}u, \end{aligned}$$

where we have used $|u_i| = \sqrt{u_i^2} \leq \sqrt{u_1^2 + \ldots + u_d^2} = \|u\|$. Since the integral and the bounds for the partial derivatives exist, there is a constant $C_1 > 0$ such that

$$|\mathbb{E}(\hat{p}_n(x)) - p(x)| \leq C_1 h_n^2.$$

□

Theorem A.8 (Variance of kernel density estimation) *Let the assumptions in Definition A.3 hold and let $\varepsilon \leq p_{\max}$. Then there exists a constant $C_2 > 0$ such that for sufficiently large n for all $x \in \mathbb{R}^d$*

$$\Pr\left(|\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))| \geq \varepsilon\right) \leq 2 \exp\left(-C_2 n h_n^d \varepsilon^2\right).$$

Proof. We have

$$\begin{aligned} \hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x)) &= \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_n^d} K\left(\frac{x - x_j}{h_n}\right) - \frac{1}{n} \sum_{j=1}^{n} \int_{\mathbb{R}^d} \frac{1}{h_n^d} K\left(\frac{x - y}{h_n}\right) p(y) \, \mathrm{d}y \\ &= \frac{1}{n} \sum_{j=1}^{n} \left(\frac{1}{h_n^d} K\left(\frac{x - x_j}{h_n}\right) - \int_{\mathbb{R}^d} \frac{1}{h_n^d} K\left(\frac{x - y}{h_n}\right) p(y) \, \mathrm{d}y \right) \end{aligned}$$

Setting

$$Z_i = \frac{1}{h_n^d} K\left(\frac{x - x_i}{h_n}\right)$$

and $Y_i = Z_i - \mathbb{E}(Z_i)$, we have

$$\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x)) = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Clearly, $Y_1, \ldots, Y_n$ are i.i.d. random variables with $\mathbb{E}(Y_1) = 0$ and $|Y_1| \leq h_n^{-d} \sup_{z \in \mathbb{R}^d} K(z)$, since $Z_i, \mathbb{E}(Z_i) \in [0, h_n^{-d} \sup_{z \in \mathbb{R}^d} K(z)]$. Furthermore,

$$\begin{aligned}
\mathbb{E}(Y_1^2) &= \mathbb{E}\left((Z_1 - \mathbb{E}(Z_1))^2\right) = \mathbb{E}\left(Z_1^2 - 2Z_1\mathbb{E}(Z_1) + (\mathbb{E}(Z_1))^2\right) = \mathbb{E}\left(Z_1^2\right) - (\mathbb{E}(Z_1))^2 \\
&\leq \mathbb{E}\left(Z_1^2\right) = \int_{\mathbb{R}^d} \left(\frac{1}{h_n^d} K\left(\frac{x-y}{h_n}\right)\right)^2 p(y)\, \mathrm{d}y \\
&\leq \frac{1}{h_n^d} \sup_{z \in \mathbb{R}^d} K(z) \int_{\mathbb{R}^d} \frac{1}{h_n^d} K\left(\frac{x-y}{h_n}\right) p(y)\, \mathrm{d}y \\
&= \frac{1}{h_n^d} \sup_{z \in \mathbb{R}^d} K(z) \mathbb{E}(\hat{p}_n(x)).
\end{aligned}$$

Applying Theorem A.4, that is Bernstein's inequality, we obtain

$$\begin{aligned}
&\Pr\left(|\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))| \geq \varepsilon\right) \\
&\quad \leq 2\exp\left(-\frac{n\varepsilon^2}{2\frac{1}{h_n^d} \sup_{z \in \mathbb{R}^d} K(z)\mathbb{E}(\hat{p}_n(x)) + h_n^{-d} \sup_{z \in \mathbb{R}^d} K(z)\varepsilon}\right) \\
&\quad = 2\exp\left(-\frac{nh_n^d\varepsilon^2}{2\sup_{z \in \mathbb{R}^d} K(z)\mathbb{E}(\hat{p}_n(x)) + \sup_{z \in \mathbb{R}^d} K(z)\varepsilon}\right)
\end{aligned}$$

For sufficiently large n we have due to Theorem A.7 $\mathbb{E}(\hat{p}_n(x)) \leq 2p_{\max}$, and thus for $\varepsilon \leq p_{\max}$

$$\Pr\left(|\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))| \geq \varepsilon\right) \leq 2\exp\left(-\frac{nh_n^d\varepsilon^2}{5p_{\max}\sup_{z \in \mathbb{R}^d} K(z)}\right)$$

Setting

$$C_2 = \frac{1}{5p_{\max}\sup_{z \in \mathbb{R}^d} K(z)},$$

we obtain the result. □

Corollary A.9 *Let the assumptions of Definition A.3 hold. For a sequence h_n let $\hat{p}_n$ be the kernel density estimator with bandwidth h_n and let ε_n be a sequence in $\mathbb{R}_{>0}$. Define the event*

$$\mathcal{D}_n = \{|\hat{p}_n(x_i) - p(x_i)| \leq \varepsilon_n \mid i = 1, \ldots, n\}.$$

If $\varepsilon_n = \varepsilon > 0$ is fixed there exist constants $\tilde{C}_1$ and $\tilde{C}_2$ such that setting the bandwidth $h_n = h \leq \tilde{C}_1\sqrt{\varepsilon}$ we obtain for sufficiently large n

$$\Pr\left(\mathcal{D}_n^c\right) \leq \exp\left(-\tilde{C}_2 n h^d \varepsilon^2\right).$$

If we set $\varepsilon_n = \varepsilon_0(\log n/n)^{2/(d+4)}$ *and* $h_n = h_0(\log n/n)^{1/(d+4)}$ *for suitable constants* $\varepsilon_0, h_0 > 0$*, then*

$$\sum_{i=1}^{\infty} \Pr\left(\mathcal{D}_n^c\right) < \infty.$$

Proof. For an arbitrary point $x \in \mathbb{R}^d$ we can split the deviation of the density estimator from the true density into bias and variance term:

$$|\hat{p}_n(x) - p(x)| \leq |\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))| + |\mathbb{E}(\hat{p}_n(x)) - p(x)|.$$

In the following we use the fact that the left term must be bounded by ε_n if both the terms on the right hand side are bounded by $\varepsilon_n/2$.
According to Theorem A.7 there exists a constant C_1 such that for sufficiently large n we have $|\mathbb{E}(\hat{p}_n(x)) - p(x)| \leq C_1 h_n^2$. That is, if $h_n^2 \leq \varepsilon_n/(2C_1)$ we have $|\mathbb{E}(\hat{p}_n(x)) - p(x)| \leq \varepsilon_n/2$ for all $x \in \mathbb{R}^d$ and for sufficiently large n.
According to Theorem A.8 there exists a constant C_2 such that for all $x \in \mathbb{R}^d$ and sufficiently large n

$$\Pr\left(|\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))| \geq \frac{\varepsilon_n}{2}\right) \leq 2\exp\left(-C_2\frac{nh_n^d\varepsilon_n^2}{4}\right).$$

Therefore, under the condition $h_n^2 \leq \varepsilon_n/(2C_1)$, for sufficiently large n

$$\Pr\left(\mathcal{D}_n^c\right) \leq 2n\exp\left(-C_2\frac{nh_n^d\varepsilon_n^2}{4}\right) \leq \exp\left(\log 2 + \log n - \frac{C_2}{4}nh_n^d\varepsilon_n^2\right).$$

If $\varepsilon_n = \varepsilon$ is fixed, we set $h_n = h$ for $0 < h \leq \sqrt{\varepsilon/(2C_1)}$ and $\tilde{C}_2 = C_2/8$ which implies $\Pr\left(\mathcal{D}_n^c\right) \leq \exp\left(-\tilde{C}_2 nh^d\varepsilon^2\right)$ for sufficiently large n.
In the other case let $h_0 > 0$ be arbitrary. We have,

$$h_n^2 = h_0^2\left(\frac{\log n}{n}\right)^{2/(d+4)} \leq \frac{h_0^2}{\varepsilon_0}\varepsilon_0\left(\frac{\log n}{n}\right)^{2/(d+4)} = \frac{h_0^2}{\varepsilon_0}\varepsilon_n,$$

that is, if we choose $\varepsilon_0 \geq 2C_1h_0^2$, the condition $h_n^2 \leq \varepsilon_n/(2C_1)$ holds for all $n \in \mathbb{N}$. We have

$$nh_n^d\varepsilon_n^2 = h_0^d\varepsilon_0^2 n\left(\frac{\log n}{n}\right)^{\frac{d}{d+4}}\left(\frac{\log n}{n}\right)^{\frac{4}{d+4}} = h_0^d\varepsilon_0^2\log n,$$

and thus for n sufficiently large,

$$\Pr\left(\mathcal{D}_n^c\right) \leq \exp\left(\log 2 + \log n - \frac{C_2}{4}h_0^d\varepsilon_0^2\log n\right) \leq \exp\left(\log n\left(2 - \frac{C_2}{4}h_0^d\varepsilon_0^2\right)\right).$$

That is, for $\varepsilon_0 \geq 4/\sqrt{C_2h_0^d}$, we have $\Pr\left(\mathcal{D}_n^c\right) \leq 1/n^2$ for sufficiently large n, which implies with Borel-Cantelli that $\mathcal{D}_n$ occurs almost surely for all but finitely many n. □

A.4 Inequalities to show Convergence

The following inequalities are used in the proof of Theorem 3.1 and 3.2 to show that the convergence of cut_{n,k_n}, vol_{n,k_n} and card_n implies the convergence of Ncut_{n,k_n} and $\text{RatioCut}_{n,k_n}$ and to show the convergence rate (and similarly for the corresponding quantities of the r-neighborhood graph). The result we will use in the end is Corollary A.13, which is proved using the Lemmas A.10 – A.12.

Lemma A.10 *Let $A, B, a, b \in \mathbb{R}$. Then we have*

$$|A + B - (a + b)| \leq |A - a| + |B - b|.$$

Proof. With the triangle inequality we have

$$|A + B - (a + b)| = |A - a + B - b| \leq |A - a| + |B - b|.$$

□

Lemma A.11 *Let $A, B, a, b \geq 0$. If $|A - a| \leq a$ then*

$$|AB - ab| \leq 2|B - b|a + |A - a|b.$$

Proof. Under the conditions above we have with the triangle inequality

$$\begin{aligned} |AB - ab| &= |(A - a + a)(B - b + b) - ab| \\ &= |(A - a)(B - b) + (A - a)b + a(B - b) + ab - ab| \\ &\leq |A - a||B - b| + |A - a|b + |B - b|a \\ &= |B - b|\,(a + |A - a|) + |A - a|b \\ &\leq 2|B - b|a + |A - a|b. \end{aligned}$$

□

Lemma A.12 *Let $A, a > 0$. If $|A - a| \leq a/2$ we have*

$$\left|\frac{1}{A} - \frac{1}{a}\right| \leq \frac{2|A - a|}{a^2}.$$

Proof. Under the condition $|A - a| \leq a/2$ we have

$$\begin{aligned} \left|\frac{1}{A} - \frac{1}{a}\right| = \left|\frac{a - A}{aA}\right| = \frac{|A - a|}{|a(A - a + a)|} = \frac{|A - a|}{|a(A - a) + a^2|} \leq \frac{|A - a|}{-a|A - a| + a^2} \\ \leq \frac{|A - a|}{a(a - |A - a|)} \leq \frac{2|A - a|}{a^2}. \end{aligned}$$

□

Corollary A.13 *Let $A, B_1, B_2, a, b_1, b_2 > 0$. If $|A-a| \leq a$, $|B_1 - b_1| \leq b_1/2$ and $|B_2 - b_2| \leq b_2/2$ we have*

$$\left| A\left(\frac{1}{B_1}+\frac{1}{B_2}\right) - a\left(\frac{1}{b_1}+\frac{1}{b_2}\right)\right| \leq \frac{4a}{b_1^2}|B_1-b_1| + \frac{4a}{b_2^2}|B_2-b_2| + \frac{b_1+b_2}{b_1 b_2}|A-a|.$$

Proof. If $|A-a| \leq a$ we have with Lemma A.11

$$\begin{aligned}\left| A\left(\frac{1}{B_1}+\frac{1}{B_2}\right) - a\left(\frac{1}{b_1}+\frac{1}{b_2}\right)\right| &\leq 2a\left|\frac{1}{B_1}+\frac{1}{B_2}-\left(\frac{1}{b_1}+\frac{1}{b_2}\right)\right| + |A-a|\left(\frac{1}{b_1}+\frac{1}{b_2}\right)\\ &\leq 2a\left|\frac{1}{B_1}-\frac{1}{b_1}\right| + 2a\left|\frac{1}{B_2}-\frac{1}{b_2}\right| + |A-a|\left(\frac{1}{b_1}+\frac{1}{b_2}\right)\end{aligned}$$

and for $|B_1 - b_1| \leq b_1/2$ and $|B_2 - b_2| \leq b_2/2$ with Lemma A.12

$$\leq \frac{4a}{b_1^2}|B_1-b_1| + \frac{4a}{b_2^2}|B_2-b_2| + \frac{b_1+b_2}{b_1 b_2}|A-a|.$$

□

A.5 Properties of hypersurfaces

Theorem A.14 (Volume of $\{x \mid \operatorname{dist}(x,S) < r\}$ for hypersurfaces S) *Let $d \geq 2$ and let S be a finite union of closed smooth hypersurfaces in $\mathbb{R}^d$ without boundaries or with smooth boundaries. Then for every constant $C > 0$ there exists $r_0 > 0$ such that for $r \leq r_0$*

$$\mathcal{L}_d\left(\{x \mid \operatorname{dist}(x,S) < r\}\right) \leq (2+C) r \mathcal{L}_{d-1}(S).$$

Proof. According to Steffen [83] a set that can be covered by a finite or countable union of smooth $(d-1)$-dimensional submanifolds in $\mathbb{R}^d$ is $(d-1)$-rectifiable. Since S is closed we have with Theorem 3.2.39 in Federer [32]

$$\lim_{r \to 0+} \frac{\mathcal{L}_d(\{x \mid \operatorname{dist}(x,S) < r\})}{2r} = \mathcal{H}_{d-1}(S),$$

where $\mathcal{H}_{d-1}$ denotes the $(d-1)$-dimensional Hausdorff measure. Under the conditions on S we have $\mathcal{H}_{d-1}(S) = \mathcal{L}_{d-1}(S)$. Thus, for every constant $C > 0$ there exists $r_0 > 0$ such that for all $r < r_0$

$$\frac{\mathcal{L}_d(\{x \mid \operatorname{dist}(x,S) < r\})}{2r} \leq \mathcal{L}_{d-1}(S) + \frac{C}{2}\mathcal{L}_{d-1}(S),$$

and thus $\mathcal{L}_d(\{x \mid \operatorname{dist}(x,S) < r\}) \leq (2+C) r \mathcal{L}_{d-1}(S)$. □

Lemma A.15 (Change of normal and geodesic distance) *Let $d \geq 2$, ∂C be a smooth hypersurface in Euclidean space $\mathbb{R}^d$ with minimal curvature radius κ and let N denote its normal vector field. Furthermore, for $x, y \in \partial C$ let $\mathrm{dist}_{\partial C}(x, y)$ denote the geodesic distance in ∂C of x and y. Then for any $x, y \in \partial C$*

$$\|N(x) - N(y)\| \leq \frac{\mathrm{dist}_{\partial C}(x, y)}{\kappa}.$$

Proof. Let γ be a unit speed geodesic in ∂C connecting x and y, that means $\gamma : [0, \mathrm{dist}_{\partial C}(x, y)] \to \partial C$, $\gamma(0) = x$, $\gamma(\mathrm{dist}_{\partial C}(x, y)) = y$ and $\|\dot{\gamma}(t)\| = 1$ for all $t \in [0, \mathrm{dist}_{\partial C}(x, y)]$. Using the Weingarten equation for Euclidean hypersurfaces (see Lee [54]), we obtain

$$N(x) - N(y) = \int_0^{\mathrm{dist}_{\partial C}(x,y)} \nabla_{\dot{\gamma}(t)} N(\gamma(t)) \, \mathrm{d}t = \int_0^{\mathrm{dist}_{\partial C}(x,y)} -s_{\gamma(t)} \, \dot{\gamma}(t) \, \mathrm{d}t,$$

where $\nabla_{\dot{\gamma}(t)} N(\gamma(t))$ denotes the derivative of N at the point $\gamma(t)$ in the direction of $\dot{\gamma}(t)$ and $s_{\gamma(t)}$ denotes the shape operator of ∂C at the point $\gamma(t)$ (see the remark after the definition of the minimal curvature radius in Section 1.4 for the geometric interpretation of the shape operator). Therefore, by the triangle inequality for integrals

$$\|N(x) - N(y)\| \leq \int_0^{\mathrm{dist}_{\partial C}(x,y)} \|s_{\gamma(t)} \, \dot{\gamma}(t)\| \, \mathrm{d}t.$$

The shape operator $s_{\gamma(t)}$ at any point $\gamma(t)$ is a self-adjoint linear transformation of the tangent space at $\gamma(t)$. Let $e_1, \ldots, e_{d-1}$ denote the eigenvalues of $s_{\gamma(t)}$. By the definition of the minimal curvature radius $\kappa \leq \min_{i=1,\ldots,d-1} |1/e_i|$. Since $\dot{\gamma}(t)$ is a unit vector in the tangent space at $\gamma(t)$ we have $\|s_{\gamma(t)} \, \dot{\gamma}(t)\| \leq 1/\kappa$. Consequently

$$\|N(x) - N(y)\| \leq \frac{1}{\kappa} \int_0^{\mathrm{dist}_{\partial C}(x,y)} \mathrm{d}t = \frac{\mathrm{dist}_{\partial C}(x, y)}{\kappa}.$$

□

Lemma A.16 (Ball of certain radius $\tilde{\kappa}$ contained in C) *Let $d \geq 2$ and let C be a compact set in $\mathbb{R}^d$ whose boundary ∂C is a smooth compact $(d-1)$-dimensional submanifold of $\mathbb{R}^d$ with minimal curvature radius $\kappa > 0$. For $x \in \partial C$ let n_x denote the normal to ∂C in x pointing towards the interior of C. Then we can find a radius $\tilde{\kappa} > 0$ such that for any point $x \in \partial C$ we have $B(x + \tilde{\kappa} n_x, \tilde{\kappa}) \subseteq C$.*

Proof. For any $x \in \partial C$ we define $O(x) = \{y \in \partial C \mid \mathrm{dist}_{\partial C}(x, y) < \pi\kappa\}$ and the function

$$\rho(x) = \inf_{y \in \partial C \setminus O(x)} \mathrm{dist}(x, y)$$

(compare also the construction in the proof of Lemma 2.22 in Hein [45]). Since $\partial C \setminus O(x)$ is compact and the Euclidean distance is continuous we can find a y for every x such

that $\operatorname{dist}(x,y) = \rho(x)$. Suppose that we can find a sequence $x_1, x_2, \ldots$ with $\rho(x_n) \to 0$ for $n \to \infty$. Then we can find a corresponding sequence $y_1, y_2, \ldots$ with $y_i \in \partial C \setminus O(x_i)$ and $\operatorname{dist}(x_i, y_i) = \rho(x_i)$. Both are sequences in the compact hypersurface ∂C and thus we can find $x, y \in \partial C$ and subsequences x_{i_l} and y_{i_l} that converge to x and y for $l \to \infty$, that is $\operatorname{dist}_{\partial C}(x, x_{i_l}) \to 0$ and $\operatorname{dist}_{\partial C}(y, y_{i_l}) \to 0$ for $l \to \infty$. We have

$$\begin{aligned}\operatorname{dist}(x,y) &\leq \operatorname{dist}(x, x_{i_l}) + \operatorname{dist}(x_{i_l}, y_{i_l}) + \operatorname{dist}(y_{i_l}, y) \\ &\leq \operatorname{dist}_{\partial C}(x, x_{i_l}) + \operatorname{dist}(x_{i_l}, y_{i_l}) + \operatorname{dist}_{\partial C}(y_{i_l}, y) \to 0\end{aligned}$$

for $l \to \infty$, which implies $x = y$ and therefore $\operatorname{dist}_{\partial C}(x,y) = 0$. But then

$$\begin{aligned}\operatorname{dist}_{\partial C}(x_{i_l}, y_{i_l}) &\leq \operatorname{dist}_{\partial C}(x_{i_l}, x) + \operatorname{dist}_{\partial C}(x,y) + \operatorname{dist}_{\partial C}(y, y_{i_l}) \\ &= \operatorname{dist}_{\partial C}(x_{i_l}, x) + \operatorname{dist}_{\partial C}(y, y_{i_l}) \to 0\end{aligned}$$

for $l \to \infty$. This is a contradiction to the fact that $\operatorname{dist}_{\partial C}(x_{i_l}, y_{i_l}) \geq \pi\kappa$ since $y_{i_l} \notin O(x_{i_l})$. Thus $\rho(x)$ must be bounded away from 0 on ∂C, that is, there exists $\rho_{\min} > 0$ with $\rho(x) \geq \rho_{\min}$ for all $x \in \partial C$.
Set $\tilde{\kappa} = \min\{\kappa, \rho_{\min}/2\}$. Let $x \in \partial C$ and remember that the normal n_x points towards the interior of C. Then $B(x + \tilde{\kappa} n_x, \tilde{\kappa}) \subseteq C$, since the interior of this ball cannot contain any point from ∂C: A point $y \in \partial C$ with $\operatorname{dist}_{\partial C}(x,y) \leq \pi\kappa$ cannot be in the interior of $B(x + \kappa n_x, \kappa) \subseteq C$ due to the curvature constraints but $B(x + \tilde{\kappa} n_x, \tilde{\kappa}) \subseteq B(x + \kappa n_x, \kappa)$. For $y \in \partial C$ with $\operatorname{dist}_{\partial C}(x,y) \geq \pi\kappa$ we have $\operatorname{dist}(x,y) \geq \rho_{\min}$, and thus y cannot be in the interior of $B(x + \tilde{\kappa} n_x, \tilde{\kappa})$ since $\tilde{\kappa} \leq \rho_{\min}/2$. Thus the interior of $B(x + \tilde{\kappa} n_x, \tilde{\kappa})$ is a subset of C; since C is closed we also have $B(x + \tilde{\kappa} n_x, \tilde{\kappa}) \subseteq C$. □

Corollary A.17 (Minimum volume of connected components of C) *Let the conditions and notations of Lemma A.16 hold. Then for any connected component $G \subseteq C$ we have $\mathcal{L}_d(G) \geq \tilde{\kappa}^d \eta_d$.*

Lemma A.18 (Finitely many surfaces with smooth boundary in $S \cap C$) *Let the conditions and notations of Lemma A.16 hold and let S be a hyperplane in $\mathbb{R}^d$ with normal n_S. Suppose further that for an $\alpha \in (0, \pi/2)$ and all $x \in S \cap \partial C$ we have $|\langle n_S, n_x \rangle| \leq \cos(\alpha)$. Then $S \cap C$ consists of finitely many connected surfaces with a smooth $(d-2)$-dimensional boundary.*

Proof. Let G be a connected component of $S \cap C$ and let $x \in G \cap \partial C$, that is, x is a point on the relative boundary of G. Let $\beta \in [\alpha, \pi/2)$ such that $|\langle n_S, n_x \rangle| = \cos(\beta)$. According to Lemma A.16 there exists a $\tilde{\kappa} > 0$, which is independent of x, such that $B(x + \tilde{\kappa} n_x, \tilde{\kappa}) \subseteq C$. Since $x \in S \cap B(x + \tilde{\kappa} n_x, \tilde{\kappa})$ and the intersection is connected $S \cap B(x + \tilde{\kappa} n_x, \tilde{\kappa}) \subseteq G$. We have $\operatorname{dist}(x + \tilde{\kappa} n_x, S) = |\langle \tilde{\kappa} n_x, n_S \rangle| = \tilde{\kappa} \cos(\beta) < \tilde{\kappa}$. Thus $S \cap B(x + \tilde{\kappa} n_x, \tilde{\kappa})$ is a $(d-1)$-dimensional unit ball with radius $\sqrt{\tilde{\kappa}^2 - \tilde{\kappa}^2 \cos^2(\beta)} = \tilde{\kappa} \sin(\beta)$. Therefore

$$\mathcal{L}_{d-1}(G) \geq \mathcal{L}_{d-1}(S \cap B(x + \tilde{\kappa} n_x, \tilde{\kappa})) \geq \tilde{\kappa}^{d-1} \sin^{d-1}(\beta)\, \eta_{d-1} \geq \tilde{\kappa}^{d-1} \sin^{d-1}(\alpha)\, \eta_{d-1}.$$

Since C is compact $\mathcal{L}_{d-1}(S \cap C)$ is finite; the $(d-1)$-dimensional area $\mathcal{L}_{d-1}(G)$ of each connected component is bounded away from zero. Thus there can exist only finitely many connected components.
The smoothness of the relative boundary of $\partial C \cap S$ in x can be shown with the implicit function theorem: Since $x \in \partial C$ we can find a tangent hyperplane T_x with normal n_x. Due to $|\langle n_S, n_x \rangle| = \cos(\beta)$ with $\beta > 0$ we have $n_S \neq n_x$ and the intersection $S \cap T_x$ is a $(d-2)$-dimensional affine subspace.
We choose an orthonormal coordinate system in T_x with its origin at x and basis vectors $u_1, \ldots, u_{d-1}$, such that $u_1, \ldots, u_{d-2}$ span $S \cap T_x$. Then $u_1, \ldots, u_{d-1}, n_x$ is an orthonormal basis of $\mathbb{R}^d$. There must be a representation $n_S = \langle n_S, n_x \rangle n_x + \langle n_S, u_{d-1} \rangle u_{d-1}$. Since all the vectors are unit vectors and $|\langle n_S, n_x \rangle| = \cos(\beta)$ we conclude that $|\langle n_S, u_{d-1} \rangle| = \sin(\beta)$. We assume without loss of generality that $\langle n_S, u_{d-1} \rangle = \sin(\beta)$.
In the above basis of $\mathbb{R}^d$ we have $S = \{v = (v_1, \ldots, v_d) \in \mathbb{R}^d \mid \langle n_S, v \rangle = 0\}$, but $\langle n_S, v \rangle = v_{d-1} \langle n_S, u_{d-1} \rangle + v_d \langle n_S, n_x \rangle = v_{d-1} \sin(\beta) + v_d \cos(\beta)$. Thus $S = \{v = (v_1, \ldots, v_d) \in \mathbb{R}^d \mid v_d = -v_{d-1} \tan(\beta)\}$ and the function $f_S : \mathbb{R}^{d-1} \to \mathbb{R}$ with $f_S(v_1, \ldots, v_{d-1}) = -v_{d-1} \tan(\beta)$ is a representation of the hyperplane S.
Due to the smoothness of ∂C there exists an open set $W \subseteq \mathbb{R}^{d-1}$ and a smooth function $f_{\partial C} : W \to \mathbb{R}$ with $\nabla f_{\partial C}(0) = 0$ that is locally a representation of ∂C.
Now we consider the difference of the two functions. Setting $f = f_{\partial C} - f_S$ we have $f(0) = 0$ and

$$\frac{\partial f(0)}{\partial u_{d-1}} = \frac{\partial f_{\partial C}(0)}{\partial u_{d-1}} - \frac{\partial f_S(0)}{\partial u_{d-1}} = -\frac{\partial f_S(0)}{\partial u_{d-1}} = \tan(\beta) \neq 0.$$

Furthermore, f has the same differentiability properties as $f_{\partial C}$, since f_S is infinitely differentiable. With the implicit function theorem we conclude that there exist open sets $U = U(0) \subseteq \mathbb{R}^{d-2}$, $V = V(0) \subseteq \mathbb{R}$ and a function $g : U \to V$ (with the same differentiability properties as f) such that $f(y, g(y)) = 0$ and for all $(y, z) \in U \times V \subseteq W$ $f(y, z) \neq 0$ if $z \neq g(y)$. Furthermore

$$\nabla g(0) = -\left(\frac{\partial f(0)}{\partial u_{d-1}}\right)^{-1} \left(\frac{\partial f(0)}{\partial u_1} \cdots \frac{\partial f(0)}{\partial u_{d-2}}\right) = -\tan(\beta) \left(\frac{\partial f(0)}{\partial u_1} \cdots \frac{\partial f(0)}{\partial u_{d-2}}\right) = 0.$$

By construction of f locally the intersection $\partial C \cap S$ are the points $(y, g(y), -g(y) \tan(\beta))$ with $y \in U$, $g(y) \in V$.
Setting $u'_{d-1} = \cos(\beta) u_{d-1} - \sin(\beta) n_x$ we have

$$\langle u'_{d-1}, n_S \rangle = \cos(\beta) \langle u_{d-1}, n_S \rangle - \sin(\beta) \langle n_x, n_S \rangle = 0,$$

and clearly $\|u'_{d-1}\| = 1$ and $\langle u'_{d-1}, u_i \rangle = 0$ for $i = 1, \ldots, d-2$. Furthermore,

$$\begin{aligned} n_x &= \langle n_x, n_S \rangle n_S + \langle n_x, u'_{d-1} \rangle u'_{d-1} = \cos(\beta) n_S + \langle n_x, \cos(\beta) u_{d-1} - \sin(\beta) n_x \rangle u'_{d-1} \\ &= \cos(\beta) n_S - \sin(\beta) u'_{d-1}, \end{aligned}$$

and

$$u_{d-1} = \langle u_{d-1}, n_S \rangle n_S + \langle u_{d-1}, u'_{d-1} \rangle u'_{d-1} = \sin(\beta) n_S + \cos(\beta) u'_{d-1}.$$

Thus, for $y \in U \subseteq S \cap T_x$

$$\begin{aligned} y + g(y)u_{d-1} - g(y)\tan(\beta)n_x &= y + g(y)\left(u_{d-1} - \tan(\beta)n_x\right) \\ &= y + g(y)\left(\sin(\beta)n_S + \cos(\beta)u'_{d-1} - \tan(\beta)\left(\cos(\beta)n_S - \sin(\beta)u'_{d-1}\right)\right) \\ &= y + g(y)u'_{d-1}\left(\cos(\beta) + \tan(\beta)\sin(\beta)\right) = y + \frac{1}{\cos(\beta)}g(y)u'_{d-1}. \end{aligned}$$

That is, in the orthonormal coordinate system of S with its origin at x and unit vectors $u_1, \ldots, u_{d-2}, u'_{d-1}$ we can represent $S \cap \partial C$ locally as the graph of the function $\tilde{g} : \mathbb{R}^{d-2} \to \mathbb{R}$ defined by $\tilde{g}(v_1, \ldots, v_{d-2}) = g(v_1, \ldots, v_{d-2})/\cos(\beta)$. Furthermore $\nabla\tilde{g}(0) = 0$, that is $S \cap T_x$ is tangential to $S \cap \partial C$ in x. □

A.6 Upper bound on η_d

Lemma A.19 *For all $d \geq 1$ we have $\eta_d \leq 6$.*

Proof. We have

$$\eta_d = \frac{\pi^{d/2}}{\Gamma(1 + d/2)} = \frac{\pi^{(d-2)/2}}{\Gamma(1 + (d-2)/2)}\frac{\pi}{d/2} = \frac{\pi^{(d-2)/2}}{\Gamma(1 + (d-2)/2)}\frac{2\pi}{d} = \eta_{d-2}\frac{2\pi}{d}.$$

So for $d < 7$ we have $\eta_d > \eta_{d-2}$, whereas for $d \geq 7$ we have $\eta_d < \eta_{d-2}$. Therefore $\eta_5 = 8\pi^2/15 \approx 5.26$ is the maximum η_d with uneven d and $\eta_6 = \pi^3/6 \approx 5.17$ is the maximum η_d with even d. □

List of basic notations used in the text

Here we present a list of the notations that are used in this thesis and that are of more than local significance. First, the general mathematical notations that are used throughout the thesis are presented, followed by more special notations in the single chapters.

$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}_{\geq 0}$	set of non-negative real numbers
$\mathbb{R}_{>0}$	set of positive real numbers
$\lvert A\rvert$	cardinality of the set A
$A \cap B$	intersection of the sets A and B
$A \cup B$	union of the sets A and B
A^c	complement of the set A
$\mathbf{1}$	vector with all entries 1, that is $\mathbf{1} = (1, \ldots, 1)'$
$\operatorname{diag}(v_1, \ldots, v_n)$	diagonal matrix with entries $v_1, \ldots, v_n$ on the diagonal
$\mathbb{1}_A(x)$	indicator function of the set A, that is $\mathbb{1}_A(x) = 1$ if $x \in A$, otherwise 0
$\langle x_1, x_2 \rangle$	Euclidean dot product of $x_1, x_2 \in \mathbb{R}^d$
$\lVert x\rVert$	Euclidean norm of $x \in \mathbb{R}^d$, i.e. $\lVert x\rVert = \sqrt{\langle x, x\rangle}$
$\lvert a\rvert$	absolute value of $a \in \mathbb{R}$
$\operatorname{dist}(x_1, x_2)$	Euclidean distance between $x_1, x_2 \in \mathbb{R}^d$, i.e. $\operatorname{dist}(x_1, x_2) = \lVert x_1 - x_2\rVert$
$\operatorname{dist}_{\partial C}(x, y)$	geodesic distance in ∂C of x and y
∂C	boundary of the set C
$\mathcal{L}$	the Lebesgue volume
$\mathcal{L}_{d-1}$	the $(d-1)$-dimensional Lebesgue measure in a $(d-1)$-dimensional affine subspace or the $(d-1)$-dimensional area of a $(d-1)$-dimensional surface
$\mathcal{L}_{d-2}$	the $(d-2)$-dimensional area of a $(d-2)$-dimensional surface
$B(x, r)$	the closed ball of radius r around $x \in \mathbb{R}^d$, that is, $B(x, r) = \{y \in \mathbb{R}^d \mid \operatorname{dist}(x, y) \leq r\}$
η_d	volume of the d-dimensional unit ball in the Euclidean metric, that is, $\eta_d = \mathcal{L}_d(B(0, 1))$
$\Pr(A)$	probability of the event A
$\mathbb{E}(U)$	expectation of the random variable U
$\operatorname{Var}(U)$	variance of the random variable U

$\mathrm{Bin}(n,p)$	discrete density of the binomial distribution with parameters n and p
$\xrightarrow{a.s.}$	almost sure convergence
$f = O(g)$	f is bounded above by g asymptotically up to a constant factor
$\nabla f(x)$	gradient of f at x
$\frac{\partial f(x)}{\partial x_i}$	partial derivative of the function f in the direction x_i
$G_{\mathrm{r}}(n,r)$	directed r-neighborhood graph
$G^u_{\mathrm{r}}(n,r)$	undirected r-neighborhood graph
$G_{\mathrm{kNN}}(n,k)$	directed k-nearest neighbor graph
$R^k(x_i)$	k-nearest neighbor radius of sample point x_i
$G_{\mathrm{sym}}(n,k)$	undirected k-nearest neighbor graph with an edge between two points if one is among the k-nearest neighbors of the other
$G_{\mathrm{mut}}(n,k)$	undirected k-nearest neighbor graph with an edge between two points if both are among the k-nearest neighbors of the other
$p(x)$	density, points are sampled from
$\hat{p}_n(x)$	density estimate at point x
μ	measure induced by the density p, that is, $\mu(A) = \int_A p(x)\,\mathrm{d}x$
n	sample size
$x_1,\ldots,x_n$	sample points
m	number of true clusters
m'	number of empirical or sample clusters
t	density level set parameter for high-density clusters
$L(t)$	t-level set of p
$C^{(1)},\ldots,C^{(m)}$	true clusters, that is for high-density clusters the connected components of $L(t)$
$\hat{C}_n^{(1)},\ldots,\hat{C}_n^{(m')}$	empirical clusters
$\tilde{C}_n^{(1)},\ldots,\tilde{C}_n^{(m')}$	sample clusters
$C_-^{(i)}(\varepsilon)$	connected component of $L(t-\varepsilon)$ containing $C^{(i)}$
$\beta_{(i)}, \tilde{\beta}_{(i)}$	probability mass of $C^{(i)}$ and $C_-^{(i)}(2\tilde{\varepsilon})$, respectively
$p_{\max}^{(i)}$	maximal value of the density in cluster $C^{(i)}$
$\rho^{(i)}$	lower bound on probability of balls of radius $u^{(i)}$ around points in $C_-^{(i)}(2\tilde{\varepsilon})$
$\kappa^{(i)}$	minimal curvature radius of the boundary $\partial C^{(i)}$
$\nu_{\max}^{(i)}$	maximal covering radius of cluster $C^{(i)}$
$Col^{(i)}(\nu)$	collar set of cluster $C^{(i)}$ for radius ν
$u^{(i)}$	lower bound on the distances between $C^{(i)}$ and other clusters

$\tilde{\varepsilon}$	parameter ε such that $\operatorname{dist}(C_-^{(i)}(2\varepsilon), C_-^{(j)}(2\varepsilon)) \geq u^{(i)}$ for all $\varepsilon \leq \tilde{\varepsilon}$
$R_{\min}^{(i)}$	minimal k-nearest neighbor radius of the sample points in cluster $C^{(i)}$
$\tilde{R}_{\max}^{(i)}$	maximal k-nearest neighbor radius of the sample points in cluster $C^{(i)}$
$G'_{\text{mut}}(n,k,t')$	the mutual kNN graphs on points with a density estimate over t'
$G'_{\text{sym}}(n,k,t')$	the symmetric kNN graph on points with a density estimate over t'
$\tilde{G}_{\text{mut}}(n,k,t',\delta)$	the graph $G'_{\text{mut}}(n,k,t')$ where connected components of less than δn points have been removed
$\tilde{G}_{\text{sym}}(n,k,t',\delta)$	the graph $G'_{\text{sym}}(n,k,t')$ where connected components of less than δn points have been removed
$\operatorname{cut}(C, V \setminus C)$	cut size of the cut defined by $(C, V \setminus C)$ in the graph $G(V,E)$
$\operatorname{vol}(C)$	volume of $C \subseteq V$ in the graph $G(V,E)$
$\operatorname{card}(C)$	number of vertices in $C \subseteq V$ in the graph $G(V,E)$
$\operatorname{Ncut}(C, V \setminus C)$	the normalized cut measure for the partition $(C, V \setminus C)$ in the graph $G(V,E)$
$\operatorname{RatioCut}(C, V \setminus C)$	the RatioCut measure for the partition $(C, V \setminus C)$ in the graph $G(V,E)$
S	hyperplane in $\mathbb{R}^d$ that defines the cuts we consider in the neighborhood graphs
H^+, H^-	halfspaces of $\mathbb{R}^d$ defined by S
$\operatorname{cut}_{n,r}(S)$	cut size of cut in $G_{\text{r}}(n,r)$ defined by S
$\operatorname{cut}_{n,k}(S)$	cut size of cut in $G_{\text{kNN}}(n,k)$ defined by S
$\operatorname{vol}_{n,r}(A)$	volume of sample points in the set A in $G_{\text{r}}(n,r)$
$\operatorname{vol}_{n,k}(A)$	volume of sample points in the set A in $G_{\text{kNN}}(n,k)$
$\operatorname{card}_n(A)$	number of sample points in the set A
$\operatorname{Ncut}_{n,r}(S)$	normalized cut of cut in $G_{\text{r}}(n,r)$ defined by S
$\operatorname{Ncut}_{n,k}(S)$	normalized cut of cut in $G_{\text{kNN}}(n,k)$ defined by S
$\operatorname{RatioCut}_{n,r}(S)$	RatioCut of cut in $G_{\text{r}}(n,r)$ defined by S
$\operatorname{RatioCut}_{n,k}(S)$	RatioCut of cut in $G_{\text{kNN}}(n,k)$ defined by S

Bibliography

[1] M. Ackerman and S. Ben-David. Measures of clustering quality: A working set of axioms for clustering. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 121–128. MIT Press, 2009.

[2] D. Angluin and L. Valiant. Fast probabilistic algorithms for Hamiltonian circuits. *Journal of Computer and System Sciences*, 18:155–193, 1979.

[3] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC 2004)*, pages 222–231. ACM, 2004.

[4] S. Arora, E. Hazan, and S. Kale. Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *46th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2005)*, pages 339–348. IEEE Computer Society, 2005.

[5] Ch. Avin and G. Ercal. On the cover time and mixing time of random geometric graphs. *Theoretical Computer Science*, 390(1-2):2–22, 2007.

[6] F. Avram and D. Bertsimas. On central limit theorems in geometrical probability. *The Annals of Applied Probability*, 3(4):1033–1046, 1993.

[7] J. Banfield and A. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

[8] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[9] C. Bettstetter. On the minimum node degree and connectivity of a wireless multihop network. In *Proceedings of the 3rd ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2002)*, pages 80–91. ACM, 2002.

[10] G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESIAM: Probability and Statistics*, 11:272–280, 2007.

[11] Y. Bilu and N. Linial. Are stable instances easy? Published electronically as arXiv:0906.3162v1, avaliable at `http://arxiv.org/abs/0906.3162`, 2009.

[12] B. Bollobas. *Random Graphs*. Cambridge University Press, Cambridge, 2001.

[13] B. Bollobas and O. Riordan. *Percolation*. Cambridge Universiy Press, 2006.

[14] S. Boucheron, G. Lugosi, and S. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.

[15] M. Brito, E. Chavez, A. Quiroz, and J. Yukich. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Statistics and Probability Letters*, 35:33–42, 1997.

[16] S. Bubeck and U. von Luxburg. Nearest neighbor clustering: A baseline method for consistent clustering with arbitrary objective functions. *Journal of Machine Learning Research*, 10:657–698, 2009.

[17] T. Bühler and M. Hein. Spectral clustering based on the graph p-Laplacian. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *Proceedings of the 26th International Conference on Machine Learning*, volume 382 of *ACM International Conference Proceedings*. ACM, 2009.

[18] C. Caroni and P. Prescott. Inapplicability of asymptotic results on the minimal spanning tree in statistical testing. *Journal of Multivariate Analysis*, 83:487–492, 2002.

[19] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[20] F. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.

[21] F. Chung. Laplacians and the cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

[22] T. Cormen, C. Leiserson, and R. Rivest. *Introduction to Algorithms*. MIT Press, 1989.

[23] A. Cuevas, M. Febrero, and R. Fraiman. Estimating the number of clusters. *Canadian Journal of Statistics*, 28, 2000.

[24] S. Dasgupta and L. Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

[25] Data Repository by Gunnar Rätsch. Available on the internet at `http://theoval.cmp.uea.ac.uk/~gcc/matlab/default.html#benchmarks`.

[26] H. Dette and N. Henze. The limit distribution of the largest nearest-neighbour link in the unit *d*-cube. *Journal of Applied Probability*, 26(1):67–80, 1989.

[27] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, 2001.

[28] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[29] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2000.

[30] R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge studies in advanced mathematics. Cambridge University Press, 1999.

[31] D. Estrin, R. Govindan, J. S. Heidemann, and S. Kumar. Next century challenges: Scalable coordination in sensor networks. In *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking (MOBICOM 1999)*, pages 263–270, 1999.

[32] H. Federer. *Geometric Measure Theory*, volume 153 of *Die Grundlehren der mathematischen Wissenschaften*. Springer, 1969.

[33] W. Fernandez de la Vega, M. Karpinski, and C. Kenyon. Approximation schemes for metric bisection and partitioning. In J. Ian Munro, editor, *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms (SODA 04)*, pages 506–515. SIAM, 2004.

[34] G. Flake, R. Tarjan, and K. Tsioutsiouliklis. Graph clustering and minimum cut trees. *Internet Mathematics*, 1(4):385–408, 2004.

[35] C. Fraley and A. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[36] M. Gaertler. Clustering. In *Network Analysis: Methodological Foundations*, volume 3418 of *Lecture Notes in Computer Science*, pages 178–215. Springer, 2005.

[37] J. M. Gonzàlez-Barrios and A. J. Quiroz. A clustering procedure based on the comparison between the k nearest neighbors graph and the minimal spanning tree. *Statistics and Probability Letters*, 62:23–34, 2003.

[38] G. Grimmett. *Percolation*. Springer, 1999.

[39] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, 2001.

[40] S. Guattery and G. Miller. On the quality of spectral separators. *SIAM Journal of Matrix Analysis and Applications*, 19(3):701–719, 1998.

[41] J. Hartigan. *Clustering algorithms*. Wiley, New York, 1975.

[42] J. Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374):388–394, 1981.

[43] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[44] X. He and P. Niyogi. Locality preserving projections. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

[45] M. Hein. *Geometric Aspects of Statistical Learning Theory*. PhD thesis, Technical University of Darmstadt, 2005.

[46] M. Hein, J.-Y. Audibert, and U. von Luxburg. From graphs to manifolds - weak and strong pointwise consistency. In P. Auer and R. Meier, editors, *Proceedings of the 18th Annual Conference on Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 470–485. Springer, 2005.

[47] M. Hein, J.-Y. Audibert, and U. von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8: 1325–1370, 2007.

[48] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[49] A. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[50] N. Jardine and R. Sibson. *Mathematical taxonomy*. Wiley, London, 1971.

[51] J. Kleinberg. An impossibility theorem for clustering. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 446–453. MIT Press, 2003.

[52] I. Kozakova, R. Meester, and S. Nanda. The size of components in continuum nearest neighbor graphs. *The Annals of Probability*, 34(2):528–538, 2006.

[53] S. S. Kunniyur and S. S. Venkatesh. Threshold functions, node isolation, and emergent lacunae in sensor networks. *IEEE Transactions on Information Theory*, 52(12): 5352–5372, 2006.

[54] J. M. Lee. *Riemannian manifolds: an introduction to curvature*. Springer, 1997.

[55] G. Lugosi. Concentration-of-measure inequalities. Lecture Notes, available at `http://www.econ.upf.es/~lugosi/anu.pdf`, 2006.

[56] M. Maier, M. Hein, and U. von Luxburg. Cluster identification in nearest-neighbor graphs. In M.Hutter, R. Servedio, and E. Takimoto, editors, *Proceedings of the 18th Conference on Algorithmic Learning Theory (ALT 2007)*, volume 4754 of *Lecture Notes in Artificial Intelligence*, pages 196–210. Springer, 2007.

[57] M. Maier, M. Hein, and U. von Luxburg. Optimal construction of k-nearest neighbor graphs for identifying noisy clusters. *Theoretical Computer Science*, 410(19): 1749–1764, 2009.

[58] M. Maier, U. von Luxburg, and M. Hein. Influence of graph construction on graph-based clustering measures. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1025–1032. MIT Press, 2009.

[59] D. Marchette. *Random Graphs for Statistical Pattern Recognition*. Wiley-IEEE, 2004.

[60] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989.

[61] C. McDiarmid and B. Reed. Concentration for self-bounding functions and an inequality of Talagrand. *Random Structures and Algorithms*, 29:549–557, 2006.

[62] M. Meila and W. Pentney. Clustering by weighted cuts in directed graphs. In *Proceedings of the Seventh SIAM International Conference on Data Mining (SDM 2007)*. SIAM, 2007.

[63] G. Miller, S. Teng, W. Thurston, and S. Vavasis. Separators for sphere-packings and nearest neighbor graphs. *Journal of the ACM*, 44(1):1–29, 1997.

[64] A. Mood, F. Graybill, and D. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, 1974.

[65] F. Morgan. *Riemannian geometry: a beginner's guide*. Jones and Bartlett Publishers, 1993.

[66] H. Narayanan, M. Belkin, and P. Niyogi. On the relation between low density separation, spectral clustering and graph cuts. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1025–1032. MIT Press, 2007.

[67] M. Penrose. *Random Geometric Graphs*. Oxford University Press, Oxford, 2003.

[68] M. Penrose. The longest edge of the random minimal spanning tree. *The Annals of Applied Probability*, 7(2):340–361, 1997.

[69] M. Penrose. A strong law for the longest edge of the minimal spanning tree. *The Annals of Probability*, 27(1):246–260, 1999.

[70] M. Penrose and J. Yukich. Central limit theorems for some graphs in computational geometry. *The Annals of Applied Probability*, 11(4):1005–1041, 2001.

[71] M. Penrose and J. Yukich. Weak laws of large numbers in geometric probability. *The Annals of Applied Probability*, 13(1):277–303, 2003.

[72] D. Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9(1): 135–140, 1981.

[73] G. Pottie and W. Kaiser. Wireless integrated network sensors. *Communications of the ACM*, 43(5):51–58, 2000.

[74] A. Rahimi and B. Recht. Clustering with normalized cuts is clustering with a hyperplane. In *Statistical Learning in Computer Vision*, 2004.

[75] B. Prakasa Rao. *Non Parametric Functional Estimation*. Academic Press, 1983.

[76] P. Santi and D. Blough. The critical transmitting range for connectivity in sparse wireless ad hoc networks. *IEEE Transactions on Mobile Computing*, 02(1):25–39, 2003.

[77] S.E. Schaeffer. Graph clustering. *Computer Science Review*, 1:27–64, 2007.

[78] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[79] B. Silverman. *Density estimation*. Chapman and Hall, 1986.

[80] D. Spielman and S. Teng. Spectral partitioning works: planar graphs and finite element meshes. In *37th Annual Symposium on Foundations of Computer Science (FOCS 1996)*, pages 96–105, 1996.

[81] A. Srivastav and P. Stangier. Algorithmic Chernoff-Hoeffding inequalities in integer programming. *Random Structures and Algorithms*, 8(1):27–58, 1996.

[82] D. Stauffer and A. Aharony. *Introduction to Percolation*. Taylor and Francis, 1994.

[83] K. Steffen. Hausdorff-Dimension, reguläre Mengen und total irreguläre Mengen. In E. Brieskorn, editor, *Felix Hausdorff zum Gedächtnis: Aspekte seines Werkes*, pages 185–228. Vieweg+Teubner, 1996.

[84] S. Vempala, R. Kannan, and A. Vetta. On clusterings - good, bad and spectral. In *Proceedings of the 41st Symposium on the Foundation of Computer Science (FOCS 2000)*, pages 367–377. IEEE Computer Society, 2000.

[85] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007.

[86] U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. In *PASCAL Workshop on Statistics and Optimization of Clustering*, 2005.

[87] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555–586, 2008.

[88] U. von Luxburg, S. Bubeck, S. Jegelka, and M. Kaufmann. Consistent minimization of clustering objective functions. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 961–968. MIT Press, 2008.

[89] J. Šíma and S. E. Schaeffer. On the NP-completeness of some graph cluster measures. In *Proceedings of the 32nd Conference on Current Trends in Theory and Practice of Informatics (SOFSEM 2006)*, volume 3831 of *Lecture Notes in Computer Science*, pages 530–537. Springer, 2006.

[90] D. Wagner and F. Wagner. Between min cut and graph bisection. In *Proceedings of the 18th International Symposium on Mathematical Foundations of Computer Science (MFCS 1993)*, pages 744–750. Springer, 1993.

[91] W. Walter. *Analysis 2*. Springer, 1995.

[92] W. Walter. *Analysis 1*. Springer, 1997.

[93] L. Wasserman. *All of Statistics. A Concise Course in Statistical Inference.* Springer, 2004.

[94] D. Werner. *Funktionalanalysis*. Springer, 2002.

[95] M.A. Wong and T. Lane. A kth nearest neighbour clustering procedure. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):362–368, 1983.