Model Reconstruction for Moment-based Stochastic Chemical Kinetics

---

A dissertation submitted towards the degree
Doctor of Engineering (Dr.-Ing.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by

ALEXANDER ANDREYCHENKO

Saarbrücken
October, 2017

Day of Colloquium	25 June 2018
Dean of the Faculty	Univ.-Prof. Dr. Sebastian Hack
Chair of the Committee	Prof. Dr. Volkhard Helms
First reviewer	Prof. Dr. Verena Wolf
Second reviewer	Prof. Dr. Luca Bortolussi
Third reviewer	Prof. Dr. Heinz Köppl
Academic Assistant	Dr. Daniel Stan

Abstract

Based on the theory of stochastic chemical kinetics, the inherent randomness and stochasticity of biochemical reaction networks can be accurately described by discrete-state continuous-time Markov chains, where each chemical reaction corresponds to a state transition of the process. However, the analysis of such processes is computationally expensive and sophisticated numerical methods are required. The main complication comes due to the largeness problem of the state space, so that analysis techniques based on an exploration of the state space are often not feasible and the integration of the moments of the underlying probability distribution has become a very popular alternative.

In this thesis we propose an analysis framework in which we integrate a number of moments of the process instead of the state probabilities. This results in a more time-efficient simulation of the time evolution of the process. In order to regain the state probabilities from the moment representation, we combine the moment-based simulation (MM) with a maximum entropy approach: the maximum entropy principle is applied to derive a distribution that fits best to a given sequence of moments.

We further extend this approach by incorporating the conditional moments (MCM) which allows not only to reconstruct the distribution of the species present in high amount in the system, but also to approximate the probabilities of species with low molecular counts.

For the given distribution reconstruction framework, we investigate the numerical accuracy and stability using case studies from systems biology, compare two different moment approximation methods (MM and MCM), examine if it can be used for the reaction rates estimation problem and describe the possible future applications.

Zusammenfassung

Basierend auf der Theorie der stochastischen chemischen Kinetiken knnen die inhrente Zufälligkeit und Stochastizität von biochemischen Reaktionsnetzwerken durch diskrete zeitkontinuierliche Markow-Ketten genau beschrieben werden, wobei jede chemische Reaktion einem Zustandsübergang des Prozesses entspricht.

Die Analyse solcher Prozesse ist jedoch rechenaufwendig und komplexe numerische Verfahren sind erforderlich. Analysetechniken, die auf dem Abtasten des Zustandsraums basieren, sind durch dessen Größe oft nicht anwendbar. Als populäre Alternative wird heute häufig die Integration der Momente der zugrundeliegenden Wahrscheinlichkeitsverteilung genutzt.

In dieser Arbeit schlagen wir einen Analyserahmen vor, in dem wir, anstatt der Zustandswahrscheinlichkeiten, zugrundeliegende Momente des Prozesses integrieren. Dies führt zu einer zeiteffizienteren Simulation der zeitlichen Entwicklung des Prozesses. Um die Zustandswahrscheinlichkeiten aus der Momentreprsentation wiederzugewinnen, kombinieren wir die momentbasierte Simulation (MM) mit Entropiemaximierung: Die Maximum-Entropie-Methode wird angewendet, um eine Verteilung abzuleiten, die am besten zu einer bestimmten Sequenz von Momenten passt. Wir erweitern diesen Ansatz durch das Einbeziehen bedingter Momente (MCM), die es nicht nur erlauben, die Verteilung der in großer Menge im System enthaltenen Spezies zu rekonstruieren, sondern es ebenso ermöglicht, sich den Wahrscheinlichkeiten von Spezies mit niedrigen Molekulargewichten anzunähern.

Für das gegebene System zur Verteilungsrekonstruktion untersuchen wir die numerische Genauigkeit und Stabilität anhand von Fallstudien aus der Systembiologie, vergleichen zwei unterschiedliche Verfahren der Momentapproximation (MM und MCM), untersuchen, ob es für das Problem der Abschätzung von Reaktionsraten verwendet werden kann und beschreiben die mögliche zukünftige Anwendungen.

Contents

Chapter 1

Introduction

Recently, the single-cell imaging techniques has become one of the primer focuses of the research community in the drug development and microbiology experiments. This also has raised interest for the influence of stochastic effects onto the living systems. These systems are successfully modeled and analyzed using the Markov chains formalism

However, one of the main problems that hinders Markov chains from being applied to any stochastic model is the existence of possibly unbounded number of system configurations. It may be partially mitigated by changing the formal representation, treating chemical species present in high amount as continuous variables and predicting their dynamics with moment closure techniques.

The hybrid approach, where low-amount species are represented by the Markov chain and the high-amount ones are approximated by moments reveals to be very beneficial. However, the moment approximation can only provide the averaged statistical metrics.

Hence, the goal of this doctoral thesis is to develop the efficient methods to approximate the underlying stochastic distribution of the Markov chain when the dynamics is described using cumulative quantities such as moments.

This work relies on the work of Abramov [3] and Hasenauer et al. [105]. We succeeded in joining the hybrid moment computation and maximum entropy reconstruction of the conditional and joint distributions for several models from stochastic chemical kinetics. It is shown that this combination of techniques provides benefits to the approximation of such stochastic systems behavior from computational point of view, where the underlying state space is discrete and too large to be treated by direct solution methods.

Here we solve the following problem: given the approximation of (finite amount of) moments at a certain time point, reconstruct the underlying discrete-state stochastic distribution, where the moment constraints come from the standard or conditional moment closure. This allows to obtain the state-wise description of the model at certain time instant without losing much accuracy and several times faster than applying direct solution approaches to solve the chemical master equation.

The structure of the thesis is as follows. The basic notation and theoretical background on the stochastic chemical kinetics is provided in Chapter 2. The moment approximation methods, namely moment closure and method of conditional moments are described in Chapter 3. The theoretical background on inverse moment problem and maximum entropy method and its application to the synthetic examples is considered in Chapter 4. The numerical analysis of the models from systems biology using the proposed combination of techniques is conducted in Chapter 5 together with the approach to parameter estimation. The details of the implementation and miscellaneous background information is given in Appendix A.

Through the thesis you find the biological models and examples defined in green and beige boxes. Each table and figure is referenced using hyperlinks and we use the symbol ↩ in the caption to return to the mention of table or figure in the text. For each reference given in the bibliography, you find the list of pages (together with link to the mention in text) where this reference is cited.

Chapter 2

Stochastic Modeling

2.1 Chemical Reaction Kinetics

The variety of mathematical models were used for a long time in science to describe the interactions between chemical compounds. Following the progress of recent developments in singe-cell imaging techniques, discrete stochastic models gained a lot of attention and have become a very popular description of biochemical reactions that take place in living organisms. We consider *chemical reaction networks* given as a set of N_R stoichiometric equations

$$R_j: \qquad \ell_{j,1}S_1 + \ldots + \ell_{j,N_S}S_{N_S} \xrightarrow{c_j} \tilde{\ell}_{j,1}S_1 + \ldots + \tilde{\ell}_{j,N_S}S_{N_S},$$

where $\ell_{j,i}, \tilde{\ell}_{j,i} \in \mathbb{N}_0$ are stoichiometric coefficients that refer to the number of molecules used up and produced by the reaction R_j and $S_1, \ldots, S_{N_S}$ refer to N_S distinct chemical species. Stochastic *reaction rate constant* [82] c_j determines the average speed of the reaction R_j and depends on the temperature, the reaction volume and the affinity of the chemical species. We associate a *change vector* $v_j = (v_{j,1}, \ldots, v_{j,N_S}) \in \mathbb{Z}^{N_S}$, $v_{j,i} = \tilde{\ell}j, i - \ell j, i$ that gives the population change caused by a single firing of the reaction R_j. The *state* x of the system is given by the *population* X_i (cf. Modeling Approaches) of each chemical species S_i such that $x = (X_1, \ldots, X_{N_S})$. Below we provide the example of the chemical reaction system that describes the simple dimerization process:

Biological model 2.1 (Simple Dimerization)**.** ↩ We consider the dimerization kinetics [238, p. 163] of a protein P in a bacterial cell. The stoichiometric equations that correspond to forward and backward reactions are

$$R_1: \quad 2P \xrightarrow{c_1} P_2,$$
$$R_2: \quad P_2 \xrightarrow{c_2} 2P,$$

where $\ell_{1,1} = \tilde{\ell}_{2,1} = 2$, $\ell_{2,2} = \tilde{\ell}_{1,2} = 1$ and $\ell_{1,2} = \ell_{2,1} = \tilde{\ell}_{1,1} = \tilde{\ell}_{2,2} = 0$. Note that we omit terms which are zero.

In the sequel we also restrict to chemical reactions that are at most bimolecular, i.e., we assume that $\sum_{i=1}^{N_S} \ell_{j,i} \in \{0, 1, 2\}$, which is a reasonable assumption because reactions

where more than two molecules have to collide can usually be decomposed into smaller ones where at most two molecules collide [83].

2.2 Modeling Approaches

We distinguish three approaches to model the chemical reaction systems. To put these systems on a common framework, we emphasize that they share a *population structure*, so that a system consists of N_S different *populations* whose amounts determine the future behavior. The state x of the system consists of quantities that are usually non-negative, and might all be discrete, real-valued or mixed. Also they might be bounded or unbounded. Usually the state of the system changes over time and time can be treated as discrete or continuous[1]. Here we consider three modeling semantics:

- deterministic semantics
- stochastic semantics
- hybrid semantics

In the deterministic setting $\vec{X}$ is a function and in stochastic and hybrid settings it corresponds to the stochastic process.

2.2.1 Deterministic Semantics

The dynamics of biochemical reactions can be modeled using the system ordinary differential equations that describe the evolution of species concentrations in time [133] (i.e., molecules per volume, commonly measured in mol/L). The state of the system at the given time instant t is given by the vector of concentrations $x = ([X_1], \ldots, [X_{N_S}])$. The state space in that case is continuous, i.e., $\vec{X} : \mathbb{R} \to \mathbb{R}^{N_S}$. We define the *deterministic propensity* $\alpha_R^{\mathtt{det}}(x)$ of a reaction R for state x as

$$\alpha_R^{\mathtt{det}}(x) = c_R^{\mathtt{det}} \cdot \prod_{i=1}^{N_S} [X_i]^{\ell_{R,i}},$$

where $c_R^{\mathtt{det}}$ differs depending on the type of the model [239]. The exponent $\ell_{R,i}$ is the number of occurences of the species in the reaction R. The definition of the rate function is given by the law of mass action [47], which serves as the basis for such models in enzyme kinetics as Hill and Michaelis-Menten equations [109, 185]. Given the vector of *initial concentrations* $\vec{X}(0)$, the future behavior is determined by the system of ODEs (*reaction rate equations*)

$$\dot{X}_i(t) = \sum_{j=1}^{N_R} v_{j,i} \cdot \alpha_R^{\mathtt{det}}(X(t)), \tag{2.1}$$

[1] In this thesis we consider only continuous-time models.

The deterministic framework does not account for possible measurement errors neither for uncertainties in states which limits its direct application to such tasks as parameter inference. Stochastic differential equations [40] framework is capable to take these considerations into an account.

Example 2.1. ↩ The ODE for the Simple Dimerization model as obtained from the CRN is constructed as follows:

$$\dot{X}(t) = [\text{-2 } 1]^T \cdot \alpha_1^{\texttt{det}t}(X(t)) + [2 \text{ -1}]^T \cdot \alpha_2^{\texttt{det}}(X(t))$$

with $\alpha_1^{\texttt{det}}(x) = c_1^{\texttt{det}} \cdot [X_1]^2$, $\alpha_2^{\texttt{det}}(x) = c_2^{\texttt{det}} \cdot [X_2]$.

This framework can be extended to account for the spatial information by considering the system to be compartmental [31, 102].

2.2.2 Stochastic Semantics

Under the certain conditions the deterministic modeling formalism is not the appropriate framework to describe the behavior up to the required accuracy level. Namely, when the number of molecules in the cell is small the role of the stochastic noise can not be ignored.

There are two possible sources of stochasticity in modeling. The first one is so-called *extrininsic noise* that is associated with the measurement error, environmental changes or interference with other processes [17, 29, 66, 71, 85, 218, 247]. The second source of stochasticity is the *intrinsic noise* that governs the stochastic evolution of the system. One simple example is the event of collision of two molecules of a gas in a large volume. Intrinsic noise can also be associated with the speed or the orientation of the molecules.

Example 2.2. ↩ Consider the simple bi-molecular chemical reaction

$$A + B \to ?$$

where a molecule of species A collides with a molecule of species B (the result of this collision is currently of no interest). Two molecules are assumed to collide if the distance is less then a reacting distance ζ. The probability of such collision in the reaction volume Ω under the assumption that we can neglect the size of both molecules and consider them as particles is given by

$$\mathbf{P}[|(x,y,z)_A - (x,y,z)_B| < \zeta],$$

where the tuple $(x,y,z)_A$ describes the coordinates of species A molecule in $\mathbb{R}^3$. This probability can be computed as the volume of the sphere of radius ζ around any of both particles divided by Ω

$$\mathbf{P}[|(x,y,z)_1 - (x,y,z)_2| < \zeta] = \frac{4}{3}\pi\zeta^3 \cdot \frac{1}{\Omega}.$$

It does not depend on time for well-stirred reaction volumes. Using this observation, one can continue with the rigorous derivation [84] of stochastic chemical kinetics theory. Here we introduce basic notions that are needed later in the course of the thesis.

We note that one can also follow the more detailed derivation where the radius of both molecules is taken into account. As before, the assumption about the thermal equilibrium (with the temperature T) is made, therefore the velocity of a molecule is follows the Boltzman distribution (with the Boltzmann constant k_B). It is important that the collision between the molecules does not always lead to the reaction, thus the corresponding probability can be computed. In this setting, the probability for the reaction between two molecules to happen in the infinitesimal time period $[t, t+dt]$ to happen is given by

$$\mathbf{P}[\text{reaction}] = \mathbf{P}[\text{collision}] \cdot \mathbf{P}[\text{reaction}|\text{collision}] = \frac{1}{\Omega}(r_1 + r_2)^2 \sqrt{\frac{8k_BT}{\pi m}} \cdot e^{\frac{-\epsilon}{k_BT}} = p^*.$$

Here, r_1 and r_2 correspond to the radii of the molecules, the average mass of molecules is denoted by $m = \frac{1}{2}(m_1 + m_2)$ and the constant ϵ is the critical energy to run the reaction when molecules collide. It gives rise to the rigorous definition of the propensity function. For example, the propensity function of the bi-molecular reaction $S_1 + S_2 \to \dots$ is given by $x_i x_j p^*$, where $x_i x_j$ corresponds to the number of possible distinct pairs of S_1 and S_2 molecules.

Other sources of the intrinsic noise are the lack of the total predictability and quantum indeterminacy [88]. For low species levels, these stochastic fluctuations play the important role and seriously affect the behavior of the system [206]. Because of these stochastic effects, the granularity of the modeling is at the molecule level and the state of the system is given by the vector $x = (x_1, \dots, x_{N_S})$, $x \in \mathbb{N}_0^{N_S}$ where x_i corresponds to the number of molecules of species S_i.

The stochastic semantics of a chemical reaction network is given by a *(homogeneous) continuous-time Markov chain (CTMC)* [83, 84]. Therefore, we construct a Markov process $\{\vec{X}(t), t \geq 0\}$ which is a family of random variables indexed by time. It satisfies the Markov property, i.e., the possible future behavior only depends on the current state

$$\begin{aligned} & \mathbf{P}[\vec{X}(t_n) = x_n \mid \vec{X}(t_{n-1}) = x_{n-1}, \dots, \vec{X}(t_0) = x_0] \\ = \; & \mathbf{P}[\vec{X}(t_n) = x_n \mid \vec{X}(t_{n-1}) = x_{n-1}] = \mathbf{P}[\vec{X}(t_n - t_{n-1}) = x_n \mid \vec{X}(0) = x_{n-1}], \end{aligned}$$

where the last equation shows the property of homogeneity such that the next state does not depend on the absolute value of time but only on the difference $t_n - t_{n-1}$.

Transitions of the Markov chain $\vec{X}$ correspond to chemical reactions and the transition rate of reaction R_j is given by

$$\lim_{\delta \to 0} \frac{1}{\delta} \mathbf{P}[\vec{X}(t+\delta) = x + v_j \mid \vec{X}(t) = x] = c_j \prod_{i=1}^{N_S} \binom{x_i}{\ell_{j,i}},$$

where the binomial coefficients describe the number of possible combinations of reactant molecules. The reaction rate constant c_j encodes the physical properties of reaction R_j and it is assumed to be constant in time (which corresponds to the homogeneity of the Markov chain). The units of c_j depend on the type of reaction. The actual value can be computed using the rate constant c_j^{det} used in the deterministic semantics [143].

In contrast to the deterministic setting, the system is in a state $x \in \mathbb{N}_0^{N_S}$ with a certain probability $\pi_x(t)$. Distribution $\pi(t)$ is called the *transient probability distribution* at time t and is defined as a row vector $\pi(t)$ such that $\pi_x(t) = \mathbf{P}[\vec{X}(t) = x] \in [0, 1]$ and $\sum_{x \in \mathbb{N}^{N_S}} \pi_x(t) = 1$.

The behavior of the CTMC is fully described by an *initial distribution* $\pi(0)$ and an *infinitesimal generator matrix* $Q = [q_{x,y}]$ where all elements $q_{x \neq y} \in \mathbb{R}_{\geq 0}$ and the diagonal entries $q_{x,x} = -\sum_{y \neq x} q_{x,y}$ for all states $x, y \in \mathbb{N}_0^{N_S}$. The transient distribution satisfies the *forward Kolmogorov differential equation* [134]

$$\frac{d}{dt}\pi(t) = \pi(t) \cdot Q. \tag{2.2}$$

This is also called the *chemical master equation (CME)* [84] in natural sciences. The stochastic *propensity* $\alpha_R^{\text{st}}(x)$ of the reaction R_j for a state x is defined as

$$\alpha_j^{\text{st}}(x) = \begin{cases} c_j^{\text{st}} \cdot \prod_{i=1}^{N_S} \binom{x_i}{\ell_{j,i}}, & \text{if } x_i \geq \ell_{j,i}, \\ 0, & \text{otherwise,} \end{cases}$$

and the elements of the generator matrix Q are given by

$$q_{x,y} = \begin{cases} \sum_{\{j \mid x + v_j = y\}} \alpha_j^{\text{st}}(x), & x \neq y, \\ -\sum_{z \neq x} q_{x,z}, & x = y. \end{cases}$$

We assume that the chemical master equation (2.2) has a unique solution at time t given the initial condition $\pi(0)$

$$\pi(t) = e^{Qt}.$$

To solve the CME (2.2) in practice, one may use the specially designed methods such as *simulation* [40, 83, 86] and *uniformization* [59, 96] as well as generally applicable analytical and numerical methods for solving ODE systems. The example of the simple CME construction is given in Example 2.3:

Example 2.3. ↩ Under the assumption that the number of protein molecules P is small inside in a cell, the dynamics of the Simple Dimerization model is well described by the Markov process $\vec{X}(t) = (X_P(t), X_{P_2}(t))$, $t \geq 0$. Here, $X_P(t)$ ($X_{P_2}(t)$) corresponds to the number of P (P_2) molecules at time t. The state space $S \subset \mathbb{N}_0^{N_S}$ and it is bounded from above, i.e., there is a finite number of possible configurations depending on the initial condition $\vec{X}(0)$. The dynamics of $\vec{X}(t)$ is governed by the infinitesimal generator matrix Q with elements

$$q_{x,y} = \begin{cases} c_1 \cdot \frac{1}{2} x_P (x_P - 1) & \text{if } x_P > 1, y_P = x_P - 2, y_{P_2} = x_{P_2} + 1, \\ c_2 \cdot x_{P_2} & \text{if } x_{P_2} > 0, y_P = x_P + 2, y_{P_2} = x_{P_2} - 1, \\ -\sum_{z \neq x} q_{x,z} & \text{if } x = y, \text{and} \\ 0 & \text{otherwise.} \end{cases}$$

Assume that the initial conditions are given by $\vec{X}(0) = (6, 0)$. Then the state space is defined as $S = \{x_1, x_2, x_3, x_4\} = \{(6, 0), (4, 1), (2, 2), (0, 3)\}$. We can construct the chemical master equation for a state $x_3 = (2, 2)$ as follows

$$\begin{aligned} \frac{d}{dt}\pi(x_3) &= \pi(x_2) \cdot q_{x_2,x_3} + \pi(x_4) \cdot q_{x_4,x_3} - \pi(x_3) \cdot (q_{x_3,x_2} + q_{x_3,x_4}) \\ &= \pi(x_2) \cdot 6c_1 + \pi(x_4) \cdot 3c_2 - \pi(x_3) \cdot (2c_2 + c_1)\,, \end{aligned}$$

where the first two terms correspond to the *inflow* of probability mass from states x_2 and x_4 and the third term is the *outflow* from state x_3 (cf. Figure 2.1). To simplify the notation, we omit the dependency of the distribution on time t. We can set up equations for all states $x \in S$ in a similar fashion. Given the initial condition $\vec{X}(0)$ we can solve CME using the technique of choice (cf. Section 2.3.2).

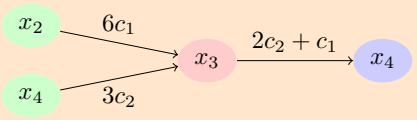


Figure 2.1: ↩ Visualization of the chemical master equation for the state x. Green states provide the inflow of the probability mass and the outflow from state x goes to red states.

If there is no prior knowledge about the bounds on molecular counts, the state space of the Markov chain is countably infinite. A tool specifically tailored to analyze such systems numerically is SHAVE [146]. It implements the sophisticated techniques such as dynamic state space truncation, error tolerance control and automated hybridization of populations. This helps to cope with very large or even infinite state spaces.

Example 2.4 (Relationship of Stochastic and Deterministic Descriptions)**.** ↩ The solution of the CME can be approximated using the deterministic approach in the limit of the number of molecules per volume. This allows to use only the mean population numbers by considering the *mean-field limit* of the system. Assume that Φ is the solution of Eq. (2.1) with the initial condition Φ_0.

Let $\vec{X}^\Omega = \vec{X}/\Omega$ be the stochastic representation of the same chemical system with $\vec{X}^\Omega(0) = \Phi_0$. For every time $t \geq 0$ it holds *almost surely* that

$$\lim_{\Omega \to \infty} \sup_{s \leq t} |\vec{X}^\Omega(s) - \Phi(s)| = 0.$$

This result holds only under certain conditions [24, 98]. The authors of [90, 143] conducted the numerical investigation of the relation between two formalisms in the simulation setting.

2.2.3 Hybrid Semantics

It is often the case for real-life biological studies that some species posses low molecular counts and others are present in high amounts. Usage of stochastic semantics in this case may result in an extremely large (or even infinite) state space which renders the solution of the corresponding CME (2.2) to be very demanding. We can treat the dynamics of the whole system deterministically which allows for faster computation but omits the stochastic effects. So-called *hybrid* approach is used to get the best from both deterministic and stochastic semantics by being able to combine the computational speed and accurately describe parts of the system highly influenced by noise. Therefore, we represent the original process $X(t)$ as a tuple $X(t) = (X_s(t), X_d(t))$, where the discrete stochastic component $X_s(t)$ describes the dynamics of small populations and the deterministic continuous component $X_d(t)$ describes large populations. We assume that the dimension of $X_s(t)$ is $\hat{n}$ and the dimension of $X_d(t)$ is $\tilde{n}$ such that $N_S = \hat{n} + \tilde{n}$. The dynamics of the stochastic component $X_s(t) \in \mathbb{N}^{\hat{n}}$ is described by a CTMC where the rates not only depend on the current state s but also on the state y of the deterministic counterpart $X_d(t) \in \mathbb{R}^{\tilde{n}}$. The entries of the time-dependent generator matrix $Q(t)$ are given by

$$q_{s,s'}(t) = \begin{cases} \sum_{\{j \mid s+v_j=s'\}} \alpha_j^{\text{hyb}}(s, y(t)), & \text{if } s \neq s' \\ -\sum_{s'' \neq s} q_{ss''}(t), & \text{if } s = s' \end{cases} \tag{2.3}$$

with propensity functions $\alpha_j^{\text{hyb}}(s, y(t))$ defined as

$$\alpha_j^{\text{hyb}}(s, y(t)) = \begin{cases} c_j^{\text{st}} \cdot \prod_{i=1}^{\hat{n}} \binom{s_i}{\ell_{j,i}} \cdot \prod_{j=1}^{\tilde{n}} y_j^{\ell_{r,j}}(t), & \text{if } s_i \geq \ell_{r,i} \\ 0, & \text{otherwise} \end{cases}$$

The transient probability distribution for $X_s(t)$ follows the *time-dependent* Kolmogorov equation $\dot{\pi}(t) = \pi(t) \cdot Q(t)$. We assume that this system of equations has a unique solution under the given initial condition $\pi(0)$. To solve this equation system the methods described in Section 2.3 can be applied as well as methods designed to analyze time-dependent continuous-time Markov chains [12, 18, 173, 227] and hybrid systems [6, 54, 163, 191]. The behavior of the continuous process $X_d(t)$ is defined by the system of ODE similar to Equation (2.1)

$$\dot{X}_d(t) = \sum_{j=1}^{\tilde{R}} v_{j,i} \cdot \alpha_j^{\text{det}}(\vec{X}(t)), \tag{2.4}$$

where $\tilde{R}$ is the number of reactions, where only species whose concentration is treated as a continuous quantity, take part.

Such a representation is especially useful if the Markov chain $X_s(t)$ only encodes the states (on/off) of a set of genes that control the production of proteins. Then each state

variable $s_1, s_2, \ldots, s_{\hat{n}}$ can take values 1 or 0 and the concentration of the corresponding proteins is controlled by the deterministic process $X_d(t)$. We can refer to states of $X_s(t)$ as to *modes* [199] of the hybrid system. In each mode, the components of $X_d(t)$ follow the system of ODE as in Equation (2.4). For a rigorous derivation of the corresponding theory, we refer to [114].

Consider the following biological model:

Biological model 2.2 (Predator-Prey)**.** ↩ This reaction network is often used to characterize the population dynamics of animals. The predator-prey model was proposed independently by Lotka and Volterra [154, 229]. The chemical reaction network is described by

$$\begin{aligned} S_1 &\xrightarrow{c_1} 2 \cdot S_1 \\ S_1 + S_2 &\xrightarrow{c_2} 2 \cdot S_2 \\ S_2 &\xrightarrow{c_3} \emptyset \end{aligned}$$

where the first reaction describes reproduction of a prey species S_1, the second reaction encodes the reproduction of predators S_2 consuming prey (food) and the last reaction relates to the natural death of the predator species.

We show how the hybrid semantics is applied in the next example:

Example 2.5 (Predator-Prey Hybrid)**.** ↩ The predator-prey model is often used to test the applicability of hybrid solution techniques since under certain parameter values it possesses high stiffness. In this example we assume that the reaction rate constants are such that species S_2 (predator) has low molecular counts and species S_1 (prey) is present in high counts. A hybrid approach leads to a definition of the process as $\vec{X}(t) = (S_1(t), S_2(t))$, where S_2 is treated stochastically and its behavior is determined by the time-dependent generator matrix $Q(t)$ with entries $q_{s_1,s_1'}(t)$ according to Equation (2.3), where $\alpha_2^{\text{hyb}}(s_1(t), s_2) = c_2^{\text{st}} \cdot s_1(t) \cdot s_2$, $\alpha_3^{\text{hyb}}(s_1(t), s_2) = c_3^{\text{st}} \cdot s_2$. The evolution of species S_1 is deterministic and follows the ODE $\dot{S}_1(t) = c_1^{\text{det}} s_1(t) - c_2^{\text{det}} s_1(t) s_2$.

The presented fixed decomposition of the state space vector onto two parts is based on dividing species into low and high abundance groups [129, 182]. There exist other approaches to divide the system into subsystems. They allow for dynamical repartitioning by exploiting the model structure. For details, we refer to [24, 40, 54, 93, 111, 231]. We emphasize the application of the *linear noise approximation*, which separately treats the noise term and the deterministic drift and provides a quasi-second order approximation of the dynamics [135, 233, 234].

2.3 Continuous-time Markov Chain Probability Distribution Computation Techniques

In the previous section we describe the three widely used semantics that allow us to select the best suited description for the given biological model. Now let us consider the stochastic semantics in more details. The main purpose when using the stochastic description is to analyze the distribution π of the CTMC $\vec{X}$. In order to analyze the system in *equilibrium* the time is set to $t \to \infty$. The corresponding distribution $\pi(\infty)$ is called *equilibrium distribution* of the Markov chain $\vec{X}$. Numerical approximation of the equilibrium distribution requires specially designed techniques [56, 100] and we do not consider this problem here. Instead, we concentrate on the approximation of transient distribution π at a certain time instant T.

2.3.1 Monte-Carlo Simulation

One of the first methods to approximate the transient distribution $\pi(T)$ is based on the Gillespie theory of chemical kinetics (therefore the name "Gillespie algorithm" occurs in the literature more often). Since for the stochastic process $\vec{X}(t)$ the Markov property holds, one can determine the time until next reaction firing. Assume that the system is in state x at time t. The time t' until the firing of any possible reaction is distributed exponentially with the rate $-\alpha_0(x) = -\sum_{j=1}^{N_R} \alpha_j(x)$, which is called the *exit rate* of the state x. We need to calculate the two following quantities:

- Time t' until the change of the state. We generate the exponentially distributed random number $t' \sim Exp(\text{-}\alpha_0(x))$ for that.
- Index j of the next reaction. The probability to fire the reaction R_j is distributed according to the discrete distribution $\alpha_1(x)/\alpha_0(x), \ldots, \alpha_{\tilde{N}_R}(x)/\alpha_0(x)$.

  ($\tilde{N}_R \leq N_R$ since not all reactions may be possible in state x).

Thus, starting from a state x at time t one needs to generate the two random variables for t' and j. The state at time $t + t'$ is then given by $x' = x + v_j$. Repeating these steps, one generates a single *simulation trajectory* until the time horizon of interest T is reached. The example is shown in Figure 2.2 for simple protein production model.

The approximation p of the transient distribution π is then obtained by repeating N_{sr} simulation runs. Transient probability of the state x is then given by $p_x(T) = 1/N_s \sum_{j=1}^{N_{sr}} \mathbb{1}_x \left(\vec{X}_j(T)\right)$, where $\vec{X}_j$ is the simulation run that ends up in the state x at time T. The mean and the variance of the stochastic process are then approximated using the formulas for *sample*

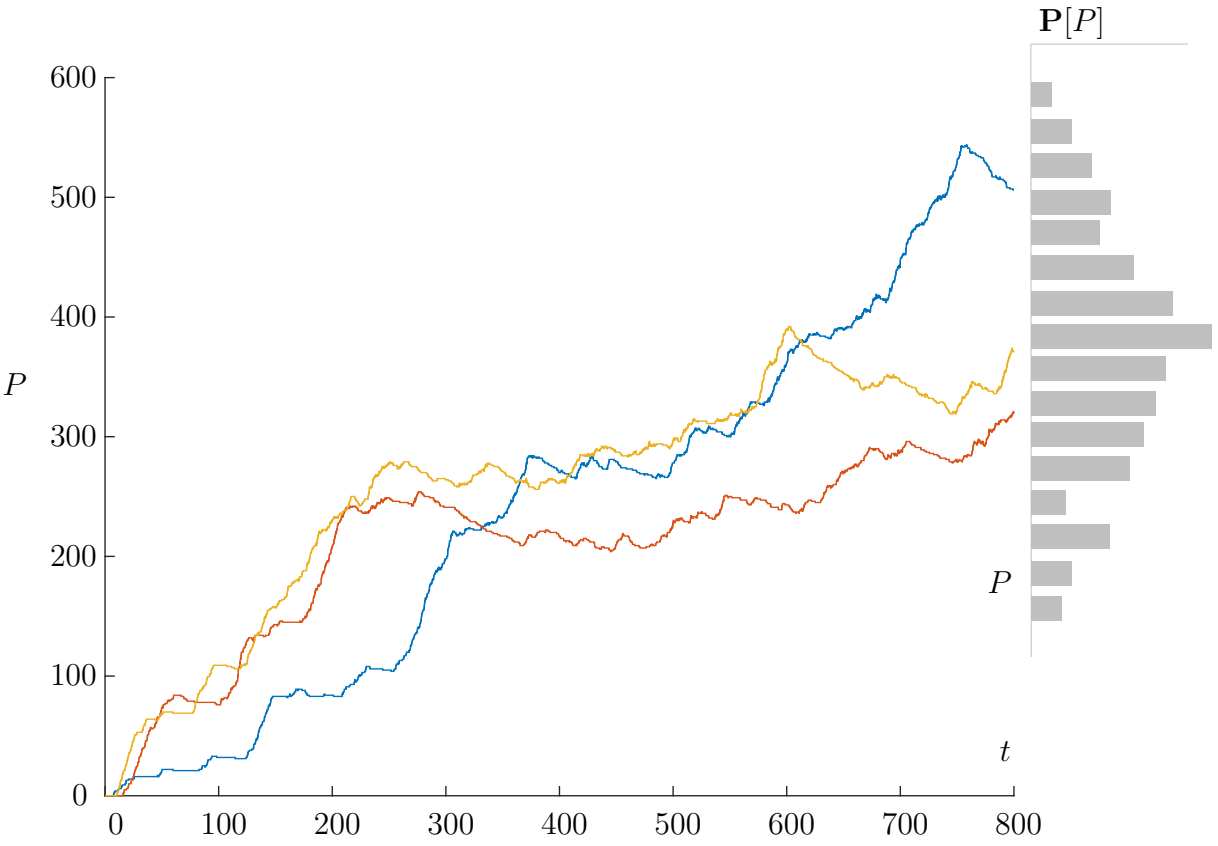


Figure 2.2: ↩ Three realizations of Monte-Carlo simulation for the simple protein production model shown as blue, red and orange lines. The time horizon of interest is 800. The example of the transient probability distribution approximation is shown as in-figure, where each simulation realization falls at the certain bin of the histogram (gray boxes) that approximates the probability distribution of interest.

mean and variance

$$E\left(\vec{X}(T)\right) = \frac{1}{N_{sr}} \sum_{j=1}^{N_{sr}} x_j(T),$$

$$\Sigma(T) = \frac{1}{N_{sr}-1} \sum_{j=1}^{N_{sr}} \left(x_j - E\left[\vec{X}(T)\right]\right) \left(x_j - E\left[\vec{X}(T)\right]\right)^T,$$

where $x_j(T)$ is a state of the Markov chain on j-th sample path at time T. These estimators for the moments and the probability distribution are unbiased and converge. The central limit theorem can be used to obtain the confidence interval for the individual probabilities. It also holds that in order to halve the length of this confidence interval one needs 4 times more simulation runs. This shows the main computational restriction of simulation-based approaches: they need a lot of runs to provide the reasonable approximations and low-variance estimates for systems with many reacting species.

There exist many approaches to mitigate the computational time issues that are based on the approximation of the exact simulations using the reduction approaches [26, 148, 202, 232] and the separation of time scales [43, 87, 111, 191].

2.3.2 Direct Solution Approaches

The main alternative to the simulation and continuous approximation methods is to cope with the CME directly. The analytical solution can be obtained only in the limited number of cases [180], for example, for systems with only zero and first order reactions [118] (however, reactions of the type $S_1 \to S_1 + S_2$ are not covered).

For most systems it is impossible to get the analytical solution of the CME. Therefore, the numerical solution methods are used such as already mentioned uniformization [59, 96, 158]. Another group of methods focuses on the direct solution of differential equations, such as *Proper Generalized Decomposition* (PGD) [10, 45], conversion into integro-differential equations [178], *Finite State Projection* (FSP) and its extensions [41, 44, 61, 71, 175, 176], tensor-based approaches [58, 62, 228] and the *Sliding Window* [110]. The extended version of the latter approach is implemented in the tool SHAVE [146].

Let us write the CME (2.2) in the following form

$$\frac{\partial \pi(x,t)}{\partial t} = \sum_{j=1}^{N_R} \left[\alpha_j(x - v_j)\pi(x - v_j, t) - \alpha_j(x)\pi(x,t)\right]. \tag{2.5}$$

It is important to note that often the number of states with $\pi(x,t) > 0$ is infinite since bounds on the population sizes are not known a priori. Thus, although in reality the molecule numbers are always finite, theoretically an infinite number of states can have positive probability. This leads to two complications compared to finite-state models. First, the limiting distribution of the Markov chain may not exist and additional conditions are necessary to ensure its existence [56]. Second, the *truncation* techniques are necessary to numerically simulate (2.5) since only a tractable number of states can be considered in each integration step. The direct numerical simulation performs well as long as the average population sizes remain small and the approximation error can be controlled by a threshold criterion.

2.3.3 Sliding Window Approach

The direct numerical simulation is based on the dynamic state space truncation developed for uniformization methods [158] and for integration schemes such as Runge-Kutta methods [13, 164]. We use the inflow-outflow form of (2.5) for the construction of the dynamic state space. The terms $\alpha_j(x - v_j)\pi(x - v_j, t)$ can be seen as the inflow to state x for reaction R_j while $\alpha_j(x,t)\pi(x,t)$ is the corresponding outflow. Let $p(x,t)$ be the approximation of $\pi(x,t)$ during the numerical integration for all x and all $t \geq 0$. Initially we set $p(x,0) = \pi(x,0)$ and during an integration step for the interval $[t, t+h)$ we start with a subset $S^{(t)}$ of states that have significant probability at time t, i.e.,

$$S^{(t)} := \{x \mid p(x,t) > \delta_1\}$$

where $\delta_1 > 0$ is a small threshold. For all states not in $S^{(t)}$ we let $p(x,t) = 0$. During the numerical integration we add new states to $S^{(t)}$ whenever they receive a significant

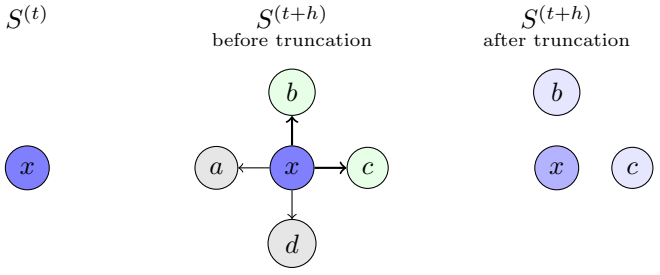


Figure 2.3: ↩ Illustration of the truncation technique. Left: the state space at time t, $S^{(t)}$. Center: all the possible states with non-zero inflow. Green states b and c obtain the inflow larger then the threshold δ, states a and d do not. Right: the state space $S^{(t+h)}$ after the truncation. Less intensive color of the state x shows that it has an outflow to the states b and c.

amount of inflow, i.e., if we use the explicit Euler method, the new state probability at time $t + h$ is calculated as

$$p(x, t + h) = p(x, t) + h \cdot \sum_{j=1}^{N_R} \left(\alpha_j(x - v_j)p(x - v_j, t) - \alpha_j(x)p(x, t)\right).$$

For a state $x \notin S^{(t)}$ this reduces to

$$p(x, t + h) = h \cdot \sum_{j=1}^{m} \alpha_j(x - v_j)p(x - v_j, t),$$

so only the inflow counts. One can iterate over all states in $S^{(t)}$ and, before integrating their probability, check whether their successors receive significant inflow. More precisely, we simply add state x to the set $S^{(t)}$ if $h \cdot \alpha_j(x - v_j)p(x - v_j, t) > \delta_2$ for some j. Here, δ_2 is again a small threshold. We then also compute $p(x, t + h)$ for this new state.

It turns out that for most example networks an accurate approximation is obtained if we work with a single threshold $\delta_1 = \delta_2 =: \delta$ and choose $\delta \in \{10^{-10}, 10^{-9} \ldots, 10^{-5}\}$. Note that the new set $S^{(t+h)}$ will then contain all states $x \in S^{(t)}$ whose probability at time $t + h$ is at least δ as well as all successors $x + v_j \notin S^{(t)}$ where $x \in S^{(t)}$ and there exists the reaction with index j such that $h \cdot \alpha_j(x - v_j)p(x - v_j, t) > \delta$ (which implies that their probability at time $t + h$ is at least δ). This process is illustrated in Figure 2.3.

Different truncation strategies are possible (e.g. choose δ_2 smaller than δ_1). However, simply adding all successors ($\delta_2 = 0$) is not efficient since often we have reversible reactions, i.e., $v_j = -v_k$ for some $j \neq k$ where one direction is much more likely then the other, say R_j. In such a case the main part of the probability mass moves in the direction of v_j and the accuracy gain in adding a successor w.r.t. v_k is not worth the effort since during the next construction of $S^{(t)}$ these successors are anyway removed from $S^{(t)}$.

This method makes the most sense to be used with the systems that possess the unbounded (or bounded, but reasonably large) state space. Consider the following biological model 2.3 of exclusive switch:

Biological model 2.3 (Exclusive Switch)**.** ↩ We consider a gene regulatory network called exclusive switch [153]. It describes the dynamics of two genes with an overlapping promoter region, and their products P_1 and P_2. Molecules of both species P_1 and P_2 are produced if no transcription factor is bound to the promotor region (region is free). However if a molecule of type P_1 (P_2) is bound to the promotor then it inhibits the expression of the other product, i.e., only molecules of P_2 (P_1) can be produced. Only one molecule can be bound to the promotor region at a time which implies that the promoter region has only three different states (free, P_1 or P_2 bound). The model has an *infinite* state space and the stoichiometric equations are

$$
\begin{array}{lll}
DNA \xrightarrow{c_1} DNA + P_1 & DNA + P_1 \xrightarrow{c_5} DNA.P_1 & \\
DNA \xrightarrow{c_2} DNA + P_2 & DNA + P_2 \xrightarrow{c_6} DNA.P_2 & DNA.P_1 \xrightarrow{c_9} DNA.P_1 + P_1 \\
P_1 \xrightarrow{c_3} \emptyset & DNA.P_1 \xrightarrow{c_7} DNA + P_1 & DNA.P_2 \xrightarrow{c_{10}} DNA.P_2 + P_2 \\
P_2 \xrightarrow{c_4} \emptyset & DNA.P_2 \xrightarrow{c_8} DNA + P_2 &
\end{array}
$$

The reaction rate constants $c_1, \ldots, c_{10}$ are given by the entries of the vector $c = (2.0, 5.0, 0.005, 0.005, 0.005, 0.002, 0.02, 0.02, 2.0, 5.0)$ and the initial conditions are such that only one DNA molecule is present in the system while the molecular counts for the rest of species are zero.

In order to illustrate the method performance, we list the size of the truncated state space together with the total loss of probability mass and show the solution for the Exclusive Switch model (cf. Example 2.6):

Example 2.6. ↩ In Figure 2.4 we plot the results of a direct numerical simulation until $t = 100$ using a dynamical state space as explained above.

The different subfigures show the marginal distributions of protein counts P_1 and P_2 when we condition on the three different states of the promotor region.

To investigate the accuracy of the obtained results we refer to the Table 2.1, where we list the amount ϵ of probability mass lost during the computation and the average size $|S|$ of the number of significant states for different thresholds δ.

Note that the probability of all states not in $S^{(t)}$ is approximated as zero. Thus, the probability loss ϵ is equal to the total approximation error (sum of all state-wise errors). If $\pi(x,t)$ is the true state probability and $p(x,t)$ is the approximated state probability then

$$\left| \sum_x \pi(x,t) - \sum_x p(x,t) \right| \leq \epsilon, \tag{2.6}$$

where we have equality at the final time instant of the computation. We find that, for instance, if we choose $\delta = 10^{-15}$ the total approximation error ϵ remains below 10^{-10}.

Recently, the authors of [136] deeply investigated the approximation error of different numerical approaches to solve the CME.

δ	$\lvert S\rvert$	ϵ	time (sec)
10^{-10}	183210	$3 \cdot 10^{-6}$	154
10^{-12}	203948	$2 \cdot 10^{-8}$	174
10^{-15}	265497	$9 \cdot 10^{-11}$	239
10^{-20}	381374	$1 \cdot 10^{-13}$	1027

Table 2.1: ↩ Dynamical state space truncation results for the exclusive switch.

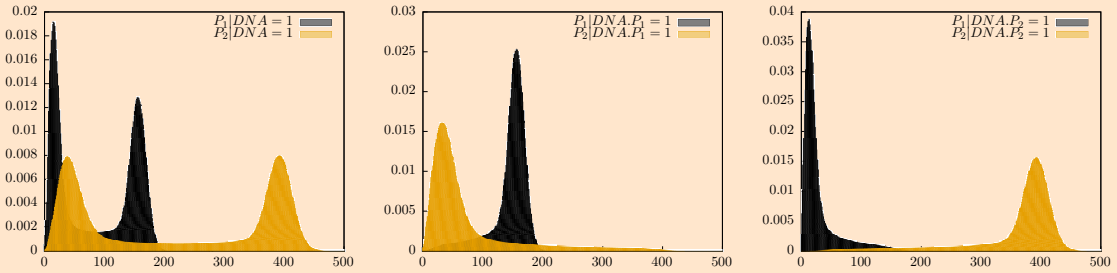


Figure 2.4: ↩ Probability (y-axis) distribution of the protein counts P_1 and P_2 (x-axis) conditioned on the events that the promotor region is free (left) bound to P_1 (center) and bound to P_2 (right) computed at time instant $t = 100$ for exclusive switch system.

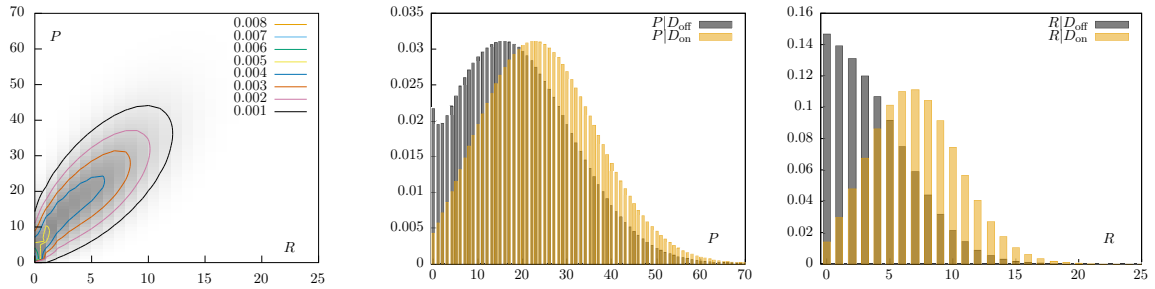
Consider the following biological model of gene expression:

Biological model 2.4 (Gene Expression)**.** ↩ We consider a simple gene expression model that describes the formation of mRNA (R) and protein (P) molecules [105]. The production of R is controlled by the state of the DNA which can be either active (D_{on}) or inactive (D_{off}). We assume a single copy of the corresponding gene, i.e., it always holds that $D_{\text{on}} + D_{\text{off}} = 1$. If $D_{\text{on}} = 1$ then mRNA molecules are synthesized and can be further translated into proteins. The proteins induce the activation of the DNA, forming a positive feedback mechanism. Moreover, mRNA and proteins can degrade. The chemical reactions are as follows

$$D_{\text{off}} \underset{\tau_{\text{off}}}{\overset{\tau_{\text{on}}}{\rightleftharpoons}} D_{\text{on}} \qquad R \xrightarrow{\gamma_r} \emptyset$$

$$D_{\text{on}} \xrightarrow{k_r} D_{\text{on}} + R \qquad P \xrightarrow{\gamma_p} \emptyset$$

$$R \xrightarrow{k_p} R + P \qquad P + D_{\text{off}} \xrightarrow{\tau_{\text{on}}^{p}} P + D_{\text{on}}$$



(a) (b) (c)

Figure 2.5: ↩ Gene expression example: two-dimensional marginal distribution of mRNA and proteins (a), conditional distributions of proteins (b) and mRNA (c) computed by directly solving the CME. Density of gray color corresponds to probability mass for (a) and for (b) and (c) y-axis corresponds to probability values.

The accuracy of the direct solution of the corresponding CME is investigated in Example 2.7.

Example 2.7. ↩ A state of the CTMC that corresponds to the gene expression model is a vector $x = (x_{D_{\text{off}}}, x_{D\text{on}}, x_R, x_P)$, where $x_{D_{\text{off}}}, x_{D\text{on}} \in \{0, 1\}$ are the populations of D_{off} and D_{on} and $x_R, x_P \in \mathbb{N}$ is the number of mRNA and protein molecules, respectively. We plot the two-dimensional marginal distributions of mRNA and proteins at time $t = 10$ in Figure 2.5 (darker points correspond to larger probability values), where we chose rate constants $(\tau_{\text{on}}, \tau_{\text{off}}, k_r, k_p, \gamma_r, \gamma_p, \tau_{on}^p) = (1, 1, 10, 1, 4, 1, 0.015)$ as in [105]. For the initial state we chose $x_0 = (1, 0, 0, 0)$. The other two plots in Figure 2.5 show the distributions of mRNA and proteins after conditioning on the two states of the DNA. The distributions were computed by integrating the CME based on a fourth-fifth order Runge-Kutta method with the adaptive step size. We chose $\delta = 10^{-15}$ as truncation threshold, yielding a total error of $\epsilon = 3 \cdot 10^{-10}$ at time instant $t = 10$.

2.3.4 Moment-based Reconstruction Techniques

Assume that is is possible to obtain the values of moments $\mu_0, \ldots, \mu_M$ of the distribution $\pi(T)$ without solving the CME directly (for example, using the moment closure methods). These moments already give us a certain information about the average values of species (via expectation) and their range (via covariances). In addition to that, we can try to solve the following problem: given the values of moments $\mu_0, \ldots, \mu_M$, approximate the corresponding distribution $p(T)$. This problem is usually called the *inverse moment problem*. For example, we can uniquely reconstruct the normal distribution $\mathcal{N}(\mu, \sigma)$ given its mean μ and variance σ in the continuous state space setting. The Poisson distribution is uniquely defined by its mean λ in the discrete state setting.

When CME for the given chemical reaction network is to be solved, there is no evidence of whether there is an analytical model that describes the transient solution properly. Therefore, the flexible models such as generalized exponential family (cf. Appendix A.5.7) can be used in order to incorporate many different behaviors such as non-symmetry and multimodality. In this thesis, we use the maximum entropy principle and the corresponding family of distributions to cope with this problem. The details on the moment approximation and the distribution reconstruction methods are given in Chapter 3.

Chapter 3

Moment Approximation Methods

The main difficulty arising in the numerical solution of the CME is the *curse of dimensionality*: each chemical species that is involved in a reaction adds one dimension to the state space of the Markov chain. Tight bounds for molecular counts are usually not known a priori and thus the size of the state space that has to be considered is extremely large or even infinite rendering a direct numerical integration of the CME (2.5) infeasible. The main idea is to concentrate on those population vectors containing a significant amount of probability mass. If in a reaction network some chemical species are highly abundant then the support of the underlying probability distribution of the process becomes very large even when insignificant parts are truncated (e.g. if we consider the sets $\{x \mid p(x) > \delta\}$ for some small δ). In such a case it is advantageous to change the representation of the distribution. The idea of *methods of moments* or *moment closure* methods is to replace the distribution of the Markov chain by its moments up to a certain finite order [5, 67, 161, 236]. It is possible to derive differential equations that can be used to approximate the time evolution of moments. For instance, if the distribution of the chain is similar to a multivariate normal distribution, one can obtain a very accurate approximation of the distribution by tracking the average molecule counts and their variances and covariances over time. For systems exhibiting more complex behavior such as oscillations or multimodality, moments of higher order are necessary for an accurate description [5]. It is, for instance, straightforward to derive a differential equation for the time evolution of the first-order moments $E[\vec{X}(t)]$. For this we derive a system of differential equations for the moments along the lines of Ale et al. [5]. For simplicity, here we restrict ourselves to the first two moments.

3.1 Moment Approximation

3.1.1 Evolution of the Mean

Consider the expectation $E\left(\vec{X}\right)$ of the random vector $\vec{X}^{(t)}$, where i-th entry denotes the expected number of molecules of type S_i at time t. The expectation is given by

$$E\left(\vec{X}\right) = \sum_{x \in \mathbb{Z}_S^N} x \cdot \mathbf{P}[\vec{X}(t) = x]$$

Let apply the test function $f : \mathbb{Z}_S^N \to \mathbb{R}_S^N$ that is independent of t to the both sides

$$E\left(f(\vec{X})\right) = \sum_{x \in \mathbb{Z}_S^N} f(x) \cdot \mathbf{P}[\vec{X}(t) = x]$$

By taking the derivatives w.r.t. time on both sides we obtain

$$\begin{aligned} \tfrac{d}{dt} E\left(f(\vec{X})\right) &= \sum_{x \in \mathbb{Z}_S^N} f(x) \cdot \tfrac{d}{dt} \mathbf{P}[\vec{X}(t) = x] \\ &= \sum_{j=1}^{N_R} E\left(\alpha_j(\vec{X}) \cdot (f(\vec{X} + v_j) - f(\vec{X}))\right). \end{aligned} \tag{3.1}$$

For $f(x) = x$ this yields a system of equations for the population means

$$\tfrac{d}{dt} E\left(\vec{X}\right) = \sum_{j=1}^{N_R} v_j \cdot E\left(\alpha_j(\vec{X})\right). \tag{3.2}$$

The expected propensity of reaction R_j is given by

$$E\left(\alpha_j(\vec{X})\right) = \begin{cases} c_j & \text{if } \alpha_j(x) = c_j, \\ c_j \cdot E\left(\vec{X}_i(t)\right) & \text{if } \alpha_j(x) = c_j \cdot x_i, \\ c_j \cdot E\left(\vec{X}_i(t) \cdot \vec{X}_k(t)\right) & \text{if } \alpha_j(x) = c_j \cdot x_i \cdot x_k, \\ \frac{1}{2} c_j \cdot \left(E\left(\vec{X}_i^2(t)\right) - E\left(\vec{X}_i(t)\right)\right) & \text{if } \alpha_j(x) = \frac{1}{2} c_j \cdot (x_i^2 - x_i). \end{cases}$$

In the third and fourth case we get a new equation for all expectations involving the product of two random variables since α_j is not linear in the elements of x. They again involve the expectations for which we get new equations, which leads to an infinite series of differential equations. However, if all propensity functions α_j are linear (first and second case), the equation (3.2) can be simplified to

$$\frac{d}{dt} E\left(\vec{X}\right) = \sum_{j=1}^{M} \vec{v}_j \cdot \alpha_j \left(E\left(\vec{X}\right)\right),$$

This gives a system of differential equations, which is identical to the reaction rate equations except that we consider populations instead of concentrations. In general, the solution

of the reaction rate equations is different from the average populations of the Markov chain because of the non-linearity of propensity functions. To illustrate the method, we introduce the simple protein production model:

Biological model 3.1 (Simple Protein Production)**.** ↩ Consider the following chemical reaction network

$$\begin{aligned} \emptyset &\xrightarrow{c_1} \text{mRNA} \\ \text{mRNA} &\xrightarrow{c_2} \text{mRNA} + \text{protein} \\ \text{mRNA} &\xrightarrow{c_3} \emptyset \\ \text{protein} &\xrightarrow{c_4} \emptyset \end{aligned}$$

It involves two chemical species which are represented by mRNA and protein and describes the over-simplified process of the protein production. More formally, we write the chemical reaction network as

$$\begin{aligned} \emptyset &\xrightarrow{c_1} R \\ R &\xrightarrow{c_2} R + P \\ R &\xrightarrow{c_3} \emptyset \\ P &\xrightarrow{c_4} \emptyset \end{aligned}$$

Now we describe the evolution of the means for this model:

Example 3.1. ↩ Assume that the initial distribution of the CTMC $\vec{X}(t)$ that corresponds to the simple protein production is given by $\pi(0) = [1, 0, \ldots]$, where the first element of the vector is $\pi_{s_0}(0) = 1$ and the initial state $s_0 = (0, 0)$. The state space of $\vec{X}(t)$ is unbounded and given by $\mathbb{Z}^2$. According to the initial distribution of the CTMC, we initialize the expected populations as $E\left(\vec{X}(0)\right) = (0, 0)$. We compute $E\left(\vec{X}\right)$ by solving

$$\begin{aligned} \tfrac{d}{dt}E\left(\vec{X}\right) &= \sum_{j=1}^{N_R} \vec{v}_j \cdot E\left(\alpha_j(\vec{X})\right) = \sum_{j=1}^{4} \vec{v}_j \cdot \alpha_j\left(E\left(\vec{X}\right)\right) \\ &= (1,0)\cdot c_1 + (0,1)\cdot c_2 \cdot E\left(\vec{X}_1(t)\right) \\ &\quad +(-1,0)\cdot c_3 \cdot E\left(\vec{X}_1(t)\right) + (0,-1)\cdot c_4 \cdot E\left(\vec{X}_2(t)\right), \end{aligned}$$

where X_1 corresponds to mRNA species and X_2 corresponds to proteins. This ODE system can be simplified to

$$\begin{aligned} \tfrac{d}{dt}E\left(\vec{X}_1(t)\right) &= c_1 - c_3 \cdot E\left(\vec{X}_1(t)\right) \\ \tfrac{d}{dt}E\left(\vec{X}_2(t)\right) &= c_2 \cdot E\left(\vec{X}_1(t)\right) - c_4 \cdot E\left(\vec{X}_2(t)\right). \end{aligned}$$

3.1.2 Evolution of the Covariance and Higher Order Moments

For most networks the system of ODEs to approximate moments is not finite but we can consider the *closure* of the system which is based on the Taylor expansion of the function $\alpha_j\left(\vec{X}(t)\right)$ about the mean $E\left(\vec{X}(t)\right)$. Let us write $\mu_i(t)$ for $E\left(\vec{X}_i(t)\right)$ and $\mu(t)$ for the vector with entries $\mu_i(t)$, $1 \leq i \leq N_S$. Then

$$\begin{aligned} E\left(\alpha_j(\vec{X})\right) &= \alpha_j(\mu) + \tfrac{1}{1!}\sum_{i=1}^{n} E\left(\vec{X}_i - \mu_i\right) \tfrac{\partial}{\partial x_i}\alpha_j(\mu) \\ &\quad + \tfrac{1}{2!}\sum_{i=1}^{n}\sum_{k=1}^{n} E\left((\vec{X}_i - \mu_i)(\vec{X}_k - \mu_k)\right) \tfrac{\partial^2}{\partial x_i \partial x_k}\alpha_j(\mu) + \dots \end{aligned}$$

where we omitted t in the equation to improve the readability. Note that $E\left(\vec{X}_i(t) - \mu_i\right) = 0$ and since we restrict to reactions that are at most bimolecular, all terms of order three and more disappear. By letting C_{ik} be the covariance $E((X_i(t) - \mu_i)(X_k(t) - \mu_k))$ we get

$$E\left(\alpha_j(\vec{X})\right) = \alpha_j(\mu) + \tfrac{1}{2}\sum_{i=1}^{N_S}\sum_{k=1}^{N_S} C_{ik}\tfrac{\partial^2}{\partial x_i \partial x_k}\alpha_j(\mu) \tag{3.3}$$

Next, we derive an equation for the covariances by first exploiting the relationship

$$\frac{d}{dt}C_{ik} = \frac{d}{dt}E\left(\vec{X}_i\vec{X}_k\right) - \frac{d}{dt}\left(\mu_i\mu_k\right) = \frac{d}{dt}E\left(\vec{X}_i\vec{X}_k\right) - \left(\frac{d}{dt}\mu_i\right)\mu_k - \mu_i\left(\frac{d}{dt}\mu_k\right) \tag{3.4}$$

and if we couple this equation with equations for the means, the only unknown term that remains is the derivative $\frac{d}{dt}E\left(\vec{X}_i\vec{X}_k\right)$ of the second moment. We can apply the same strategy as before by using Eq. (3.1) and doing the Taylor expansion about the mean for the test function $f_j(x) := \alpha_j(x)x_i$ for the corresponding terms

$$\tfrac{d}{dt}E\left(\vec{X}_i\vec{X}_k\right) = \sum_{j=1}^{N_R}\left(v_{j,i}v_{j,k}E\left(\alpha_j(\vec{X})\right) + v_{j,k}E\left(\alpha_j(\vec{X})\vec{X}_i\right) + v_{j,i}E\left(\alpha_j(\vec{X})\vec{X}_k\right)\right), \tag{3.5}$$

where $v_{j,i}$ and $v_{j,k}$ are the corresponding entries of the vector v_j. We can use Eq. (3.3) for the term $E\left(\alpha_j(\vec{X})\right)$, while the terms $E\left(\alpha_j(\vec{X})\vec{X}_i\right)$ and $E\left(\alpha_j(\vec{X})\vec{X}_k\right)$ have to be replaced by the corresponding Taylor series about the mean. Let $f_j(x) := \alpha_j(x)x_i$. Similar to Eq. (3.3), we get that $E\left(\alpha_j(\vec{X})\vec{X}_i\right)$ equals

$$\begin{aligned} E\left(\alpha_j(\vec{X})\vec{X}_i\right) &= \alpha_j(\vec{\mu})\mu_i + \tfrac{1}{1!}\sum_{i=1}^{N_S} E\left(\vec{X}_i - \mu_i\right)\tfrac{\partial}{\partial x_i}f_j(\vec{\mu}) \\ &\quad + \tfrac{1}{2!}\sum_{i=1}^{N_S}\sum_{k=1}^{N_S} E\left((\vec{X}_i - \mu_i)(\vec{X}_k - \mu_k)\right)\tfrac{\partial^2}{\partial x_i \partial x_k}f_j(\vec{\mu}) + \dots \end{aligned} \tag{3.6}$$

Here, it is important to note that moments of order three come into play, since derivatives of order three of $f_j(x) = \alpha_j(x)x_i$ may be nonzero. It is possible to take these terms into account by deriving additional equations for moments of order three and higher. Obviously, these equations will then include moments of even higher order such that theoretically we end up with an infinite system of equations. However, a popular strategy is to *close* the

equations by assuming that all moments of order $> M$ that are centered around the mean are equal to zero. E.g. if we choose $M = 2$, then we can simply use the approximation

$$E\Big(\alpha_j(\vec{X})X_i\Big) \approx \alpha_j(\vec{\mu})\mu_i + \frac{1}{2!}\sum_{i=1}^{N_S}\sum_{k=1}^{N_S} E\Big((\vec{X}_i - \mu_i)(\vec{X}_k - \mu_k)\Big)\frac{\partial^2}{\partial x_i \partial x_k} f_j(\vec{\mu}).$$

This approximation is then inserted into Eq. (3.5) and the result is used to replace the term $\frac{d}{dt}E\Big(\vec{X}_i\vec{X}_k\Big)$ in Eq. (3.4). Finally, we can integrate the time evolution of the means and that of the covariances and variances.

The moments of higher order can be approximated using this approach by choosing the appropriate test function $f(x)$. To truncate possibly infinite system of ODEs we apply low dispersion closure (cf. Section 3.1.3).

To illustrate the method, we consider the following examples for the simple dimerization (cf. Example 3.2), gene expression (cf. Example 3.3) and bursty protein production (cf. Example 3.4) reaction systems.

Example 3.2 (Moment Closure for Simple Dimerization Model)**.** ↩ Assuming that all central moments of order three and higher are equal to zero, we get the following equations for the means, variances and the covariance of the species

$$\begin{aligned}
\tfrac{d}{dt}\mu_1 &= -c_1\mu_1\left(\mu_1 - 1\right) - c_1 C_{1,2} + 2c_2\mu_2 \\
\tfrac{d}{dt}\mu_2 &= \tfrac{1}{2}c_1\mu_1\left(\mu_1 - 1\right) + \tfrac{1}{2}c_1 C_{1,2} - c_2\mu_2 \\
\tfrac{d}{dt}C_{1,1} &= -2c_1\mu_1^3 + 4c_1\mu_1^2 - 2c_1\mu_1 + 4c_2\mu_2 + 4c_2\mu_1\mu_2 \\
\tfrac{d}{dt}C_{1,2} &= -\tfrac{3}{2}c_1 C_{1,1} + \mu_1\left(-c_1(\mu_1 - 1) - \tfrac{1}{2}c_1\mu_1 C_{1,1} + c_2\mu_1^2\right) \\
&\quad + \mu_2\left(-2c_2 - c_2\mu_1 + c_1 C_{1,1}\right) \\
\tfrac{d}{dt}C_{2,2} &= \tfrac{3}{2}c_1 C_{1,1} + \tfrac{1}{2}c_1\mu_1(\mu_1 - 1) + \mu_2(c_2 - c_1 C_{1,1})
\end{aligned}.$$

Here we denote the expectations of species P (P_2) by μ_1 (μ_2), variances are given by $C_{1,1}$, $C_{2,2}$ and the covariance between P and P_2 is $C_{1,2}$. In the equations we omit t to improve readability.

Example 3.3 (Moment Closure for Gene Expression Model)**.** ↩ We apply the moment closure technique described above to the gene expression system. When we consider only the moments up to second order, the corresponding equations for the average number of molecules are given by

$$\frac{d}{dt}\mu_{D_{\text{off}}} = \tau_{\text{off}}\mu_{D_{\text{on}}} - E\left(\tau_{\text{on}}^{p} X_{D_{\text{off}}} X_P\right) \qquad \frac{d}{dt}\mu_R = k_r\mu_{D_{\text{on}}} - \gamma_r\mu_R$$
$$\frac{d}{dt}\mu_{D_{\text{on}}} = \tau_{\text{on}}\mu_{D_{\text{off}}} + E\left(\tau_{\text{on}}^{p} X_{D_{\text{off}}} X_P\right) \qquad \frac{d}{dt}\mu_P = k_p\mu_R - \gamma_p\mu_P,$$

where $\mu_{D_{\text{off}}}, \mu_{D_{\text{on}}}$ are the expected numbers of D_{off} and D_{on} respectively, and μ_R, μ_P are the expected numbers of mRNA and proteins.

Biological model 3.2 (Bursty Protein Production)**.** ↩ Consider a simple model of gene expression with enzymatic degradation [219]

$$\varnothing \xrightarrow{k_0} M \xrightarrow{k_1} \varnothing$$
$$M \xrightarrow{k_2} M + P$$
$$P + E \underset{k_4}{\overset{k_3}{\rightleftharpoons}} C \xrightarrow{k_5} E.$$

The detailed description of the corresponding biological functions can be found in [77, 195], where it is noted that the active protein degradation influences the protein distribution such that they are skewed towards high molecule numbers. Assuming binding and unbinding of P and E are fast compared to protein degradation, the authors of [77, 222] simplify this chemical reaction network in the following way

$$P \xrightarrow{f(x_P)} \varnothing,$$

with $f(x_P) = \frac{\Omega v_M x_P}{\Omega K_M + x_P}$, where x_P is the number of proteins and $K_M = \frac{k_4+k_5}{k_3}$ is the Michaelis-Menten constant. Here rates are set to $k_0 = 8$, $k_1 = 10$, $k_2 = 100$, $v_M = 100$, $K_M = 20$ and the volume of the system is set to $\Omega = 1$. The reaction rates are chosen such that the proteins are produced in bursts of size $b = k_2/k_1$. The complete chemical reaction network with the Michaelis-Menten type kinetics for P, $f(x_P) = \frac{v_M x_P}{K_M + x_P}$ is given by

$$\varnothing \xrightarrow{k_0} M \xrightarrow{k_1} \varnothing$$
$$M \xrightarrow{k_2} M + P$$
$$P \xrightarrow{f(x_P)} \varnothing.$$

Example 3.4 (Moment Closure for Bursty Protein Production). ↩ Assuming that all central moments of order three and higher are equal to zero, we get the following equations for the means and covariances of the species

$$\begin{aligned}
\tfrac{d}{dt}\mu_M &= \Omega k_0 - k_1\mu_M \\
\tfrac{d}{dt}\mu_P &= k_1\mu_M - f(\mu_P) - \tfrac{1}{2}\tfrac{\partial^2 f(\mu_P)}{\partial {\mu_P}^2} C_{P,P} \\
\tfrac{d}{dt}C_{M,M} &= k_0 + k_1\mu_M - 2k_1 C_{M,M} \\
\tfrac{d}{dt}C_{M,P} &= -k_1 C_{M,P} + k_2 C_{M,M} - \left(\tfrac{\partial f(\mu_P)}{\partial \mu_P}\right) C_{M,P} \\
\tfrac{d}{dt}C_{P,P} &= 2k_2 C_{M,P} + f(\mu_P) + \left(\tfrac{1}{2}\tfrac{\partial^2 f(\mu_P)}{\partial {\mu_P}^2} - 2\,\tfrac{\partial f(\mu_P)}{\partial \mu_P}\right) C_{P,P},
\end{aligned} \tag{3.7}$$

where the expectations of the species are given by μ_M and μ_P, variances are given by $C_{M,M}$ and $C_{P,P}$ and the covariance between M and P is denoted by $C_{M,P}$. The derivatives of the rate function $f(x_P)$ are given by $\frac{\partial f(\mu_P)}{\partial \mu_P} = \frac{v_M K_M}{(K_M + x_P)^2}$ and $\frac{\partial^2 f(\mu_P)}{\partial {\mu_P}^2} = -2\frac{v_M K_M}{(K_M + x_P)^3}$.

3.1.3 Other Closure Techniques

Possibly infinite system of differential equations obtained using moment closure approach needs to be truncated. In the course of this thesis we rely on the closure method where we assume that all central moments of the order $> M$ are zero. It is referred in the literature as to *low dispersion closure*.

Here, we provide the general idea of the moment closure and short description of some of the existing techniques. Given the order vector of non-negative integers $\mathbf{I} = (I_1, \ldots, I_{N_S})$, the uncentered moment of $\vec{X}$ is given by $\mu^{(\mathbf{I})} = E\left(\vec{X}^{(I_1)}\vec{X}^{(I_2)}\cdots\vec{X}^{(I_{N_S})}\right)$. The *order* of this moment is $|\mathbf{I}| = \sum_{j=1}^{N_S} I_j$. The evolution of the vector μ of moments of order $\leq k$ is governed by the exact k-th order dynamics [200]

$$\frac{d}{dt}\mu = A\mu + B\bar{\mu}, \tag{3.8}$$

where $\mu \in \mathbb{R}^K$ and $K \leq k$ since there are many moments of the same order. The vector $\bar{\mu} \in \mathbb{R}^{\bar{K}}$ contains moments of order $> k$ (k is called the order of truncation). The term $B\bar{\mu}$ appears to be non-zero when there are bi-molecular reactions in chemical reaction network so that the system of ODEs is *open* since it depends on $\bar{\mu}$. Using the *moment closure* procedure, we can approximate the dynamics of μ by the one of ν:

$$\frac{d}{dt}\nu = A\nu + B\phi(\nu), \tag{3.9}$$

where $\nu \in \mathbb{R}^K$ and $\phi(\nu)$ is the finite approximation of $\bar{\mu}$. The idea of the moment closure techniques is to find such closure function $\phi(\nu)$ that the solution ν in (3.8) is nicely approximates to the solution μ in (3.9).

Various methods can be used to close the system of equations, such as derivative matching and zero cumulants closures [112, 167, 200, 236] as well as beta-binomial [139, 140] (that addresses the problem of highly skewed population distributions), maximum entropy [203] and Dirac [108] distribution assumptions.

Mean-Field. The evolution of the mean of populations is analyzed using mean-field approach. The higher-order moments $E\left(\vec{X}_i\vec{X}_j\cdots\vec{X}_k\right)$ are approximated using the independence assumption, i.e., $E\left(\vec{X}_i\vec{X}_j\cdots\vec{X}_k\right) = E\left(\vec{X}_i\right)E\left(\vec{X}_j\right)\cdots E\left(\vec{X}_k\right)$, so that the covariances are ignored. When the amount of molecules is high, this introduces the accurate approximation for the expected populations. However, for some models (such as circadian clock) it does not perform well.

Zero Cumulants. It is assumed that the populations are distributed according to the multivariate normal, so that all the moments of order > 2 can be expressed in terms of means and covariances according to Isserlis theorem [117].

Derivative Matching. The assumption is made that the populations are distributed according to the log-normal distribution [200]. It provides the straightforward multiplicative way of closing higher-order moments.

Methods based on the *mass fluctuation* analysis are considered in [22, 89]. Recently, Lakatos et al. presented the method [145] to conduct the closure using multidimensional normal, log-normal and gamma distributions. Other group of methods are based on the *system size expansion* [94, 194, 221] (that takes its roots in Van Kampen's work [126]). The comparison of the mean-field, normal closure, min-normal and log-normal closures for the performance models is provided in [97]. The dynamic switching between different closure methods (namely, derivative matching, zero cumulants, zero variance and low dispersion) for the parameter identification task is considered in [35]. It can also happen that the system of ODEs is stiff and it requires the specialized solvers to be applied. The authors of [249] propose to use so-called convergent moment to mitigate the stiffness problem.

Most of the above mentioned closures were recently implemented in the tool CERENA [128]. The (non-complete) list of the available software to conduct the moment closure analysis is given in Appendix A.4.

3.1.4 Accuracy of Moment Closure Approximation

For non-complicated systems such as the simple dimerization we find that the approximation provided by the moment closure method is very accurate even if only 2 moments are considered. In general, however, numerical results show that the approximation tends to become worse if systems exhibit complex behavior such as multi-stability or oscillations. Increasing the number of moments typically improves the accuracy [5] but it may happen that the resulting equations become very stiff [67].

Grima has investigated the accuracy of the approximation for $|\mathbf{I}| = 2$ and $|\mathbf{I}| = 3$ by a comparison with the system size expansion of the master equation [94]. He found that

for monostable systems with large volumes, the approximation of the means $E[\vec{X}(t)]$ has a relative error that scales as $\Omega^{-|\mathbf{I}|}$, while the relative errors of the variances and covariances scale as $\Omega^{-(|\mathbf{I}|-1)}$, $|\mathbf{I}| \in \{2, 3\}$. Detailed comparison of existing methods was recently presented by Schnoerr et al. [194]. The authors of [147] provide the theoretical and numerical results about the errors that come from the truncation of moment ODE system when standard moment closure approach is applied. For small volumes or systems with multiple modes, however, only numerical evaluations of the accuracy are available [5, 67], where the approximated moments are compared to statistical estimates based on Monte Carlo simulations of the process. Recently, it was shown that it is possible to derive the exact lower and upper bounds on the moments in equilibrium state [76].

Here we investigate the accuracy of the low dispersion closure using several biological models. In Example 3.5 we consider the simple dimerization:

Example 3.5 (Accuracy of Moment Closure for Simple Dimerization)**.** ↩ The comparison between the moments computed using direct CME integration and obtained using moment closure technique for the simple dimerization model is provided in Table 3.1, where N_{eq} denotes the number of equations in the ODE system. The integration of the moment equations takes less than one second for this example. The initial protein numbers are chosen as $P = 301$ and $P_2 = 0$ and we consider the system at time $t = 20$ [5]. We find that the moment closure approximation provides very accurate results.

$\|\mathbf{I}\|$	N_{eq}	$\epsilon_{\|\mathbf{I}\|=1}$	$\epsilon_{\|\mathbf{I}\|=2}$	$\epsilon_{\|\mathbf{I}\|=3}$	$\epsilon_{\|\mathbf{I}\|=4}$	$\epsilon_{\|\mathbf{I}\|=5}$
2	5	0.001754	0.003495	-	-	-
3	9	0.001752	0.003492	0.005215	-	-
4	14	0.001743	0.003465	0.005211	0.006907	-
5	20	0.001721	0.003418	0.005183	0.006901	0.008555

Table 3.1: ↩ Errors of the moment closure approximation for the dimerization chemical reaction network.

The first column in Table 3.1 refers to the highest order $|\mathbf{I}|$ of the moments that were considered during the computation, i.e., all central moments of higher order are approximated with zero. In addition we list the relative errors of the means ($\epsilon_{|\mathbf{I}|=1}$) and the moments of the higher order k ($\epsilon_{|\mathbf{I}|=k}$) using the error norm

$$\epsilon_{|\mathbf{I}|=k} = \max_{i \in \{1,\dots,N_S\}} \frac{|\hat{m}_i^{(k)} - m_i^{(k)}|}{m_i^{(k)}}. \tag{3.10}$$

Here, $\hat{m}_i^{(k)}$ and $m_i^{(k)}$ are the values of moments $E(X_i^k)$ computed using the moment closure method and obtained via direct numerical simulation and the maximum is taken over the chemical species. The order vector is $\mathbf{I} = (0, \dots, 0, k, 0, \dots, 0)$ both for $\hat{m}_i^{(k)}$ and $m_i^{(k)}$. The second column in tables refers to the number of equations N_{eq} that were integrated for the moment closure method.

Let us introduce another model of multi-attractor for which we obtain the larger equation system that governs the dynamics of moments:

Biological model 3.3 (Multi-attractor Model). ↩ The multi-attractor model [251] consists of 23 chemical reactions (listed in Appendix A.7) and describes the dynamics of three genes and the corresponding proteins. The proteins *PaxProt*, *MAFAProt* and *DeltaProt* are able to bind to the promotor regions of the DNA and activate or suppress the production of other proteins. The model is infinite in three dimensions.

Example 3.6. ↩ We consider the accuracy of the moment closure approximation (cf. Table 3.2) in the same way as for the previous example but list the running time in addition (third column). The values of stochastic reaction constants are chosen as $c_p = 5, c_d = 0.1, c_b = 1.0, c_u = 1.0$ and we consider the system at time $t = 10$. As initial conditions we assumed one molecule for all DNA-like species ($PaxDna = 1$, $MAFADna = 1$, $DeltaDna = 1$) and the molecular counts for the remaining species are 0.

We find that the moments obtained via the moment closure approximation are accurately approximated. For instance, the average number of *MAFADna* is approximated as 19.719 while the result of the direct numerical simulation gives 19.544. Note that it takes 20634 seconds to finish the numerical simulation (the size of the truncated state space $|S| = 7736339$) whereas the moment closure approximation takes only 3649 seconds.

$\|\mathbf{I}\|$	N_{eq}	time (sec)	$\epsilon_{\|\mathbf{I}\|=1}$	$\epsilon_{\|\mathbf{I}\|=2}$	$\epsilon_{\|\mathbf{I}\|=3}$	$\epsilon_{\|\mathbf{I}\|=4}$	$\epsilon_{\|\mathbf{I}\|=5}$
2	104	2	0.043987	0.077507	-	-	-
3	559	40	0.043987	0.067790	0.104288	-	-
4	2379	443	0.043987	0.058938	0.082293	0.096345	-
5	8567	3649	0.043987	0.037542	0.066227	0.056258	0.110358

Table 3.2: ↩ Moment closure approximation results for the multi-attractor network.

Yet another model to be considered is the exclusive switch that exhibits bimodal behavior for certain parameter regions:

Example 3.7. ↩ The reaction rate constants and initial conditions are chosen as in the description of the system. Again, we first consider the accuracy of the moment closure approximation (cf. Table 3.3) at time $t = 100$ in the same way as for the previous examples. As also noted by Grima, the error of the moments that have the highest order which have been considered during the computation, is rather high [94]. Thus, in the moment closure approximation we have to consider at least all moment up to order five to accurately estimate the moments up to order four.

$\|\mathbf{I}\|$	N_{eq}	time (sec)	$\epsilon_{\|\mathbf{I}\|=1}$	$\epsilon_{\|\mathbf{I}\|=2}$	$\epsilon_{\|\mathbf{I}\|=3}$	$\epsilon_{\|\mathbf{I}\|=4}$	$\epsilon_{\|\mathbf{I}\|=5}$
2	20	< 1	0.004555	0.194240	-	-	-
3	55	< 1	0.004555	0.026281	0.060490	-	-
4	125	2	0.004555	0.020493	0.028242	0.136965	-
5	251	6	0.004555	0.017774	0.027933	0.026724	0.015461

Table 3.3: ↩ Moment closure approximation results for the exclusive switch.

Let us also investigate the accuracy of the moment closure approximation for gene expression system:

Example 3.8. ↩ We compare the approximated moments with those computed via a direct numerical integration of the CME (Table 3.4). We consider the following three cases. The moment closure approximation is carried out using all moments up to order 4, 6, and 8. For each case we list the number of moment equations, the running time, and the relative errors in the first four moments (columns 4-7).

$\|\mathbf{I}\|$	N_{eq}	time (sec)	$\epsilon_{\|\mathbf{I}\|=1}$	$\epsilon_{\|\mathbf{I}\|=2}$	$\epsilon_{\|\mathbf{I}\|=3}$	$\epsilon_{\|\mathbf{I}\|=4}$	$\epsilon_{\|\mathbf{I}\|=5}$
4	70	1	$8 \cdot 10^{-6}$	$8.3 \cdot 10^{-5}$	$9.6 \cdot 10^{-5}$	$8.24 \cdot 10^{-4}$	-
6	209	25	$2 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$1 \cdot 10^{-5}$	$3.6 \cdot 10^{-5}$	$8.5 \cdot 10^{-6}$
8	494	3726	$1 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$2 \cdot 10^{-6}$	$4 \cdot 10^{-6}$	$1.3 \cdot 10^{-6}$

Table 3.4: ↩ Moment closure approximation results for the gene expression system.

3.2 Method of Conditional Moments

In many chemical reaction networks we find a joint distribution for which a representation in terms of the moments is not ideal. For instance, many networks describing gene regulatory processes contain binding events where the number of binding sites is very low (often there is just a single binding site). Then the size of the populations of the corresponding chemical species is bounded by the number of binding sites. For instance, in case of the gene expression model we only have two possibilities: the DNA is either in the on or in the off state. Then the marginal probability distribution consists of a small number of discrete probabilities and the joint distribution is typically divided into several modes that correspond to the different binding states. In such cases it is obvious that for the small populations a moment representation is not adequate compared to considering the discrete mode probabilities. This, however, implies that one has to consider conditional moments (conditioned on the mode) for the remaining (large) populations. Here the focus is on a comparison of the underlying probability distributions that are reconstructed from the moments. In the work of Hasenauer et al. [105] equations have been derived for the integration of the mode probabilities over time (where the mode corresponds to the state of the small populations, e.g. state of binding sites, etc.) and equations to integrate the

moments of the large populations, conditioned on the mode. We refer to this as the MCM approach and in the numerical results presented in the following, we reconstruct the joint distribution based on both the MM approach and the MCM approach to see whether the reconstructed joint distributions are more accurate if conditional moments are used.

We first decompose the chemical populations described by $\vec{X}(t)$ into small and large populations. Here we assume that this decomposition is static. However, it is obvious that during the integration over time, we can (after reconstructing the joint distribution) choose a different decomposition for the remaining time. From what size on a population should be considered as small typically depends on the amount of the main memory that is available and on the maximum order of the moments that we consider for the large populations. Note that considering conditional moments yields a smaller number of equations if the order of the considered moments is high. The reason is that the number of equations for representing the dynamics of the small populations does not increase as the order of considered conditional moments increases. Also, for many systems the decomposition is obvious, as the small populations are exactly those that have a maximum size of, say, less than 10 (because they represent binding sites) and the large populations count protein numbers which may become rather high. Recently, it was shown by Ruess [187] that the size of the ODE system can be reduced even further by considering only those configurations of the system that are realistic (reachable).

Formally, we write the random vector $\vec{X}(t)$ at time t as $\vec{X}(t) = \left(\vec{Y}(t), \vec{Z}(t)\right)$, where $\vec{Y}(t)$ corresponds to the small, and $\vec{Z}(t)$ to the large populations. Similarly, we write $x = (\vec{y}, \vec{z})$ for the states of the process and $v_j = (\hat{v}_j, \tilde{v}_j)$ for the change vectors, $j \in \{1, \ldots, N_R\}$. Again, the first component refers to the small and the second component to the large populations. The CME 2.5 becomes

$$\frac{\partial \pi(\vec{y},\vec{z})}{\partial t} = \sum_{j=1}^{N_R} (\alpha_j(\vec{y}-\hat{v}_j, \vec{z}-\tilde{v}_j)\pi(\vec{y}-\hat{v}_j, \vec{z}-\tilde{v}_j) - \alpha_j(\vec{y},\vec{z})\pi(\vec{y},\vec{z})) \tag{3.11}$$

where we omitted the time parameter t to improve readability. Next, we sum over all possible $\vec{z}$ to get the time evolution of the marginal distribution $\hat{\pi}(\vec{y}) = \sum_{\vec{z}} \pi(\vec{y},\vec{z})$ of the small populations:

$$\begin{aligned} &\frac{\partial}{\partial t}\hat{\pi}(\vec{y}) = \sum_{\vec{z}} \sum_{j=1}^{N_R} \alpha_j(\vec{y}-\hat{v}_j, \vec{z}-\tilde{v}_j)\pi(\vec{y}-\hat{v}_j, \vec{z}-\tilde{v}_j) - \sum_{\vec{z}} \sum_{j=1}^{m} \alpha_j(\vec{y},\vec{z})\pi(\vec{y},\vec{z}) = \\ &\sum_{j=1}^{N_R} \hat{\pi}(\vec{y}-\hat{v}_j) E\Big[\alpha_j(\vec{y}-\hat{v}_j, \vec{Z}) \mid Y = \vec{y}-\hat{v}_j\Big] - \sum_{j=1}^{N_R} \hat{\pi}(\vec{y}) E\Big[\alpha_j(\vec{y},\vec{Z}) \mid \vec{Y} = \vec{y}\Big] \end{aligned} \tag{3.12}$$

Note that in this small master equation that describes the change of the mode probabilities over time, the sum runs only over those reactions that modify $\vec{y}$, since for all other reactions the terms cancel out. Moreover, on the right side we have only mode probabilities of neighboring modes and conditional expectations of the continuous part of the reaction rate. For the latter, we can use a Taylor expansion about the conditional population means. Similar to Eq. (3.6), this yields an equation that involves the conditional means and centered conditional moments of second order (variances and covariances). Thus, in order to close the system of equations, we need to derive equations for the time evolution

of the conditional means and centered conditional moments of higher order. Since the mode probability $\pi(\vec{y})$ may become zero, we first derive an equation for the evolution of the partial means (conditional means multiplied by the probability of the condition)

$$\begin{aligned}\frac{\partial}{\partial t}\left(E[\vec{Z} \mid \vec{y}]\ \pi(\vec{y})\right) = \sum_{\vec{z}} \vec{z}\frac{\partial}{\partial t}\pi(\vec{y},\vec{z}) = \sum_{j=1}^{N_R} E[(\vec{Z}+\tilde{v}_j)\alpha_j(\vec{y}-\hat{v}_j,\vec{Z}) \mid \vec{y}-\hat{v}_j]\ \pi(\vec{y}-\tilde{v}_j) \\ - \sum_{j=1}^{N_R} E[\vec{Z}\alpha_j(\vec{y},\vec{Z}) \mid \vec{y}]\ \pi(\vec{y}),\end{aligned}$$

where in the second line we applied Eq. (3.11) and simplified the result. The conditional expectations $E[(\vec{Z}+\tilde{v}_j)\alpha_j(\vec{y}-\hat{v}_j,\vec{Z}) \mid \vec{y}-\hat{v}_j]$ and $E[\vec{Z}\alpha_j(\vec{y},\vec{Z}) \mid \vec{y}]$ are then replaced by their Taylor expansion about the conditional means such that the equation involves only conditional means and higher centered conditional moments [105]. For higher centered conditional moments, similar equations can be derived.

If all centered conditional moments of order higher than k are assumed to be zero, the result is a closed system of differential algebraic equations (algebraic equations are obtained whenever a mode probability $\pi(\vec{y})$ is equal to zero). However, it is possible to transform the system of differential algebraic equations into a system of (ordinary) differential equations after truncating modes with insignificant probabilities. Then we can get an accurate approximation of the solution after applying standard numerical integration methods. We construct the ODE system using the tool SHAVE which implements the truncation based approach and solves it using MATLAB's `ode23` solver with the default error tolerance settings. Therefore, the overall moments are computed as follows

$$\begin{aligned}E\left[\vec{Y}_i\right] &= \sum_{\vec{y}} \vec{y}_i \pi(\vec{y}) \\ E\left[\vec{Z}_i\right] &= \sum_{\vec{y}} E\left[\vec{Z}_i|\vec{y}\right] \pi(\vec{y})\end{aligned}$$

We consider the application of method of conditional moments to the gene expression model in the next example:

Example 3.9 (MCM applied to Gene Expression model)**.** ↩ We apply the method of conditional moments to the gene expression system. The modes of the system are then given by the state of the DNA. The equations for the mode probabilities (p_{off}, p_{on}) and the expected number of mRNA ($\mu_{R,\text{off}}$, $\mu_{R,\text{on}}$) and proteins ($\mu_{P,\text{off}}$, $\mu_{P,\text{on}}$) are as follows:

$$\begin{aligned}\frac{d}{dt}p_{\text{off}} &= \tau_{\text{on}}\, p_{\text{on}} - (\tau_{\text{off}} + \tau^p_{\text{on}}\, \mu_{P,\text{off}})p_{\text{off}} & \frac{d}{dt}p_{\text{on}} &= (\tau_{\text{off}} + \tau^p_{\text{on}}\mu_{P,\text{off}})p_{\text{off}} - \tau_{\text{on}}\, p_{\text{on}} \\ \frac{d}{dt}\left(\mu_{R,\text{off}}\, p_{\text{off}}\right) &= -\gamma_r \mu_{R,\text{off}}\, p_{\text{off}} & \frac{d}{dt}\left(\mu_{R,\text{on}}\, p_{\text{on}}\right) &= (k_r - \gamma_r \mu_{R,\text{on}})p_{\text{on}} \\ \frac{d}{dt}\left(\mu_{P,\text{off}}\, p_{\text{off}}\right) &= (k_p \mu_{R,\text{off}} - \gamma_p \mu_{P,\text{off}})p_{\text{off}} & \frac{d}{dt}\left(\mu_{P,\text{on}}\, p_{\text{on}}\right) &= (k_p \mu_{R,\text{on}} - \gamma_p \mu_{P,\text{on}})p_{\text{on}}\end{aligned}$$

We computed the conditional moments and conditional probabilities over time by considering moments up to the order of 4, 6, and 8. For these three cases the number of equations, when compared to the method of moments (MM), are as follows (however, can be further decreased [187]):

moment order M	4	6	8
N_{eq}, MM	69	209	494
N_{eq}, MCM	30	56	90

The relative errors $\epsilon_{|\mathbf{I}|=k}$ (3.10) of the results of the method of conditional moments (MCM) are given in Table 3.5, where we again compared to the results obtained via a direct numerical solution.

Our experiments show that the MCM performs much faster (due to the smaller number of equations) and still yields accurate approximation of the moments. For the chosen set of parameters the MCM tends to provide a better approximation for higher moments, whereas the MM approach is more accurate for lower moments when the same number of moments is considered. For example, in the case of 6 moments the maximum relative error for the first moments computed by the MM approach is $2 \cdot 10^{-6}$, compared to $3.2 \cdot 10^{-5}$ when computed using the MCM. At the same time, the maximum relative errors of the sixth moments are $6.5 \cdot 10^{-4}$ and $2 \cdot 10^{-4}$ for the MM and the MCM respectively. Note that the (unconditional) moments for the MCM are computed via multiplication of the conditional moments with the mode probabilities and sum over all possible conditions.

We also consider another set of parameters for the gene expression kinetics. The rate constants are chosen $(\tau_{\text{on}}, \tau_{\text{off}}, k_r, k_p, \gamma_r, \gamma_p, \tau_{on}^p) = (0.05, 0.05, 10, 1, 4, 1, 0.015)$ as in [105]. For the initial states we simply use $x_{0,1} = (1, 0, 4, 10)$ and $x_{0,2} = (1, 0, 4, 10)$ with probabilities $P(x_{0,1}) = 0.7$ and $P(x_{0,1}) = 0.3$. The comparison of the moment values at time instant $t = 10$ reveals that the MCM provides a much better approximation both for high and low order moments as opposed to the first parameter set. For instance, in the case of 6 moments the maximum relative error for the first moments computed by the MM approach is 0.14 whereas in the MCM approach the error is $7.5 \cdot 10^{-5}$. The maximum relative error of the sixth moments for the MM approach is 0.28 compared to 0.02 using the MCM.

$\lvert\mathbf{I}\rvert$	N_{eq}	time (sec)	ϵ_{mode}	$\epsilon_{\lvert\mathbf{I}\rvert=1}$	$\epsilon_{\lvert\mathbf{I}\rvert=2}$	$\epsilon_{\lvert\mathbf{I}\rvert=3}$	$\epsilon_{\lvert\mathbf{I}\rvert=4}$
4	30	1	$7 \cdot 10^{-6}$	$1 \cdot 10^{-5}$	$2.86 \cdot 10^{-4}$	$1.12 \cdot 10^{-3}$	$6.98 \cdot 10^{-3}$
6	56	2	$6 \cdot 10^{-6}$	$3.2 \cdot 10^{-5}$	$5.9 \cdot 10^{-5}$	$6.8 \cdot 10^{-5}$	$2.18 \cdot 10^{-4}$
8	90	9	$2 \cdot 10^{-6}$	$4.2 \cdot 10^{-5}$	$6.2 \cdot 10^{-5}$	$7.7 \cdot 10^{-5}$	$9.1 \cdot 10^{-5}$

Table 3.5: ↩ Conditional moment closure approximation results for the gene expression system.

The application of MCM to the exclusive switch model is considered in the following example:

Example 3.10 (MCM applied to Exclusive Switch model)**.** ↩ We apply the method of conditional moments to the exclusive switch system. This system possesses three modes that correspond to three possible states of the promoter, i.e., $p_1 = P(DNA = 1)$, $p_2 = P(DNA.P_1 = 1)$ and $p_3 = P(DNA.P_2 = 1)$. We first calculate the number of equations to be solved when the standard moment closure is applied and compare that to the size of ODE system solved when MCM is used in the table below.

moment order M	4	6	8
N_{eq}, MM	125	461	1286
N_{eq}, MCM	45	84	135

As expected, for the system with more species, the benefit of applying the MCM is more obvious. We also list the relative error $\epsilon_{|\mathbf{I}|=k}$ (3.10) of the results of MCM applied to the exclusive switch system in Table 3.6, where we again compared to the results obtained via a direct numerical solution. The relative error of the approximation gradually decreases as more moments are taken into the consideration. We emphasize that the computation time for the case of 8 moments is 83 seconds whereas the application of MM is almost infeasible (the corresponding computation takes around 60 hours).

$\lvert\mathbf{I}\rvert$	N_{eq}	time (sec)	$\epsilon_{\lvert\mathbf{I}\rvert=1}$	$\epsilon_{\lvert\mathbf{I}\rvert=2}$	$\epsilon_{\lvert\mathbf{I}\rvert=3}$	$\epsilon_{\lvert\mathbf{I}\rvert=4}$
4	45	5	$1.0 \cdot 10^{-5}$	$3.7 \cdot 10^{-5}$	$4.9 \cdot 10^{-5}$	$1.5 \cdot 10^{-4}$
6	84	20	$2.4 \cdot 10^{-6}$	$3.9 \cdot 10^{-6}$	$5.1 \cdot 10^{-6}$	$6.4 \cdot 10^{-6}$
8	135	83	$7.8 \cdot 10^{-7}$	$1.4 \cdot 10^{-6}$	$2.0 \cdot 10^{-6}$	$1.9 \cdot 10^{-6}$

Table 3.6: ↩ Conditional moment closure approximation results for the exclusive switch system.

Chapter 4

Inverse Moment Problem

Moment closure methods can be used to approximate the moments of a stochastic dynamical system over time. The numerical integration of the corresponding ODE system is usually faster than a direct integration of the transient probability distribution or an estimation of the moments based on Monte-Carlo simulations of the system. However, if one is interested in certain events and only the moments of the distribution are known, the corresponding probabilities are not directly accessible and have to be reconstructed based on the moments. Here, we shortly review standard approaches to reconstruct one-dimensional marginal probability distributions $\pi_i(x_i, t) = \mathbf{P}[X_i(t) = x_i]$ of a Markov chain $\vec{X}$ that describes the dynamics of a chemical reaction network. Assume that a finite number of moments of the i-th molecular population at a fixed time t is given. Then the corresponding probability distribution is in general *not uniquely determined* since there can be several distributions that fulfill the moment constraints. For instance, the exponential distribution maximizes the entropy among all continuous distributions on $[0, \infty)$ with the same mean. Similarly, the normal distribution is chosen among all continuous distributions if mean and variance are known. In order to choose one distribution from this set, we apply the *maximum entropy* principle. In this way we minimize the amount of prior information and avoid any other latent assumption about the distribution.

Here, we use the moments up to order M to obtain the one- and two-dimensional marginal probability distributions of a reaction network. The reconstruction of the distributions of higher dimension is more involved and hence the advanced numerical techniques must be applied [3].

Having its roots in statistical mechanics and thermodynamics [119], the maximum entropy approach was successfully applied to solve moment problems in the field of climate prediction [2, 131, 186], nonlinear chaotic systems [207], econometrics [27, 130, 224, 240], speech processing [123], image processing [69, 70, 113, 230], finance [39, 210], performance analysis [99, 214] and many others [32, 64, 193].

4.1 Classical Moment Problem

In many problems in physics and finance it is easier to obtain the statistical data (in terms of moments) of the underlying probability distribution than the distribution itself. Therefore, it is the important task to understand how to reconstruct the unknown probability measure from the partial information contained in the finite number of moments. This problem was formulated by Stieltjes:

> Given the moments $\mu_k = \int_a^b x^k g(x)\,\mathrm{d}x$, $k = 0, 1, \ldots, M$, recover the function $g(x)$.

In the univariate case, the moment problem is subdivided into the following subproblems according to the values of a and b:

- $a = 0$, $b < \infty$ (usually, $b = 1$): Hausdorff moment problem [138] ([107])
- $a = 0$, $b = \infty$: Stieltjes moment problem [208]
- $a = -\infty$, $b = \infty$: Hamburger moment problem [103]

These are difficult inverse problems that usually lead to the (numerical) solution of ill-posed systems of equations [19]. In the setting of the stochastic chemical kinetics, we restrict ourselves to the Stieltjes or Hausdorff moment problems. More precisely, we consider the *truncated* versions of those since moment closure methods are capable to provide accurate approximations only to the first $M + 1$ moments of the distribution, $\mu = (\mu_0, \ldots, \mu_M)$ (within the reasonable computational time).

Here, we present the summary of the solution existence results for the truncated Stieltjes problem as given in [75]. The moment space $\mathcal{M}^{(M+1)}$ consists of all vectors $\mu \in \mathbb{R}_{\geq 0}^{(M+1)}$ such that there exists at least one probability distribution function $g(x)$ with moments μ. The moment space $\mathcal{M}^{(M+1)}$ is a convex subset (not necessarily closed or bounded [49]). The following cases are distinguished:

- the point $\mu \in \mathcal{M}^{(M+1)} \setminus \partial\mathcal{M}^{(M+1)}$ belongs to the interior of the moment space. The corresponding problem possesses infinitely many solutions and it is called *indeterminate*,
- the point $\mu \in \partial\mathcal{M}^{(M+1)}$ belongs to the boundary of the moment space. The corresponding problem possesses a unique solution and it is called *determinate*.

The Hankel determinants $|H_{2k}^{(0)}|$ and $|H_{2k+1}^{(1)}|$ are given by

$$|H_{2k}^{(0)}| = \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_k \\ \vdots & & & \vdots \\ \mu_{k-1} & \mu_k & \cdots & \mu_{2k-1} \\ \mu_k & \mu_{k+1} & \cdots & \mu_{2k} \end{vmatrix}, \qquad |H_{2k+1}^{(1)}| = \begin{vmatrix} \mu_1 & \mu_2 & \cdots & \mu_{k+1} \\ \vdots & & & \vdots \\ \mu_k & \mu_{k+1} & \cdots & \mu_{2k} \\ \mu_{k+1} & \mu_{k+2} & \cdots & \mu_{2k+1} \end{vmatrix}. \tag{4.1}$$

The necessary and sufficient conditions for solution existence (given that moment vector $\mu \in \mathbb{R}^{(M+1)}_{\geq 0}$ belongs to the Stieltjes moment space, $\mu \in \mathcal{M}^{(M+1)} \cup \partial\mathcal{M}^{(M+1)}$) to the Stieltjes moment problem are provided by the following inequalities [212]

$$\begin{aligned} |H^{(0)}_s| &> 0, \qquad 2 \leq 2s < M+1 \\ |H^{(1)}_s| &> 0, \qquad 3 \leq 2s+1 < M+1. \end{aligned}$$

These conditions are often formulated in terms of monotonically increasing sequences [48, 137, 209, 237].

The conditions for solution existence in case of Hausdorff are studied in details in the PhD thesis of Tari [214]. Here we provide only the theorem [107] for the Hausdorff moment problem: a necessary and sufficient condition that $\mu = (\mu_0, \ldots, \mu_M)$ are moments of a distribution supported over the interval $[0, 1]$ is that all the differences δ^k defined as

$$\delta^k \mu_l = \mu_l - \binom{k}{1}\mu_{l+1} + \binom{k}{2}\mu_{l+2} + \cdots + (-1)^k \mu_{l+k}$$

are non-negative, i.e., $\delta^k \mu_l \geq 0$, $k, l = 0, \ldots, M$.

Results on convergence properties of the constructed solution (using not only the maximum entropy approach) can be found in [27, 46, 165, 169, 171, 172, 213]. Here we use the results given in [162], where the authors consider the sequence of functions $g_N(x)$ that converge to $g(x)$ as N goes to infinity, i.e., $\lim\limits_{N\to\infty} g_N(x) = g(x)$. The point-wise convergence is very strong assumption, so the authors stick to the weaker convergence in averages, i.e., $\lim\limits_{N\to\infty} \int_a^b f_k(x) g_N(x)\,\mathrm{d}x = \int_a^b f_k(x) g(x)\,\mathrm{d}x = \mu_k$, where $f_k(x)$ is a basis function to generate moments, for example $f_k(x) = x^k$. If one applies the maximum entropy principle, the sequence of approximations $g_N(x)$ is the least biased one. It is important to note that most results are derived under the assumption that the number of moment constraints tends to infinity, $M \to \infty$. Here we always consider the finite number of moments since we propose the numerical procedure that can be applied on the real data, therefore only the limited amount of statistical information is available (the problem of selecting the certain finite subset of moments from the infinite sequence is solved for the case of fractional moments in [179]).

The extensions to multiple dimensions are non-trivial and still are not elaborated to the same level as for univariate case [124, 132, 170, 183].

4.2 Maximum Entropy Reconstruction

4.2.1 Maximum Entropy Principle

The principle of maximum entropy provides a solution to the classical moment problem of probability distribution reconstruction that satisfies the given moment constraints.

The principle can be stated as follows [119, 120]: consider the *discrete* random variable X that has a finite (or countable infinite) set of possible states $S = \{s_0, s_1, \ldots, s_n, \ldots\}$ with probabilities $p(x) = \mathbf{P}[X = x]$, $x \in S$. The only assumptions given are $M + 1$ constraints

$$\begin{aligned} \sum_{x \in S} p(x) = 1, \quad p_i \geq 0, \\ \sum_{x \in S} f_k(S) p(x) = \mu_k, \quad 1 \leq k \leq M < \infty, \end{aligned} \tag{4.2}$$

where μ_k are the moments defined for a set of suitable functions $f_k(S)$ that represent the known information (generally, they may depend not only on the current state x but also on the other states, therefore it is given as $f_k(S)$, not $f_k(x)$). The number of constraints is usually less than the number of possible states, therefore there are an infinite number of distributions $\mathcal{G}$ that satisfy these constraints. The maximum entropy principle states that the minimally prejudiced distribution is the one that maximizes the entropy

$$H(p) = -\sum_{x \in S} p(x) \cdot \ln [p(x)] \tag{4.3}$$

while satisfying the provided constraints. That is, the distribution of choice is

$$q = \arg\max_{p \in \mathcal{G}} H(p), \tag{4.4}$$

where $\mathcal{G}$ denotes the family of all possible distributions.

As originally stated by Jaynes [119], the maximum entropy distribution is *"uniquely determined as the one which is maximally noncommittal with regards to missing information, and that it agrees with what is known, but expresses maximum uncertainty with respect to all other matters."* He analyzed the distribution of gas molecules and showed that the maximization of the statistical mechanic entropy is equal to maximization of Shannon's information entropy.

Example 4.1 (MaxEnt Exponential Distribution)**.** ↩ Consider one-dimensional discrete distribution p over the set of states $S = \{0, 1, \ldots, n\}$. Assume that the only given constraint is the expectation μ_1. The constraints are then given by:

$$\sum_{i=0}^{n} p_i = \mu_0 = 1, \quad \sum_{i=0}^{n} i \cdot p_i = \mu_1, \quad p_i \geq 0.$$

The maximum entropy distribution is then given by (4.6) as $p_i = \exp(-\lambda_0 - \lambda_1 i) =$

$\alpha\beta^i$, $\alpha = \exp(-\lambda_0)$, $\beta = \exp(-\lambda_1)$. In this notation, the constraints can be expressed as:

$$\sum_{i=0}^{n} \alpha\beta^i = \mu_0 = 1, \quad \sum_{i=0}^{n} i\alpha\beta^i = \mu_1, \quad p_i \geq 0.$$

We can see that $\mu_1 = \mu_1 \cdot \mu_0$, therefore $\sum_{i=0}^{n} i\alpha\beta^i = \mu_1 \sum_{i=0}^{n} \alpha\beta^i$ and $\sum_{i=0}^{n}(i-\mu_1)\beta^i = 0$. The coefficients β^i are positive, $(0-\mu_1) < 0$ and there exist such k that $(k-\mu_1) > 0$, so there is only one change in sign. Therefore, there exist a unique real root to this set of equations according to Descartes' rule of signs, which corresponds to the exponential distribution.

In the rare case when the number of provided constraints is equal to the number of states $|S|$, the probability distribution function can be uniquely determined as shown in Example 4.2.

Example 4.2 (Moment Generating Function)**.** ↩ Consider one-dimensional discrete distribution p over the finite set of states $S = \{s_0, s_1, \ldots, s_n\}$ and the moment generating function g. Then the function g is uniquely determined by p and vice versa [95, Theorem 10.2].

In case of *continuous* random variables, the maximum entropy distribution maximizes the Shannon's entropy (4.3) which is defined relative to the uniform measure

$$H(p) = -\int p(x) \cdot \ln\left[p(x)\right] \, \mathrm{d}x,$$

and the constraints are given as

$$\int p(x)\,\mathrm{d}x = 1, \quad \int f_k(x)p(x)\,\mathrm{d}x = \mu_k, \quad 1 \leq k \leq M < \infty, \tag{4.5}$$

where $f_k(x)$ are the basis functions and the constraints are $E[f_k(X)] = \mu_k$. For both the discrete and continuous random variables, the general form of solution is given by (Appendix A.5.1)

$$p(x) = \exp\left(-\sum_{k=0}^{M} \lambda_k f_k(x)\right), \tag{4.6}$$

where λ_k is the Lagrangian multiplier for the k-th moment constraint. Here we use the *monomial basis* functions $f_k(x) = x^k$, so that the solution (4.6) is $p(x) = \exp\left(-\sum_{k=0}^{M} \lambda_k x^k\right)$, whereas extensions to other basis functions are possible. Theorem 11.1.1 in [53, p. 267] states that the density $p(x)$ is a unique maximizer of the entropy $H(p)$.

Generally, there is no analytical solution given by (4.6) for $M > 2$, therefore we use numerical approaches to solve the nonlinear constrained optimization problem (4.4).

The existence conditions for various modifications of the maximum entropy problem are given in [124]. The extension to the case of inequality constraints is given in [181]. The multidimensional maximum entropy theory is addressed in [9, 65, 242]. We also note that the entropy value $H(p)$ can be estimated indirectly without reconstructing the underlying distribution p ([101] and references therein).

There is a possibility to choose the entropy functional which is different from Boltzmann-Shannon entropy $H(p)$ considered here (for example, Burg entropy). The generalized results on the existence of solution and comparison of the numerical results can be found in [37].

4.2.2 Dual Approach in Maximum Entropy Problem

The objective function (4.3) is strictly concave with the constraints (4.5) therefore we can transform the constrained primal optimization problem into an unconstrained one using the *dual approach*. The authors of [4] were the first to apply the duality approach to entropy maximization. Generally, one uses Kuhn-Tucker theorem [142] to prove the duality relation. For each constraint (given by μ_k) we introduce the Lagrange multiplier λ_k and the Lagrangian functional is defined as follows

$$\mathcal{L}(p,\lambda) = H(p) - \sum_{k=0}^{M} \lambda_k \left(\textstyle\sum_x x^k p(x) - \mu_k\right).$$

It is possible to show (cf. Appendix A.5.1) that maximizing the Lagrangian $\mathcal{L}$ gives a solution to the constrained maximum entropy problem. The variation of the functional $\mathcal{L}$ according to the unknown distribution $q(x)$ provides the general form of $q(x)$

$$\frac{\partial \mathcal{L}}{\partial q(x)} = 0 \implies q(x) = \exp\left(-1 - \sum_{k=0}^{M} \lambda_k x^k\right) = \frac{1}{Z}\exp\left(-\sum_{k=1}^{M} \lambda_k x^k\right),$$

where

$$Z = e^{1+\lambda_0} = \sum_x \exp\left(-\sum_{k=1}^{M} \lambda_k x^k\right) \tag{4.7}$$

is a normalization constant. The dual function is defined as $\Psi(\lambda) = \mathcal{L}(q,\lambda)$ and

$$\Psi(\lambda) = \ln Z + \sum_{k=1}^{M} \lambda_k \mu_k. \tag{4.8}$$

Therefore, the *unconstrained* optimization problem that has to be solved is stated as follows (cf. Appendix A.5.2)

$$\lambda^* = \arg\min_{\lambda \in \mathbb{R}^M} \Psi(\lambda). \tag{4.9}$$

The corresponding maximum entropy distribution is given by $q(x,\lambda^*) = \exp\big(-\sum_{k=0}^{M} \lambda_k^* f_k(x)\big)$.

The dimensionality of the primal problem (4.4) is $|S|$ (whenever there is a finite number of states) and the dimensionality of the dual problem is M. The authors of [217] were the first to formally prove that the dual function $\Psi(\lambda)$ (before, it was usually addressed as *potential function* [4]) is the formal convex dual of the problem (4.9) using both the arguments of Lagrange saddle-point conditions and geometric programming. They also raise the discussion about the nature of the dual function: because of the different dimensionality, the values of the dual function do not describe the entropy but measure the bias (prejudice) contained in the information. Theoretical results on the duality relationship in maximum entropy problems can be found in [20, 36, 57, 106].

If the solution λ^* if found using the first M as constraints, the moments of the order $l > M$ can be found as functions of first M moments and elements of λ^* [190, 250]

$$(l+1)\mu_l - 1 + \sum_{k=1}^{M} k\lambda_k \left(\mu_k - \mu_{k+l}\right) = 0,$$

so that the maximum entropy can be used as a moment closure technique. This is used in [203] where the values of the higher order moments are approximated via the summation over the state space using the maximum entropy probability mass function, i.e., $\mu_l = \sum_x x^l q(x)$, $l > M$.

4.3 Numerical Approach to solve Maximum Entropy Problem

Normally, for the number of constraints $M \geq 3$ it is impossible to analytically find the maximum entropy distribution. Therefore, most of the literature on the application of maximum entropy formalism considers the numerical approximation methods. Among the most known are work of Agmon et al. [4], Bandyopadhyay [25], Mead et al. [162], Ormoneit et al. [177], Wu et al. [240, 241]. For our implementation we adopt the approach of Abramov [1–3] which combines several techniques to overcome the numerical instability of the dual minimization problem. This approach is the most up-to-date technique that is adopted for various applications [8, 92, 116, 213].

4.3.1 Minimization of the Dual Function

We consider the minimization of the dual function $\Psi(\lambda)$ (4.8). Assume that no change of basis is performed and the given data are algebraic moment constraints μ_k computed with respect to monomials $f_k(x) = x^k$. Given an approximation $\lambda^{(\ell)} = (\lambda_1^{(\ell)}, \ldots, \lambda_M^{(\ell)})$ of the vector $\lambda = (\lambda_1, \ldots, \lambda_M)$ in the ℓ-th iteration (we omit λ_0 according to Eq. (4.8), where it is a part of the normalization constant Z), the elements of the gradient vector are computed as $\partial\Psi/\partial\lambda_i \approx \mu_i - \frac{1}{Z}\widetilde{\mu}_i$ (cf. Appendix A.5.2). The entries of the Hessian matrix H are approximated by $H_{ij} \approx 1/Z^2\,(\tilde{\mu}_{i+j}Z - \tilde{\mu}_i\tilde{\mu}_j)$. Note that the normalization constant Z and the approximated moments $\widetilde{\mu}_i$ are updated in each step of the iteration,

i.e., they depend on ℓ but we omit the superscript ℓ here. We use the `minFunc` method from the numerical minimization package `minFunc2012` [192] with the default settings where we chose $\lambda^{(0)} = (0, \ldots, 0)$ as an initial starting point (other initialization strategies are considered in Appendix A.5.5). In case of standard Newton algorithm, the iterations are as follows

$$\lambda^{(\ell+1)} = \lambda^{(\ell)} - \zeta^{(\ell)} \cdot \left(H^{-1} \, \partial\Psi / \partial\lambda \right) |_{\lambda=\lambda^{(\ell)}},$$

where $\zeta^{(\ell)}$ is a stepping distance parameter. Since for the systems that we consider, the dual function is convex, there exists a unique minimum $\lambda^* = (\lambda_1^*, \ldots, \lambda_M^*)$ where all first derivatives are zero and where the Hessian is positive definite.

The dimensionality of the optimization problem is M due to the fact that λ_0^* can be calculated as $\lambda_0^* = \ln[Z] - 1$. To find the minimum of the dual function, we use the Newton-based methods which might fail due to an ill-conditioned Hessian matrix [4, 8]. To speed up the convergence, Abramov [2] uses the Sherman-Morrison formula [196] to obtain the inverse of pseudo-Hessian. The authors of [243] use the fast Fourier transform (FFT) to evaluate the Sherman-Morrison formula so that the computational cost of one iteration is reduced from $\mathcal{O}(n^3)$ (in traditional Newton method) to $\mathcal{O}(n \log n)$ floating point operations. The authors of [8] apply two different stopping criterion for the Newton iteration to better render the nature of the underlying problem. To achieve better numerical stability, other functions f_k of random variables need to be considered (other than monomials $f_k(x) = x^k$ that generate algebraic moments) such as Chebyshev polynomials [25], Fup basis functions [91], monomials of non-integer powers [101], or Lagrange interpolation polynomials [226]. The transformation from algebraic moments to Chebyshev moments is given in [60] where shifted Chebyshev polynomials of the first kind are used as basis functions. The authors of [34] use up to 100 moments in their case studies (for example, for a uniform distribution $p_6(x)$). The magnitude order difference between μ_0 and μ_{100} in this case is only 10^4, much smaller than in case of standard monomial basis functions.

Note that the above approach provides a reasonable approximation of the individual probabilities only in the region where the main part of the probability mass is located. In order to accurately approximate the tails of the distribution, special methods have been developed [75, 152].

4.3.2 Preconditioning and Minimization of Dual Function

We demonstrate the main steps of the Abramov's algorithm by considering one-dimensional maximum entropy reconstruction problem for continuous distributions. The adopted outline of the algorithm [3] is given below, where we assume that original constraints (4.2) are such that $f_k(x) = x^k$ since we consider the monomial basis functions in the course of this thesis.

1. Approximate the support of the probability distribution $D = \text{supp}(p)$ such that $|D| < \infty$;
2. Convert input constraints μ_k into $\tilde{\mu}_k$ such that $\tilde{\mu}_1 = 0$, $\tilde{\mu}_2 = 1$ (zero mean and unity variance). Transform the support D of the distribution correspondingly;

3. Choose the new basis using the set of orthogonal independent polynomials $J_k(x)$ of order M;
4. Initialize the Lagrange multipliers γ_k in the new basis $v_k(x)$;
5. Reorthogonalize the monomial basis with respect to new basis $v_k(x)$;
6. Perform the numerical minimization $\gamma^* = \arg\min \Psi_v(\gamma)$ of the dual function;
7. Convert the solution γ^* to λ^* in the original basis.

The details for the preconditioning step 2 are given in Appendix A.5.3 (p. 101) and for the basis orthogonalization - in Appendix A.5.3 (p. 102).

In comparison to the original algorithm [3], we do not proceed with re-orthogonalization on each iteration of the optimization procedure. The approach of Alldredge et al. [8] is more advanced since re-orthogonalization is done as soon as the condition number of the Hessian becomes high, i.e., when the problem becomes ill-conditioned (the theoretical results on the condition number of the Hessian matrix can be found in [165]) which provides the better convergence properties. Here, we perform the orthogonalization only once in the beginning, since the further re-orthogonalizations does not give the better accuracy nor the smaller running time.

The first attempts to implement the preconditioning for the maximum entropy problem with 4 parameters were done in [28] where special attention is devoted to the problem of equivalent moment transformations. It is pointed out that the strictest possible interpretation of the maximum entropy formalism does not permit for any moment transformations. However, since such transformations introduce negligible errors, one may use them in implementation purposes. We are using approximations of the moments in the first place, so the errors introduced by the preconditioning are negligible in comparison to errors from moment closure techniques.

For the sake of computational speed we can stick to the algorithm that does not involve any preconditioning, but still able to provide the reasonable approximations. The comparison of approximation accuracy is given in Section 4.3.5.

The stochastic process $\vec{X}(t)$ has discrete state space, therefore we modify the above-stated algorithm such that it can be applied to the moments obtained with standard moment closure technique or method of conditional moments and allows us to reconstruct the underlying discrete distribution.

The core of the algorithm is minimization of the dual function (4.8). The expressions for gradient vector and Hessian matrix of the dual function can be obtained analytically (cf. Appendix A.5.2), therefore we can use any Newton-based approach that makes use of the first and second derivatives of the goal function such as the standard Newton method [4], Broyden-Fletcher-Goldfarb-Shanno (BFGS) procedure [3, 42] and damped Newton method [8]. In our implementation in `MATLAB` [160] we use the numerical optimization package `minFunc` [192] that chooses between several techniques depending on the properties of the provided Hessian matrix.

4.3.3 Distribution Support Approximation

In order to approximate the moments in (4.5), we need to sum (integrate) over all possible states in the state space S which might be infinite. Instead, during the iterative procedure, we consider a subset $D = \{x_L, \ldots, x_R\} \subset \mathbb{N}_0$ ($D = (x_L, x_R) \subset \mathbb{N}_0$) that contains the main part of the probability mass [215] (here we consider only one-dimensional state spaces S). We have to find the appropriate values for x_L and x_R, since the iteration of type (4.3.1) might fail to converge if the chosen value of x_R is very large (and if $x_L = 0$) as the conditional number of the Hessian (or modified Hessian) matrix is very large in this case. One of the simple ways to regularize the Hessian matrix is to use Levenberg-Marquardt formula [149, 156, 184], where the modified Hessian $\hat{H}$ is given by $\hat{H} = (H + \gamma \cdot \mathrm{diag}(H))$ and γ is the damping parameter that has to be properly chosen [225], however more advanced methods are implemented in `minFunc2012` [192] package. Thus, we make use of the results in [215] to find a region of the state space that contains the main part of the probability mass. We consider the roots of the function

$$\Delta^0(w) = \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_k \\ \vdots & & & \vdots \\ \mu_{k-1} & \mu_k & \cdots & \mu_{2k-1} \\ 1 & w & \cdots & w^k \end{vmatrix}, \tag{4.10}$$

where $k = \lfloor \frac{M}{2} \rfloor$ and M is even. Let $W = \{w_1, \ldots, w_k\}$ be the set of solutions for $\Delta^0(w) = 0$, where $w_1 < \ldots < w_k$ are real and simple roots [74] The set $D^{(0)} = \{x_L^{(0)}, \ldots, x_R^{(0)}\}$ with $x_L^{(0)} = \lfloor w_1 \rfloor$ and $x_R^{(0)} = \lceil w_k \rceil$ is used as an initial guess for the approximated support when we start the optimization procedure.

We can also account for the case of an odd number of moment constrains. In addition to the function $\Delta^0(w)$ defined in Eq. (4.10), we also consider the function $\Delta^1(\eta)$

$$\Delta^1(\eta) = \begin{vmatrix} \mu_1 - w_1\mu_0 & \mu_2 - w_1\mu_1 & \cdots & \mu_z - w_1\mu_{z-1} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{z-1} - w_1\mu_{z-2} & \mu_z - w_1\mu_{z-1} & \cdots & \mu_{2z-2} - w_1\mu_{2z-3} \\ 1 & \eta & \cdots & \eta^{z-1} \end{vmatrix},$$

where $z = \lfloor \frac{M}{2} \rfloor + 1$ and w_1 is the smallest root of the equation $\Delta^0(w) = 0$. Again, let $W = \{w_1, \ldots, w_k\}$ be the set of the solutions of $\Delta^0(w) = 0$ and $H = \{\eta_1, \ldots, \eta_z\}$ be the set of solutions of $\Delta^1(\eta) = 0$, where all the elements of W and H are real and simple. The first approximation for the truncated support of the distribution is then given by the set $D^{(0)}\left(x_L^{(0)}, x_R^{(0)}\right) = \{x_L^{(0)}, \ldots, x_R^{(0)}\}$ with $x_L^{(0)} = \lfloor \min(w_1, \eta_1) \rfloor$ and $x_R^{(0)} = \lceil \max(w_k, \eta_z) \rceil$. We extend the support until the relative change of the dual function becomes smaller than the threshold δ_Ψ

$$\left| \frac{\Psi(\lambda^{(\ell-1)}) - \Psi(\lambda^{(\ell)})}{\Psi(\lambda^{(\ell)})} \right| < \delta_\Psi. \tag{4.11}$$

Figure 4.1: ↩ Extension of one-dimensional support approximation done by adding two states (green) on the left and on the right.

If the inequality is not satisfied, we extend the support by adding new states in each iteration

$$D^{(\ell+1)}\left(x_L^{(\ell+1)}, x_R^{(\ell+1)}\right) = \{\max(0, x_L^{(\ell)} - 1), \ldots, x_R^{(\ell)} + 1\}. \tag{4.12}$$

Various strategies of support extension can be applied: more states can be added to the set $D^{(\ell+1)}$ on each step and asymmetric treatment of x_L and x_R is possible. This is shown in Figure 4.1, where the support $D^{(\ell)}$ is shown with blue circles and the new states are shown by green circles.

The current approach addresses the solution of the two paired problems, namely the choice of the finite approximation D for the possibly infinite support $\mathrm{supp}(p)$ of the underlying distribution and the numerical minimization of the dual function Ψ given D

$$\lambda^* = \arg \min_{\substack{\lambda \in \mathbb{R}^M \\ x \in D^*}} \Psi(\lambda, D^*) = \arg \min_{\lambda \in \mathbb{R}^M} \left(\ln \left[\sum_{x \in D^*} \exp\left(-\sum_{k=1}^{M} \lambda_k x^k \right) \right] + \sum_{k=1}^{M} \lambda_k \mu_k \right),$$

where D^* is an approximation of the support $\mathrm{supp}(q(x, \lambda^*))$ such that (4.11) holds,

$$D^* = \min_{\ell} \{ D^{(\ell)} \Big| \left| \Psi(\lambda^{(\ell-1)}, D^{(\ell-1)}) - \Psi(\lambda^{(\ell)}, D^{(\ell)}) / \Psi(\lambda^{(\ell)}, D^{(\ell)}) \right| < \delta_\Psi \}.$$

The final results λ^* and D^* of the iteration yields the distribution $\tilde{q}(x)$ that approximates the marginal distribution of interest

$$\tilde{q}(x) = \begin{cases} \exp(-1 - \sum_{k=0}^{M} \lambda_k^* x^k), & x \in D^* \\ 0, & x \notin D^* \end{cases}$$

We consider the result of the application of the algorithm to the mixture of two normal distributions $p_8(x)$ in the Example 4.3.

Example 4.3. ↩ We consider the approximation of the support using the iterative procedure for mixture of two normal distributions p_8 (cf. Section 4.3.5)

$$p_8(x) = w_1 \cdot f_{\mathcal{N}(s_1,\sigma_1)}(x) + w_2 \cdot f_{\mathcal{N}(s_2,\sigma_2)}(x),$$

where s_1 = -3, $s_2 = 10$, $\sigma_1 = 1$, $\sigma_2 = 2$, $w_1 = 0.3$, $w_2 = 0.7$ and $x \in \mathbb{R}$. This is the example of numerical solution to the Hamburger moment problem. Assume that the highest order of moment constraints is $M = 4$, i.e., the constraints for the maximum entropy problem are given by the vector $\mu = (\mu_0, \ldots, \mu_4)$. The approximation of the support is based on both $\Delta^0(w)$ and $\Delta^1(\eta)$. We show the obtained reconstructions for the initial

approximation of support ($D^{(0)} = [-5.9, 10.9]$) the reconstruction after two iterations ($D^{(2)} = [-6.7, 11.8]$) and the final reconstruction ($D^* = D^{(7)} = [-11.6, 16.7]$) obtained with the convergence threshold for the dual function $\delta_\Psi = 10^{-4}$ in Figure 4.2.

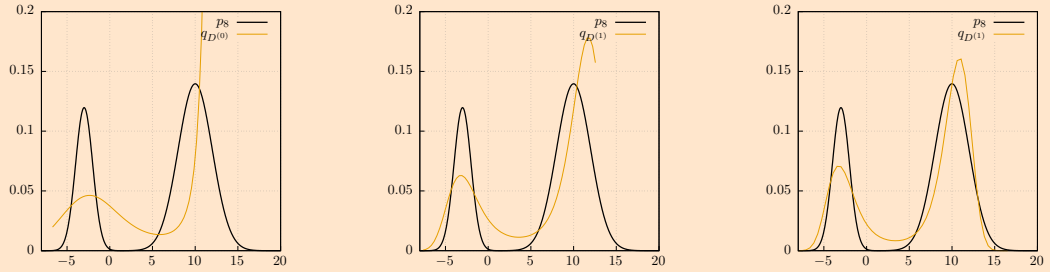


Figure 4.2: ↩ Reconstruction of distribution p_8 with different support approximations. Black curve corresponds to distribution p_8 and yellow curves correspond to reconstructions.

4.3.4 Numerical Approach for the Two-dimensional Maximum Entropy Problem

In the previous sections we only addressed the one-dimensional maximum entropy problem. Here we consider the modification of the maximum entropy problem in case of two-dimensional distributions. The constraints are given by the sequence of non-central moments $E\left(X_x^r X_y^l\right) = \mu_{r,l}$, $0 \leq r + l \leq M$, and the set $\mathcal{G}_2$ of all two-dimensional discrete distributions that satisfy the following constraints

$$\sum_{x,y \in S} x^r y^l g(x,y) = \mu_{r,l}, \quad 0 \leq r + l \leq M.$$

Here $X_\cdot$ and $X_\circ$ correspond to the populations of two different species, i.e., to the two distinct elements of the random vector $\vec{X}(t) = (X_1(t), \ldots, X_n(t))$ at some fixed time instant t. The optimization problem (4.4) is formulated as

$$q = \arg\max_{p \in \mathcal{G}^2} H(p) = \arg\max_{p \in \mathcal{G}_2} \left(-\sum_{x,y} p(x,y) \ln p(x,y) \right),$$

where we seek the distribution $p \in \mathcal{G}_2$ that maximizes the entropy $H(p)$. The general form of the solution for the maximum entropy problem is given by

$$q(x,y) = \exp(-1 - \sum_{0 \leq r+l \leq M} \lambda_{r,l} x^r y^l) = \frac{1}{Z} \exp(- \sum_{1 \leq r+l \leq M} \lambda_{r,l} x^r y^l), \tag{4.13}$$

where the normalization constant Z is calculated as

$$Z = e^{1+\lambda_{0,0}} = \textstyle\sum_{x,y} \exp(- \sum_{1 \leq r+l \leq M} \lambda_{r,l} x^r y^l). \tag{4.14}$$

We solve the optimization problem numerically, similar to the one-dimensional case. The vector $\lambda^{(\ell)} = (\lambda_{0,1}, \lambda_{1,0}, \ldots, \lambda_{0,M}, \lambda_{M,0})$ is an approximation of the vector λ in Eq. (4.13). The elements of the gradient vector are computed as

$$\frac{\partial \Psi}{\partial \lambda_{r,l}} \approx \mu_{r,l} - \frac{\widetilde{\mu}_{r,l}}{Z},$$

where $\widetilde{\mu}_{r,l}$ is approximated by

$$\widetilde{\mu}_{r,l} = \sum_{x,y} x^r y^l \exp(- \sum_{1 \le r+l \le M} \lambda_{r,l} x^r y^l). \tag{4.15}$$

Here $r, l \in \{0, \ldots, 2M\}$ and the sum is taken over all $(x, y) \in \mathbb{N}_0^2$. Finally, the elements of the Hessian matrix are computed as

$$H_{r+u,l+v} = \frac{\partial^2 \Psi}{\partial \lambda_{r,l} \partial \lambda_{u,v}} \approx \frac{Z \cdot \widetilde{\mu}_{r+u,l+v} - \widetilde{\mu}_{r,l} \widetilde{\mu}_{u,v}}{Z^2},$$

where $0 \le r + l \le M, 0 \le u + v \le M$. Using the numerical minimization procedure, the vector $\lambda^* = \left(\lambda^*_{0,1}, \lambda^*_{1,0}, \ldots, \lambda^*_{0,M}, \lambda^*_{M,0}\right)$ is found. The dimensionality of the optimization problem is $0.5\,(M^2 + 3M)$, and $\lambda^*_{0,0}$ can be calculated from (4.14) as $\lambda^*_{0,0} = \ln Z - 1$. In comparison to the one-dimensional case, the range of the values of $\widetilde{\mu}_{r,l}$ becomes wider due to the larger dimensionality, so that the conditional number of the Hessian matrix is even higher and the iteration might fail.

To approximate the moment values in (4.15) we truncate the infinite support similarly to the one-dimensional case. As an initial approximation for the support D_{xy} we use the results of the corresponding approximations $D^*_{x,\alpha}$ and $D^*_{y,\alpha}$ obtained for marginal one-dimensional distributions, $D^{(0)}_{xy} = D^*_{x,\alpha} \times D^*_{y,\alpha}$. Here α corresponds to the amount of probability mass which is set to 0.95 for all considered case studies. The region that contains at least α of probability mass is computed using Algorithm 7. The condition (4.11) is checked and if not satisfied, the support is extended according to the selected strategy. The trivial strategy is to add two discrete states in each dimension at the iteration $\ell + 1$ as follows

$$\begin{aligned} D^{(\ell+1)}_{xy} &\left(x_L^{(\ell+1)}, x_R^{(\ell+1)}; y_L^{(\ell+1)}, y_R^{(\ell+1)}\right) \\ &= \{\max(0, x_L^{(\ell)}\text{-}1), \ldots, x_R^{(\ell)}\text{+}1\} \times \{\max(0, y_L^{(\ell)}\text{-}1), \ldots, y_R^{(\ell)}\text{+}1\} \end{aligned}$$

The example of such extension is shown in Figure 4.3 where green circles correspond to the newly added states.

As before, we seek for the solution of the two paired problems: the approximation of the support and the dual function minimization. The final results λ^* and D^*_{xy} of the iteration yields the distribution $\tilde{q}(x, y)$ that approximates two-dimensional distribution of interest

$$\tilde{q}(x, y) = \begin{cases} \exp(-1 - \sum_{0 \le r+l \le M} \lambda^*_{r,l} x^r y^l), & (x, y) \in D^*_{xy} \\ 0, & (x, y) \notin D^* \end{cases}$$

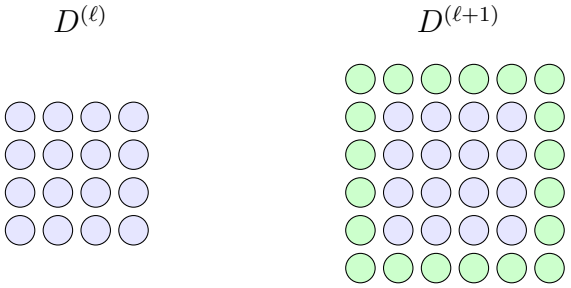


Figure 4.3: ↩ Extension of two-dimensional support approximation, where newly added states are shown with green circles.

The algorithm of the iteration procedure is given in Algorithm 2 (Algorithm 3 in case of two-dimensional distribution), where the subset of moments used for the reconstruction is truncated in case if the current trial was not successful (because the numerical minimization fails or no distribution is found that suffice the validity criteria according to Appendix A.5.3). The adaptive support optimization procedure is given in Algorithm 4, where it is assumed that the preconditioning is applied. If it does not, the corresponding step is omitted. The influence of the basis orthogonalization is visible only when we define the goal function to be minimized: if orthogonalization is applied, the corresponding dual function is given by (A.8). Therefore, it is possible define the dual function to be minimized in several ways:

- without any precomputation or preconditioning
- with precomputation (according to Appendix A.5.3)
- with preconditioning (centralization, scaling and rotation) and partial precomputation
- with preconditioning (centralization, scaling and rotation, basis orthogonalization)

4.3.5 Maximum Entropy for Continuous Distributions, Numerical Results

Here, we analyze the numerical properties of maximum entropy procedure in case of *one-dimensional continuous distributions*. Consider the numerical minimization of the dual function $\Psi(\lambda)$ that corresponds to the primal problem (4.4) and constraints (4.5). The initial approximation for the parameter vector λ is chosen as $\lambda = (\lambda_1, \ldots, \lambda_M) = (0, \ldots, 0)$ and λ_0 is calculated after the dual function is minimized. Similar to the discrete case in (4.7), we obtain

$$\lambda_0 = \ln Z - 1, \quad Z = \int\limits_{x \in D^*} \exp\left(-\sum_{k=1}^{M} \lambda_k x^k\right) \mathrm{d}x.$$

We consider the following case studies, where the first three distributions are representatives of the generalized exponential family:

1. Exponential distribution $p_1(x) = 5\exp(-5x) = \exp\left(-(-\ln 5 + 5x)\right)$, $x \in \mathbb{R}_{\geq 0}$
2. Normal distribution $p_2(x) = f_{\mathcal{N}(-1,\sqrt{2})}(x) = {}^1\!/\!_{\sqrt{2\pi}} \exp(-{}^1\!/\!_2(x+1)^2) = \exp(\ln {}^1\!/\!_{\sqrt{2\pi}} - {}^1\!/\!_2 - x - {}^1\!/\!_2 x^2)$, $x \in \mathbb{R}$
3. Quartic exponential distribution $p_3(x) = \exp(-2.82 + 1.5x + 1.5x^2 - 0.1x^3 - 0.5x^4)$, $x \in \mathbb{R}$
4. Triangular distribution $p_4(x) = 1 - |x|$, $|x| < 1$
5. Shifted symmetric exponential type distribution $p_5(x) = {}^\beta\!/\!_{2\Gamma(1/\beta)} \exp(-|x-s|^\beta)$, $\beta = 1.3$, $s = 7$, $x \in \mathbb{R}$
6. Uniform distribution $p_6(x) = {}^1\!/\!_{b-a}$, $b = 30$, $a = 0$, $x \in [a, b]$
7. Generalized Pareto distribution, heavy tailed $p_7(x) = {}^1\!/\!_\sigma \left(1 + {}^{kx}\!/\!_\sigma\right)^{-1-1/k}$, $\sigma = 20$, $k = 0.1$, $x \in \mathbb{R}_{\geq 0}$
8. Mixture of two normal distributions
   $p_8(x) = w_1 \cdot f_{\mathcal{N}(s_1,\sigma_1)}(x) + w_2 \cdot f_{\mathcal{N}(s_2,\sigma_2)}(x) = w_1 \cdot {}^1\!/\!_{\sqrt{2\pi\sigma_1}} \exp(-{}^1\!/\!_{2\sigma_1^2}(x-s_1)^2) + w_2 \cdot {}^1\!/\!_{\sqrt{2\pi\sigma_2}} \exp(-{}^1\!/\!_{2\sigma_2^2}(x-s_2)^2)$, s_1 = -3, $s_2 = 10$, $\sigma_1 = 1$, $\sigma_2 = 2$, $w_1 = 0.3$, $w_2 = 0.7$, $x \in \mathbb{R}$
9. Mixture of three normal distributions
   $p_9(x) = w_1 f_{\mathcal{N}(s_1,\sigma_1)}(x) + w_2 f_{\mathcal{N}(s_2,\sigma_2)}(x) + w_3 f_{\mathcal{N}(s_3,\sigma_3)}(x)$, $s_1 = -3$, $s_2 = 10$, $s_3 = 20$, $\sigma_1 = 1$, $\sigma_2 = 2$,$\sigma_3 = 3$, $w_1 = 0.3$, $w_2 = 0.4$, $w_3 = 0.3$ $x \in \mathbb{R}$

Reconstruction of p_1. To obtain the reconstruction of p_1, the first 2 moments are used, namely $\mu_0 = 1$ and $\mu_1 = 0.2$. The dual function that is minimized is given by

$$\Psi(\lambda) = \ln Z + \lambda_1\mu_1 = \ln(e^{1+\lambda_0}) + \lambda_1\mu_1 = \ln\left(\sum_{x \in D^*} \exp(-\lambda_1 x)\right) + \lambda_1\mu_1.$$

	λ_0	λ_1	λ_0	λ_1
δ_Ψ	10^{-4}		10^{-6}	
p_1	-1.609438	5.0	-1.609438	5.0
$\tilde{q}_1$	-1.608986	4.997488	-1.609434	4.999981
$\epsilon_\lambda^{\text{rel}}$	$2.8 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$	$2.2 \cdot 10^{-6}$	$3.8 \cdot 10^{-6}$
x_R^*	1.98		3.07	
ℓ^*	48		57	
$\epsilon_{\mu_1}^{\text{rel}}$	$6.0 \cdot 10^{-6}$		$5.4 \cdot 10^{-7}$	

Table 4.1: ↩ Reconstruction of the exponential distribution p_1 with $\delta_\Psi = 10^{-4}$ and $\delta_\Psi = 10^{-6}$.

The results are given in Table 4.1, where $\epsilon_{\lambda_i}^{\text{rel}} = |\lambda_i^* - \lambda_i / \lambda_i|$ is the relative error of i-th element of the coefficients vector, $\epsilon_{\mu_k}^{\text{rel}} = |\mu_k^* - \mu_k / \mu_k|$ is the relative error of k-th element of the moments vector and ℓ^* is the number of iterations for the support search procedure. The moments μ^* are computed as $\mu_k^* = \int_{x \in D^*} x^k \tilde{q}(x, \lambda^*)\, \mathrm{d}x$. The preconditioning techniques are not used in these experiments. For the support extension the following strategy is used: at each iteration we extend the length of the support by 5% but not more than by 1.0

$$D^{(\ell+1)} = \left(x_L^{(\ell)} - \delta, x_R^{(\ell)} + \delta \right), \qquad \delta = \min(0.05 \cdot |D^{(\ell)}|, 1).$$

The initial approximation of the support is computed using only $\Delta^0(w)$

$$\Delta^0(w) = \begin{vmatrix} 1 & 0.2 \\ 1 & w \end{vmatrix} = w - 0.2,$$

and $W = \{0.2\}$, $x \in \mathbb{R}_{\geq 0}$, thus we initialize the support search with $D^{(0)} = [0, 0.2]$.

We observe that decreasing the threshold δ_Ψ results in the significant accuracy gain in determining λ (from the order of 10^{-4} to 10^{-6}) but requires a bit more support extension iterations. The approximation of the expectation μ_1 also becomes more accurate.

Reconstruction of p_2. Results for the normal distribution p_2 are given in Table 4.2, where we used first 3 moments for the reconstruction, namely $\mu = (1, -1, 2)$. In the table, D^* denotes the final approximation of the support $\text{supp}(p_2)$ and $\max\limits_{i \in \{0,1,2\}} \epsilon_{\mu_i}^{\text{rel}}$ denotes the maximum relative distance between moment values. The initial approximation of the support is given by

$$\Delta^0(w) = \begin{vmatrix} 1 & -1 \\ 1 & w \end{vmatrix} = w + 1, \quad \Delta^1(\eta) = \begin{vmatrix} 0 & 1 \\ 1 & \eta \end{vmatrix} = -1,$$

therefore we modify $\Delta^1(\eta)$ assuming that $w_1 = 0$

$$\Delta^1(\eta) = \begin{vmatrix} -1 & 2 \\ 1 & \eta \end{vmatrix} = -\eta - 2,$$

to obtain at least some reasonable approximation of the support. Thus, we have $D^{(0)} = [-2, -1]$. In both cases it would be more preferable to use other support approximation methods. In this case, the standard heuristics $[\mu_1 - 3\sqrt{\mu_2}, \mu_1 + 3\sqrt{\mu_2}]$ gives better results. For the normal distribution p_2 only 2 iterations are needed if the interval $D^{(0)} = [x_L^{(0)}, x_R^{(0)}] \approx [-5.24, 3.24]$ is used as an initial approximation for $\mathrm{supp}(p_2)$ with $\delta_\Psi = 10^{-4}$ (4 iterations with $\delta_\Psi = 10^{-6}$).

The measure of the overall agreement between the reference normal distribution p_2 and the reconstruction is also listed in Table 4.2 and is denoted by $||\epsilon||_V$. This is *total variation distance* [150] (adopted by Philipp Thomas in [16, Section 5.3]). Here we use it for continuous distributions as

$$||\epsilon||_V = \frac{100\%}{2} \int_{x_L}^{x_R} |p(x) - q(x)| \, \mathrm{d}x. \tag{4.16}$$

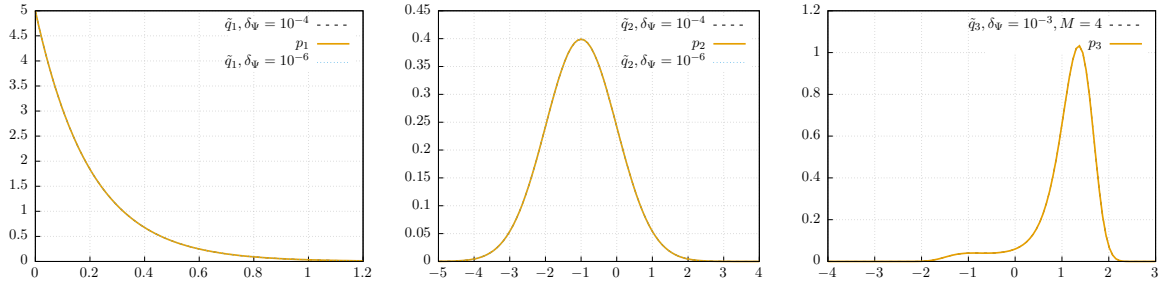
It can be interpreted as "the maximum percentage difference between the probabilities of all possible events assigned by the two distributions which achieves its maximum (100% error) when the distributions do not overlap". We consider it as a percentage representation of the standard L_1 distance between any two distributions [53, p. 369, (11.132)]

We observe that the distance $||\epsilon||_V$ decreases by two orders of magnitude when δ_Ψ is decreased, however such result is obtainable only in very limited number of cases, such as this well-behaving synthetic case study. Both distributions p_1 and p_2 together with reconstructions $\tilde{q}_1$, $\tilde{q}_2$ are shown in Figure 4.4 (4.4a, 4.4b), however the curves are visually indistinguishable from each other since the approximation of λ vector is accurate in both cases. In all figures in this section the ground truth distribution is shown by yellow solid curve and the reconstructions are given by black and blue dashed lines.

Reconstruction of p_3. In case of quartic exponential distribution we investigate how number of moment constraints influences the accuracy of the reconstruction (results are given in Table 4.3). When only first 4 moments are provided, i.e., only the moments

	λ_0	λ_1	λ_2	λ_0	λ_1	λ_2
δ_Ψ		10^{-4}			10^{-6}	
p_2	1.418939	1.0	0.5	1.418939	1.0	0.5
$\tilde{q}_2$	1.418913	0.999866	0.499943	1.418913	0.999866	0.499943
$\epsilon_\lambda^{\mathrm{rel}}$	$1.8 \cdot 10^{-5}$	$1.3 \cdot 10^{-4}$	$1.1 \cdot 10^{-4}$	$9.6 \cdot 10^{-8}$	$1.1 \cdot 10^{-6}$	$9.9 \cdot 10^{-7}$
D^*		[-6.42, 3.42]			[-7.46, 4.46]	
ℓ^*		25			27	
$\max_i \epsilon_{\mu_i}^{\mathrm{rel}}$		$9.0 \cdot 10^{-6}$			$1.5 \cdot 10^{-7}$	
$||\epsilon||_V$, %		$2.8 \cdot 10^{-3}$			$2.4 \cdot 10^{-5}$	

Table 4.2: ↩ Reconstruction of normal distribution with $\delta_\Psi = 10^{-4}$ and $\delta_\Psi = 10^{-6}$.



(a) (b) (c)

Figure 4.4: ↩ Reconstruction of exponential p_1 (a), normal p_2 (b) and quartic exponential p_3 (c) distributions.

$\mu_0, \mu_1, \mu_2, \mu_3$, the reconstruction with the threshold $\delta_\Psi = 10^{-4}$ is not even possible. The process converges if the threshold is set up to $\delta_\Psi = 10^{-3}$ but the corresponding distance from the true distribution is $||\epsilon||_V = 24\%$.

The best reconstruction (with respect to $||\epsilon||_V$) is obtained when $M = 5$ is used and it becomes worse if we add more constraints (the constraints were generated using simple integration, $\mu_k = \int_{-\infty}^{\infty} x^k \, p_3(x) \, \mathrm{d}x$). It shows that the dimensionality of the optimization problem have the strong influence on the accuracy of the reconstruction.

The difficulty of the numerical optimization is often associated with the condition number of the Hessian matrix [81] of the second derivatives (A.7). The condition number

$$\kappa(H) = {}^{|\lambda_{\max}(H)|}\!/_{|\lambda_{\min}(H)|},$$

i.e., the ratio of maximal and minimal eigenvalues of the Hessian matrix $H^{(\ell^*)}$ on the last iteration ℓ^* is also shown in Table 4.3. It gradually increases when additional constraints are added. We plot the distribution p_3 together with the reconstruction $\tilde{q}_3$ obtained with $M = 4$ in Figure 4.4c.

We note that the authors of [248] provide a proof of the unique solution existence for the case of quartic exponential distribution and the special numerical treatment of the maximum entropy optimization problem is given in [177], where the exponential function computation is optimized in order to avoid large numerical precision errors.

Reconstruction of p_4. The triangular distribution p_4 cannot be represented well using the general density shape (4.6). We list the distance $||\epsilon||_V$ between the reconstructed and

M	3	4	5	6	8	10	14
$\|\|\epsilon\|\|_V$,%	-	$6.6 \cdot 10^{-5}$	$5.6 \cdot 10^{-5}$	$3.7 \cdot 10^{-3}$	$2.4 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	$3.1 \cdot 10^{-3}$
$\kappa(H)$	-	$5.4 \cdot 10^{2}$	$4.1 \cdot 10^{3}$	$3.3 \cdot 10^{4}$	$1.5 \cdot 10^{6}$	$1.4 \cdot 10^{8}$	$7.8 \cdot 10^{11}$

Table 4.3: ↩ Reconstruction of quartic exponential distribution p_3.

M	4	6	8	10	14
$\|\|\epsilon\|\|_V$,%	4.2	1.3	1.4	$6.6 \cdot 10^{-1}$	$4.1 \cdot 10^{-1}$
$\kappa(H)$	$7.4 \cdot 10^{1}$	$2.8 \cdot 10^{3}$	$7.9 \cdot 10^{4}$	$2.8 \cdot 10^{6}$	$2.9 \cdot 10^{9}$
$\kappa(H^{\perp})$	$2.4 \cdot 10^{3}$	$1.3 \cdot 10^{4}$	$2.5 \cdot 10^{5}$	$7.5 \cdot 10^{5}$	$1.8 \cdot 10^{5}$

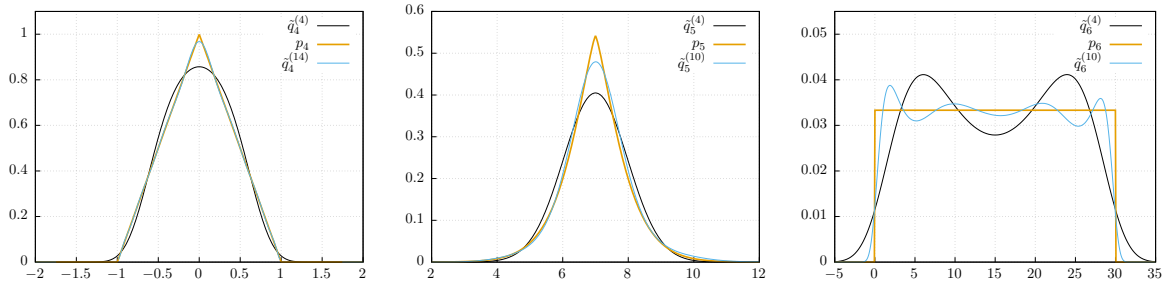
Table 4.4: ↩ Reconstruction of triangular distribution p_4.

original distribution and the conditional number $\kappa(H)$ of the Hessian matrix H in Table 4.4. From here on we use the convergence threshold for dual function $\delta_\Psi = 10^{-4}$, otherwise we indicate it explicitly. The approximation of the support with 5 moment constraints ($M = 4$) uses both

$$\Delta^0(w) = \begin{vmatrix} 1 & 0 & 1/6 \\ 0 & 1/6 & 0 \\ 1 & w & w^2 \end{vmatrix}, \quad \Delta^1(\eta) = \begin{vmatrix} -w_1 & 1/6 - w_1 & -w_1 \\ 1/6 & 0 & 1/15 \\ 1 & \eta & \eta^2 \end{vmatrix},$$

thus, the initial approximation of the support is given by $D^{(0)} = [x_L^{(0)}, x_R^{(0)}] \approx [-0.454, 0.880]$. We also apply the preconditioning technique [3] and list the conditional number $\kappa(H^{\perp})$ of the modified Hessian matrix $H^{\perp}$ calculated at the last iteration of the process. Please note that the preconditioning procedure runs in the beginning of each dual function minimization (with the given approximation of the support). It does not give any improvements in terms of distribution distance $||\epsilon||_V$, however it may mitigate problems associated with high condition number. We observe that it does not make sense to apply these procedures when condition number $\kappa(H)$ is smaller than 10^6. We show the distribution p_4 together with reconstructions obtained with $M = 4$, $\tilde{q}_4^{(4)}$ and $M = 14$, $\tilde{q}_4^{(14)}$ in Figure 4.5a where it can be seen that the plot of $\tilde{q}_4^{(14)}$ is much closer to the original distribution.

Reconstruction of p_5. The reconstruction of the shifted exponential type distribution reveals to be a complicated numerical task. The results of the reconstruction are provided in Table 4.5. The preconditioning technique is applied only partially in this case: only centralization, rotation and scaling steps are performed. If orthogonalization is also applied,



(a) (b) (c)

Figure 4.5: Reconstruction of triangle p_4 (a) ↩ , symmetric exponential type p_5 (b) ↩ and uniform p_6 (c) ↩ distributions.

M	4	6	8	10	14
$\|\|\epsilon\|\|_V$,%	8.1	4.1	4.0	2.9	7.1
$\kappa(H)$	$4.4 \cdot 10^{8}$	$4.0 \cdot 10^{16}$	$3.1 \cdot 10^{19}$	$8.4 \cdot 10^{22}$	$7.0 \cdot 10^{37}$
$\kappa(H^{\circ})$	$6.0 \cdot 10^{5}$	$4.9 \cdot 10^{5}$	$1.7 \cdot 10^{11}$	$3.9 \cdot 10^{12}$	$1.9 \cdot 10^{26}$

Table 4.5: ↩ Reconstruction of shifted exponential type distribution p_5.

either we obtain the worse results (which require longer computational time) or numerical minimization procedure does not converge.

In Table 4.5 we list the results obtained both without and with the preconditioning (the latter corresponds to the Hessian matrix H°). We observe that the preconditioning gives the significant decrease (up to 10 orders of magnitude) in condition number of the Hessian matrix. Interestingly, the condition number $\kappa(H^{\circ})$ decreases when we use $M = 6$ in comparison to the case $M = 4$. The reason is the higher number of approximation support iterations with $M = 4$, namely $\ell^* = 77$, $D^* = [-18.30, 32.53]$ and the dual function converges slow in λ in this case. On the other hand, with $M = 6$ the convergence is faster, $\ell^* = 18$, $D^* = [2.09, 12.22]$.

The reconstructions with $M = 4$, $\tilde{q}_5^{(4)}$ (maximal $||\epsilon||_V$) and $M = 10$, $\tilde{q}_5^{(10)}$ (minimal $||\epsilon||_V$) are shown in Figure 4.5b, where one can observe, that even the best possible reconstruction is not representing the peak value of the true distribution accurately. Both reconstructions of p_4 and p_5 are worse in terms of the distribution distance $||\epsilon||_V$ in comparison to previous case studies.

Reconstruction of p_6. The reconstruction of the uniform distribution p_6 on $[0, 30]$ is also a complicated numerical problem because of the discontinuities present in the distribution density. The results are listed in Table 4.6, where we do not show those obtained with the number of constraints more than 10, $M > 10$ since the numerical optimization does not converge. We plot the two reconstructions (with $M = 4$ and $M = 10$) in Figure 4.5c. All obtained reconstructions demonstrate the oscillatory pattern.

Reconstruction of p_7. The reconstruction of the generalized Pareto distribution p_7 takes many iterations for the support approximation. Another problem is the accuracy of moment approximations, for which the standard `integral` function of `MATLAB` is used.

M	4	6	8	10
$\|\|\epsilon\|\|_V$,%	10	6.6	4.8	3.6
$\kappa(H)$	$1.2 \cdot 10^{11}$	$6.4 \cdot 10^{17}$	$1.3 \cdot 10^{23}$	$1.7 \cdot 10^{28}$

Table 4.6: ↩ Reconstruction of uniform distribution p_6.

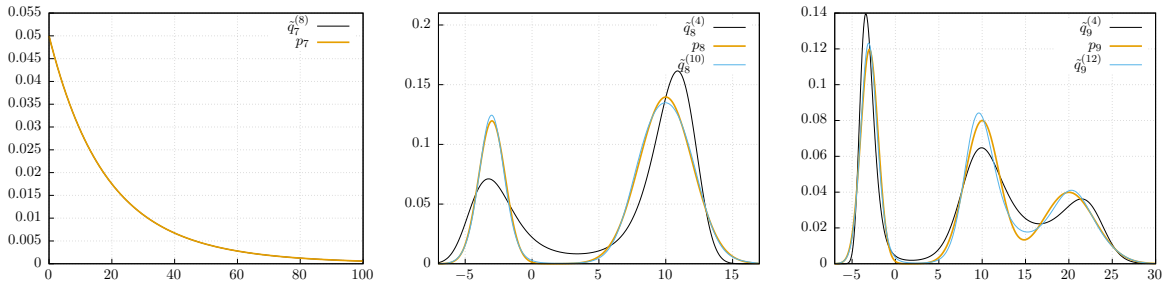
To compute the moments of the theoretical distribution we apply this function over the support $\text{supp}(p_7)$, $\text{supp}(p_7) = \mathbb{R}_{\geq 0}$ in this case

The problem of the moment value approximation and properties of Gauss quadrature and its modifications are well studied [8, 151, 166, 235]. Recently, the problem of approximation the continuous distribution by discrete ones was considered in [213], where the error bounds are provided together with the convergence analysis. Special approach to the numerical approximation of the exponential function computation for the case of quartic exponential reconstruction is given in [177], where it is decomposed into the product of several exponentials with exponents of the lower magnitude order.

In Table 4.7 we list the obtained results where each reconstruction takes a high number of iterations to converge in support. When we use more than $M > 8$ moments, the process does not converge even when preconditioning is applied. However, the obtained reconstructions reveal to be accurate for such heavy-tailed distribution. Only the distribution distance $||\epsilon||_V$ is provided since other metrics are not informative in this case. We plot the best obtained reconstruction $\tilde{q}_7^{(8)}$ together with the ground truth distribution in p_7 Figure 4.6a.

Reconstruction of p_8. The reconstruction of the mixture of two Gaussian distributions shows the capability of the maximum entropy family to reproduce bimodal distributions. Though the results (listed in Table 4.7) indicate the lower accuracy than in the previously considered case studies, such reconstruction still can be used to get the insight to the underlying stochastic process. The case of 4 moment constraints ($M = 3$) does not provide the reasonable reconstruction: the corresponding maximum entropy distribution is $q(x) = \exp(-\lambda_0 - \lambda_1 x - \lambda_2 x^2 - \lambda_3 x^3)$ which can not create bimodal densities (another term $\lambda_4 x^4$ must present in order to allow for bimodality [51]). We plot the mixture distribution p_8 together with reconstructions obtained with $M = 4$ and $M = 10$ in Figure 4.6b.

Reconstruction of p_9. Here we show the reconstruction of trimodal density given by the mixture of three normal distributions. The task of reconstructing such distributions given the limited amount of information is a difficult numerical problem. Therefore, the distance $||\epsilon||_V$ is larger than in the previous case studies. It was possible to obtain the



(a) (b) (c)

Figure 4.6: Reconstruction of generalized Pareto distribution p_7 (a) ↩ , mixture of two Gaussians p_8 (b) ↩ and mixture of three Gaussians p_9 (c). ↩

M		3	4	6	8	10
$\|\|\epsilon\|\|_V$,%	p_7	$2.2 \cdot 10^{-1}$	$3.2 \cdot 10^{-1}$	$3.3 \cdot 10^{-1}$	$1.1 \cdot 10^{-2}$	-
	p_8	48	19	9.2	4.2	3.6
	p_9	40	36	21	12	6.2

Table 4.7: Reconstruction of generalized Pareto distribution p_7 ↩ and mixture of two p_8 ↩ and three p_9 ↩ normal distributions.

convergence only with increased threshold $\delta_\Psi = 5 \cdot 10^{-4}$ and M is even number (or $M = 3$). According to [51], the finite moments of the corresponding densities exist only when M is even and $\lambda_M < 0$. In case of $M = 7$ the numerical optimization procedure was unable to find such a combination of support and parameter vector that satisfies the conditions of distribution validity that we imply (cf. Appendix A.5.3). It is possible to properly reconstruct the three-modal distributions with the maximum entropy distributions of at least 6th order, i.e., at least 7 moment constraints have to be provided ($M = 6$). Please note that if 13 moment constraints ($M = 12$) are used, the corresponding distance further decreases to 3.7%. The results are provided in Table 4.7. We plot the ground truth distribution p_9 together with reconstructions for cases $M = 8$ and $M = 12$ in Figure 4.6c.

Convergence and error control. All the considered case studies show that the reconstructions obtained with the help of maximum entropy principle are able to describe distributions of many shapes using only the limited amount of available data. Generally, the more degrees of freedom are used (larger value of M), the better the reconstruction is. We also observe that the exponential functions with polynomials of even degrees are more easy to fit to the moment constraints; odd order polynomials may give the unbounded rise of the approximated moments, therefore the problem of proper support approximation becomes more difficult.

However, there is no guarantee that the numerical procedure converges for the given moment constraints and no a priori evidence that the reconstruction is accurate enough (generally there is no ground truth solution to compare against). One possible way to partially cope with this problem is to repeat the process varying M and check for similarities in results (both in terms of support approximation D^* and solution vector λ^*).

We also consider the possibility to track the behavior of the dual function. We list the minimal obtained values of the dual function $\Psi(\lambda^*, D^*)$ for the reconstruction of distributions p_6, p_8 and p_9 in Table 4.8. We observe that the smaller value of the dual function corresponds to the smaller value of the distance $||\epsilon||_V$ for all 3 distributions. In the same time, if we consider the exponential type distribution p_5, the value of $||\epsilon||_V$ increases from 2.9% ($M = 10$) to 7.1% ($M = 14$). It corresponds to the increase of the dual function value from 1.3944 to 1.3965. For this simple case studies we can use the number of moment constraints $M + 1$ that corresponds to the reconstruction with the minimum reachable value of Ψ. Unfortunately, this rule of thumb does not hold in general and can not be applied to all problems where the ground truth distribution is unknown and a certain evidence of the approximation accuracy is required.

M		4	6	8	10
$\Psi(\lambda^*, D^*)$	p_6	3.4653	3.4345	3.4217	3.4210
	p_8	2.6620	2.5359	2.5260	2.5175
	p_9	3.5259	3.2391	3.1286	3.0918

Table 4.8: ↩ The value of dual function $\Psi(\lambda^*, D^*)$ for the reconstruction distributions p_6, p_8 and p_9.

The authors of [34] show that the *root mean square* (RMS) deviation Δ_2 can be used to measure the quality of the approximation when analytical solution is unknown. It is defined as

$$\Delta_2(M) = \left[\frac{1}{n_g}\sum_{i=1}^{n_g}\left(\tilde{q}_{M+\Delta M}(x_i) - \tilde{q}_M(x_i)\right)^2\right]^{1/2},$$

where n_g is the number of points used in quadrature approximation of integrals and ΔM is the It is shown that the deviation Δ_2 decreases in the number of moments M almost exponentially. We can consider the generalization $\Delta_{2,c}$ where we do not take the quadrature size into the account as follows

$$\Delta_{2,c}(M) = \left[\int_{x\in D^*}\left(\tilde{q}_{M+\Delta M}(x) - \tilde{q}_M(x)\,\mathrm{d}x\right)^2\right]^{1/2}.$$

However, the instability of the numerical procedure does not allow us to use high number of moments and we can not conduct the proper investigation $\Delta_{2,c}$ applicability. And more than that, the moment closure procedure that is used to approximate moments of the Markov chain becomes stiff as the number of moments is higher than 8 for most systems (cf. Section 3.1.4). Therefore, the only possible way to verify the quality of the reconstruction "on-the-fly" is to reconstruct the distribution with increasing number of moments until the optimization procedure is stable and check for the similarity of the obtained distributions (for instance, using the distance defined by (4.16)).

4.3.6 Maximum Entropy for Discrete Distributions, Numerical Results

Here, we analyze the numerical properties of maximum entropy reconstruction procedure in case of *one-dimensional discrete distributions*. The set of constraints for the the primal problem (4.4) is given by (4.2). The initial approximation for the parameter vector λ is chosen as $\lambda = (\lambda_1, \dots, \lambda_M) = (0, \dots, 0)$, λ_0 is calculated after the dual function is minimized. Using (4.7), we obtain

$$\lambda_0 = \ln Z - 1, \quad Z = \sum_{x \in D^*} \exp\left(- \sum_{k=1}^{M} \lambda_k x^k \right).$$

We consider the following case studies:

1. Poisson distribution, $p_1(x) = {}^{a^k}/_{k!} \exp(-ax)$, $a = 5$, $x \in \mathbb{N}$
2. Hypergeometric distribution, $p_2(x) = \frac{\binom{K}{x}\binom{L-K}{n-x}}{\binom{L}{n}}$, $L = 400$, $K = 100$, $N = 20$, $x \in \{\max(0, n + K - N), \dots, \min(n, K)\}$
3. Discrete uniform distribution, $p_3(x) = {}^{1}/_{N} \mathbb{1}_{(1,\dots,N)}$, $N = 10$, $x \in \{1, \dots, N\}$
4. Mixture of Poisson and binomial distribution with high probability mass at 0, $p_4(x) = w_1 {}^{a^k}/_{k!} \exp(-ax) + w_2 \binom{N}{x} p^x (1-p)^{N-x}$, $w_1 = 0.3$, $w_2 = 0.7$, $a = 1$, $N = 20$, $p = 0.5$, $x \in \mathbb{N}$
5. Mixture of two binomial distributions, $p_5(x) = w_1 \binom{N}{x} p_1^x (1-p_1)^{N-x} + w_2 \binom{N}{x} p_2^x (1-p_2)^{N-x}$, $w_1 = 0.4$, $w_2 = 0.6$, $N = 50$, $p_1 = 0.1$, $p_2 = 0.5$, $x \in \{0, \dots, N\}$
6. Mixture of three binomial distributions, $p_6(x) = \sum_{i=1}^{3} w_i \binom{N}{x} p_i^x (1-p_i)^{N-x}$, $w = (0.3, 0.4, 0.3)$, $N = 50$, $p = (0.1, 0.5, 0.7)$, $x \in \{0, \dots, N\}$

The problem of discrete distribution reconstruction is similar to the reconstruction of continuous distribution. The approximation of the support is done in a similar fashion (iterative extension), however in the discrete setting the minimal possible increment corresponds to exactly one state:

$$x_L^{(\ell+1)} = \max(0, x_L^{(\ell)} - 1) \text{ or } x_R^{(\ell+1)} = x_R^{(\ell)} + 1.$$

Due to this discrete nature of support, there is no need to approximate the value of the integrals (4.5) using quadrature formulas that may include many points of discretization. Instead, we can directly apply the explicit summation over the approximated support $D^{(\ell)}$ that does not include any additional discretization step (A.5)

$$\widetilde{\mu}_i = \sum_{x \in D^{(\ell)}} x^i \exp\left(- \sum_{k=1}^{M} \lambda_k^{(\ell)} x^k \right), \quad i = 1, \dots, 2M.$$

M	4	6	8	10
$\kappa(H)$	$1.2 \cdot 10^{12}$	$7.6 \cdot 10^{19}$	$8.5 \cdot 10^{24}$	$2.6 \cdot 10^{30}$
$\kappa(H^\circ)$	$1.8 \cdot 10^{2}$	$9.6 \cdot 10^{3}$	$5.1 \cdot 10^{6}$	$2.1 \cdot 10^{7}$
$\kappa(H^\perp)$	$5.9 \cdot 10^{1}$	$1.1 \cdot 10^{2}$	$2.6 \cdot 10^{6}$	$1.2 \cdot 10^{3}$

Table 4.9: ↩ Comparison of condition number of Hessian matrix after application of preconditioning for p_6.

Also, the computational time of the reconstruction can be significantly improved by precomputing the values of the exponential function for a certain subset of states $\tilde{S} \in \mathbb{N}$. The normalization constant Z (4.7) can be computed using the tables of values for $x^2, \ldots, x^M$ and at least $(M-2) \cdot |\tilde{S}|$ exponentiation operations can be avoided during each computation of the dual function. The computational time can be similarly saved when moments $\widetilde{\mu}$ are approximated, thus another $(M-1) \cdot |\tilde{S}|$ exponentiation operations of type x^k can be avoided during the computation of the gradient and the Hessian of the dual function. We select the subset of $\mathbb{N}$ to be $\tilde{S} = \{0, \ldots, 500\}$ or $\tilde{S} = \{0, \ldots, 1000\}$ but this is not a restriction: as soon as support approximation goes over $\tilde{S}$, the values of x powers can be computed by usual exponentiation.

The preconditioning techniques can be applied to decrease the condition number of the Hessian matrix $\kappa(H)$. In this case, however, there is no possibility to precompute the values of polynomial functions since it is exactly the setting of the continuous maximum entropy problem. One possible way to combine the benefits of both approaches is to first try the reconstruction based completely on discrete support approximation and apply centralization and orthogonalization approaches only if it fails to converge. We compare the condition number of the Hessian matrix for the case of three binomial distributions mixture p_6 and report results in Table 4.9, where it can be seen that the effect of preconditioning is drastic: applying centralization and rotation (which is normalization of the variance to 1 in one-dimensional case) leads to the decrease of condition number $\kappa(H^\circ)$ for at least 10 orders of magnitude. Interestingly, the orthogonalization does not significantly improves the condition number $\kappa(H^\perp)$ further. Please note that the comparison of the Hessian matrices H, H° and $H^\perp$ is done using the same approximation of support D^* for the fixed order of the problem M (including more states in the set D^* usually further increases the condition number). The similar comparison in case of Chebyshev and Legendre moments application is given in [60].

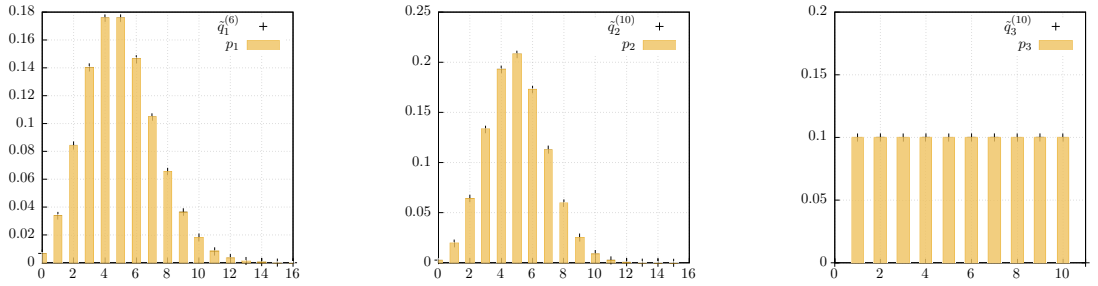
The results of the reconstruction for discrete distributions are given in Table 4.10, where we use the convergence threshold for dual function $\delta_\Psi = 10^{-4}$. We observe that it is generally beneficial to include more constraints and calculate the maximum entropy distribution with more degrees of freedom to obtain the smaller values of the total variation distance $||\epsilon||_V$ defined (for finite or countable set [150]) as L_1 norm

$$||\epsilon||_V = \frac{100\%}{2} \sum_{x \in S} |p(x) - q(x)|. \tag{4.17}$$

M		2	4	6	8	10
$\|\|\epsilon\|\|_V$,%	p_1	4.9	$7.6 \cdot 10^{-1}$	$2.1 \cdot 10^{-1}$	$8.7 \cdot 10^{-2}$	$6.8 \cdot 10^{-2}$
	p_2	2.9	$4.1 \cdot 10^{-1}$	$1.1 \cdot 10^{-1}$	$2.6 \cdot 10^{-2}$	$6.6 \cdot 10^{-3}$
	p_3	18	10	5.7	1.2	$3.7 \cdot 10^{-5}$
	p_4	34	12	3.4	2.6	$9.1 \cdot 10^{-1}$
	p_5	44	16	6.1	3.4	1.9
	p_6	36	24	8.6	6.9	2.1

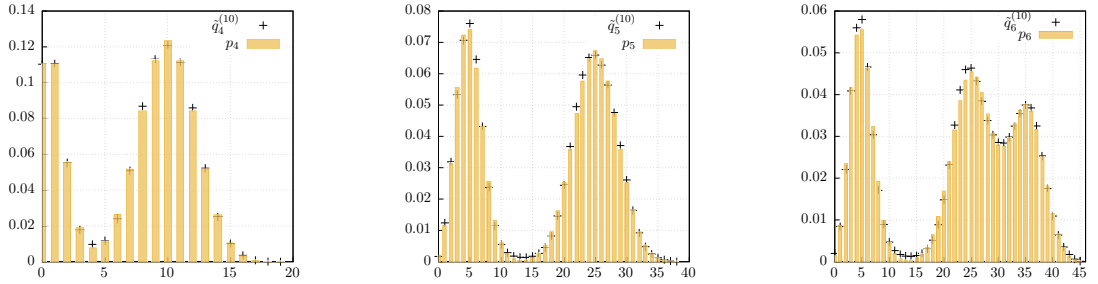
Table 4.10: ↩ Distance $||\epsilon||_V$ between the theoretical distribution and corresponding reconstruction depending on the order M of the highest moment constraint.

We plot the best obtained reconstructions of p_2, p_3, p_3 in Figure 4.7 and of p_4, p_5, p_6 in Figure 4.8, where the ground truth distributions are represented using yellow bars and reconstructions are shown as black crosses. We observe that for all considered case studies, the shape of the theoretical distribution is well captured and the distance $||\epsilon||_V$ decreases in most cases when more moment constraints are considered.



(a) (b) (c)

Figure 4.7: ↩ Reconstruction of Poisson p_1 (a), hypergeometric p_2 (b) and uniform p_3 (c) distributions.



(a) (b) (c)

Figure 4.8: ↩ Reconstruction of discrete distribution mixtures: Poisson and binomial p_4 (a), two binomial p_5 (b) and three binomial p_6 (c) distributions.

Chapter 5

Model Reconstruction with Maximum Entropy Approach

In this chapter we demonstrate the numerical results of the maximum entropy approach when it is applied to the moments of species population counts for a given chemical reaction network. To determine the quality of the probability distribution reconstruction we compare maximum entropy distributions to those obtained via a direct numerical simulation. Thus, we consider only systems of such sizes where a direct numerical simulation is possible (cf. Table 2.1). Clearly, for systems with large population sizes a direct numerical simulation is not feasible while the running time of the moment closure approximation is independent of the population sizes.

In order to distinguish errors that are introduced by the moment closure approximation from errors introduced by the reconstruction, we also compare the obtained moments with those computed based on the distributions obtained via direct numerical simulation. Moreover, we also apply the maximum entropy approach to the more accurate approximation of the moments obtained via direct numerical simulation.

5.1 Maximum Entropy and Method of Moments

We consider the problem of model reconstruction for the given chemical reaction network. We assume that the dynamics of this reaction network is governed by the continuous-time Markov chain $\{\vec{X}(t), t \geq 0\}$ and the corresponding chemical master equation (2.5). The sliding window method can be applied to obtain the approximation of the transient distribution π at the time horizon of interest T. However, this is often a computationally difficult task (if feasible at all), thus we may want to use the moment approximation methods to describe the dynamics in terms of moments of the corresponding distribution. The ODE system for moments is usually much smaller than the one of CME, and easier to solve.

If there is a need to reconstruct the underlying distribution from moments to get more insights in the stochastic process, one may try to solve the classical moment problem.

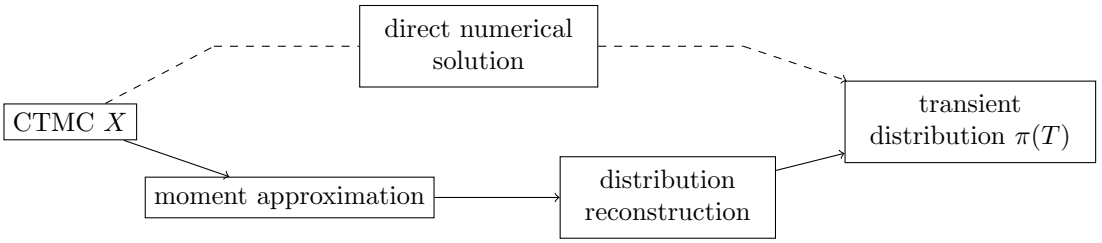


Figure 5.1: ↩ Techniques to approximate the transient distribution.

Here, we use the maximum entropy approach to reconstruct the marginal distributions $\pi_i(x_i, T) = \mathbf{P}[X_i(T) = x_i]$ of a Markov chain $\vec{X}$ at time T. The workflow that we follow in shown by solid line in Figure 5.1.

The values of moments $\mu_0, \ldots, \mu_M$ are approximated by the moment closure method. We use the tool `SHAVE` to export moment closure ODE system as a `MATLAB` file. The ODE system is then solved using `ode45` method with default precision settings (`AbsTol` $= 10^{-6}$, `RelTol` $= 10^{-3}$; if other precision settings or another solver are used, it is noted additionally). For each experiment we use the moments of the highest order M, however moment closure technique computes all the moments up to order $M + 1$. That is, all the moments $\mu_0, \ldots, \mu_{M+1}$ are approximated and the moment of the highest order is ignored in the maximum entropy reconstruction, i.e., reconstruction procedure uses only $\mu_0, \ldots, \mu_M$. We ignore the moment μ_{M+1} due to the high sensitivity of the reconstruction procedure even to the small relative error in the approximation of constraints.

Simple Dimerization model. We reconstruct the marginal distributions of the species P and P_2 and compare with those obtained using the direct numerical simulation (where we chose $\delta = 10^{-15}$ yielding a total approximation error of $\epsilon = 5 \cdot 10^{-15}$, see also (2.6)). For instance, to reconstruct the distribution of P we used the sequence of moments $\mu_0, \ldots, \mu_M$, where $\mu_k = E(X_1^k(t))$ and $X_1(t)$ represents the number of molecules of type P at time t.

In Table 5.1 we show how accurate is the approximation $q(x)$ of individual probabilities $\pi_\cdot(x)$ by calculating the distance (4.17)

$$||\epsilon||_V = \frac{100\%}{2} \sum_{x \in D^*} |\pi_\cdot(x) - q(x)|,$$

where $\pi_\cdot(x)$ is the "true" probability of having x molecules of type i at time T (obtained via the direct numerical simulation of the CME) and $q(x)$ is the value obtained from the combination of moment closure approximation and maximum entropy reconstruction. The distance is calculated for all states in the support approximation D^* obtained with $\delta_\Psi = 10^{-4}$.

Please note that the support of distribution for protein P contains either only even or odd non-negative integers (depending on the initial condition). Here we use the initial condition $P = 301$, $P_2 = 0$, therefore non-zero probability mass of marginal distribution for P is possible only for odd non-negative integers. It requires the corresponding change

M		2	4	6	8
$\|\|\epsilon\|\|_V$,%	P	1.8	6.0	6.0	6.1
	P^{st}	1.8	6.0	2.0	2.8
	P_2	0.4	$3.9 \cdot 10^{-2}$	$4.5 \cdot 10^{-2}$	$7.5 \cdot 10^{-2}$
	P_2^{st}	0.4	$8.3 \cdot 10^{-2}$	$9.4 \cdot 10^{-2}$	1.9

Table 5.1: ⟵ Maximum entropy reconstruction results for the dimerization network.

in the support search procedure since this information can not be derived directly from the moments. To proceed with this derivation automatically, the symbolic approaches of state space construction [58, 144] or the graph reconstruction [211] methods can be applied. In Figure 5.2 we plot the distributions of P and P_2 where we use yellow bars for the distribution obtained via direct numerical simulation and black crosses for the reconstructed distribution. For the order of the moments that were considered, we used for both species the order with which the best approximation was obtained ($M = 2$, $\tilde{q}_P^{(2)}$ for P and $M = 4$, $\tilde{q}_{P_2}^{(4)}$ for P_2).

In addition, we list the approximation error for the case where the maximum entropy reconstruction is applied to the moments calculated from the results of the direct numerical simulation, $\mu_k = \sum_{x \in S} x^k \pi_\cdot(x)$. The corresponding rows in the table are denoted by P^{st} and P_2^{st}. Surprisingly, the reconstruction based on those is worse for the distribution of P_2 and $M = 4$, $M = 6$ which shows that the reconstruction error comes mainly from the numerical procedure of maximum entropy, not from the moment closure approximation in this case. We observe that the maximum entropy method provides the least error if three moment constrains ($M = 2$) are used to reconstruct the marginal distribution for protein P and five moment constraints ($M = 4$) for protein P_2. However, the reconstruction is very accurate in all cases and we suppose that the reason why the distance measure does not decrease when more moments are considered is that the shape of the maximum entropy distribution is already nicely described using the lower-order maximum entropy distribution. Increasing the order further forces the numerical optimization to make the coefficients that correspond to the high order monomials very small (instead of setting them to zero) therefore introducing additional approximation error.

Multi-attractor model. We consider the reconstruction of the marginal distribution of *PaxProt*, *MAFAProt* and *DeltaProt* in the multi-attractor model. The results are given in Table 5.2 for all three proteins. We compare the results with the solution of the direct numerical simulation (where we chose $\delta = 10^{-15}$ yielding a total approximation error of $\epsilon = 6 \cdot 10^{-10}$). We see that the error is minimal when we consider $M = 3$ (4 moment constraints) if moments are obtained using the moment closure approximation. The use of more moments does not increase the accuracy further. We assume that this artifacts arise due to the rapid change of the distribution behavior in the region $\{0, \ldots, 5\}$.

For this case study we used the moment closure with moments up to the order 6. Unfortunately, it was impossible to apply the low dispersion closure of the higher order due to the large size of ODE system. The reconstruction based on the moments calculated from the

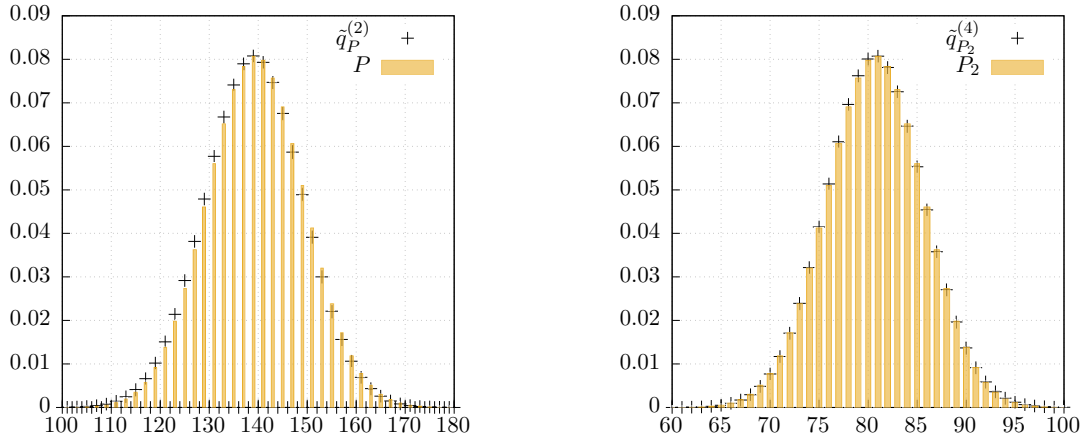
	species \ M	2	3	4	5
$\|\|\epsilon\|\|_V$,%	$MAFAProt$	11	2.9	7.8	8.2
	$MAFAProt^{\text{st}}$	11	1.7	1.4	1.1
	$DeltaProt$	4.36	4.41	5.9	5.9
	$DeltaProt^{\text{st}}$	7.0	3.7	2.0	1.5
	$PaxProt$	13.0	9.4	16	16
	$PaxProt^{\text{st}}$	9.9	9.4	16	18

Table 5.2: ↩ Maximum entropy reconstruction results for the multi-attractor model.

direct numerical solution gives better results in case of *MAFAProt* and provides slightly different reconstruction for the other two proteins.

The best obtained reconstructions (with $M = 3$, $\tilde{q}^{(3)}_{MAFAProt}$ for *MAFAProt*, $M = 3$, $\tilde{q}^{(3)}_{DeltaProt}$ for *DeltaProt* and $M = 3$, $\tilde{q}^{(3)}_{PaxProt}$ for *PaxProt*) are plotted in Figure 5.3, where we use $\delta_\Psi = 10^{-5}$.

Bursty protein production model. We compute an approximation of the moments of species M and P up to order 6 using the moment closure equations given in Example 3.4. The moments of P are then used to reconstruct the marginal probability distribution of P. The statistical distance $||\epsilon||_V$ is listed in Table 5.4 for different values of moment constraints M. We observe that using more moment constraints makes the reconstruction more accurate. The reconstructions obtained with $M = 3$, $\tilde{q}^{(3)}_P$ and $M = 5$, $\tilde{q}^{(5)}_P$ are shown in Figure 5.4a using solid lines together with the plot of the absolute error ϵ (Figure 5.4b) and the result of direct numerical simulation of the CME is shown using yellow bars. While for most parts of the support the distribution is accurately reconstructed even



(a) (b)

Figure 5.2: ↩ Maximum entropy reconstruction of marginal probability distributions of the protein counts P (a), and P_2 (b) at time instant $t = 20$ for simple dimerization system.

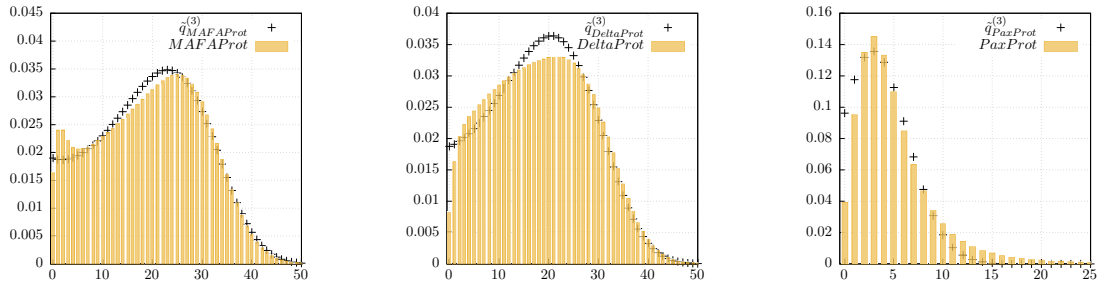
M	metric	$\delta_\Psi = 10^{-2}$	$\delta_\Psi = 5 \cdot 10^{-3}$	$\delta_\Psi = 10^{-3}$	$\delta_\Psi = 10^{-4}$
4	x_R^*	612	674	861	1141
	$\|\|\epsilon\|\|_V$	19.1	19.3	19.5	19.5
	tnc (sec)	4.4	5.0	6.3	8.9
	twc (sec)	4.6	5.2	6.6	9.1
6	x_R^*	579	651	1532	4586
	$\|\|\epsilon\|\|_V$	5.3	5.7	7.5	8.4
	tnc (sec)	3.8	4.6	18.2	178.1
	twc (sec)	3.5	4.1	14.3	69.7

Table 5.3: ↩ The influence of δ_Ψ on the support approximation, accuracy and running time in case of bursty protein production model. The row tnc/twc correspond to the running time when the precomputation of exponential function is not used/used.

with $M = 3$, the method is less precise when considering the probability of small copy numbers of P. Taking more moment constraints into an account increases the accuracy, in particular in the region with more than 100 protein molecules. We may use more moments to improve the reconstruction, for instance, up to the order 7. However, in this case the moment equations become stiff and the numerical integration fails. This happens due to the combination of highly nonlinear derivatives of the rate function $f(x_P) = \frac{\Omega v_M x_P}{\Omega K_M + x_P}$ with large values of the higher order moments.

To obtain these results, we used the thresholds $\delta_\Psi = 10^{-5}$ and $\delta_D = 10^{-5}$ (as defined in Appendix A.5.3). In Table Table 5.3 we investigate the influence of δ_Ψ on the running time, support approximation and accuracy. It can be observed that decreasing the δ_Ψ leads to the larger support (larger values of x_R) and, therefore, the larger running time. The running time for this case study is much larger than for the rest of models and the positive effect of exponential function precomputation (Appendix A.5.3) becomes visible for $M > 3$ and large support.

Please note that we approximated the moments at the quasi-steady state of the system, therefore we track the convergence of the moment values (or, equally, check whether



(a) (b) (c)

Figure 5.3: ↩ Maximum entropy reconstruction of marginal probability distributions of the protein counts $MAFAProt$ (a), $DeltaProt$ (b) and $PaxProt$ (c) at time instant $t = 10$ for multi-attractor model.

M	2	3	4	5
$\|\|\epsilon\|\|_V$,%	8.9	5.8	4.7	2.1

Table 5.4: ↩ Maximum entropy reconstruction results for the bursty protein production network.

the magnitude of each moment derivative is close enough to zero) while integrating the moment closure ODE system.

The numerical results for the considered case studies show that the proposed combination of methods has many advantages. It is a fast and surprisingly accurate way of obtaining the distribution of the system at specific points in time and therefore well suited for computationally expensive tasks such as the approximation of event probabilities.

We compare the reconstruction based on the moments approximated using standard moment closure (MM) and computed using the method of conditional moments (MCM) in Section 5.2.

5.2 Maximum Entropy and Method of Conditional Moments

In this section we present numerical results of the maximum entropy reconstruction when it is applied to the moments of a reaction network approximated using both method of moments (MM) and method of conditional moments (MCM). The MCM method allows one to keep the original probability-based representation for some of the chemical populations. This makes sense for those populations that are very small. For instance, a population may represent the binding of a transcription factor to a promoter region

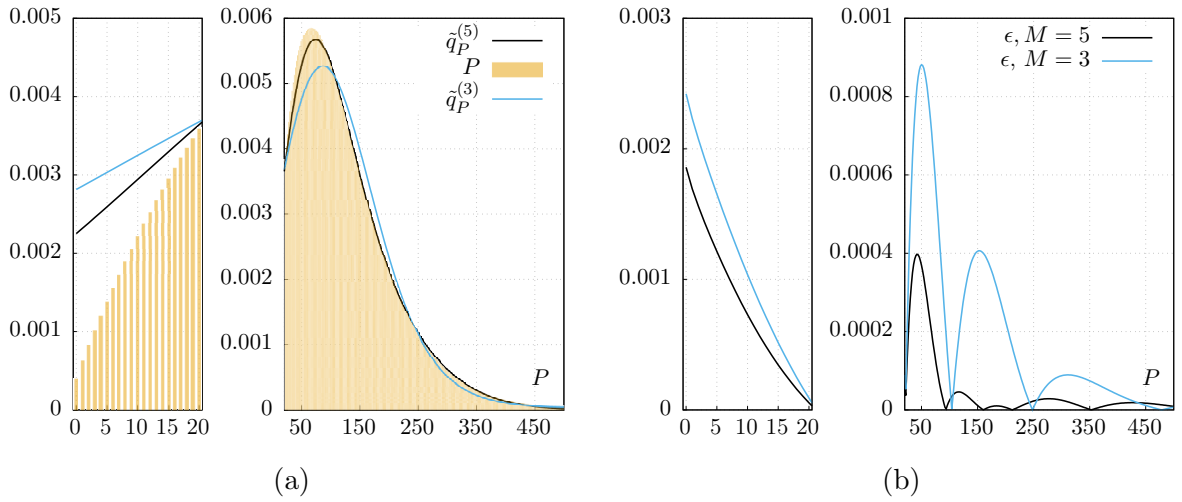


Figure 5.4: ↩ Maximum entropy reconstructions of marginal probability distributions of the protein P (a) and the absolute error (b).

where only the values 0 (promoter free) and 1 (transcription factor bound to promoter) are possible. Similarly, in many systems one has to deal at the same time with very large and very small populations (small populations may refer to species with approximately ten molecules at most) with significant probability.

Clearly, for such cases a pure moment-based description is not appropriate and it is usually better to integrate over time the probabilities of having $0, 1, \ldots, 10$ molecules. In Hasenauer et al. [105] a hybrid integration scheme for the moments of large populations conditioned on the actual molecule numbers of small populations is derived from the CME. In [223] the authors approximate the distribution of gene expression products conditioned on the state of the promoter using a linear noise approximation, where they make a quasi-steady state assumption for the reactions involving highly abundant species and the result is given as a closed form expression. In [16] the authors apply system size expansion of higher order to derive the analytic expression that describes the underlying distribution. In [105] the MM and MCM are compared in terms of accuracy, but the comparison is limited to the moment values.

Our main focus is on investigating whether the MCM approach is more suitable for the reconstruction than the MM approach. We reconstruct one- and two-dimensional marginal distributions in three different ways, in the following referred to as *weighted sum MCM*, *joint MCM* and *MM*. The last case refers to a reconstruction based on the moments obtained from the MM approach. Two former approaches refer to the reconstructions based on the MCM approach where *weighted sum* means that we first reconstruct conditional distributions and derive the full marginal by multiplying with the probability of the corresponding condition (e.g. gene is active or not) and summing up. In contrast, *joint MCM* means that we use the unconditional moments (approximated by the product of conditional moments and the probability of the condition) for the reconstruction. In addition, we reconstruct the conditional distributions from the conditional moments and compare them to the conditional distributions obtained via a direct numerical integration of the CME.

We illustrate the *weighted sum MCM* method for one-dimensional discrete distributions through the following example:

Example 5.1. ↩ We consider the gene expression model where we reconstruct the marginal distribution of protein molecules $\mathbf{P}[\vec{X}_P(t) = x] = \pi_{\vec{X}_P}(x, t)$. The moments $\mu_k = E\left(\vec{X}_P^k\right)$ and the corresponding conditional moments are obtained using the MCM and MM equations, for $k = 0, \ldots, M+1$. In the case of *joint MCM* and *MM* we use the first M moments' values as constraints in (4.2) and solve the maximum entropy optimization problem (4.4). In both cases, the solution is given by a pair (λ^*, D^*) of the parameter vector λ^* and the truncated support D^*. The corresponding reconstructed distribution is defined as

$$\tilde{q}(x) = \begin{cases} \exp(-1 - \sum_{k=0}^{M} \lambda_k^* x^k), & x \in D^* \\ 0, & x \notin D^*. \end{cases}$$

In order to apply the *weighted sum MCM*, we reconstruct the conditional distribution from the sequences $\mu_{P_{\text{off}},k}$ and $\mu_{P_{\text{on}},k}$ that approximate the conditional moments $E\big(X_P^k|D_{\text{off}}=1\big)$ and $E\big(X_P^k|D_{\text{on}}=1\big)$. Here, X_P corresponds to the number of proteins and the condition $D_{\text{off}}=1$ ($D_{\text{on}}=1$) refers to the state of the gene. These sequences of moments are obtained using the MCM approach together with the approximation of the mode probabilities p_{off} and p_{on} (cf. Example 3.9). We solve the maximum entropy problem for each moment sequence and the reconstruction of marginal unconditional distribution is given by

$$\tilde{q}_{wsMCM}(x)=\begin{cases} p_{\text{off}}\,\tilde{q}_{\text{off}}(x), & x\in D^*_{P_{\text{off}}}\setminus D^*_{P_{\text{on}}}\\ p_{\text{on}}\,\tilde{q}_{\text{on}}(x), & x\in D^*_{P_{\text{on}}}\setminus D^*_{P_{\text{off}}}\\ p_{\text{off}}\,\tilde{q}_{\text{off}}(x) & \\ \quad +p_{\text{on}}\,\tilde{q}_{\text{on}}(x), & x\in D^*_{P_{\text{off}}}\cap D^*_{P_{\text{on}}}, \end{cases}$$

where $\tilde{q}_{\text{off}}(x)$ and $\tilde{q}_{\text{on}}(x)$ are the reconstructions of the conditional distributions.

Gene expression model. For the gene expression reaction network the moment-based analysis using MM and the MCM approach is very accurate, but the MCM approach has lower relative errors for high moments (see [105]).

We first consider the one-dimensional marginal distributions of R and P. We then compare the approximation errors of all three reconstruction methods at time $t=10$ and provide the results in Table 5.5. There, the first two columns refer to the approximation error of the conditional distributions for protein (mRNA) denoted by $P|D_{\text{off}}$ ($R|D_{\text{off}}$) and $P|D_{\text{on}}$ ($R|D_{\text{on}}$). The last three columns refer to the reconstructions of the marginal distribution obtained using *weighted sum MCM*, *joint MCM* and *MM*, respectively.

We find that the reconstruction is accurate for all $M>2$ and the best result for the distribution of mRNA (R) is obtained when the *joint MCM* method is applied with $M=7$ yielding a statistical percentage error of 0.003%. Thus, for this distribution an accurate approximation of the unconditional moments yields the best reconstruction since here the conditional distributions are slightly harder to reconstruct than the (unconditional) marginal distributions (approximation error of *weighted sum MCM* is 0.03%). The distribution of protein molecules P is reconstructed most accurately when the *joint MCM* or *MM* is applied with $M=3$ while reconstructions using the moments obtained with MCM are slightly worse. When we use moments of order $M>3$ for reconstruction, *MM* usually gives worse results than *joint MCM*. The approximation error of the distribution reconstructed using *weighted sum MCM* is of the same order as of the conditional distributions and it does not significantly increase. It shows that rather high under- and overapproximations of conditional distributions can still provide an accurate approximation of the marginal distribution after summation and weighting. We note that the probabilities p_{off} and p_{on} of modes D_{off} and D_{on} are (approximately) 0.44 and 0.56 (cf. Example 3.9), therefore the conditional distributions bring equally important contribution to the final result of the reconstruction. The relative approximation error for mode probabilities computed with MCM is of order 10^{-5}.

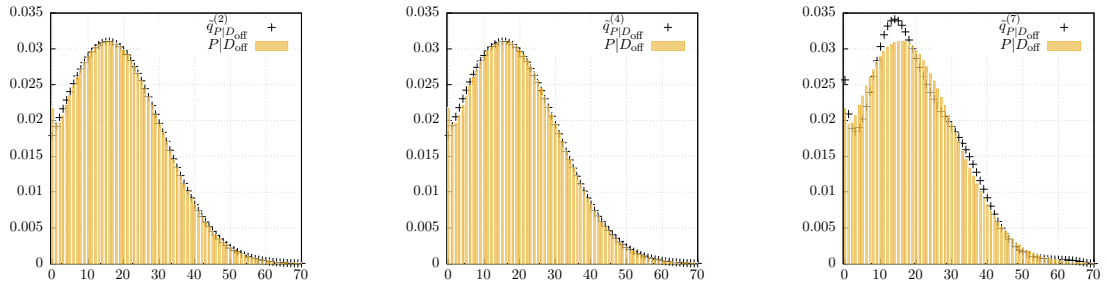
We note that the reconstruction results are generally quite similar for the approaches that are based on an approximation of the conditional and unconditional moments. However, the MCM approach has the advantage that the distribution of species such as DNA is very accurate, since it is directly available and is not reconstructed from the moments. A moment-based approach such as MM would need large number of moments for an accurate reconstruction [25] of such distributions with very small support.

The sensitivity of the optimization procedure can also influence the final result. The reconstruction that uses the fewer degrees of freedom can provide an accurate solution since the distribution of the simple shape is able to explain the main behavior. At the same time, adding more moments into the consideration allows one to capture more details, but it may change the reconstruction drastically due to the sensitivity, and the corresponding approximation error can become larger.

To the best of our knowledge, there exist no criteria that provide the number of moments that have to be considered such that adding more information does not significantly change the maximum entropy reconstruction. We show that in Figure 5.5, where we plot the reconstructions of the conditional distribution $P|D_{\text{off}}$ and use $M \in \{2, 4, 7\}$. The reconstruction using $M = 7$ moments has the largest approximation error, but it is able to capture the complex nature of the distribution by treating the point $P = 0$ differently (however, this results in the effect similar to overfitting, cf. Figure 5.5c).

Reconstruction using moments of the order higher than $M > 7$ is not reasonable: the approximation error of both MM and MCM becomes high and the results are less accurate. In order to solve the moment closure system for the case $M = 7$ (where we truncate at the order 8, i.e., all the central moments of order ≤ 9 are set to 0), we use the solver `ode23s` in order to partially remedy the stiffness of the ODE system. The moment approximation with MCM method remains fast enough with $M = 7$ to provide the computational time gain in comparison to the direct simulation of the CME, while MM becomes very slow (running time is 6 minutes) and the difference in reconstruction results in not large.

In Figure 5.6 we plot the best reconstructions (shown as crosses) of the marginal distribution of protein and mRNA molecules obtained with *joint MCM*, $M = 3$ and $M = 7$ respectively.

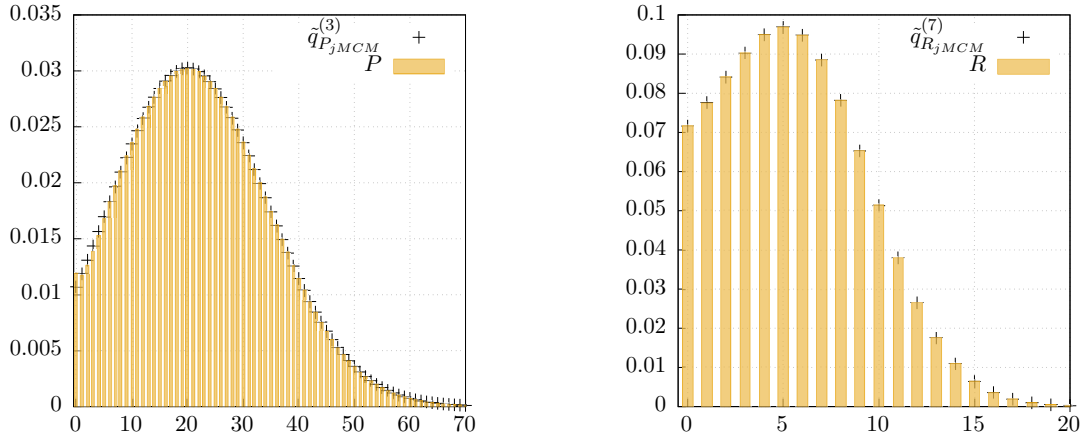


(a) (b) (c)

Figure 5.5: ↩ Approximation of the conditional distribution of protein $P|D_{\text{off}}$. The number of moments used for reconstruction is $M = 2$ (a), $M = 4$ (b) and $M = 7$ (c).

	M	$P\|D_{\text{off}}$	$P\|D_{\text{on}}$	P_{wsMCM}	P_{jMCM}	P_{MM}
$\|\|\epsilon\|\|_V$	2	$5.5 \cdot 10^{-1}$	1.9	1.1	$7.2 \cdot 10^{-1}$	$7.1 \cdot 10^{-1}$
	3	$8.7 \cdot 10^{-1}$	$9.3 \cdot 10^{-1}$	$5.2 \cdot 10^{-1}$	$2.6 \cdot 10^{-1}$	$2.6 \cdot 10^{-1}$
	4	$7.1 \cdot 10^{-1}$	$7.1 \cdot 10^{-1}$	$3.4 \cdot 10^{-1}$	$2.7 \cdot 10^{-1}$	$4.1 \cdot 10^{-1}$
	5	1.4	$6.2 \cdot 10^{-1}$	$7.6 \cdot 10^{-1}$	$2.8 \cdot 10^{-1}$	$4.5 \cdot 10^{-1}$
	6	1.8	$6.6 \cdot 10^{-1}$	$6.0 \cdot 10^{-1}$	$3.1 \cdot 10^{-1}$	$6.3 \cdot 10^{-1}$
	7	4.1	$6.9 \cdot 10^{-1}$	1.9	$3.5 \cdot 10^{-1}$	$8.7 \cdot 10^{-1}$
		$R\|D_{\text{off}}$	$R\|D_{\text{on}}$	R_{wsMCM}	R_{jMCM}	R_{MM}
$\|\|\epsilon\|\|_V$	2	1.1	1.8	$7.3 \cdot 10^{-1}$	2.3	2.4
	3	$3.8 \cdot 10^{-1}$	$8.8 \cdot 10^{-1}$	$3.6 \cdot 10^{-1}$	$9.6 \cdot 10^{-1}$	$9.6 \cdot 10^{-1}$
	4	$1.8 \cdot 10^{-1}$	$7.3 \cdot 10^{-1}$	$3.4 \cdot 10^{-1}$	$1.6 \cdot 10^{-1}$	$3.5 \cdot 10^{-1}$
	5	$1.7 \cdot 10^{-1}$	$5.9 \cdot 10^{-1}$	$3.8 \cdot 10^{-1}$	$6.0 \cdot 10^{-2}$	$1.0 \cdot 10^{-1}$
	6	$2.7 \cdot 10^{-1}$	$3.9 \cdot 10^{-1}$	$1.0 \cdot 10^{-1}$	$6.4 \cdot 10^{-2}$	$2.9 \cdot 10^{-1}$
	7	$4.3 \cdot 10^{-1}$	$2.5 \cdot 10^{-1}$	$3.1 \cdot 10^{-1}$	$3.0 \cdot 10^{-2}$	$2.6 \cdot 10^{-1}$

Table 5.5: ↩ Maximum entropy reconstruction results for the gene expression network.



(a) (b)

Figure 5.6: ↩ Gene expression: reconstruction (crosses) of the marginal distribution of proteins (a) and mRNA (b) compared to the CME solution (bars).

The reconstruction of the *two-dimensional* marginal distributions also requires solving the maximum entropy problem but with larger dimensionality. However, we solve the same moment closure system, so the additional complexity comes only from the reconstruction process itself. We demonstrate the differences that arise in the two-dimensional case with the Example 5.2:

Example 5.2. ↩ We again consider the gene expression model where we reconstruct the two-dimensional marginal distribution of mRNA and protein molecules $\mathbf{P}[\vec{X}_{R,P}(t) = (x, y)] = \pi_{\vec{X}_{R,P}}(x, y, t)$. We first approximate the probabilities of modes $p_{\text{off}} = \mathbf{P}[X_{D_{\text{off}}} = 1]$ and $p_{\text{on}} = \mathbf{P}[X_{D_{\text{on}}} = 1]$ (cf. Eq. 3.12). In addition, the conditional moments

$$\mu_{\text{off};r,l} = E\big(X_R^r X_P^l | X_{D_{\text{off}}} = 1\big)$$
$$\mu_{\text{on};r,l} = E\big(X_R^r X_P^l | X_{D_{\text{on}}} = 1\big)$$

are approximated for $0 \leq r + l \leq M + 1$. The constraints (4.3.4) for the maximum entropy problem are given by the elements of these moment sequences for $0 \leq r+l \leq M$ and the corresponding solutions of the optimization problem are given by the two pairs $(\lambda^*_{\text{off}}, D^*_{\text{off}})$ and $(\lambda^*_{\text{on}}, D^*_{\text{on}})$. The reconstructed distribution is given by

$$\tilde{q}_{wsMCM}(x, y) = \begin{cases} p_{\text{off}}\ \tilde{q}_{\text{off}}(x, y), & (x, y) \in D^*_{\text{off}} \setminus D^*_{\text{on}} \\ p_{\text{on}}\ \tilde{q}_{\text{on}}(x, y), & (x, y) \in D^*_{\text{on}} \setminus D^*_{\text{off}} \\ p_{\text{off}}\ \tilde{q}_{\text{off}}(x, y) + p_{\text{on}}\ \tilde{q}_{\text{on}}(x, y), & (x, y) \in D^*_{\text{off}} \cap D^*_{\text{on}}, \end{cases}$$

where $\tilde{q}_{\text{off}}(x, y) = \exp(-1 - \sum_{1 \leq r+l \leq M} \lambda^*_{\text{off},r,l} x^r y^l)$ and
$\tilde{q}_{\text{on}}(x, y) = \exp(-1 - \sum_{1 \leq r+l \leq M} \lambda^*_{\text{on},r,l} x^r y^l)$.

To approximate the moment values in the two-dimensional case (4.15), we truncate the infinite support and consider the subset $D^*_{xy} = D^*_x \times D^*_y$, where D^*_x and D^*_y are the approximations of the support of the corresponding one-dimensional marginal distributions $\mathbf{P}[X_\cdot(t) = x]$ and $\mathbf{P}[X_\circ(t) = y]$. Again, we choose D^*_{xy} such that the relative change of the dual function (4.11) becomes smaller than the threshold δ_Ψ. The approximation $\tilde{q}(x, y)$ of the marginal distribution $p_{\cdot,\circ}(x, y, t) = \mathbf{P}[X_\cdot(t) = x, X_\circ(t) = y]$ is then defined by the result λ^* of the iteration procedure such that $p_{\cdot,\circ}(x, y, t) \approx \tilde{q}(x, y)$ if $(x, y) \in D^*_{xy}$ and $p_{\cdot,\circ}(x, y, t) \approx 0$ if $(x, y) \notin D^*_{xy}$.

We list the approximation error for the two-dimensional marginal distributions of R and P of the gene expression network (both conditional and unconditional) in Table 5.6. For the sake of readability we denote the reconstructed distribution by $\tilde{q}$ in the following tables. For instance, the approximation of the joint marginal distribution of R and P under the condition $D_{\text{off}} = 1$ is denoted by $\tilde{q}_{D_{\text{off}}}$.

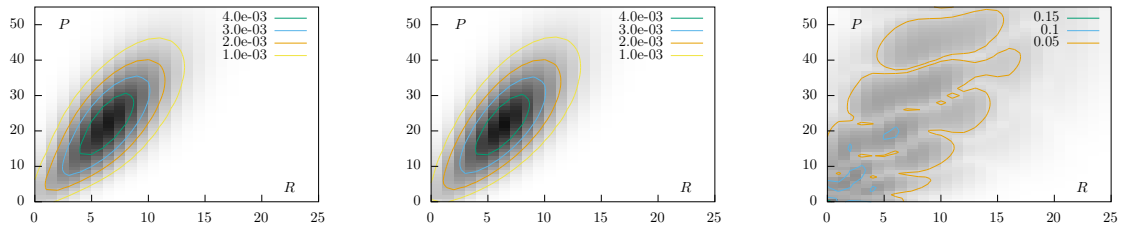
We observe that the relative error of the reconstruction decreases when higher-order moments are used. The minimum value of error $||\epsilon||_V = 2.1\%$ for the marginal reconstruction is obtained using $M = 6$ in *weighted sum MCM*. While all three methods have similar

	M	$\tilde{q}_{D_{\text{off}}}$	$\tilde{q}_{D_{\text{on}}}$	$\tilde{q}_{wsMCM}$	$\tilde{q}_{jMCM}$	$\tilde{q}_{MM}$
$\|\|\epsilon\|\|_V$	2	12	5.1	6.9	8.3	8.2
	3	9.1	2.5	4.9	6.3	6.3
	4	6.0	1.8	3.4	4.8	4.7
	5	5.3	1.5	2.8	4.0	4.1
	6	3.9	1.2	2.1	3.0	2.8
	7	5.8	1.1	2.7	3.0	-

Table 5.6: ↩ Maximum entropy reconstruction of two-dimensional marginal distribution of mRNA and protein for the gene expression network.

errors, the MCM approach uses fewer moment equations. The best reconstruction of the conditional distribution with $D_{\text{on}} = 1$ is shown in Figure 5.7, together with the corresponding solution of the CME and the approximation error $\epsilon_4(x, y) = |\tilde{q}_{D_{\text{on}}}(x, y) - p_{D_{\text{on}}}(x, y)|^{\frac{1}{4}}$. We take the power $1/4$ of the norm in L^1 in order to better distinguish very small numbers; for instance, the contour value 0.05 refers to the L^1 distance $6.25 \cdot 10^{-6}$. Please note that the dual function convergence threshold is chosen as $\delta_\Psi = 10^{-2}$.

We also consider another set of parameters for the gene expression kinetics (cf. Example 3.9). Because of the complex shape of conditional distributions (more than a half of total probability mass corresponds to state 0) the reconstruction is less accurate than in the previous case. It was not possible to reconstruct the distribution $R_{D_{\text{off}}}$ for $M = \{5, 6, 7\}$ (with the value of threshold $\delta_D = 10^{-3}$, as defined in (A.9)) and we use moments up to $M = 3$ for the approximation. However, it was still possible to reconstruct both marginal distributions and the best result is obtained using *weighted sum MCM*. We plot the conditional $P_{D_{\text{off}}}$, $R_{D_{\text{off}}}$ (in Figure 5.8) and marginal distributions of P and R (in Figure 5.9) together with the reconstructions (for conditional distributions we show the reconstructions obtained with $M = 4$ and $M = 7$, and for marginal distributions we plot the *weighted sum MCM* approximations calculated with $M = 7$ for P and with $M = 3$ for R). We observe that such distributions with a lot of probability mass at 0 are more difficult to reconstruct. It is possible to address this feature using maximum entropy reconstruction



(a) (b) (c)

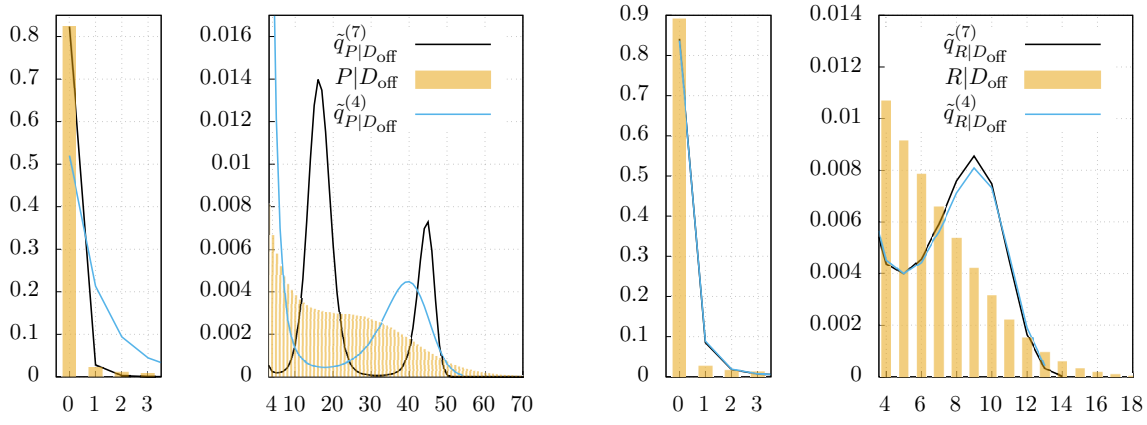
Figure 5.7: ↩ Two-dimensional conditional distribution of the proteins and mRNA (gene expression), where $D_{\text{on}} = 1$. We plot the reconstructed distribution (a), the solution of the CME (b) and the approximation error ϵ_4 (c).

	M	$P\|D_{\text{off}}$	$P\|D_{\text{on}}$	P_{wsMCM}	P_{jMCM}	P_{MM}
$\|\|\epsilon\|\|_V$	2	66	6.9	28	47	34
	3	62	5.0	26	34	32
	4	36	6.0	16	31	32
	5	32	3.8	13	16	25
	6	33	5.6	16	14	46
	7	9.3	12	9.7	10	26
		$R\|D_{\text{off}}$	$R\|D_{\text{on}}$	R_{wsMCM}	R_{jMCM}	R_{MM}
$\|\|\epsilon\|\|_V$	2	28	2.6	13	40	33
	3	4.6	1.2	1.4	21	27
	4	7.6	1.5	5.0	20	24
	5	7.2	$9.1 \cdot 10^{-1}$	5.0	7.4	17
	6	7.2	$6.8 \cdot 10^{-1}$	5.2	5.8	23
	7	7.2	$7.3 \cdot 10^{-1}$	5.2	4.2	13

Table 5.7: ↩ Maximum entropy reconstruction for gene expression network (second parameter set).

only using the approximation of moments computed using MCM. The corresponding numerical results are listed in Table 5.7.

Exclusive switch model. Next, we address the accuracy of the reconstruction of conditional and marginal distributions of the exclusive switch model. In Table 5.8 the approximation error $||\epsilon||_V$ is listed for the conditional distributions for both proteins where we condition on the three possible states of the promoter, i.e., $DNA = 1$, $DNA.P_1 = 1$ or $DNA.P_2 = 1$.

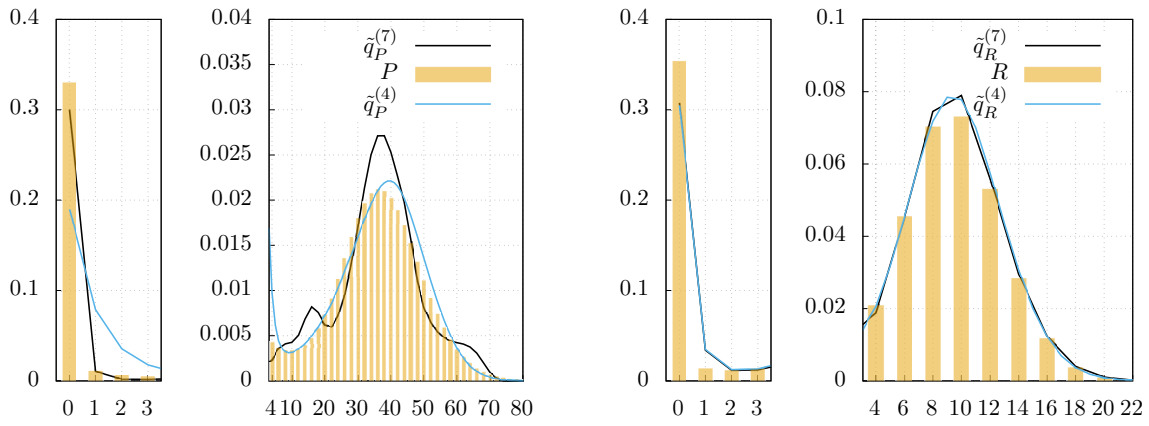


(a) (b)

Figure 5.8: ↩ Reconstruction of conditional distributions $P_{D_{\text{off}}}$ (a) and $R_{D_{\text{off}}}$ (b) for gene expression model.

We observe that the approximation error is minimal for both proteins P_1 and P_2 when the *weighted sum MCM* approach is applied. Here we do not use $M = 2$ since the corresponding reconstructions are very inaccurate, as well as with $M = 3$ (the results are provided for the clarity of comparison). Thus, for the exclusive switch system it is advantageous to approximate the marginal distributions by first reconstructing the conditional distributions and computing the weighted sum. In most cases the error decreases when more information about the moments is used. Because of the complex bi-modal shape of the distributions it is beneficial to include higher order moments as constraints.

In Figure 5.10 we show the reconstructions of both conditional (a, b) and marginal (c) distributions of P_1. Here, the *weighted sum MCM* was used with $M = 7$ to reconstruct the marginal distribution. We also plot the absolute approximation error ϵ_{P_1} in Figure 5.11 for the two conditional distributions ($DNA.P_1 = 1$ and $DNA.P_2 = 1$) and the marginal distribution, where we observe the oscillatory pattern. For example, the absolute error ϵ_{P_1} for the marginal distribution of P_1 is calculated as $\epsilon_{P_1}(x) = |q_{P_1}(x) - \pi_{P_1}(x)|$.



(a) (b)

Figure 5.9: ⟵ Reconstruction of marginal distributions of proteins P (a) and mRNA R (b) for gene expression model.

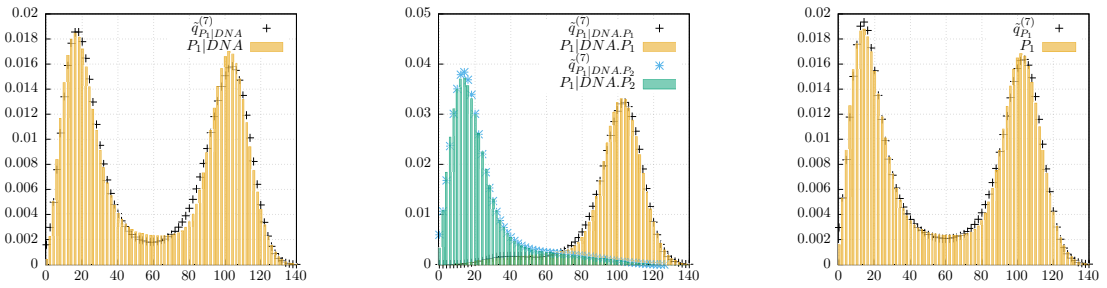
	M	$P_1\|DNA$	$P_1\|DNA.P_1$	$P_1\|DNA.P_2$	$P_{1,wsMCM}$	$P_{1,jMCM}$	$P_{1,MM}$
	3	32	13	21	15	30	29
	4	5.1	6.5	4.4	4.6	5.3	29
$\|\|\epsilon\|\|_V$	5	3.6	4.0	3.1	3.3	4.0	6.2
	6	4.6	4.9	4.8	3.8	4.0	4.1
	7	3.6	3.0	1.5	2.1	3.8	15
		$P_2\|DNA$	$P_2\|DNA.P_1$	$P_2\|DNA.P_2$	$P_{2,wsMCM}$	$P_{2,jMCM}$	$P_{2,MM}$
	3	40	22	26	21	37	37
	4	8.7	4.4	8.8	6.6	9.1	36
$\|\|\epsilon\|\|_V$	5	7.7	4.5	6.8	5.4	8.4	8.3
	6	5.7	3.4	6.3	4.6	4.9	6.2
	7	5.3	2.3	6.3	4.3	5.1	6.2

Table 5.8: ↩ Approximation errors for the distribution of proteins P_1 and P_2 in exclusive switch model.

We also consider the conditional and marginal two-dimensional distributions of proteins P_1 and P_2 in the following example:

Example 5.3. ↩ We consider the reconstruction of two-dimensional marginal distribution $P(X_{P_1} = x, X_{P_2}(t) = y)$ of proteins P_1 and P_2. We first approximate the mode probabilities $p_1 = P(DNA = 1)$, $p_2 = P(DNA.P1 = 1)$ and $p_3 = P(DNA.P2 = 1)$ (cf. Eq. 3.12). In addition, the conditional moments

$$\mu_{1;r,l} = E\big(X_{P_1}^r X_{P_2}^l | DNA = 1\big)$$
$$\mu_{2;r,l} = E\big(X_{P_1}^r X_{P_2}^l | DNA.P1 = 1\big)$$
$$\mu_{3;r,l} = E\big(X_{P_1}^r X_{P_2}^l | DNA.P2 = 1\big)$$



(a) (b) (c)

Figure 5.10: ↩ Exclusive switch: approximations of the conditional distributions of protein P_1 where $DNA = 1$ (a), $DNA.P_1 = 1$, $DNA.P_2 = 1$ (b) and the reconstruction of the marginal distribution (c). The solution of the CME is plotted with yellow bars and the reconstructions are plotted with black crosses (green bars and blue stars are used for the conditional distribution (b) where $DNA.P_2 = 1$). The reconstruction of the marginal distribution (c) is obtained using *weighted sum MCM* with $M = 7$.

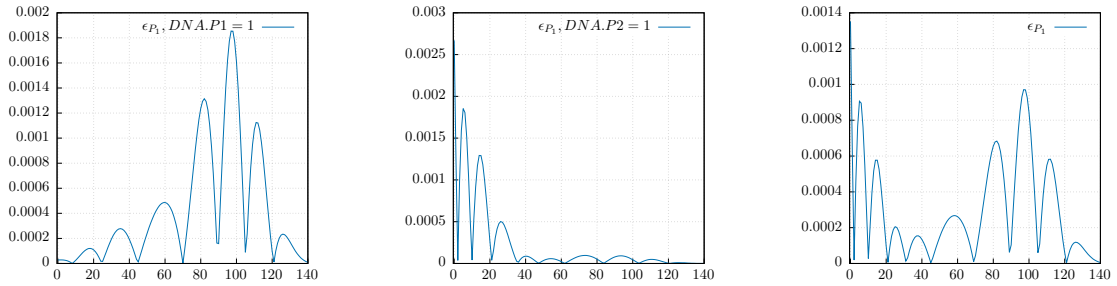
are approximated for $0 \leq r+l \leq M+1$, where $DNA = 1$ refers to the case where promoter is free and $DNA.P1 = 1$ ($DNA.P2 = 1$) corresponds to the case where a molecule of type P_1 (type P_2) is bound to the promoter. The constraints (4.3.4) for the maximum entropy problem are given by the elements of these three sequences for $0 \leq r+l \leq M$ and the corresponding solutions of the optimization problem are given by the pairs (λ_i^*, D_i^*), $i = \{1,2,3\}$. Then the joint reconstructed distribution is given by

$$\tilde{q}_{wsMCM}(x,y) = \begin{cases} p_1\,\tilde{q}_1(x,y), & (x,y) \in D_1^* \setminus (D_2^* \cup D_3^*) \\ p_2\,\tilde{q}_2(x,y), & (x,y) \in D_2^* \setminus (D_1^* \cup D_2^*) \\ p_3\,\tilde{q}_3(x,y), & (x,y) \in D_3^* \setminus (D_1^* \cup D_2^*) \\ \sum_{i=1}^{2} p_i\,\tilde{q}_i(x,y), & (x,y) \in (D_1^* \cap D_2^*) \setminus D_3^*, \\ \sum_{i=\{1,3\}} p_i\,\tilde{q}_i(x,y), & (x,y) \in (D_1^* \cap D_3^*) \setminus D_2^*, \\ \sum_{i=2}^{3} p_i\,\tilde{q}_i(x,y), & (x,y) \in (D_2^* \cap D_3^*) \setminus D_1^*, \\ \sum_{i=1}^{3} p_i\,\tilde{q}_i(x,y), & (x,y) \in D_1^* \cap D_2^* \cap D_3^*, \end{cases}$$

where $\tilde{q}_i(x,y) = \exp(-1 - \sum_{1 \leq r+l \leq M} \lambda_{r,l}^* x^r y^l)$.

Numerical results are given in Table 5.10. Again, we condition on the state of the promoter region, e.g. $\tilde{q}_{DNA.P_1}$ corresponds to the joint distribution of proteins P_1 and P_2 when $DNA.P_1 = 1$. It can be seen that the most accurate reconstructions are obtained with $M = 7$ when *weighted sum MCM* is applied. The *MM* method does not provide the same level of accuracy nor is compatible with respect to the running time: the solution of ODE system for moment approximation is much slower in comparison to MCM. As for the gene expression system, the convergence threshold for the support approximation procedure is chosen $\delta_\Psi = 10^{-2}$. We show the reconstructions of three conditional distributions in Figure 5.12 where the plots refer to the conditions (from left to right) $DNA = 1$, $DNA.P_1 = 1$ and $DNA.P_2 = 1$.

The reconstruction of the marginal distribution obtained using *weighted sum MCM* (*joint MCM*) is shown together with the approximation error in Figure 5.13 (Figure 5.14) . In



(a) (b) (c)

Figure 5.11: ↩ The absolute approximation error of the conditional distributions of protein P_1 where $DNA.P_1 = 1$ (a), $DNA.P_2 = 1$ (b) and the approximation error of the marginal distribution reconstruction (c) for the exclusive switch model.

model	precomputation	4	5	6	7	8
gene expression	tnc	3.1	4.7	7.2	11.4	36
	twc	2.5	2.6	2.6	2.6	2.6
exclusive switch	tnc	6.5	12.8	28.0	44.8	> 300
	twc	2.4	5.4	9.9	11.8	29.5

Table 5.9: ↩ The running time of the reconstruction in two-dimensional case for gene expression and exclusive switch case.

Figure 5.15 we also plot the marginal distributions of P_1 and P_2 calculated via marginalization of the two-dimensional reconstruction. For example, the distribution $\tilde{q}_{P_1,jMCM}(x)$ is calculated as $\tilde{q}_{P_1,jMCM}(x) = \sum_{y \in D^*_{P_2}} \tilde{q}_{jMCM}(x,y)$ for all $x \in D^*_{P_1}$.

The reconstruction process for the exclusive switch model usually takes more time than for gene expression model because of a much larger support.

The running time for all one-dimensional experiments (except for bursty protein production) do not exceed 2 second; it only gets to 4 seconds if the procedure has to restart because it can not converge for the given set of moment constraints. We list the running time for the two-dimensional reconstructions both for gene expression (fast kinetics) and exclusive switch system in Table 5.9, where we assume compare the case when the values of the exponential function are already precomputed (twc) or not (tnc) according to A.5.3. We can observe that the running time depends on overall support approximation and precomputation is a very effective way to shorten the running time.

	M	$\tilde{q}_{DNA}$	$\tilde{q}_{DNA.P_1}$	$\tilde{q}_{DNA.P_2}$	$\tilde{q}_{wsMCM}$	$\tilde{q}_{jMCM}$	$\tilde{q}_{MM}$
$\|\|\epsilon\|\|_V$	3	37	19	25	31	34	36
	4	11	7.3	11	22	12	13
	5	10	5.7	8.3	21	11	12
	6	6.6	3.9	6.2	5.8	5.6	7.4
	7	6.5	2.8	4.8	3.4	5.0	18

Table 5.10: ↩ The approximation error for the reconstruction of two-dimensional distributions of proteins P_1 and P_2 in exclusive switch model.

Cooperative self-activation of the gene expression. We consider another biological model describing the dynamics of a single gene inducing its own expression.

Biological model 5.1 (Cooperative self-activation of gene expression)**.** We consider the biological system where protein P can bind to the two independent promotor sites therefore inducing the activation the gene G (reactions I). As a result, there are the three gene states G, G^* and G^{**}, corresponding to zero, one or two activators bound. Result of the translation is denoted by M and can occur via one of the three reactions II. The translation process resulting in protein formation is described by the last three reactions III.

$$G + P \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} G^*, \quad G^* + P \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} G^{**}, \tag{I}$$

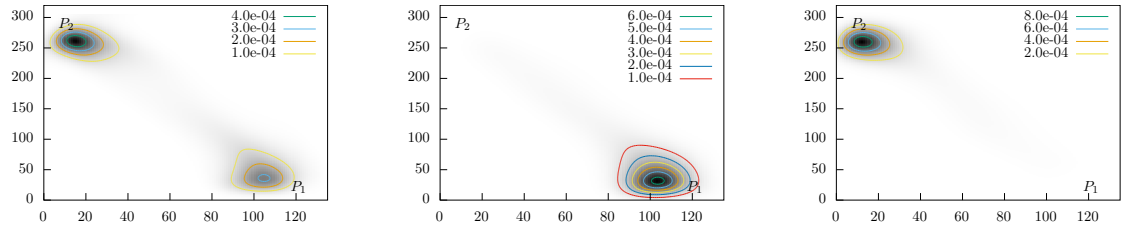
$$G \xrightarrow{k_G} G + M, \quad G^* \xrightarrow{k_{G^*}} G^* + M, \quad G^{**} \xrightarrow{k_{G^{**}}} G^{**} + M, \tag{II}$$

$$M \xrightarrow{k_3} M + P, \quad M \xrightarrow{k_4} \varnothing, \quad P \xrightarrow{k_5} \varnothing, \tag{III}$$

We consider two sets of parameters leading to moderate (k_A) and low (k_B) protein population counts such that $k_i = (k_1, k_{-1}, k_2, k_{-2}, k_G, k_{G^*}, k_{G^{**}}, k_3, k_4, k_5)$ for $i \in \{A, B\}$. Parameter sets are given by $k_A = (5 \cdot 10^{-4}, 3 \cdot 10^{-3}, 5 \cdot 10^{-4}, 2.5 \cdot 10^{-2}, 4, 12, 24, 1200, 300, 1)$ and $k_B = (5{\cdot}10^{-4}, 2{\cdot}10^{-4}, 5{\cdot}10^{-4}, 2{\cdot}10^{-3}, 4, 60, 160, 30, 300, 1)$.

As before, the values of the conditional moments $E\big(X_P^k | G = 1\big)$, $E\big(X_P^k | G^* = 1\big)$ and $E\big(X_P^k | G^{**} = 1\big)$ of P are approximated using MCM and used as constraints in the maximum entropy reconstruction. The marginal distribution of P is reconstructed by applying the law of total probability $\tilde{q}(x) = \sum_{i=\{1,2,3\}} p_i \cdot \tilde{q}_i(x)$, where $p_1 = \mathbf{P}[G = 1]$, $p_2 = \mathbf{P}[G^* = 1]$ and $p_3 = \mathbf{P}[G^{**} = 1]$ and $\tilde{q}_i(x)$ are the corresponding conditional distributions.

The results of the reconstruction are listed in Table 5.11 where P^{k_A} (P^{k_B}) denotes the distribution of proteins when the set of parameters k_A (k_B) is used. For the case of moderate protein populations (parameter set k_A) it is more reasonable to apply *weighted sum MCM*: the reconstruction even with $M = 2$ is more accurate that when *joint MCM*



(a) (b) (c)

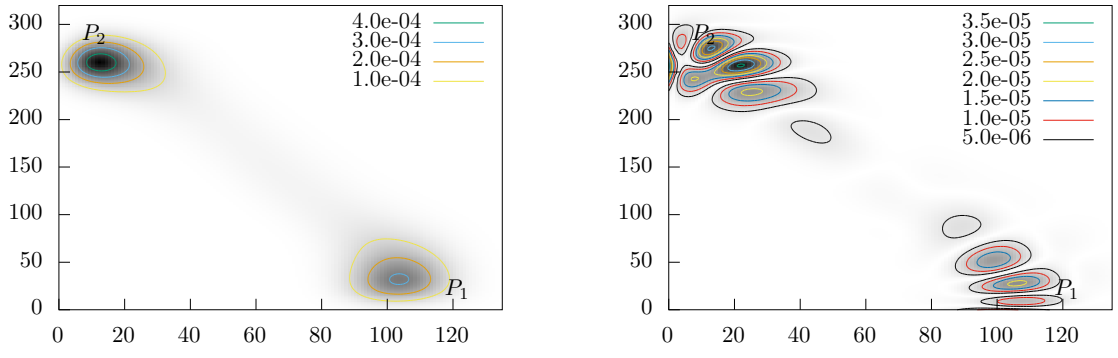
Figure 5.12: ↩ Exclusive switch: the approximations of the conditional distributions of proteins P_1 and P_2 where $DNA = 1$ (a), $DNA.P_1 = 1$ (b) and $DNA.P_2 = 1$ (c) . The reconstructions are obtained using $M = 7$.

is applied with $M = 7$. In the case of low protein populations (parameter set k_B) it is also obvious that application of *weighted sum MCM* provides much better results. Please note that in this case we do not apply MM method to approximate the moments since the corresponding ODE system is very stiff and takes extremely long time to converge to a steady state. To solve the system of conditional moments, we use `ode15s` solver from `MATLAB` with the default accuracy settings. We consider the system at quasi-steady state, therefore we integrate the ODE system until the moment values converge and do not change much.

We plot the reconstructions for both P^{k_A} (P^{k_B}) in Figure 5.16 where we use $M = 3$ and $M = 7$. As we can see, the multimodality of the distribution is well captured and the approximation is accurate when $M = 7$.

Bursty protein production model. The proposed method can be also applied to bursty protein production model, however this computation has not been implemented. The model possesses the countably infinite number of modes (where each mode corresponds to the number of species M molecules), whereas for systems considered before the number of modes is finite. This can be mitigated if we select only those M counts that provide the mode probability larger than a certain threshold, i.e., set of modes $\tilde{M} = \{M | p_M > \epsilon_{\text{mode}}\}$, where $p_M = \sum_{x_P \in [0,\infty]} \mathbf{P}[(X_M, X_P) = (x_M, x_P)]$ and $0 < \epsilon_{\text{mode}} \ll 1$. Then the distribution of the protein P can be approximated using *weighted sum MCM* as $\tilde{q}_P(x_P) = \sum_{x_M \in \tilde{M}} p_{x_M} \tilde{q}(x_M, x_P)$.

For all considered experiments we apply the maximum entropy principle using the moments approximated using both MM and MCM and found that the integration of the moment equations and the reconstruction of the distributions is several orders of magnitude faster than a direct numerical integration of the CME. By increasing the population sizes,

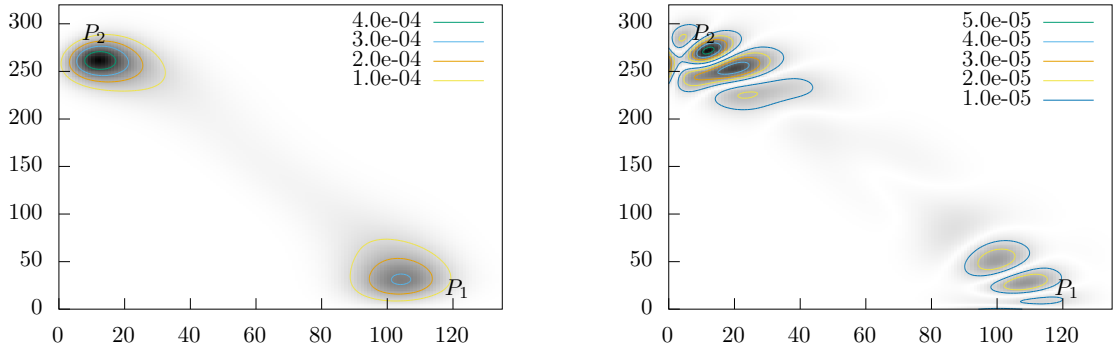


(a) (b)

Figure 5.13: ↩ Exclusive switch: the reconstruction of the marginal distribution of proteins P_1 and P_2 obtained using *weighted sum MCM* with $M = 7$ moments (a) and the corresponding approximation error (b).

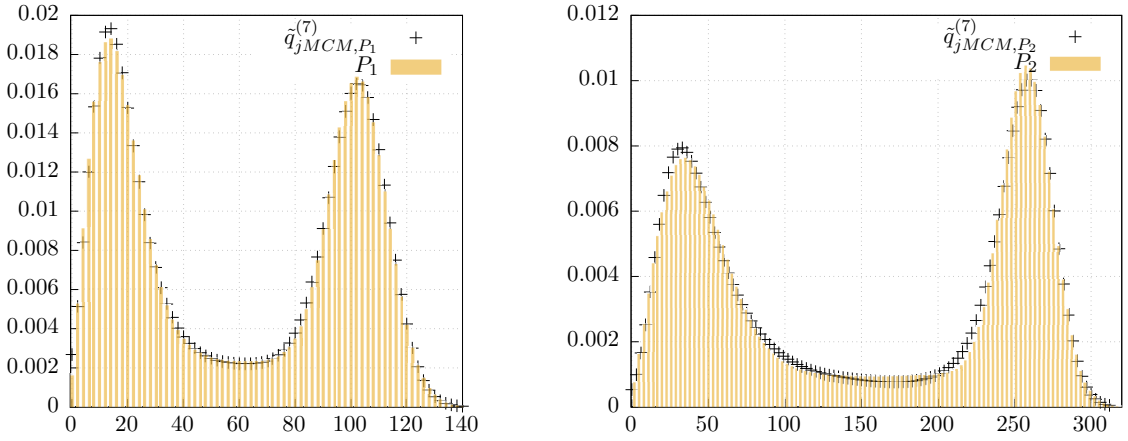
the speed-up can be made arbitrarily large. However, the accuracy of the reconstructed distributions is not always satisfying and simply increasing the order of the considered moments, from a certain order on, does not improve the results. This is due to the fact that the optimization procedure becomes numerically unstable. In particular, if the distributions have a complex shape (such as multimodal distributions or distributions with high probability mass concentrated in states with small population counts) the reconstruction may become inaccurate.

We propose to reconstruct the *conditional* probability distributions that result from the MCM approach and multiply with the probability of the condition (i.e., marginal distributions of small populations) to obtain the full distribution. This has several advantages



(a) (b)

Figure 5.14: ↩ Exclusive switch: the reconstruction of the marginal distribution of proteins P_1 and P_2 obtained using *joint MCM* with $M = 7$ moments (a) and the corresponding approximation error (b).



(a) (b)

Figure 5.15: ↩ Exclusive switch: marginal distributions of proteins P_1 (a) and P_2 (b) calculated via marginalization of the *joint MCM* reconstruction of two-dimensional distribution with $M = 7$.

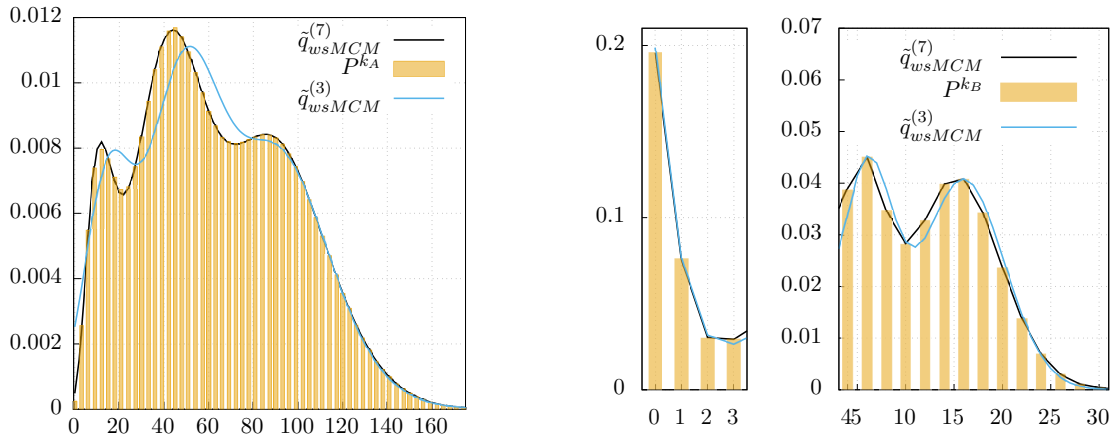
	M	$P^{k_A}_{wsMCM}$	$P^{k_A}_{jMCM}$	$P^{k_B}_{wsMCM}$	$P^{k_B}_{jMCM}$
$\|\|\epsilon\|\|_V$	2	4.7	7.4	27	25
	3	6.9	7.7	2.7	23
	4	1.2	5.7	1.4	10
	5	$8.7 \cdot 10^{-1}$	5.5	$4.2 \cdot 10^{-1}$	8.9
	6	$4.9 \cdot 10^{-1}$	5.1	$2.6 \cdot 10^{-1}$	8.6
	7	$3.2 \cdot 10^{-1}$	5.0	$1.4 \cdot 10^{-1}$	8.0

Table 5.11: ↩ Self-activating gene model: reconstruction of the marginal distribution of proteins P using *weighted sum MCM* and *joint MCM*.

compared to the MM reconstruction. First, in the MM approach the reconstruction of the probability distributions of small populations is usually very imprecise, while in the MCM approach these distributions are directly available. Second, the conditional distributions in the MCM are often less complex (e.g. unimodal instead of bimodal) and thus easier to reconstruct. Finally, as the maximal order M of considered moments increases, the MCM requires fewer variables (conditional moments and conditional probabilities) than in the MM approach. This is because the number of conditional probabilities is fixed while the number of considered (conditional) moments grows exponentially in M (recently, Ruess showed [187] that the number of equations in the moment closure system can be further reduced).

From a computational point of view, moment closure based reconstruction is limited by the order of moments that can be numerically integrated due to the stiffness of the ODE system. Such difficulties are encountered when the moments of order $M = 9$ are to be approximated.

Nevertheless, we find that the information provided by this smaller set of variables is as good or even better for reconstructing the original distributions. Thus, we argue that the



(a) (b)

Figure 5.16: ↩ Self-activating gene model: reconstruction of the marginal distribution of proteins P^{k_A} and P^{k_B} obtained using *weighted sum MCM* with $M = 3$ and $M = 7$.

MCM approach should be preferred when deriving the probability of certain events from the moment values.

5.3 Maximum Entropy and Parameter Estimation

5.3.1 Maximum Likelihood Based Approach

The experimental imaging techniques (such as fluorescent microscopy) are able to count the molecular populations in a living cell. If small populations are present, stochastic models can be used to describe the dynamics of the biological system using continuous-time Markov chain. However, this is possible only if the reaction rate constants c_j are known. Often it is not the case and c_j must be estimated based on time-series data of the molecular populations. This procedure is computationally intensive and require to solve the CME for each set of the parameter estimates. The corresponding maximum likelihood based method that allows to estimate rate constants, initial populations and parameters from the noisy population measurements is described in [13].

The method given in [13] is however limited to systems with relatively small state space and may suffer from the state space explosion. In order to mitigate this, we can apply the moment approximation methods such as MM or MCM and use the moment data to obtain the reaction rate constant estimates for systems with even high population counts. Here, we adopt the maximum likelihood method such that the moment approximation is used as an initial data for the maximum entropy reconstruction. The reconstructed distribution is then used to compute the likelihood of the given data.

Assume that $\mathbf{O} = \{O_1, \ldots, O_K\}$ is a sequence of observations taken at a certain time point T. Each observation $O_i = (O_{i,1}, \ldots, O_{i,N_S})$ is a vector of population counts for species where some species may not be observed. We also assume that the chemical reaction network is known and we need to estimate rate constants $c = (c_1, \ldots, c_{N_R})$. In order to proceed with estimation of the vector c, the maximum likelihood estimation (MLE) approach can be used where we maximize the likelihood of the given data $\mathbf{O}$:

$$(c_1, \ldots, c_{N_R}) = \arg \max_{c_1,\ldots,c_{N_R}} \mathcal{L}(\mathbf{O}),$$

where we omit the dependency of the likelihood function $\mathcal{L}$ on the reaction rate constants for clarity. We can use the values of moments in two ways to estimate the rate constants:

1. Use approximated moments and the moments computed from data and minimize the distance. The most recent application of this principle was proposed in [155] where the generalized method of moments (GMM) is used to approximate the moments of the biological model with complex dynamics.

2. Reconstruct the distribution from the moments and maximize the likelihood computed with respect to this reconstruction.

Here we apply the second approach where we use the maximum entropy method to reconstruction the unknown distribution. The first approach has been successfully applied before: the authors of [168] approximate the likelihood by a normal distribution for the cases where dynamics can be described well using first and second moments (i.e. using the expectations and the covariance matrix). However, in the presence of multimodality [246] such approximations become inaccurate. Several moment closure schemes are applied in [35] where Monte-Carlo simulations are combined with moment approximations and the likelihood is approximated by a normal distribution using the first four moments. Similar approach was applied before to solve the problem of the optimal experiment design in [189]. The maximization of likelihood (minimization of negative likelihood) can be done much more efficiently if the derivatives of the likelihood function w.r.t. reaction rate constants $\partial\mathcal{L}/\partial c_i$ are known [141]. Recently, Fröhlich et al. presented the method based on system size expansion [72]. The overview of the moment-based parameter inference techniques for biological systems can be found in [141, 188].

We assume that the likelihood function can be approximated using the maximum entropy principle. Therefore, the likelihood of data $\mathbf{O}$ (without noise) is given by

$$\mathcal{L}(\mathbf{O}) = \prod_{O_k \in \mathbf{O}} \mathbf{P}[\vec{X}(t) = O_k] \approx \prod_{O_k \in \mathbf{O}} \tilde{q}(O_k, \lambda^{*,k}),$$

where $\tilde{q}$ is the maximum entropy reconstruction. To simplify the numerical maximization procedure, we apply the logarithm to obtain the log-likelihood of the data

$$\ln \mathcal{L}(\mathbf{O}) = \sum_{O_k \in \mathbf{O}} \ln \mathbf{P}[\vec{X}(t) = O_k] \approx \sum_{O_k \in \mathbf{O}} \ln \tilde{q}(O_k, \lambda^{*,k}).$$

This function is however nonlinear and may possess several local optima, therefore the global optimization algorithm shall be applied. If available, the gradient vector (and the Hessian matrix) may be provided to the local gradient-based minimizer in order speed up the numerical optimization procedure [13]

$$\frac{\partial \ln \mathcal{L}}{\partial c_i}(\mathbf{O}) = \sum_{O_k \in \mathbf{O}} \frac{\partial}{\partial c_i} \ln \mathbf{P}[\vec{X}(t) = O_k] = \sum_{O_k \in \mathbf{O}} \frac{\frac{\partial}{\partial c_i}\mathbf{P}[\vec{X}(t) = O_k]}{\mathbf{P}[\vec{X}(t) = O_k]},$$

$$\frac{\partial}{\partial c_i}\mathbf{P}[\vec{X}(t) = O_k] \approx \frac{\partial}{\partial c_i}\tilde{q}(O_k, \lambda^{*,k}) = \frac{\partial}{\partial c_i} \exp\left(-\sum_{l=0}^{M} \lambda_l^{*,k} O_k^l\right),$$

$$\frac{\partial}{\partial c_i} \exp\left(-\sum_{l=0}^{M} \lambda_l^{*,k} O_k^l\right) = \exp\left(-\sum_{l=0}^{M} \lambda_l^{*,k} O_k^l\right) \cdot \left(-\sum_{l=0}^{M} \left[\frac{\partial \lambda_l^{*,k}}{\partial c_i}\right] O_k^l\right),$$

where in the second equation we do not put restrictions on the dimensionality of the observation O_k, i.e., the sum $\left(-\sum_{l=0}^{M} \lambda_l^{*,k} O_k^l\right)$ can be considered as in (4.7) or in (4.13) and can be generalized to any number of species N_S. Here, we assume that the observation is one-dimensional ($N_S = 1$) for the sake of readability.

Please note that it is impossible to obtain the derivatives $\frac{\partial \lambda_l^*}{\partial c_i}$ in a closed form since the coefficients λ are the result of the optimization procedure which is not described in

terms of functional dependency between the set of approximated moments μ_l and λ_l (the corresponding derivation can be, however, conducted for the case of generalized exponential distributions, cf. Appendix A.5.7). Alternatively, we can again apply the principle of maximum entropy and try to reconstruct the function $\partial/\partial c_i \mathbf{P}[\vec{X}(t) = O_k]$ (which is not a distribution, so it does not possess the corresponding properties). Therefore, we need to approximate the derivatives of the moments with respect to reaction rate constants $\partial\mu_l/\partial c_i$ and consider the following moment problem

$$\begin{aligned} \frac{\partial}{\partial c_i} \sum_{x \in S} x^l p(x) &= \frac{\partial \mu_l}{\partial c_i}, \\ \sum_{x \in S} x^l r(x) &= \nu_{l,i}, \end{aligned} \tag{5.1}$$

where we denote $\frac{\partial p(x)}{\partial c_i} = r_i(x)$ and $\frac{\partial \mu_l}{\partial c_i} = \nu_{l,i}$. The resulting maximum entropy is then

$$\sum_{x \in S} x^l r_i(x) = \nu_{l,i} \quad 1 \leq l \leq M < \infty, \quad 1 \leq i \leq N_R.$$

In order to obtain the values $\nu_{l,i}$, the moment closure system has to be extended to include the derivatives of moments w.r.t. rate constants as well. Instead of considering only the dynamics of moments

$$\frac{d}{dt}\mu^{(\mathbf{I}_l)} = \mathbf{F}(\mu^{(\mathbf{I}_1)}, \mu^{(\mathbf{I}_2)}, \ldots; t), \tag{5.2}$$

the dynamics of the derivatives shall also be included into the ODE system

$$\frac{d}{dt}\left(\frac{\partial}{\partial c_i}\mu^{(\mathbf{I}_l)}\right) = \mathbf{G}\left(\mu^{(\mathbf{I}_1)}, \mu^{(\mathbf{I}_2)}, \ldots, \frac{\partial \mu^{(\mathbf{I}_1)}}{\partial c_i}, \frac{\partial \mu^{(\mathbf{I}_2)}}{\partial c_i}, \ldots; t\right) \tag{5.3}$$

for all order vectors $\mathbf{I}$ and rate constants c_i. Obviously, both ODE systems $\mathbf{F}$ and $\mathbf{G}$ can be considered as functions of reaction rate constants c_i. Consider the following example where we extend the moment closure system for simple dimerization model:

Example 5.4 (Extended Moment Closure for Simple Dimerization Model)**.** ↩ We consider the dynamics of the derivative of covariance $C_{2,2}$ in Equation 3.2 w.r.t. to rate constants. Recall that the time of derivative $C_{2,2}$ is given by (where we assume the low dispersion closure of the order 2)

$$\frac{d}{dt}C_{2,2} = \frac{3}{2}c_1 C_{1,1} + \frac{1}{2}c_1\mu_1(\mu_1 - 1) + \mu_2(c_2 - c_1 C_{1,1})$$

Now we extend it by

$$\begin{aligned} \tfrac{d}{dt}\left(\tfrac{\partial C_{2,2}}{\partial c_1}\right) &= \tfrac{3}{2}\left(C_{1,1} + C_{1,1}^{(c_1)}\right) + \tfrac{1}{2}\left(\mu_1^2 + 2C_1\mu_1\mu_1^{(c_1)} - \mu_1 - c_1\mu_1^{(c_1)}\right) + \mu_2 C_{1,1} \\ &\quad -c_1\left(\mu_2 C_{1,1}^{(c_1)} + \mu_2^{(c_1)} C_{1,1}\right) \\ \tfrac{d}{dt}\left(\tfrac{\partial C_{2,2}}{\partial c_2}\right) &= \mu_2 + c_2\mu_2^{(c_2)}, \end{aligned}$$

where we denote the derivatives of moments w.r.t. to c_i by $\mu^{(c_i)}$ and $C^{(c_i)}$.

Both ODE systems **F** and **G** are coupled and need to be solved simultaneously w.r.t. time t.

However, the moment problem (5.1) does not allow for a direct solution. The classical moment problem is formulated as

$$\mu_l = \int_{-\infty}^{\infty} x^l \, \mathrm{d}\rho(x),$$

where $\rho(x)$ is a measure on $\mathbb{R}$. It is impossible to formulate the problem (5.1) in the same way since there is no reason for the derivative function $\partial/\partial_{c_i} p(x) = r_i(x)$ to be non-negative and satisfy the properties of the measure. Here, we use the numerical approximation of the derivatives instead and $\partial/\partial_{c_i}\mathcal{L}$ is calculated using the finite differences

$$\nabla\mathcal{L} \approx \left[\frac{\mathcal{L}(c_1 + \delta_{c_1}) - \mathcal{L}(c_1 - \delta_{c_1})}{2\delta_{c_1}}, \ldots, \frac{\mathcal{L}(c_{N_R} + \delta_{c_{N_R}}) - \mathcal{L}(c_{N_R} - \delta_{c_{N_R}})}{2\delta_{c_{N_R}}} \right].$$

The absence of the closed-form solution for the gradient and the Hessian of the likelihood function limits the ability to estimate the quality of the obtained estimates, i.e., there is no direct way to get the estimate of the variance of the inferred rate constant. Again, here we use the numerical approximation of the Hessian matrix $\hat{H}$ and the covariance matrix is given by $\hat{\Sigma} = \hat{H}^{-1}$ [63]. The standard deviation of ith parameter estimation can be computed as $\hat{\sigma}_i = \sqrt{\hat{\Sigma}_{i,i}} = \sqrt{\hat{H}^{-1}_{i,i}}$.

Example 5.5 (Parameter Estimation for Exclusive Switch)**.** ↩ We consider the problem of reaction rate constants estimation for the exclusive switch system given several sample data sets **O**, each consisting of 10000 samples. The aim is to estimate the reaction rate constants c_1 and c_2 whereas the rest are fixed. We assume that only protein counts P_1 and P_2 are observable. The log-likelihood of the data **O** is given by

$$\ln \mathcal{L}(\mathbf{O}) = \sum_{O_k \in \mathbf{O}} \ln \mathbf{P}[\vec{X}(t) = O_k] \approx \sum_{O_k \in \mathbf{O}} \ln \tilde{q}(O_k, \lambda^*),$$

where $\tilde{q}(O_k, \lambda^*)$ corresponds to the maximum entropy reconstruction of the marginal distribution of P_1 and P_2. The initial parameter guess can be obtained using one of the moment-matching parameter inference methods [35, 168, 189]. Here we use the result of GMM-based parameter estimation [155] to bound the search space of the parameters as $\bar{c}_i \pm z_\alpha \hat{\sigma}_{c_i}$, where $\bar{c}_i$ is the expectation of the estimator for c_i, $\hat{\sigma}_{c_i}$ is the standard deviation and z_α is the z-value corresponding to the fixed level of confidence α. Unfortunately, the results of the numerical experiments reveal that the parameter estimates does not become closer to its true values, therefore it is better to stick to the original estimates from GMM. One may try to get the better estimates by running

the parameter inference procedure described in [13] if the size of the state space is manageable.

Another way to address the problem of the derivative approximation is to apply the moment recursive relations of the generalized exponential family. For instance, the vector of parameters β of the density $f_G(\beta)$ of type G with the support $[0, \infty]$ can be obtained by simply solving a linear system of equations $\beta = M^{-1}\alpha$, where both vector α and matrix M are the functions of non-central moments μ_l (cf. Equation A.13). The parameter vector β is then a function of moments μ_l, i.e., $\beta = M^{-1}\alpha$, therefore we can compute derivatives $\partial/\partial c_i \beta = \partial/\partial c_i \left(M^{-1}\alpha\right)$ for all $i \in 1, \ldots, N_R$. The derivatives $\partial/\partial c_i \mathbf{P}[\vec{X}(t) = O_k]$ can be approximated by $\partial/\partial c_i f_G(O_k, \beta^*)$. Both the vector of parameters β^* is computed via $\beta^* = \tilde{M}^{-1}\tilde{\alpha}$, where both the matrix $\tilde{M}$ and the vector $\tilde{\alpha}$ are computed using the moments obtained with MM or MCM method extended with ODE system $\mathbf{G}$ (5.3) that tracks the dynamics of moment derivatives.

For the distribution $f(x, \beta)$ of the generalized exponential family the derivatives $\partial/\partial c_i f(O_k, \beta^*)$ can be computed as

$$\frac{\partial}{\partial c_i} f(O_k, \beta^*) = \frac{\partial}{\partial \beta} f(O_k, \beta^*) \cdot \frac{\partial}{\partial c_i} \beta(\vec{\mu}) = \frac{\partial}{\partial \beta} f(O_k, \beta^*) \cdot \frac{\partial}{\partial c_i} \left(M^{-1}(\vec{\mu})\ \alpha(\vec{\mu})\right)$$
$$\frac{\partial}{\partial c_i} \left(M^{-1}(\vec{\mu})\ \alpha(\vec{\mu})\right) = \frac{\partial}{\partial c_i}\vec{\mu}(c) \cdot \left(\alpha(\vec{\mu})\frac{\partial}{\partial \vec{\mu}} M^{-1}(\vec{\mu}) + M^{-1}(\vec{\mu})\frac{\partial}{\partial \vec{\mu}}\alpha(\vec{\mu})\right),$$

where $\frac{\partial}{\partial c_i}\vec{\mu}(c) = \mathbf{G}^{(c_i)}\left(\mu^{(\mathbf{I}_1)}, \mu^{(\mathbf{I}_2)}, \ldots, \frac{\partial}{\partial c_i}\mu^{(\mathbf{I}_1)}, \partial/\partial c_i \mu^{(\mathbf{I}_2)}, \ldots\right)$ is the solution of the extended moment closure system (5.3).

Similarly, we can further extend this approach to obtain the Hessian matrix of the likelihood function as a closed form expression. In order to do so, we need the approximation of the second derivatives of the moments. The moment closure system $(\mathbf{F}, \mathbf{G})$ (5.2)-(5.3) can be extended with the equation system $\mathbf{H}$ defined as

$$\frac{d}{dt}\left(\frac{\partial^2}{\partial c_i \partial c_j}\mu^{(\mathbf{I}_l)}\right) = \mathbf{H}\left(\mu^{(\mathbf{I}_1)}, \mu^{(\mathbf{I}_2)}, \ldots, \frac{\partial \mu^{(\mathbf{I}_1)}}{\partial c_i}, \frac{\partial \mu^{(\mathbf{I}_2)}}{\partial c_i}, \ldots \frac{\partial^2 \mu^{(\mathbf{I}_1)}}{\partial c_i \partial c_j}, \frac{\partial^2 \mu^{(\mathbf{I}_2)}}{\partial c_i \partial c_j}, \ldots; t\right) \tag{5.4}$$

and the second derivatives of the distribution reconstruction can be computed as

$$\frac{\partial^2}{\partial c_i \partial c_j} f(O_k, \beta^*) = \frac{\partial^2}{\partial \beta^2} f(O_k, \beta^*) \cdot \frac{\partial}{\partial c_i}\beta(\vec{\mu}) \cdot \frac{\partial}{\partial c_j}\beta(\vec{\mu}) + \frac{\partial}{\partial \beta} f(O_k, \beta^*) \cdot \frac{\partial^2}{\partial c_i \partial c_j}\beta(\vec{\mu}),$$

where the elements of the Hessian matrix come into the second derivative of the function $\beta(\vec{\mu}) = M^{-1}(\vec{\mu})\alpha(\vec{\mu})$. The second derivative of the likelihood function is given by

$$\frac{\partial^2 \ln \mathcal{L}(\mathbf{O})}{\partial c_i \partial c_j} = \sum_{O_k \in \mathbf{O}} \frac{\frac{\partial^2}{\partial c_i \partial c_j} p_k p_k - \frac{\partial}{\partial c_i} p_k \frac{\partial}{\partial c_j} p_k}{p_k^2}, \tag{5.5}$$

where we denote $\mathbf{P}[\vec{X}(t) = O_k]$ by p_k.

The given method can be generalized to noisy observations as well [13]. For that we assume that each observation $O_k \in \mathbf{O}$ can be represented as $O_k = X_k + \epsilon_k$, where the error terms ϵ_k are independent and identically normally distributed with zero mean and standard deviation σ. In this case, the log-likelihood of the data is given by

$$\begin{aligned}
\ln \mathcal{L}(\mathbf{O}) &= \sum_{O_k \in \mathbf{O}} \ln \bar{\mathcal{L}}\left(O_k\right), \\
\bar{\mathcal{L}}\left(O_k\right) &= \sum_{x \in D} f_\epsilon \left(O_k | \vec{X} = x\right) \mathbf{P}[\vec{X} = x] \\
\bar{\mathcal{L}}\left(O_k\right) &\approx \sum_{x \in D^*} f_\epsilon \left(O_k | \vec{X} = x\right) f(x, \beta^*) \\
f_\epsilon \left(O_k | \vec{X} = x_k\right) &= \phi_\sigma \left(O_k - x_k\right) = \phi_\sigma \left(\Delta_k\right),
\end{aligned}$$

where $\phi_\sigma(x)$ denotes the probability density function of the normal distribution with standard deviation σ. We can derive the first and second derivative in a similar fashion as before

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}}{\partial c_i} &= \sum_{O_k \in \mathbf{O}} \frac{\partial}{\partial c_i} \ln \bar{\mathcal{L}}(O_k) = \sum_{O_k \in \mathbf{O}} \frac{\frac{\partial}{\partial c_i} \bar{\mathcal{L}}(O_k)}{\bar{\mathcal{L}}(O_k)}, \\
\frac{\partial}{\partial c_i} \bar{\mathcal{L}}(O_k) &\approx \sum_{x \in D^*} \phi_\sigma \left(\Delta_k\right) \frac{\partial}{\partial c_i} f(x, \beta^*), \\
\frac{\partial^2 \ln \mathcal{L}}{\partial c_i \partial c_j} &= \sum_{O_k \in \mathbf{O}} \frac{\frac{\partial^2}{\partial c_i \partial c_j} \bar{\mathcal{L}}(O_k) \bar{\mathcal{L}}(O_k) - \frac{\partial}{\partial c_i} \bar{\mathcal{L}}(O_k) \frac{\partial}{\partial c_j} \bar{\mathcal{L}}(O_k)}{\bar{\mathcal{L}}^2(O_k)}, \\
\frac{\partial^2}{\partial c_i \partial c_j} \bar{\mathcal{L}}(O_k) &\approx \sum_{x \in D^*} \phi_\sigma \left(\Delta_k\right) \frac{\partial^2}{\partial c_i \partial c_j} f(x, \beta^*),
\end{aligned}$$

where we need to solve the system $(\mathbf{F}, \mathbf{G}, \mathbf{H})$ defined by (5.2),(5.3),(5.4). The weighting with the normal distribution over the approximated state space D^* does not have to be complete: it might well be the case that values of the normal distribution $\phi_\sigma(\Delta_k)$ become negligible providing additional truncation bound and therefore shortening the number of terms in sum.

The proposed algorithm is given in Algorithm 1, where we assume that all reaction rate constants are to be estimated (we denote the coupled system ODE system for moment closure by $(\mathbf{F}, \mathbf{G}, \mathbf{H})$). The function **ReconstructDistribution** may refer to any kind of distribution reconstruction technique such as maximum entropy or generalized exponential family based. The function **UpdateParameters** denotes any kind of gradient-based search method such as Newton [196], Levenberg-Marquardt [174] or limited memory BFGS [42]. Some of those methods can also make use of the Hessian matrix of second derivatives to speed up the convergence to the optimum. If closed form expressions are not available (as in case of maximum entropy reconstruction), the corresponding finite difference approximations may be used.

The third derivatives of the likelihood function can be also analytically derived. However, the corresponding optimization methods [11, 33, 52, 125] are not widely applied. In case of maximum entropy reconstruction, the application of third derivatives tensor to the minimization of dual function (implemented in [192]) does not give any gain in terms of convergence speed.

5.3.2 Maximum Relative Entropy

The alternative approach to the parameter estimation procedure is based on the usage of the maximum relative entropy (MrE) method designed in order to incorporate two types of constraints: the *moment constraints* of type (4.2) and *data constraints*. The method is explained in details in [78–80] and often applied in physics and economics [21, 240].

The authors emphasize that the MrE procedure is fully compatible with Bayesian inference method and one exactly follows the Bayesian inference in the absence of moment constrains. Special attention is made on how and in which order to apply the data constraints (observation data) and moment constraints. In the scope of chemical reaction network analysis this method can be helpful to estimate the parameters in situation when the excessive simulation is not possible but moments of the species quantities may be obtained reasonably fast.

Chapter 6

Conclusion

The realistic biological systems that are subject to the stochastic effects may reach over the bounds of the discrete state space that can be handled reasonably well by current implementations of chemical master equation solvers. The state space explosion currently is (and will) remain one of the main obstacles to cope with such systems in purely stochastic and most accurate way. This motivates the need to use moment-based approaches that are capable of describing the same behaviour in more condensed form.

This doctoral thesis is focused on how to recover the discrete-state representation of the system given the moment description. This is the well known inverse moment problem for which the mathematical foundations are rigorously investigated. However, this is one of the first approaches to apply these methods to the systems with (possibly unbounded) discrete state space and distributions of molecule numbers with relatively complicated shape (successfully decomposed onto simpler ones using the conditional representation).

Moment-based representation of the behaviour allows for the fast computation of main cumulative measures but it may suffer from the stiffness problems, that is, the asymmetry in the system evolution caused by large difference in the number of molecules of different species. Recently introduced method of conditional moments allows to mitigate this problem by decomposing the system onto several interconnected subsystems with respect to the number of modes. Each mode corresponds to the configuration of slowly-changing species in the system. In such way, the hybrid description is introduced, where fast-changing parts are described by moments and slow-changing parts are treated in the stochastic fashion, by a chemical master equation.

We applied the maximum entropy method to solve the inverse moment problem when both moment-based approaches are used: the pure moment closure and the method of conditional moments. The comparison reveals that it is beneficial to use the conditional moments in terms of accuracy and computation time. This maximum entropy based reconstruction technique may also be applied to estimate the unknown parameters of the systems. However, due to the instability of the optimization procedure and high sensitivity to the noise, estimation of the reaction rate constants from the experimental data reveals to be not realistic.

Future Work. The two major problems of the presented technique are the sensitivity of the optimization procedure to the noise and dimensionality limit (we applied it to one- and two-dimensional distribution). To mitigate the first problem, other preconditioning techniques may applied, mainly working with transformation of the moment space. Being able to compensate for the instability, the second problem is technically solvable (as shown by Abramov [3] and recently by Hao [104]). Solving both problems would allow for the fast change of the state space representation: while the population number of the certain species stays low, its dynamics is controlled by the chemical master equation. As soon as it grows over a certain threshold, its representation can be changed to the moment-based one and the other way around if significantly decreases. It would also mitigate the problems arising while applying the maximum entropy reconstruction to the parameter estimation task.

Chapter 7

Algorithms

Here, we list all algorithms introduced in the course of this thesis.

Algorithm 1: ↩Parameter estimation using distribution reconstruction

---

Input: $\mathbf{O}$: observation samples, $R_1, \ldots, R_{N_R}$: chemical reaction network, $(c_1^{(0)}, \ldots, c_{N_R}^{(0)})$: initial approximation for the search procedure, C: the search space for parameters, $C \subset \mathbb{R}^{N_R}$, $\mathbf{F}$: moment closure system for the moment approximation, $\mathbf{G}$: moment closure system for the moment derivatives approximation, $\vec{\mu}_{t=0}$: initial values of moments at time $t = 0$, M: truncation order for the moment closure, T: time instant at which samples $\mathbf{O}$ are obtained

Output: $\hat{c} = (\hat{c}_1, \ldots, \hat{c}_{N_R})$: estimation of reaction rate constants

$\iota := 0$;

do

Solve moment closure ODE system:

$$[\vec{\mu}, \nabla_{\vec{\mu}}, \mathcal{H}_{\vec{\mu}}] := \int_0^T (\mathbf{F}, \mathbf{G}, \mathbf{H})\,(c_1^{(\iota)}, \ldots, c_{N_R}^{(\iota)}; \vec{\mu}_{t=0}; t)\,\mathrm{d}t \ ;$$

Reconstruct the distribution: $f(x, \beta) := \mathit{ReconstructDistribution}(\vec{\mu})$;

Compute first derivatives of the reconstruction: $f'_{c_i}(x, \beta) = \frac{\partial}{\partial c_i} f(x, \beta)$;

Compute second derivatives of the reconstruction: $f''_{c_i, c_j}(x, \beta) = \frac{\partial^2}{\partial c_i \partial c_j} f(x, \beta)$;

Compute the log-likelihood: $l^{(\iota)} := \sum_{O_k \in \mathbf{O}} \ln f(O_k, \beta)$;

Compute first derivatives of the log-likelihood:

$$\frac{\partial}{\partial c_i} l^{(\iota)} = \sum_{O_k \in \mathbf{O}} \frac{f'_{c_i}(O_k, \beta)}{f(O_k, \beta)}, \qquad \nabla_{l^{(\iota)}} = \left(\frac{\partial}{\partial c_1} l^{(\iota)}, \ldots, \frac{\partial}{\partial c_{N_R}} l^{(\iota)} \right) \ ;$$

Compute second derivatives of the log-likelihood:

$$\frac{\partial^2}{\partial c_i \partial c_j} l^{(\iota)} = \sum_{O_k \in \mathbf{O}} \frac{f''_{c_i, c_j}(O_k, \beta) f(O_k, \beta) - f'_{c_i}(O_k, \beta) f'_{c_j}(O_k, \beta)}{f(O_k, \beta)}, \qquad \left(\mathcal{H}_{l^{(i)}}\right)_{i,j} = \frac{\partial^2}{\partial c_i \partial c_j} l^{(\iota)} \ ;$$

$\iota := \iota + 1$;

Get the new estimate of parameters:

$$(c_1^{(\iota)}, \ldots, c_{N_R}^{(\iota)}) := \mathit{UpdateParameters}\ (l^{(i)}, \nabla_{l^{(i)}}, \mathcal{H}_{l^{(i)}}) \ ;$$

while *maximum likelihood is not found*;

Algorithm 2: ↩ Main algorithm of distribution reconstruction (one-dimensional case), *ExternalMaxent1D*

Input: $\vec{\mu}$: set of moment constraints, M: number of moment constraints, $\mho_{1D}$: optimization parameters

Output: (λ^*, D^*): solution vector of parameters and the corresponding support

Choose number of moments currently used for the reconstruction

$\dot{M} = M$

do

  Choose set of moments currently used for the reconstruction

  $\vec{\mu}_\cdot = \{\mu_0, \mu_1, \ldots, \mu_{\dot{M}}\}$

  Get the initial approximation of the support (Section 4.3.3)

  $\{x_L, \ldots, x_R\} =$ *ComputeSupportApproximation* $(\vec{\mu}_\cdot)$

  Run the adaptive support minimization of the dual function

  $(\lambda^*, D^*, \flat) =$ *AdaptiveSupportOptimization* $(\vec{\mu}_\cdot, \{x_L, \ldots, x_R\}, \mho_{1D})$

  if $\flat \neq$ ***true*** **then**

    $\dot{M} = \dot{M} - 1$

    $\mho_{1D}.\delta_d = 2 \cdot \mho_{1D}.\delta_d$

  else

    break

  end

while $\dot{M} > 1$;

Algorithm 3: ↩ Main algorithm of distribution reconstruction (two-dimensional case)

Input: $\vec{\mu}$: set of moment constraints, M: the highest order of available moment constraints, $\mho_{1D}$: optimization parameters for one-dimensional reconstructions, $\mho_{2D}$: optimization parameters for two-dimensional reconstruction,

Output: (λ^*, D^*_{xy}): solution vector of parameters and the corresponding support

Choose the highest order of moments currently used for the reconstruction

$\dot{M} = M$

Reconstruct the distribution for both species

$(\lambda^*_x, D^*_x) =$ *ExternalMaxent1D* $(\vec{\mu}_x, \mho_{1D})$

$(\lambda^*_y, D^*_y) =$ *ExternalMaxent1D* $(\vec{\mu}_y, \mho_{1D})$

Get the initial approximation for the two-dimensional support

$\bar{D}_x =$ *ComputeMainProbMassRegion* $(D^*_x, \mho_{1D}.\delta_F)$

$\bar{D}_y =$ *ComputeMainProbMassRegion* $(D^*_y, \mho_{1D}.\delta_F)$

$\bar{D}_{xy} = \bar{D}_x \times \bar{D}_y$

do

  Choose set of moments currently used for the reconstruction

  $\vec{\mu}_\cdot = \{\mu_{0,0}, \ldots, \mu_{i,j}\}$, $i + j \leq \dot{M}$

  Run the adaptive support minimization of the dual function

  $\left(\lambda^*, D^*_{xy}, \flat\right) =$ *AdaptiveSupportOptimization* $(\vec{\mu}_\cdot, \bar{D}_{xy}, \mho_{2D})$

  if $\flat \neq$ ***true*** **then**

    $\dot{M} = \dot{M} - 1$

  else

    break

  end

while $\dot{M} > 1$;

Algorithm 4: ↩ Dual function minimization with adaptive support approximation, *AdaptiveSupportOptimization*

Input: $\vec{\mu}$: set of moment constraints, D: distribution approximation, $\mho$: optimization parameters

Output: (λ^*, D^*): solution vector of parameters and the corresponding support, $\flat$: flag that shows whether the optimization procedure finished successfully

Initialize the number of support search iterations (ℓ)
$\ell = 0$
Initialize the number of iterations where minimization is unsuccessful ($\ell^{\triangledown}$)
$\ell^{\triangledown} = 0$
do
  $(\Psi^{\ddagger}, \lambda^{\ddagger}, \natural) =$ *Preconditioning and dual function minimization* $(\vec{\mu}, D)$
  if $\natural =$ ***true*** **then**
    $\flat = CheckSolutionValidity(\Psi, \tilde{q}(\lambda^{\ddagger}, D), D, \mho.\delta_{\bar{s}}, \mho.\delta_D, \mho.\delta_F, \mho.\bar{N}_{\texttt{peaks}})$
    if $\flat =$ ***true*** **then**
      if (4.11) *is satisfied with* $\mho.\delta_{\Psi}$ **then**
        return $(\lambda^{\ddagger}, D, \textbf{true})$
      else
        $\flat =$ **false**
      end
    end
  else
    $\ell^{\triangledown} = \ell^{\triangledown} + 1$
    if $\ell^{\triangledown} \geq \mho.\ell^{\triangledown}$ **then**
      return $(\emptyset, \emptyset, \textbf{false})$
    end
  end
  if $\ell < \mho.\ell$ **then**
    $\ell = \ell + 1$
  else
    return $(\lambda^{\ddagger}, D, \textbf{true})$
  end
while $\flat =$ ***false***;

Algorithm 5: ↩ Orthogonalization of basis in one-dimensional case

Input: D: approximation of the support on the current iteration, $\{\mu_0, \ldots, \mu_M\}$: set of moment constraints

Output: $\vec{\gamma}^{(0),\perp}$: initial approximation of the solution in orthogonalized basis, $P^{\perp}$: matrix with coefficients of orthogonalized basis, $\Psi(\vec{x}, \vec{\gamma})$: dual function in orthogonalized basis

Initial solution approximation

$\vec{\gamma}^{(0)} = (10^{-10}, 0, \ldots, 0)^T$;

Basis orthogonalization

$P = I$;

$P^{\perp} = ModifiedGramSchmidtWithReorthogonalization(D, P, \vec{\gamma}^{(0)}, P)$;

$\vec{\gamma}^{(0)} = (P^{\perp})^{-1} \cdot P \cdot \vec{\gamma}^{(0)}$;

Dual function in the new basis

$$\Psi(\vec{x}, \vec{\gamma}) = \ln \sum_x \exp\left(\sum_{k=1}^{K} \gamma_k v_k(\vec{x})\right) + \sum_{k=1}^{K} \gamma_k v_k(\vec{\mu});$$

with derivatives defined by $(\nabla\Psi)_k = Q_{\vec{\gamma}}(v_k) - v_k(\vec{\mu})$, $H_{kl} = Q_{\vec{\gamma}}(v_k v_l)$

Algorithm 6: ↩ Definition of distribution reconstruction validity

Input: Ψ: dual function on the current iteration, q: distribution reconstruction on the current iteration, D: approximation of the support on the current iteration, $\delta_{\bar{s}}$: threshold for $\bar{s}_q$ criterion, δ_D: dual function maximum to tail value threshold, δ_F: threshold for the cumulative probability mass, $\bar{N}_{\texttt{peaks}}$: maximum allowed number of peaks in the support region with main part of probability mass

Output: **true** for "valid" reconstruction, **false** for "invalid" reconstruction

Verify that the values of dual function are finite

if $\exists x \in D : \Psi(x) = \infty \vee \dfrac{\partial\Psi(x)}{\partial x} = \infty \vee \dfrac{\partial^2\Psi(x)}{\partial x^2} = \infty$ **then**

| **return false**

end

Verify that the values of distribution reconstruction are finite and (A.9) holds

if $\exists x \in D : q(x) = \infty \vee {}^{\max}{}_{x \in D^{(\ell)}} \Psi(\lambda^{(\ell)})(x) / \Psi(\lambda^{(\ell)})(x_R^{\ell}) > \delta_D$, **then**

| **return false**

end

$\bar{D} =$ *ComputeMainProbMassRegion* (D, δ_F)

Compute $\bar{s}_q$ for $\bar{D} = [x_L, x_R]$

$N_{\texttt{peaks}} =$ *ComputeNumberOfPeaks* $(\bar{D}, q)$

if $\bar{s}_q > \delta_{\bar{s}} \vee N_{peaks} > \bar{N}_{peaks}$ **then**

| **return false**

end

return true

Algorithm 7: ↩ *ComputeMainProbMassRegion*, computes the region where given amount of probability mass is located using the initial approximation of support

Input: q: distribution reconstruction on the current iteration, D: approximation of the support on the current iteration, δ_F: threshold for the cumulative probability mass,

Output: $\bar{D}$, the region of support where δ_F of probability mass is located

$(x_L, x_R) = D$

$x_L^* = x_L$

$x_R^* = \max\{x | \sum_{x_L}^{x_R^*} q(x) \leq \delta_F\}$

$\bar{D} = (x_L^*, x_R^*)$

return $\bar{D}$

Algorithm 8: ↩ *ComputeNumberOfPeaks*, computes number of peaks in the provided region of support

Input: q: distribution reconstruction on the current iteration, $\bar{D}$: approximation of the support on the current iteration

Output: $N_{\texttt{peaks}}$, number of peaks in $\bar{D}$

Compute $\frac{\partial q(t)}{\partial t}|_{t=x}$ for all $x \in \bar{D}$

Define the specific discrete sign function $sgn(x) = \begin{cases} 0, & \frac{\partial q(t)}{\partial t}|_{t=x} < 0 \\ 1, & \frac{\partial q(t)}{\partial t}|_{t=x} > 0 \end{cases}$ with the discrete derivative $sgn'_+(x) = sgn(x+1) - sgn(x)$

$N_{\texttt{peaks}} = \sum_{x \in \bar{D}} |sgn'_+(x)|$

return $N_{\texttt{peaks}}$

Algorithm 9: ↩ Preconditioning procedure

Input: $\{\mu_0, \ldots, \mu_M\}$: set of moment constraints, D: approximation of the support

Output: $\{\mu_{0,c,r}, \ldots, \mu_{M,c,r}\}$: transformed moments, $D_{c,r}$: transformed support, A: transformation matrix

Centralization

$\{\mu_{0,c}, \ldots, \mu_{M,c}\} = \mathit{ConvertRawToCentralMoments}\ (\{\mu_0, \ldots, \mu_M\})$;

$D_c = \mathit{CentralizeSupport}\ (D, \mu_1)$;

Rotation and scaling

$(\{\mu_{0,c,r}, \ldots, \mu_{M,c,r}\}, A) = \mathit{RescaleCentralMoments}\ (\{\mu_{0,c}, \ldots, \mu_{M,c}\})$;

$D_{c,r} = \mathit{RescaleCentralizedSupport}\ (D_c, A)$;

Algorithm 10: ↩ Revert the preconditioning procedure

Input: $\left(\lambda^*_{0,c,r}, \ldots, \lambda^*_{M,c,r}\right)$: solution obtained using the preconditioning, A: rotation matrix, μ_1: expectation(s) in the original coordinates

Output: $(\lambda^*_0, \ldots, \lambda^*_M)$: solution in the original basis

Revert rotation and scaling

$$\left(\lambda^*_{0,c}, \ldots, \lambda^*_{M,c}\right) = \textit{ConvertTransformedSolutionToCentral}\left(\left(\lambda^*_{0,c,r}, \ldots, \lambda^*_{M,c,r}\right), A\right) ;$$

Revert centralization

$$(\lambda^*_0, \ldots, \lambda^*_M) = \textit{ConvertCentralizedSolutionToOriginalBasis}\left(\left(\lambda^*_{0,c,r}, \ldots, \lambda^*_{M,c,r}\right), \mu_1\right) ;$$

Algorithm 11: ↩ Preconditioning and dual function minimization

Input: $\{\mu_0, \ldots, \mu_M\}$: set of moment constraints, D: approximation of the support

Output: $(\lambda^*_0, \ldots, \lambda^*_M)$: solution in the original basis, $\natural$: flag that shows whether the minimization procedure convergences

$$(\{\mu_{0,c,r}, \ldots, \mu_{M,c,r}\}, D_{c,r}, A) = \textit{Preconditioning procedure}\ (\{\mu_0, \ldots, \mu_M\}, D) ;$$

$$\left((\lambda^*_{0,c,r}, \ldots, \lambda^*_{M,c,r}), \natural\right) = \textit{MinimizeDualFunction}\ (\{\mu_{0,c,r}, \ldots, \mu_{M,c,r}\}, D_{c,r}) ;$$

$$(\lambda^*_0, \ldots, \lambda^*_M) = \textit{Revert the preconditioning procedure}\left(\left(\lambda^*_{0,c,r}, \ldots, \lambda^*_{M,c,r}\right), A, \mu_1\right) ;$$

Appendix A

Supplementary Information

A.1 Derivation of the Chemical Master Equation

The chemical master equation (CME) (2.2) describes the change of probability mass $\mathbf{P}[\vec{X}(t) = x]$ for each discrete state x in the state space (in this section we denote $\vec{X}(t) = \vec{X}_t$). Let us determine the state probabilities for $\vec{X}$ at time t

$$\mathbf{P}[\vec{X}(t) = x] = \sum_{\vec{y} \in \mathbb{Z}_S^N} \mathbf{P}[\vec{X}(t) = x \mid \vec{X}_0 = \vec{y}] \cdot \mathbf{P}[\vec{X}_0 = \vec{y}],$$

where the initial distribution $\mathbf{P}[\vec{X}_0 = \vec{y}], \vec{y} \in \mathbb{Z}_S^N$ is given and propensity functions $\alpha_1, \ldots, \alpha_M$ are known. For an infinitesimal time step of length Δ, we have

$$\begin{aligned} \mathbf{P}[\vec{X}_{t+\Delta} = x] &= \mathbf{P}[\vec{X}_{t+\Delta} = x \mid \vec{X}_t = x] \cdot \mathbf{P}[\vec{X}_t = x] \\ &+ \sum_{\substack{j=1 \\ x - v_j \geq 0}}^{N_R} \mathbf{P}[\vec{X}_{t+\Delta} = x \mid \vec{X}_t = x - v_j] \cdot \mathbf{P}[\vec{X}_t = x - v_j], \end{aligned}$$

where the first term relates to the probability of staying in the state x (no reaction occurs) and the second one relates to the probability of changing the state from $x - \vec{v}_j$ to x (reaction R_j occurred). Using the notation of propensity functions we obtain the following

$$\begin{aligned} \mathbf{P}[\vec{X}_{t+\Delta} = x] &= \left(1 - \textstyle\sum_1^{N_R} \alpha_j(x) \cdot \Delta\right) \cdot \mathbf{P}[\vec{X}_t = x] \\ &+ \sum_{\substack{j=1, \\ x - \vec{y} \geq 0}}^{N_R} \alpha_j(x - v_j) \cdot \Delta \cdot \mathbf{P}[\vec{X}_t = x - v_j] \end{aligned}$$

We can divide both parts on the Δ which is infinitesimal and thus obtain the derivative of the state probability

$$\frac{d}{dt}\mathbf{P}[\vec{X}_t = x] = \lim_{\Delta \to 0} \frac{\mathbf{P}[\vec{X}_{t+\Delta} = x] - \mathbf{P}[\vec{X}_t = x]}{\Delta} = \\ = -\sum_{j=1}^{N_R} \alpha_j(x) \cdot \mathbf{P}[\vec{X}_t = x] + \sum_{\substack{j=1, \\ x - v_j \geq 0}}^{N_R} \alpha_j(x - v_j) \cdot \mathbf{P}[\vec{X}_t = x - v_j].$$

This differential equation describes the system of the coupled ODEs for the probability for every state x and probabilities for states $\vec{y} = x - v_j$ are needed to compute the right-hand side. Given an initial probability distribution, the solution of the CME are the probabilities $\mathbf{P}[\vec{X}_t = x]$ for all states x. The intuitive meaning of the CME is that the derivative of the probability of state x is the *difference* between the probability inflow and outflow. The states are seen as nodes in a flow network and their probability is the amount of fluid, which moves through the network according to the propensities. The example of CME construction is given in Example 2.3.

A.2 Software to solve the CME numerically

Tool name	Language	Reference
SHAVE	C++, Lua	[146]
cmepy	Python	[61]
FSP	Matlab	[175]
Implicit Euler	Matlab	[121]
ExpoKit	Python	[198]
TT-Toolbox adaptation	Python	[127]

A.3 Transformation between central and non-central moments

Let us denote the $|\mathbf{I}|$-th central moment by $\bar{\mu}^{(|\mathbf{I}|)}$, where $\mathbf{I}$ is the order vector such that $\mathbf{I} = (I_1, \ldots, N_S)$. The central moment is given by [23, 122]

$$\bar{\mu}^{(|\mathbf{I}|)} = E\left[\prod_{i=1}^{N_S} \left(\vec{X}_i - E\left(\vec{X}_i\right)\right)^{I_i}\right]$$

It can be expressed in terms of non-central moments μ using the binomial expansions as follows

$$\bar{\mu}^{(|\mathbf{I}|)} = \sum_{i_1=0}^{I_1} \cdots \sum_{i_{N_S}=0}^{I_{N_S}} (-1)^{|\mathbf{I}|} \binom{I_1}{i_1} \cdots \binom{I_{N_S}}{i_{N_S}} \cdot E\left(\vec{X}_1\right)^{i_1} \cdots E\left(\vec{X}_{N_S}\right)^{i_{N_S}} \cdot \mu^{(\mathbf{I}-\mathbf{i})}, \qquad \text{(A.1)}$$

where $\mathbf{i} = (i_1, i_2, \ldots, i_{N_S})$. Non-central moments can also be expressed in terms of central moments as follows

$$\begin{aligned}\mu^{(|\mathbf{I}|)} &= E\left[\prod_{i=1}^{N_S}\left(\vec{X}_i\right)^{I_i}\right] = E\left[\prod_{i=1}^{N_S}\left\{\left(\vec{X}_i - E\left[\vec{X}_i\right]\right) + E\left[\vec{X}_i\right]\right\}^{I_i}\right] \\ &= \sum_{i_1=0}^{I_1}\cdots\sum_{i_{N_S}=0}^{I_{N_S}}\binom{I_1}{i_1}\cdots\binom{I_{N_S}}{i_{N_S}}\cdot E\left(\vec{X}_1\right)^{i_1}\cdots E\left(\vec{X}_{N_S}\right)^{i_{N_S}}\cdot\bar{\mu}^{(\mathbf{I}-\mathbf{i})}.\end{aligned} \tag{A.2}$$

A.4 Moment Closure Software

Tool name	Language	Reference
Moment Closure	Mathematica	[159]
StochDynTools	Maple	[112]
SHAVE	C++, Lua	[146]
MOCA	Mathematica	[194]
MEANS	Python	[68]
MomentClosure	Python	[167]
MultiKin	Matlab	[201, 205]
CERENA	Matlab	[128]

We note that the authors of CERENA [128, p. 3] also provide the detailed comparison to other software packages.

A.5 Maximum Entropy Formalism

A.5.1 General Solution to the Maximum Entropy Functional

We consider the optimization problem (4.4)

$$q = \arg\max_{p\in\mathcal{G}} H(p),$$

given the constraints (4.2)

$$\begin{aligned}\sum_{x\in S} p(x) = 1, \quad p_i \geq 0 \\ \sum_{x\in S} f_k(S)p(x) = \mu_k, \quad 1 \leq k \leq M < \infty,\end{aligned}$$

and entropy defined according to (4.3)

$$H(p) = -\sum_{x\in S} p(x)\cdot\ln\left(p(x)\right).$$

Here we assume that the functions f_k depend only on the current state x and $f_0(x) = 1$, so the constraints are given by

$$\sum_{x \in S} f_k(x) p(x) = \mu_k, \quad 0 \leq k \leq M < \infty.$$

In order to minimize the entropy functional, we consider the following Lagrangian where we do not specify the functions $f_k(x)$

$$\mathcal{L}(p, \lambda) = H(p) - \sum_{k=0}^{M} \lambda_k \left(\sum_{x \in S} f_k(x) p(x) - \mu_k \right).$$

Derivation with respect to the function $p(x)$ leads to

$$\frac{\partial \mathcal{L}}{\partial p} = - \left(\sum_{x \in S} \ln[p(x)] + 1 \right) - \sum_{k=0}^{M} \lambda_k \sum_{x \in S} f_k(x).$$

Setting this derivative to zero results in

$$\frac{\partial \mathcal{L}}{\partial p} = 0 \iff - \left(\sum_{x \in S} \ln[p(x)] + 1 \right) = \sum_{k=0}^{M} \lambda_k \sum_{x \in S} f_k(x).$$

This relation holds for every state $x \in S$, and if the sum $\sum_{x \in S} f_k(x)$ converges,

$$\ln[p(x)] = - \sum_{k=0}^{M} \lambda_k f_k(x) - 1.$$

This gives the general form of solution

$$q(x) = \exp \left(-1 - \sum_{k=0}^{M} \lambda_k f_k(x) \right), \tag{A.3}$$

where we can also define the normalization constant $Z(x)$ such that

$$q(x) = \frac{1}{Z(x)} \exp \left(- \sum_{k=1}^{M} \lambda_k f_k(x) \right), \quad Z = e^{1+\lambda_0} = \sum_{x \in S} \exp \left(- \sum_{k=1}^{M} \lambda_k f_k(x) \right).$$

The solution q obviously depends both on the state x and the parameter vector λ, so we can also write $q(x, \lambda)$ for clarity. It was shown in [4] that the solution to the optimization problem (4.4) can be found via minimization of the dual function $\Psi(\lambda)$. It is defined as

	primal	dual
problem	$\arg\max_{p\in\mathcal{G}} H(p)$	$\arg\min_{\lambda\in\mathbb{R}^M} \Psi(\lambda)$
description	maximum entropy	maximum likelihood
type of optimization	constrained	unconstrained
solution and search domain	$p \in \mathcal{G}$	$\lambda \in \mathbb{R}^M$

Table A.1: ↩ The duality of maximum entropy and relation to the maximum likelihood.

$\Psi(\lambda) = \mathcal{L}(q, \lambda)$ and

$$\begin{aligned}\Psi(\lambda) &= \ln Z + \frac{1}{Z}\sum_{k=1}^{M}\lambda_k \sum_{x\in S} f_k(x)\, e^{-\sum_{k=1}^{M}\lambda_k f_k(x)} - \sum_{k=0}^{M}\lambda_k\left(\sum_{x\in S}\frac{f_k(x)}{Z}e^{-\sum_{k=1}^{M}\lambda_k f_k(x)} - \mu_k\right)\\ &= \ln Z + \sum_{k=1}^{M}\lambda_k\mu_k.\end{aligned}$$

Therefore, the *unconstrained* optimization problem that has to be solved is stated as follows

$$\lambda^* = \arg\min_{\lambda\in\mathbb{R}^M} \Psi(\lambda), \tag{A.4}$$

and the corresponding maximum entropy distribution is

$$q(x, \lambda^*) = \exp\left(-1 - \sum_{k=0}^{M}\lambda_k^* f_k(x)\right).$$

In this thesis we mostly use the expression $q(x, \lambda^*) = \exp\left(-\sum_{k=0}^{M}\lambda_k^* f_k(x)\right)$, where we assume the proper transformation of λ_0 [91].

The authors of [30] establish the connection between the maximum entropy and the maximum likelihood framework (for the domain of the natural language processing). Here we present slightly modified table of correspondences [30, Table 1]. In [30], the authors address the dual problem as maximum likelihood, however this is due to definition of dual function, which is the same as $(-\Psi(\lambda))$ in our case. The extended results on the duality of entropy optimization problems in case of inequality constraints are given in [181].

A.5.2 Minimization of the Dual Function

The convexity property of the dual function $\Psi(\lambda)$ allows us to postulate that there exists a unique solution to the optimization problem (A.4). We first compute the gradient $\partial\Psi/\partial\lambda$

of the dual function. The i-th element of the gradient vector is given by

$$\frac{\partial \Psi}{\partial \lambda_i} = \mu_i - \underbrace{\frac{-f_i(x)}{Z} e^{-\sum_{k=1}^{M} \lambda_k f_k(x)}}_{=\tilde{\mu}_i} = \mu_i - \frac{1}{Z}\tilde{\mu}_i, \quad i = 1, \ldots, M,$$

where $\tilde{\mu}_i$ is the approximation of i-th moment given the vector of parameters λ,

$$\widetilde{\mu}_i = \sum_{x \in S} x^i \exp\left(-\sum_{k=1}^{M} \lambda_k x^k\right), \quad i = 1, \ldots, 2M. \tag{A.5}$$

In case of continuous random variables the moments are approximated as

$$\widetilde{\mu}_i = \int_S x^i \exp\left(-\sum_{k=1}^{M} \lambda_k x^k\right) \mathrm{d}x, \quad i = 1, \ldots, 2M. \tag{A.6}$$

The elements of the Hessian matrix H are given by

$$H_{ij} = \frac{\partial^2 \Psi}{\partial \lambda_i \partial \lambda_j} = \frac{\partial \left(\mu_i - \frac{1}{Z}\tilde{\mu}_i\right)}{\partial \lambda_j} = \frac{1}{Z^2}\left(\tilde{\mu}_{i+j} Z - \tilde{\mu}_i \tilde{\mu}_j\right), \quad i, j = 1, \ldots, M \tag{A.7}$$

It can be proven [243, Proposition 2.3] [162, Lemma 1] [7, Section 3] that the Hessian matrix H is positive definite. Therefore, if there exists the stationary point $\bar{\lambda}$ where the gradient is zero $\frac{\partial \Psi}{\partial \lambda}|_{\lambda=\bar{\lambda}} = \mathbf{0}$, it must be a unique absolute minimum. However, there is no guarantee that the feasible solution to the problem (4.9) exists, since the convexity of Ψ is established without considering the properties of the constraint set μ_k. The generalization of this result to Burg entropy is given in [38].

A.5.3 Remarks about Numerical Optimization Procedure

Here we want to address several issues of the numerical minimization of the dual function that may play an important role when it comes to computationally intensive tasks such as parameter estimation (cf. Section 5.3).

Preconditioning: centralization and rotation. As mentioned in Section 4.3.2, we apply the preconditioning procedure similar to the one in [3]. The centralization and rotation of the support corresponds to the step 2 of the procedure. Centralization step corresponds to converting the raw moments into central moments such that $\mu_{1,c} = 1$ for all species (as defined in Section A.3) and to the scaling of the support $D = \{x_L, \ldots, x_R\}$ as $D_c = \{x_L - \mu_1, \ldots, x_R - \mu_1\}$. In the two-dimensional case, this reads as $D_{xy,c} = \{x_L - \mu_{x,1}, \ldots, x_R - \mu_{x,1}\} \times \{y_L - \mu_{y,1}, \ldots, y_R - \mu_{y,1}\}$.

The rotation and scaling step is defined such that the variance $\mu_{2,r} = 1$ or the matrix of second moments is an identity matrix I. The corresponding transformations are defined for an arbitrary dimensionality in [3]. In Algorithm 9 we show these preconditioning that are performed before the optimization.

It should be noted that this procedure also require to transform the solution vector λ correspondingly to a new basis and back, after the optimization has been performed. Therefore, to verify that the support approximation procedure is always performed in the original basis, the transformations are reverted with inverse transformations as given in Algorithm 10. This process is illustrated in Figure A.1, where states are shown by blue circles. The solution scheme that makes use of the preconditioning is shown in Algorithm 11. We consider this step in the following example:

Example A.1. ↩ We consider the reconstruction of two-dimensional protein distribution in exclusive switch model at the time $t = 60$ using *joint MCM* method. We approximate the initial support for D_{P_1,P_2} as in Section 4.3.4 by $D_{P_1,P_2} = \{0,\dots,117\} \times \{0,\dots,280\}$. The condition number of the Hessian matrix is $\kappa\left(H_{D_{P_1,P_2}}\right) = 4.3 \cdot 10^{23}$. The first moments of the proteins are given by $\mu_{1,P_1} = 60.19$ and $\mu_{1,P_2} = 152.29$. The support after applying the centralization is given by $D_{P_1,P_2,c} = \{-60.19,\dots,56.81\} \times \{-152.29,\dots,127.70\}$. The matrix M_2 of second moments reads as

$$M_2 = \begin{pmatrix} 1718.12 & -4063.01 \\ -4063.01 & 10502.28 \end{pmatrix}.$$

The transformation matrix A is defined as $A = \left(VD^{1/2}\right)^{-1}$, where D is a diagonal matrix with eigenvalues of M_2 and the matrix V contains the corresponding right eigenvectors (such that $AV = VD$). In this case, the transformation matrix reads as

$$A = \begin{pmatrix} -0.0826 & -0.0324 \\ -0.0033 & 0.0085 \end{pmatrix}.$$

We apply the transformation to each point (x_c, y_c) of the centralized support $D_{P_1,P_2,c}$ as $(x_{c,r}, y_{c,r}) = A(x_c, y_c)$ which results in the transformed support $D_{P_1,P_2,c,r}$. For example, the point $(-60.19, -152.29)$ is mapped to $(9.90, -1.09)$ and point $(56.81, 127.70)$ is mapped to $(-8.83, 0.90)$. The condition number of the Hessian is $\kappa\left(H_{D_{P_1,P_2,c,r}}\right) = 6.6 \cdot 10^{10}$.

Please note that all three support approximations D_{P_1,P_2}, $D_{P_1,P_2,c}$ and $D_{P_1,P_2,c,r}$ are treated as discrete sets (not as continuous subsets of $\mathcal{R}^2$) and the number of elements stays the same. So, in this case the maximum entropy distribution is approximated in 33158 points. The set $D_{P_1,P_2,c,r}$ contains the same amount of elements but the coordinates of each point are usually real numbers.

Preconditioning: basis orthogonalization. We consider the orthogonalization of the problem basis. The original basis is given by the set of monomials with the corresponding Lagrange multipliers $\{\vec{x}^{\vec{i}}, \lambda_{\vec{i}}\}$, $\vec{i} \in \mathbb{Z}^{N_S}$, $0 \leq |\vec{i}| \leq M$. The orthogonalized basis is given by the set polynomials of order M and the corresponding Lagrange multipliers $\{v_k(\vec{x}), \gamma_k\}$,

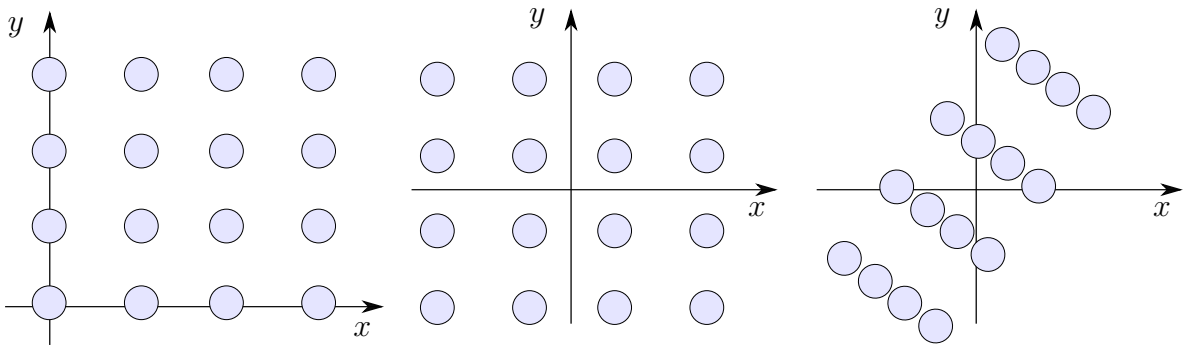


Figure A.1: ↩ Centralization, rotation and scaling of two-dimensional support.

$1 \leq k \leq K$, where $K = M + 1$ in one-dimensional case and $K = {}^{(M+2)(M+1)}/_{2}$ in two-dimensional case. The dual function Equation 4.8 in the new basis is given by

$$\Psi(\vec{x}, \vec{\gamma}) = \ln \sum_{x} \exp\left(\sum_{k=1}^{K} \gamma_k v_k(\vec{x})\right) + \sum_{k=1}^{K} \gamma_k v_k(\vec{\mu}). \tag{A.8}$$

The solution of the dual problem is then given by

$$\tilde{q}_p(\vec{x}, \vec{\gamma}) = \exp\left(\sum_{k=1}^{K} \gamma_k v_k(\vec{x})\right)$$

and the corresponding support approximation D. There exist many well-established systems of orthogonal basis functions, among which shifted Chebyshev [25] and Lagrange interpolation polynomials (with suitably spaced roots) [226] have been previously applied for the solution of inverse moment problem. Here, we construct the system of K general orthogonal polynomial functions.

The potential benefit of using the orthogonal basis comes from the basic formulation of Newton optimization

$$\begin{aligned}
\vec{\gamma}^{(\ell+1)} &= \vec{\gamma}^{(\ell)} - \xi_\ell \left(H^{-1} \cdot \nabla\Psi\right)|_{\vec{\gamma}^{(\ell)}} \\
(\nabla\Psi)_k &= \frac{\partial\Psi}{\partial\gamma_k} = Q_{\vec{\gamma}}(v_k) - v_k(\vec{\mu}) \\
H_{kl} &= \frac{\partial^2\Psi}{\partial\gamma_k\partial\gamma_l} = Q_{\vec{\gamma}}(v_k v_l),
\end{aligned}$$

where the functional Q is defined on all real-valued polynomial functions $v \in \mathcal{P}_{N_S,K}$, $v : \mathbb{R}^{N_S} \mapsto \mathbb{R}$ such that $Q : \mathbb{R}^K \times \mathcal{P}_{N_S,K} \mapsto \mathbb{R}$ and

$$Q_{\vec{\gamma}}(v) = \sum_{x \in D} v(\vec{x}) \cdot \tilde{q}(\vec{x}, \vec{\gamma})\, \mathrm{d}\vec{x}.$$

For the orthogonal system of functions $Q_{\vec{\gamma}}(v_k v_l) = \sum\limits_{x \in D} v_k(\vec{x}) v_l(\vec{x}) \cdot \tilde{q}(\vec{x}, \vec{\gamma})\, \mathrm{d}\vec{x} = \delta_{kl}$, therefore $H_{kl} = \delta_{kl} = H_{kl}^{-1}$ and $H = I$. The Newton iteration becomes

$$\vec{\gamma}^{(\ell+1)} = \vec{\gamma}^{(\ell)} - \xi_\ell \left(\nabla \Psi\right)|_{\vec{\gamma}^{(\ell)}}.$$

Note that the optimization problem remains convex under such transformation

$$\vec{w}^T H \vec{w} = \vec{w}^T \vec{w} \geq 0.$$

Here we consider the orthogonalization of basis only for the one-dimensional maximum entropy problem. The initial basis $\{1, x, x^2, \ldots, x^M\}$ corresponds to the identity matrix I, where the coefficients of the function k correspond to kth column. To obtain the orthogonal system $P^\perp$, we apply the modified Gram-Schmidt with re-orthogonalization [55]. In order to apply it properly, we need to define the initial approximation of the solution vector $\vec{\gamma}^{(0)} = \vec{\lambda}^{(0)}$ on each iteration of the support extension, $\vec{\gamma}^{(0)} = (10^{-10}, 0, \ldots, 0)$. It is required by the definition of functional Q, where $\tilde{q}$ shall not be degenerate.

The solution vector $\vec{\gamma}$ must be re-computed with respect to the new basis $P^\perp$ such that the functional Q does not change

$$\exp\left(\sum_{k=1}^{K} \gamma_k^{(0),\perp} P_k^\perp\right) = \exp\left(\sum_{k=1}^{K} \gamma_k^{(0)} P_k\right)$$

$$\sum_{k=1}^{K} \gamma_k^{(0),\perp} P_k^\perp = \sum_{k=1}^{K} \gamma_k^{(0)} P_k$$

where $P_k^\perp$ and P_k denote the kth column of the orthogonalized and initial basis, i.e., the vector corresponding to kth basis function $v_k(x)$. In the matrix-vector form, it is given by

$$\gamma_1^{(0),\perp} \begin{pmatrix} p_{11}^\perp \\ \vdots \\ p_{1M}^\perp \end{pmatrix} + \gamma_2^{(0),\perp} \begin{pmatrix} p_{21}^\perp \\ \vdots \\ p_{2M}^\perp \end{pmatrix} + \ldots + \gamma_K^{(0),\perp} \begin{pmatrix} p_{K1}^\perp \\ \vdots \\ p_{KM}^\perp \end{pmatrix} =$$

$$\gamma_1^{(0)} \begin{pmatrix} p_{11} \\ \vdots \\ p_{1M} \end{pmatrix} + \gamma_2^{(0)} \begin{pmatrix} p_{21} \\ \vdots \\ p_{2M} \end{pmatrix} + \ldots + \gamma_K^{(0)} \begin{pmatrix} p_{K1} \\ \vdots \\ p_{KM} \end{pmatrix},$$

$$\begin{bmatrix} \gamma_1^{(0),\perp} p_{11}^\perp + \gamma_2^{(0),\perp} p_{21}^\perp + \ldots + \gamma_K^{(0),\perp} p_{K1}^\perp \\ \gamma_1^{(0),\perp} p_{12}^\perp + \gamma_2^{(0),\perp} p_{22}^\perp + \ldots + \gamma_K^{(0),\perp} p_{K2}^\perp \\ \cdots \\ \gamma_1^{(0),\perp} p_{1M}^\perp + \gamma_2^{(0),\perp} p_{2M}^\perp + \ldots + \gamma_K^{(0),\perp} p_{KM}^\perp \end{bmatrix} = \begin{bmatrix} \gamma_1^{(0)} p_{11} + \gamma_2^{(0)} p_{21} + \ldots + \gamma_K^{(0)} p_{K1} \\ \gamma_1^{(0)} p_{12} + \gamma_2^{(0)} p_{22} + \ldots + \gamma_K^{(0)} p_{K2} \\ \cdots \\ \gamma_1^{(0)} p_{1M} + \gamma_2^{(0)} p_{2M} + \ldots + \gamma_K^{(0)} p_{KM} \end{bmatrix}.$$

The right part equals $P \cdot \vec{\gamma}^{(0)}$. If we define it by $b^{(0)} = P \cdot \vec{\gamma}^{(0)}$, the equation reads as

$$\begin{bmatrix} p_{11}^{\perp} & p_{21}^{\perp} & \dots & p_{K1}^{\perp} \\ p_{12}^{\perp} & p_{22}^{\perp} & \dots & p_{K2}^{\perp} \\ \dots & \dots & \dots & \dots \\ p_{1M}^{\perp} & p_{2M}^{\perp} & \dots & p_{KM}^{\perp} \end{bmatrix} \begin{pmatrix} \gamma_1^{(0),\perp} \\ \gamma_2^{(0),\perp} \\ \vdots \\ \gamma_K^{(0),\perp} \end{pmatrix} = \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_K^{(0)} \end{pmatrix},$$

and the representation of solution coefficients $\vec{\gamma}$ in the basis is defined by $\vec{\gamma}^{(0),\perp} = (P^{\perp})^{-1} \cdot b^{(0)} = (P^{\perp})^{-1} \cdot P \cdot \vec{\gamma}^{(0)}$.

After defining the orthogonalized basis, the dual function minimization is performed. Though theoretically $H = I$, in our case studies this equality holds only approximately. Therefore, we can not directly apply the simplified form of the Newton method and still need to compute the inversion of the Hessian matrix H^{-1} (for the considered number of moment constraints M, this operation remains computationally cheap). In contrast to the approach of Alldredge [8], we do not perform the re-orthogonalization on each iteration of the Newton procedure. Instead, we follow the original approach of Abramov and the Hessian matrix loses its property $H = I$ while iterating. Still, it allows for drastic reduction of the condition number of the Hessian (cf. Table 4.9).

After the minimum $\vec{\gamma}^{*,\perp}$ of the dual function $\Psi(\vec{x}, \vec{\gamma})$ is found for the current support approximation, the solution vector $\vec{\gamma}^{*}$ needs to be represented in the original monomial basis via

$$\gamma_1^{*,\perp} \begin{pmatrix} p_{11}^{\perp} \\ p_{12}^{\perp} \\ \vdots \\ p_{1M}^{\perp} \end{pmatrix} + \ldots + \gamma_K^{*,\perp} \begin{pmatrix} p_{K1}^{\perp} \\ p_{K2}^{\perp} \\ \vdots \\ p_{KM}^{\perp} \end{pmatrix} = \lambda_1^{*} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \ldots + \lambda_K^{*} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

therefore $P^{\perp} \cdot \vec{\gamma}^{*,\perp} = I \cdot \vec{\lambda}^{*}$ and $\vec{\lambda}^{*} = (-1)P^{\perp} \cdot \vec{\gamma}^{*}$ (up to a sign). Orthogonalization preconditioning procedure is shown in Algorithm 5. Note that in case of continuous distribution reconstruction, the additional correction of λ_0 is also needed, i.e.,

$$\lambda_0^{*} = \left((-1)P^{\perp} \cdot \vec{\gamma}^{*}\right)_0 - \ln \sqrt{\frac{1}{\mu_2}},$$

where $\sqrt{\frac{1}{\mu_2}}$ plays a role of transformation matrix in one-dimensional case.

If the basis is orthogonalized, the values of the exponential function can not be easily precomputed once for all the iterations (as it is shown in Appendix A.5.3). Moreover, these computations become the bottleneck of the whole iteration procedure together with the inner product functional Q (and the corresponding running time is up to 6 larger than the one of the version without preconditioning but with exponential values precomputation).

Example A.2. ↩ We consider the reconstruction of one-dimensional protein distribution in exclusive switch model at the time $t = 60$ given $M = 6$. The original monomial support $1, x, \ldots, x^M$ is characterized by the matrix $P = I$. For the approximate of support given by $D = \{15, \ldots, 110\}$, the orthogonalized support (obtained by applying the modified Gram-Schmidt with re-orthogonalization [55]) is characterized by the matrix $P^{\perp}$:

$$P^{\perp} = \begin{bmatrix} 0.1021 & -0.2302 & 0.4664 & -0.9637 & 2.0227 & -4.2813 \\ 0 & 0.0037 & -0.01858 & 0.06315 & -0.1845 & 0.4992 \\ 0 & 0 & 0.0001 & -0.0011 & 0.0054 & -0.0206 \\ 0 & 0 & 0 & 6.108\,9 \cdot 10^{-6} & -6.318\,9 \cdot 10^{-5} & 0.0004 \\ 0 & 0 & 0 & 0 & 2.527\,6 \cdot 10^{-7} & -3.279\,1 \cdot 10^{-6} \\ 0 & 0 & 0 & 0 & 0 & 1.049\,3 \cdot 10^{-8} \end{bmatrix}.$$

As an example, the third column corresponds to the basis function $v_3(x) = 0.4664 - 0.01858x + 0.0001x^2$.

Further preconditioning of moment constraints Abramov [2] introduces one more step of moment preconditioning performed using additional scalar re-scaling:

$$\tilde{\mu}_I = \alpha^{-|I|}\mu_I,$$

for $0 \leq |I| \leq M$ such that the set of moments $\tilde{\mu}$ is the same as of the standard normal distribution. To obtain such a transformation, the rescaling parameter α must be $\alpha = [(2M-1)!!]^{1/2M}$. However, this transformation does not give noticeable benefit for our case studies. This approach shall provide more visible difference in results for multiple dimensions (more than 2).

Precomputation of exponential function. The proposed method to solve the moment problem using the entropy maximization extensively uses the values of the function $f_{\exp,1D}(x) = \exp(-\sum_{k=1}^{M} \lambda_k x^k)$ (one-dimensional case) or $f_{\exp,2D}(x,y) = \exp(-\sum_{1 \leq r+l \leq M} \lambda_{r,l} x^r y^l)$ (two-dimensional case). Analyzing the performance of the implementation, we realized that most of the time is actually spent to compute these values. In order to optimize the program execution, we precompute the values of these functions once and later we just re-use them. For all the case studies, the main part of probability mass is located in a finite part of the support. Therefore, we precompute the values as shown in Table A.2. In one-dimensional case it results in the table of values similar to Table A.3. The construction is similar to Vandermonde matrix. The positive effect is more obvious during the parameter estimation procedure since reconstruction is then conducted many times.

Please note, that is preconditioning techniques (Section 4.3.2) are used, the given precomputation approach can not be used since the transformed support may not be a subset of $\mathbb{Z}$. In order to partially address this problem, we may apply the precomputation (similar to the one provided above) on each iteration of the support transformation. However, for

function	support D	M	running time (s)
$f_{\text{exp},1D}(x)$	$[0, 1000]$	20	0.02
$f_{\text{exp},2D}(x, y)$	$[0, 1000] \times [0, 1000]$	20	115

Table A.2: ↩ Parameters of the exponential function precomputation.

the considered case studies this approach does not provide any benefit in terms of running time.

Right tail behavior. For most case studies we can over-approximate the support by considering $\mathbb{R}$ (for continuous distributions) or $\mathbb{N}$ (for discrete distributions), which we can directly use to approximate the values of moments $\tilde{\mu_k}$ (A.5), (A.6). However, it is only possible if the underlying distribution vanishes in the infinity, i.e., the limit $\lim_{x \to \pm\infty} q(x) = 0$. It does not have to hold for each iteration i that $\lim_{x \to 0} q_i(x) = 0$ in case when the state space is $\mathbb{N}$ or $\mathbb{R}_{\geq 0}$. Therefore, we seek for the best approximation of the support. We take the safe strategy and use the "average-sized" initial approximation of the support [215] to rely on the convergence criteria in dual function minimization (4.11). However, it may happen that the convergence does not hold and the approximation of the distribution becomes large so that the approximated moments grow too fast and the whole procedure fails. To prevent this, the additional external loop with the increasing convergence tolerance δ_Ψ may be added, where we increase it by a certain factor each time the procedure does not converge (cf. Algorithm 2).

Since we require $\lim_{x \to +\infty} q(x) = 0$, we do not consider case studies similar to the examples studied in [34] such as $p(x) = {}^3\!/_2\sqrt{x}$, $x \in [0, 1]$. This would require to soften the stopping criteria which results in the slower convergence (or even the total absence of the convergence). Here we aim at the reconstruction of distributions taking roots in systems biology, where such polynomial functions are almost never observed.

Given odd M value, it is not obvious how to properly define the right boundary of the approximated support D^* as described in Section 4.3.3 using the iteration (4.11). It may happen that the values of dual function start to grow near the right boundary and do not

power \ x	0	1	2	3	...
0	1	1	1	1	...
1	0	1	2	3	...
2	0	1	4	9	...
3	0	1	8	27	...
...	...	...	...	...	...

Table A.3: ↩ Precomputed values in one-dimensional case.

diminish (with respect to the theoretical requirement $\lim_{x\to+\infty} q(x) = 0$). This is due to the shape of the polynomial function $r(x) = -\sum_{k=1}^{M} \lambda_k x^k$ for odd M.

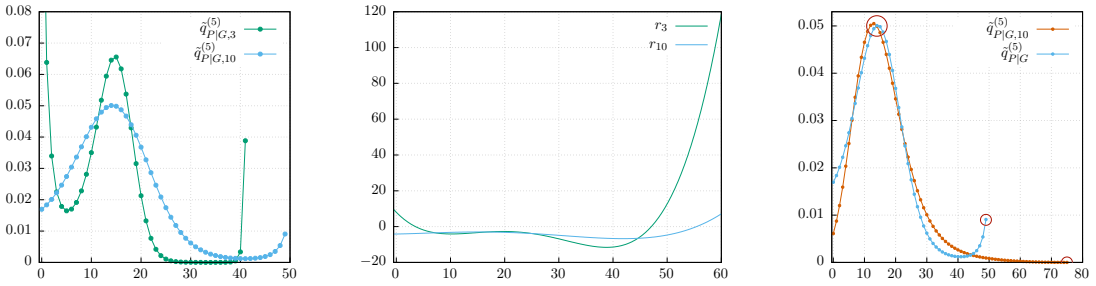
As an example, we consider the cooperative self-activation of gene expression model with the parameter set k_A, namely the conditional distribution $P|G = 1$ as shown in Figure A.2a (line is shown for better visibility). Here, the reconstruction $\tilde{q}^{(5)}_{P|G,\ell}$ and the corresponding polynomial function $r_\ell(x) = -\sum_{k=1}^{M} \lambda_k^{(\ell)} x^k$, corresponds to the ℓth iteration. It can be seen that the polynomials function start to grow in the region where the tail of the distribution is located. In order to compensate for such behavior one may apply the same truncation strategy as for the N-type generalized exponential family (cf. Example A.3) to restrict the support only to positive values. However, here we implement the strategy that does not change the λ vector directly. Instead, we compare the maximum reachable value of the dual function $\max_{x\in D^{(\ell)}} \Psi(\lambda^{(\ell)})$ on the current iteration ℓ to the value at the right tail $\Psi(\lambda^{(\ell)})(x_R^\ell)$. The iteration continues till the corresponding ratio is smaller than a threshold δ_D

$$\frac{\max_{x\in D^{(\ell)}} \Psi(\lambda^{(\ell)})(x)}{\Psi(\lambda^{(\ell)})(x_R^\ell)} < \delta_D, \tag{A.9}$$

where $\delta_D = 10^{-4}$ for most considered case studies. In Figure A.2c the corresponding points are highlighted for both the final reconstruction $\tilde{q}^{(5)}_{P|G}$ and the reconstruction obtained on 10th iteration $\tilde{q}^{(5)}_{P|G,10}$.

Additionally, we add the heuristic rule that controls the smoothness of the obtained reconstruction in the support region $\bar{D} \subset D$ where most of the probability mass is located. For most case studies this threshold for cumulative probability mass δ_F is set as $\delta_F = 0.95$. The maximum squared second derivative of the distribution is computed, where it is considered in a time-series data fashion

$$\bar{s}_q = \max_{x\in\bar{D}} \left(\frac{\mathrm{d}^2 q(x)}{\mathrm{d}x^2} \right)^2$$



(a) (b) (c)

Figure A.2: ↩ The reconstruction of $P|G = 1$ distribution, where green (blue) dots (curves are added for clarity) correspond to the 3rd (10th) iteration of the iteration procedure (a) and the corresponding polynomial functions $r(x)$ (b). The reconstruction obtained on 10th iteration (blue) together with the final reconstruction (brown), where the maximum reachable value and the value of the dual function at x_R^ℓ are highlighted using red circles.

The distribution is considered to be "smooth enough" if the value of $\bar{s}$ is smaller than a certain threshold $\delta_{\bar{s}}$ (for most of experiments it is defined as $\delta_{\bar{s}} = 10^{-2}$) or the number of peaks is smaller than $N_{\texttt{peaks}}$ (defined as $\bar{N}_{\texttt{peaks}} = 3$). The resulting procedure of ensuring the reconstruction "validity" is given in Algorithm 6.

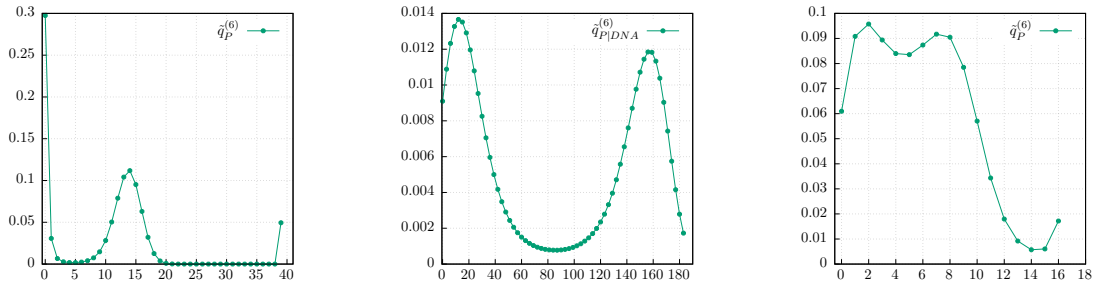
To compute the region where certain amount δ_F of probability mass is located, the Algorithm 7 (*ComputeMainProbMassRegion*) is applied, where probability mass is accumulated starting from the left boundary of the given approximation of the support. The examples of "invalid" distributions recognized by the provided algorithm are given in Figure A.3. The first two distributions violate the condition (A.9), whenever the third additionally violates the second derivative (having 3 peaks). The procedure to control the number of peaks is given in the Algorithm 8.

Another possible way to cope with such "incorrect" distributions where tail seem to grow unlimitedly is to truncate it and rescale the solution vector to compensate for the truncated probability mass (similar to Equation A.3).

A.5.4 On Error Measurement.

In order to compare the reconstructed distribution against the solution obtained using direct numerical simulation, we apply the distance defined by (4.16) and (4.17). The following distance functions can also be applied:

1. Chebyshev distance: $||\epsilon||_\infty = \max\limits_{x \in D} |\pi(x) - q(x)|$,
2. Chebyshev relative distance: $||\epsilon||_\infty^{\%} = \max\limits_{x \in D^*} \left| \frac{q(x) - \pi(x)}{\pi(x)} \right|$
3. L_1 norm: $||\epsilon||_1 = \sum\limits_{x \in D} |\pi(x) - q(x)|$,
4. χ^2 distance: $\chi^2(\pi, q) = \sum_x \frac{(\pi(x) - q(x))^2}{q(x)}$



(a) (b) (c)

Figure A.3: ↩ The reconstruction of $P|G = 1$ (in cooperative self-activation of gene expression model, set of parameters k_A) (a), $P_1|DNA$ (in exclusive switch model) (b), and $R|D_{\text{off}}$ (in simple gene expression model, fast kinetics) (c) distribution. These distributions are recognized by Algorithm 6 to be "invalid". The line is shown for better clarity.

However, they do not provide the result which is general enough to directly verify that a certain distribution is more preferable than the other [14, 15] (in those papers, we used the first three given distance functions). If no ground-truth information is available, we can stick to the techniques described in Section 4.3.5. To this end, we cite the paragraph from P. Biswas and A. K. Bhattacharya [34, p.3] on the approach to error analysis:

> "One needs to supply additional information to choose a suitable solution from an ensemble of solutions that satisfy the given moment constraints. The maximum entropy (ME) ansatz constructs the least biased solution that maximizes the entropy associated with the density and is consistent with the given moments. The accuracy of the reconstructed solution can be measured by varying the number of moments. A comparison with the exact solution (if available) would reveal to what extent the ME solution matches with the exact solution. For an unknown function with a finite set of moments, the quality of the ME solution may be judged by the proximity of the input (exact) moments to the output (approximated) moments resulting from the reconstructed distribution. By increasing the number of moments one can systematically improve the quality of the solution. It should, however, be noted, that for a function with a complicated structure, the convergence of the first few moments does not guarantee its accurate reproduction. The ME solution in this case may not represent the exact solution, but is still correct as far as the maximum entropy principle is concerned."

A.5.5 Initialization of Numerical Optimization Procedure for Dual Function

In order to proceed with any numerical optimization algorithm, the initial approximation of the solution λ needs to be defined. Along with uniform zero initialization, we may use several strategies.

A.5.5.1 One-Dimensional Reconstruction

Normal distribution. We can assume that the initial approximation $\tilde{q}^{(0)}(x)$ is given by the normal distribution:

$$\tilde{q}^{(0)}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2}\right\}.$$

The corresponding initialization is $\tilde{\lambda}^{(0)} = \left(\lambda_1^{(0)}, \lambda_2^{(0)}\right)$, where $\lambda_1^{(0)} = \frac{\mu_1}{\mu_2}$ and $\lambda_2^{(0)} = \frac{-1}{2\mu_2}$. However, this initial approximation results in a very slow convergence, therefore zero initial approximation is more effective.

Authors of [197] also investigated the investigated the convergence of the algorithm varying the initial condition. Theoretically, the maximum entropy problem shall be solvable from

any starting values but for them the convergence was sensitive to the choice of initial point (fast or slow convergence, or no convergence at all). They propose to use the normal distribution initialization and for their case studies it provided better convergence speed.

Generalized Maximum Entropy, N_k family distribution. The generalized maximum entropy reconstruction can be used to provide the initial approximation to the inverse moment problem. This is given by the type N approximation $\lambda_1 = \beta_0, \lambda_2 = \frac{1}{2}\beta_1, \ldots, \lambda_M = \frac{1}{M}\beta_{M-1}$, where β is the solution of the linear system (A.12) defined using the moment constraints as described in Section A.5.7. Such an initialization does not, however, provide any benefit in convergence speed. Instead, it usually breaks the optimization procedure such that it is not able to converge at all.

A.5.5.2 Two-Dimensional Reconstruction

Normal distribution. We can assume that the initial approximation $\tilde{q}^{(0)}(x, y)$ is given by two-dimensional normal distribution:

$$\tilde{q}^{(0)}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\},$$

where $\rho = \frac{\mu_{1,1}}{\sigma_x\sigma_y}$ is the correlation between the species. The corresponding initial approximation is given by $\tilde{\lambda}^{(0)} = -\frac{1}{2(1-\rho^2)}\left(\lambda_{1,0}^{(0)}, \lambda_{0,1}^{(0)}, \lambda_{2,0}^{(0)}, \lambda_{0,2}^{(0)}, \lambda_{1,1}^{(0)}\right)$, $\lambda_{1,0}^{(0)} = -2\left(\frac{mu_x}{\sigma_x^2} + \frac{\rho\mu_y}{\sigma_x\sigma_y}\right)$, $\lambda_{2,0}^{(0)} = \frac{1}{\sigma_x^2}$, $\lambda_{1,1}^{(0)} = \frac{2\rho}{\sigma_x\sigma_y}$, $\lambda_{0,1}^{(0)} = -2\left(\frac{mu_y}{\sigma_y^2} + \frac{\rho\mu_x}{\sigma_x\sigma_y}\right)$, $\lambda_{0,2}^{(0)} = \frac{1}{\sigma_y^2}$ (the rest of terms goes into the normalization constant). This initialization gives the result which is worse than the standard zero initialization.

Separability assumption. We can assume that the initial approximation $\tilde{q}^{(0)}(x, y)$ is separable, i.e., all covariances are ignored:

$$\begin{aligned}
\tilde{q}^{(0)}(x,y) &= \exp\Big[\lambda_{0,0} + \lambda_{1,0}x + \lambda_{2,0}x^2 + \ldots + \lambda_{M,0}x^M \\
&\qquad + \lambda_{1,1}xy + \lambda_{2,1}x^2y + (\text{mixed terms}) + \ldots \\
&\qquad + \lambda_{0,1}y + \lambda_{0,2}y^2 + \ldots + \lambda_{0,M}y^M\Big] \\
&\approx \exp\Big[\lambda_{0,0} + \lambda_{1,0}x + \lambda_{2,0}x^2 + \ldots + \lambda_{M,0}x^M \\
&\qquad + \lambda_{0,1}y + \lambda_{0,2}y^2 + \ldots + \lambda_{0,M}y^M\Big] \\
&= \tilde{q}_x(x) \cdot \tilde{q}_y(y),
\end{aligned}$$

where coefficients of $\tilde{q}_x(x)$ and $\tilde{q}_y(y)$ are obtained from one-dimensional reconstructions and $\lambda_{0,0}$ is re-scaled accordingly (the multiplier (-1) is omitted for the sake of readability). Not surprisingly this approach does not give any reasonable speed-up. Moreover, it usually initializes the optimization procedure wrongly such that it takes extremely long time to converge to the solution.

A.5.6 General Moment Functions for Maximum Entropy

In the course of this thesis we work only with monomial basis functions to obtain moments of the random variable, i.e., the set of constraints is generated using $f_k(X) = X^k$. Therefore, the corresponding moments are computed as $E\left[X^k\right] = \sum\limits_{x \in S} x^k p(x)$ or $E\left[X^k\right] = \int\limits_{x \in S} x^k p(x)\,\mathrm{d}x$ depending on the type of the random variable. Below we provide the table of some continuous maximum entropy distributions [204, Table 1] obtained for a specific type of the moment-generating function $f_k(X)$, where the standard normalization condition $\int\limits_{x \in S} p(x) = 1$ is not included for the sake of readability:

$\{x \mid f(x) > 0\}$	$f_k(X)$	MaxEnt distribution
(a, b)	$-$	uniform
$(0, 1)$	$\ln X$, $\ln(1 - X)$	beta
$(0, \infty)$	X	exponential
$(0, \infty)$	$X^\beta (\beta \neq 1)$, $\ln X$	Weibull
(a, ∞), $a > 0$	$\ln X$	Pareto
$(-\infty, \infty)$	$\lvert X \rvert$	Laplace
$(-\infty, \infty)$	X^2	normal (mean $= 0$)
$(-\infty, \infty)$	X, X^2	normal
$(-\infty, \infty)$	$\ln(1 + X^2)$	generalized Cauchy
$(-\infty, \infty)$	X, $\ln\left(1 + e^{-\lambda X}\right)$	generalized logistic
$(-\infty, \infty)$	X, $e^{-\lambda X}$	generalized extreme value

A.5.7 Maximum Entropy and Generalized Exponential Family

Consider the following non-linear diffusion

$$\mathrm{d}x_t = \mu(x_t)\,\mathrm{d}t + \sigma(x_t)\,\mathrm{d}w_t, \tag{A.10}$$

where w_t is a standard Wiener process. The following notation is used: $2\mu(x) = g(x) - v'(x)$, $\sigma^2(x) = v(x)$. Cobb, Koppstein and Chen consider in [50, 51] the general family of non-mixture multimodal densities that are stationary density functions of (A.10). These densities are applied in catastrophe theory where multiple modes can correspond to multiple stable states in dynamical systems. This family of densities is given by

$$f_k(x, \beta) = \xi(\beta) \cdot \exp\left(- \int \left[{}^{g_k(x)}/_{v(x)} \right] \mathrm{d}x \right), \tag{A.11}$$

where $\xi(\beta)$ is the normalization constant, $g_k(x) = \beta_0 + \beta_1 x + \ldots + \beta_k x^k$, $k > 0$ and the vector of parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$. The domain of $f_k(x)$ is the open interval for which $v(x)$ is positive. The distribution (4.6) is a special case of $f_k(x)$.

The authors of [51] consider the four classes of densities:

Type of density	$v(x)$	Support
N	1	$-\infty < x < \infty$
G	x	$0 < x < \infty$
I	x^2	$0 < x < \infty$
B	$x(1-x)$	$0 < x < 1$

We can show that the maximum entropy distribution (4.6) can be represented as N-type density

$$p(x) \quad = \quad \exp\left(-\sum_{k=0}^{M} \lambda_k x^k\right) = \exp\left(-\lambda_0\right) \cdot \exp\left((-1)(\lambda_1 x + \ldots + \lambda_M x^M)\right)$$

and the density $N_M(x) = \xi(\beta) \cdot \exp\left(-\int g_k(x)\,\mathrm{d}x\right)$, where $g_{M+1}(x)$ is defined as

$$\int g(x)\,\mathrm{d}x = \lambda_1 x + \ldots + \lambda_M x^M \Rightarrow g(x) = \lambda_1 + 2\lambda_2 x + \ldots + M\lambda_M x^{M-1}.$$

Therefore, the coefficients β are defined as $\beta_0 = \lambda_1, \beta_1 = 2\lambda_2, \ldots, \beta_{M-1} = M\lambda_M$ and the normalization constant $\xi(\beta) = \exp\left(-\lambda_0\right)$. Therefore, the maximum entropy density can be considered as a stationary density of a non-linear diffusion process. The densities of types G, I and B can also be obtained by applying the maximum entropy principle under certain types of constraints [248, Table 5].

The exponential family of distributions allows to determine the parameters $\beta_0, \ldots, \beta_k$ by simply solving a linear system of equations

$$\beta = M^{-1}\alpha, \tag{A.12}$$

where both matrix M and vector α are linear functions of the non-central moments μ_k. The exact form of these functions depends on the function $v(x)$ which constitutes the principal form of the distribution. Usually, for models in systems biology we consider only non-negative populations. Therefore we are interested in the first three principal forms (we apply truncation and rescaling to the density of type N in order to obtain the support $\mathrm{supp}(f_k) = [0, \infty]$).

The truncation of the support is conducted as follows. Assume that the initial approximation of the support is $D^{(0)} = \mathrm{supp}(f_k) = [-\infty, \infty]$. We simply ignore the part of the reconstruction defined on the negative part of support $[-\infty, 0)$ and compute the normalization constant $\xi(\beta)$ as $\xi(\beta) = 1/\exp\left(-\int_0^\infty g_k(x)\,\mathrm{d}x\right)$. We illustrate the application of the truncation procedure to the normal distribution in the following example:

Example A.3. ↩ Consider the normal distribution f_1 which can be described as a generalized exponential distribution of N type. It is shown in Figure A.4 (left), where $Z_{1,-} = \int_{-\infty}^{0} f_1(x)\,\mathrm{d}x$ and $Z_{1,+} = \int_{0}^{\infty} f_1(x)\,\mathrm{d}x$ are the cumulative probability mass on the negative and on the positive parts of the domain correspondingly. The maximum reachable value of the probability density is denoted by $\hat{f}_1 = \max f_1$. The same distribution with the truncated part defined on the negative domain is shown in Figure A.4 (center). However, the cumulative probability mass in this case is $Z_{1,+} < 1$. Therefore, the re-scaling is needed to construct the proper distribution f_2. The corresponding normalization constant ξ for the distribution f_2 is defined by $\xi = 1/Z_{1,+}$. The resulting truncated distribution is shown in Figure A.4 (right), where $\hat{f}_2 = \max f_2 > \hat{f}_1$ and $Z_{2,+} = \int_{0}^{\infty} f_2(x)\,\mathrm{d}x = 1$. The truncation procedure does not influence other parameters of the distribution, it only changes the normalization constant.

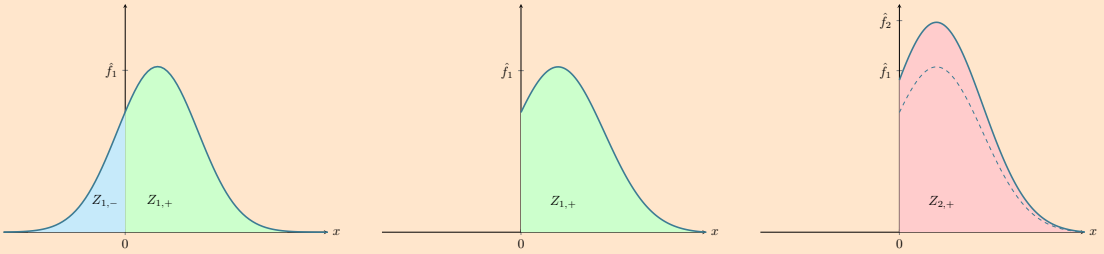


Figure A.4: The truncation of the negative part of the support. The original distribution is shown on the left, the truncated distribution before re-scaling is shown in the center and the resulting distribution is shown on the right. On all three subfigures the values on y-axis correspond to the value of probability distribution function. ↩

For example, for the form G we consider the following linear system (for $M = 2$)

$$\begin{bmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \mu_0 \\ 2\mu_1 \\ 3\mu_2 \end{bmatrix} \tag{A.13}$$

Solving the above stated linear system with respect to β, we obtain the parameters describing the distribution given a number of its non-central moments μ_k. Therefore, it can be considered as a fast alternative to maximum entropy reconstruction technique that does not require to solve the optimization problem. For instance, the authors of [250] use the similar approach to approximate the equilibrium distribution of the stochastic hybrid system. The convergence results for N type of distributions can be found in [165].

Transformation from continuous to discrete distribution. The given approach assumes that the state space of the reconstructed distributions is continuous. However, for the case studies considered in Section 4.3.6, the state space is usually given as a subset of $\mathbb{Z}^{N_S}$. Therefore, there is a need to transform the continuous reconstructed distribution (for instance, given by the generalized exponential Equation A.11). This family of transformations (w.r.t. parameter $0 < \delta < 1$) is defined as follows:

$$\tilde{q}(x) = \begin{cases} \frac{1}{\delta}\int_0^{\delta} f_k(x,\beta)\,\mathrm{d}x & \text{if } x = 0 \\ \int_{x-\delta}^{x+\delta} f_k(x,\beta)\,\mathrm{d}x & \text{if } x > 0 \end{cases}, \quad x \in \mathbb{Z}_+. \tag{A.14}$$

It is illustrated in Figure A.5 (right plot), where the value of the discretized distribution $\tilde{q}$ at point x corresponds to the highlighted area under the curve. The generalized exponential functions of both N and G types are smooth (in case if the reconstruction was successful), therefore the choice of δ mainly influences the accuracy of the approximation in the region where number of molecules is small (less than 10). However, to the best of our knowledge, there is no systematic way to choose δ for a given model. This choice does not influence the result drastically, so we can stick to "common-sense" guess $\delta = 0.5$. The provided discretization method is heuristic but it serves well for the given purpose of numerical approximation. We refer to the detailed rigorous derivation of the connection between discrete and continuous probability distributions provided in the PhD thesis of P. Thomas, [220], page. 54, Result 4.5.

Example A.4. ↩ We consider the reconstruction of the condition protein distribution $P|D_{\text{off}}$ in the gene expression model and the conditional protein distribution $P_2|DNA$ in the exclusive switch model using generalized exponential function of type N for the case of $M = 7$ moments. The results are provided in Table A.4, where it can be seen that the choice of δ influences both the approximation error for $P = 0$, $|p(0)-\tilde{q(0)}|/p(0)$ and the overall error $||\epsilon||_V$.

We plot the best reconstructions (with respect to the distance $||\epsilon||_V$) in Figure A.5, where for the distribution of $P_2|DNA$ (center) the bi-modality property is reflected in the reconstruction, however the accuracy is still low. The distribution of $P|D_{\text{off}}$ is reconstructed accurately in for most of the support but the approximation of the probability for $P = 0$ is not as accurate. This is the common problem of the considered (N and G types) generalized exponential based reconstruction approaches.

δ	error at $P = 0$	$\|\|\epsilon\|\|_V$	error at $P_2 = 0$	$\|\|\epsilon\|\|_V$
0.2	$3.64 \cdot 10^{-1}$	7.14	2.20	39.08
0.4	$3.74 \cdot 10^{-1}$	7.03	2.24	39.10
0.5	$3.80 \cdot 10^{-1}$	7.00	2.27	39.11
0.6	$3.85 \cdot 10^{-1}$	6.98	2.29	39.12
0.8	$3.96 \cdot 10^{-1}$	6.96	2.34	39.13

Table A.4: Comparison of the approximation results under the different choice of δ parameter value. ↩

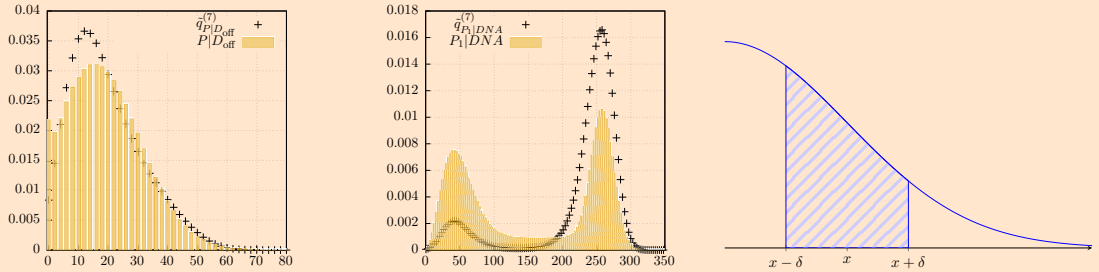


Figure A.5: ↩ The reconstruction of the protein distribution: P in gene expression model (left) and P_2 in exclusive switch model (center) in case of $M = 7$. Right plot demonstrates the integration of the continuous distribution with respect to Equation A.14. The reconstruction is shown with black crosses, whereas the result of direct solution is shown with yellow bars.

Parameter inference procedure. Consider again the parameter estimation problem formulated in Section 5.3. Let us explicitly state the difference between applying the maximum entropy and the generalized exponential approach to that problem.

The main difference is that the explicit computation of derivatives $\frac{\partial \lambda_l}{\partial c_i}$ is not possible in case of maximum entropy approach, where λ_l is one of the coefficients of the reconstructed distribution and c_i is one of the parameters to be estimated. In contrast, when applying the generalized exponential based approach, these derivatives can be symbolically computed. Therefore, the functional dependency between the input parameters c and the coefficients λ can be established. In case of maximum entropy, the procedure is based on numerical optimization which is a kind of "black box" where there exist no (known) functional dependency between c and λ, and only numerical approximations can be used to approximate the derivatives $\frac{\partial \lambda_l}{\partial c_i}$ and the derivatives of the likelihood function.

A.6 Orthogonal Polynomials with Respect to a Moment Functional

Define the complex-valued functional $\mathcal{L}$ 8on the linear space of all algebraic polynomials $\mathcal{P}$. The system of monomials is defined as $U = 1, x, x^2, \ldots$ and the functional $\mathcal{L}$ takes the values on U that are moments, i.e., $\mathcal{L}(x^k) = \mu_k$. A sequence of polynomials $\{P_n(x)\}_{n=0}^{\infty}$ is a formal orthogonal polynomial sequence with respect to a moment functional $\mathcal{L}$, if for all $i, j \in \mathbb{N}_0$

1. $P_i(x)$ is a polynomial of degree i
2. $\mathcal{L}[P_i(x)P_j(x)] = 0$ for all $i \neq j$
3. $\mathcal{L}[P_i(x)P_i(x)] \neq 0$.

If the third condition does not hold, the functional $\mathcal{L}$ is called quasi-definite, otherwise it is called positive definite [157]. The sequence of orthogonal polynomials with respect to the functional $\mathcal{L}$ exists iff. the Hankel determinants

$$\Delta_n = \begin{vmatrix} \mu_0 & \mu_1 & \cdots & \mu_{n-1} \\ \mu_1 & \mu_2 & \cdots & \mu_n \\ \vdots & \vdots & & \vdots \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-2} \end{vmatrix} \neq 0$$

for all $n \in \mathbb{N}$.

These systems are widely mostly in image processing. They include Legendre [73, 179], Zernike [216], Gaussian-Hermite [245], Bessel-Fourier [244], Gegenbauer [115] and other (cf. [113] and references therein) systems of orthogonal polynomials.

A.7 Reactions of the multi-attractor model

The multi-attractor model involves the three species *MAFAProt*, *DeltaProt*, and *PaxProt* that represent the proteins of the three genes and it involves ten species that represent the state of the genes: *PaxDna*, *MAFADna*, *DeltaDna*, *PaxDnaDeltaProt*, *MAFADnaPaxProt*, *MAFADnaMAFAProt*, *MAFADnaDeltaProt*, *DeltaDnaPaxProt*, *DeltaDnaMAFAProt*, *DeltaDnaDeltaPr*
The chemical reactions are as follows:

$$
\begin{aligned}
PaxDna &\xrightarrow{c_p} PaxDna + PaxProt \\
PaxProt &\xrightarrow{c_d} \emptyset \\
PaxDna + DeltaProt &\xrightarrow{c_b} PaxDnaDeltaProt \\
PaxDnaDeltaProt &\xrightarrow{c_u} PaxDna + DeltaProt \\
MAFADna &\xrightarrow{c_p} MAFADna + MAFAProt \\
MAFAProt &\xrightarrow{c_d} \emptyset \\
MAFADna + PaxProt &\xrightarrow{c_b} MAFADnaPaxProt \\
MAFADnaPaxProt &\xrightarrow{c_u} MAFADna + PaxProt \\
MAFADnaPaxProt &\xrightarrow{c_p} MAFADnaPaxProt + MAFAProt \\
MAFADna + MAFAProt &\xrightarrow{c_b} MAFADnaMAFAProt \\
MAFADnaMAFAProt &\xrightarrow{c_u} MAFADna + MAFAProt \\
MAFADnaMAFAProt &\xrightarrow{c_p} MAFADnaMAFAProt + MAFAProt \\
MAFADna + DeltaProt &\xrightarrow{c_b} MAFADnaDeltaProt \\
MAFADnaDeltaProt &\xrightarrow{c_u} MAFADna + DeltaProt \\
DeltaDna &\xrightarrow{c_p} DeltaDna + DeltaProt \\
DeltaProt &\xrightarrow{c_d} \emptyset \\
DeltaDna + PaxProt &\xrightarrow{c_b} DeltaDnaPaxProt \\
DeltaDnaPaxProt &\xrightarrow{c_u} DeltaDna + PaxProt \\
DeltaDnaPaxProt &\xrightarrow{c_p} DeltaDnaPaxProt + DeltaProt \\
DeltaDna + MAFAProt &\xrightarrow{c_b} DeltaDnaMAFAProt \\
DeltaDnaMAFAProt &\xrightarrow{c_u} DeltaDna + MAFAProt \\
DeltaDna + DeltaProt &\xrightarrow{c_b} DeltaDnaDeltaProt \\
DeltaDnaDeltaProt &\xrightarrow{c_u} DeltaDna + DeltaProt \\
DeltaDnaDeltaProt &\xrightarrow{c_p} DeltaDnaDeltaProt + DeltaProt
\end{aligned}
\tag{A.15}
$$

Abbreviations

CDF	**C**umulative **D**istribution **F**unction
CME	**C**hemical **M**aster **E**quation
CRN	**C**hemical **R**eaction **N**etwork
CTMC	**C**ontinuous **T**ime **M**arkov **C**hain
ODE	**O**rdinary **D**ifferential **E**quation
PDF	**P**robability **D**ensity **F**unction
SSA	**S**tochastic **S**imulation **A**lgorithm
MC	**M**oment **C**losure
ME	**M**aximum **E**ntropy
MM	**M**ethod (of) **M**oments
MCM	**M**ethod (of) **C**onditional **M**oments
wMCM	**W**eighted sum **MCM**-based reconstruction

List of Biological Models

Bibliography

[1] R. V. Abramov. An improved algorithm for the multidimensional moment-constrained maximum entropy problem. *Journal of Computational Physics*, 226 (1):621–644, 2007. doi: 10.1016/j.jcp.2007.04.026. 40

[2] R. V. Abramov. The multidimensional moment-constrained maximum entropy problem: A BFGS algorithm with constraint scaling. *Journal of Computational Physics*, 228(1):96–108, 2009. doi: 10.1016/j.jcp.2008.08.020. 34, 41, 106

[3] R. V. Abramov et al. The multidimensional maximum entropy moment problem: A review of numerical methods. *Communications in Mathematical Sciences*, 8(2): 377–392, 2010. doi: 10.4310/CMS.2010.v8.n2.a5. 1, 34, 40, 41, 42, 52, 89, 101

[4] N. Agmon, Y. Alhassid, and R. D. Levine. An algorithm for finding the distribution of maximal entropy. *Journal of Computational Physics*, 30(2):250–258, 1979. doi: 10.1016/0021-9991(79)90102-5. 39, 40, 41, 42, 99

[5] A. Ale, P. Kirk, and M. P. H. Stumpf. A general moment expansion method for stochastic kinetic models. *The Journal of Chemical Physics*, 138(17):174101, 2013. doi: 10.1063/1.4802475. 19, 26, 27

[6] A. Alfonsi, E. Cancès, G. Turinici, B. D. Ventura, and W. Huisinga. Adaptive simulation of hybrid stochastic and deterministic models for biochemical systems. *ESAIM: Proceedings and Surveys*, 14:13, 2005. doi: 10.1051/proc:2005001. 9

[7] Y. Alhassid, N. Agmon, and R. D. Levine. An upper bound for the entropy and its applications to the maximal entropy problem. *Chemical Physics Letters*, 53(1): 22–26, 1978. doi: 10.1016/0009-2614(78)80380-7. 101

[8] G. W. Alldredge, C. D. Hauck, D. P. O'Leary, and A. L. Tits. Adaptive change of basis in entropy-based moment closures for linear kinetic equations. *Journal of Computational Physics*, 258:489–508, 2014. doi: 10.1016/j.jcp.2013.10.049. 40, 41, 42, 54, 105

[9] C.-G. Ambrozie. Multivariate truncated moments problems and maximum entropy. *Analysis and Mathematical Physics*, 3(2):145–161, 2013. doi: 10.1007/s13324-012-0052-3. 39

[10] A. Ammar, E. Cueto, and F. Chinesta. Reduction of the chemical master equation for gene regulatory networks using proper generalized decompositions. *International Journal for Numerical Methods in Biomedical Engineering*, 28(9):960–973, 2012. doi: 10.1002/cnm.2476. 13

[11] A. Anandkumar and R. Ge. Efficient approaches for escaping higher order saddle points in non-convex optimization. *arXiv*, 2016. 87

[12] A. Andreychenko, P. Crouzen, L. Mikeev, and V. Wolf. On-the-fly uniformization of time-inhomogeneous infinite Markov population models. *Electronic Proceedings in Theoretical Computer Science*, 57:1–15, 2011. doi: 10.4204/EPTCS.57.1. 9

[13] A. Andreychenko, L. Mikeev, D. Spieler, and V. Wolf. Approximate maximum likelihood estimation for stochastic chemical kinetics. *EURASIP Journal on Bioinformatics and Systems Biology*, 2012(1):9, 2012. doi: 10.1186/1687-4153-2012-9. 13, 81, 82, 85, 86

[14] A. Andreychenko, L. Mikeev, and V. Wolf. Model Reconstruction for Moment-Based Stochastic Chemical Kinetics. *ACM Trans. Model. Comput. Simul.*, 25(2): 12:1–12:19, May 2015. doi: 10.1145/2699712. 110

[15] A. Andreychenko, L. Mikeev, and V. Wolf. Reconstruction of multimodal distributions for hybrid moment-based chemical kinetics. *Journal of Coupled Systems and Multiscale Dynamics*, 3(2):156–163, June 2015. doi: 10.1166/jcsmd.2015.1073. 110

[16] A. Andreychenko, L. Bortolussi, R. Grima, P. Thomas, and V. Wolf. Distribution approximations for the chemical master equation: Comparison of the method of moments and the system size expansion, pages 39–66. Contributions in Mathematical and Computational Sciences. Springer International Publishing, 2017. doi: 10.1007/978-3-319-45833-5_2. 50, 66

[17] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics*, 149 (4):1633–1648, 1998. 5

[18] M. Arns, P. Buchholz, and A. Panchenko. On the numerical analysis of inhomogeneous continuous-time Markov chains. *INFORMS Journal on Computing*, 22(3): 416–432, 2009. doi: 10.1287/ijoc.1090.0357. 9

[19] G. A. Athanassoulis and P. N. Gavriliadis. The truncated Hausdorff moment problem solved by using kernel density functions. *Probabilistic Engineering Mechanics*, 17 (3):273–291, 2002. doi: 10.1016/S0266-8920(02)00012-7. 35

[20] C. Auyeung and R. Mersereau. A dual approach to signal restoration. In *Acoustics, Speech, and Signal Processing*, pages 1326–1329. IEEE, 1989. doi: 10.1109/ICASSP.1989.266681. 40

[21] M. Avellaneda. Minimum-Relative-Entropy calibration of asset-pricing models. *International Journal of Theoretical and Applied Finance*, 01(04):447–472, 1998. doi: 10.1142/S0219024998000242. 87

[22] P. Azunre, C. Gómez-Uribe, and G. Verghese. Mass fluctuation kinetics: analysis and computation of equilibria and local dynamics. *IET Systems Biology*, 5(6):325–335, 2011. doi: 10.1049/iet-syb.2011.0013. 26

[23] N. Balakrishnan, N. L. Johnson, and S. Kotz. A note on relationships between moments, central moments and cumulants from multivariate distributions. *Statistics & Probability Letters*, 39(1):49–54, 1998. doi: 10.1016/S0167-7152(98)00027-3. 97

[24] K. Ball, T. G. Kurtz, L. Popovic, and G. Rempala. Asymptotic analysis of multiscale approximations to reaction networks. *The Annals of Applied Probability*, 16(4):1925–1961, 2006. doi: 10.1214/105051606000000420. 9, 10

[25] K. Bandyopadhyay, A. K. Bhattacharya, P. Biswas, and D. Drabold. Maximum entropy and the problem of moments: A stable algorithm. *Physical Review E*, 71 (5):057701, 2005. doi: 10.1103/PhysRevE.71.057701. 40, 41, 68, 103

[26] M. Barrio, A. Leier, and T. T. Marquez-Lago. Reduction of chemical reaction networks through delay distributions. *The Journal of Chemical Physics*, 138(10): 104–114, 2013. doi: 10.1063/1.4793982. 12

[27] A. R. Barron and C.-H. Sheu. Approximation of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991. doi: 10.1214/aos/1176348252. 34, 36

[28] P. C. Basu and A. B. Templeman. An efficient algorithm to generate maximum entropy distributions. *International Journal for Numerical Methods in Engineering*, 20(6):1039–1055, 1984. doi: 10.1002/nme.1620200607. 42

[29] B. S. Bayati. Quantifying uncertainty in the chemical master equation. *The Journal of Chemical Physics*, 146(24):244103, 2017. doi: 10.1063/1.4986762. 5

[30] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996. 100

[31] D. S. Bernstein and D. C. Hyland. Compartmental modeling and power flow analysis for state space systems. In *Proceedings of the 30th IEEE Conference on Decision and Control*, volume 2, pages 1607–1612, 1991. doi: 10.1109/CDC.1991.261676. 5

[32] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, 2012. doi: 10.1073/pnas.1118633109. 34

[33] F. Biglari, M. A. Hassan, and W. J. Leong. New quasi-Newton methods via higher order tensor models. *Journal of Computational and Applied Mathematics*, 235(8): 2412–2422, 2011. doi: 10.1016/j.cam.2010.10.041. 87

[34] P. Biswas and A. K. Bhattacharya. Function reconstruction as a classical moment problem: A maximum entropy approach. *Journal of Physics A: Mathematical and Theoretical*, 43(40):405003, 2010. doi: 10.1088/1751-8113/43/40/405003. 41, 56, 107, 110

[35] S. Bogomolov, T. A. Henzinger, A. Podelski, J. Ruess, and C. Schilling. Adaptive moment closure for parameter inference of biochemical reaction networks. *Biosystems*, 149(9308):77–89, 2015. doi: 10.1007/978-3-319-23401-4_8. 26, 82, 84

[36] J. M. Borwein and A. S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal of Control and Optimization*, 29(2):325–338, 1991. doi: 10.1137/0329017. 40

[37] J. M. Borwein, R. K. Goodrich, and M. A. Limber. A comparison of entropies in the underdetermined moment problem. *Submitted for publication*, 1993. 39

[38] J. M. Borwein, A. Lewis, M. Limber, and D. Noll. Maximum entropy reconstruction using derivative information part 2: computational results. *Numerische Mathematik*, 69(3):243–256, 1995. doi: 10.1007/s002110050090. 101

[39] P. W. Buchen and M. Kelly. The maximum entropy distribution of an asset inferred from option prices. *Journal of Financial and Quantitative Analysis*, 31(1):143–159, 1996. doi: 10.2307/2331391. 34

[40] K. Burrage, T. Tian, and P. Burrage. A multi-scaled approach for simulating chemical reaction systems. *Progress in biophysics and molecular biology*, 85(2):217–234, 2004. doi: 10.1016/j.pbiomolbio.2004.01.014. 5, 7, 10

[41] K. Burrage, M. Hegland, S. Macnamara, and R. Sidje. A Krylov-based finite state projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems. In *Mam 2006 : Markov Anniversary Meeting: an international conference to celebrate the 150th anniversary of the birth of A.A. Markov*, pages 21–38, Charleston, South Carolina, 2006. Boston Books. 13

[42] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi: 10.1137/0916069. 42, 86

[43] Y. Cao, D. T. Gillespie, and L. R. Petzold. Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics*, 124(4):044109, 2006. 12

[44] Y. Cao, A. Terebus, and J. Liang. State space truncation with quantified errors for accurate solutions to discrete chemical master equation. *Bulletin of Mathematical Biology*, 78(4):617–661, 2016. doi: 10.1007/s11538-016-0149-1. 13

[45] C. Chancellor, A. Ammar, F. Chinesta, M. Magnin, and O. Roux. Linking discrete and stochastic models: The chemical master equation as a bridge between process hitting and proper generalized decomposition. In *Computational Methods in Systems Biology*, number 8130 in Lecture Notes in Computer Science, pages 50–63. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-40708-6_5. 13

[46] J. D. Chandler. Moment problems for compact sets. *Proceedings of the American Mathematical Society*, 104(4):1134–1140, 1988. doi: 10.1090/S0002-9939-1988-0942632-2. 36

[47] V. Chellaboina, S. Bhat, W. Haddad, and D. Bernstein. Modeling and analysis of mass-action kinetics. *IEEE Control Systems Magazine*, 29(4):60–78, 2009. doi: 10.1109/MCS.2009.932926. 4

[48] H. Choi and F. Jafari. Positive definite Hankel matrix completions and Hamburger moment completions. *Linear Algebra and its Applications*, 489:217–237, 2016. doi: 10.1016/j.laa.2015.10.008. 36

[49] E. Cobb and B. Harris. The characterization of the solution sets for generalized reduced moment problems and its application to numerical integration. *SIAM Review*, 8(1):86–99, 1966. doi: 10.1137/1008007. 35

[50] L. Cobb. The multimodal exponential families of statistical catastrophe theory. In C. Taillie, G. P. Patil, and B. A. Baldessari, editors, *Statistical Distributions in Scientific Work*, number 79 in NATO Advanced Study Institutes Series, pages 67–90. Springer Netherlands, 1981. doi: 10.1007/978-94-009-8549-0_4. 112

[51] L. Cobb, P. Koppstein, and N. H. Chen. Estimation and moment recursion relations for multimodal distributions of the exponential family. *Journal of the American Statistical Association*, 78(381):124–130, 1983. doi: 10.2307/2287118. 54, 55, 112, 113

[52] G. Corradi. Higher order derivatives in optimization methods. *Optimization*, 37(1): 41–49, 1996. doi: 10.1080/02331939608844195. 87

[53] T. M. Cover and J. A. Thomas. Elements of information theory. Wiley-Interscience, 2006. 38, 50

[54] A. Crudu, A. Debussche, and O. Radulescu. Hybrid stochastic simplifications for multiscale gene networks. *BMC Systems Biology*, 3(1):89, 2009. doi: 10.1186/1752-0509-3-89. 9, 10

[55] A. Dax. A modified Gram–Schmidt algorithm with iterative orthogonalization and column pivoting. *Linear algebra and its applications*, 310(1-3):25–42, 2000. doi: 10.1016/S0024-3795(00)00022-7. 104, 106

[56] T. Dayar, H. Hermanns, D. Spieler, and V. Wolf. Bounding the equilibrium distribution of Markov population models. *Numerical Linear Algebra with Applications*, 18(6):931–946, 2011. doi: 10.1002/nla.795. 11, 13

[57] A. Decarreau, D. Hilhorst, C. Lemaréchal, and J. Navaza. Dual methods in entropy maximization. Application to some problems in crystallography. *SIAM Journal on Optimization*, 2(2):173–197, 1992. doi: 10.1137/0802010. 40

[58] S. Derisavi, H. Hermanns, and W. H. Sanders. Optimal state-space lumping in Markov chains. *Information Processing Letters*, 87(6):309–315, September 2003. doi: 10.1016/S0020-0190(03)00343-0. 13, 62

[59] F. Didier, T. A. Henzinger, M. Mateescu, and V. Wolf. Fast adaptive uniformization of the chemical master equation. In *High Performance Computational Systems Biology*, pages 118–127. IEEE, 2009. doi: 10.1109/HiBi.2009.23. 7, 13

[60] J. Ding, N. H. Rhee, and C. Zhang. On polynomial maximum entropy method for classical moment problem. *Advances in Applied Mathematics and Mechanics*, 8(1): 117–127, 2016. doi: 10.4208/aamm.2014.m504. 41, 58

[61] K. N. Dinh and R. B. Sidje. Understanding the finite state projection and related methods for solving the chemical master equation. *Physical Biology*, 13(3):035003, 2016. doi: 10.1088/1478-3975/13/3/035003. 13, 97

[62] S. Dolgov and B. N. Khoromskij. Tensor-product approach to global time-space-parametric discretization of chemical master equation. *Preprint, Max-Planck-Institut für Mathematik in den Naturwissenschaften*, 68, 2012. 13

[63] V. G. Dovì, O. Paladino, and A. P. Reverberi. Some remarks on the use of the inverse Hessian matrix of the likelihood function in the estimation of statistical properties of parameters. *Applied Mathematics Letters*, 4(1):87–90, 1991. doi: 10.1016/0893-9659(91)90129-J. 84

[64] W. Dreyer, M. Junk, and M. Kunik. On the approximation of the Fokker-Planck equation by moment systems. *Nonlinearity*, 14(4):881, 2001. doi: 10.1088/0951-7715/14/4/314. 34

[65] N. Ebrahimi, E. S. Soofi, and R. Soyer. Multivariate maximum entropy identification, transformation, and dependence. *Journal of Multivariate Analysis*, 99(6):1217–1231, 2008. doi: 10.1016/j.jmva.2007.08.004. 39

[66] M. B. Elowitz. Stochastic gene expression in a single cell. *Science*, 297(5584): 1183–1186, 2002. doi: 10.1126/science.1070919. 5

[67] S. Engblom. Computing the moments of high dimensional solutions of the master equation. *Applied Mathematics and Computation*, 180(2):498–515, 2006. doi: 10.1016/j.amc.2005.12.032. 19, 26, 27

[68] S. Fan, Q. Geissmann, E. Lakatos, S. Lukauskas, A. Ale, A. C. Babtie, P. D. W. Kirk, and M. P. H. Stumpf. MEANS: python package for Moment Expansion Approximation, iNference and Simulation. *Bioinformatics*, 2016. doi: 10.1093/bioinformatics/btw229. 98

[69] J. Flusser, B. Zitova, and T. Suk. *Moments and moment invariants in pattern recognition*. John Wiley & Sons, 2009. 34

[70] J. Flusser, S. Farokhi, C. Hoschl, T. Suk, B. Zitova, and M. Pedone. Recognition of images degraded by Gaussian blur. *IEEE Transactions on Image Processing*, 25(2): 790–806, 2016. doi: 10.1109/TIP.2015.2512108. 34

[71] Z. Fox and B. Munsky. Stochasticity or noise in biochemical reactions. *arXiv*, 2017. 5, 13

[72] F. Fröhlich, P. Thomas, A. Kazeroonian, F. J. Theis, R. Grima, and J. Hasenauer. Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLOS Computational Biology*, 12(7):e1005030, 2016. doi: 10.1371/journal.pcbi.1005030. 82

[73] B. Fu, J. Zhou, Y. Li, G. Zhang, and C. Wang. Image analysis by modified Legendre moments. *Pattern Recognition*, 40(2):691–704, feb 2007. doi: 10.1016/j.patcog.2006.05.020. 117

[74] P. Gavriliadis and G. Athanassoulis. Moment information for probability distributions, without solving the moment problem, II: Main-mass, tails and shape approximation. *Journal of Computational and Applied Mathematics*, 229(1):7–15, 2009. doi: 10.1016/j.cam.2008.10.011. 43

[75] P. Gavriliadis and G. Athanassoulis. The truncated Stieltjes moment problem solved by using kernel density functions. *Journal of Computational and Applied Mathematics*, 236(17):4193–4213, 2012. doi: 10.1016/j.cam.2012.05.015. 35, 41

[76] K. R. Ghusinga, C. A. Vargas-Garcia, A. Lamperski, and A. Singh. Exact lower and upper bounds on stationary moments in stochastic biochemical systems. *Physical Biology*, 14(4):04LT01, 2017. doi: 10.1088/1478-3975/aa75c6. 27

[77] E. Giampieri, M. De Cecco, D. Remondini, J. Sedivy, and G. Castellani. Active degradation explains the distribution of nuclear proteins during cellular senescence. *PloS One*, 10(6):e0118442, 2015. doi: 10.1371/journal.pone.0118442. 24

[78] A. Giffin. From physics to economics: An econometric example using maximum relative entropy. *Physica A: Statistical Mechanics and its Applications*, 388(8): 1610–1620, 2009. doi: 10.1016/j.physa.2008.12.066. 87

[79] A. Giffin and A. Caticha. Updating probabilities with data and moments. In *AIP Conference Proceedings*, volume 954, pages 74–84. AIP Publishing, 2007. doi: 10.1063/1.2821302.

[80] A. Giffin, C. Cafaro, and S. A. Ali. Application of the maximum relative entropy method to the physics of ferromagnetic materials. *Physica A: Statistical Mechanics and its Applications*, 455:11–26, 2016. doi: 10.1016/j.physa.2016.02.069. 87

[81] J. Gill and G. King. What to do when your Hessian is not invertible: Alternatives to model respecification in nonlinear estimation. *Sociological Methods & Research*, 33(1), 2004. doi: 10.1177/0049124103262681. 51

[82] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4): 403–434, 1976. doi: 10.1016/0021-9991(76)90041-3. 3

[83] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977. doi: 10.1021/j100540a008. 4, 6, 7

[84] D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A Statistical Mechanics and its Applications*, 188:404–425, 1992. doi: 10.1016/0378-4371(92)90283-V. 6, 7

[85] D. T. Gillespie. The chemical Langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000. doi: 10.1063/1.481811. 5

[86] D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716, 2001. doi: 10.1063/1.1378322. 7

[87] D. T. Gillespie and L. R. Petzold. Improved leap-size selection for accelerated stochastic simulation. *The Journal of Chemical Physics*, 119(16):8229–8234, 2003. doi: 10.1063/1.1613254. 12

[88] D. T. Gillespie and L. R. Petzold. Numerical simulation for biochemical kinetics. In *System Modeling in Cellular Biology*, pages 331–354. The MIT Press, 2006. doi: 10.7551/mitpress/9780262195485.003.0016. 6

[89] C. A. Gómez-Uribe and G. C. Verghese. Mass fluctuation kinetics: Capturing stochastic effects in systems of chemical reactions through coupled mean-variance computations. *The Journal of Chemical Physics*, 126(2):024109, 2007. doi: 10.1063/1.2408422. 26

[90] D. Gonze, J. Halloy, and A. Goldbeter. Deterministic versus stochastic models for circadian rhythms. *Journal of biological physics*, 28(4):637–653, 2002. doi: 10.1023/A:1021286607354. 9

[91] H. Gotovac and B. Gotovac. Maximum entropy algorithm with inexact upper entropy bound based on Fup basis functions with compact support. *Journal of Computational Physics*, 228(24):9079–9091, 2009. doi: doi:10.1016/j.jcp.2009.09.011. 41, 100

[92] J. Goutsias and G. Jenkinson. Markovian dynamics on complex reaction networks. *Physics Reports*, 529(2):199–264, 2013. doi: 10.1016/j.physrep.2013.03.004. 40

[93] M. Griffith, T. Courtney, J. Peccoud, and W. H. Sanders. Dynamic partitioning for hybrid simulation of the bistable HIV-1 transactivation network. *Bioinformatics*, 22(22):2782–2789, 2006. doi: 10.1093/bioinformatics/btl465. 10

[94] R. Grima. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 136(15):154105, 2012. doi: 10.1063/1.3702848. 26, 28

[95] C. M. Grinstead and J. L. Snell. Introduction to probability. American Mathematical Society, 2 revised edition, 1997. 38

[96] D. Gross and D. R. Miller. The randomization technique as a modeling tool and solution procedure for transient Markov processes. *Operations Research*, 32(2): 343–361, 1984. doi: 10.1287/opre.32.2.343. 7, 13

[97] M. C. Guenther, A. Stefanek, and J. T. Bradley. *Moment Closures for Performance Models with Highly Non-linear Rates*, pages 32–47. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-36781-6_3. 26

[98] M. L. Guerriero and J. K. Heath. Computational modeling of biological pathways by executable biology. *Methods in enzymology*, 487:217–251, 2011. doi: 10.1016/B978-0-12-381270-4.00008-1. 9

[99] S. Guiasu. Maximum entropy condition in queueing theory. *The Journal of the Operational Research Society*, 37(3):293–301, 1986. doi: 10.2307/2582209. 34

[100] A. Gupta, C. Briat, and M. Khammash. A scalable computational framework for establishing long-term behavior of stochastic reaction networks. *PLOS Computational Biology*, 10(6):e1003669, 2014. doi: 10.1371/journal.pcbi.1003669. 11

[101] H. Gzyl, P. N. Inverardi, and A. Tagliani. Entropy and density approximation from Laplace transforms. *Applied Mathematics and Computation*, 265:225–236, 2015. doi: 10.1016/j.amc.2015.05.020. 39, 41

[102] W. M. Haddad, V. Chellaboina, and Q. Hui. Nonnegative and compartmental dynamical systems. Princeton University Press, 2010. 5

[103] H. Hamburger. Über eine erweiterung des Stieltjesschen Momentenproblems. *Mathematische Annalen*, 81(2):235–319, 1920. doi: 10.1007/BF01564869. 35

[104] W. Hao and J. Harlim. An Equation-By-Equation method for solving the multidimensional moment constrained maximum entropy problem. *arXiv*, 2017. 89

[105] J. Hasenauer, V. Wolf, A. Kazeroonian, and F. J. Theis. Method of conditional moments (MCM) for the chemical master equation. *Journal of Mathematical Biology*, 69(3):687–735, 2014. doi: 10.1007/s00285-013-0711-5. 1, 16, 17, 29, 31, 32, 66, 67

[106] C. Hauck, C. Levermore, and A. Tits. Convex duality and entropy-based moment closures: Characterizing degenerate densities. *SIAM Journal on Control and Optimization*, 47(4):1977–2015, 2008. doi: 10.1137/070691139. 40

[107] F. Hausdorff. Momentprobleme für ein endliches Intervall. *Mathematische Zeitschrift*, 16(1):220–248, 1923. doi: 10.1007/BF01175684. 35, 36

[108] K. Hausken and J. F. Moxnes. A closure approximation technique for epidemic models. *Mathematical and Computer Modelling of Dynamical Systems*, 16(6):555–574, 2010. doi: 10.1080/13873954.2010.496149. 26

[109] R. Heinrich and S. Schuster. *The Regulation of Cellular Systems*. Springer US, 1996. doi: 10.1007/978-1-4613-1161-4. 4

[110] T. A. Henzinger, M. Mateescu, and V. Wolf. Sliding window abstraction for infinite Markov chains. In *Computer Aided Verification*, volume 5643, pages 337–352. Springer Berlin Heidelberg, 2009. doi: 10.1007/978-3-642-02658-4_27. 13

[111] B. Hepp, A. Gupta, and M. Khammash. Adaptive hybrid simulations for multiscale stochastic reaction networks. *The Journal of Chemical Physics*, 142(3):034118, 2015. doi: 10.1063/1.4905196. 10, 12

[112] J. Hespanha. Moment closure for biochemical networks. In *3rd International Symposium on Communications, Control and Signal Processing*, pages 142–147, 2008. doi: 10.1109/ISCCSP.2008.4537208. 26, 98

[113] B. Honarvar, R. Paramesran, and C.-L. Lim. Image reconstruction from a complete set of geometric and complex moments. *Signal Processing*, 98:224–232, 2014. doi: 10.1016/j.sigpro.2013.11.037. 34, 117

[114] G. Horton, V. G. Kulkarni, D. M. Nicol, and K. S. Trivedi. Fluid stochastic Petri nets: Theory, applications, and solution techniques. *European Journal of Operational Research*, 105(1):184–201, 1998. doi: 10.1016/S0377-2217(97)00028-3. 10

[115] K. M. Hosny. Image representation using accurate orthogonal Gegenbauer moments. *Pattern Recognition Letters*, 32(6):795–804, 2011. doi: 10.1016/j.patrec.2011.01.006. 117

[116] M. Huang, J. He, and X. Guan. Probabilistic inference of fatigue damage propagation with limited and partial information. *Chinese Journal of Aeronautics*, 28(4):1055–1065, 2015. doi: 10.1016/j.cja.2015.06.017. 40

[117] L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1):134–139, 1918. doi: 10.2307/2331932. 26

[118] T. Jahnke and W. Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54(1):1–26, 2006. doi: 10.1007/s00285-006-0034-x. 13

[119] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106 (4):620–630, 1957. doi: 10.1103/PhysRev.106.620. 34, 37

[120] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939 – 952, 1982. doi: 10.1109/PROC.1982.12425. 37

[121] G. Jenkinson and J. Goutsias. Numerical integration of the master equation in some models of stochastic epidemiology. *PLOS ONE*, 7(5):e36160, 2012. doi: 10.1371/journal.pone.0036160. 97

[122] N. L. Johnson, S. Kotz, and N. Balakrishnan. Discrete multivariate distributions. Wiley-Interscience, 1 edition, 1997. 97

[123] R. Johnson and J. Shore. Which is the better entropy expression for speech processing: -S log S or log S? *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(1):129–137, 1984. doi: 10.1109/TASSP.1984.1164296. 34

[124] M. Junk. Maximum entropy for reduced moment problems. *Mathematical Models and Methods in Applied Sciences*, 10(7):1001–1025, 2000. doi: 10.1142/S0218202500000513. 36, 39

[125] R. Kalaba and A. Tishler. On the use of higher order derivatives in optimization using Lagrange's expansion. *Nonlinear Analysis: Theory, Methods & Applications*, 7(10):1149–1161, 1983. doi: 10.1016/0362-546X(83)90024-X. 87

[126] N. G. v. Kampen. A power series expansion of the master equation. *Canadian Journal of Physics*, 39(4):551–567, 1961. doi: 10.1139/p61-056. 26

[127] V. Kazeev, M. Khammash, M. Nip, and C. Schwab. Direct solution of the chemical master equation using quantized tensor trains. *PLoS Computational Biology*, 10(3): e1003359, 2014. doi: 10.1371/journal.pcbi.1003359. 97

[128] A. Kazeroonian, F. Fröhlich, A. Raue, F. J. Theis, and J. Hasenauer. CERENA: ChEmical REaction network analyzer - A toolbox for the simulation and analysis of stochastic chemical kinetics. *PLOS ONE*, 11(1):e0146732, 2016. doi: 10.1371/journal.pone.0146732. 26, 98

[129] T. R. Kiehl, R. M. Mattheyses, and M. K. Simmons. Hybrid simulation of cellular behavior. *Bioinformatics*, 20(3):316–322, 2004. doi: 10.1093/bioinformatics/btg409. 10

[130] Y. Kitamura and M. Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4):861–874, 1997. doi: 10.2307/2171942. 34

[131] R. Kleeman. Measuring dynamical prediction utility using relative entropy. *Journal of the Atmospheric Sciences*, 59(13):2057–2072, 2002. 34

[132] C. Kleiber and J. Stoyanov. Multivariate distributions and the moment problem. *Journal of Multivariate Analysis*, 113:7–18, 2013. doi: 10.1016/j.jmva.2011.06.001. 36

[133] E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach, and R. Herwig. Systems biology: A Textbook. Wiley-VCH Verlag GmbH & Co. KGaA, 1 edition, 2009. 4

[134] A. Kolmogoroff. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104:415–458, 1931. doi: 10.1007/BF01457949. 7

[135] M. Komorowski, J. Mikisz, and M. P. H. Stumpf. Decomposing noise in biochemical signaling systems highlights the role of protein degradation. *Biophysical Journal*, 104(8):1783–1793, 2013. doi: 10.1016/j.bpj.2013.02.027. 10

[136] K. Kormann and S. MacNamara. Error control for exponential integration of the master equation. *arXiv*, 2016. 16

[137] M. G. Krein and A. A. Nudelman. The Markov moment problem and extremal problems. American Mathematical Society, 1977. 36

[138] M. G. Krein and A. A. Nudelman. The Markov moment problem and extremal problems: ideas and problems of P. L. Cebysev and A. A. Markov and their further development. Translations of mathematical monographs. American Mathematical Society, 1977. 35

[139] I. Krishnarajah, A. Cook, G. Marion, and G. Gibson. Novel moment closure approximations in stochastic epidemics. *Bulletin of Mathematical Biology*, 67(4):855–873, 2005. doi: 10.1016/j.bulm.2004.11.002. 26

[140] I. Krishnarajah, G. Marion, and G. Gibson. Novel bivariate moment-closure approximations. *Mathematical Biosciences*, 208(2):621–643, 2007. doi: 10.1016/j.mbs.2006.12.002. 26

[141] P. Kügler. Moment fitting for parameter inference in repeatedly and partially observed stochastic biological models. *PLoS ONE*, 7(8), 2012. doi: 10.1371/journal.pone.0043001. 82

[142] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, California, 1951. University of California Press. 39

[143] U. Kummer, B. Krajnc, J. Pahle, A. K. Green, C. J. Dixon, and M. Marhl. Transition from stochastic to deterministic behavior in calcium oscillations. *Biophysical Journal*, 89(3):1603–1611, 2005. doi: 10.1529/biophysj.104.057216. 7, 9

[144] M. Kwiatkowska, G. Norman, and D. Parker. PRISM 4.0: Verification of probabilistic real-time systems. In *Proceedings of 23rd International Conference on Computer Aided Verification (CAV'11)*, volume 6806 of *LNCS*, pages 585–591. Springer, 2011. doi: 10.1007/978-3-642-22110-1_47. 62

[145] E. Lakatos, A. Ale, P. D. W. Kirk, and M. P. H. Stumpf. Multivariate moment closure techniques for stochastic kinetic models. *The Journal of Chemical Physics*, 143(9):094107, 2015. doi: 10.1063/1.4929837. 26

[146] M. Lapin, L. Mikeev, and V. Wolf. SHAVE: Stochastic hybrid analysis of markov population models. In *Proceedings of the 14th International Conference on Hybrid Systems: Computation and Control*, HSCC '11, pages 311–312. ACM, 2011. doi: 10.1145/1967701.1967746. 8, 13, 97, 98

[147] C. H. Lee, K.-H. Kim, and P. Kim. A moment closure method for stochastic reaction networks. *The Journal of Chemical Physics*, 130(13):134107, 2009. doi: 10.1063/1.3103264. 27

[148] A. Leier, M. Barrio, and T. T. Marquez-Lago. Exact model reduction with delays: closed-form distributions and extensions to fully bi-directional monomolecular reactions. *Journal of The Royal Society Interface*, 11(95):20140108, 2014. doi: 10.1098/rsif.2014.0108. 12

[149] K. Levenberg. A method for the solution of certain problems in least squares. *Quarterly of applied mathematics*, 2:164–168, 1944. 43

[150] D. Levin, Y. Peres, and E. Wilmer. *Markov Chains and Mixing Times*, volume 58. American Mathematical Society, 2008. doi: 10.1090/mbk/058. 50, 58

[151] G. Li and K. Zhang. A combined reliability analysis approach with dimension reduction method and maximum entropy method. *Structural and Multidisciplinary Optimization*, 43(1):121–134, 2010. doi: 10.1007/s00158-010-0546-2. 54

[152] B. G. Lindsay and P. Basak. Moments determine the tail of a distribution (but not much else). *The American Statistician*, 54(4):248–251, 2000. doi: 10.2307/2685775. 41

[153] A. Loinger, A. Lipshtat, N. Q. Balaban, and O. Biham. Stochastic simulations of genetic switch systems. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 75(2):021904, 2007. doi: 10.1103/PhysRevE.75.021904. 15

[154] A. J. Lotka and F. Weiling. Elements of mathematical biology. *Biometrische Zeitschrift*, 7(3):208–208, 1965. doi: 10.1002/bimj.19650070323. 10

[155] A. Lück and V. Wolf. Generalized method of moments for estimating parameters of stochastic reaction networks. *arXiv*, 2016. 81, 84

[156] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. doi: 10.1137/0111030. 43

[157] G. Mastroianni and G. Milovanovic. Interpolation processes: Basic theory and applications. Springer Science & Business Media, 2008. doi: 10.1007/978-3-540-68349-0. 117

[158] M. Mateescu, V. Wolf, F. Didier, and T. A. Henzinger. Fast adaptive uniformisation of the chemical master equation. *IET Systems Biology*, 4(6):441–452, 2010. doi: 10.1049/iet-syb.2010.0005. 13

[159] T. Matis and I. Guardiola. Achieving moment closure through cumulant neglect. *The Mathematica Journal*, 12, 2010. doi: 10.3888/tmj.12-2. 98

[160] MATLAB Users Guide. The Mathworks. *Inc., Natick, MA*, 5:333, 1998. 42

[161] D. A. McQuarrie, C. J. Jachimowski, and M. E. Russell. Kinetics of small systems. II. *The Journal of Chemical Physics*, 40(10):2914–2921, 1964. doi: 10.1063/1.1724926. 19

[162] L. R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *Journal of Mathematical Physics*, 25(8):2404, 1984. doi: 10.1063/1.526446. 36, 40, 101

[163] S. Menz, J. C. Latorre, C. Schtte, and W. Huisinga. Hybrid stochastic–deterministic solution of the chemical master equation. *Multiscale Modeling & Simulation*, 10(4): 1232–1262, 2012. doi: 10.1137/110825716. 9

[164] L. Mikeev, W. Sandmann, and V. Wolf. Efficient calculation of rare event probabilities in markovian queueing networks. In *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools*, pages 186–196. ICST, 2011. doi: 10.4108/icst.valuetools.2011.245597. 13

[165] M. Milev, P. N. Inverardi, and A. Tagliani. Moment information and entropy evaluation for probability densities. *Applied Mathematics and Computation*, 218(9): 5782–5795, 2012. doi: 10.1016/j.amc.2011.11.093. 36, 42, 114

[166] A. C. Miller and T. R. Rice. Discrete approximations of probability distributions. *Management Science*, 29(3):352–362, 1983. 54

[167] P. Milner, C. S. Gillespie, and D. J. Wilkinson. Moment closure approximations for stochastic kinetic models with rational rate laws. *Mathematical Biosciences*, 231(2): 99–104, 2011. doi: 10.1016/j.mbs.2011.02.006. 26, 98

[168] P. Milner, C. S. Gillespie, and D. J. Wilkinson. Moment closure based parameter inference of stochastic kinetic models. *Statistics and Computing*, 23(2):287–295, 2013. doi: 10.1007/s11222-011-9310-8. 82, 84

[169] R. M. Mnatsakanov. Hausdorff moment problem: Reconstruction of distributions. *Statistics & Probability Letters*, 78(12):1612–1618, 2008. doi: 10.1016/j.spl.2008.01.011. 36

[170] R. M. Mnatsakanov. Moment-recovered approximations of multivariate distributions: The Laplace transform inversion. *Statistics & Probability Letters*, 81(1):1–7, 2011. doi: 10.1016/j.spl.2010.09.011. 36

[171] R. M. Mnatsakanov and A. S. Hakobyan. Recovery of distributions via moments. In *Institute of Mathematical Statistics Lecture Notes - Monograph Series*, pages 252–265. Institute of Mathematical Statistics, 2009. 36

[172] R. M. Mnatsakanov and K. Sarkisian. A note on recovering the distributions from exponential moments. *Applied Mathematics and Computation*, 219(16):8730–8737, 2013. doi: 10.1016/j.amc.2013.02.057. 36

[173] A. P. A. v. Moorsel and K. Wolter. Numerical solution of non-homogeneous Markov processes through uniformization. In *Proceedings of the 12th European Simulation Multiconference on Simulation - Past, Present and Future*, pages 710–717. SCS Europe, 1998. 9

[174] J. J. Mor. The Levenberg-Marquardt algorithm: Implementation and theory. In *Numerical Analysis*, pages 105–116. Springer, Berlin, Heidelberg, 1978. doi: 10.1007/BFb0067700. 86

[175] B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 124(4):044104, 2006. doi: 10.1063/1.2145882. 13, 97

[176] B. Munsky and M. Khammash. A multiple time interval finite state projection algorithm for the solution to the chemical master equation. *Journal of Computational Physics*, 226(1):818–835, 2007. doi: 10.1016/j.jcp.2007.05.016. 13

[177] D. Ormoneit and H. White. An efficient algorithm to compute maximum entropy densities. Taylor & Francis, 1999. doi: 10.1080/07474939908800436. 40, 51, 54

[178] M. Pájaro, A. A. Alonso, I. Otero-Muras, and C. Vázquez. Stochastic modeling and numerical simulation of gene regulatory networks with protein bursting. *Journal of Theoretical Biology*, 421:51–70, 2017. doi: 10.1016/j.jtbi.2017.03.017. 13

[179] R. C. Papademetriou. Reconstructing with moments. In *11th IAPR International Conference on Pattern Recognition*, volume 3, pages 476–480. IEEE, 1992. doi: 10.1109/ICPR.1992.202028. 36, 117

[180] P. R. Parthasarathy and R. Sudhesh. Exact transient solution of a state-dependent birth-death process. *Journal of Applied Mathematics and Stochastic Analysis*, pages 1–16, 2006. doi: 10.1155/JAMSA/2006/97073. 13

[181] V. Preda and C. Balcau. Maxentropic reconstruction of some probability distributions with linear inequality constraints. In *Proceedings of the 7-th Balkan Conference on Operational Research*, pages 151–159, 2007. 39, 100

[182] J. Puchaka and A. M. Kierzek. Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophysical Journal*, 86(3):1357–1372, 2004. doi: 10.1016/S0006-3495(04)74207-1. 10

[183] M. Putinar and K. Schmüdgen. Multivariate determinateness. *arXiv*, 2008. 36

[184] A. Ranganathan. The Levenberg-Marquardt algorithm, 2004. 43

[185] B. M. Rao, D. A. Lauffenburger, and K. D. Wittrup. Integrating cell-level kinetic modeling into the design of engineered protein therapeutics. *Nature Biotechnology*, 23(2):191–194, 2005. doi: 10.1038/nbt1064. 4

[186] M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653–1660, 2002. 34

[187] J. Ruess. Minimal moment equations for stochastic models of biochemical reaction networks with partially finite state space. *The Journal of Chemical Physics*, 143 (24):244103, 2015. doi: 10.1063/1.4937937. 30, 32, 80

[188] J. Ruess and J. Lygeros. On the use of the moment equations for parameter inference, control and experimental design in stochastic biochemical reaction networks. In *Computational Methods in Systems Biology*, number 8130 in Lecture Notes in Computer Science, pages 1–4. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-40708-6_1. 82

[189] J. Ruess and J. Lygeros. Moment-based methods for parameter inference and experiment design for stochastic biochemical reaction networks. *ACM Transactions on Modeling and Computer Simulation*, 25(2):8:1–8:25, 2015. doi: 10.1145/2688906. 82, 84

[190] A. Rusu. A survey on the Hausdorff moment problem and entropy approach. *BALCOR 2011*, page 261, 2011. 40

[191] H. Salis and Y. Kaznessis. Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *The Journal of chemical physics*, 122(5): 54103, 2005. doi: 10.1063/1.1835951. 9, 12

[192] M. Schmidt. minFunc: unconstrained differentiable multivariate optimization in Matlab, 2005. http://www.cs.ubc.ca/schmidtm/Software/minFunc.html. 41, 42, 43, 87

[193] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087): 1007–1012, 2006. doi: 10.1038/nature04701. 34

[194] D. Schnoerr, G. Sanguinetti, and R. Grima. Comparison of different moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 143(18):185101, 2015. doi: 10.1063/1.4934990. 26, 27, 98

[195] V. Shahrezaei and P. S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008. doi: 10.1073/pnas.0803850105. 24

[196] A. Sherman. On Newton-iterative methods for the solution of systems of nonlinear equations. *SIAM Journal on Numerical Analysis*, 15(4):755–771, 1978. doi: 10.1137/0715050. 41, 86

[197] J. N. Siddall and Y. Diab. The use in probabilistic design of probability curves generated by maximizing the Shannon entropy function constrained by moments. *Journal of Engineering for Industry*, 97(3):843–852, 1975. doi: 10.1115/1.3438691. 110

[198] R. B. Sidje. Expokit: A software package for computing matrix exponentials. *ACM Transactions on Mathematical Software*, 24(1):130–156, 1998. doi: 10.1145/285861.285868. 97

[199] A. Singh and J. P. Hespanha. Stochastic hybrid systems for studying biochemical processes. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1930):4995–5011, 2010. doi: 10.1098/rsta.2010.0211. 10

[200] A. Singh and J. Hespanha. Lognormal moment closures for biochemical reactions. In *2006 45th IEEE Conference on Decision and Control*, pages 2063–2068, 2006. doi: 10.1109/CDC.2006.376994. 25, 26

[201] P. Smadbeck and Y. N. Kaznessis. Efficient moment matrix generation for arbitrary chemical networks. *Chemical Engineering Science*, 84:612–618, 2012. doi: 10.1016/j.ces.2012.08.031. 98

[202] P. Smadbeck and Y. Kaznessis. Stochastic model reduction using a modified Hill-type kinetic rate law. *The Journal of Chemical Physics*, 137(23):234109, 2012. doi: 10.1063/1.4770273. 12

[203] P. Smadbeck and Y. N. Kaznessis. A closure scheme for chemical master equations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(35):14261–14265, 2013. doi: 10.1073/pnas.1306481110. 26, 40

[204] E. S. Soofi, N. Ebrahimi, and M. Habibullah. Information distinguishability with application to analysis of failure data. *Journal of the American Statistical Association*, 90(430):657–668, 1995. doi: 10.2307/2291079. 112

[205] V. Sotiropoulos and Y. N. Kaznessis. Analytical derivation of moment equations in stochastic chemical kinetics. *Chemical Engineering Science*, 66(3):268–277, 2011. doi: 10.1016/j.ces.2010.10.024. 98

[206] R. Srivastava, L. You, J. Summers, and J. Yin. Stochastic vs. deterministic modeling of intracellular viral kinetics. *Journal of Theoretical Biology*, 218(3):309–321, 2002. 6

[207] W.-H. Steeb, F. Solms, and R. Stoop. Chaotic systems and maximum entropy formalism. *Journal of Physics A: Mathematical and General*, 27(12):L399, 1994. doi: 10.1088/0305-4470/27/12/001. 34

[208] T.-J. Stieltjes. Recherches sur les fractions continues. *Annales de la Facult des sciences de Toulouse : Mathmatiques*, 8(4):J1–J122, 1894. 35

[209] J. Stoyanov. Krein condition in probabilistic moment problems. *Bernoulli*, 6(5): 939–949, 2000. doi: 10.2307/3318763. 36

[210] M. J. Stutzer. Simple nonparametric approach to derivative security valuation. SSRN Scholarly Paper, Social Science Research Network, 1998. 34

[211] N. Sulaimanov and H. Koeppl. Graph reconstruction using covariance-based methods. *EURASIP Journal on Bioinformatics and Systems Biology*, 2016(1):19, 2016. doi: 10.1186/s13637-016-0052-y. 62

[212] J. D. Tamarkin, J. A. Shohat, and J. D. Tamarkin. The problem of moments. American mathematical society, 1943. 36

[213] K. Tanaka and A. Toda. Discretizing distributions with exact moments: Error estimate and convergence analysis. *SIAM Journal on Numerical Analysis*, 53(5): 2158–2177, 2015. doi: 10.1137/140971269. 36, 40, 54

[214] A. Tari. *Moments based bounds in stochastic models*. PhD thesis, Budapest University of Technology and Economics, Department of Telecommunications, 2005. 34, 36

[215] A. Tari, M. Telek, and P. Buchholz. *A simplified moment-based estimation method for extreme probabilities, infinite and positive cases*, volume 7. United Kingdom Simulation Society, 2006. 43, 107

[216] M. R. Teague. Image analysis via the general theory of moments. *JOSA*, 70(8): 920–930, aug 1980. doi: 10.1364/JOSA.70.000920. 117

[217] A. B. Templeman and L. Xingsi. Entropy duals. *Engineering Optimization*, 9(2): 107–119, 1985. doi: 10.1080/03052158508902506. 40

[218] M. Thattai and A. v. Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, 2001. doi: 10.1073/pnas.151588598. 5

[219] P. Thomas, H. Matuschek, and R. Grima. Computation of biochemical pathway fluctuations beyond the linear noise approximation using iNA. In *2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1–5, 2012. doi: 10.1109/BIBM.2012.6392668. 24

[220] P. Thomas. Systematic approximation methods for stochastic biochemical kinetics. *Doctoral dissertation*, 2015. 115

[221] P. Thomas and R. Grima. Approximate probability distributions of the master equation. *Physical Review E*, 92(1):012120, 2015. doi: 10.1103/PhysRevE.92.012120. 26

[222] P. Thomas, A. V. Straube, and R. Grima. Communication: limitations of the stochastic quasi-steady-state approximation in open biochemical reaction networks. *The Journal of Chemical Physics*, 135(18):181103, 2011. doi: 10.1063/1.3661156. 24

[223] P. Thomas, N. Popovi, and R. Grima. Phenotypic switching in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 111(19):6994–6999, 2014. doi: 10.1073/pnas.1400049111. 66

[224] A. A. Toda. Existence of a statistical equilibrium for an economy with endogenous offer sets. *Economic Theory*, 45(3):379–415, 2010. doi: 10.1007/s00199-009-0493-6. 34

[225] M. K. Transtrum and J. P. Sethna. Improvements to the Levenberg-Marquardt algorithm for nonlinear least-squares minimization. *arXiv*, 2012. 43

[226] I. Turek. A maximum-entropy approach to the density of states within the recursion method. *Journal of Physics C: Solid State Physics*, 21(17):3251, 1988. doi: 10.1088/0022-3719/21/17/014. 41, 103

[227] N. M. van Dijk. Uniformization for nonhomogeneous Markov chains. *Operations Research Letters*, 12(5):283–291, 1992. doi: 10.1016/0167-6377(92)90086-I. 9

[228] H. D. Vo and R. B. Sidje. An adaptive solution to the chemical master equation using tensors. *The Journal of Chemical Physics*, 147(4):044102, 2017. doi: 10.1063/1.4994917. 13

[229] V. Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–560, 1926. doi: 10.1038/118558a0. 10

[230] C. Wachinger and N. Navab. Entropy and Laplacian images: Structural representations for multi-modal registration. *Medical Image Analysis*, 16(1):1–17, 2012. doi: 10.1016/j.media.2011.03.001. 34

[231] H. Wagner, M. Moller, and K. Prank. COAST: Controllable approximative stochastic reaction algorithm. *The Journal of Chemical Physics*, 125(17):174104, 2006. doi: 10.1063/1.2361284. 10

[232] S. Waldherr and B. Haasdonk. Efficient parametric analysis of the chemical master equation through model order reduction. *BMC Systems Biology*, 6(1):1–12, 2012. doi: 10.1186/1752-0509-6-81. 12

[233] E. W. J. Wallace, D. T. Gillespie, K. R. Sanft, and L. R. Petzold. Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *IET systems biology*, 6(4):102–115, 2012. doi: 10.1049/iet-syb.2011.0038. 10

[234] E. W. J. Wallace. A simplified derivation of the linear noise approximation. *arXiv*, 2010. 10

[235] J. C. Wheeler. Modified moments and Gaussian quadratures. *Rocky Mountain Journal of Mathematics*, 4(2):287–296, 1974. doi: 10.1216/RMJ-1974-4-2-287. 54

[236] P. Whittle. On the use of the normal approximation in the treatment of stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 19(2): 268–281, 1957. 19, 26

[237] D. V. Widder. The Laplace transform. Princeton University Press, 1941. 36

[238] D. J. Wilkinson. Stochastic modelling for systems biology. CRC Press, 2 edition, 2011. doi: 10.1186/1475-925X-5-64. 3

[239] O. Wolkenhauer, M. Ullah, W. Kolch, and K.-H. Cho. Modeling and simulation of intracellular dynamics: Choosing an appropriate framework. *IEEE Transactions on Nanobioscience*, 3(3):200–207, 2004. doi: 10.1109/TNB.2004.833694. 4

[240] X. Wu. Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics*, 115(2):347–354, 2003. doi: 10.2139/ssrn.369680. 34, 40, 87

[241] X. Wu. A weighted generalized maximum entropy estimator with a data-driven weight. *Entropy*, 11(4):917–930, 2009. doi: 10.3390/e11040917. 40

[242] X. Wu. Exponential series estimator of multivariate densities. *Journal of Econometrics*, 156(2):354–366, 2010. doi: 10.1016/j.jeconom.2009.11.005. 39

[243] Z. Wu, G. N. Phillips, Jr., R. Tapia, and Y. Zhang. A fast Newton algorithm for entropy maximization in phase determination. *SIAM Review*, 43(4):623–642, 2001. doi: 10.1137/S0036144500371737. 41, 101

[244] B. Xiao, J.-F. Ma, and X. Wang. Image analysis by Bessel–Fourier moments. *Pattern Recognition*, 43(8):2620–2629, 2010. doi: 10.1016/j.patcog.2010.03.013. 117

[245] B. Yang and M. Dai. Image analysis by Gaussian–Hermite moments. *Signal Processing*, 91(10):2290–2303, 2011. doi: 10.1016/j.sigpro.2011.04.012. 117

[246] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl. Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences*, 109(21):8340–8345, 2012. doi: 10.1073/pnas.1200161109. 82

[247] C. Zechner, F. Wadehn, and H. Köppl. Sparse learning of markovian population models in random environments. In *Proceedings of the 19th IFAC World Congress*, volume 19, pages 1723–1728. 19th World Congress The International Federation of Automatic Control, 2014. doi: 10.3929/ethz-a-010362110. 5

[248] A. Zellner and R. A. Highfield. Calculation of maximum entropy distributions and approximation of marginal posterior distributions. *Journal of Econometrics*, 37(2): 195–209, 1988. doi: 10.1016/0304-4076(88)90002-4. 51, 113

[249] J. Zhang, Q. Nie, and T. Zhou. A moment-convergence method for stochastic analysis of biochemical reaction networks. *The Journal of Chemical Physics*, 144 (19):194109, 2016. doi: 10.1063/1.4950767. 26

[250] J. Zhang, L. DeVille, S. Dhople, and A. D. Domínguez-García. A maximum entropy approach to the moment closure problem for stochastic hybrid systems at equilibrium. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 747–752. IEEE, 2014. doi: 10.1109/CDC.2014.7039471. 40, 114

[251] J. X. Zhou, L. Brusch, and S. Huang. Predicting pancreas cell fate decisions and reprogramming with a hierarchical multi-attractor model. *PLOS ONE*, 6(3):e14752, 2011. doi: 10.1371/journal.pone.0014752. 28