THE UNIVERSITY
of EDINBURGH

Approximation methods and inference for stochastic biochemical kinetics

David Schnoerr

Doctor of Philosophy
University of Edinburgh
2016

Abstract

Recent experiments have shown the fundamental role that random fluctuations play in many chemical systems in living cells, such as gene regulatory networks. Mathematical models are thus indispensable to describe such systems and to extract relevant biological information from experimental data. Recent decades have seen a considerable amount of modelling effort devoted to this task. However, current methodologies still present outstanding mathematical and computational hurdles. In particular, models which retain the discrete nature of particle numbers incur necessarily severe computational overheads, greatly complicating the tasks of characterising statistically the noise in cells and inferring parameters from data. In this thesis we study analytical approximations and inference methods for stochastic reaction dynamics.

The chemical master equation is the accepted description of stochastic chemical reaction networks whenever spatial effects can be ignored. Unfortunately, for most systems no analytic solutions are known and stochastic simulations are computationally expensive, making analytic approximations appealing alternatives. In the case where spatial effects cannot be ignored, such systems are typically modelled by means of stochastic reaction-diffusion processes. As in the non-spatial case an analytic treatment is rarely possible and simulations quickly become infeasible. In particular, the calibration of models to data constitutes a fundamental unsolved problem.

In the first part of this thesis we study two approximation methods of the chemical master equation; the chemical Langevin equation and moment closure approximations. The chemical Langevin equation approximates the discrete-valued process described by the chemical master equation by a continuous diffusion process. Despite being frequently used in the literature, it remains unclear how the boundary conditions behave under this transition from discrete to continuous variables. We show that this boundary problem results in the chemical Langevin equation being mathematically ill-defined if defined in real space due to the occurrence of square roots of negative expressions. We show that this problem can be avoided by extending the state space from real to complex variables. We prove that this approach gives rise to real-valued moments and thus admits a probabilistic interpretation. Numerical examples demonstrate better accuracy of the developed complex chemical Langevin equation than various real-valued implementations proposed in the literature.

Moment closure approximations aim at directly approximating the moments of a process, rather then its distribution. The chemical master equation gives rise to an

infinite system of ordinary differential equations for the moments of a process. Moment closure approximations close this infinite hierarchy of equations by expressing moments above a certain order in terms of lower order moments. This is an ad hoc approximation without any systematic justification, and the question arises if the resulting equations always lead to physically meaningful results. We find that this is indeed not always the case. Rather, moment closure approximations may give rise to diverging time trajectories or otherwise unphysical behaviour, such as negative mean values or unphysical oscillations. They thus fail to admit a probabilistic interpretation in these cases, and care is needed when using them to not draw wrong conclusions.

In the second part of this work we consider systems where spatial effects have to be taken into account. In general, such stochastic reaction-diffusion processes are only defined in an algorithmic sense without any analytic description, and it is hence not even conceptually clear how to define likelihoods for experimental data for such processes. Calibration of such models to experimental data thus constitutes a highly non-trivial task. We derive here a novel inference method by establishing a basic relationship between stochastic reaction-diffusion processes and spatio-temporal Cox processes, two classes of models that were considered to be distinct to each other to this date. This novel connection naturally allows to compute approximate likelihoods and thus to perform inference tasks for stochastic reaction-diffusion processes. The accuracy and efficiency of this approach is demonstrated by means of several examples.

Overall, this thesis advances the state of the art of modelling methods for stochastic reaction systems. It advances the understanding of several existing methods by elucidating fundamental limitations of these methods, and several novel approximation and inference methods are developed.

Lay Summary

The number of molecules of many chemical species in living cells, such as proteins, vary randomly in time. Such random fluctuations have been found to play a crucial role in many biological processes in living cells. An important example is the process of gene expression, which denotes the production of proteins from genes. Gene expression is a fundamental process in biology, because it enables cells to transform static information encoded in the DNA into dynamically varying protein concentrations which in turn enables cells to respond to and process external stimuli. Experiments have shown that random fluctuations in gene expression can lead to dramatically differing behaviours of genetically identical cells. Accordingly, recent years have seen a rapid improvement of experimental techniques that allow to measure such effects. In order to be able to interpret the experimental data and to understand the underlying processes, however, mathematical methods are needed that allow to model such processes. In this thesis, we study several existing methods and develop new ones for modelling the effect of random fluctuations in biological systems.

Chemical reactions between small populations of molecules are the major source of random fluctuations in cells. As these molecules move randomly in the cell driven by Brownian diffusion, reactions occur randomly with certain probabilities whenever two reactant molecules collide. Under certain conditions the spatial positions of molecules can be ignored and the system dynamics can be described by simpler models.

In the first part of this thesis, we study two such methods: the chemical Langevin equation and moment closure approximations. We find that both methods have severe mathematical problems and show that they can give rise to unphysical results. Moment closure approximations, for example, may lead to negative molecule numbers, which clearly does not make sense physically. For the chemical Langevin equation, we propose a modified version and show that this new method does not suffer from these problems. Moreover, we find that the developed method is more accurate than other versions proposed in the literature for several example systems.

In the second part of this work we consider the case where spatial positions of molecules have to be taken into account. Such systems are typically modelled as stochastic reaction-diffusion processes. We consider the problem of estimating unknown parameters of a model from experimental data. Our main result is the derivation of a novel connection between stochastic reaction-diffusion processes and so-called Cox processes. The latter constitute a popular class of models from statistics used to describe realisations of random spatio-temporal processes. The established connection

provides the basis for a novel method for parameter estimation for stochastic reaction-diffusion processes. We demonstrate the accuracy and efficiency of our method for several examples.

Acknowledgements

First of all I would like to thank my supervisors Ramon Grima and Guido Sanguinetti for their outstanding supervision of my PhD project. Both of them warmly welcomed and introduced myself to the field. They always took the time for intensive discussions and thus guided my research with many creative ideas, while giving me enough freedom to develop and realise my own ideas. I appreciate very much their excellent support and cooperation.

I thank Peter Swain and Nikola Popovic not only for joining my PhD committee, but particularly for many interesting discussions and the instructive feedback.

I would like to thank the BBSRC for funding this PhD project, which gave me a lot of freedom and enabled me to participate in several conferences and summer schools.

My special thanks go to Philipp Thomas for countless hours of discussions and refreshing breaks from work. There were almost no questions on stochastic processes that he could not help me with and I learned a lot from him.

I owe particular thanks to Botond Cseke for an exciting collaboration on expectation propagation and the resulting publication. But more than that, I was always able to ask Botond for advice on inference methods. He taught me a lot.

Claudia Cianci deserves my special thanks for a fruitful collaboration and the resulting publication, as well as for refreshing distractions from work in various coffee and lunch breaks.

I would like to thank further former and current lab members for sharing their knowledge with me and for creating a pleasant working atmosphere, including Daniel Trejo Banos, Ronald Begg, Giulio Caravagna, Anastasis Georgoulas, Abhishek Gupta, Yuanhua Huang, Van Anh Huynh-Thu, Harriet Jones, Andreas Kapouranis, Tom Mayo, Michalis Michaelides, Dimitrios Milios, Andrea Ocone, Gabriele Schweikert, Alina Selega, Stephen Smith and Edward Wallace.

Special thanks go to my close friends in Edinburgh for being extremely supportive and making my time in Edinburgh an exciting experience, particularly Jenny Lawy, Leila

Bright, Marina Calvo, Karina Banda, Adrian Martinez, Marisa de Andrade and Elco Bakker.

My flatmates Cameron Malcom and Mary-Jane Fitzsimons deserve my special thanks for supporting me during my PhD in various ways. They made me feel home in our flat and regularly provided pleasant distractions from work.

I am very happy to have the best friends in the world back home, and I would like to thank all of them for the support I received and for the exciting holidays I had whenever I was back in Heidelberg, especially Wayne Götz, Ann-Katrin Unglert, Fabian Bläß, Lionel Evrard, Bernhard Schmitzer, Fabian Bergermann, Mareike Janßen, Franziska Hentzschel, Konrad Heimpel, Lena Flecken and Julia Patzelt.

I am particularly thankful for having such amazing parents, Barbara and Christoph. They were extremely supportive throughout my PhD and helped me in many ways. Particular thanks go to the coolest siblings in the world, Florian, Anna Lena and Johannes, and also to Maya for supporting me and for making sure I had relaxing holidays whenever I was back home. Last but not least I would like to thank my grandparents, my aunts, my uncles and my cousins who contributed as well to making this PhD project a successful experience.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

This thesis contains work of the following publications:

- **David Schnoerr**, Ramon Grima, and Guido Sanguinetti. "Cox process representation and inference for stochastic reaction-diffusion processes." *Nature Communications* 7:11729, 2016.
- **David Schnoerr**, Guido Sanguinetti, and Ramon Grima. "Comparison of different moment-closure approximations for stochastic chemical kinetics." *The Journal of chemical physics*, 143(18):185101, 2015.
- **David Schnoerr**, Guido Sanguinetti, and Ramon Grima. "Validity conditions for moment closure approximations in stochastic chemical kinetics." *The Journal of chemical physics*, 141(8):084103, 2014.
- **David Schnoerr**, Guido Sanguinetti, and Ramon Grima. "The complex chemical Langevin equation." *The Journal of chemical physics*, 141(2):024103, 2014.

The major part of research and the major part of writing of all four publications has been conducted by the candidate. The two co-authors contributed by guiding the research and the publication writing process, as well as giving feedback and corrections for the publications.

(*24/07/2016, David Schnoerr*)

List of Abbreviations

CME	chemical master equation
SSA	stochastic simulation algorithm
MA	moment closure approximation
CLE	chemical Langevin equation
CFPE	chemical Fokker-Planck equation
RE	rate equations
LNA	linear noise approximation
PR	Poisson representation
SRDP	stochastic reaction-diffusion process
RDME	reaction-diffusion master equation

Contents

Chapter 1

Introduction

1.1 Motivation and Overview

Understanding the functioning of living cells and biological organisms at the system level has gained increasing focus in recent years and defines a key research programme for the next decades [1]. Experimental techniques are developing at breathtaking speed producing a wealth of data at finer and finer resolutions. Biologists are now for instance able to obtain time-course data of certain components in single cells simultaneously across hundreds of cells [2]. An important example are fluorescent mRNA or protein molecules. Even more than that, scientists are now able to modify living cells, either genetically or otherwise, to gain deeper insights into their functioning [3] or to produce specific desirable behaviours [4, 5].

However, such experimental data does not by itself reveal the function of such biological systems. The underlying processes typically involve large numbers of interacting components giving rise to highly complex behaviour. Moreover, experimental data are generally corrupted by measurement noise and incomplete, thus posing the mathematical and statistical challenge to infer the relevant biological information from such measurements.

We focus here on mathematical and statistical modelling of biological systems in which *random fluctuations* of molecule numbers play an important role. Recent experiments have shown this to be the case in many processes, gene expression being a prominent example. One of the first experiments studying such random fluctuations was performed by Elowitz et al. in 2002 [6]. The authors observed a large heterogeneity in the expression levels of certain genes across an *Escherichia coli* colony. They found that these differences between different cells originate to a large extent from

random, or *intrinsic*, fluctuations of the protein numbers in time, rather than only from physiological differences between cells, which one may naively expect. Since then scientists have found that the role of intrinsic noise is crucial for many biological processes. Not only can intrinsic noise drastically change the quantitative behaviour of systems, but it also gives rise to novel qualitative phenomena, such as noise-induced oscillations [7], which cannot be explained in a deterministic setting. While under certain conditions the spatial distribution of the components in the cell can be ignored, which significantly reduces the complexity of possible models, these conditions are often not met in reality, making spatial descriptions indispensable. In [8], for example, the authors showed that a certain protein in *Escherichia coli* undergoes spatial stochastic switching between two sides of the cell, thereby directing the location of cell division. Such effects can clearly only be described by spatial models.

The question arises how one can mathematically model the effects of random fluctuations in chemical reaction networks. Optimally, one would like to know the probability distribution of the system to be in a certain state for all times. As it turns out, under certain conditions for which spatial effects can be ignored, an exact analytic equation exists governing the time evolution of this probability distribution: the *chemical master equation* (CME). By solving the CME, one may thus in principle extract all the relevant information about the dynamics of a given system. In practice, however, solving the CME is a difficult task, and analytic solutions are known only for the most simple systems.

One possibility to circumvent this problem is to use stochastic simulations. The latter exactly simulate the underlying stochastic process, thereby drawing exact samples from the CME's solution. Unfortunately, however, this procedure is computationally expensive and quickly becomes infeasible for larger systems and/or inference tasks. Significant effort has thus been spent in the development of approximation methods to the CME. We study two such approximation methods in this thesis. One class of methods approximates the actual stochastic process underlying the CME. This in turn leads to approximations of the probability distribution. An example which we study here is the *chemical Langevin equation* (CLE). Another class of methods aims at directly approximating the statistical moments of the CME's solution rather than the full process. One example are so-called *moment closure approximations* (MAs). We study a certain class of MAs in this work.

The *chemical Langevin equation* (CLE) and its corresponding *chemical Fokker-Planck equation* (CFPE) constitute a popular diffusion approximation of the CME [9].

It approximates the discrete process underlying the CME by a continuous diffusion process. The continuous variables describing this process satisfy a coupled set of stochastic differential equations, which are often more efficient to simulate than the CME. The CLE has been used successfully in the literature [10, 11] and is subject of ongoing research [12, 13].

However, it remains unclear how the boundary behaviour of the CME translates to the CLE formalism. The CME has a natural boundary at zero molecule numbers, which manifests itself in the probability to find a negative molecule number to be zero for all times. To this date, it is not clear how this boundary behaves when approximating the discrete process by a continuous one in the CLE formalism. As it turns out, this approximation leads to fundamental problems of the CLE due to the occurrence of square roots of negative expression for which the CLE is not defined since it is only defined for real numbers. We term this problem the *breakdown* of the CLE. Despite being noticed in the literature [14, 15], the breakdown has not been studied thoroughly up to date, and several open questions remain.

First of all, there exist different representations of the CLE, and it is not clear if a representation can be chosen that avoids breakdown, or if the breakdown is actually intrinsic to the CLE and thus independent of the chosen representation. If the latter is the case, one may want to develop modified versions of the CLE that avoid breakdown. Several such modifications of the CLE have been proposed in the literature that aim at keeping the CLE real-valued, such as imposing artificial boundaries enforcing positivity or omitting some noise terms. However, all these methods are ad hoc without a microscopic justification, and it remains unclear how such modifications influence the accuracy of the CLE. The question hence arises if there exists a more natural, microscopically justified version of the CLE that avoids breakdown.

The second type of approximation method that we study in this work are *moment closure* approximations (MAs) [16, 17]. MAs aim at approximating the first few moments of the solution of the CME. The latter gives rise to an infinite hierarchy of coupled ordinary differential equations for the moments. The class of MAs that we consider in this thesis close this infinite hierarchy of equations by expressing all moments above a certain order in terms of lower order moments. This is an ad hoc approximation lacking a systematic justification, and hence little is known about the validity and accuracy of MAs. While there exist studies of the quantitative accuracy of MAs [18, 19], a more fundamental question which has not been studied to-date is: when can we trust MAs to lead to physically meaningful results? By “physically meaningful”

we mean positive mean concentrations and positive even central moments of the fluctuations in molecule numbers. We perform a numerical case study in this work that aims at answering this question.

As mentioned before, the conditions underlying the CME description are often not fulfilled in reality, and spatial positions and movement of molecules have to be taken into account. Such systems are then typically modelled by *stochastic reaction-diffusion processes* (SRDPs), in which each single molecule is simulated in space and time, undergoing Brownian diffusion and with chemical reactions happening under certain rules when molecules collide. Simulations of such systems are computationally expensive and quickly become infeasible for larger systems. In contrast to the non-spatial case and its CME description discussed before, there generally does not exist an analytic description of SRDPs. It is therefore not even conceptually clear how to define a likelihood for such processes, making the calibration of models to data extremely difficult. SRDPs are not only used for chemical kinetics, but in various other scientific fields, such as ecology [20] or social sciences [21]. With an increasing amount of data becoming available in many of these fields, the calibration of models to data constitutes an outstanding problem. The development of an efficient and accurate inference method for such processes is one of the goals of this thesis.

1.2 Related work

The chemical Langevin equation

The CFPE was first derived by Kramers [22] and Moyal [23] by means of the so-called *Kramers Moyal expansion*. A more intuitive approach, which leads directly to the CLE, was given seven decades later by Gillespie [9]. The breakdown of the CLE due to the occurrence of square roots of negative expressions has first been studied in [24] and [14]. The authors show for some simple reaction systems that the CLE can lead to negative molecule numbers and may break down. The authors do not study different forms of the CLE, however, and the question thus remains if different representations of the CLE may avoid breakdown. Wilkie and Wong proposed a modified version of the CLE, which avoids breakdown by globally removing some noise terms [24]. Using such a global modification for fixing a local boundary problem is criticised in [15], where the authors show that this modification leads to significant quantitative inaccuracies for one example system. Dana and Raha proposed another modified ver-

sion [25]. Here breakdown is avoided by setting a noise term to zero whenever the expression in its corresponding square root becomes negative. While this local modification seems more appropriate than the global one proposed in [24], it still lacks a microscopic justification.

Moment closure approximations

The idea of MA methods goes back more than six decades [16, 17, 26], but has only been popularised in the context of biochemical reaction systems in recent years [27, 28]. These references all use what we call here the *normal moment closure*. Several other MA methods have been proposed in the literature, such as the here studied *log-normal* MA [29], the *Poisson* MA [30] and *central-moment-neglect* (CMN) MA [31]. There exist few studies that investigate the quantitative accuracy of MA methods, including [18, 19]. However, to our knowledge, no studies are available that consider the validity of MAs, i.e., if they always give physically meaningful results that admit a probabilistic interpretation.

Stochastic reaction-diffusion processes

SRDPs were originally introduced in the context of coagulation of colloids by Smoluchowski in 1916 [32, 33]. The Doi picture of SRDPs, which we will adopt in this work, has been developed by Doi in 1976 [34, 35]. As far as we are aware, the few attempts at statistical inference for SRDPs either used simulation-based likelihood free methods [20], inheriting the intrinsic computational difficulties, or abandoned the SRDP framework by adopting a coarse space discretisation [36], or neglecting the individual nature of molecules using a linear noise approximation [37].

1.3 Contribution

The complex chemical Langevin equation

Our results on the breakdown problem of the CLE presented in this thesis are based on reference [38]. Our first result is that the breakdown cannot be avoided for most chemical reaction systems independently of the choice of the CLE's noise matrix. This means that the CLE as presented in the literature is *mathematically ill-defined for most reaction systems*.

A fundamental property of the CLE is that its predictions of the *mean and variance agree exactly with the corresponding CME results for linear reaction systems*. We show here that naive implementations imposing artificial boundaries to enforce positivity, which is typically done in the literature, *violate this agreement for linear systems*. We find that the same is true for the two modified versions of the CLE mentioned above.

Next, we show that the problem of breakdown can be avoided by *extending the CLE's state space from real to complex numbers*. This novel *complex CLE* does not suffer from breakdown, and we prove that its *moments are real-valued* and hence admit a probabilistic interpretation. Moreover, we show that the complex CLE gives rise to real-valued autocorrelation functions and power spectra, and that it allows to compute first-passage times. We further show that the complex CLE *restores the CLE's exactness for mean and variance for linear reaction systems*. We find that it gives significantly more accurate results than the different real-valued versions mentioned above, as well as the second-order normal MA and the so-called *linear noise approximation*, for two non-linear example systems.

Our results thus suggest that the different real-valued CLEs should be used with care and that the complex CLE developed here constitutes a useful alternative tool for the study of stochasticity in chemical reaction systems.

Validity of moment closure approximations

Our analysis of the validity of MAs is based on references [39] and [40]. First, we formulate a set of validity conditions for MAs which guarantee physically meaningful results. Next, we give a detailed study of the second-order normal MA for a deterministically monostable system with one species. By "deterministically monostable" we mean that the deterministic rate equations possess one positive stable fixed point. We find that the normal MA indeed fails to give physically meaningful results whenever the system size is small enough. We find that this is also true for the other studied MAs and for other deterministically monostable systems: the MAs *fail to give physically meaningful results below a certain critical system volume*. Below this critical volume the MAs may give rise to *divergent trajectories* or *negative means or variances*, and hence *do not admit a probabilistic interpretation*.

Next, we study the second-order normal, Poisson, log-normal and central-moment-neglect (CMN) MAs with respect to their validity and their quantitative accuracy (as compared to exact CME results) for three more complex systems: a deterministically

bistable system, i.e., a system whose rate equations have two positive stable fixed points, a deterministically ultrasensitive and a deterministically oscillatory system. For all three systems, we find that, as for the previously studied deterministically monostable system, the MAs fail to give physically meaningful results below a certain critical volume. Moreover, we find for the deterministically bistable and deterministically oscillatory systems that *the MAs are valid only for an intermediate range of system sizes*, with unphysical multistability and unphysical oscillations, respectively, for large system sizes. For the deterministically ultrasensitive system, we find that the *Poisson and log-normal MAs are not uniquely defined*, a flaw not shared by the normal and CMN-MAs. We show that this non-uniqueness is due to a conservation law in particle numbers. We find that the wrong choice of the different versions violates certain symmetries of the system, while the normal and CMN-MAs conserve these symmetries.

Overall, we find that the normal MA has a higher range of validity than the other MAs, and that it is always uniquely defined in contrast to the Poisson and log-normal MAs. In terms of quantitative accuracy as compared to the exact CME results, we find no significant difference between the different MA methods. Our results thus suggest that the normal MA is favourable over the other methods.

We present here also the Mathematica package *MOCA* which allows the automated derivation and numerical analysis of the four studied MA methods in a graphical user interface. It is applicable to non-polynomial and time-dependent reaction rates and thus applicable to virtually any reaction system. Moreover, it allows the user to define novel MA methods. We believe that these features make MOCA a useful software tool for the chemical kinetics community.

Inference for stochastic reaction-diffusion processes

Reference [41] is the basis of our results on inference methods for SRDPs. As we mentioned before, inference for SRDPs is a highly non-trivial task and barely any methods exist in the literature. We develop a novel inference method in this work by establishing a novel connection between SRDPs and so-called *spatio-temporal point processes*.

Our starting point is the simple observation that, at any fixed time point, an SRDP creates random point patterns in space. Consequently, if we measure an SRDP at discrete time points, we obtain a sequence of random spatial point patterns. *Spatio-temporal point processes*, a class of models widely used for inference tasks in statistics, describe exactly such time-sequences of spatial point patterns. The question thus arises

if there exists a connection between these two types of models. In the literature, they have been considered distinct to each other so far. However, as it turns out, there indeed exists such a connection between SRDPs and so-called *Cox processes*, the latter being a particularly popular type of point processes.

By using the so-called *Poisson representation* and reaction-diffusion master equation, we are able to show that the single time probability distributions of a certain class of SRDPs are *exactly the same as the ones of a Cox process*, and that the same is true for other SRDPs in an approximate sense. We derive a stochastic partial differential equation for the intensity of the corresponding Cox process. To our knowledge, we derive here for the first time a mathematical connection between Cox processes and SRDPs. Since the Cox process likelihood is a well-known object, we can use the derived connection to (approximately) compute single-time-point likelihoods.

By applying a certain type of mean-field approximation, we can go one step further and use the established connection to *approximate multi-time-point likelihoods for SRDPs*. This in turn allows us to use statistical inference methods for Cox processes to perform inference for SRDPs. We find our approach to be highly accurate and efficient for several examples.

1.4 Organisation

This thesis is structured as follows. Chapter 2 presents basic mathematical concepts that form the basis of the subsequent chapters. We start by reviewing the origins and importance of intrinsic noise in biological systems in Section 2.1. Next, we give a brief introduction to the theory of stochastic processes in Section 2.2 with particular focus on Markov jump and diffusion processes. Next, we introduce the concept of stochastic chemical kinetics in a non-spatial setting in Section 2.3. We introduce the CME and the stochastic simulation algorithm, as well as several approximation methods of the CME, including the CLE and MAs, and introduce the Poisson representation. In Section 2.4, we then relax the conditions underlying the non-spatial CME description and introduce SRDPs as a spatial description of chemical reaction systems. After giving a microscopic description of SRDPs we present the reaction-diffusion master equation as a coarse-grained approximation of SRDPs. Next, we introduce Poisson and Cox processes and discuss some of their properties in Section 2.5.

Chapter 3 focuses on the CLE. First, we analyse the breakdown problem of the real-valued CLE in Section 3.1. Next, we derive the complex CLE and prove its basic

properties in Section 3.2. We then study its accuracy and compare it to different real-valued CLEs for two example systems in Section 3.3.

In Chapter 4 we analyse the validity of MA methods. We start by stating the general problem in Section 4.1. After formulating validity conditions for MAs that guarantee physically meaningful results in Section 4.2, we analyse the normal MA with respect to these conditions for a simple one-species system in Section 4.3. A numerical comparison of the normal, Poisson, log-normal and CMN-MAs for three more complex systems is then conducted in Section 4.4. Finally, we present the software package MOCA in Section 4.5.

Chapter 5 is devoted to the development of an efficient and accurate inference method for SRDPs. First, we classify different reaction types with respect to the behaviour of their Poisson representation in Section 5.2. Next, we derive approximations for reaction types whose Poisson representation is complex-valued, leading to a real-valued Poisson representation for all reaction systems. We then apply this real-valued Poisson representation to the reaction-diffusion master equation and subsequently take the continuum limit in Section 5.3. In Section 5.4 we show that this amounts to a Cox process representation of SRDPs which in turn allows us to derive a novel inference method for SRDPs. We demonstrate the accuracy and efficiency of our approach for three example systems in Section 5.5.

Finally in Chapter 6 we conclude and point out possible future research directions. In order not to disrupt reading, some technical derivations are presented in the appendix.

Chapter 2

Preliminaries

In this chapter we introduce some well-known methods and concepts that our results in the following chapters will be based on. First, we explain how stochastic effects arise in biological systems in Section 2.1. We then give a brief mathematical introduction to the theory of stochastic processes in Section 2.2. Next, in Section 2.3 we show how stochastic processes can be used to model stochasticity in chemical reaction systems and discuss the most general non-spatial description of such systems, the *chemical master equation*. We present several approximation methods of the latter, as well as an alternative formulation, before treating the more general, spatial case of *stochastic reaction-diffusion processes* in Section 2.4. Finally, in Section 2.5, we introduce *spatio-temporal point processes*, a class of models from statistics which seems to be unrelated to the previously mentioned fields, but which will prove to be crucial for our analysis of stochastic reaction-diffusion processes later on.

2.1 Stochasticity in biological systems

Stochastic effects play an important role in many chemical reaction networks in living cells. Examples are enzyme catalysed processes, transduction of external signals to the interior of cells or the process of gene expression, to name just a few. We explain here the emergence and importance of stochasticity in biochemical networks on the example of gene expression.

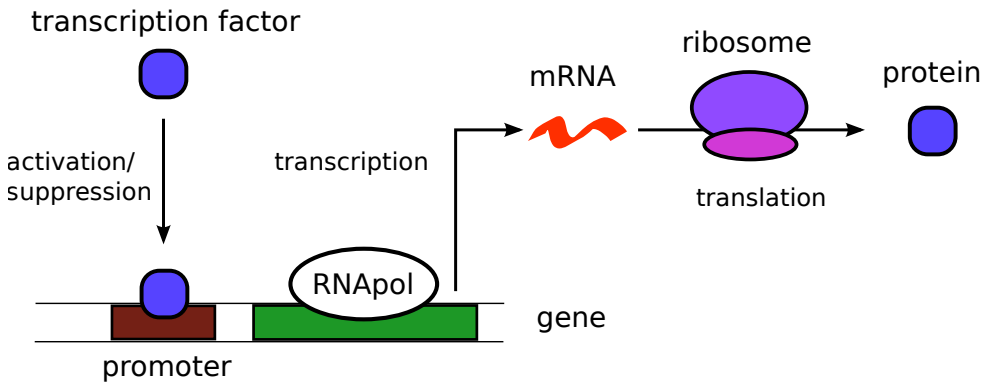

Figure 2.1: Illustration of gene expression. Transcription is conducted by an RNA polymerase enzyme (RNApol) that binds to the gene and produces an mRNA molecule from the gene's DNA code. The mRNA is then translated into proteins by ribosomes.

2.1.1 Gene expression

The term "gene expression" denotes the process of protein production from a gene and is illustrated in Figure 2.1. The process includes two main steps; *transcription* during which mRNA molecules are produced, and *translation* during which protein molecules are produced. The transcription process involves the enzyme RNA polymerase. For the mechanism to become initiated, an RNA polymerase enzyme has to bind to the beginning sequence of a gene. It then slides along the gene and produces an mRNA molecule that represents the DNA code of the gene. The RNA polymerase moves around randomly in the cell, which can be approximately described as Brownian diffusion. This means that the RNA polymerase binding to the gene is a stochastic event that happens randomly in time. As it turns out, not only the binding of the RNA polymerase to the gene, but also the sliding along the gene happens stochastically. Therefore, the production of mRNA molecules is a stochastic process.

The production of protein molecules from mRNA during translation is conducted by ribosome molecules, which are RNA and protein complexes. The ribosomes and mRNA diffuse in the cell and hence meet randomly before transcription can occur. Transcription is thus also a stochastic process. Similarly, the degradation of mRNA molecules and proteins is conducted by certain enzymes and thus happens stochastically.

Gene expression is frequently regulated by certain types of proteins, so-called *transcription factors*. Transcription factors are gene specific and bind to the *promoter* which is a part of the gene upstream of the DNA sequence that becomes transcribed. Once bound to the promoter, the transcription factor modifies the rate at which RNA

polymerase molecules bind to the gene and hence the rate of transcription. A transcription factor can either increase or decrease the binding rate and thus either enhance or suppress gene expression.

We refer to fluctuations that arise due to the inherent stochasticity of chemical processes as described above as *intrinsic noise*. Differences in molecule numbers of a certain species, say proteins, between different cells can be due to this intrinsic noise. However, such differences can also origin from other effects, such as physiological differences between cells or differing environmental conditions. For example, the numbers of RNA polymerase or ribosomes may differ between different cells, or different cells may be exposed to varying nutrient concentrations due to environmental fluctuations. Such differences that are not due to the stochasticity of chemical reactions are referred to as *extrinsic noise*.

2.1.2 Experimental evidence

As explained above, stochastic fluctuations are inherent to biochemical processes such as gene regulation. The question hence arises: what effects do these fluctuations have on the functioning of cells and which processes dominate the emergence of these fluctuations? Another important question is how the effects of intrinsic and extrinsic noise compare to each other. As it turns out, both types of noise can have significant effects on biochemical processes as has been found in various experimental studies.

One of the first experiments that aimed at separating the effects of intrinsic and extrinsic noise on gene expression has been conducted by Elowitz et al. in 2002 [6]. The authors observed large variations in the expression levels of certain genes across an *Escherichia coli* colony, and showed that both intrinsic and extrinsic noise can contribute significantly to these variations.

[6] and other early studies in prokaryotes suggested that fluctuations are mainly due to small-number fluctuations in mRNA numbers, giving rise to *translational bursts* in protein numbers [42]. Since then, however, it has been found that also *transcriptional bursts* in mRNA numbers can have significant effects in prokaryotes [43, 44]. Transcriptional bursts have also been found to largely govern fluctuations in gene expression in eukaryotes [45].

In [44] the authors measured the expression levels of 43 genes in *Escherichia coli* under various different environmental conditions. They observed that the variance of protein numbers scales roughly like the mean protein numbers, and their analysis sug-

gests that this scaling is likely due to promoter fluctuations. Moreover, the authors show that for intermediate abundance proteins, the intrinsic fluctuations are comparable or even larger than extrinsic fluctuations. Similarly, in a large scale study of 2500 genes in yeast, it was found that fluctuations are dominated by transcriptional bursts [46]. In addition, the authors found that fluctuations of a gene are strongly correlated with regulating transcription factors. In another large scale study on *Escherichia coli* in [47], the authors found that the distributions of all proteins can be approximated by geometric distributions, and that fluctuations are dominated by extrinsic noise at high expression levels.

The reported fluctuations can have significant influence on the functional behaviour of cells. A particular important example are *stochastic cell fate decisions* [48]. Differing cell fate decisions of genetically identical cells are believed to be beneficial for a population of cells experiencing strongly fluctuating environments [48].

There are a large number of other experimental studies demonstrating the importance of intrinsic noise in biochemical networks. The question arises if it is possible to mathematically model the dynamics of such systems. To answer this question, we will first give a brief introduction to the theory of stochastic processes in the next section, before discussing mathematical models for stochastic chemical networks in Section 2.3.

2.2 Stochastic processes

Processes that experience random fluctuations as described in the previous section are typically modelled by *stochastic processes*. We only consider continuous-time stochastic processes in this work. A continuous-time stochastic process is a time-dependent random variable $\mathbf{X}(t)$. Given some values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ of $\mathbf{X}(t)$ at time points $t_1, t_2, \ldots, t_n$, we assume that the joint probability density

$$p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \ldots; \mathbf{x}_n, t_n) \tag{2.1}$$

exists. If such a distribution is defined for all possible values of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and $t_1, t_2, \ldots, t_n$ for all $n \in \mathbb{N}$, the stochastic process is completely characterised.

We consider here a particular class of stochastic processes called *Markov processes*.

2.2.1 Markov processes

Intuitively, the Markov property states that the dynamics of a stochastic process only depends on the present state of the process rather than on its past. In other words, Markov processes are memoryless. If we assume the time-ordering $\tau_1 \leq \tau_2 \leq \ldots \tau_m \leq t_1 \leq \ldots t_n$, the Markov property can be formulated as

$$p(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1 | \mathbf{z}_m, \tau_m; \ldots; \mathbf{z}_1, \tau_1) = p(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1 | \mathbf{z}_m, \tau_m). \tag{2.2}$$

Assuming that $t_1 \leq t_2 \leq t_3$, we can use this to show

$$\begin{aligned} p(\mathbf{x}_3, t_3; \mathbf{x}_2, t_2; \mathbf{x}_1, t_1) &= p(\mathbf{x}_3, t_3 | \mathbf{x}_2, t_2; \mathbf{x}_1, t_1) p(\mathbf{x}_2, t_2; \mathbf{x}_1, t_1) \\ &= p(\mathbf{x}_3, t_3 | \mathbf{x}_2, t_2) p(\mathbf{x}_2, t_2 | \mathbf{x}_1, t_1) p(\mathbf{x}_1, t_1). \end{aligned} \tag{2.3}$$

It is straightforward to generalise this to show that any joint distribution can be written as the product of transition probabilities of the form $p(\mathbf{x}_2, t_2 | \mathbf{x}_1, t_1)$ times the marginal distribution of the initial state, i.e., $(t_1 \leq t_2 \leq \ldots \leq t_n)$

$$p(\mathbf{x}_n, t_n; \ldots; \mathbf{x}_1, t_1) = p(\mathbf{x}_1, t_1) \prod_{i=1}^{n-1} p(\mathbf{x}_{i+1}, t_{i+1} | \mathbf{x}_i, t_i). \tag{2.4}$$

This means that a Markov process is fully characterised by its initial distribution and its transition probabilities.

2.2.2 The differential Chapman-Kolmogorov equation

By integrating both sides of Eq. (2.3) over x_2 and dividing by $p(\mathbf{x}_1, t_1)$ we obtain the *Chapman-Kolmogorov equation*

$$p(\mathbf{x}_3, t_3 | \mathbf{x}_1, t_1) = \int d\mathbf{x}_2 p(\mathbf{x}_3, t_3 | \mathbf{x}_2, t_2) p(\mathbf{x}_2, t_2 | \mathbf{x}_1, t_1). \tag{2.5}$$

Let us now consider a n-dimensional process and write $\mathbf{x} = (x_1, \ldots, x_N)$. Under certain assumptions the Chapman-Kolmogorov equation in Eq. (2.5) can be written in a differential form. Specifically, we require that functions $W(\mathbf{z}|\mathbf{x}, t), A_i(\mathbf{x}, t)$ and $B_{ij}(\mathbf{x}, t), i, j = 1, \ldots, n$ exist, such that the following relations hold for all $\varepsilon > 0$:

$$\lim_{dt \to 0} \frac{p(\mathbf{z}, t + dt | \mathbf{x}, t)}{dt} = W(\mathbf{z}|\mathbf{x}, t), \tag{2.6a}$$

$$\lim_{dt \to 0} \frac{1}{dt} \int_{|\mathbf{z}-\mathbf{x}|<\varepsilon} d\mathbf{z} (x_i - z_i) p(\mathbf{z}, t + dt | \mathbf{x}, t) = A_i(\mathbf{x}, t) + O(\varepsilon), \tag{2.6b}$$

$$\lim_{dt \to 0} \frac{1}{dt} \int_{|\mathbf{z}-\mathbf{x}|<\varepsilon} d\mathbf{z} (x_i - z_i)(x_j - z_j) p(\mathbf{z}, t + dt | \mathbf{x}, t) = B_{ij}(\mathbf{x}, t) + O(\varepsilon). \tag{2.6c}$$

Requirement (2.6a) has to be true uniformly in $\mathbf{x}, \mathbf{z}$ and t for $|\mathbf{x}-\mathbf{z}| < \varepsilon$, while requirements (2.6a) and (2.6b) have to be true uniformly in $\mathbf{z}, t$ and ε. Using these relations it is possible to derive the *differential Chapman-Kolmogorov equation* [49]:

$$\frac{\partial}{\partial t}p(\mathbf{y},t|\mathbf{x},t') = \int d\mathbf{x}'[W(\mathbf{y}|\mathbf{x}',t)p(\mathbf{x}',t|\mathbf{x},t') - W(\mathbf{x}'|\mathbf{y},t)p(\mathbf{y},t|\mathbf{x},t') \\ - \sum_{i=1}^{N}\frac{\partial}{\partial y_i}[A_i(\mathbf{y},t)p(\mathbf{y},t|\mathbf{x},t')] + \frac{1}{2}\sum_{i,j=1}^{N}\frac{\partial^2}{\partial y_i \partial y_j}[B_{ij}(\mathbf{y},t)p(\mathbf{y},t|\mathbf{x},t')]. \tag{2.7}$$

The first line describes discrete jumps between two states $\mathbf{x}$ and $\mathbf{x}'$ with *transition rates* or *propensity functions* $W(\mathbf{x}'|\mathbf{x},t)$. The second line describes continuous changes of the process, and $\mathbf{A}(\mathbf{x},t)$ and $\mathbf{B}(\mathbf{x},t)$ are called *drift vector* and *diffusion matrix*, respectively. Note that the Markov property manifests itself in $W(\mathbf{x}'|\mathbf{x},t)$, $\mathbf{A}(\mathbf{x},t)$ and $\mathbf{B}(\mathbf{x},t)$ being functions of the current state and time only, and not of any past time points.

Eq. (2.7) only has a probabilistic interpretation if a non-negative solution exists. It can be shown that this is the case if the following conditions are met [50]: $W(\mathbf{x}'|\mathbf{x},t)$ has to be non-negative, $\mathbf{B}(\mathbf{x},t)$ has to be positive semi-definite, appropriate boundary conditions have to be specified, and the following *initial condition* has to be satisfied,

$$p(\mathbf{y},t|\mathbf{x},t) = \delta(\mathbf{y}-\mathbf{x}). \tag{2.8}$$

Let us next study two different types of stochastic processes which arise as special cases of Eq. (2.7) and which we will use throughout this thesis.

2.2.3 Markov jump processes

Consider the case where $\mathbf{A}(\mathbf{x},t) = 0$ and $\mathbf{B}(\mathbf{x},t) = 0$. In this case Eq. (2.7) becomes a so-called *master equation*:

$$\frac{\partial}{\partial t}p(\mathbf{y},t|\mathbf{x},t') = \int d\mathbf{x}'\left[W(\mathbf{y}|\mathbf{x}',t)p(\mathbf{x}',t|\mathbf{x},t') - W(\mathbf{x}'|\mathbf{y},t)p(\mathbf{y},t|\mathbf{x},t')\right]. \tag{2.9}$$

It can be shown that the corresponding process has a finite probability to remain in a certain initial state for a finite time. This means that realisations of the process are locally constant functions with discontinuous jumps [49]. Hence the name *Markov jump processes*. Let $R(\mathbf{x},t+\Delta t,t)$ be the probability that the process remains in state $\mathbf{x}$ for some time interval Δt with initial time t. Assuming that the transition rates

$W(\mathbf{x}|\mathbf{x}',t) = W(\mathbf{x}|\mathbf{x}')$ do not depend on time, it can be shown [49] that $R(\mathbf{x},t+\Delta t,t)$ takes the form

$$R(\mathbf{x},t+\Delta t,t) = \exp(-\lambda\Delta t), \quad \lambda = \int d\mathbf{x}\, W(\mathbf{x}|\mathbf{x}'). \tag{2.10}$$

This means that the jump times are exponentially distributed which can be used to simulate the process.

Note that although a Markov jump process described by a master equation as in Eq. (2.9) only undergoes discrete jumps, its state space can generally be continuous. However, many Markov jump processes of interest do only live on an integer-valued state space. In this case the integral in Eq. (2.9) becomes a sum and the master equation can be written as

$$\frac{\partial}{\partial t}p(\mathbf{y},t|\mathbf{x},t') = \sum_{\mathbf{x}'} d\mathbf{x}'[W(\mathbf{y}|\mathbf{x}',t)p(\mathbf{x}',t|\mathbf{x},t') - W(\mathbf{x}'|\mathbf{y},t)p(\mathbf{y},t|\mathbf{x},t')]. \tag{2.11}$$

2.2.4 Diffusion processes

Consider next the case of vanishing transition rates, i.e., $W(\mathbf{x}|\mathbf{x}',t) = 0$. In this case one obtains a so-called *diffusion process* and Eq. (2.7) becomes a *Fokker-Planck equation*,

$$\frac{\partial}{\partial t}p(\mathbf{y},t|\mathbf{x},t') = -\sum_{i=1}^{N}\frac{\partial}{\partial y_i}\left[A_i(\mathbf{y},t)p(\mathbf{y},t|\mathbf{x},t')\right] + \frac{1}{2}\sum_{i,j=1}^{N}\frac{\partial^2}{\partial y_i \partial y_j}\left[B_{ij}(\mathbf{y},t)p(\mathbf{y},t|\mathbf{x},t')\right]. \tag{2.12}$$

The first term is called *drift term* and describes deterministic motion of the process. The second term is called *diffusion term* and describes fluctuations around the deterministic motion. $B_{ij}(\mathbf{y},t)$ has to be positive semi-definite and symmetric.

The Wiener process

A simple but important one-dimensional example is the *Wiener process* which is obtained by $A = 0$ and $B = 1$ leading to

$$\frac{\partial}{\partial t}p(w,t|w_0,t_0) = \frac{1}{2}\frac{\partial^2}{\partial w^2}p(w,t|w_0,t_0). \tag{2.13}$$

This is solved by a normal distribution:

$$p(w,t|w_0,t_0) = \frac{1}{\sqrt{2\pi(t-t_0)}}\exp\left(-\frac{(w-w_0)^2}{2(t-t_0)}\right), \tag{2.14}$$

with mean and variance given by

$$\langle w \rangle = w_0, \tag{2.15}$$

$$\langle (w - w_0)^2 \rangle = t - t_0. \tag{2.16}$$

The Wiener process is also referred to as *Brownian motion* or *Gaussian white noise*.

The Langevin equation

It can be shown that any Fokker-Planck equation of the form in Eq. (2.12) is equivalent to an *Ito stochastic differential equation* (SDE) or *Langevin equation*. In one dimension, the latter reads

$$dx(t) = A(x(t),t)dt + \sqrt{B(x(t),t)}dW(t), \tag{2.17}$$

where $W(t)$ is a Wiener process whose distribution satisfies Eq. (2.13). The SDE in Eq. (2.17) is defined in the sense that its solutions satisfy

$$x(t) = x(t_0) + \int_{t_0}^{t} dt' A(x(t'),t')dt' + \int_{t_0}^{t} \sqrt{B(x(t'),t')}dW(t'). \tag{2.18}$$

The last term denotes an *Ito stochastic integral* [49]. An easy method to simulate an SDE of the form in Eq. (2.17) is the *Euler-Maruyama algorithm* which discretises time into intervals Δt and updates the process iteratively as

$$x(t+\Delta t) = x(t) + A(x(t),t)\Delta t + \sqrt{B(x(t),t)\Delta t}\; dw, \quad dw \sim \mathcal{N}(0,1), \tag{2.19}$$

where $\mathcal{N}(0,1)$ is a normal distribution with mean 0 and variance 1. It can be shown [49] that the distribution of a process described by Eq. (2.17) agrees exactly with the solution of the corresponding Fokker-Planck equation in (2.12). One can thus interpret the SDE in Eq. (2.17) as a generator of realisations of the stochastic process described by the corresponding Fokker-Planck equation.

Similarly, one can define an SDE for the general, multi-dimensional case. The SDE corresponding to the Fokker-Planck equation in (2.12) is given by

$$d\mathbf{x}(t) = \mathbf{A}(\mathbf{x}(t),t)dt + \mathbf{C}(\mathbf{x}(t),t)d\mathbf{W}(t), \quad \mathbf{C}\mathbf{C}^T = \mathbf{B}, \tag{2.20}$$

where $d\mathbf{W}(t) = (dW_1(t), \ldots, dW_m(t))$, m is the number of columns of $\mathbf{C}$ and the $W_i(t)$ are independent Wiener processes. Note that in general different choices for the matrix $\mathbf{C}$ exist. It can be shown, however, that such different choices lead to the same transition probabilities, which means that they are equivalent [49].

2.3 Stochastic chemical kinetics

After having given a formal mathematical introduction to stochastic processes in Section 2.2, let us now consider how the presented methods can be used to model the dynamics of stochastic chemical reaction networks.

2.3.1 Chemical reaction networks

As discussed in Section 2.1, gene expression is an example of a biological process for which stochastic fluctuations are important. The underlying mechanisms of such processes are typically complicated procedures involving several different types of molecules and physical operations. Instead of modelling all these mechanisms explicitly, we model them as single stochastic events. In the case of transcription, for example, a gene produces mRNA molecules at random points in time. We call such events *chemical reactions* and use an abstract notation to represent them. For a gene G that produces an mRNA molecule M, for example, we write

$$G \longrightarrow G + M. \tag{2.21}$$

A set of chemical species together with a set of chemical reactions constitutes what we call a *chemical reaction network*.

General chemical reaction networks

Let us next introduce the notation for representing general chemical reaction networks. A general chemical reaction system consists of some chemical species $X_i,\ i = 1,...,N$, that interact via some chemical reactions as

$$\sum_{i=1}^{N} s_{ij} X_i \xrightarrow{k_j} \sum_{i=1}^{N} r_{ij} X_i, \quad j = 1,\dots,R, \tag{2.22}$$

where the *stoichiometric coefficients* s_{ij} and r_{ij} are non-negative integer numbers denoting numbers of reactant and product molecules, respectively, and k_r is the macroscopic reaction rate constant of the r^{th} reaction. We say that the r^{th} reaction is "of order m" if $\sum_{i=1}^{N} s_{ir} = m$, i.e., if it involves m reactant molecules. We further call a reaction "bimolecular" if $m = 2$ and a system "linear" if $m \leq 1$ for all reactions in the system.

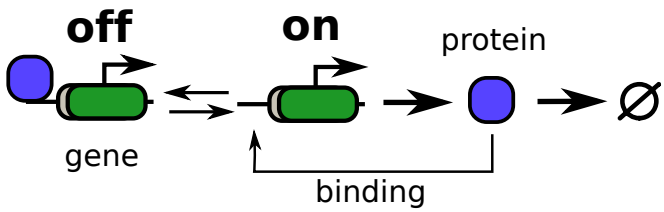

Figure 2.2: Illustration of a gene expression system with negative feedback loop. When the gene is in the "on" state it produces proteins of a certain type. The protein can decay or bind to the gene. In the bound state the gene is in the "off" state, i.e., no protein is produced. The protein hence suppresses its own production. The corresponding reactions are given in Eq. (2.23).

Example

As an example, consider the gene system in Figure 2.2. This system can be viewed as a simplified version of the gene expression process described in Figure 2.1: we replace the process of transcription (gene produces an mRNA molecule) and translation (mRNA molecule produces a protein) by a single reaction in which the gene directly produces the protein. This is a valid approximation if the time scales of transcription and decay of mRNA are much faster than all other time scales in the system.

In the system in Figure 2.2, the protein can bind to the gene and thus acts as its own transcription factor. In the bound state G^{off}, the gene does not produce any protein. The protein thus suppresses its own production, which means the system is an example of a *negative feedback loop*. The corresponding reactions in the notation of Eq. (2.22) read

$$G^{\text{on}} \xrightarrow{k_1} G^{\text{on}} + P, \quad G^{\text{on}} + P \underset{k_3}{\overset{k_2}{\rightleftharpoons}} G^{\text{off}}, \quad P \xrightarrow{k_4} \varnothing, \tag{2.23}$$

where we call the gene in the *on* and off state and the protein $G^{\text{on}}, G^{\text{off}}$ and P rather than X_1, X_2 and X_3 for notational clarity. In our nomenclature, $G^{\text{on}} + P \to G^{\text{off}}$ is a second-order/bimolecular reaction, while the other three reactions are of first order/linear. By "$P \to \varnothing$" we indicate that P leaves the system under consideration. This could either mean that P physically leaves the compartment under consideration, or that it becomes converted into different types of chemical species that are not included in our model.

Modeling

In the most general description of chemical reaction processes, we assume that the X_i molecules in Eq. (2.22) undergo Brownian diffusion in a closed compartment and

that the reactions in Eq. (2.22) happen stochastically under certain rules. Bimolecular reactions happen with certain probabilities whenever two reactant molecules collide. We thus have two sources of stochasticity for bimolecular reactions; Brownian diffusion leading to stochastically occurring collisions; and the collisions themselves stochastically leading to reaction events. The most general models of such systems typically rely on time discretisation and simulations of each single molecule in space, with chemical reactions happening with certain probabilities under certain rules. Under certain conditions, however, simplified descriptions can be employed making a spatial description and the simulation of single particles obsolete. We will next first discuss the situation where such conditions are met and deal with the more general spatial case later in Section 2.4.

2.3.2 The chemical master equation

We now consider a general chemical reaction system as in Eq. (2.22) in a closed compartment, and assume well-mixed and dilute conditions. By "well-mixed" we mean that the diffusion of particles in the compartment constitutes the fastest time scale of the system, in the sense that the expected travelled distance of each molecule between reactive collisions is much larger then the length scale of the compartment. By "dilute" we mean that the combined volume of all the considered molecules is much smaller than the total volume, which means that the volume of the molecules can be ignored. If these two conditions are met, it can be shown [51] that the state of the system at any time is fully determined by the state vector $\mathbf{n} = (n_1, \ldots, n_N)$, where n_i is the molecule number of species X_i. In particular, the spatial locations and diffusion of molecules do not have to be modelled. The dynamics of the system can then be modelled as a *Markov jump process* (c.f. Section 2.2.3). Since chemical reactions always lead to integer-valued changes in molecule counts, the state space of the process is given by the non-negative integers. Accordingly, the transition probabilities of the process satisfy a master equation of the form in Eq. (2.11). Since in the case of chemical reaction systems, the propensity functions do not depend on time and since only a finite number of transitions are possible - the chemical reactions - the master equation takes a particular simple form and is called the *chemical master equation* (CME)[51]:

$$\partial_t P(\mathbf{n},t) = \sum_{r=1}^{R} f_r(\mathbf{n}-\mathbf{S}_r)P(\mathbf{n}-\mathbf{S}_r,t) - \sum_{r=1}^{R} f_r(\mathbf{n})P(\mathbf{n},t), \tag{2.24}$$

where $P(\mathbf{n},t)$ is the probability of the system to be in state $\mathbf{n}$ at time t and we implicitly assume conditioning on some initial state $\mathbf{n}_0$ at some initial time t_0. $\mathbf{S}_r$ in Eq. (2.24) is a vector whose entries correspond to the r^{th} column of the *stoichiometric matrix* defined as $S_{ij} = r_{ij} - s_{ij}$, and we call the $f_r(\mathbf{n})$ *microscopic propensity functions*. Specifically, $f_r(\mathbf{n})dt$ gives the probability for the r^{th} reaction to happen in an infinitesimal time interval dt. In the case of *mass-action kinetics* the propensity functions are defined as

$$f_r(\mathbf{n}) = k_r \Omega \prod_{k=1}^{N} \frac{n_k!}{(n_k - s_{kj})! \Omega^{s_{kj}}}, \tag{2.25}$$

where Ω is the volume of the system and k_r is the macroscopic reaction rate constant of the r^{th} reaction. Eq. (2.25) can be derived from combinatorial considerations. Propensity functions that are not of mass-action kinetics type as in Eq. (2.25) are frequently used in the literature, such as hill-type functions for modelling the dependence of a gene's expression level on its transcription factor. Such propensity functions typically arise in reduced models where an effective reaction replaces several microscopic reactions. In the gene expression system mentioned before, for instance, the binding of the transcription factor to the promoter of the gene is not modelled explicitly, but the effect of its concentration included in the modified propensity function of the expression reaction. In this thesis, we only consider example systems with propensity functions of mass-action kinetics type. However, most of the derived results are valid for more general, analytic propensity functions.

In the context of chemical reactions, the CME was first proposed by McQuarrie [52] and later derived from physical principles by Gillespie [51]. Analytic solutions of the CME are known only for few simple systems. Therefore, one generally has to rely on stochastic simulations or on approximations of the CME.

Example

As an example, consider the gene system in Figure 2.2 with reactions given in Eq. (2.23). If we order the reactions according to the rate constants in Eq. (2.23), and the species as G^{on}, P and G^{off} with particle numbers n_1, n_2 and n_3, respectively, the matrices $\mathbf{s}$ and $\mathbf{r}$ with coefficients defined in Eq. (2.22) read

$$\mathbf{s} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \qquad \mathbf{r} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{2.26}$$

Accordingly, the stoichiometric matrix $\mathbf{S} = \mathbf{r} - \mathbf{s}$ and propensity vector $\mathbf{f}(\mathbf{n}) = (f_1(\mathbf{n}), \ldots, f_R(\mathbf{n}))^T$, with the f_r defined in Eq. (2.25), are given by

$$\mathbf{S} = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 1 & -1 & 1 & -1 \\ 0 & 1 & -1 & 0 \end{pmatrix}, \quad \mathbf{f}(\mathbf{n}) = (k_1 n_1, \frac{k_2}{\Omega} n_1 n_2, k_3 n_3, k_4 n_2)^T, \tag{2.27}$$

and it is straightforward to write down the corresponding CME using Eq. (2.24). However, note that the system has a conservation law in particle numbers which we can use to find a simplified description by reducing the number of variables: the total number of genes in the "on" state and genes in the "off" state is constant, i.e., $n_1 + n_3 = const$. Let as assume that the constant is equal to unity, i.e., that there is only one gene in the system. We can thus reduce the system to a two species system by using $n_3 = 1 - n_1$. The matrices $\mathbf{s}$ and $\mathbf{r}$ for the reduced system are simply obtained from Eq. (2.28) by dropping the last row,

$$\mathbf{s} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \qquad \mathbf{r} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}, \tag{2.28}$$

and the stoichiometric matrix and propensity vector of the reduced system read accordingly

$$\mathbf{S} = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 1 & -1 & 1 & -1 \end{pmatrix}, \quad \mathbf{f}(\mathbf{n}) = (k_1 n_1, \frac{k_2}{\Omega} n_1 n_2, k_3 (1 - n_1), k_4 n_2)^T. \tag{2.29}$$

The corresponding CME becomes (c.f. Eq. (2.24))

$$\begin{aligned} P(n_1, n_2, t) = {} & k_1 n_1 P(n_1, n_2 - 1, t) + \frac{k_2}{\Omega}(n_1 + 1)(n_2 + 1) P(n_1 + 1, n_2 + 1, t) \\ & + k_3 (2 - n_1) P(n_1 - 1, n_2 - 1, t) + k_4 (n_2 + 1) P(n_1, n_2 + 1, t) \\ & - (k_1 n_1 + \frac{k_2}{\Omega} n_1 n_2 + k_3 (1 - n_1) + k_4 n_2) P(n_1, n_2, t). \end{aligned} \tag{2.30}$$

We will work with the reduced system in the following. Despite having a relatively simple system here with effectively only two species, no time-dependent solution for its CME in Eq. (2.30) has been derived to our knowledge. A solution in steady state has been derived in [53], but for most other systems not even a steady-state solution is available. Therefore, one generally needs to rely on stochastic simulations or approximations of the CME to study the behaviour of such systems.

2.3.3 The stochastic simulation algorithm

The stochastic simulation algorithm (SSA) is a popular Monte Carlo method that allows to simulate exact realisations of the stochastic process described by the CME in Eq. (2.24). It was first proposed in the context of chemical kinetics by Gillespie [54, 55], and several variants have been proposed in the literature [56]. As explained in Section 2.2.3, the waiting times of reaction events are exponentially distributed if the propensity functions of the master equation do not explicitly depend on time. Since this is the case in the CME, it is straightforward to simulate the occurrences of chemical reactions and to update the state vector accordingly. This is the basic idea underlying the SSA and its variants. One example is the so-called "direct method" [54], which samples the time step τ for the next reaction to happen from an exponential distribution with mean given by the inverse of the sum of all propensity functions evaluated at the current state, i.e.,

$$\tau \sim \exp\left(-\tau \sum_{r=1}^{R} f_r(\mathbf{n})\right) \sum_{r=1}^{R} f_r(\mathbf{n}), \tag{2.31}$$

if $\mathbf{n}$ is the current state of the system. Which reaction happens is then determined by using that the probability of the r^{th} reaction to happen is proportional to $f_r(\mathbf{n})$.

Unfortunately, the applicability of the SSA is severely limited due to its computational cost. Since each and every reaction event is simulated explicitly, the SSA becomes computationally expensive even for systems with few species. This is particularly the case if the molecule numbers have large fluctuations or if many reactions happen per unit time. In the first case a large number of sample paths have to be simulated to obtain statistically accurate results, whereas in the second case single simulations become expensive since the time between reaction events becomes small. The fact that exact simulations are computationally expensive makes analytic approximations of the CME appealing.

2.3.4 Moment closure approximations

One class of such approximation methods are so-called *moment closure approximations* (MA). MAs aim at approximating the first few moments of a process satisfying the CME.

Moment equations

Starting from the CME in Eq. (2.24) one can derive time evolution equations for the moments of its solution. To obtain the time evolution equation for the moment $\langle n_i \dots n_l \rangle$ we multiply Eq. (2.24) by $n_i \dots n_l$ and sum over all molecule numbers, leading to

$$\partial_t \langle n_i \dots n_l \rangle = \sum_{r=1}^{R} \langle (n_i + S_{ir}) \dots (n_l + S_{lr}) f_r(\mathbf{n}) \rangle - \sum_{r=1}^{R} \langle n_i \dots n_l f_r(\mathbf{n}) \rangle. \tag{2.32}$$

Here, $\langle \cdot \rangle$ denotes the expectation with respect to the solution $P(\mathbf{n},t)$ of the CME in Eq. (2.24). For moments of up to order two Eq. (2.32) becomes

$$\partial_t \langle n_i \rangle = \sum_{r=1}^{R} S_{ir} \langle f_r(\mathbf{n}) \rangle, \tag{2.33}$$

$$\partial_t \langle n_i n_j \rangle = \sum_{r=1}^{R} \left[S_{jr} \langle n_i f_r(\mathbf{n}) \rangle + S_{ir} \langle f_r(\mathbf{n}) n_j \rangle + S_{ir} S_{jr} \langle f_r(\mathbf{n}) \rangle \right]. \tag{2.34}$$

We see that unless all $f_r(\mathbf{n})$ are zeroth or first-order polynomials in $\mathbf{n}$, i.e., unless the system is linear without any bimolecular or higher order reactions, the evolution equation of a certain moment depends on higher order moments. This means that the moment equations of different orders are coupled to each other, leading to an infinite hierarchy of coupled equations with can obviously not be solved.

General formulation

The idea underlying the class of moment closure approximations that we will study in this work is to express all moments above a certain order M as functions of lower-order moments. The latter is typically done by assuming the distribution of the system to have a particular functional form, for example a normal distribution. This decouples the equations of the moments up to order M from higher-order moments, which in turn allows to (numerically) solve the resulting moment equations. We refer to such a moment closure as "MA of order M". Such MA methods thus amount to solving a finite system of ordinary differential equations (ODEs) for which efficient and accurate numerical algorithms exist. Solving MAs numerically is thus generally orders of magnitude faster than using stochastic simulations of the CME. Let

$$y_{i_1,\ldots,i_k} = \langle n_{i_1} \ldots n_{i_k} \rangle, \tag{2.35}$$

$$z_{i_1,\ldots,i_k} = \begin{cases} \langle (n_{i_1} - y_{i_1}) \ldots (n_{i_k} - y_{i_k}) \rangle & \text{if} \quad k \geq 2, \\ y_{i_1} & \text{if} \quad k = 1, \end{cases} \tag{2.36}$$

$$c_{i_1,\ldots,i_k} = \partial_{s_{i_1}} \ldots \partial_{s_{i_k}} g(s_1, \ldots, s_N)|_{s_1,\ldots,s_N=0}, \tag{2.37}$$

denote the raw or "normal" moments, central moments and cumulants of order k, respectively. We call $y_{i_1,\ldots,i_k}$ a "diagonal moment" if $i_l = i_m$ for all $l, m \in \{1, \ldots, k\}$, and a "mixed moment" otherwise, and similarly for central moments and cumulants. $g(s)$ in Eq. (2.37) is the cumulant generating function defined as

$$g(s_1, \ldots, s_N) = \log\langle \exp(s_1 n_1 + \ldots + s_N n_N) \rangle. \tag{2.38}$$

We note that all three types of moments are respectively invariant under permutations of their indices. Therefore, only one representative combination of each permutation class has to be considered. Taking this symmetry into account significantly reduces the number of variables and moment equations. We adopt here the convention that the indices are ordered from small to large, i.e., for a moment $y_{i_1,\ldots,i_k}$ we have $i_1 \leq i_2 \leq \ldots \leq i_k$. Expressing the moment closure functions in terms of cumulants rather than raw moments often gives shorter expressions. The equations for the cumulants can then be rearranged to give equations for the raw moments.

Some popular moment closure methods

Four popular moment closures methods are

- "Normal moment closure" (also called "cumulant neglect moment closure" in the literature): all cumulants above order M are set to zero, i.e.,

$$c_{i_1,\ldots,i_k} = 0, \quad \text{for} \quad k > M. \tag{2.39}$$

- "Poisson moment closure": the cumulants of a one-dimensional Poisson distribution are all equal to the mean value. We assume here the multivariate distribution to be a product of uni-variate Poisson distributions. Accordingly, for the Poisson MA of order M we set all diagonal cumulants to the corresponding mean

and all mixed cumulants to zero, i.e.,

$$c_{i_1,\ldots,i_k} = y_i, \quad \text{for} \quad k > M \quad \text{and} \quad i_1,\ldots,i_k = i, \quad \text{for some} \quad i \in \{1,\ldots,N\}, \tag{2.40}$$

$$c_{i_1,\ldots,i_k} = 0, \quad \text{for} \quad k > M \quad \text{and} \quad i_m \neq i_n \quad \text{for some} \quad m,n \in \{1,\ldots,N\}. \tag{2.41}$$

- "Log-normal moment closure": let $\mathbf{m}$ and $\mathbf{S}$ be the mean vector and covariance matrix of a multi-dimensional normal random variable. Then the logarithm of the latter has a multivariate log-normal distribution and its moments can be expressed in terms of $\mathbf{m}$ and $\mathbf{S}$ as [57]

$$y_{i_1,\ldots,i_k} = \exp\left(\mathbf{v}^T\mathbf{m} + \frac{1}{2}\mathbf{v}^T\mathbf{S}\mathbf{v}\right), \quad \text{for} \quad k > M, \tag{2.42}$$

  where $\mathbf{v} = (g_1,\ldots,g_N)$, and g_m is the number of i_j's having the value m. This allows to express $\mathbf{m}$ and $\mathbf{S}$ in terms of the first two moments y_i and $y_{i,j}$ which then in turn allows to express higher-order moments in terms of y_i and $y_{i,j}$, too.

- "Central-moment-neglect (CMN) moment closure": all central moments above order M are set to zero:

$$z_{i_1,\ldots,i_k} = 0, \quad \text{for} \quad k > M. \tag{2.43}$$

Each of the closure equations can be used to express all raw moments above order M in terms of lower order moments and thus close the moment equations according to the corresponding MA. We note that the normal MA, Poisson MA and CMN-MA can be equivalent depending on the reaction system and closure order.

The normal moment closure dates back more than sixty years [16, 17, 26]. It has been introduced to the field of biochemical reactions only in recent years [27], however, and is probably the most commonly used one. The Poisson and log-normal moment closure techniques have been proposed in [30] and [29], respectively. In [58] it has been shown that the assumption of a log-normal distribution is equivalent to a "derivative matching" closure. The CMN-MA is also called "low dispersion moment closure" in the literature [31].

Example

Let us again consider the gene system in Figure 2.2 with reactions given in Eq. (2.23). The corresponding stoichiometric matrix and propensity vector are given in Eq. (2.29).

Using these in Eqs. (2.33) and (2.34) and using the notation introduced in Eq. (2.35) one obtains

$$\partial_t y_1 = -\frac{k_2}{\Omega} y_{1,2} + k_3(1 - y_1), \tag{2.44}$$

$$\partial_t y_2 = k_1 y_1 - \frac{k_2}{\Omega} y_{1,2} + k_3(1 - y_1) - k_4 y_2, \tag{2.45}$$

$$\partial_t y_{1,1} = \frac{k_2}{\Omega}(-2y_{1,1,2} + y_{1,2}) + k_3(-2y_{1,1} + y_1 + 1), \tag{2.46}$$

$$\partial_t y_{1,2} = k_1 y_{1,1} + \frac{k_2}{\Omega}(-y_{1,1,2} - y_{1,2,2} + y_{1,2}) + k_3(-y_{1,1} - y_{1,2} + y_2 + 1) - k_4 y_{1,2}, \tag{2.47}$$

$$\partial_t y_{2,2} = k_1(2y_{1,2} + y_1) + \frac{k_2}{\Omega}(-2y_{1,2,2} + y_{1,2}) + k_3(-2y_{1,1} - y_1 + 2y_2 + 1) + k_4(-2y_{2,2} + y_2). \tag{2.48}$$

Note that the equations for the first moments in Eqs. (2.44) and (2.45) depend on the second moment $y_{1,2}$, and that the equations for the second moments (2.46)-(2.48) depend on the third moments $y_{1,1,2}$ and $y_{1,2,2}$. Similarly, it is easy to see that moment equations of any order depend on higher order moments, which means that we have an infinite system of coupled equations. Note that all terms in Eqs. (2.44)-(2.48) depending on higher order moments are proportional to the rate constant k_2 of the bimolecular reaction in Eq. (2.23), illustrating that the moment equations decouple in the absence of bimolecular (and higher order) reactions. This could be achieved here by setting $k_2 = 0$ for which the moment equations would decouple and could thus be solved numerically.

Let us now apply the normal moment closure defined in Eq. (2.39) to second order, i.e., we close the moment equations in (2.44)-(2.48) by setting the third order cumulants $c_{1,1,2}$ and $c_{1,2,2}$ to zero. To this end, we express $c_{1,1,2}$ and $c_{1,2,2}$ in terms of raw moments, set them to zero and rearrange in terms of $y_{1,1,2}$ and $y_{1,2,2}$, leading to

$$y_{1,1,2} = 2y_1 y_{1,2} + y_2 y_{1,1} - 2y_2 y_1^2, \tag{2.49}$$

$$y_{1,2,2} = 2y_2 y_{1,2} + y_1 y_{2,2} - 2y_1 y_2^2. \tag{2.50}$$

Using these expressions in Eqs. (2.44)-(2.48) the equations decouple from higher order moments. We give here the resulting equations in terms of central rather than raw

moments (c.f. Eq. (2.36))

$$\partial_t z_1 = -\frac{k_2}{\Omega}(z_{1,2} + z_1 z_2) + k_3(1 - z_1), \tag{2.51}$$

$$\partial_t z_2 = k_1 z_1 - \frac{k_2}{\Omega}(z_{1,2} - z_1 z_2) + k_3(1 - z_1) - k_4 z_2, \tag{2.52}$$

$$\partial_t z_{1,1} = \frac{k_2}{\Omega}(-2z_2 z_{1,1} - 2z_1 z_{1,2} + z_{1,2} + z_1 z_2) + y_{1,2}) + k_3(-2z_{1,1} - z_1 + 1), \tag{2.53}$$

$$\partial_t z_{1,2} = k_1 z_{1,1} + \frac{k_2}{\Omega}(-z_2 z_{1,1} - z_1 z_{1,2} - z_2 z_{1,2} - z_1 z_{2,2} + + z_{1,2} + z_1 z_2) \tag{2.54}$$

$$+ k_3(z_{1,2} - z_{1,1} - z_1 + 1) - k_4 z_{1,2}, \tag{2.55}$$

$$\partial_t z_{2,2} = k_1(2z_{1,2} + z_1) + \frac{k_2}{\Omega}(-2z_1 z_{2,2} - 2z_2 z_{1,2} + z_{1,2} + z_1 z_2) \tag{2.56}$$

$$+ k_3(-2z_{1,2} - z_1 + 1) + k_4(-2z_{2,2} + z_2). \tag{2.57}$$

2.3.5 The chemical Langevin equation

Suppose now that we are not only interested in approximating the moments of the CME solution as in the previous section, but that we would like to approximate the whole process and its distribution. The *chemical Langevin equation* (CLE) and the corresponding *chemical Fokker-Planck equation* (CFPE) constitute a popular diffusion approximation of the CME that aims at exactly that. Kramers and Moyal derived the latter by applying a Taylor expansion to the CME which upon truncation leads to a partial differential equation approximation of the CME [22, 23].

The chemical Fokker-Planck equation

Letting the variables in the CME in Eq. (2.24) become continuous and performing a Taylor expansion to second order in the first term of the r.h.s. of Eq. (2.24), we obtain a Fokker-Planck equation of the same form as in Eq. (2.12):

$$\partial_t P(\mathbf{x},t) = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i}\left[A_i(\mathbf{x})P(\mathbf{x},t)\right] + \frac{1}{2}\sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\left[B_{ij}(\mathbf{x})P(\mathbf{x},t)\right], \tag{2.58}$$

where we denote the continuous variable corresponding to species X_i by x_i. Note that the drift vector $\mathbf{A}(\mathbf{x})$ and diffusion matrix $\mathbf{B}(\mathbf{x})$ in Eq. (2.58) do not depend on time. Eq. (2.58) is called the *chemical Fokker-Planck equation* (CFPE). Note that whereas the state variables denote discrete molecule numbers in the CME, they denote continuous real numbers in the CFPE. It has been shown that the differences between the predictions of the two descriptions tend to zero in the limit of large molecule numbers

[59]. The drift vector $\mathbf{A}$ and diffusion matrix $\mathbf{B}$ are given by

$$A_i(\mathbf{x}) = \sum_{r=1}^{R} S_{ir} f_r(\mathbf{x}), \tag{2.59}$$

$$B_{ij}(\mathbf{x}) = \sum_{r=1}^{R} S_{ir} S_{jr} f_r(\mathbf{x}). \tag{2.60}$$

The chemical Langevin equation

As explained in Section 2.2.4, a Fokker-Planck equation of the form in (2.58) is equivalent to a Langevin equation of the form

$$d\mathbf{x} = \mathbf{A}(\mathbf{x})dt + \mathbf{C}(\mathbf{x})d\mathbf{W}, \qquad \mathbf{C}(\mathbf{x})\mathbf{C}(\mathbf{x})^T = \mathbf{B}(\mathbf{x}), \tag{2.61}$$

where $\mathbf{W}$ is a multi-dimensional Wiener process. By "equivalent" we mean that simulating Eq. (2.61) corresponds to drawing samples from the solution of Eq. (2.58). In the context of the CFPE, Eq. (2.61) is called the *chemical Langevin equation* (CLE). Note that the domain of both the CFPE and of the CLE is (implicitly) assumed to be that of real numbers.

Generally there exist different choices for $\mathbf{C}(\mathbf{x})$ in Eq. (2.61) corresponding to different factorisations of the matrix $\mathbf{B}(\mathbf{x})$; these lead to as many different representations of the CLE. A commonly used choice, following the seminal paper by Gillespie [9], is $C_{ir}(\mathbf{x}) = S_{ir}\sqrt{f_r(\mathbf{x})}$ leading to a CLE of the form

$$dx_i = \sum_{r=1}^{R} S_{ir} f_r(\mathbf{x})dt + \sum_{r=1}^{R} S_{ir}\sqrt{f_r(\mathbf{x})}dW_r. \tag{2.62}$$

This representation of the CLE is the most commonly used one in the literature, and we shall hence call it the "standard form" of the CLE throughout the rest of this thesis.

Properties

For a function $g(\mathbf{x})$, one can derive an ordinary differential equation for the time evolution of its expectation value $\langle g(\mathbf{x})\rangle$ from the CFPE in Eq. (2.58) by multiplying the latter with $g(\mathbf{x})$ and integrating over all $\mathbf{x}$ [49]. Alternatively, one can use Ito's formula to derive an equation for the time evolution of $g(\mathbf{x})$ from the CLE in Eq. (2.61) and subsequently take the average [49]. The equations derived from either the CFPE and CLE are identical. In particular, they depend only on $\mathbf{B}(\mathbf{x}) = \mathbf{C}(\mathbf{x})\mathbf{C}(\mathbf{x})^T$ and are thus independent of the particular choice for $\mathbf{C}(\mathbf{x})$. In this sense, the different choices for $\mathbf{C}(\mathbf{x})$ are often claimed to be equivalent in the literature [60].

When setting $g(\mathbf{x}) = x_i \dots x_l$ we obtain evolution equations for the moments of the solution of the CFPE. Importantly, it turns out that the equations for moments of up to order two are *exactly the same as the corresponding equations derived from the CME in Eqs.* (2.33) *and* (2.34). Note however that since they are generally coupled to higher order moments for which the evolution equations derived from the CFPE and CME do not agree, the first two moments (and higher order moments) of the CFPE do generally not agree with the ones of the CME. However, since the moment equations decouple for linear systems as pointed out in Section 2.3.4, we obtain the important result that *the moments of up to order two of the CFPE and CME agree exactly for linear reaction systems*.

The CFPE in Eq. (2.58) is generally a non-linear Fokker-Planck equation and no analytic solutions are known. However, since efficient stochastic simulations of the corresponding CLE in Eq. (2.61) exist, it is often more efficient to simulate the latter than the CME. However, simulations of the CLE will generally still be computationally significantly more expensive than numeric solutions of moment closure approximations introduced in Section 2.3.4. The advantage over moment closure approximations is that the CLE approximates the whole stochastic process and its distribution, rather than only the first few moments of the process.

Example

Let us come back to the gene expression system in Figure 2.2 with reactions in Eq. (2.23) and consider the corresponding CFPE and CLE. Using the stoichiometric matrix and propensity vector in Eq. (2.29) we obtain for the drift vector and diffusion matrix defined in Eq. (2.59) and Eq. (2.60), respectively,

$$\mathbf{A}(\mathbf{x}) = \begin{pmatrix} -\frac{k_2}{\Omega} x_1 x_2 + k_3(1-x_1) \\ k_1 x_1 - \frac{k_2}{\Omega} x_1 x_2 + k_3(1-x_1) - k_4 x_2 \end{pmatrix}, \tag{2.63}$$

$$\mathbf{B}(\mathbf{x}) = \begin{pmatrix} \frac{k_2}{\Omega} x_1 x_2 + k_3(1-x_1) & \frac{k_2}{\Omega} x_1 x_2 + k_3(1-x_1) \\ \frac{k_2}{\Omega} x_1 x_2 + k_3(1-x_1) & k_1 x_1 + \frac{k_2}{\Omega} x_1 x_2 + k_3(1-x_1) + k_4 x_2 \end{pmatrix}, \tag{2.64}$$

where x_1 and x_2 are the (continuous) particle numbers of G^{on} and P, respectively. To obtain the CLE in Eq. (2.61), we have to compute $\mathbf{C}$, which is the square root of $\mathbf{B}$ and thus generally not uniquely defined. In the standard form given in Eq. (2.62) it reads

$$C_{ij}(\mathbf{x}) = \begin{pmatrix} 0 & -\sqrt{\frac{k_2}{\Omega} x_1 x_2} & \sqrt{k_3(1-x_1)} & 0 \\ \sqrt{k_1 x_1} & -\sqrt{\frac{k_2}{\Omega} x_1 x_2} & \sqrt{k_3(1-x_1)} & -\sqrt{k_4 x_2} \end{pmatrix}. \tag{2.65}$$

We thus obtain the CLE

$$dx_1 = (-\frac{k_2}{\Omega}x_1x_2 + k_3(1-x_1))dt - \sqrt{\frac{k_2}{\Omega}x_1x_2}dW_2 + \sqrt{k_3(1-x_1)}dW_3, \tag{2.66}$$

$$\begin{aligned} dx_2 = {} & (k_1x_1 - \frac{k_2}{\Omega}x_1x_2 + k_3(1-x_1) - k_4x_2)dt \\ & + \sqrt{k_1x_1}dW_1 - \sqrt{\frac{k_2}{\Omega}x_1x_2}dW_2 + \sqrt{k_3(1-x_1)}dW_3 - \sqrt{k_4x_2}dW_4, \end{aligned} \tag{2.67}$$

where the W_i are independent Wiener processes. Note that it does not make a difference if one changes the signs in front of the square roots in Eq. (2.65), or equivalently in front of the noise terms in Eq. (2.66), as long as one does so simultaneously for each occurrence of a specific term, i.e., changes the sign of whole columns in Eq. (2.65). To see that such changes are equivalent note that the diffusion matrix $\mathbf{B} = \mathbf{C}\mathbf{C}^T$ of the CFPE is invariant under such changes. This can also be seen directly from the CLE in Eqs. (2.66) and (2.67) since the Wiener processes are symmetric with zero mean.

2.3.6 System size expansion

Let us next consider another popular approximation of the CME, the *system size expansion* (SSE). The SSE is a perturbative expansion of the CME in the inverse system size originally developed by van Kampen [61, 62]. The idea is to separate concentrations into a deterministic part, given by the solution of the macroscopic *rate equations*, and a part describing the fluctuations about the deterministic part.

Deterministic rate equations

The starting point for the SSE are the macroscopic rate equations. For a chemical reaction system as in Eq. (2.22) these read

$$\partial_t \phi_i = \sum_{r=1}^{R} S_{ir} f_r^{(0)}(\boldsymbol{\phi}), \tag{2.68}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$, ϕ_i is the concentration of species X_i and $f_r^{(0)}(\boldsymbol{\phi})$ is the *macroscopic propensity function* of the r^{th} reaction. $f_r^{(0)}(\boldsymbol{\phi})$ is given in terms of the microscopic propensity function defined in Eq. (2.25) as

$$f_r^{(0)}(\boldsymbol{\phi}) = \lim_{\Omega\to\infty} \frac{f_r(\Omega\boldsymbol{\phi})}{\Omega}, \tag{2.69}$$

where Ω is the system size. For monostable systems, i.e., systems whose rate equations possess exactly one positive stable fixed point, the mean value of the CME converges to

the rate equations result in the limit of infinite system sizes $\Omega \to \infty$, and the fluctuations around the mean converge to zero. In this sense the rate equations in Eq. (2.68) can be thought of as a macroscopic description of chemical reaction systems.

General derivation

The system size expansion makes the ansatz to split the instantaneous particle numbers n_i in the CME into a deterministic part and a fluctuating part as

$$\frac{n_i}{\Omega} = \phi_i + \Omega^{-1/2}\varepsilon_i, \quad i = 1, \ldots, N, \tag{2.70}$$

where ϕ_i is the solution of the deterministic rate equations in Eq. (2.68) and we introduced the new variables ε_i representing fluctuations about the deterministic mean. For the system size expansion to be applicable we have to assume that the microscopic propensity functions in the CME can be expanded as

$$f_r(\mathbf{n}) = \Omega \sum_{i=0}^{\infty} \Omega^{-i} f_r^{(i)}\left(\tfrac{\mathbf{n}}{\Omega}\right). \tag{2.71}$$

If we assume mass-action kinetics for which the propensity functions take the form in Eq. (2.25), it is easy to see that such an expansion always exists. Note that the first term in the sum in Eq. (2.71) is just the macroscopic propensity function of the r^{th} reaction already defined in Eq. (2.69). Next, we perform a transformation of variables from the n_i variables to the new ε_i variables. For the distribution $\Pi(\boldsymbol{\varepsilon}, t)$ of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)$ this means

$$\Pi(\boldsymbol{\varepsilon}, t) = \Omega^{N/2} P(\Omega\boldsymbol{\phi} + \Omega^{1/2}\boldsymbol{\varepsilon}, t). \tag{2.72}$$

We have now all the ingredients we need to perform the system size expansion. Plugging Eqs. (2.70), (2.71) and (2.72) into the CME in Eq. (2.24) and performing a Taylor expansion around $\boldsymbol{\varepsilon} = 0$ one obtains an expansion of the CME in powers of $\Omega^{-1/2}$.

The linear noise approximation

If we truncate this expansion to zeroth order we obtain the so-called *linear noise approximation* (LNA) [61, 62],

$$\partial_t \Pi(\boldsymbol{\varepsilon}, t) = \Big[-\sum_{i=1}^{N} \frac{\partial}{\partial \varepsilon_i} \sum_{j=1}^{N} J_i^j \varepsilon_j + \frac{1}{2} \sum_{i,j=1}^{N} D_{ij} \frac{\partial}{\partial \varepsilon_i} \frac{\partial}{\partial \varepsilon_j} \Big] \Pi(\boldsymbol{\varepsilon}, t) + O(\Omega^{-1/2}), \tag{2.73}$$

where we have defined

$$D_{ij} = \sum_{r=1}^{R} S_{ir} S_{jr} f_r^{(0)}(\boldsymbol{\phi}), \quad J_i^l = \frac{\partial}{\partial \phi_l} \sum_{r=1}^{R} S_{ir} f_r^{(0)}(\boldsymbol{\phi}). \tag{2.74}$$

Note that both $\mathbf{J}$ and $\mathbf{D}$ depend on the solution $\boldsymbol{\phi}$ of the rate equations in (2.68) and are thus generally time dependent. Looking at Eq. (2.68) we find that $\mathbf{J}$ in Eq. (2.74) is just the Jacobian of the rate equations. Eq. (2.73) is a linear Fokker-Planck equation and is hence solved by a multivariate normal distribution under appropriate initial conditions. By multiplying Eq. (2.73) with ε_i and $\varepsilon_i \varepsilon_j$ and integrating over all $\boldsymbol{\varepsilon}$, one obtains ODEs for the first and second moments, $\langle \varepsilon_i \rangle$ and $\langle \varepsilon_i \varepsilon_j \rangle$, respectively. By doing so one finds that if the mean is initially zero, $\langle \varepsilon_i \rangle|_{t=0} = 0$, it remains zero for all times. Since $\boldsymbol{\varepsilon}$ describes the fluctuations around $\boldsymbol{\phi}$, it is reasonable to assume this to be the case. The solution of Eq. (2.73) is thus a multivariate normal distribution with zero mean. Since $\boldsymbol{n}$ and $\boldsymbol{\varepsilon}$ are related by a linear transformation given in Eq. (2.70), the distribution of $\boldsymbol{n}$ is also given by a multivariate normal distribution. The mean of the latter satisfies the rate equations in Eq. (2.68) and the covariance $\boldsymbol{\Sigma}$ fulfils

$$\partial_t \boldsymbol{\Sigma} = \boldsymbol{J}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{J}^T + \Omega^{-1}\boldsymbol{D}. \tag{2.75}$$

The linear noise approximation thus describes the lowest order fluctuations about the deterministic mean.

For linear reaction systems, the mean and variance predicted by the LNA agree exactly with the mean and variance predicted by the CME. For non-linear systems, however, this is generally not the case, and it has been shown that the deviations of the LNA as compared to the CME can become quite significant for realistic systems [63, 64]. It is therefore appealing to consider higher order corrections beyond the LNA.

Higher order corrections

For the derivation of higher order corrections of the system size expansion, we simply need to keep higher order terms in the expansion of the CME beyond the LNA in Eq. (2.73). The resulting PDEs can generally not be solved analytically anymore. However, it is straightforward to derive ODEs for the moments of the system accurate to the corresponding order. If one includes the leading order correction to the mean concentrations given by the rate equations, the corresponding equations have been called “effective mesoscopic rate equations” [63], and the next leading order corrections to the variance “Inverse Omega Square” method [65]. We will not use these

higher order corrections in this work and refer the interested reader to the references [63] and [65].

Properties

As long as one is only interested in the LNA or higher order corrections to the moments, rather than the distributions of higher order truncations, the system size expansion amounts to the solution of finite sets of ODEs. Efficiency wise, it is thus comparable with moment closure approximations introduced in Section 2.3.4 and hence generally superior to stochastic simulations of the CME or of the chemical Langevin equation introduced in Section 2.3.5.

Since the system size expansion is an expansion around the deterministic mean, it cannot be used for deterministically multistable systems, i.e., systems whose rate equations have two or more positive stable fixed points, unless one is only interested in the short-time behaviour of a process.

We note that for truncations of higher orders than zero, the resulting PDEs corresponding to Eq. (2.73) involve higher order derivatives and thus have no probabilistic interpretation due to non positive-definite solutions [66]. It can however be shown that the moment equations for the system size expansion can be derived directly from the moment equations of the CME, thereby avoiding this flaw of negative distributions.

Example

Let us consider the rate equations and LNA for the gene system in Figure 2.2. Using the definition of the macroscopic propensity functions in Eq. (2.69) together with Eq. (2.29), we obtain

$$\mathbf{f}^{(0)}(\boldsymbol{\phi}) = (k_1\phi_1, k_2\phi_1\phi_2, k_3(\tfrac{1}{\Omega} - \phi_1), k_4\phi_2)^T. \tag{2.76}$$

Using Eqs. (2.68) and (2.29) we obtain the rate equations for this system,

$$\partial_t\phi_1 = -k_2\phi_1\phi_2 + k_3(\tfrac{1}{\Omega} - \phi_1), \tag{2.77}$$

$$\partial_t\phi_2 = k_1\phi_1 - k_2\phi_1\phi_2 + k_3(\tfrac{1}{\Omega} - \phi_1) - k_4\phi_2. \tag{2.78}$$

For the LNA defined in Eq. (2.73) we need the matrices $\mathbf{J}$ and $\mathbf{D}$ defined in Eq. (2.74), for which we obtain

$$\mathbf{D} = \begin{pmatrix} k_2\phi_1\phi_2 + k_3(\frac{1}{\Omega} - \phi_1) & k_2\phi_1\phi_2 + k_3(\frac{1}{\Omega} - \phi_1) \\ k_2\phi_1\phi_2 + k_3(\frac{1}{\Omega} - \phi_1) & k_1\phi_1 + k_2\phi_1\phi_2 + k_3(\frac{1}{\Omega} - \phi_1) + k_4\phi_2 \end{pmatrix}, \tag{2.79}$$

$$\mathbf{J} = \begin{pmatrix} -k_2\phi_2 - k_3 & -k_2\phi_1 \\ k_1 - k_2\phi_1 - k_3 & -k_2\phi_1 - k_4 \end{pmatrix}. \tag{2.80}$$

Note that $\mathbf{D}$ and $\mathbf{J}$ are functions of the time-dependent solutions ϕ_1 and ϕ_2 of the rate equations in (2.77) and (2.78). The solution of the LNA is a normal distribution in $\mathbf{n} = (n_1, n_2)$. Its mean is obtained by (numerically) solving the rate equations in (2.77) and (2.78), and the covariance by subsequently solving Eq. (2.75) using Eqs. (2.79) and (2.80).

2.3.7 The Poisson representation

After having discussed several approximation methods of the CME in the previous sections, let us next discuss an elegant (exact) alternative formulation of the CME, the "Poisson representation" (PR). The PR was first derived by Gardiner in 1977 [67]. Despite its elegance it has rarely been used for actual applications in the literature. It will however prove to be indispensable for our work in Chapter 5.

General formulation

The basic ansatz underlying the PR is to write the solution $P(\mathbf{n},t)$ of the CME as a mixture of Poisson distributions [67],

$$P(\mathbf{n},t) = \int d\mathbf{u}\, \mathcal{P}(n_1;u_1)\ldots\mathcal{P}(n_N;u_N)p(\mathbf{u},t), \quad u_i \in \mathbb{C}, \tag{2.81}$$

where $\mathbf{u} = (u_1,\ldots,u_N)$ and $\mathcal{P}(n_i;u_i) = (e^{-u_i}u_i^{n_i})/n_i!$ is a Poisson distribution in n_i. The u_i are generally complex-valued and the integrals in Eq. (2.81) run over the whole complex plane for each u_i. Starting from the CME, one can derive a time evolution equation for the *generating function* of $P(\mathbf{n},t)$ [49]. Using the ansatz in Eq. (2.81) in this equation leads to [67]

$$\partial_t p(\mathbf{u},t) = \sum_{r=1}^{R} \Omega k_r \left(\prod_{i=1}^{N} \left(1 - \frac{\partial}{\partial u_i}\right)^{r_{ir}} - \prod_{i=1}^{N} \left(1 - \frac{\partial}{\partial u_i}\right)^{s_{ir}} \right) \prod_{j=1}^{N} \Omega^{-s_{jr}} u_j^{s_{jr}} p(\mathbf{u},t). \tag{2.82}$$

We note that for the derivation of Eq. (2.82), an integration by parts has to be performed and a corresponding surface term needs to be dropped. The latter is only possible if

$p(\mathbf{u},t)$ is sufficiently compact and has to be checked a posteriori in general. Note that the PDE in Eq. (2.82) involves derivatives of finite order but may involve higher orders than two. In this case, Pawula's theorem tells us that $p(\mathbf{u},t)$ can generally become negative which means that Eq. (2.82) does not admit a probabilistic interpretation [66]. However, if we consider only reactions satisfying

$$\sum_i s_{ir} \leq 2, \quad \sum_i r_{ir} \leq 2, \quad r = 1,\ldots,R, \tag{2.83}$$

i.e., a maximum of two reactant and a maximum of two product particles, respectively, no derivatives of higher order than two arise and Eq. (2.82) simplifies to a Fokker-Planck equation of the form

$$\partial_t p(\mathbf{u},t) = -\sum_{i=1}^{N} \frac{\partial}{\partial u_i}\left[A_i(\mathbf{u})p(\mathbf{u},t)\right] + \frac{1}{2}\sum_{i,j=1}^{N} \frac{\partial}{\partial u_i}\frac{\partial}{\partial u_j}\left[B_{ij}(\mathbf{u})p(\mathbf{u},t)\right]. \tag{2.84}$$

The drift vector $\mathbf{A}(\mathbf{u})$ and diffusion matrix $\mathbf{B}(\mathbf{u})$ are given by

$$A_i(\mathbf{u}) = \sum_{r=1}^{R} S_{ir} g_r(\mathbf{u}), \tag{2.85}$$

$$B_{ij}(\mathbf{u}) = \sum_{r=1}^{R} g_r(\mathbf{u})(r_{ir}r_{jr} - s_{ir}s_{jr} - \delta_{i,j}S_{ir}), \tag{2.86}$$

$$g_r(\mathbf{u}) = \Omega k_r \prod_{j=1}^{N} \Omega^{-s_{jr}} u_j^{s_{jr}}, \tag{2.87}$$

where $\delta_{i,j}$ denotes the Kronecka delta, and we have introduced the *PR propensity functions* $g_r(\mathbf{u})$. We note that if a the r^{th} reaction does not have two or more reactant molecules of the same species, i.e., $s_{ir} \leq 1, i = 1,\ldots,N$, the PR propensity $g_r(\mathbf{u})$ is simply obtained from the CME propensity defined in Eq. (2.25) by replacing $n_i \to u_i$. A Fokker-Planck equation of the form in (2.84) is equivalent to the Langevin equation

$$d\mathbf{u} = \mathbf{A}(\mathbf{u})dt + \mathbf{C}(\mathbf{u})d\mathbf{W}, \quad \mathbf{C}\mathbf{C}^T = \mathbf{B}, \tag{2.88}$$

where $\mathbf{W}$ is a l-dimensional Wiener process and l is the number of columns of $\mathbf{C}$. By "equivalent" we mean that simulating Eq. (2.88) corresponds to drawing samples from the solution of Eq. (2.84).

Properties

Depending on the reactions in the system, the diffusion matrix $\mathbf{B}$ may be zero, in which case the Langevin equation in Eq. (2.88) reduces to a set of deterministic ODEs. On the

other hand, depending on the reactions, $\mathbf{B}(\mathbf{u})$ does not always remain positive semi-definite which means that $\mathbf{C}\mathbf{C}^T = \mathbf{B}$ cannot be fulfilled for real $\mathbf{u}$ (see Appendix A.2 for a proof). This in turn means that Eq. (2.84) is not a proper Fokker-Planck equation in real variables, but has to be extended to complex-valued space.

Note also that despite Eqs. (2.84) and (2.88) looking similar to the chemical Fokker-Planck equation and chemical Langevin equation in Eqs. (2.58) and (2.61), respectively, they have fundamentally different interpretations: the u_i variables in the Poisson representation are auxiliary variables with no real physical interpretation. In the CFPE and CLE in Eqs. (2.58) and (2.61), in contrast, the x_i variables represent physical quantities, namely (continuous) molecule numbers.

An important property of the PR is that the mean values of the u_i variables *agree exactly* with the mean values of the n_i variables. More generally, one can derive an exact relationship between the moments of the u_i variables and the moments of the n_i variables [49]. This means that if one computes moments in PR space up to a certain order, one can easily translate them into moments of molecule numbers, without having to know the whole distribution of the u_i variables.

Example

Let us next derive the PR Langevin equation for the gene system in Figure 2.2. The PR propensity vector $\mathbf{g}$, drift vector $\mathbf{A}$ and diffusion matrix $\mathbf{B}$ are defined in Eqs. (2.85)-(2.87). Using Eqs. (2.28) and (2.29), we obtain

$$\mathbf{g} = (k_1 u_1, \frac{k_2}{\Omega} u_1 u_2, k_3(1-u_1), k_4 u_2)^T, \tag{2.89}$$

$$\mathbf{A} = \begin{pmatrix} -\frac{k_2}{\Omega} u_1 u_2 + k_3(1-u_1) \\ k_1 u_1 - \frac{k_2}{\Omega} u_1 u_2 + k_3(1-u_1) - k_4 u_2 \end{pmatrix}, \tag{2.90}$$

$$\mathbf{B} = (k_1 u_1 - \frac{k_2}{\Omega} u_1 u_2 + k_3 u_3) \times \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{2.91}$$

Since the matrix in the last equation has the eigenvalues $+1$ and -1, $\mathbf{B}$ cannot be positive semi-definite, which means that the PR becomes complex-valued. For the PR Langevin equation in Eq. (2.88) we have to take the square root of $\mathbf{B}$. One possible choice is

$$\mathbf{C} = \frac{1}{2}\sqrt{k_1 u_1 - \frac{k_2}{\Omega} u_1 u_2 + k_3 u_3} \begin{pmatrix} 1+i & 1-i \\ 1-i & 1+i \end{pmatrix}. \tag{2.92}$$

Note that $\mathbf{C}$ cannot be chosen real since $\mathbf{B}$ is not positive semi-definite (see Appendix A.2 for a proof).

2.4 Stochastic reaction-diffusion processes

Let us next consider the case when the "well-mixed" condition underlying the previous section does not hold, i.e., if the diffusion of particles is not fast. In this case the spatial positions and diffusion of particles have to be taken into account, and the non-spatial CME description introduced in Section 2.3.2 is no longer valid. Such systems are typically modelled as *stochastic reaction-diffusion processes* (SRDPs). We will in the following first discuss an algorithmic description of such processes and subsequently a coarse-grained approximation thereof.

2.4.1 Brownian dynamics in the Doi model

The standard approach to describe spatial chemical reaction systems relies on Brownian dynamics simulations. The idea is to discretise time into steps dt, to simulate the diffusion of each individual particle in space and to let reactions happen under certain rules with certain probabilities during each time step. Let us start by considering the spatial diffusion of particles.

Diffusion

Consider a single particle in an $M-$dimensional spatial domain $V \subset \mathbb{R}^M$ and let $p_{\text{diff}}(x,t|x^0,t^0)$ be the probability to find the particle at position x at time t if it was initially at position x^0 at time t^0. Then $p_{\text{diff}}(x,t|x^0,t^0)$ fulfils the diffusion equation

$$\partial_t p_{\text{diff}}(x,t|x^0,t^0) = D\Delta p_{\text{diff}}(x,t|x^0,t^0), \tag{2.93}$$

where D is the microscopic diffusion constant and $\Delta = \partial^2/\partial x_1^2 + \ldots + \partial^2/\partial x_M^2$ is the $M-$dimensional Laplace operator. If the system is unbounded, i.e., $V = \mathbb{R}^M$, the solution of Eq. (2.93) is given by a product of normal distributions with variance $2Ddt$, one for each spatial component:

$$p_{\text{diff}}(x,t^0+dt|x^0,t^0) = \prod_{i=1}^{M} \mathcal{N}(x_i;x_i^0,2Ddt), \tag{2.94}$$

$$\mathcal{N}(x;m,V) = \frac{1}{\sqrt{2\pi V}} e^{-\frac{(x-m)^2}{2V}}, \tag{2.95}$$

where we set $t = t^0 + dt$. In particular, this means that the diffusion in different spatial (orthogonal) directions is independent of each other. We can thus use Eq. (2.94) to

simulate the diffusion of particles in space for discrete time steps dt by sampling the traveled distance from a normal distribution,

$$x_i(t+dt) = x_i(t) + dx, \quad dx \sim \mathcal{N}(x; 0, 2Ddt), \tag{2.96}$$

independently for each particle and each spatial dimension $i = 1, \ldots, M$. The mean square displacement is accordingly given by $\langle dx^2 \rangle = 2Ddt$. If the system is (spatially) unbounded, Eq. (2.96) is exact for any dt in the sense that it draws samples from Eq. (2.94). However, if the system is bounded with say reflecting boundary conditions, the solution of Eq. (2.93) is no longer given by Eq. (2.94). For time steps dt small enough such that the expected overlap of Eq. (2.94) with the boundary is small, however, Eq. (2.94) can be expected to give a good approximation of the solution of Eq. (2.93). Roughly speaking, this is the case if the root of the mean square displacement, $\sqrt{\langle dx^2 \rangle} = \sqrt{2Ddt}$, is much smaller than the length scale of the system. Therefore, as long as we choose the time step dt sufficiently small, we can use Eq. (2.96) together with appropriate boundary conditions to simulate the diffusion of particles. If the updated position of a particle according to Eq. (2.96) lies outside of the compartment, we have to apply certain boundary constraints. In the case of reflecting boundaries, for example, we simply reflect the particle at the boundary back into the compartment.

Zeroth order reactions

Let us next consider zeroth order reactions, i.e., such with zero reactant molecules, such as $\varnothing \xrightarrow{c} X$. This could for example model a random input of particles into a system, or the particle could become created by other species in the system which are not modelled explicitly. Such a reaction may occur homogeneously across the whole system or be space dependent. Independently of that, for the purpose of simulation, one first has to decide if the reaction occurs in the time step dt or not. If c is the propensity of the reaction, then $c\delta t$ is the probability for it to happen in an infinitesimally small time interval δt. If we choose our time step dt small enough, cdt will therefore be a good approximation of the probability for the reaction to happen in dt. We implicitly make the assumption here that the reaction can occur only once in each time step which is obviously only valid for small enough dt. Specifically, the waiting time distribution of the reaction to happen is given by the exponential $\exp(-cdt)$ with expected value $1/c$, and we require $dt \ll 1/c$. In summary, the reaction occurs with probability cdt in a time step of length dt. If the reaction occurs, we sample from its spatial distribution to determine the position where it occurs.

First order reactions

First order reactions, i.e., such with one reactant molecule, are treated in a similar way. As an example, consider the decay reaction $X \xrightarrow{c} \varnothing$. One possibility to simulate such a reaction is to decide for each X particle separately if the reaction occurs with probability cdt. This is quite inefficient, however, because one needs to simulate a random number for each X molecule in the system in each time step. Alternatively, one can decide if the reaction occurs at all for any of the X molecules in the system with probability $\#X \times cdt$, where $\#X$ is the current number of X molecules in the system. As for zeroth order reactions, this approach is only valid if the probability for a reaction to happen in dt is small and the probability for the reaction to happen more then once thus negligible. If the reaction does occur in a given time step, we decide randomly for which of the X molecules it occurs. Since the propensity $\#X \times cdt$ now depends on the number of X molecules in the system, it is difficult to verify these criteria a priori, and they may have to be verified a posteriori.

Bimolecular reactions

Finally, let us consider bimolecular reactions, i.e., such with two reactant molecules, which introduce explicit spatial interactions between particles. The particular form of the interactions needs to be specified, and there are different ways to do so. We will work in the Doi picture in this work [34, 35], in which molecules are modelled as point particles and bimolecular reactions occur with a certain probability per unit time whenever the distance between two reactant molecules is smaller than a certain reaction range. Let us take the dimerisation $X + X \to \varnothing$ as an example, and let r and c be the corresponding reaction range and reaction rate constant, respectively. In each simulation step, one needs to find all pairs of X molecules that are separated from each other by less than r. For each such pair, the reaction then occurs with probability cdt. Here, the time step dt needs to be small enough such that the probability of a molecule to react with more than one other molecule is negligible. As before, this generally needs to be verified a posteriori.

Additional remarks

Note that certain reactions may require additional spatial rules to be specified. For the reaction $X_1 \to X_2 + X_2$, for example, one needs to specify where the two product molecules should be placed relatively to the position of the reacting X_1 molecule. One

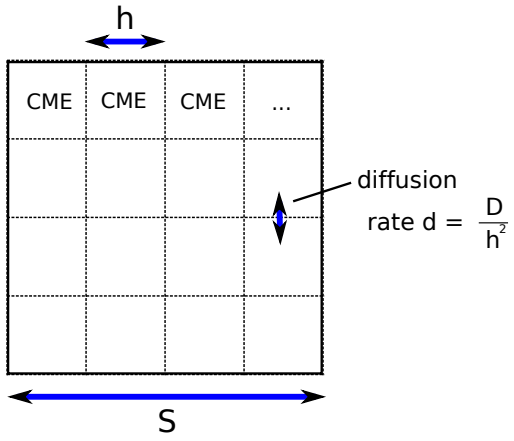

Figure 2.3: Visualisation of the RDME for a two-dimensional square system of edge length S. The RDME splits up space into square compartments of edge length h and assumes well-mixed and dilute conditions in each compartment. The dynamics in each compartment can therefore be described by a CME of the form in Eq. (2.24). The CMEs of the single compartments are coupled via diffusion, which is modelled as linear reactions with rate $d = D/h^2$, where D is the microscopic diffusion rate constant.

possibility is to create both X_2 particles at the exact position of X_1. Another possibility would be to create them at a fixed distance opposite of each other relative to the X_1 molecule at a random angle. If bimolecular reactions are involved and if they are diffusion limited, such different definitions can have dramatic effects on the dynamics of the system. However, the algorithm described above that allows to decide if a reaction happens in each step is valid independently of such additional rules. The latter just need to be implemented "on top" of the former.

For each of the reaction types, we formulated conditions that the time interval dt needs to satisfy for the algorithm to be valid. In a system with more than one reaction, one may obtain additional requirements. For example, dt has to be small enough such that a single X_1 molecule has negligible probability to undergo both the reactions $X_1 \to X_2$ and $X_1 \to X_3$ in one time step. It is not straightforward to verify all such criteria explicitly. In practise, it is generally easier to converge the quantity of interest, e.g. the mean particle number of a certain species, with respect to dt.

Since the system has to be simulated for small time steps dt and every molecule has to be modelled explicitly, such simulations in continuous-space are computationally extremely expensive, and several modified simulation methods have been proposed that aim at lower computational costs [68, 69]. An alternative method which we discuss

next relies on the discretisation of space.

2.4.2 The reaction-diffusion master equation

The *reaction-diffusion master equation* (RDME) can be viewed as a coarse-grained approximation of stochastic reaction-diffusion processes in continuous space which we discussed in Section 2.4.1. The basic idea is to discretise space into compartments and assume well-mixed and dilute conditions in each compartment. For simplicity, we consider a system in an M-dimensional cubic volume $V = [0,S]^M$ of edge length S and discretise it into cubic compartments of edge length h. We denote by $L = (S/h)^M$ the number of compartments and label each compartment with an index $i \in \{1,\dots,L\}$. We note that it is straightforward to generalise the following considerations to more complicated geometries for both the system V and the space discretisation. The RDME framework assumes that the system within each compartment is well-mixed and dilute, which means that the dynamics in each compartment can be described by a chemical master equation of the form in Eq. (2.24). The diffusion of particles is modelled by linear reactions between adjacent compartments with rate constant $d = D/h^2$, where D is the microscopic diffusion constant. Figure 2.3 visualises the idea of the RDME for a two-dimensional system.

Let $\mathbf{n} = (n_1^1,\dots,n_N^1,\dots,n_1^L,n_N^L)$ denote the state of the system, where n_i^l is the copy number of species X_i in compartment l, and let $\mathbf{n}^l = (n_1^l,\dots,n_N^l)$ denote the state vector of the l^{th} compartment. The time evolution of the probability $P(\mathbf{n},t)$ to be in state $\mathbf{n}$ at time t fulfils the RDME [70],

$$\begin{aligned}\partial_t P(\mathbf{n},t) = &\sum_{l=1}^{L}\sum_{m\in\mathcal{N}(l)}\sum_{i=1}^{N} d_i[(n_i^m+1)P(\mathbf{n}+\delta_i^m-\delta_i^l,t)-n_i^l P(\mathbf{n},t)]\\ &+\sum_{l=1}^{L}\sum_{r=1}^{R}[f_r(\mathbf{n}^l-\mathbf{S}_r)P(\mathbf{n}-\mathbf{S}_r^l,t)-f_r(\mathbf{n}^l)P(\mathbf{n},t)],\end{aligned} \tag{2.97}$$

where $\mathcal{N}(l)$ denotes all adjacent compartments of compartment l, $f_r(\mathbf{n}^l)$ is the propensity function of the r^{th} reaction, δ_i^l is a vector of length $N\times L$ with the entry corresponding to species i in box l equal to 1 and all other entries equal to zero and $\mathbf{S}_r^l$ is a vector of length $N\times L$ with the entries corresponding to the l^{th} compartment equal to $\mathbf{S}^r$ and zero otherwise. The latter denotes the r^{th} row of the stoichiometric matrix $\mathbf{S}$. Note that the propensity functions $f_r(\mathbf{n}^l)$ only depend on the state vector of single compartments. Importantly, if a propensity function f_r has a volume dependence, as do propensity functions of mass-action kinetics type defined in Eq. (2.25), for example, the volume

of the corresponding compartment has to be taken here, rather than the volume of the whole system. The first term in Eq. (2.97) describes the diffusion of particles, while the second term is a sum over CME type of terms for each compartment.

The RDME can be seen as a coarse-grained approximation of the Doi model [71]. For a linear reaction system, i.e., in the absence of bimolecular reactions, it constitutes a perturbative approximation of the corresponding continuous system, in the sense that the RDME converges to the latter in the continuum limit $h \to 0$. In the presence of bimolecular reactions, however, this is not the case in systems with more than one spatial dimension [71]. The reason is that in the RDME framework bimolecular reactions only happen if both reactant molecules are positioned in the same compartment. As the compartment size approaches zero, the probability to find two molecules in the same compartment converges to zero and no bimolecular reactions occur, independently of the scaling of the propensity function with the compartment volume [71]. This leads to a lower bound on the compartment spacing [72].

As for the CME, there are generally no analytic solutions known for the RDME and one has to rely on further approximations or stochastic simulations. For the purpose of simulations, the RDME has the advantage over Brownian dynamics simulations that there is no need for time discretisation or the simulation of individual particles. By treating molecules of the same species positioned in different compartments as different species, one can write the RDME in Eq. (2.97) as a CME. Therefore, the stochastic simulation algorithm, as well as the approximation methods of the CME introduced in Section 2.3, can be applied to the RDME. Using the stochastic simulation algorithm or an advanced version thereof (c.f. Section 2.3.2) directly on the RDME is generally highly inefficient, however, since the sparse structure of the RDME is not taken into account, and significant effort has been spent in the literature on the development of more efficient simulation methods for the RDME [73, 74].

2.5 Poisson and Cox processes

Let us now leave the field of chemical reaction systems and consider a class of models that seems to be quite unrelated at first sight; so-called *spatio-temporal point processes*. The latter are typically used empirically to perform inference tasks for systems with spatial (temporal) events for which no microscopic description exists, such as conflict events in social sciences for example [75]. Even though stochastic reaction-diffusion processes (SRDPs) (Section 2.4) and point processes have both been applied

to the same type of problems, for example in epidemiology [76, 77], they are widely perceived as conceptually distinct. However, we will show in Chapter 5 that the two approaches are intimately related.

2.5.1 Definition

A *(spatial) Poisson process* on a spatial region $\mathcal{D}$ of arbitrary dimension defines a measure on countable unions of zero-dimensional subsets (points) of $\mathcal{D}$. A Poisson process is often characterised by an intensity function $u : \mathcal{D} \to \mathbb{R}^+$ giving the probability density of finding a point in an infinitesimal region around some $x \in \mathcal{D}$. Now let $N(A)$ denote the number of points in a subregion $A \subset \mathcal{D}$. Then $N(A)$ is a Poisson random variable with mean given by the integral of u over A:

$$p(N(A) = n) = \mathcal{P}(n; u_A), \quad u_A = \int_{u_A} dx\, u(x), \tag{2.98}$$

where $\mathcal{P}(n; u_A)$ is a Poisson distribution in n with mean u_A.

A *(spatial) Cox process* is a generalisation of a Poisson process and also called "doubly stochastic process", in the sense that the intensity function u is itself a random process, and conditioned on the intensity u a Cox process reduces to a Poisson process. The distribution of the number of points in a subregion $A \subset \mathcal{D}$ is thus a mixture of Poisson distributions,

$$p(N(A) = n) = \int du_A \mathcal{P}(n; u_A) p(u_A). \tag{2.99}$$

Since we are interested in dynamical systems, we will assume time-dependent intensity functions $u : \mathcal{D} \times \mathcal{T} \to \mathbb{R}^+$, where $\mathcal{T}$ is a finite real interval denoting time. We then require that for any fixed time point $t \in \mathcal{T}$ the process is a spatial Poisson (Cox) process with intensity $u(\cdot, t)$. In the case of a Poisson (Cox) process, the intensity u may for example be defined as the solution of a PDE (SPDE). Note that, strictly speaking, our definition does not describe "spatio-temporal point processes" in the conventional sense, which would be defined as point processes on the product space $\mathcal{D} \times \mathcal{T}$.

2.5.2 Numerical solution of (S)PDEs via basis projection

Let us next discuss a method to numerically solve (S)PDEs, which we will make use of in the context of (Cox) Poison processes in Chapter 5. The basic idea is to project an infinite dimensional (S)PDE in continuous space onto a finite set of spatial basis functions. This gives rise to a finite set of coupled (S)ODEs. For simplicity, we confine the

following presentation to a one-dimensional and one-species system. It is straightforward to extend the presented methods to multi-dimensional and multi-species systems. Consider an SPDE of the form

$$du(x,t) = A(x,t) + \sqrt{C(x,t)}dW(x,t), \tag{2.100}$$

where $A(x,t)$ and $C(x,t)$ are polynomials in $u(x,t)$ with potentially space-dependent coefficients, and $W(x,t)$ is a spatio-temporal Wiener process. We approximate $u(x,t)$ by a linear-combination of a finite set of basis functions $\{\phi_i(x)\}_{i=1}^n$,

$$u(x,t) = \sum_{i=1}^{n} c_i(t)\phi_i(x), \tag{2.101}$$

where we have introduced the time-dependent coefficients $c_i(t)$. Inserting this ansatz into Eq. (2.100), multiplying from the left with ϕ_j and integrating over x, it can be shown that the parameter vector $\mathbf{c}(t) = (c_1(t), \dots c_n(t))$ fulfils

$$d\mathbf{c}(t) = \Phi^{-1}\langle\phi|A\rangle dt + \sqrt{\Phi^{-1}\langle\phi|C|\phi\rangle\Phi^{-1}}d\mathbf{W}, \tag{2.102}$$

where $\mathbf{W}$ is a n-dimensional temporal Wiener process and we have defined

$$\langle\phi_i|f\rangle = \int dx\ \phi_i(x)f(x,t), \tag{2.103}$$

$$\langle\phi_i|f|\phi_j\rangle = \int dx\ \phi_i(x)f(x,t)\phi_j(x), \tag{2.104}$$

$$\langle\phi|f\rangle_i = \langle\phi_i|f\rangle, \tag{2.105}$$

$$\langle\phi|f|\phi\rangle_{ij} = \langle\phi_i|f|\phi_j\rangle, \tag{2.106}$$

$$\Phi_{ij} = \langle\phi_i|\phi_j\rangle, \tag{2.107}$$

for a general function $f(x,t)$. Eq. (2.102) constitutes a finite set of coupled SDEs. In the case of a Poisson process, i.e., $C(x,t) = 0$, Eq. (2.102) reduces to a finite set of ODEs and can hence be integrated numerically. In the case of a Cox process with $C(x,t) \neq 0$, Eq. (2.102) contains non-vanishing noise terms. If the system is linear, i.e., $A(x,t)$ is linear in $u(x,t)$ and $C(x,t)$ independent of $u(x,t)$, the system of SDEs is solved by a multivariate normal distribution whose mean and covariance can be obtained from direct numerical integration. If $C(x,t)$ also depends linearly on $u(x,t)$, the mean and covariance can still be obtained from direct numerical integration, but the solution of Eq. (2.102) is not a multivariate normal distribution anymore. A simple approximation is given by approximating the solution by a multivariate normal distribution with the corresponding mean and covariance. If $A(x,t)$ or $C(x,t)$ is a second or higher order

polynomial in $u(x,t)$, the moment equations couple to higher order equations, and further approximations may have to be applied. For the methods employed in this work this will not be necessary, however.

Locally constant non-overlapping basis functions

As an example, consider locally constant non-overlapping step functions for the ϕ_i. For a one-dimensional system in the interval $[0,1]$, for example, we define n basis functions as

$$\psi(x) = \begin{cases} 1 & 0 \le x \le \frac{1}{n}, \\ 0 & \text{otherwise}, \end{cases} \tag{2.108}$$

$$\phi_i(x) = \psi(x-(i-1)/n) \quad \text{for} \quad i = 1,\dots n. \tag{2.109}$$

The corresponding overlap matrix and diffusion operator read

$$\Phi = \langle\phi|\phi\rangle = \frac{1}{n}\mathbb{1}^{n\times n}, \tag{2.110}$$

$$\langle\phi|\Delta|\phi\rangle = n\begin{pmatrix} -1 & 1 & & & & & & \\ 1 & -2 & 1 & & & & & \\ & 1 & -2 & \ddots & & & & \\ & & & \ddots & & & & \\ & & & \ddots & -2 & 1 & & \\ & & & & 1 & -2 & 1 & \\ & & & & & 1 & -1 & \end{pmatrix}, \tag{2.111}$$

where $\mathbb{1}^{n\times n}$ is the n-dimensional unity matrix and Δ is the Laplace operator.

2.5.3 The likelihood function

Point processes are typically used to perform inference from experimental data. Suppose we have some spatial point measurements $\mathbf{x} = (\mathbf{x}_0,\dots,\mathbf{x}_n)$ of the state $\mathbf{z} = (\mathbf{z}_0,\dots,\mathbf{z}_n)$ of the system at discrete time points $t_0,\dots,t_n$. Suppose further that we model the system by means of a Poisson (Cox) process with intensity $u(x,t)$ which we obtain by solving a (S)PDE via a basis projection approximation as defined in Eq. (2.101). This means that $u(x,t)$ can be represented by the coefficient vector $\mathbf{c}(t) = (c_1(t),\dots c_n(t))$. We can thus identify $\mathbf{z}_i = \mathbf{c}_i = \mathbf{c}(t_i)$. Let Θ represent the used model. For a Poisson process, the likelihood of one measurement $\mathbf{x}_i$ at time t_i is given by [78]

$$p(\mathbf{x}_i|\Theta) = = \prod_{s\in\mathbf{x}_i} u(s,t_i)e^{-\int dx u(x,t_i)}. \tag{2.112}$$

In the case of a Cox process, for which $u(x,t)$ is stochastic, Eq. (2.112) has to be averaged accordingly with respect to the distribution of $u(x,t)$.

To perform inference we need to compute the likelihood $p(\mathbf{x}|\Theta)$ of the full data given the model. The likelihood factorises as

$$p(\mathbf{x}|\Theta) = p(\mathbf{x}_0)\prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{x}_{i-1},\dots,\mathbf{x}_0), \tag{2.113}$$

where we suppressed conditioning on Θ on the right hand side for notional convenience and will implicitly assume all distributions to be conditioned on Θ in the following. The factors in Eq. (2.113) can be written as

$$\begin{aligned} p(\mathbf{x}_i|\mathbf{x}_{i-1},\dots\mathbf{x}_0) &= \int d\mathbf{c}_i d\mathbf{c}_{i-1} p(\mathbf{x}_i|\mathbf{c}_i)p(\mathbf{c}_i|\mathbf{c}_{i-1})p(\mathbf{c}_{i-1}|\mathbf{x}_{i-1},\dots\mathbf{x}_0,) \\ &= \int d\mathbf{c}_i p(\mathbf{x}_i|\mathbf{c}_i)p(\mathbf{c}_i|\mathbf{x}_{i-1},\dots\mathbf{x}_0), \end{aligned} \tag{2.114}$$

and can be computed iteratively as follows. Assume that we know the *posterior* $p(\mathbf{c}_{i-1}|\mathbf{x}_{i-1},\dots\mathbf{x}_0)$ at time step $i-1$, and that we can solve the system forward in time to obtain the *predictive distribution* $p(\mathbf{c}_i|\mathbf{x}_{i-1},\dots\mathbf{x}_0) = \int d\mathbf{c}_{i-1} p(\mathbf{c}_i|\mathbf{c}_{i-1})p(\mathbf{c}_{i-1}|\mathbf{x}_{i-1},\dots\mathbf{x}_0)$ of time step i. The posterior of time step i is then obtained by performing the *Bayesian measurement update*

$$\begin{aligned} p(\mathbf{c}_i|\mathbf{x}_i,\dots\mathbf{x}_0) &= \frac{p(\mathbf{x}_i|\mathbf{c}_i,\mathbf{x}_{i-1},\dots\mathbf{x}_0)p(\mathbf{c}_i|\mathbf{x}_{i-1},\dots\mathbf{x}_0)}{p(\mathbf{x}_i|\mathbf{x}_{i-1},\dots\mathbf{x}_0)} \\ &= \frac{p(\mathbf{x}_i|\mathbf{c}_i)p(\mathbf{c}_i|\mathbf{x}_{i-1},\dots\mathbf{x}_0)}{p(\mathbf{x}_i|\mathbf{x}_{i-1},\dots\mathbf{x}_0)}, \end{aligned} \tag{2.115}$$

where we have used the Markov property to obtain the second line. We observe that the likelihood contribution in Eq. (2.113) at time step i is just the normalisation of the posterior at that time step. Therefore, if we can solve the system forward in time and perform the measurement update in Eq. (2.115), we can compute the likelihood iteratively. This procedure is called "filtering".

In the case of a Poisson process this procedure is particularly simple: since the intensity and thus the $\mathbf{c}_i$ are deterministic, the predictive distributions $p(\mathbf{c}_i|\mathbf{x}_{i-1},\dots\mathbf{x}_0)$ are delta functions, and hence the measurement updates in Eq. (2.115) become trivial. Effectively, one only needs to solve the system forward in time over the whole time interval, compute the likelihoods of single measurement time points as in Eq. (2.112) and take the product.

In the case of a Cox process, however, the measurement update in Eq. (2.115) is generally not trivial, and the posterior in Eq. (2.115) is in fact intractable in most cases. In the systems we study in this thesis, the solution of the SDE in Eq. (2.102) is

either a multivariate normal distribution or we approximate it by a multivariate normal distribution. In this case it is convenient to project the posteriors at different time steps back to multivariate normal distributions during the filtering procedure. A popular method to achieve this is the *Laplace approximation* [79], which approximates the posterior by a Gaussian centred at the posterior's maximum and with covariance being the negative Hessian of the posterior in the maximum.

Assuming that we can (approximately) compute the likelihood $p(\mathbf{x}|\Theta)$, the most straightforward way to perform e.g. parameter inference is to optimise $p(\mathbf{x}|\Theta)$ with respect to the parameters of the model Θ. We will adopt this approach in this work, but would like to point out that the computation of the likelihood is the basis for most more advanced inference schemes such as Bayesian methods [79].

We have now established all the machinery that we need in the following chapters, and are hence ready to present the results of this thesis.

Chapter 3

The complex chemical Langevin equation

In this Chapter we study the chemical Langevin equation (CLE) and chemical Fokker-Planck equation (CFPE), with particular regard to their boundary condition problem. We introduced the CLE/CFPE as a diffusion approximation of the chemical master equation (CME) in Section 2.3.5. Here, after defining the breakdown problem, we show that the CLE is mathematically not always well-defined due to the occurrence of square roots of negative expressions in Section 3.1. We call this phenomenon the *breakdown* of the CLE. Next, we show in Section 3.2 that the breakdown can be avoided by extending the state space of the CLE from real-valued to complex-valued numbers. In Section 3.3 we demonstrate that the resulting "complex CLE" is significantly more accurate than other modified versions of the CLE that have been proposed in the literature to keep the state space real-valued. Finally, we conclude in Section 3.4.

3.1 Breakdown of the chemical Langevin equation

In Section 2.3.5 we introduced the CLE/CFPE as a diffusion approximation of the CME, which can be obtained by letting the variables in the CME become continuous and performing a Taylor expansion to second order. This gives the CFPE

$$\partial_t P(\mathbf{x},t) = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i}\left[A_i(\mathbf{x})P(\mathbf{x},t)\right] + \frac{1}{2}\sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j}\left[B_{ij}(\mathbf{x})P(\mathbf{x},t)\right], \tag{3.1}$$

$$A_i(\mathbf{x}) = \sum_{r=1}^{R} S_{ir} f_r(\mathbf{x}), \quad B_{ij}(\mathbf{x}) = \sum_{r=1}^{R} S_{ir}S_{jr} f_r(\mathbf{x}), \tag{3.2}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ and x_i is the continuous molecule number variable of species X_i. As explained in Section 2.3.5, the CFPE in Eq. (3.1) is equivalent to the CLE

$$d\mathbf{x} = \mathbf{A}(\mathbf{x})dt + \mathbf{C}(\mathbf{x})d\mathbf{W}, \qquad \mathbf{C}(\mathbf{x})\mathbf{C}(\mathbf{x})^T = \mathbf{B}(\mathbf{x}), \tag{3.3}$$

where $\mathbf{W}$ is a multi-dimensional Wiener process. Generally there exist different choices for $\mathbf{C}(\mathbf{x})$ in Eq. (3.3). One such choice which we call the "standard form" of the CLE reads

$$dx_i = \sum_{r=1}^{R} S_{ir} f_r(\mathbf{x})dt + \sum_{r=1}^{R} S_{ir}\sqrt{f_r(\mathbf{x})}dW_r. \tag{3.4}$$

Note that the CLE/CFPE has continuous variables in contrast to the discrete variables of the CME. It is easy to see that the CME has a natural boundary at zero molecule numbers. Specifically, if the probability of finding a negative number of particles is zero initially, it remains zero for all times. It is not clear how this boundary condition behaves when approximating the discrete process described by the CME by a continuos diffusion process satisfying the CLE/CFPE in Eqs. (3.1) and (3.3), respectively. As pointed out in Chapter 1, it is known that the x_i variables can become negative for some systems, which can lead to negative propensity functions $f_r(\mathbf{x})$ and thus to square roots of negative expressions in the CLE. Since the CLE is defined for real-valued variables, it is not defined for such occurrences. We call the occurrence of square roots of negative expressions *breakdown* of the CLE. Let us next study this breakdown behaviour for some example systems.

3.1.1 Linear reaction systems

We start by considering two simple linear reaction systems. Recall that for linear systems, the CLE's predictions for the mean and variance of the process *agree exactly with the mean and variance predicted by the CME* (c.f. Section 2.3.5). This fundamental property of the CLE can be derived by comparing the moment equations derived from the CLE and CME, respectively. One may thus require any implementation or modified version of the CLE to satisfy this property.

Example (i): birth-death process

We start by considering a simple linear reaction system with one species

$$\varnothing \underset{c_2}{\overset{c_1}{\rightleftharpoons}} X. \tag{3.5}$$

The CLE for this system reads

$$dx = (\Omega c_1 - c_2 x)dt + \mathbf{C}(x)d\mathbf{W}, \tag{3.6}$$

where $\mathbf{C}(x)\mathbf{C}(x)^T = \mathbf{B} = \Omega c_1 + c_2 x$ and Ω is the compartment volume. We consider two forms of the CLE: the standard form with $C_{11} = \sqrt{\Omega c_1}$ and $C_{12} = -\sqrt{c_2 x}$ and a possible alternative form where $\mathbf{C}(x) = \sqrt{\Omega c_1 + c_2 x}$. We rescale time as $\tau = t c_2$ and define $k = \Omega c_1 / c_2$. Note that rescaling time also rescales the noise terms since from Ito calculus we have $\langle dW(\tau)\rangle = \sqrt{d\tau} = \sqrt{c_2 dt} = \sqrt{c_2}\langle dW(t)\rangle$ [49]. The two CLEs are then respectively given by

$$dx = (k - x)d\tau + \sqrt{k}dW_1 - \sqrt{x}dW_2, \tag{3.7}$$

$$dx = (k - x)d\tau + \sqrt{k + x}dW_1. \tag{3.8}$$

We first consider the standard CLE given by Eq. (3.7). Assume we start with a positive $x > 0$ at $\tau = 0$. The noise terms can drive the system towards $x = 0$. For $x = 0$ the second noise term vanishes and the drift becomes positive. However, due to the first noise term, the variable x becomes negative with a finite probability in a finite time interval and the CLE breaks down.

Next consider the alternative form of the CLE as given by Eq. (3.8). This CLE breaks down for $x < -k$. However, since the diffusion term vanishes for $x = -k$ and the drift term becomes $2k > 0$, the region $x < -k$ is not accessible and this CLE does hence not break down (note that since one has to numerically integrate the CLE with a finite time step, breakdown may still occur, but this purely numeric effect vanishes in the limit of infinitesimally small time steps).

We note that other alternative forms of the CLE than the one considered are possible by choosing different factorisations $\mathbf{C}$ of the Fokker-Planck diffusion matrix $\mathbf{B}$ in Eq. 3.3. However, it is easy to verify that all other possible choices of $\mathbf{C}$ give rise to CLEs for which the argument in the square roots becomes negative for $x < 0$ (as for the standard form of the CLE) or for $x < -k$ (as for the alternate form given by Eq. (3.8)). Hence the two cases considered above provide a complete picture of the breakdown phenomenon.

We call the implementations of the CLEs in Eq. (3.7) and Eq. (3.8), CLE-R1 and CLE-R2, respectively, and simulate them using the standard Euler-Maruyama algorithm [80]. For the CLE-R1, we impose a reflecting boundary at $x = 0$ to avoid the breakdown of the CLE for finite times. The simulation parameters are the time step of the Euler-Maruyama algorithm ($\delta\tau$), the time after which steady state is assumed to be

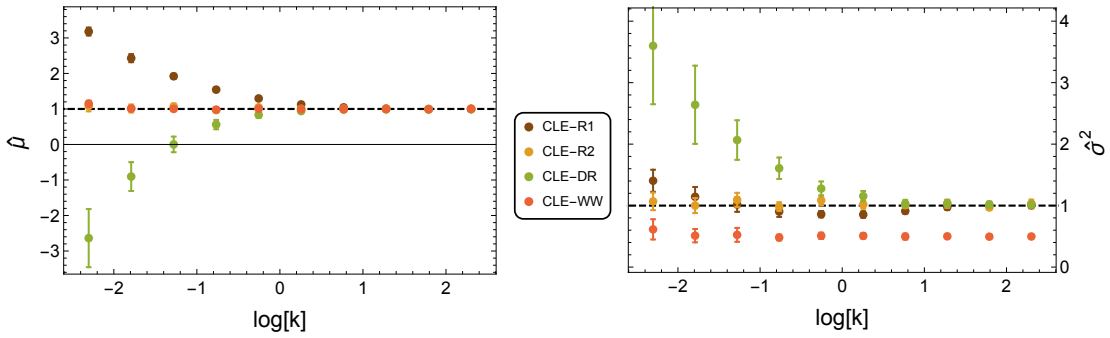

Figure 3.1: The normalised mean $\hat{\mu}$ and normalised variance $\hat{\sigma}^2$ as a function of the non-dimensional parameter k for the various CLEs of the simple production-decay reaction system given by scheme (3.5). CLE-R1 denotes the CLE in Eq. (3.7) with reflective boundary condition, CLE-R2 denotes the CLE in Eq. (3.8), CLE-WW is the modified CLE approach in [24] and CLE-DR is the modified CLE approach in [25]. The normalisation involves dividing the means and variances obtained from the simulations by the exact analytic results: $\mu = \sigma^2 = k = \Omega c_1/c_2$. Only the CLE-R2 agrees with the analytic result (black dashed line) for all k. The simulation parameters are $\delta\tau = 10^{-5}, \Delta\tau = 1, N = 10^3$ (see main text for discussion of these parameters and for the method used to calculate the moments from the CLEs).

achieved ($\Delta\tau$) and the number of samples (N). Moments are calculated from a single time trajectory by averaging over the fluctuating variables at time points $\Delta\tau$, $2\Delta\tau$, ..., $N\Delta\tau$; this procedure is repeated ten times leading to ten independent estimates for the moments - the average over the estimates and the standard deviation about these averages are what is plotted in the figures. This simulation protocol is followed throughout the rest of this Chapter.

Figure 3.1 shows the results for the mean number of X molecules and the variance of fluctuations about this mean in steady-state conditions normalised by the analytic results (mean concentration = variances of fluctuations = k for a birth-death process simulated using the CME or the CLE) as a function of k. Both methods give the correct result for large values of k. This is because a large k value corresponds to a large input-to-decay ratio and thus a large mean value which implies a small probability of the number of molecules becoming negative. With decreasing k, the discrepancy between the two CLEs becomes evident: the CLE-R1 gives the wrong moments, whereas the CLE-R2 agrees with the analytic result. This is clearly due to the fact that the CLE-R1

imposes an artificial boundary to avoid the breakdown of the CLE whereas the CLE-R2 naturally does not suffer from any breakdown. Our results imply that *different versions of the CLE with different noise terms are not necessarily equivalent in terms of their boundary behaviour*.

Figure 3.1 shows as well the results of the modified CLE methods of Wilkie and Wong (CLE-WW) [24] and of Dana and Raha (CLE-DR) [25] applied to the reaction scheme (3.5). The latter method becomes accurate in the macroscopic limit (the limit of large k) while the former method (CLE-WW) is accurate only in the mean concentration but gives an incorrect variance of fluctuations for all values of k. The CLE-WW does not converge to the correct results in the macroscopic limit because it postulates a global change to the diffusion terms of the CLE (the deletion of some of these terms) to fix the breakdown problem which is localised to the boundary of zero molecule numbers. On the other hand, the CLE-DR modifies the diffusion coefficients only in breakdown regions of the parameter space setting them to zero, and hence it necessarily becomes accurate in the macroscopic limit. Because of these reasons, in the rest of this chapter we shall compare our results only with those of the CLE-DR.

Hence it is clear that even for the simple example considered here, the methods which artificially correct for the breakdown of the CLE (CLE-R1 with reflection boundary conditions, CLE-DR and CLE-WW) lead to an inequivalence between the CLE's predictions for the first two moments and those of the CME. Equivalence can be restored, in this case, by choosing an alternative CLE representation (CLE-R2) which naturally does not break down at any point in time. We note that the alternative CLE representation is consistent with a drift-diffusion process which can access real values of x larger than $-k$; the probabilistic interpretation of the CFPE is also consistent with such a process since the diffusion scalar $B = \Omega c_1 + c_2 x$ of the CFPE is positive for $x > -k$. Hence one can state that for this example it is possible to find a well-defined CLE representation in real space because the drift-diffusion process describing the chemical reaction lives on the real domain. We next consider a linear multi-species system to check whether one can always find a representation of the CLE in real space which does not suffer breakdown.

Example (ii): Production followed by isomerisation

We consider the following system of linear reactions involving two distinct species

$$\varnothing \underset{c_4}{\overset{c_1}{\rightleftharpoons}} X_1 \underset{c_3}{\overset{c_2}{\rightleftharpoons}} X_2. \tag{3.9}$$

We rescale time as $\tau = c_4 t$ and define $k_1 = \Omega c_1/c_4, k_2 = c_2/c_4, k_3 = c_3/c_4$. The standard form of the CLE for the reaction system (3.9) (denoted as CLE-R1) reads

$$dx_1 = (k_1 - k_2 x_1 + k_3 x_2 - x_1)d\tau + \sqrt{k_1}dW_1 - \sqrt{k_2 x_1}dW_2 + \sqrt{k_3 x_2}dW_3 - \sqrt{x_1}dW_4, \quad (3.10)$$

$$dx_2 = (k_2 x_1 - k_3 x_2)d\tau + \sqrt{k_2 x_1}dW_2 - \sqrt{k_3 x_2}dW_3. \quad (3.11)$$

When one of the variables becomes zero, some noise terms become zero and some remain finite and thus the noise can drive the system to negative values of the variables which then leads to breakdown. A possible alternative form is given by the Langevin equation (denoted as CLE-R2)

$$dx_1 = (-x_1 + k_1 - k_2 x_1 + k_3 x_2)d\tau + \sqrt{y_1}dW_1 + \sqrt{y_2}dW_2, \quad (3.12)$$

$$dx_2 = (k_2 x_1 - k_3 x_2)d\tau - \sqrt{y_2}\, dW_2, \quad (3.13)$$

where we have defined

$$y_1 = x_1 + k_1, \quad (3.14)$$

$$y_2 = k_2 x_1 + k_3 x_2. \quad (3.15)$$

The CLE-R2 breaks down if y_1 or y_2 become negative. To check whether this can occur, we transform the CLE-R2 to the new variables y_1 and y_2. For this purpose, we express x_1 and x_2 in terms of y_1 and y_2 as

$$x_1 = y_1 - k_1, \quad (3.16)$$

$$x_2 = \frac{1}{k_3}(y_2 + k_2(k_1 - y_1)). \quad (3.17)$$

Using Ito's formula [49], we find that the CLE-R2 in the new variables reads

$$dy_1 = (2k_1 - y_1(2k_2 + 1) + 2k_1 k_2 + y_2)d\tau + \sqrt{y_1}dW_1 + \sqrt{y_2}dW_2, \quad (3.18)$$

$$dy_2 = (2k_1 k_2(1 + k_2 - k_3) + k_2 y_1(2k_3 - 2k_2 - 1) + y_2(k_2 - k_3))d\tau + k_2\sqrt{y_1}dW_1 + (k_2 - k_3)\sqrt{y_2}dW_2. \quad (3.19)$$

Consider the case $y_2 = 0, y_1 > 0$. The CLE-R2 then reads

$$dy_2 = (2k_1 k_2(1 + k_2 - k_3) + k_2 y_1(2k_3 - 2k_2 - 1))d\tau + k_2\sqrt{y_1}dW_1. \quad (3.20)$$

Clearly the diffusion term can drive the system to negative values of y_2 and hence to breakdown. Interestingly, breakdown can also occur because the drift becomes negative for positive y_1. For example for $2k_3 - 2k_2 - 1 \neq 0$ and $(1 + k_2 - k_3)/(2k_3 - 2k_2 - 1) < 0$,

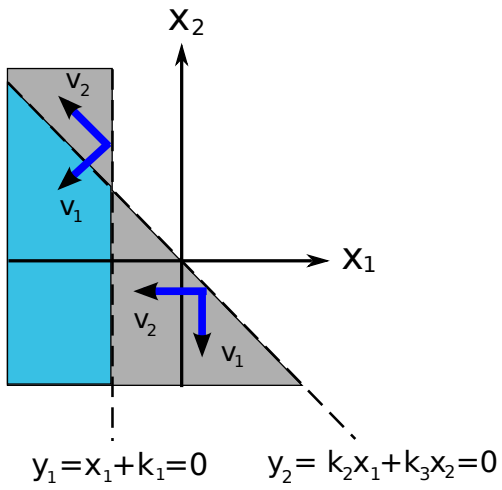

Figure 3.2: Graphical representation of the state space of the two-dimensional reaction system in (3.9). The dashed lines indicate the boundaries where either y_1 or y_2 become zero. The grey area corresponds to the part of the state space where the eigenvalues λ_1 and λ_2 of the CFPE diffusion matrix are negative and positive, respectively. The blue shaded area represents the region of space where both eigenvalues are negative. Thus the diffusion matrix is not positive semi-definite in the grey and blue shaded areas. The blue arrows represent the eigenvectors for the case $y_1 = 0, y_2 > 0$ and the case $y_2 = 0, y_1 > 0$. Since the eigenvector of the non-vanishing eigenvalue is not parallel to the boundary, there is a non-vanishing noise component orthogonal to the boundary that can drive the system to breakdown.

the drift becomes negative for $y_1 < 2k_1(k_3 - k_2 - 1)/(2k_3 - 2k_2 - 1)$, which is possible under the constraint $y_1 > 0$. Similarly it is easy to show that for the case $y_1 = 0, y_2 > 0$ the diffusion term can drive the system to breakdown (breakdown due to the drift term is here not possible because the drift is always positive).

To gain insight into the underlying reason for breakdown, we next consider the diffusion matrix of the CFPE. Using Eq. (3.2) we find that the diffusion matrix is given by

$$\mathbf{B} = \begin{pmatrix} k_1 + x_1 + k_2 x_1 + k_3 x_2 & -k_2 x_1 - k_3 x_2 \\ -k_2 x_1 - k_3 x_2 & k_2 x_1 + k_3 x_2 \end{pmatrix} = \begin{pmatrix} y_1 + y_2 & -y_2 \\ -y_2 & y_2 \end{pmatrix}. \tag{3.21}$$

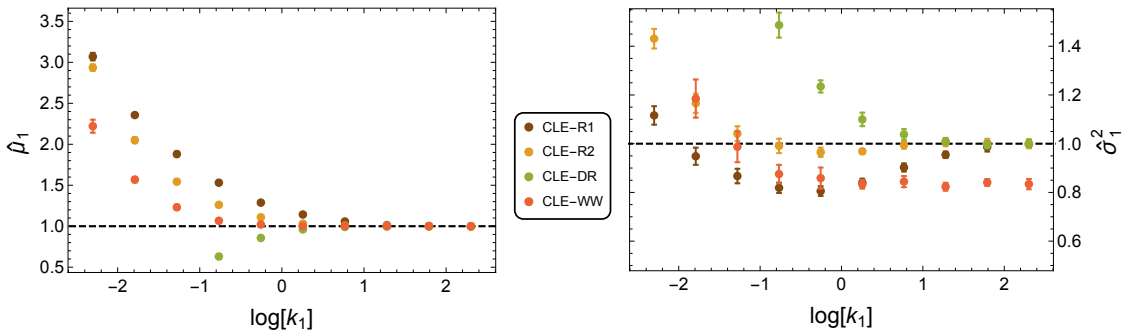

Figure 3.3: Normalised mean and variance of the implementations CLE-R1, CLE-R2, CLE-DR and CLE-WW for species X_1 of the linear reaction system (3.9) as a function of $k_1 = \Omega c_1/c_4$. All other parameters are set to unity. Reflection boundary conditions at zero molecule numbers and at $y_1 = y_2 = 0$ are respectively imposed on CLE-R1 and CLE-R2 to avoid their breakdown. The values are normalised by the exact analytic expression for the moments obtained from the CME: $\mu_1 = \sigma_1^2 = k_1$. A large k_1 thus corresponds to a large mean value. The dashed line represents the exact value. Since the system is linear, the CLE should reproduce the exact result. The large deviations for small values of k_1 thus clearly indicate that the imposed reflection boundaries distort the moments. Simulation details for all implementations are $\delta\tau = 10^{-4}, \Delta\tau = 1, N = 10^4$.

Its eigenvalues and corresponding eigenvectors are given by

$$\lambda_1 = \frac{1}{2}\left(y_1 + 2y_2 - \sqrt{y_1^2 + 4y_2^2}\right), \quad \lambda_2 = \frac{1}{2}\left(y_1 + 2y_2 + \sqrt{y_1^2 + 4y_2^2}\right), \tag{3.22}$$

$$v_1 = \left(-\frac{-y_1 + \sqrt{y_1^2 + 4y_2^2}}{2y_2}, -1\right)^T, \quad v_2 = \left(-\frac{y_1 + \sqrt{y_1^2 + 4y_2^2}}{2y_2}, 1\right)^T. \tag{3.23}$$

Inspection of these equations shows that the eigenvalue λ_1 becomes negative if y_1 or y_2 become negative, i.e., the positive semi-definite form of the diffusion matrix, which is a necessary requirement of any Fokker-Planck equation, cannot be maintained. Hence it follows that breakdown of the CLE-R2 is due to the fact that the drift-diffusion process has a finite probability of accessing a region of space (y_1 or y_2 negative) where the diffusion matrix of the CFPE is not positive semi-definite, i.e., there is no probabilistic interpretation of a drift-diffusion process in the real domain which describes the reaction system in Eq. (3.9). Therefore, *the breakdown of the CLE-R2 is independent of the particular choice of* $\mathbf{C}$ *and thus the same problem is manifest for all possible choices of* $\mathbf{C}$*, for all possible Langevin equation representations of the CFPE.*

This intrinsic breakdown of the CFPE can also be understood as follows. Figure 3.2

shows the eigenvectors corresponding to $y_1 = 0, y_2 > 0$ and $y_2 = 0, y_1 > 0$. The eigenvector corresponding to λ_2 in the case $y_1 = 0, y_2 > 0$ is $v_2 = (-1, 1)^T$. This is clearly not parallel to the boundary $y_1 = x_1 + k_1 = 0$ which is parallel to $(0, 1)^T$. There is thus always a non-vanishing noise component orthogonal to the boundary which implies that the noise can drive the system across the boundary leading to breakdown of the CFPE. A similar conclusion follows for the case $y_2 = 0, y_1 > 0$.

We note that the connection between the form of the diffusion matrix and the breakdown properties of the CLE is not specific to this example. It can be generally proved for all chemical systems that if the diffusion matrix **B** is not positive semi-definite then the matrix **C** cannot be real, i.e., the CLE necessarily breaks down due to square roots of negative arguments. A proof of this result can be found in Appendix A.2.

Correcting the breakdown by imposing artificial reflective boundaries introduces significant errors. The results of such simulations - the mean and variance of species X_1 for CLE-R1 and CLE-R2 - are shown in Figure 3.3. The results are normalised with the exact analytic results obtained by solving the CME for the reaction system in Eq. (3.9) (this leads to mean = variance = k_1). Both CLEs show significant deviations from the exact result for small values of k_1, i.e, for small values of the average number of molecules of X_1. As for the previous example of a simple birth-death process, it is found that these significant deviations from the exact CME result cannot be eliminated using CLEs with modified propensities, i.e., using the methods of Wilkie and Wong [24] or of Dana and Raha [25].

3.1.2 Bimolecular reaction systems

In the previous section we saw for a linear system with one species that there exists a representation of the CLE which avoids breakdown and which recovers the equivalence of the CLE and CME results for the mean and variance of the process for linear systems. In contrast, for a simple linear two-species system we found that this is not the case, i.e., that the CLE breaks down for all representations and the equivalence of the CLE and CME for linear systems is violated. With the intuitive eigenvector picture in mind (as illustrated in Figure 3.2), *we expect the CLE to break down for most multi-dimensional systems*, since there is no reason why the eigenvectors of the diffusion matrix should in general be parallel to the boundary separating the regions in state space where the diffusion matrix is positive semi-definite and where it is not.

As an example, a detailed breakdown analysis of a three variable CLE describing a

reaction which is catalysed by two enzymes can be found in Appendix A.1. We show there that all representations of the CLE do break down. For the two systems that we will study in Section 3.3, one of which is the example system of Chapter 2 introduced in Figure 2.2, one can easily perform a similar analysis leading to the same conclusion. The intrinsic reason for the breakdown is found to be as for linear systems - namely that the diffusion matrix of the associated CFPE loses its positive semi-definite form for points in real space which can be reached by the drift-diffusion process described by the CLE/CFPE.

3.2 The complex chemical Langevin equation

In the previous section we have shown that the commonly employed CLE generally suffers from breakdown at finite times due to the occurrence of negative arguments in square roots. This problem can be alleviated by imposing reflecting boundary conditions or by a variety of other modified versions of the CLE. However, as we have shown, these procedures introduce inaccuracies in the CLE predictions. Foremost amongst such is the inequivalence between the modified CLE predictions and those of the CME for the mean and variance of the process for linear systems.

The state space of the CLE is generally taken to be the real domain since molecule numbers are real, and this has been an assumption in our derivations of the previous section as well. However as we show in this section, the breakdown can be avoided by working directly with a complex extension of the CLE; we show here that this restores the equivalence of the CLE and CME predictions for the mean and variance for linear reactions and gives accurate results for bimolecular systems. Although the molecule numbers are generally complex, we show that the *complex CLE* predicts real-valued moments and hence admits a physical interpretation. The only properties used for this derivation are the analyticity and behaviour under complex continuation of the drift and diffusion terms in the CFPE - hence the generality of our approach. We also show that the expectation of any analytic function in $\mathbb{R}^N$ and the autocorrelation functions and power spectra predicted by the complex CLE are real-valued functions. Finally we discuss the method by which the complex CLE can be used to simulate first passage times.

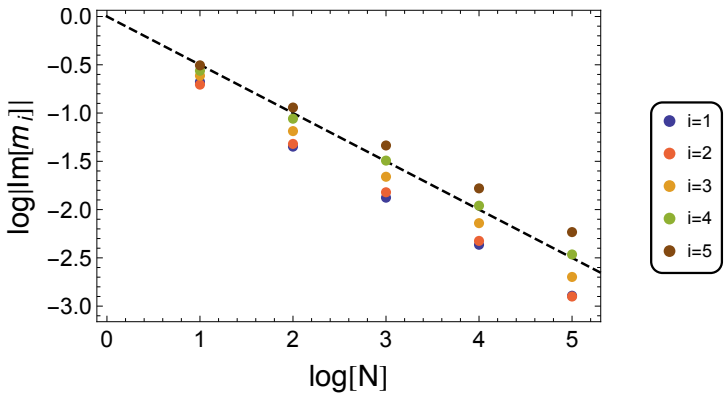

Figure 3.4: Absolute value of the imaginary parts of the first five moments of species X_1 in reaction (3.9) as a function of the number of samples N in CLE-C simulations. The i^{th} moment is given by m_i. Each value is normalised by the absolute value of the corresponding moment. The points of each moment can be approximately fitted by a straight line (black dashed curve) with a slope of $-1/2$. This implies that the normalised imaginary parts decay as $\sim 1/\sqrt{\mathrm{N}}$ and thus converge to zero in the limit of an infinite number of samples. The simulation parameters are $k_1 = 0.3, k_2 = k_3 = 1$, $\delta\tau = 10^{-3}, \Delta\tau = 1$.

3.2.1 Definition and properties

We start by defining the complex CLE:

Definition 3.1. *The complex CLE is defined by the very same equation as the real-valued CLE in Eq.* (3.3)*, but on a complex-valued state space.*

This means in particular that for real and positive initial conditions the complex CLE reduces to the real-valued CLE for certain reaction systems, such as the one-species linear system studied in Section 3.1.1. As we have argued in Section 3.1.2, however, we expect the complex CLE to actually become complex-valued for most multi-species systems, in which case it does not agree with the real-valued CLE.

The complex CLE is well-defined and does not suffer form any breakdown by construction. However, the question arises if the complex CLE allows a probabilistic interpretation since complex-valued molecule numbers are physically not meaningful. As it turns out it does indeed admit a probabilistic interpretation since we can show the following result:

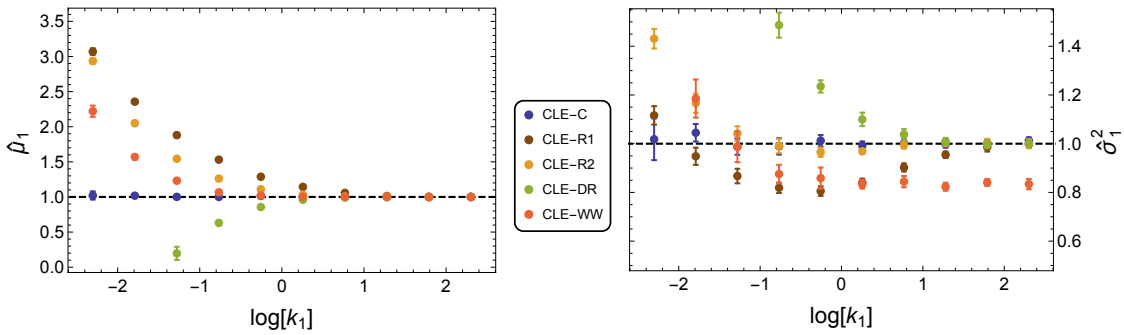

Figure 3.5: Testing the accuracy of the complex CLE (CLE-C). This is the same plot as in Figure 3.3, but with the corresponding results of the CLE-C included. The CLE-C gives the correct mean and variance of species X_1, i.e, agrees with the CME, within sampling error. The CLE-C's superior accuracy over that of the real-valued CLEs stems from the fact that the CLE-C does not suffer from breakdown and that hence it does not require the imposition of artificial boundaries (as necessary for the real-valued CLE-R1 and CLE-R2). The used simulation parameters are $\delta\tau = 10^{-3}, \Delta\tau = 1, N = 10^4$.

Theorem 3.1. *For real initial conditions, the complex CLE defined in Definition 3.1 has real-valued moments for all times. More generally, consider a real-valued function $g(\mathbf{x})$ analytic in $\mathbb{R}^N$ that can be globally represented as a power series. When continued to complex variables, the expectation of $g(\mathbf{z})$ with respect to the distribution predicted by the complex CLE is real-valued for all times.*

We give the proof in the next section. So despite single realisations of the complex CLE being complex-valued, the moments become real-valued in the ensemble average. This is verified numerically in Figure 3.4 for the two-species linear system defined in Eq. (3.9). The Figure shows the imaginary part of the first five moments of species X_1 as a function of the number of simulated samples used to compute the moments. We find that the imaginary part converges to zero in the limit of an infinite number of samples.

Theorem 3.1 implies

Corollary 3.1. *The complex CLE fulfils the property of exact mean and variance of fluctuations for linear reaction systems.*

This follows directly from the agreement of the moment equations of up to order two as derived from the CLE and the CME as shown in Section 2.3.5, together with the fact that the complex CLE is well-defined without any imposed boundaries and that its moments are real-valued. This result is verified in Figure 3.5, which is the same as

Figure 3.3 but with the results of the complex CLE included (CLE-C). We find that, in contrast to the various real-valued versions of the CLE, the complex CLE's prediction for the mean and variance agree with the corresponding CME results.

Other quantities that are frequently of interest are autocorrelation functions and power spectra. For these we can prove:

Theorem 3.2. *Autocorrelation functions and power spectra predicted by the complex CLE are real-valued for all times.*

To summarise, the complex CLE is well-defined for all reaction networks, and gives rise to real-valued moments, real-valued autocorrelation functions and real-valued power spectra.

3.2.2 Proof of main results

Before proving the theorems of the previous section, let us write the complex CLE in terms of real-valued expressions. For the purpose of presentation, we use the CLE in the standard form in the following. We note, however, that the complex CLE and the following proofs are independent of the chosen representation. Writing $\mathbf{z} = (z_1, \ldots, z_N)^T$, where z_i is the complex variable corresponding to the particle number of species X_i, the complex CLE reads

$$d\mathbf{z} = \mathbf{A}(\mathbf{z})dt + \mathbf{C}(\mathbf{z})d\mathbf{W}. \tag{3.24}$$

By writing $z_j = x_j + iy_j$ this equation can be split up into coupled Langevin equations for the real parts x_j and imaginary parts y_j. By relabelling the variables as $(w_1, \ldots, w_{2N})^T = (x_1, \ldots, x_N, y_1, \ldots, y_N)^T$ we can write the equations in the form

$$d\mathbf{w} = \mathcal{A}dt + \mathcal{C}d\mathcal{W}, \quad \mathcal{C}\mathcal{C}^{\mathbf{T}} = \mathcal{B}. \tag{3.25}$$

Here, we have defined

$$\mathcal{A} = (A_1^x, \ldots, A_N^x, A_1^y, \ldots, A_N^y)^T, \quad \mathcal{C} = \begin{pmatrix} C^x \\ C^y \end{pmatrix}, \quad d\mathcal{W} = (dW_1, \ldots, dW_R)^T. \tag{3.26}$$

$$\mathcal{B} = \mathcal{C}\mathcal{C}^T = \begin{pmatrix} C^x(C^x)^T & C^x(C^y)^T \\ C^y(C^x)^T & C^y(C^y)^T \end{pmatrix} = \begin{pmatrix} \mathcal{B}^{xx} & \mathcal{B}^{xy} \\ \mathcal{B}^{yx} & \mathcal{B}^{yy} \end{pmatrix}. \tag{3.27}$$

The superscripts x and y denote the real and imaginary part of a function $f(\mathbf{x}, \mathbf{y}) = f^x(\mathbf{x}, \mathbf{y}) + if^y(\mathbf{x}, \mathbf{y})$.

The new diffusion matrix $\mathcal{B}$ is symmetric; this follows from the fact that $\mathcal{B}^{xx}$ and $\mathcal{B}^{yy}$ are symmetric while $\mathcal{B}^{xy} = (\mathcal{B}^{yx})^T$. Moreover, we find that $\mathcal{B}$ is positive semi-definite since $\mathcal{C}$ is real in $\mathbb{R}^{2N}$ (see Appendix A.2). The FPE corresponding to the complex CLE reads

$$\partial_t P(\mathbf{w},t) = -\sum_{i=1}^{2N} \partial_i \left[\mathcal{A}_i(\mathbf{w},t)P(\mathbf{w},t)\right] + \frac{1}{2}\sum_{i,j=1}^{2N} \partial_i\partial_j \left[\mathcal{B}_{ij}(\mathbf{w},t)P(\mathbf{w},t)\right]. \tag{3.28}$$

For proving the theorems of the previous section we need the following lemma describing an important property of the probability distribution of the process:

Lemma 3.1. *The solution $P(\mathbf{x},\mathbf{y},t) = P(\mathbf{w},t)$ of the FPE in Eq.* (3.28) *is invariant under the joint reflection of the imaginary variables, $\mathbf{y} \to -\mathbf{y}$, i.e.,*

$$P(\mathbf{x},-\mathbf{y},t) = P(\mathbf{x},\mathbf{y},t). \tag{3.29}$$

Proof. We first rewrite the FPE in Eq. (3.28) in the form

$$\begin{aligned}\partial_t P(\mathbf{x},\mathbf{y},t) = [&-\sum_{i=1}^{N}(\partial_{x_i}A_i^x(\mathbf{x},\mathbf{y}) + \partial_{y_i}A_i^y(\mathbf{x},\mathbf{y})) \\ &+ \frac{1}{2}\sum_{i,j=1}^{N}(\partial_{x_i}\partial_{x_j}\mathcal{B}_{ij}^{xx}(\mathbf{x},\mathbf{y}) + \partial_{y_i}\partial_{y_j}\mathcal{B}_{ij}^{yy}(\mathbf{x},\mathbf{y}) + 2\partial_{x_i}\partial_{y_j}\mathcal{B}_{ij}^{xy}(\mathbf{x},\mathbf{y}))]P(\mathbf{x},\mathbf{y},t),\end{aligned} \tag{3.30}$$

where we have used $\partial_{x_i}\partial_{y_j}\mathcal{B}_{ij}^{xy} = \partial_{y_i}\partial_{x_j}\mathcal{B}_{ij}^{yx}$, which can easily be verified from the definition of $\mathcal{B}$ in Eq. (3.27). Since A_i is a polynomial with real coefficients, it fulfils $A_i(\bar{\mathbf{z}}) = \overline{A_i(\mathbf{z})}$, or written in terms of the $\mathbf{x}$ and $\mathbf{y}$ variables $A_i(\mathbf{x},-\mathbf{y}) = \overline{A_i(\mathbf{x},\mathbf{y})}$. For its real and imaginary parts this implies

$$\begin{aligned}A_i^x(\mathbf{x},-\mathbf{y}) &= A_i^x(\mathbf{x},\mathbf{y}), \\ A_i^y(\mathbf{x},-\mathbf{y}) &= -A_i^y(\mathbf{x},\mathbf{y}).\end{aligned} \tag{3.31}$$

Using that f_r are polynomials in the molecule numbers (c.f. Eq. (2.25)) and the symmetry properties of the complex square root, i.e. $\sqrt{\bar{\mathbf{z}}} = \overline{\sqrt{\mathbf{z}}}$, we find $\sqrt{f_r(\bar{\mathbf{z}})} = \sqrt{\overline{f_r(\mathbf{z})}} = \overline{\sqrt{f_r(\mathbf{z})}}$. Using $\mathbf{C}$ in the standard form, $C_{ir} = \sum_r S_{ir} f_r^{1/2}$, we find that $C_{ir}(\bar{z}) = \overline{C_{ir}(\mathbf{z})}$. The real and imaginary parts thus obey $C_{ij}^x(\mathbf{x},-\mathbf{y}) = C_{ij}^x(\mathbf{x},\mathbf{y})$ and $C_{ij}^y(\mathbf{x},-\mathbf{y}) = -C_{ij}^y(\mathbf{x},\mathbf{y})$, respectively. Using these properties and the definition of $\mathcal{B}_{ij}$ given in Eq. (3.27), it is straightforward to verify that

$$\begin{aligned}\mathcal{B}_{ij}^{xx}(\mathbf{x},-\mathbf{y}) &= \mathcal{B}_{ij}^{xx}(\mathbf{x},\mathbf{y}), \\ \mathcal{B}_{ij}^{yy}(\mathbf{x},-\mathbf{y}) &= \mathcal{B}_{ij}^{yy}(\mathbf{x},\mathbf{y}), \\ \mathcal{B}_{ij}^{yx}(\mathbf{x},-\mathbf{y}) &= -\mathcal{B}_{ij}^{yx}(\mathbf{x},-\mathbf{y}).\end{aligned} \tag{3.32}$$

Using Eqs. (3.31) and (3.32), we find that the FPE in Eq. (3.30) is invariant under the joint reflection of the imaginary variables, $\mathbf{y} \to -\mathbf{y}$:

$$\begin{aligned}\partial_t P(\mathbf{x},-\mathbf{y},t) = [&-\sum_{i=1}^{N}(\partial_{x_i}A_i^x(\mathbf{x},-\mathbf{y}) + \partial_{-y_i}A_i^y(\mathbf{x},-\mathbf{y})) \\ &+\frac{1}{2}\sum_{i,j=1}^{N}(\partial_{x_i}\partial_{x_j}\mathcal{B}_{ij}^{xx}(\mathbf{x},-\mathbf{y}) + \partial_{-y_i}\partial_{-y_j}\mathcal{B}_{ij}^{yy}(\mathbf{x},-\mathbf{y}) + 2\partial_{x_i}\partial_{-y_j}\mathcal{B}_{ij}^{xy}(\mathbf{x},-\mathbf{y}))]P(\mathbf{x},-\mathbf{y},t)\end{aligned} \tag{3.33}$$

$$\begin{aligned}= [&-\sum_{i=1}^{N}(\partial_{x_i}A_i^x(\mathbf{x},\mathbf{y}) + \partial_{y_i}A_i^y(\mathbf{x},\mathbf{y})) \\ &+\frac{1}{2}\sum_{i,j=1}^{N}(\partial_{x_i}\partial_{x_j}\mathcal{B}_{ij}^{xx}(\mathbf{x},\mathbf{y}) + \partial_{y_i}\partial_{y_j}\mathcal{B}_{ij}^{yy}(\mathbf{x},\mathbf{y}) + 2\partial_{x_i}\partial_{y_j}\mathcal{B}_{ij}^{xy}(\mathbf{x},\mathbf{y}))]P(\mathbf{x},-\mathbf{y},t).\end{aligned} \tag{3.34}$$

Since the initial condition is always given by a symmetric probability distribution (the imaginary parts are necessarily zero since the initial specification is in terms of real-valued molecule numbers), it follows that the probability distribution solution of the FPE in Eq. (3.30) remains symmetric for all times, $P(\mathbf{x},\mathbf{y},t) = P(\mathbf{x},-\mathbf{y},t)$, which concludes the proof. □

Lemma 3.1 allows us to prove Theorem 3.1:

Proof of Theorem 3.1. Consider a general moment $\langle z_1^{m_1} z_2^{m_2} \dots z_N^{m_N}\rangle, m_1, \dots m_N \in \mathbb{N}$, of the complex variables $z_i = x_i + iy_i$:

$$\begin{aligned}\langle z_1^{m_1} \dots z_N^{m_N}\rangle &= \int dz_1 \dots dz_N \; z_1^{m_1} \dots z_N^{m_N} P(z,t) \\ &= \int dx_1 \dots dx_N dy_1 \dots dy_N \; (x_1+iy_1)^{m_1} \dots (x_N+iy_N)^{m_N} P(x,y,t).\end{aligned} \tag{3.35}$$

Each summand of the imaginary part of the product $(x_1+iy_1)^{m_1} \dots (x_N+iy_N)^{m_N}$ can be written in the form $x_1^{m_1-k_1} \dots x_N^{m_N-k_N} y_1^{k_1} \dots y_N^{k_N}$, with $k_i \in \mathbb{N}, k_i \le m_i$ for $i = 1, \dots N$, with odd $\sum_{i=1}^N k_i$, i.e., the exponents of the y_i sum to an odd integer. The term $x_1^{m_1-k_1} \dots x_N^{m_N-k_N} y_1^{k_1} \dots y_N^{k_N}$ thus changes its sign under $\mathbf{y} \to -\mathbf{y}$, and since the probability distribution is symmetric in $\mathbf{y}$ according to Lemma 3.1, it follows that the imaginary part of the integral in Eq. (3.35) vanishes. This means that moments of the complex variables z_i are real-valued which proves the first statement of Theorem 3.1.

Next, suppose we are interested in the moments of a general real-valued function $g(\mathbf{x})$. Suppose g is analytic in $\mathbb{R}^N$ and that it can be globally represented as a power series. This implies that it can be analytically continued to $\mathbb{C}^N$. Since $g(\mathbf{x})$ is real-valued

the coefficients of its power series about a real point are real, too. This means that g fulfils $g(\bar{\mathbf{z}}) = \overline{g(\mathbf{z})}$, and since the probability is symmetric under $\mathbf{z} \to \bar{\mathbf{z}}$, the expectation of the imaginary part of $g(\mathbf{z})$ vanishes. This means that the expectation of $g(\mathbf{z})$ is real-valued, which concludes the proof. □

Similarly, we can now prove Theorem 3.2

Proof of Theorem 3.2. The autocorrelation matrix for a homogeneous process can be computed by [49]

$$G(\tau) = \langle \mathbf{z}(\tau)\mathbf{z}^T(0)\rangle = \lim_{T\to\infty} \frac{1}{T}\int_0^T dt\ \mathbf{z}(t+\tau)\mathbf{z}^T(t). \tag{3.36}$$

In terms of probability densities, it can be written as

$$G(\tau) = \int d\mathbf{z}_\tau d\mathbf{z}_0\ \mathbf{z}_\tau \mathbf{z}_0^T P(\mathbf{z}_\tau, \tau; \mathbf{z}_0, 0), \tag{3.37}$$

where we defined $\mathbf{z}_t = \mathbf{z}(t)$. According to Lemma 3.1, the solution of the FPE corresponding to the complex CLE is symmetric under the reflection of the imaginary variables, $\mathbf{y} \to -\mathbf{y}$, under appropriate initial conditions. It follows that transition probabilities and hence joint probability distributions have this property, too. Writing $\mathbf{z}_t = \mathbf{x}_t + i\mathbf{y}_t$, we have

$$\mathrm{Im}[G_{ij}(\tau)] = \int d\mathbf{x}_\tau d\mathbf{y}_\tau d\mathbf{x}_0 d\mathbf{y}_0\ ((\mathbf{x}_\tau)_i(\mathbf{y}_0)_j + (\mathbf{y}_\tau)_i(\mathbf{x}_0)_j) P(\mathbf{x}_\tau, \mathbf{y}_\tau, \tau, \mathbf{x}_0, \mathbf{y}_0, 0). \tag{3.38}$$

The integrand is an odd function under the joint reflection $\mathbf{y}_\tau \to -\mathbf{y}_\tau, \mathbf{y}_0 \to -\mathbf{y}_0$, which means $\mathrm{Im}[G_{ij}(\tau)] = 0$, i.e., the correlation matrix G is real. For a homogeneous process it further fulfils $G(-\tau) = G(\tau)$ by construction. This means that the power spectrum, which is simply the Fourier transform of the autocorrelation matrix [49], is a real function given by

$$S(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\tau e^{-i\omega\tau} G(\tau). \tag{3.39}$$

We thus have shown that the autocorrelation function and power spectrum are real-valued functions.

□

3.2.3 First passage times

Another physical quantity that is often of interest is the first passage time, i.e., the mean time the state vector $\mathbf{x} = (x_1, \ldots, x_N)$ takes to reach a particular value. For example, one

may want to know the time it takes a certain number of protein molecules of some species to be produced. Say the molecule number of this species is labeled x_i; then the first passage time can be computed from the complex CLE by calculating the average time it takes for the real part of x_i to achieve a certain value, i.e., we simply leave the imaginary parts unbounded.

3.3 Examples

Next, we demonstrate the accuracy of the complex CLE for two systems of biochemical importance. In both cases we find the complex CLE's accuracy to be much higher than the accuracy of the conventional real-valued CLE as well as the accuracy of the linear noise approximation (LNA) and the normal moment closure approximation of second order (2MA) introduced in Sections 2.3.6 and 2.3.4, respectively.

A genetic negative feedback loop

First, we consider the gene system introduced in Figure 2.2 with reactions in Eq. (2.23). The CME for this system has recently been solved exactly in steady state [53]. The CLE in standard form is given in Eqs. (2.66) and (2.67). We implement the CLE in three different ways: the naive implementation enforcing reflective boundary conditions, i.e., $0 < x_1 < 1$ and $x_2 > 0$, such that the terms under the square roots in the standard form of the CLE in Eqs. (2.66) and (2.67) remain positive (CLE-R), the complex version of the CLE (CLE-C) and the modified CLE of Dana and Raha (CLE-DR) [25]. The simulations utilise the parameter set $k_1 = 10, k_3 = 0.5, k_4 = 1$. Figure 3.6 shows the normalised mean and the normalised variance of the protein and gene as a function of the dimensionless parameter $\sigma_b = k_2/(\Omega k_4)$. This can be viewed as varying the bimolecular reaction constant k_2 for fixed volume Ω or equivalently as varying the volume of the system for fixed k_2. We observe that the predictions of the the CLE-R and CLE-DR are considerably more inaccurate than those of the other methods (CLE-C, LNA and 2MA). In particular, they are significantly more inaccurate than the CLE-C, showing that enforcing artificial boundary conditions and modifying noise terms can have large effects on the method's accuracy. The difference in the accuracy of the CLE-C, LNA and 2MA is much less pronounced. The CLE-C is slightly more accurate than the 2MA, which in turn is slightly more accurate than the LNA (see insets of Figure 3.6).

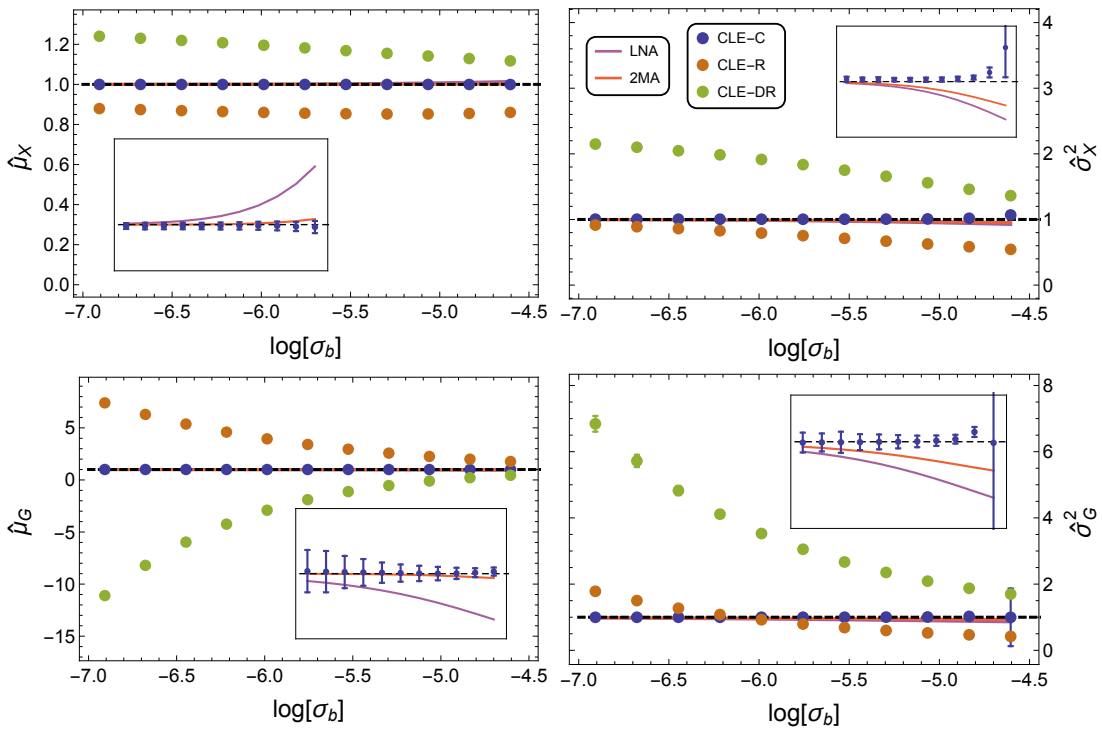

Figure 3.6: Normalised steady-state mean numbers of protein $\hat{\mu}_P$ and bound gene $\hat{\mu}_G$, and their corresponding variances $\hat{\sigma}^2_P$ and $\hat{\sigma}^2_G$, as a function of the non-dimensional parameter σ_b (a measure of binding affinity of the protein to the gene), for the genetic negative feedback loop in Figure 2.2. The values are normalised by the exact values obtained from the CME [53]. We find that the CLE-R (CLE in standard form with artificial reflective boundaries to avoid breakdown) and CLE-DR (a modified CLE proposed in [25]) give significantly worse results than the CLE-C. The accuracy of the latter and of the conventional LNA and the 2MA approximations are comparable. The simulation parameters are as follows. For the CLE-R and CLE-DR: $\delta\tau = 10^{-4}, \Delta\tau = 10, N = 10^4$. For the CLE-C: $\delta\tau = 10^{-4}, \Delta\tau = 10, N = 10^5$.

The Michaelis-Menten reaction with substrate input

Next, we consider the Michaelis-Menten reaction with substrate input

$$\varnothing \xrightarrow{c_4} S, \quad S + E \underset{c_2}{\overset{c_1}{\rightleftharpoons}} ES \xrightarrow{c_3} E + X, \tag{3.40}$$

where E is the free enzyme, ES is the enzyme-substrate complex, S is the substrate and X is the product. The number of enzyme molecules is fixed to one. The system has a steady state in the substrate concentration if $\alpha \equiv c_4\Omega/c_3 < 1$, which simply means that the input rate must be slower than the maximum turnover rate. We derived an exact steady-state solution of the CME for this reaction which has not been reported in the

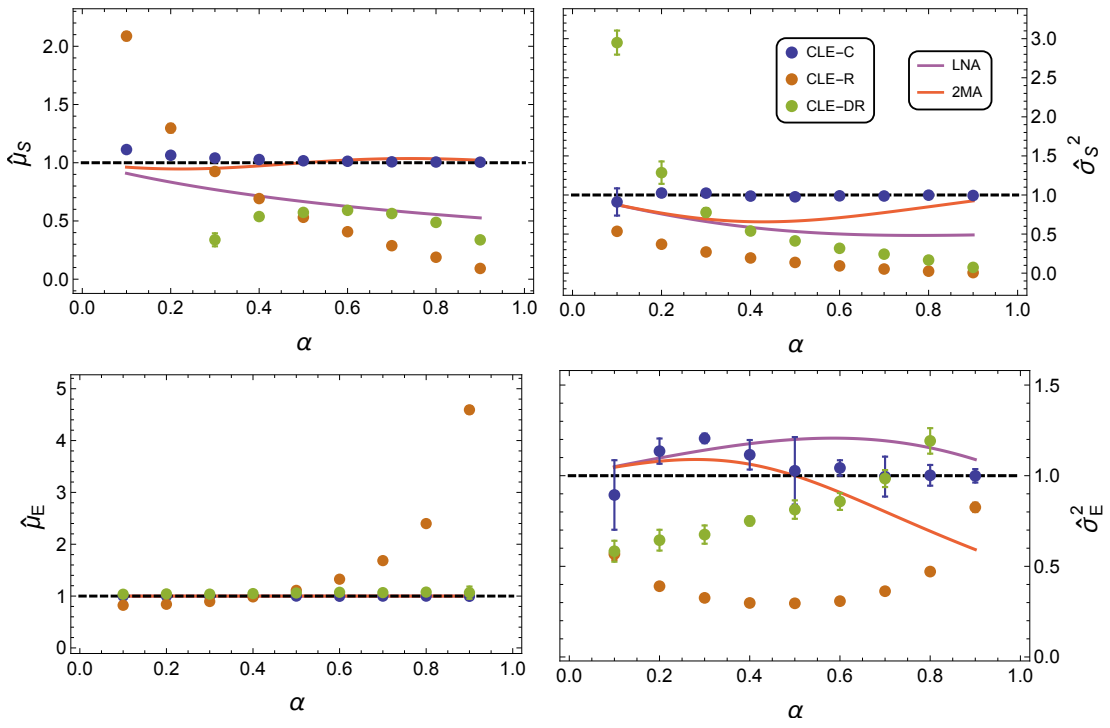

Figure 3.7: Normalised steady-state mean number of molecules of substrate $\hat{\mu}_S$ and enzyme $\hat{\mu}_E$, and their corresponding variances $\hat{\sigma}_S^2$ and $\hat{\sigma}_E^2$, as a function of the non-dimensional parameter α (a measure of saturation), for the enzyme reaction system in Eq. (3.40). The values are normalised by the exact values obtained from the CME which are derived in Appendix A.3. We find that the CLE-R (CLE in standard form with artificial reflective boundaries to avoid breakdown) and CLE-DR (a modified CLE proposed in [25]) give significantly worse results than the CLE-C. The latter is also significantly more accurate than both the conventional LNA and the 2MA approximations. The simulation parameters are as follows. For the CLE-C: $\delta\tau = 10^{-4}, N = 10^5$. $\Delta\tau$ scales like α^4 from $5-45$ for $\alpha = 0.1-0.9$; for the CLE-R and CLE-DR: $\delta\tau = 10^{-4}$, $\Delta\tau = 10$ and $N = 10^4$.

literature to our knowledge. We present the full derivation in Appendix A.3, where we obtain expressions for $P_0(n,\tau)$ and $P_1(n,\tau)$ - the probability of having n substrate molecules at time τ given 0 and 1 free enzyme molecules, respectively. These are given by

$$P_0(n) = \frac{C''}{kn!}\left(\frac{k_4}{k_3}\right)^{n+1}\frac{\Gamma(n+k+1)}{\Gamma(k)}\,{}_1F_1\left[-n;-(k+n);k_3\right], \tag{3.41}$$

$$P_1(n) = \frac{C''}{n!}\left(\frac{k_4}{k_3}\right)^{n}\frac{\Gamma(k+n)}{\Gamma(k)}\,{}_1F_1\left[-n;-(k+n-1);k_3\right], \tag{3.42}$$

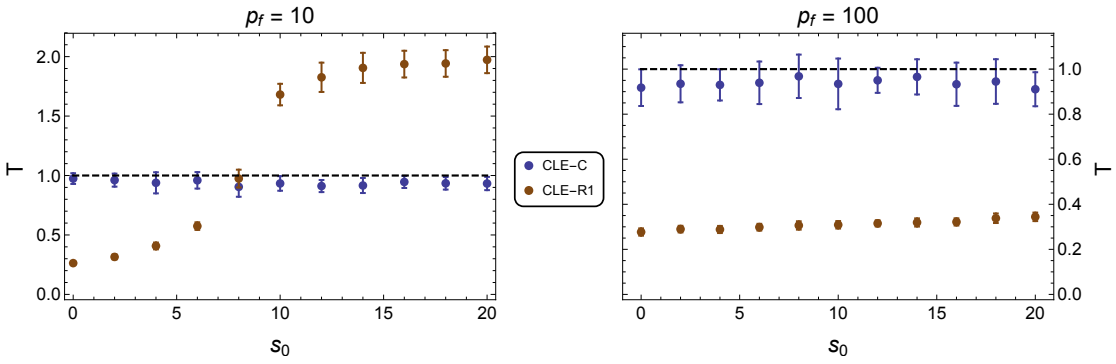

Figure 3.8: Normalised mean first passage time T for a number p_f of product molecules to be produced, as a function of the initial substrate concentration s_0, for the enzyme reaction system (3.40). The values are normalised by the exact values corresponding to the CME obtained by stochastic simulations using the SSA. We find that the CLE-R (CLE in standard form with artificial reflective boundaries to avoid breakdown) gives generally worse results than the CLE-C. For the simulation time step we used $\delta\tau = 10^{-3}$ for the CLE-C and CLE-R. The number of samples drawn were 10^3 and 10^2 for $p_f = 10$ and 100 respectively.

where $c = c_1/\Omega$, $k_2 = c_2/c, k_3 = c_3/c, k_4 = \Omega c_4/c$, $k = k_2 + k_4$, $k_{34} = k_3/k_4$, $C'' = e^{-k_4}(k_{34} - 1)^{k+1}/k_{34}^{k+1}$, Ω is the compartment volume, Γ is the gamma function and ${}_1F_1$ is the confluent hypergeometric function. All moments of the the substrate and enzyme molecule numbers can thus be computed exactly without the need for stochastic simulations.

The CLE in standard form (and the time rescaled by c_1/Ω) is given by

$$dx_1 = (\Omega c_4 + c_2(1-x_2) - \frac{c_1}{\Omega}x_1x_2)dt + \sqrt{\Omega c_4}dW_1 + \sqrt{c_2(1-x_2)}dW_2 - \sqrt{\frac{c_1}{\Omega}x_1x_2}dW_3,$$
$$dx_2 = ((c_2+c_3)(1-x_2) - \frac{c_1}{\Omega}x_1x_2)dt + \sqrt{c_2(1-x_2)}dW_2 - \sqrt{\frac{c_1}{\Omega}x_1x_2}dW_3 + \sqrt{c_3(1-x_2)}dW_4, \tag{3.43}$$

where x_1 is the number of substrate molecules and x_2 is the number of free enzyme molecules. Here, we have used the conservation law between the free enzyme and complex molecules, which implies that the number of complex molecules can be written as $1 - x_2$.

For simulations, we choose physiologically realistic values for the rate constants [81, 82]: $c_1 = 2 \times 10^6 (Ms)^{-1}$, $c_2 = 1s^{-1}$, and $c_3 = 1s^{-1}$. We choose the volume to be $\Omega = 10^6 M^{-1}$ which corresponds to a spherical submicron compartment of roughly 150 nm diameter. The dimensionless parameter $\alpha \equiv c_4\Omega/c_3$ is varied over the whole interval

$[0,1]$ admitting a steady state through modification of the value of the input rate c_4. As for the gene system, we implement the CLE in three different ways: the naive implementation enforcing reflective boundary conditions (CLE-R), i.e., $0 < x_1$ and $0 < x_2 < 1$, the complex version of the CLE (CLE-C) and the modified CLE of Dana and Raha (CLE-DR) [25], and compare these with the corresponding LNA and 2MA results.

Figure 3.7 shows the mean and variance of both enzyme and substrate species obtained from the different methods normalised by the corresponding CME value (as determined from the exact solution - see Appendix A.3) as a function of α. First of all, we observe that the CLE-C's predictions are significantly more accurate than those of the CLE-R and CLE-DR, again showing how enforcing positivity or modifying noise terms can give rise to significant inaccuracies. Moreover, we find that the CLE-C gives significantly more accurate results than the LNA and 2MA as well (only for the variance for the latter).

Next, we consider the following first passage time problem. We want to compute the mean time it takes to produce a certain number of product molecules as a function of the initial substrate numbers. We explained in Section 3.2.1 how the complex CLE can be used to simulate first passage times. Figure 3.8 shows the mean first passage time T for the catalytic reaction to produce $p_f = 10$ and 100 product molecules. The values are normalised by those obtained from stochastic simulations using the SSA. The rate constants c_1, c_2 and c_3 and the volume Ω are chosen as before and in addition we set $c_4 = 10^5 Ms^{-1}$. We observe that the CLE-R gives much larger deviations from the SSA result than the CLE-C, which once more shows that enforcing positivity can lead to inaccurate results.

3.4 Conclusion

Although the CLE is a popular and convenient approximation method of the chemical master equation, the problem of its boundary behaviour has been rarely studied in the literature, despite well-known problems arising for small molecule numbers. Here, we have shown that the CLE is in general mathematically ill-defined if defined in real space due to the occurrence of square roots of negative expressions in finite time with finite probability for most reactions systems, Moreover, we have shown that this breakdown can generally not be avoided by different choices of the noise matrix in the CLE. We have found that this breakdown is due to the stochastic process reaching regions in state space for which the diffusion matrix of the CFPE is not positive semi-definite. We

have also shown that, due to their different breakdown behaviours, different choices of the noise matrix are generally not equivalent as often claimed in the literature.

The naive way to avoid breakdown is to enforce artificial boundary conditions enforcing positive variables. We have found that such an implementation, as well as other alternative formulations of the CLE proposed in the literature that modify noise terms, lead to highly inaccurate predictions for the mean and variance of the process as compared to the exact CME result. In particular, they violate the fundamental property of agreement of CLE and CME predictions for the mean and variance for linear reaction systems. Next, we have shown that the problem of breakdown can be solved by extending the state space of the CLE to complex-valued variables. This new CLE is always well-defined and does not suffer from breakdown. Its physical interpretation stems from the fact that it predicts real-valued moments, as well as real-valued autocorrelation functions, power spectra and first-passage times. We have also shown that the complex CLE restores the agreement of the CLE and CME predictions for the mean and variance for linear systems.

As we have seen, simulation using the real-valued CLE requires the use of methods to artificially correct for its breakdown near the zero molecule number boundary, and hence the apparent inaccuracy of the CLE comes from the use of these methods as well as from its intrinsic assumption of continuous molecule numbers. When the CLE is considered in complex space, we found it to be remarkably accurate even for chemical systems with species in very low molecule numbers, such as the two bimolecular examples studied in Section 3.3, where the numbers of gene and enzyme were just one. This suggests that inaccuracies of the CLE may often not be due to its assumption of continuous molecule numbers but rather due to the methods used to avoid breakdown.

The complex CLE involves the simulation of twice the number of coupled stochastic differential equations as the conventional real-valued CLE, and hence it is typically found that more samples are needed to obtain accurate estimates of the moments. Moreover, in some cases, such as the gene and enzyme examples in Section 3.3, we found that to guarantee numerical stability it was necessary to take a smaller time step size for the complex CLE than for the conventional CLE. Probably this restriction can be lifted or eased by use of more sophisticated stochastic differential equation simulation methods than the simple Euler-Maruyama method used here (see [80] for a broad discussion of available methods). Another disadvantage of the complex CLE is that, since its variables are complex-valued, it is not possible to directly obtain the probability distribution over the real variables. The latter would be the object needed if

one wanted to approximate the solution of the CME. However, whenever the distribution predicted by the complex CLE is narrow in imaginary directions in comparison to real directions, a good estimate of the distribution over real-valued variables can be obtained by projecting the distribution over the complex-valued variables onto the real axes.

We note that even using the Euler-Maruyama implementation, the complex CLE is computationally advantageous compared to the stochastic simulation algorithm whenever one is simulating systems characterised by many reactions per unit time and relatively few species. The complex CLE was found to achieve a high accuracy over a broad range of molecule numbers suggesting that it could be a novel useful tool for modelling stochasticity in chemical reaction networks.

Chapter 4

Validity of moment closure approximations

In the previous chapter we found that the chemical Langevin equation may become mathematically ill-defined due to the absence of natural boundary conditions for the diffusion approximation, a phenomenon we termed *breakdown*. We have shown that ad hoc methods for fixing the breakdown can lead to inaccurate predictions for mean concentrations and variance of fluctuations as compared to corresponding chemical master equation (CME) results. In the present chapter, we perform a somewhat similar analysis of the moment closure approximations (MAs) introduced in Section 2.3.4. We try to answer the question if MAs are mathematically always well-defined and if they always give physically meaningful predictions that admit a probabilistic interpretation.

To this end, after revising the definition of moment closure approximations in Section 4.1, we clarify what we mean by "physically meaningful predictions", and we formulate a set of validity conditions for MAs that ensure feasibility of these predictions in Section 4.2. Next, we study the normal MA in detail for a simple one-species system. We find that the MA indeed fails to give physically meaningful results depending on the parameters, and a detailed steady-state and stability analysis is provided to elucidate this breakdown behaviour. In Section 4.4 we then compare the normal, Poisson, log-normal and central-moment-neglect (CMN) MAs for three more complex reaction systems, with respect to their breakdown behaviour and their quantitative accuracy. Next, we introduce the software package *MOCA* in Section 4.5 which allows the automated derivation and analysis of the studied MA methods. Finally, we conclude in Section 4.6.

4.1 Moment closure approximations

Recall that by multiplying the CME in Eq. (2.24) by $n_i \dots n_l$ and summing over all $n_i, \dots, n_l$ one obtains an ODE for the moment $\langle n_i \dots n_l \rangle$, where $\langle \cdot \rangle$ denotes the expectation with respect to the solution of the CME. Up to order two these read

$$\partial_t \langle n_i \rangle = \sum_{r=1}^{R} S_{ir} \langle f_r(\mathbf{n}) \rangle, \tag{4.1}$$

$$\partial_t \langle n_i n_j \rangle = \sum_{r=1}^{R} \left[S_{jr} \langle n_i f_r(\mathbf{n}) \rangle + S_{ir} \langle f_r(\mathbf{n}) n_j \rangle + S_{ir} S_{jr} \langle f_r(\mathbf{n}) \rangle \right]. \tag{4.2}$$

For a bimolecular reaction the propensity function $f_r(\mathbf{n})$ is a polynomial of order two in the n_i, which means that the ODEs for the second moments in Eq. (4.2) depend on third order moments. This in turn means that the set of ODEs in Eqs. (4.1) and (4.2) is not closed (and similarly for reactions of order three or higher). The same is true for higher order moments, i.e., their time evolution equations depend on higher order moments. This means that we have an infinite system of coupled ODEs which hence cannot be solved.

The MA methods we study here express moments above a certain order in terms of lower order moments which allows to close the equations to that order. This is often done by assuming the solution of the CME to have a particular functional form. We study here the *normal MA*, which sets all cumulants above a certain order to zero corresponding to a normal distribution, the *Poisson MA*, which sets all diagonal cumulants to the mean and all non-diagonal cumulants to zero, the *log-normal MA* which assumes a multivariate log-normal distribution to close the equations, and the *central-moment-neglect MA* (CMN-MA) which sets all central moments above a certain order to zero (c.f. Eqs. (2.39)-(2.43)).

These MAs are non-systematic, ad hoc approximations, in the sense that they are not defined as a perturbative expansion in a small parameter or similarly. It is thus not clear a priori, and has not been studied so far to our knowledge, if MAs always give physically meaningful results admitting a probabilistic interpretation. We next clarify what we mean by "physically meaningful".

4.2 Validity conditions

Let us formulate the requirements that we pose on MA methods to be valid in a more systematic way. To this end we define a set of criteria which guarantee physically

meaningful predictions of MA approximations and which we will repeatedly use in the rest of this chapter. We only consider systems for which the CME has a globally attractive steady-state solution.

Note that negative mean values of particle numbers are physically meaningless and central moments of probability distributions are always non-negative. Therefore, the first obvious requirement that we pose for MAs to be valid, is that the *mean values, the variance and all higher-order central moments should be non-negative for all times*, provided they are chosen accordingly initially.

Moments of a distribution are defined as expectations of polynomials in the variables with respect to the distribution. Now if the CME has a globally attractive steady-state solution, this means that the moments in steady state are uniquely defined. Note that this is also true if the steady-state solution of the CME is multimodal or if the deterministic rate equations are multistable. We thus require a valid MA to *have exactly one positive and globally attractive fixed point*. This means that time trajectories should converge to the same positive fixed-point for all initial conditions. By "positive fixed point" we mean a fixed point with non-negative mean value and non-negative central moments.

Consider a system for which the deterministic rate equations show sustained oscillations. In this case, single-time trajectories of the SSA may show sustained oscillations. However, since independent trajectories get out of phase as time progresses, the moments of the ensemble, which are the moments of the solution of the CME, always show either damped or over-damped oscillations. We hence require valid MAs to *not exhibit sustained oscillations*. Note that this requirement is actually already contained in the previous one since the existence of a globally attractive fixed point excludes sustained oscillations. We treat the case of sustained oscillations explicitly anyway for clarification.

In summary, we require valid MAs to fulfil the following conditions, provided the CME has a globally attractive steady-state solution:

- Mean values and central moments are non-negative for all times.
- The MAs have a globally attractive positive fixed point.
- The MAs do not exhibit sustained oscillations.

Note that the first condition is obviously a sensible condition for all systems, also if the CME does not have a globally attractive fixed point.

4.3 Validity analysis of normal MA for a simple example

In this section, we analyse the second-order normal MA with respect to the validity conditions described in the previous section for a simple one-species reaction system. We will find that the method does not always fulfil the validity conditions. We give a detailed analysis of the method's validity for this simple example here before giving a quantitative comparison of the different MA methods mentioned above for several more complex systems in the following section.

Consider a model of bursty gene expression followed by a post-translational protein dimerisation reaction:

$$\varnothing \xrightarrow{c_1 q_m} mP, \quad m \in \mathbb{N}\backslash\{0\}, \quad P + P \xrightarrow{c_2} \varnothing, \tag{4.3}$$

where P is the protein species, c_2 is the dimerisation rate constant and m is the burst size. Experimental [83] and theoretical [84, 85] evidence indicates a geometric distribution $q_m = p(1-p)^{m-1}$ with constant parameter p ($p < 1$) as an appropriate model for bursting; $c_1 q_m$ are then the rates at which bursts of size m are created. The production step can be viewed as either an infinite number of input reactions, or equivalently as a single input reaction with input size m being a random variable. Note that $1/p$ is equal to the mean burst size. Note also that in the limit of $p \to 1$, the expression is non-bursty and the set of protein production reactions in (4.3) reduces to the single reaction $\varnothing \xrightarrow{c_1} P$.

We rescale time as $\tau = tc_2/\Omega$ and define the dimensionless constant $k = \Omega^2 c_1/c_2$, where Ω is the volume of the system. It is easy to show that the rate equations for this system have a unique positive fixed point which is globally attractive for all k; exact stochastic simulations using the stochastic simulation algorithm also show that the CME has a stationary solution for all values of k.

Using Eqs. (4.1) and (4.2), we can derive the time-evolution equations for the first moment $\langle n \rangle$ and the second moment $\langle n^2 \rangle$ which read

$$\partial_\tau \langle n \rangle = -2\langle n^2 \rangle + 2\langle n \rangle + k\langle m \rangle, \tag{4.4}$$

$$\partial_\tau \langle n^2 \rangle = -4\langle n^3 \rangle + 8\langle n^2 \rangle - 4\langle n \rangle + 2\langle n \rangle k\langle m \rangle + k\langle m^2 \rangle, \tag{4.5}$$

where $\langle m \rangle = 1/p$ and $\langle m^2 \rangle = (2-p)/p^2$ (these follow from the definition of q_m). These equations are then closed by using the normal MA (setting the third cumulant of n to

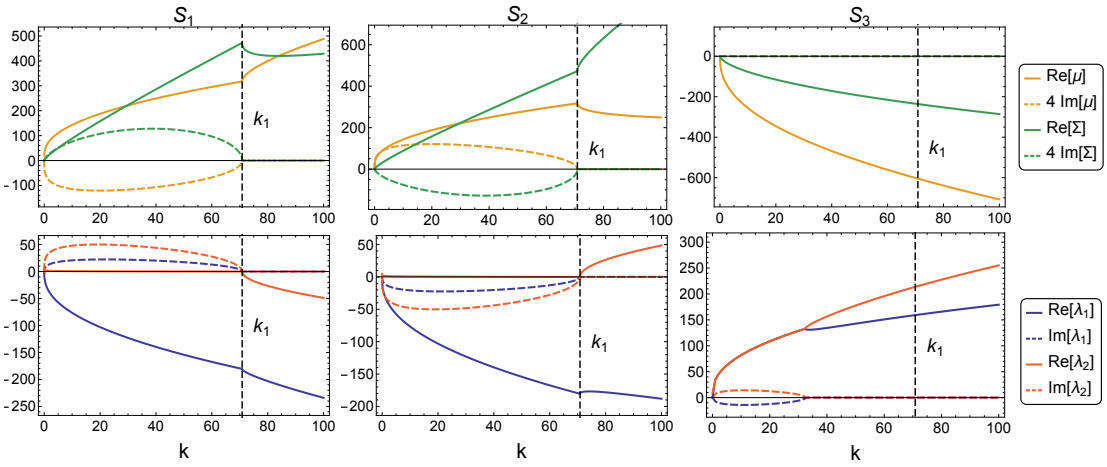

Figure 4.1: Stability of the second-order normal MA for the bursty gene expression system in Eq. (4.3). The three plots in the upper panel show the real and imaginary parts of the mean protein number μ and the variance Σ in protein number fluctuations of the three fixed points S_i for the normal MA in Eqs. (4.6) and (4.7) as a function of k for a mean burst size of 20 (i.e., $p = 1/20$). The three plots in the lower panel show the corresponding real and imaginary parts of the eigenvalues of the Jacobian, λ_1 and λ_2, respectively. We find that S_2 and S_3 are always unstable, while S_1 is stable and the mean and variance are positive, if and only if $k > k_1$.

zero), leading to:

$$\partial_\tau \mu = \langle m \rangle k + 2\mu - 2\mu^2 - 2\Sigma, \tag{4.6}$$

$$\partial_\tau \Sigma = \langle m^2 \rangle k + 4\mu(\mu - 1) + 8(1-\mu)\Sigma, \tag{4.7}$$

where $\mu = \langle n \rangle$ and $\Sigma = \langle n^2 \rangle - \langle n \rangle^2$ are the mean and variance in protein numbers.

Setting the left hand side of Eqs. (4.6) and (4.7) to zero, and solving simultaneously, one finds that there are three possible solutions which we call $S_i = (\mu_i, \Sigma_i), i = 1, \ldots, 3$. Figure 4.1 shows the real and imaginary parts of the mean and variance of these three fixed points, as well as the real and imaginary parts of the corresponding eigenvalues of the Jacobian of Eqs. (4.6) and (4.7) for the case $p = 1/20$ (this corresponds to a mean burst size of 20 which has been measured experimentally for gene expression [83]). By inspection of Figure 4.1, we see that of the three possible steady-state solutions only S_1 is physically admissible; this is since it is the only steady-state solution which displays a positive mean and variance of molecule numbers and which is locally stable (negative real part of the eigenvalues of the Jacobian). However note that these properties only manifest for k larger than a certain critical value $k = k_1 \approx 70$. This would lead one

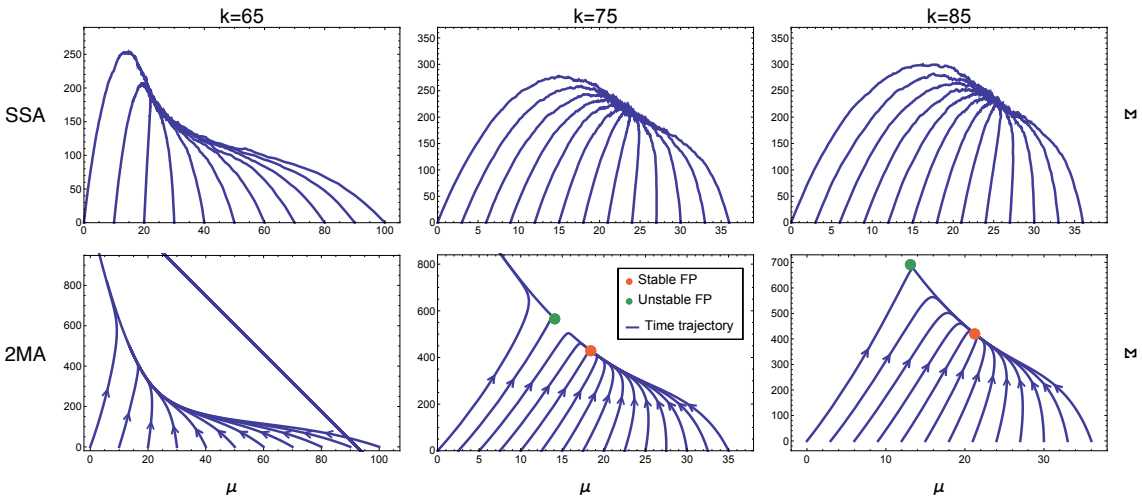

Figure 4.2: Time trajectories of the second-order normal MA (2MA) and of the ensemble-averaged SSA in the μ - Σ plane for different values of k for the bursty gene expression system in Eq. (4.3). The figure shows time trajectories for deterministic initial conditions. The red and green dots show the stable and unstable fixed points (FPs) of the normal MA equations, respectively. While for $k = 65$ there is no fixed point in the positive orthant and the time trajectories diverge for all initial conditions, for $k = 75$ most initial conditions lead to trajectories converging to a unique stable fixed point. However, a small initial mean value still leads to divergence. Finally, for $k = 85$ all initial conditions lead to trajectories converging to a unique fixed point with positive mean and variance. In contrast, the CME has a stationary solution for all values of k. The mean burst size is 20 ($p = 1/20$).

to surmise that the CME has a stationary solution only for k greater than this critical value. However, as noted earlier this is not the case: the CME has a stationary solution for all values of k. These results taken together imply that *the normal MA does not give a physically meaningful steady-state solution for all values of k*.

Next, we study the time evolution of the moments. Figure 4.2 shows the numerically integrated time trajectories for several initial conditions for three different k values and for $p = 1/20$. We find that for $k < k_2 \approx 1.2k_1 \approx 85$, *some of the trajectories diverge as time goes to infinity* which is unphysical since the CME has a stable fixed point. This instability manifests for all deterministic initial conditions for $k < k_1$ and for initial conditions characterised by a small initial mean number of protein molecules for $k_1 \leq k < k_2$. For $k \geq k_2$, however, the trajectories converge to the stable fixed point and are non-negative at all times. We verified this numerically for initial conditions up

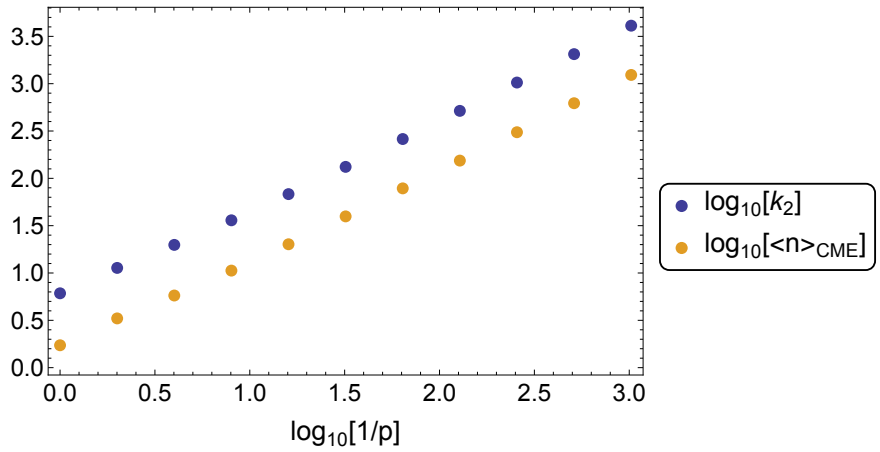

Figure 4.3: Critical values $(k_2, \langle n \rangle_{\text{CME}})$ of the normal MA as a function of mean burst size $1/p$ for the reaction system in Eq. (4.3). Both curves are monotonically increasing with increasing $1/p$ implying that the critical molecule number above which the normal MA equations lead to physically meaningful results increases with the burstiness in protein expression.

to $\mu(\tau = 0) = 10^6$. Since $k = \Omega^2 c_1/c_2$, increasing k is equivalent to increasing the system volume Ω while keeping the rate constants c_1 and c_2 constant. *Hence in coincidence with the steady-state analysis above, we find that the normal MA only gives physically admissible solutions for system volumes above a certain critical threshold.* We shall refer to the requirement of physically meaningful convergent time trajectories as the "time-dependent criterion" and the requirement of a positive stable fixed point as the "steady-state criterion". The fulfilment of the steady-state criterion is obviously a necessary condition for the time-dependent criterion. However, since we have $k_1 < k_2$, we find that it is not a sufficient condition, i.e., even despite the existence of a positive stable fixed point, there may exist time trajectories that diverge or converge to unphysical values.

The analysis described above was specifically for the case of $p = 1/20$. Qualitatively similar results are found for all values of p, i.e, there exist critical p-dependent values k_1 and k_2 with $k_2 \geq k_1$, such that for $k > k_1$ the steady-state criterion is satisfied and for $k > k_2$ the time-dependent criterion is satisfied. Figure 4.3 shows the p-dependence of the critical value k_2 as well as the p-dependence of the corresponding mean particle number $\langle n \rangle_{\text{CME}}$. Note that both k_2 and $\langle n \rangle_{\text{CME}}$ increase with $1/p$ implying that the larger the burstiness in protein expression, the larger the critical molecule number above which the normal MA gives physically meaningful results. For $p = 1/20$,

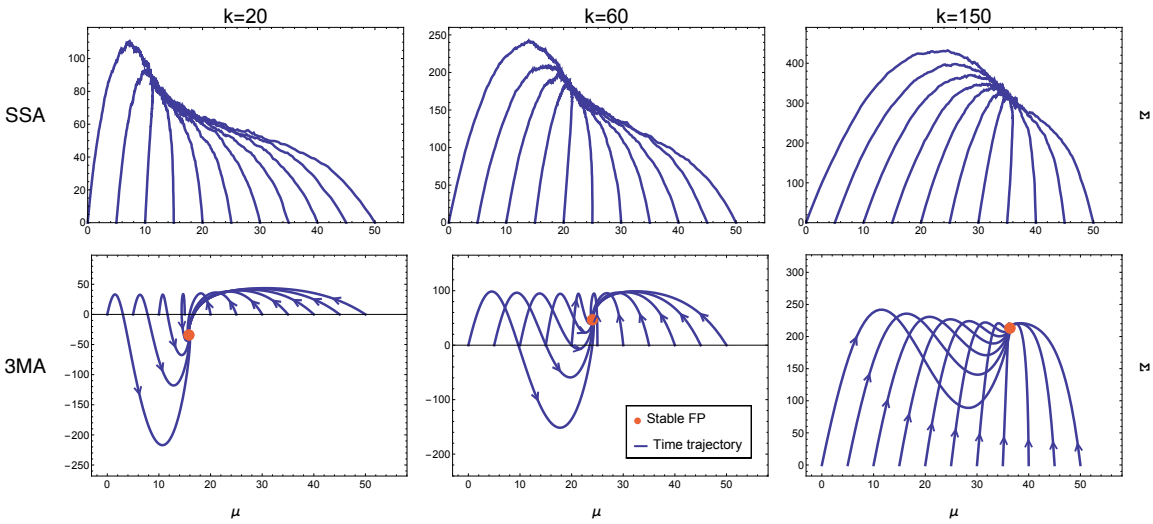

Figure 4.4: Time trajectories of the third-order normal MA (3MA) and the ensemble-averaged SSA in the μ - Σ plane for different values of k for the bursty gene expression system in Eq. (4.3). For $k = 20$ there is an unphysical stable fixed point of the MA equations with negative variance. For $k = 60$ the stable fixed point has positive mean and variance but the variance of some trajectories becomes negative as the fixed point is approached. Finally, for $k = 150$ all trajectories converge to a physically admissible steady state and are physically meaningful (positive mean and variance at all times). In contrast, the CME has a stationary solution for all values of k. The mean burst size is 20 ($p = 1/20$).

we had found before that $k_2 = 85$ which corresponds to a mean steady-state protein number of $\langle n \rangle_{CME} \approx 25$, i.e., *the normal MA equations for a mean protein burst size of 20 give physically meaningful results for the time evolution of the system only when the number of protein molecules in steady state exceeds 25*. It is well known that protein numbers in cells can be very small, even of the order of a few molecules and hence our results show that one must be careful when using moment closure approximations to understand cell level phenomena.

In Figure 4.2 we have shown that the normal MA can give rise to diverging time trajectories despite the existence of a positive stable fixed point. To illustrate the possibility of qualitatively different unphysical behaviour we performed the same analysis for the third order normal MA (3MA). Figure 4.4 shows the 3MA analog of the second-order normal MA time evolution analysis shown in Figure 4.2. The similarity between the two figures is evident: the time evolution criteria are only satisfied for k greater than a certain critical value (≈ 100). However, we find a quite different unphysical

behaviour here: rather than divergent trajectories as for the second-order normal MA, we find here that the trajectories either converge to a fixed point with negative variance or run through negative values for some finite time intervals before converging to a positive fixed point.

In summary, our analysis of this section shows that the normal MA for a gene circuit involving a bimolecular reaction gives physically meaningful results only for systems above a certain critical volume. We did not present the corresponding results for the Poisson, log-normal and CMN-MAs, but only want to report here that these methods behave qualitatively exactly the same.

4.4 Analysis and comparison of different moment closure methods

In the previous section we found that the second-order normal MA gives physically meaningful results only above a certain critical system size for a simple one-species bimolecular reaction system that is deterministically monostable. In the present section, we analyse the second-order normal MA, as well as the second-order Poisson, log-normal and CMN-MAs for systems with different deterministic behaviours: we first analyse a system whose deterministic rate equations are bistable. Next, we investigate how well the MA methods can capture the influence of noise in a protein-phosphorylation system whose deterministic system shows ultrasensitivity. And finally, we use the MAs to study the role of stochasticity in a system whose deterministic system is oscillating and which becomes entrained by an external force. The different MA methods were defined in Eqs. (2.39)-(2.43).

4.4.1 A deterministic bistable system

Consider the minimal elementary reaction system whose rate equations show bistability [86]

$$\varnothing \xrightarrow{k_0} X, \quad Y \xrightarrow{k_1} 2X, \tag{4.8}$$

$$2X \xrightarrow{k_2} X+Y, \quad X+Y \xrightarrow{k_3} Y, \tag{4.9}$$

$$X \xrightarrow{k_4} \varnothing. \tag{4.10}$$

We added the first reaction to the ones given in [86] to prevent the stochastic system from having an absorbing state for zero molecule numbers. Depending on the pa-

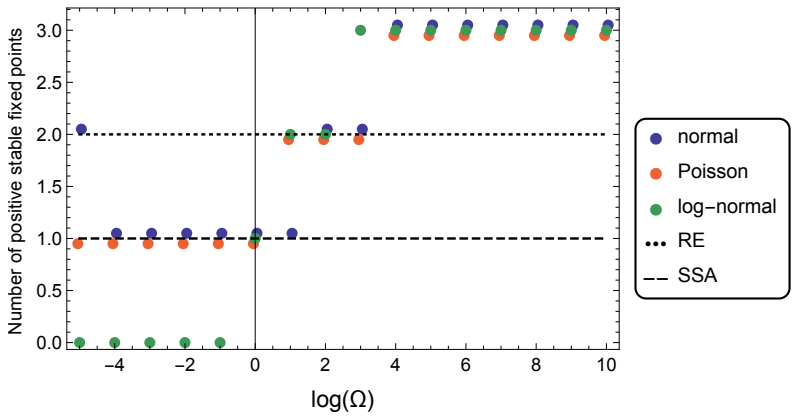

Figure 4.5: Number of positive stable fixed points as a function of the volume Ω on log-scale obtained from steady-state analysis for the bistable reaction system in Eqs. (4.8)-(4.10) for the parameters $k_0 = 1, k_1 = 1, k_2 = 5, k_3 = 0.2$ and $k_4 = 5$. We shift the points slightly to make coinciding points distinguishable. We find that all three MAs give a physical result of a single positive stable fixed point only on an intermediate range of volumes. The latter is significantly smaller for the log-normal MA than for the normal and Poisson MAs.

rameter values the deterministic rate equations become bistable for this system. All parameter sets used in this section are chosen such that this is the case. Since the reactions in Eqs. (4.8)-(4.10) are of order two or lower, their propensity functions are polynomials of up to order two in the species variables. This means that the time evolution equations of the second-order moments depend on the third-order moments, but not on higher-order moments. We thus have to express the third-order moments in terms of first and second-order moments to close the equations to second order. Recall that the second-order normal and CMN-MAs set all cumulants and central moments above order two to zero, respectively (c.f. Eqs. (2.39) and (2.43)). Since the third-order cumulant and third-order central moment are identical, the second-order normal MA and CMN-MA are equivalent for the reaction system in Eqs. (4.8)-(4.10). This is of course a general result, i.e., for chemical reaction systems with elementary reactions and mass-action kinetics (i.e., reactions up to order two and polynomial propensity functions), the second-order normal MA and second-order CMN-MA are identical.

We thus analyse the normal, Poisson and log-normal MAs here.

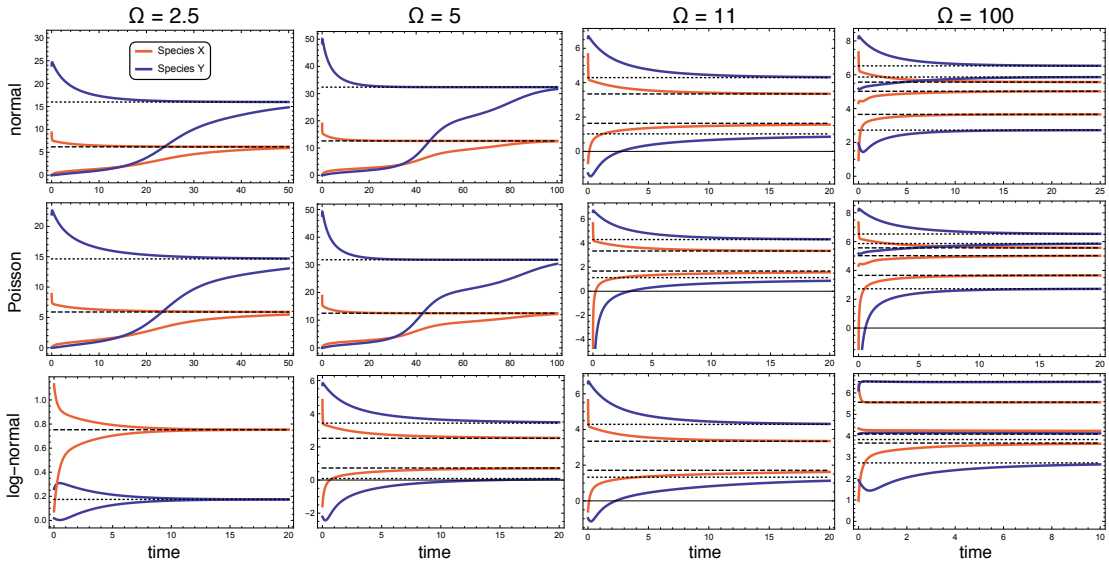

Figure 4.6: Time trajectories for the bistable reaction system in Eqs. (4.8)-(4.10) for different volumes Ω and different initial conditions for the parameters $k_0 = 1, k_1 = 1, k_2 = 5, k_3 = 0.2$ and $k_4 = 5$. The dashed and dotted lines indicate the respective positive and stable fixed points of species X and Y. Depending on the volume, the MAs have one, two or three positive stable fixed points.

Validity

Qualitatively, we find a similar behaviour for the three different MA methods. As for the deterministically monostable system studied in Section 4.3, we find that the MAs show unphysical behaviour *below* a certain critical volume. However, in contrast to Section 4.3, we find that the MAs become bistable *above* a second, larger critical volume, and thus fail to give a physical result since moments of a distribution are uniquely defined. We thus find that *for the studied deterministically bistable system, the MAs fulfil the validity conditions only in an intermediate range of system volumes*.

The bistability of the MAs for large volumes may led one to the interpretation that the MAs approach the (bistable) rate equations for large volumes. Interestingly, however, we find that when increasing the volume further all three MAs become *tristable*, i.e., have three positive stable fixed points, see Figure 4.5. This means the MAs have more positive stable fixed points than the rate equations here, the latter being bistable independent of the volume, and thus the MAs have no physical interpretation anymore whatsoever. In [18] it has been shown that for monostable systems, the normal MA becomes equivalent to the rate equations for the mean values in the limit of large vol-

parameters					normal		Poisson		log-normal	
k_0	k_1	k_2	k_3	k_4	$\log(\Omega_1)$	$\log(\Omega_2)$	$\log(\Omega_1)$	$\log(\Omega_2)$	$\log(\Omega_1)$	$\log(\Omega_2)$
0.5	2	2	0.5	2	< -11	2.4	< -11	2.4	0.75	1.0
0.5	4	1	0.25	2	-5.4	-0.59	< -11	-0.11	-1.9	0.58
1	4	1	0.5	2	< -11	-0.75	-2.5	0.96	0.06	0.49
2	4	2	0.5	4	-4.7	1.6	< -11	1.7	0.05	0.25
0.25	4	1	1	1	< -11	2.2	-1.8	2.3	0.73	1.4
1/3	3	3	1/3	3	-4.9	0.59	< -11	0.57	-1.5	1.3
5	5	1	0.2	5	-5.8	-0.34	< -11	-0.25	-1.7	-0.74
0.2	1	1	0.2	1	-4.2	1.3	< -11	1.4	-0.13	0.85
1	1	5	0.2	5	-4.4	1.0	< -11	0.72	-0.38	0.85
0.2	5	5	0.2	5	-6.0	0.40	< -11	0.40	-3.1	2.0

Table 4.1: Range of validity in the volume Ω on logarithmic scale for different parameter sets for the bistable reaction system in Eqs. (4.8)-(4.10). Ω_1 and Ω_2 denote the left and right end of the validity interval, respectively. We have only checked for fixed points down to a volume of e^{-11}. The term "< -11" thus indicates that the lower boundary of the corresponding validity interval is smaller than e^{-11}.

umes. One can easily show that the result also applies to the Poisson, log-normal and CMN-MA. Here, we find numerically that the tristability remains for volumes up to 10^{10}, which suggests that the convergence of the MAs to the REs in the limit of large volumes does not hold for deterministic bistable systems. Figure 4.6 shows the time trajectories for the MAs for different volumes, verifying that the MAs can indeed have one, two or three positive stable fixed points depending on the volume.

Table 4.1 lists the endpoints of the validity interval for the MAs for ten different parameter sets on logarithmic scale. Figure 4.7 visualises these. We observe that the log-normal MA has a much smaller validity range than the other two MAs. The normal and Poisson MA most of the time have a similar upper bound while the lower bound is generally smaller for the Poisson MA. We thus find that in terms of validity, the log-normal MA performs significantly worse than the other two MA schemes for the reaction system studied here.

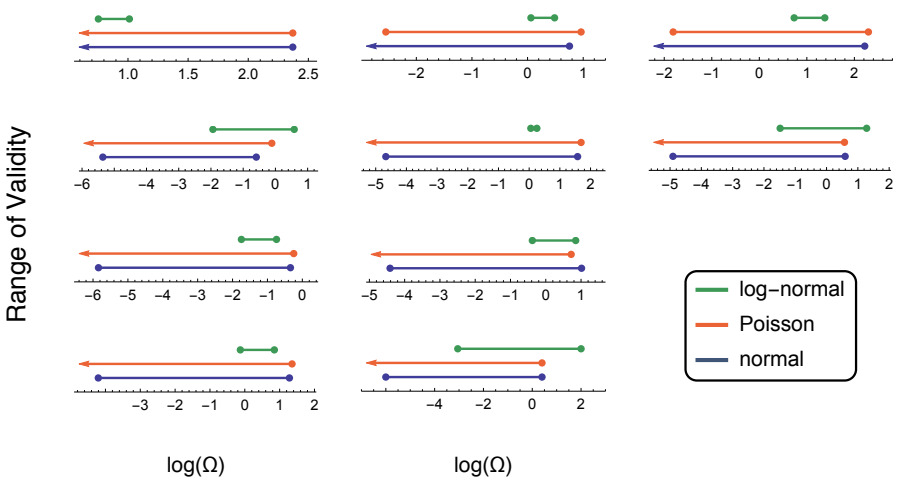

Figure 4.7: Visualisation of validity intervals shown in Table 4.1 on logarithmic scale in the volume for the same ten parameter sets as used in the table. For a lower bound smaller than e^{-11} the lines have an arrow pointing to the left. We find that the log-normal MA's range of validity is significantly smaller than that of the normal and Poisson MAs.

Accuracy

We next compare the prediction of the different MA schemes and of the rate equations for the mean copy numbers of species X and species Y in steady state with results obtained from exact stochastic simulations using the SSA. The latter have been performed using the software package iNA [87]. Figure 4.8 shows the mean values of species X as a function of the volume for the ten parameter sets used in Table 4.1. The corresponding figures for species Y look very similar and are not shown here. The result is divided by the corresponding SSA result. The range of volumes shown corresponds roughly to the validity range of the normal and Poisson MA. We observe here again that the MAs become bistable for larger volumes and that the validity interval of the log-normal MA is significantly smaller than the one of the normal and Poisson MA.

We find that the MAs overestimate the mean copy numbers and that the deviation from the SSA result increases for decreasing volumes. Where two or all three MAs are valid and thus comparable, the accuracy of the different methods is similar, with the log-normal MA being slightly more accurate than the other two and the normal MA being slightly more inaccurate than the Poisson MA. Note, however, that for most parameter sets the log-normal MA's range of validity is significantly smaller than that of the other MAs.

For large volumes, the MAs have two positive stable fixed points converging to

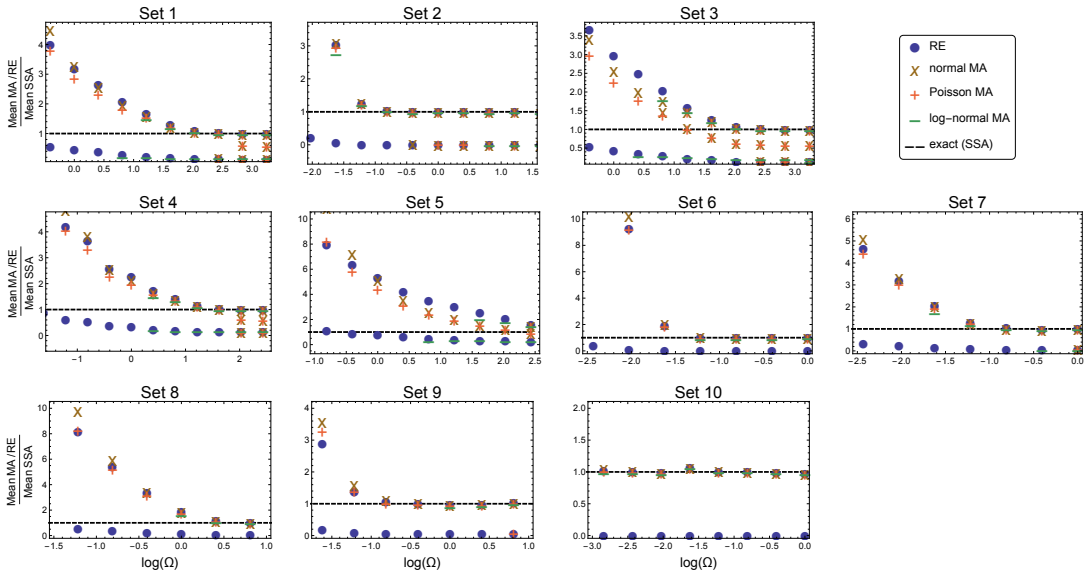

Figure 4.8: Mean value of species X in steady state obtained from moment closures and rate equations as a function of volume Ω on logarithmic scale for the bistable reaction system in Eqs. (4.8)-(4.10). The parameter sets are the same as in Table 4.1. The values are divided by the corresponding result obtained from stochastic simulations using the SSA. The horizontal dashed line thus indicates the exact value. For the SSA result 10^4 samples were simulated for each point.

the two positive stable fixed points of the rate equations. The exact result obtained from SSA simulations agrees with the larger of these two fixed points. The third fixed point of the MAs for large volumes seems to always lie between the two of the rate equations. While it lies exactly in the middle for the normal and Poisson MA, it is very close to the lower one for the log-normal MA. We find the same behaviour for all parameter sets. Note though that this can not be seen for all parameter sets in Figure 4.8 due to the small plot range.

4.4.2 A deterministic ultrasensitive system

Next, we study an enzyme catalysed protein-phosphorylation system with reactions

$$P + E_1 \underset{d_1}{\overset{a_1}{\rightleftharpoons}} E_1P \xrightarrow{k_1} P^* + E_1, \tag{4.11}$$

$$P^* + E_2 \underset{d_2}{\overset{a_2}{\rightleftharpoons}} E_2P^* \xrightarrow{k_2} P + E_2. \tag{4.12}$$

This system shows ultrasensitivity for certain parameter values [88], namely when the enzymes are saturated, i.e., when most enzymes are on average in the complex state.

Here, P and P^* denote the non-phosphorylated and phosphorylated forms of the protein, respectively, E_1 and E_2 the phosphorylating and de-phosphorylating enzymes, respectively, and E_1P and E_2P^* the respective protein-enzyme-complexes. In [88] the authors studied the dependence of the ratio of phosphorylated to non-phosphorylated proteins as a function of w_1/w_2 with $w_1 = k_1E_1^t$ and $w_2 = k_2E_2^t$ in the deterministic setting, where E_1^t and E_2^t are the conserved total numbers of the respective enzymes in the system. Assuming a Hill-type response curve, the corresponding Hill coefficient is often used to quantify the steepness of the response. The authors speak of an "ultrasensitive response" whenever the response is steeper than a Michaelis-Menten response, i.e., has a Hill coefficient of larger than unity.

We study here the effect of noise on the ultrasensitive response and again compare moment closure results with SSA simulations. The latter have been performed using the software package iNA [87]. First, however, we describe a surprising non-uniqueness of the Poisson and log-normal MA and study the validity of the different MA schemes. As we have explained below Eq. (4.10), the second-order normal and second-order CMM-MA are identical for elementary reaction systems with mass-action kinetics. Since this is the case here, we only study the normal, Poisson and log-normal MAs in the following.

Non-uniqueness for reduced systems

The studied reaction system in Eqs. (4.11) and (4.12) has six species: P, P^*, E_1, E_2, E_1P, E_2P^*, and three conservation laws: the total number of proteins and the total numbers of the respective enzymes, i.e., $P+P^*+E_1P+E_2P^*$, E_1+E_1P and $E_2+E_2P^*$, are conserved, where we use the symbol for the species also as the corresponding molecule number variable in a slight abuse of notation. The conservation laws allow one to reduce the system to three variables, which is obviously of computational advantage. There are two ways of obtaining the reduced moment closure equations: arguably, the standard approach would be to start from the reduced CME, compute the reduced moment equations and subsequently apply the moment closure. Alternatively, one may start from the full CME, compute the moment closure equations and afterwards reduce the equations by taking the conservation laws into account. One may expect, or require, the two approaches for a sensible moment closure scheme to be equivalent. It is easy to show that this is indeed the case for the normal and CMN moment closures. However, we find here that this is not the case for the Poisson and log-normal MA. We thus conclude that *the Poisson and log-normal MAs are generally not uniquely defined*

if one reduces a system according to conservation laws in molecule numbers, a clear flaw of these methods. The reason for the non-uniqueness of the MA equations is that while the moment equations depend on *diagonal* higher-order moments if one starts from a reduced CME, no such dependence is found if the MA equations are derived from the full CME. While the normal and CMN-MAs treat diagonal and non-diagonal moments equivalently, the Poisson and log-normal MAs do not do so, thus leading to non-uniqueness. We explain this in more detail in Appendix B.

One consequence of this non-uniqueness is that certain symmetries of the system are broken. Looking at the reaction system in Eqs. (4.11) and (4.12), one sees that the system is symmetric under exchanging species labels and reaction constants, $P \leftrightarrow P^*$ and $E_1 \leftrightarrow E_2$ and $a_1 \leftrightarrow a_2, d_1 \leftrightarrow d_2$ and $k_1 \leftrightarrow k_2$. This means that for $a_1 = a_2, d_1 = d_2$ and $k_1 = k_2$ the mean values of P and P^*, E_1 and E_2, as well as E_1P and E_2P^* should be respectively equal. We find that this is indeed the case for the normal and CMN moment closure, and also for the Poisson and log-normal MAs if one derives the equations starting from the full CME. If one applies the Poisson and log-normal MAs to the reduced CME, however, *they do break the symmetry*.

We conclude that one should be careful when using the Poisson or log-normal MA for systems with conservation laws. In case the MAs are non-unique it is favourable to first derive the MAs before applying the conservation laws. In the following we will study the opposite cases, i.e., if the Poisson and log-normal MA are applied to the reduced CME, which would normally be the standard approach.

Validity

As in [88] we define $w_1 = k_1 E_1^t$ and $w_2 = k_2 E_2^t$. The authors in [88] studied the dependence of the fraction of the protein number in the phosphorylated state as a function of w_1/w_2 using deterministic rate equations. The authors call this response "ultrasensitive" whenever it is steeper than a Michaelis-Menten response, meaning a Hill-coefficient larger than one. Here, we would like to study the effect of noise on the response and investigate how different moment closures perform for this system. To this end, we compute the mean value of the phosphorylated protein P^* in steady state on a grid in w_1/w_2 with all the other parameters fixed and fit a Hill function $(w_1/w_2)^{n_H}/(K_d + (w_1/w_2)^{n_H})$ to the result, where K_d and n_H are the dissociation constant and the Hill coefficient, respectively.

We find that the normal MA and rate equations are valid for all w_1/w_2 for all chosen

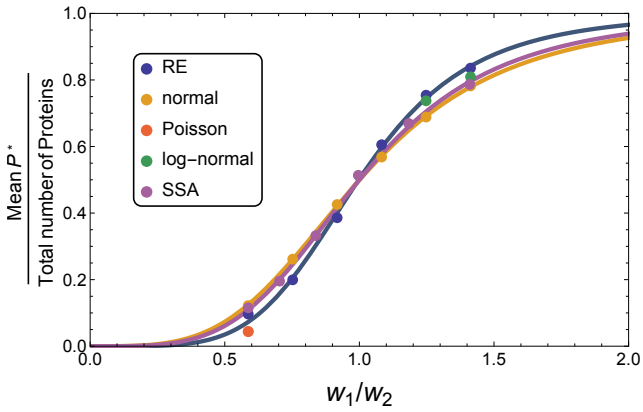

Figure 4.9: Fraction of mean phosphorylated protein in steady state as a function of w_1/w_2 for the protein phosphorylation system in Eqs. (4.11) and (4.12). The blue and orange curve are Hill-functions fitted to the points of the RE and normal MA, respectively. The Poisson and log-normal MAs have only few positive stable fixed points in the response region making a sensible fit impossible. The used parameters are $a_1 = a_2 = 5, d_1 = d_2 = 1, k_1 = k_2 = 1, \Omega = 1, E_1^t = E_2^t = 7$ and $P^t = 15$, where E_1^t, E_2^t and P^t are the total number of enzyme E_1, the total number of enzyme E_2 and the total number of proteins in the system, respectively. For the SSA result 10^4 samples were simulated for each point.

parameter sets, whereas the Poisson and log-normal MA are not valid for certain parameter regimes, i.e., they do not always have a positive stable fixed point. Figure 4.9 visualises the fitting procedure for one parameter set. While the rate equations and normal MA are stable on the whole considered response region in w_1/w_2, the Poisson and log-normal MAs are unstable for the major part of the region. We obtain only one and two values in the response region, respectively. The Poisson and log-normal MAs thus do not allow a sensible estimate of the response-steepness via a fit of a Hill-function.

Figure 4.10 visualises the validity of the rate equations, normal, Poisson and log-normal MAs as a function of the total enzyme number and w_1/w_2 for five different parameter sets. The figure indicates where the methods have a positive stable fixed point and where not. In addition, when a positive stable fixed points exists, we solve the time-dependent MAs with the initial condition being the fixed point of the rate equations for the corresponding parameters, and the figure indicates the regions where these diverge despite the existence of a positive stable fixed point. This thus indicates the sensitivity of the different methods to initial conditions. While the rate equations and normal MA are stable and the time trajectories converge everywhere, the Poisson

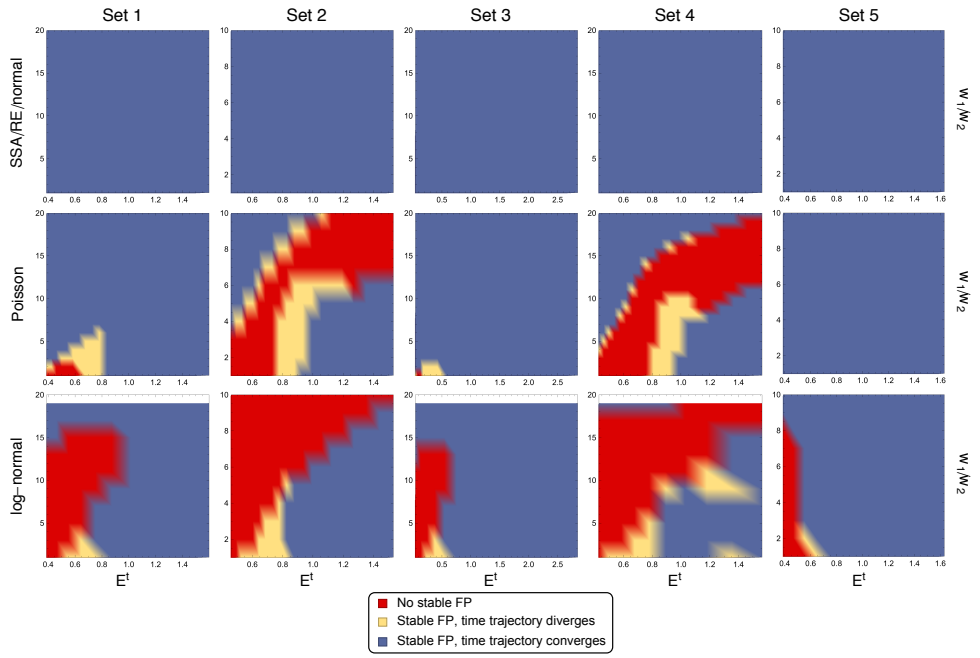

Figure 4.10: Validity of different MAs as a function of the total enzyme numbers $E_1^t = E_2^t = E^t$ and of w_1/w_2 for the protein phosphorylation system in Eqs. (4.11) and (4.12) for five different parameter sets. If we write (a, d, k, P^t, Ω) with $a_1 = a_2 = a$, $d_1 = d_2 = d$ and $k_1 = k_2 = k$, where P^t is the total protein number and Ω is the volume, the parameter sets are given by Set 1 $=(1,1,1,25,0.3)$, Set 2 = $(5,1,1,15,1)$, Set 3 = $(5,2,2,25,1)$, Set 4 = $(10,1,1,25,1)$ and Set 5 = $(1,1,1,20,1)$. The red regions indicate that the methods have no positive stable fixed point. The blue regions indicate where a positive stable fixed points exists *and* the time trajectories converge with initial condition being the fixed point of the rate equations. The yellow regions show where the time trajectories diverge despite the existence of a positive stable fixed point, which means that the fixed point is only locally attractive.

and log-normal MA do so only in subregions of the parameter space. Note that we do not make any statements about *unstable* fixed points here since we investigated the convergence of time trajectories only for one fixed initial condition. The divergence of the time trajectories in the yellow region suggest that there exists an unstable positive fixed point, but the same might be true in some parts of the blue region despite the convergence of time trajectories.

It is important to mention that, despite being valid for all investigated parameter regimes here, the normal MA does not remain valid if the volume is decreased sufficiently enough (as do the other MA methods). So as for the previous example systems, there is a critical volume below which the MAs fail to give physical predictions of the

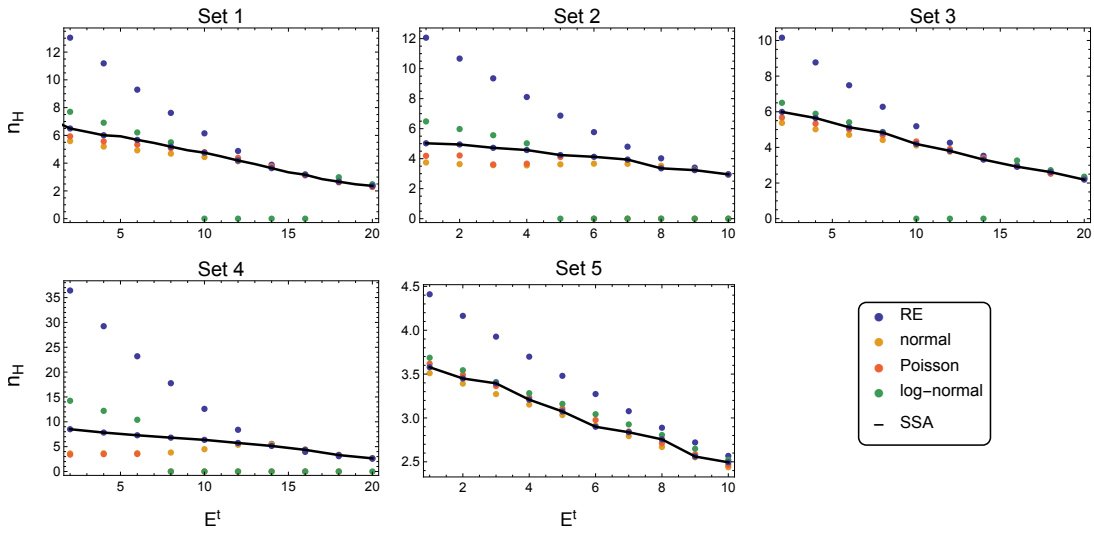

Figure 4.11: The Hill coefficient as a function of total enzyme number for the five different parameter sets introduced in Figure 4.10 for the protein phosphorylation system in Eqs. (4.11) and (4.12). The SSA result is shown as a solid black line. As explained in the main text, for some parameter values the Poisson and log-normal MA do not allow to estimate a Hill function due to instability. In such cases we set the Hill coefficient to zero. For the SSA result 10^4 samples were simulated for each point.

system.

In conclusion, we find that the normal MA performs significantly better than the Poisson and log-normal MA for the studied system in terms of validity.

Accuracy

Next, we compare the Hill coefficient obtained from the different methods with the results obtained from SSA simulations as a function of the total enzyme number E^t for the five parameter sets defined in the caption of Figure 4.10. The SSA simulations were performed using the software package iNA [87]. If a method did not allow to estimate a Hill coefficient for some E^t because its validity range was too small, we set the corresponding value to zero. Figure 4.11 illustrates the results. First of all, we find that the rate equations overestimate the Hill coefficient for all E^t, with a larger deviation for small E^t, which means that the noise in the system significantly reduces the steepness of the response. For small E^t the Hill coefficient estimated from the rate equations becomes up to four times larger then the SSA result (Set 4 in Figure 4.11). Whenever they allow to estimate a Hill coefficient, the moment closure approximations are more

accurate than the rate equations. While the normal and Poisson MAs underestimate the response, i.e., overestimate the influence of noise, the log-normal overestimates the response. In terms of accuracy, the three methods perform very similarly, the Poisson MA perhaps being slightly more accurate than the other two methods. However, this slightly higher accuracy of the Poisson MA does not overcome its disadvantage of instability described before.

4.4.3 A deterministic oscillatory system

Next, we study the Brusselator, a well-known deterministic oscillating chemical system given by [89, 90]

$$2X + Y \xrightarrow{c_1} 3X, \quad X \xrightarrow{c_2} Y, \quad \varnothing \underset{c_4}{\overset{c_3}{\rightleftharpoons}} X. \tag{4.13}$$

Depending on the parameter values, the deterministic rate equations show sustained oscillations, damped oscillations or over-damped oscillations. Single SSA trajectories may show sustained oscillations, while ensemble averages of the SSA always show damped or over-damped oscillations (unless an external, time-dependent input is applied) due to the dephasing of independent trajectories. Therefore, in the absence of an external force, a MA can only be interpreted as a valid moment approximation if its trajectories show damped or no oscillations. Here, we want to first study the validity of the different MA methods for different parameter sets, and then analyse their behaviour if the system becomes entrained by an external force. Note that the first reaction in (4.13) is trimolecular, which means that the corresponding propensity function is of third order in the molecule numbers (c.f. Eq. (2.25)). The time evolution equation of the second-order moments thus depend on the third and fourth-order moments (c.f. Eq. (4.2)). Therefore, since the fourth-order central moments and fourth-order cumulants are not identical (in contrast to the third-order ones), the normal and CMN-MAs are not equivalent for the reaction system in Eq. (4.13) and we hence analyse all four MAs separately in the following.

Validity

We study here the validity of the MAs for three different parameter sets defined in the caption of Figure 4.13. Not surprisingly, as for the previously studied systems, we find that the MAs show unphysical behaviour *below* a certain critical volume. In contrast to

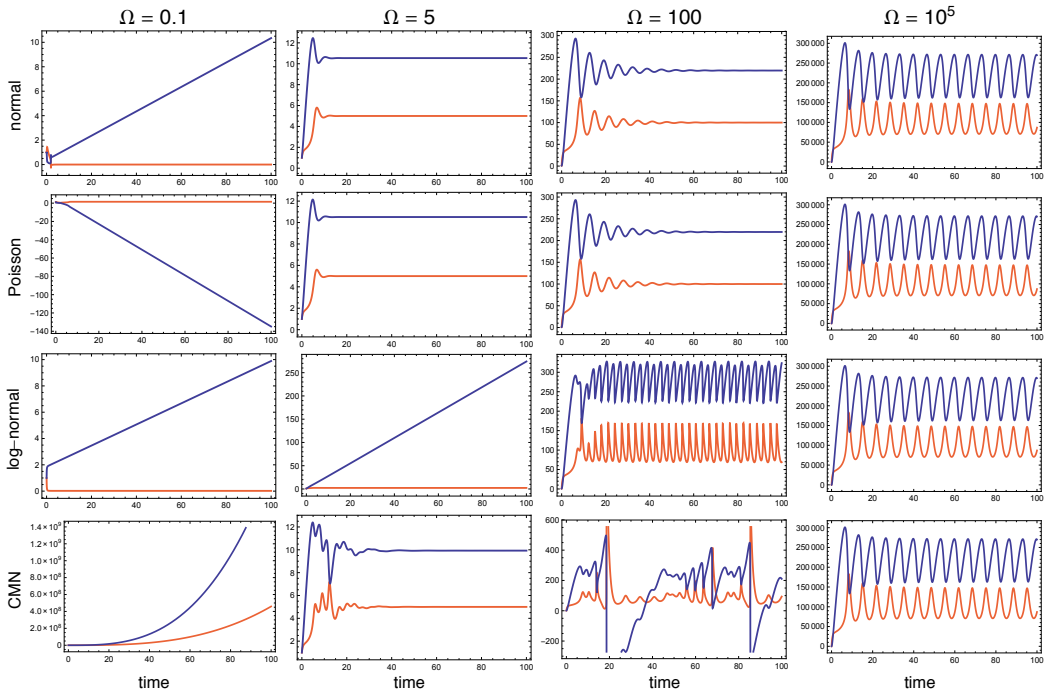

Figure 4.12: Time trajectories of the mean of species X (blue line) and Y (orange line) for several volumes for the Brusselator system in Eq. (4.13) for the parameters $(c_1, c_2, c_3, c_4) = (0.9, 2, 1, 1)$. While the normal, Poisson and CMN-MAs give physically meaningful results, i.e., damped oscillations, for an intermediate range of volumes, the log-normal MA fails to do so for all volumes. To minimise the possibility of numerical effects we computed the shown results using the ODE integration methods “Adams", “Backward Differentiation Formula", “Explicit Runge Kutta", “Implicit Runge Kutta", “Explicit Midpoint" and “Stiffness Switching" and varied the step sizes of the numerical integrator over several orders or magnitude, all giving the same results.

the two previous cases, however, we find here that the MAs show sustained oscillations and therefore unphysical behaviour *above* a second, larger critical volume. So similar to the deterministically bistable system in Section 4.4.1, we find that *the MA methods are valid only for an intermediate range of volumes*. Surprisingly, however, for the log-normal we can not find such an intermediate regime, i.e., the two critical volumes seem to coincide.

Figure 4.12 shows the time trajectories of the moments for the different MAs for four different volumes for one fixed parameter set. While the normal, Poisson and CMN-MAs diverge for small volumes, are monostable for intermediate volumes and show sustained oscillations for large volumes, the log-normal switches directly from divergent to oscillatory behaviour. We estimated the range of validity for the three

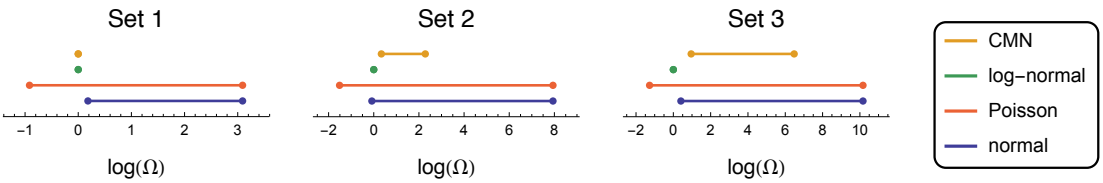

Figure 4.13: Range of validity for the Brusselator system in Eq. (4.13) for three different parameter sets as a function of the volume Ω on logarithmic scale. The used parameters for (c_1, c_2, c_3, c_4) are Set 1 = $(1, 3, 0.9, 1)$, Set 2 = $(0.9, 2, 1, 1)$ and Set 3 = $(1, 2, 1, 1.5)$. If the range of validity has length zero we plot a single point at zero. By "range of validity" we mean the range of volumes for which the MAs give physically meaningful (i.e., non-negative and converging) time trajectories.

different parameter sets for fixed initial conditions of unity for the mean values of both species and zero variance. Figure 4.13 shows the ranges of validity on logarithmic scale in the volume. While the Poisson and normal MA have a finite range of volumes where they lead to physically meaningful results for all parameter sets, the CMN-MA has a vanishing one for one parameter set and the log-normal for all parameter sets.

System with entrainment

In systems biology it is frequently of interest to study systems where one or several propensity functions are time dependent. For example, circadian oscillators are often modelled by a deterministic oscillatory system with an imposed periodic propensity function modelling the influence of an external light input [91, 92, 93]. Here, we want to study the performance of the different MA schemes for such a system in the stochastic setting. To this end, we modify the rate constant c_2 of the second reaction in Eq. (4.13) such that it varies over time from 0.5 to 1.5 times the chosen mean value in a sinusoidal way, i.e., $c_2(t) = c_2^0 \times (1 + \frac{1}{2}\sin(\omega t))$ where c_2^0 is the fixed mean value of c_2 and the frequency ω of the sine curve is chosen to be the oscillation frequency of the deterministic system. After ten periods, we switch off the time dependence and fix c_2 to its mean value. Since we have a time-dependent propensity function here, we cannot use the SSA to simulate the system. We therefore use Extrande, a recently developed exact MC method to sample from the solution of CMEs with time-dependent rate functions [94].

Figure 4.14 shows the time trajectories for the rate equations, Extrande simulations

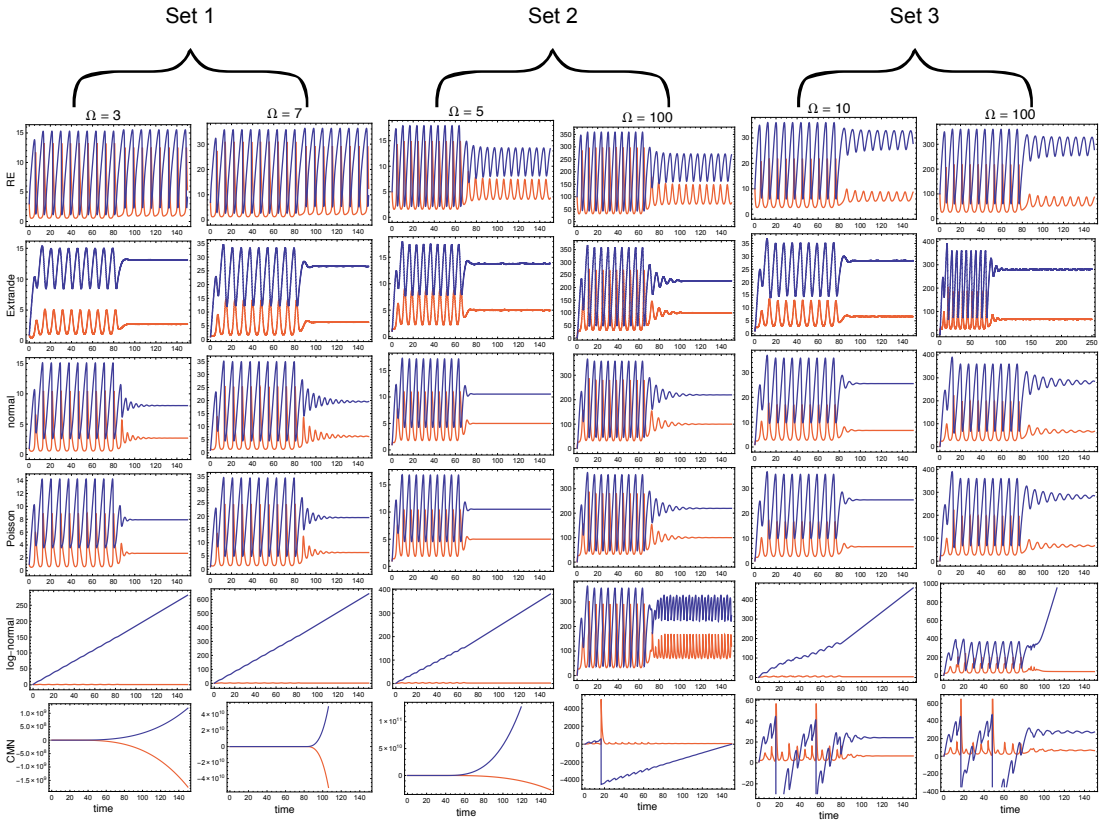

Figure 4.14: Time trajectories for the Brusselator system in Eq. (4.13) for the three parameter sets defined in the caption of Figure 4.13 with entrainment for two different volumes for each parameter set. The red and blue lines denote the mean values of species X and Y, respectively. The external input gets switched on at time $t = 0$ and switched off after ten oscillation periods of the deterministic system (which depends on the given parameter set). For the Extrande result we simulated 10^5 samples for Set 1 and 10^4 samples for Set 2 and Set 3, respectively.

and the different MA methods. We find that the rate equations show sustained oscillations after entrainment, while the Extrande results show damped or over-damped oscillations. The normal and Poisson MA behave qualitatively the same way as the Extrande and are thus valid moment approximations for the chosen parameter values. Quantitatively they differ quite significantly from the Extrande result, however. They underestimate the mean values, show oscillations with larger amplitudes during entrainment and a weaker damping after entrainment. Looking at Figure 4.14 one finds that these effects are stronger for the respective smaller volume for each parameter set. The normal and Poisson MA thus *underestimate* the influence of noise here. The log-normal and CMN-MAs fail everywhere to provide a physical result. For the former

this may to be expected, since it also failed to do so in the case without entrainment. Interestingly, however, the CMN-MA is invalid even for parameters for which it is valid in the case without external input. Overall, the normal and Poisson MA seem to perform significantly better for this system than the log-normal and CMN-MA.

4.5 MOCA - A software package for moment closure approximations

The derivation of the moment equations from the CME and the subsequent application of moment closures is conceptually a straightforward task. Practically, however, it becomes extremely cumbersome if more than one species is involved and if one considers higher-order MAs. Suppose for example a system of three species for which we want to compute the fourth-order normal MA equations. Taking symmetries into account, this leads to 34 moment equations which have to be derived from the CME. These will have to be closed, and several fifth-order moments (and potentially higher-order moments) will have to be replaced in terms of lower-order moments. Obviously, this task quickly becomes unfeasible to do manually. Moreover, the numerical analysis of MA equations is not straightforward, and there is no user-friendly software package available allowing non-expert users to derive and analyse MAs.

To our knowledge, there are three software packages available in the literature for moment closures: the Matlab toolbox StochDynTools [95] which allows the derivation of MA equations using several different closure schemes for mass-action chemical systems, i.e., those with polynomial propensity functions; the Python package MomentClosure [96] which allows the same but only for the normal moment closure and has the facility to export the MA equations to a Maple file for further analysis; and a Matlab toolbox presented in [97] which allows to use the normal moment closure to second order for mass-action chemical systems. For the application of all three packages, the user needs to be familiar with the respective programming language and the numerical analysis is not automated.

Here we present the Mathematica package *MOCA* (moment closure analysis) which was used for the presented numerical analysis. MOCA significantly extends the applicability and functionality of the three software packages mentioned above. It implements the investigated four moment closure approximations, as well as deterministic rate equations, in a graphical user interface and is freely available via the website [98]

```
SetDirectory[NotebookDirectory[]];
<< MOCA.m
nS = 2;
stochMatrix = {{{0, -1, 1, 0}, {1, -1, 1, -1}}};
parameters = {k1, k2, k3, k4};
propensity = {k1 x1, k2 x1 x2, k3 (1 - x1), k4 x2};
```

Figure 4.15: MOCA input for time-independent propensity functions for the gene system in Figure 2.2. The first two lines do not need to be modified. They set the directory of the file and load the package MOCA.m. The following lines define the number of species, stoichiometric matrix, parameters and propensity functions of the system, respectively. We have set the volume to $\Omega = 1$ here.

or in the supplemental material of [40]. In contrast to other available moment closure software packages [95, 96, 97], MOCA does not only derive the closure equations but also automatically performs numerical analysis of the derived equations, making the methods available to non-expert users. The results are automatically visualised and can be exported to various formats.

We show here some functionalities of MOCA by applying it to our running example of Chapter 2; the gene system in Figure 2.2 with reactions in Eq. (2.23).

Applicability

MOCA extends the applicability over existing moment closure packages to

- non-polynomial propensity functions
- time-dependent propensities functions
- propensities defined on discrete time points (e.g. measured fluctuating external parameter)

Note that while non-polynomial propensities can often give a useful description of a system, they should really be interpreted as an effective approximate description of a set of elementary reactions, valid only under certain conditions [99]. For these type of propensities the software applies a Taylor expansion of the propensity around the mean value to a specified order as proposed in [100]. These different features make MOCA applicable to virtually any reaction system with arbitrary propensity functions.

In addition to the different moment closure methods described above, MOCA allows the user to define his own moment closure method, providing an easy way to develop novel moment closure schemes.

User input

To use the package, the file MOCA.m needs to be placed in the same folder as the Mathematica notebook that will be used for the analysis. Figure 4.15 shows an example input for the corresponding notebook for the reaction system in Figure 2.2. The first two lines, which set the path and load the package, respectively, have to be executed without any modification. Next, the number of species and the stoichiometric matrix have to be specified and assigned to the variables **nS** and **stochMatrix**, respectively, as depicted in the third and fourth line in Figure 4.15. The propensity vector and stoichiometric matrix are given in Eq. (2.29). The number of species **nS** has to be equal to the number of rows of **stochMatrix**. Next, the parameter vector **parameters** and the propensity vector called **propensity** need to be specified, as done in the fifth and sixth input lines in Figure 4.15.

The species variables have to be denoted by an "x" with the species index as a subscript. All terms in the propensity function that are not species variables or numerical values have to be listed as a parameter in **parameters**. This is all the input needed if dealing with time-independent propensity functions and when using the GUI. Note that the propensities do not need to be of mass-action, i.e., polynomial type, they can have any analytical form.

For using the coding version of MOCA, deterministic rate equations and time-dependent propensity functions, as well as for the definition of new moment closure methods, see the corresponding tutorial files in the supplemental material [98].

Analysis - the graphical user interface

There are four functions available within a GUI. They simply need to be typed into the notebook and evaluated to open the corresponding GUI:

- **DeriveEquations**: derives the MA equations for central moments for general parameters and allows to assign numerical values to the parameters.
- **SteadyState**: numerically searches for positive and stable fixed points of the MA equations.

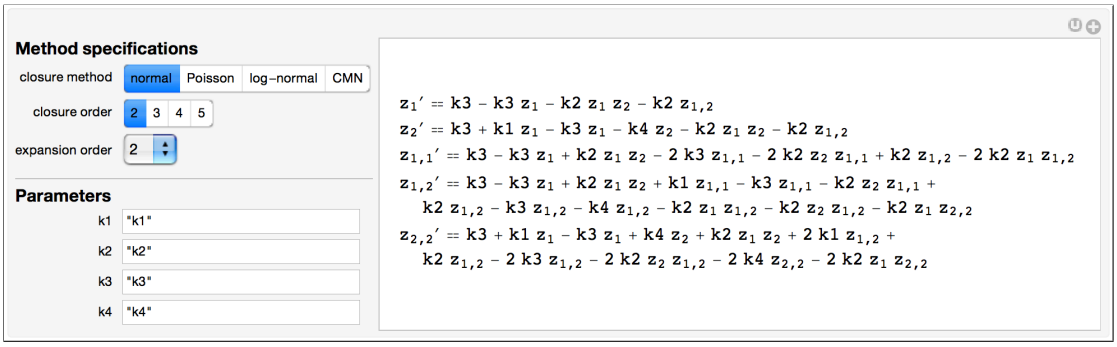

Figure 4.16: GUI for deriving MA equations with MOCA for the gene system in Figure 2.2. After defining the system as in Figure 4.15 the command **SteadyState** has to be evaluated in the notebook for the GUI to appear. The user can choose the closure method, closure order, expansion order and specify parameter values. For changes to apply the user needs to press the little “update button" in the top right corner.

- **SteadyStateVaryParameter**: same as **SteadyState** but with one parameter varied over a grid specified by the user. The resulting table can be exported into a “CSV” (“Comma-separated values”) file.
- **TimeTrajectory**: solves MA equations numerically in time for numerical parameter values and plots the result. The result can be exported as a figure to various formats or evaluated on a grid in time and stored in a “CSV” file.

Figure 4.16 shows the GUI that appears after typing and evaluating **DeriveEquations**. The user can interactively choose a moment closure method, the closure order as well as the expansion order. By “expansion order" we mean the expansion of the propensity functions around the mean value as proposed in [100]. This is only necessary for non-polynomial rate functions. For exclusively polynomial rate functions, the expansion does not make a difference as long as its order is equal to or higher than the maximum order of the propensity polynomials. Finally, it is possible to assign numerical values to the parameters. The equations only become updated when the small “update bottom" in the top right corner is clicked. This is also true for the functions described in the following, i.e., changes in the input are only applied after clicking the “update bottom".

The function **SteadyState** allows to numerically compute positive stable fixed points of the MA equations. It has the same input parameters as the function **DeriveEquations** described before, with the difference that the parameters have an initial numerical value. For some parameter values, the method cannot find a positive and stable fixed point. However, this does not necessarily mean that the numerical algorithm

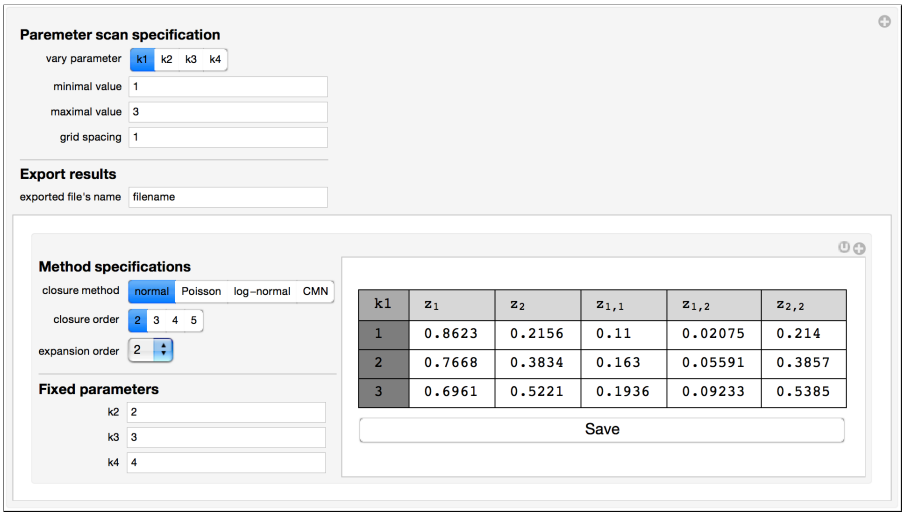

Figure 4.17: GUI corresponding to the command **SteadyStateVaryParameter** in MOCA for the gene system in Figure 2.2. The table shows positive stable fixed points obtained by varying one parameter over a specified grid.

fails. Earlier in this chapter we showed that MA equations can indeed have no positive and stable fixed point for certain bimolecular reaction system (even though the SSA and rate equations do have positive stable fixed points). We also showed that MAs can have more than one positive stable fixed point, in which case the **SteadyState** function may give more than one result.

Similarly, the function **SteadyStateVaryParameter** searches for positive stable fixed points but varies a user specified parameter over a user specified grid. The corresponding GUI is shown in Figure 4.17. The resulting table can be exported to a text file in "CSV" format.

The final function **TimeTrajectory** solves the MA equations numerically in time and plots the result. Figure 4.18 shows the corresponding GUI. In addition to method specifications and values for parameters, the user can specify initial conditions for the mean values of the species (higher-order central moments are set to zero initially, i.e., deterministic initial conditions), the final time point and the plot order specifying up to which order moments should be plotted. The result can either be exported as a figure to various formats or into a "CSV" text file where the solution is evaluated on a time grid with user-specified time spacing.

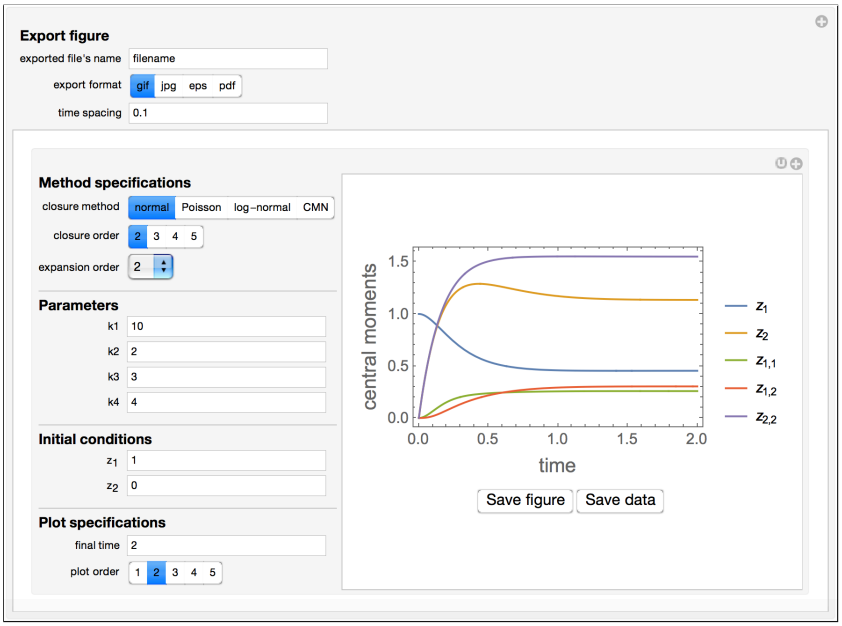

Figure 4.18: GUI for solving and visualising the MA equations numerically in time using the **TimeTrajectory** command of MOCA. In addition to the method specifications, the user can specify initial conditions for the mean values, the final time point as well as up to which order moments should be plotted. The result can be exported as a figure or into a "CSV" file evaluated on a time grid.

Coding commands

The GUI commands described above are also available as Mathematica functions allowing more experienced Mathematica users a more flexible application of the methods. See the example files on [98] or in the supplemental material of [40] for details on how to use these functions.

4.6 Conclusion

In this chapter, we studied several moment closure approximation schemes for the moment equations of the CME in terms of their physical validity and their quantitative accuracy. For non-linear systems, we found that the MAs give physically meaningful results only above a certain critical volume if the system is deterministically monostable, and only for intermediate volumes if the system is not deterministically monostable.

While we found no significant difference in quantitative accuracy between the four MAs, the ranges in parameter space for which the MAs gave physically meaningful

results were significantly larger for the normal MA suggesting that the normal MA is favourable over the other studied methods. We emphasise that these results are exclusively based on numerical analysis and although we confirmed the results for a wide range of parameter sets and several example systems, we cannot expect all of them to hold in general for all parameter sets or chemical reaction systems. This is particularly true for the results on quantitative accuracy. In [101], for example, it has been found for a single parameter set for one chemical reaction system that the log-normal MA is significantly more accurate than the normal MA.

Our analysis revealed a number of non-trivial, surprising aspects of the limiting behaviour of MA equations. We found that all four MAs give physically meaningful results only *above a certain critical system volume* for deterministically monostable systems. Below this threshold, the equations give rise to diverging trajectories or otherwise unphysical behaviour, such as negative mean values or variances and hence do not admit a probabilistic interpretation. For a deterministically bistable and a deterministic oscillatory system, we found that the MAs give physically meaningful results only for *an intermediate range of system volumes*, showing unphysical multistability or unphysical oscillatory behaviour above this regime. Surprisingly, we found for an enzyme-catalysed reaction, that *the Poisson and log-normal MAs were not uniquely defined*. Our analysis suggests that this may indeed be generally the case for systems with conservation laws, a flaw not shared by the other two MAs.

In conclusion, our results suggest that one should be careful when using MAs, since one may be led to physically wrong conclusions otherwise. Our results do not favour one MA over the others in terms of accuracy, but suggest that the normal MA is favourable over the other MAs due to its significantly larger ranges of validity

Finally, we presented the software package MOCA which was used for the numerical analysis of the various MAs. MOCA allows to derive and analyse moment closure approximations for systems with polynomial, non-polynomial as well as time-dependent propensities. MOCA implements the four MA methods studied here as well as user-defined moment closure schemes and automatises the numerical analysis. It allows non-expert users to apply moment closure methods in a user-friendly graphical user interface. We believe that these features make MOCA a useful software tool for studying stochastic reaction systems.

Chapter 5

Cox process representation of stochastic reaction-diffusion processes

In the preceding two chapters we studied approximate methods for the non-spatial description of chemical kinetics, i.e., of the chemical master equation (CME). In the present chapter, we relax the "well-mixed" assumption underlying the CME, which means that the spatial locations and diffusion of particles have to be modelled. In this case the kinetics can be described by *stochastic reaction-diffusion processes* (SRDPs) which we introduced in Section 2.4. Deriving an efficient and accurate inference method for such processes is the goal of this chapter.

To this end we make use of the *Poisson representation* (PR) which we introduced in Section 2.3.7. As explained there, the PR can become complex-valued depending on the involved reactions. After revising SRDPs and discussing the problem of inference for such processes in the next section, we give a classification of different reaction types according to their PR behaviour in Section 5.2, and subsequently introduce approximated versions of certain reaction types. These approximations lead to a PR that is real-valued for all reaction systems. Next, we apply the resulting real-valued PR to the *reaction-diffusion master equation* (RDME) in Section 5.3 and study its continuum limit. Next, we show in Section 5.4 that the derived equations give rise to a novel representation of SRDPs in terms of Cox processes (c.f. Section 2.5), and show how this representation naturally leads to a novel inference method. We demonstrate the efficiency and accuracy of our method for several examples in Section 5.5. Finally, we conclude in Section 5.6.

5.1 Stochastic reaction-diffusion processes

5.1.1 Brownian dynamics

SRDPs are generally defined in an algorithmic way as Brownian dynamics simulations. We employ the Doi picture of SRDPs here in which particles are modelled as point particles and bimolecular reactions happen with a certain probability per unit time whenever the distance between two molecules is less than a certain reaction range.

For simulations, time is discretised into small intervals, and each particle in the system and the occurrences of chemical reactions are simulated explicitly in each time step. The particles undergo Brownian diffusion and chemical reactions happen with certain probabilities per time step. For details see Section 2.4.1.

As we explained in Section 2.4.1, the time step discretisation has to be chosen small enough to fulfil several requirements. For example, if the system is not infinite but has certain boundaries, the time step has to be small enough to avoid unphysical boundary effects. Another example is that the probability that one particle participates in more than one chemical reaction in one time step should be negligible. These requirements generally make small time steps necessary, and since every single particle has to be simulated explicitly, such simulations are computationally extremely expensive. This is particularly true whenever bimolecular reactions are involved. In this case, the distances between all particles that may react with each other have to be computed in each time step. With increasing particle numbers, this becomes extremely laborious.

To get a feeling for the computational complexity of SRDP simulations, consider the stochastic simulation algorithm (SSA) introduced in Section 2.3.3. The SSA simulates the stochastic process underlying the CME description, i.e., in the case where a non-spatial description is valid. In this case the simulation of single particles and the discretisation of time are both not necessary. Rather, the state of the system is simply given by the total particle counts for all species, and time is propagated by simulating the occurrences of reactions. The latter is possible in this case since inter-reaction times are exponentially distributed and hence can be simulated easily. This is computationally obviously much less expensive then the simulation of SRDPs. But even for the SSA, the simulation of larger systems becomes expensive, and in particular inference tasks can become infeasible. It is therefore evident that inference attempts for SRDPs based on naive Brownian simulations are likely to fail.

More importantly than these computational difficulties, it is not even clear conceptually how to define a likelihood for SRDPs. While it is obvious how to define a

likelihood for non-spatial systems described by the CME (namely by simply evaluating the CME's solution at the measured number of molecules), this is not the case for SRDPs. It is thus not even clear conceptually how to perform inference for SRDPs.

5.1.2 The reaction diffusion master equation

Let us next briefly review the RDME formalism introduced in Section 2.4.2. The RDME can be viewed as a coarse-grained approximate description of SRDPs. We consider here the case where space is discretised into cubic compartments of edge length h. The considerations in the following can easily be generalised to more general discretisations. The RDME assumes the system to be well-mixed and dilute in each compartment. The dynamics in each compartment can hence be modelled by means of a CME. The diffusion of particles between compartments is modelled as linear reactions between adjacent compartments with rate constant $d = D/h^2$, where D is the microscopic diffusion constant.

Now let L be the number of compartments, $\mathcal{N}(l)$ denote all adjacent compartments of the l^{th} compartment, $f_r(\mathbf{n})$ be the propensity function of reaction r, δ_i^l be a vector of length $N \times L$ with the entry corresponding to species X_i in the l^{th} compartment equal to 1 and all other entries zero and $\mathbf{S}_r^l$ be a vector of length $N \times L$ with the entries corresponding to the l^{th} compartment equal to $\mathbf{S}_r$ and zero otherwise, where $\mathbf{S}_r$ is the r^{th} row of the stoichiometric matrix $\mathbf{S}$. Let further $\mathbf{n} = (n_1^1, \ldots, n_N^1, \ldots, n_1^L, \ldots, n_N^L)$ denote the state of the system, where n_i^l is the copy number of species X_i in the l^{th} compartment, and let $\mathbf{n}^l = (n_1^l, \ldots, n_N^l)$ denote the state vector of the l^{th} compartment. The RDME then governs the time evolution of the probability $P(\mathbf{n},t)$ to be in state $\mathbf{n}$ at time t and reads

$$\begin{aligned}\partial_t P(\mathbf{n},t) = &\sum_{l=1}^{L} \sum_{m \in \mathcal{N}(l)} \sum_{i=1}^{N} d_i [(n_i^m + 1) P(\mathbf{n} + \delta_i^m - \delta_i^l, t) - n_i^l P(\mathbf{n},t)] \\ &+ \sum_{l=1}^{L} \sum_{r=1}^{R} [f_r(\mathbf{n}^l - \mathbf{S}_r) P(\mathbf{n} - \mathbf{S}_r^l, t) - f_r(\mathbf{n}^l) P(\mathbf{n},t)].\end{aligned} \tag{5.1}$$

The first term describes the diffusion of particles between adjacent compartments and the second term is a sum of CME-type terms for the single compartments. The RDME has the advantage over SRDPs that data likelihoods are in principle straightforward to compute given the RDME's solution. The latter is however not known for most systems. Stochastic simulations of the RDME are generally more efficient than for SRDPs, but the computational costs are still significant. In particular, they are gener-

ally significantly more expensive than SSA simulations for corresponding non-spatial systems.

A disadvantage of the RDME is that the discretisation of space is a non-trivial task. Importantly, the discretisation cannot be made arbitrarily small since this leads to the disappearance of bimolecular reactions. This happens because in the RDME framework two particles can only react with each other if they are positioned in the same compartment. However, the probability for the latter to be the case converges to zero in the continuum limit, which in turn leads to bimolecular reaction rates converging to zero.

These reasons taken together show that inference for SRDPs using the RDME is a difficult and computationally expensive task.

5.2 Real-valued Poisson representation

Consider the PR which we introduced in Section 2.3.7 as an exact reformulation of the CME. It translates the system from discrete particle numbers in the CME to continuous variables. This is achieved by making the following ansatz for the solution $P(\mathbf{n},t)$ of the CME:

$$P(\mathbf{n},t) = \int d\mathbf{u}\, \mathcal{P}(n_1;u_1)\ldots\mathcal{P}(n_N;u_N)p(\mathbf{u},t), \quad u_i \in \mathbb{C}, \tag{5.2}$$

where $\mathbf{u} = (u_1,\ldots,u_N)$ and $\mathcal{P}(n_i;u_i) = (e^{-u_i}u_i^{n_i})/n_i!$ is a Poisson distribution in n_i with mean u_i. If all reactions in the system involve at the most two reactant and two product molecules, the distribution $p(\mathbf{u},t)$ satisfies the following FPE,

$$\partial_t p(\mathbf{u},t) = -\sum_{i=1}^{N}\frac{\partial}{\partial u_i}\left[A_i(\mathbf{u})p(\mathbf{u},t)\right] + \frac{1}{2}\sum_{i,j=1}^{N}\frac{\partial}{\partial u_i}\frac{\partial}{\partial u_j}\left[B_{ij}(\mathbf{u})p(\mathbf{u},t)\right]. \tag{5.3}$$

The drift vector $\mathbf{A}(\mathbf{u})$ and diffusion matrix $\mathbf{B}(\mathbf{u})$ are given by

$$A_i(\mathbf{u}) = \sum_{r=1}^{R} S_{ir}g_r(\mathbf{u}), \tag{5.4}$$

$$B_{ij}(\mathbf{u}) = \sum_{r=1}^{R} g_r(\mathbf{u})(r_{ir}r_{jr} - s_{ir}s_{jr} - \delta_{i,j}S_{ir}), \tag{5.5}$$

$$g_r(\mathbf{u}) = \Omega k_r \prod_{j=1}^{N}\Omega^{-s_{jr}}u_j^{s_{jr}}, \tag{5.6}$$

where $\delta_{i,j}$ denotes the Kronecka delta. The Fokker-Planck equation in (5.3) is equivalent to the Langevin equation

$$d\mathbf{u} = \mathbf{A}(\mathbf{u})dt + \mathbf{C}(\mathbf{u})d\mathbf{W}, \quad \mathbf{C}\mathbf{C}^T = \mathbf{B}, \tag{5.7}$$

Table 5.1: Classification of different types of reactions w.r.t. to their Poisson representation. If different types of reactions are present in a system, the PR typically behave like the reaction of highest type.

reaction types				PR
Type	stoichiometry	description	examples	
I	$\sum_i s_{ir} \leq 1$ $\sum_i r_{ir} \leq 1$	zero or one reactant and product molecules	$\varnothing \to A$ $A \to \varnothing$ $A \to B$	real and determ.
II	$\sum_i s_{ir} \leq 1$ $r_{ir} = 2$ for one i and zero otherwise	zero or one reactant; two *identical* product molecules	$\varnothing \to A+A$ $A \to A+A$ $B \to A+A$	real and stoch.
III	$\sum_i s_{ir} \leq 1$ $r_{ir} = r_{jr} = 1$ for two $i \neq j$ and zero otherwise	zero or one reactant; two *non-identical* product molecules	$\varnothing \to A+B$ $A \to A+B$ $A \to B+C$	complex and stoch.
IV	$\sum_i s_{ir} = 2$ $\sum_i r_{ir} \leq 2$	two reactant molecules	$A+A \to \ldots$ $A+B \to \ldots$	complex and stoch.

where $\mathbf{W}$ is an l-dimensional Wiener process and l is the number of columns of $\mathbf{C}$. An important property of the PR which we will use in the following is the fact that the mean values of the u_i variables in PR space agree exactly with the mean values of the n_i variables of the CME.

As we demonstrated in Section 2.3.7, the PR can be deterministic, i.e., $\mathbf{B} = 0$, depending on the reaction system, and its variables can be real- or complex-valued. We will next give a characterisation of different reaction types with respect to their PR behaviour.

5.2.1 Classification of reactions

Table 5.1 shows a classification of different types of elementary reactions in terms of the behaviour of the corresponding PR Langevin equation. These are deduced by inspection of the PR diffusion matrix in Eq. (5.5). We note that this strict classification

only holds if the considered reaction is the only reaction in the system. If there are several reactions happening, the system typically behaves as the entry in Table 5.1 corresponding to the reaction of highest type.

The behaviours of the PR are quite intuitive: for reactions of Type I, it is well-known that fluctuations are Poissonian, which manifests itself in a deterministic PR. Note that if the Poisson representation is stochastic and real-valued, the probability distribution of the molecule numbers in Eq. (5.2) is a real-valued mixture of Poisson distributions, for which it is well known that the resulting distributions are super-Poissonian. For reactions of Type II, for which fluctuations are super-Poissonian, the PR is therefore real and stochastic. It is easy to see that reactions of Type III and IV cannot be represented in this way: a zeroth or first order reaction with two *non-identical* product molecules, i.e., of Type III, imposes a constraint on the particle numbers. For the reaction $\varnothing \to A + B$ for example, the particle numbers of species A and B differ by a constant integer number (determined by the initial condition). Conditioned on the molecule number of A, the distribution of B is a delta function which obviously has sub-Poissonian fluctuations, and can thus not be represented by a real-valued Poisson mixture. Consequently, the PR has to be complex-valued. Bimolecular reactions give rise to similar constraints or otherwise lead to sub-Poissonian fluctuations which means that their PR has to be complex-valued.

5.2.2 Linearisation of Type III and IV reactions

We now introduce approximations of Type III and IV reactions that lead to real-valued PRs. Consider first reactions of Type IV, where two molecules react with each other. We approximate this type of reactions in a *mean-field* type of sense: we replace the direct interaction of the two molecules by two unimolecular reactions whose propensity functions depend on the *mean value* of the respective other species. For instance, the reaction

$$A + B \xrightarrow{k} \varnothing, \quad f(\mathbf{n}) = \frac{k}{\Omega} n_B n_A, \tag{5.8}$$

becomes replaced by the two reactions

$$A \xrightarrow{k\langle n_B\rangle/\Omega} \varnothing, \quad f(\mathbf{n}) = \frac{k}{\Omega}\langle n_B\rangle n_A, \tag{5.9}$$

$$B \xrightarrow{k\langle n_A\rangle/\Omega} \varnothing, \quad f(\mathbf{n}) = \frac{k}{\Omega}\langle n_A\rangle n_B, \tag{5.10}$$

where $\langle n_A\rangle$ and $\langle n_B\rangle$ denote the mean values of the molecule numbers of species A and B, respectively, and Ω is the volume of the system. The reactions (5.9) and (5.10)

Table 5.2: Reactions of Types III and IV and their approximate reactions. The corresponding propensities in PR space for the approximate system are obtained by replacing n_A and n_B with u_A and u_B, respectively.

original		approximation	
reaction	propensity	reaction	propensity
$A+B\to\ldots$	kn_An_B/Ω	$A\to\ldots$ $B\to\ldots$	$k\langle n_B\rangle n_A/\Omega$ $k\langle n_A\rangle n_B/\Omega$
$A+A\to\ldots$	$kn_A(n_A-1)/\Omega$	$A\to\ldots$	$k\langle n_A\rangle n_A/\Omega$
$A\to A+B$	kn_A	$\varnothing\to B$	$k\langle n_A\rangle$
$\varnothing\to B+C$	Ωk	$\varnothing\to B$ $\varnothing\to C$	Ωk Ωk
$A\to B+C$	kn_A	$A\to B$ $A\to C$	kn_A kn_A

correspond to linear reactions with one reactant and zero product molecules. The corresponding PR is therefore real and deterministic according to Table 5.1. Since the mean values of the corresponding PR variables, say u_A and u_B, are equal to the means of n_A and n_B, i.e., $\langle u_A\rangle = \langle n_A\rangle$ and $\langle u_B\rangle = \langle n_B\rangle$, respectively, we simply rescale the rate constants in PR space by $\langle u_A\rangle/\Omega$ and $\langle u_B\rangle/\Omega$, respectively. Specifically, if there are no other reactions happening in the system, the PR Langevin equations read

$$du_A = -\frac{k}{\Omega}\langle u_B\rangle u_A dt, \tag{5.11}$$

$$du_B = -\frac{k}{\Omega}\langle u_A\rangle u_B dt. \tag{5.12}$$

In particular, if there are no other reactions happening in the system, the PR is deterministic and we can write $\langle u_A\rangle = u_A$ and $\langle u_B\rangle = u_B$. Consider next a bimolecular reaction with two identical reactant molecules,

$$A+A \xrightarrow{k} \varnothing, \quad f(\mathbf{n}) = \frac{k}{\Omega}n_A(n_A-1). \tag{5.13}$$

For such type of reactions, we replace the interaction of A with itself by the interaction of A with its mean,

$$A \xrightarrow{k\langle n_A\rangle/\Omega} \varnothing, \quad f(\mathbf{n}) = \frac{k}{\Omega}\langle n_A\rangle n_A. \tag{5.14}$$

In PR space, this leads to the Langevin equation for A,

$$du_A = -\frac{k}{\Omega}\langle u_A\rangle u_A dt. \tag{5.15}$$

Consider next reactions of Type III (c.f. Table 5.1). The reaction

$$A \xrightarrow{k} A + B, \quad f(\mathbf{n}) = kn_A, \tag{5.16}$$

for example, can be approximated in a similar fashion as the bimolecular reactions before: we replace the dependence of the creation of B molecules on A molecules by a dependence on the mean of the later,

$$\varnothing \xrightarrow{k\langle n_A\rangle} B, \quad f(\mathbf{n}) = k\langle n_A\rangle. \tag{5.17}$$

For the other two Type III reactions,

$$\varnothing \xrightarrow{k} B + C, \quad f(\mathbf{n}) = k\Omega, \tag{5.18}$$

$$A \xrightarrow{k} B + C, \quad f(\mathbf{n}) = kn_A, \tag{5.19}$$

we have to decouple the productions of B and C, i.e., which can be achieved by approximating the reactions by

$$\varnothing \xrightarrow{k} B, \quad \varnothing \xrightarrow{k} C, \quad f(\mathbf{n}) = k\Omega, \tag{5.20}$$

$$A \xrightarrow{k} B, \quad A \xrightarrow{k} C, \quad f(\mathbf{n}) = kn_A. \tag{5.21}$$

While the reactions in Eqs. (5.18) and (5.19) correlate the molecule numbers of species B and C, we have effectively decorrelated B and C by introducing the reactions in Eqs. (5.20) and (5.21).

Table 5.2 summarises the approximations for all reactions of Type III and IV. Note that depending on the reaction, a combination of the proposed approximations may have to be performed, for example for the reactions $A + B \to A + C$ and $A + B \to C + D$.

5.3 Real-valued Poisson representation in space

We now aim at applying the real-valued PR to the RDME and subsequently taking the continuum limit. The RDME is given in Eq. (5.1).

We start by considering the first term in Eq. (5.1) which describes diffusion. Since different species do not interact with each other here, we can consider a single species system, say species X_1, for which the first term in Eq. (5.1) reduces to

$$\partial_t P(\mathbf{n},t) = \sum_{l=1}^{L} \sum_{m\in\mathcal{N}(l)} d[(n^m+1)P(\mathbf{n}+\delta^m-\delta^l,t) - n^l P(\mathbf{n},t)], \tag{5.22}$$

where $\mathbf{n} = (n^1, \ldots, n^L)$, n^l is the number of X_1 particles in the l^{th} compartment, δ^m is a vector with a one in the m^{th} entry and zero otherwise, and the sum over m runs over all neighbouring compartments of the l^{th} compartment. These are reactions of Type I (c.f. Table 5.1) and we use the PR without any approximations. The corresponding Langevin equation (which is deterministic in this case) reads

$$du^l = D\frac{2Mu^l - \sum_{m\in\mathcal{N}(l)} u^m}{h^2}dt, \quad l = 1, \ldots, L, \tag{5.23}$$

where M is the spatial dimension of the system and $D = dh^2$ the microscopic diffusion constant. Since the sum over m runs over all adjacent compartments of the l^{th} compartment, the fraction in Eq. (5.23) is just the discretised version of the Laplace operator $\Delta = \partial_1^2 + \ldots + \partial_M^2$. Introducing a discretised density field $u(x^l) = u^l/h^M$, where x^l is the centre of the l^{th} compartment, and taking the continuum limit of (5.23) we get the PDE

$$du(x,t) = D\Delta u(x,t)dt, \tag{5.24}$$

which is just the diffusion equation for the field $u(x,t)$.

Consider next the second term in Eq. (5.1) which describes chemical reactions. Since reactions only occur within compartments, we can treat the compartments independently of each other. For a single compartment, the second term in Eq. (5.1) then reduces to the CME. Here, however, we first apply the approximations discussed in the previous section for reactions of Type III and IV (c.f. Table 5.1). The corresponding propensity functions in PR space are given in the last column of Table 5.2 after replacing the n_i variables by their corresponding u_i variables. This leads to a real-valued PR. Since in the applied approximations, only reactions with two identical product molecules lead to stochastic terms in the PR, the PR Langevin equation simplifies to

$$du_i = \sum_{r=1}^{R} S_{ir} g_r(\mathbf{u})dt + \sum_{r'} \sqrt{2g_{r'}(\mathbf{u})}dW_{r'}, \tag{5.25}$$

where the sum over r' only runs over reactions with two X_i product molecules. This is a direct consequence of the approximation applied to Type III and Type IV reactions. The propensities $g_r(\mathbf{u})$ are obtained by replacing the n_i variables with u_i variables and Ω with h^M in the expressions in the last column of Table 5.2. Reintroducing the label l denoting the compartment number in Eq. (5.25), and the species label i in Eq. (5.24), we can add the contributions of diffusion and chemical reactions to obtain

$$du_i^l = D_i\frac{2Mu_i^l - \sum_{m\in\mathcal{N}(l)} u_i^m}{h^2}dt + \sum_{r=1}^{R} S_{ir} g_r(\mathbf{u}^l)dt + \sum_{r'} \sqrt{2g_{r'}(\mathbf{u}^l)}dW_{r'}^l. \tag{5.26}$$

If we again define the discretised density field $u_i(x^l) = u_i^l/h^M$, where x^l is the centre of the l^{th} compartment, and $dW_r(x^l) = dW_r^l/\sqrt{h^M}$, we can take the continuum limit of Eq. (5.26) leading to

$$\begin{aligned} du_i(x,t) = [D_i \Delta u_i(x,t) + \sum_{r=1}^{R} S_{ir} g_r(\mathbf{u}(x,t))]dt \\ + \sum_{r'} \sqrt{2 g_{r'}(\mathbf{u}(x,t))} dW_{r'}(x,t). \end{aligned} \tag{5.27}$$

Here, $W_{r'}$ are independent spatio-temporal Wiener processes. Note that the sum in the second line only runs over reactions with two X_i product molecules. The $g_r(\mathbf{u}(x,t))$ are the propensity functions in PR space which are obtained by applying the mean-field approximation described in the previous section. They are not functions of single PR variables anymore, but rather functions of the space-dependent intensity field vector $\mathbf{u}(x,t) = (u_1(x,t), \ldots, u_N(x,t))$. The $g_r(\mathbf{u}(x,t))$ are obtained by taking the corresponding propensity function $f_r(\mathbf{n})$ in real space, replacing $n_i \to u_i(x,t)$ and $\langle n_i \rangle \to \langle u_i(x,t) \rangle$ and omitting Ω factors. The latter get absorbed in the definition of the intensity fields defined below Eq. (5.26). For Type III and Type IV reactions, the linearised versions of the reactions and corresponding $f_r(\mathbf{n})$ have to be taken. These are given in Table 5.2. Note that $\langle u_i(x,t) \rangle$ does *not* denote a spatial averaging, but rather an expectation locally in space. The $u_i(x,t)$ are stochastic fields and $\langle u_i(x,t) \rangle$ denotes the expectation locally in space with respect to their measure, which means that $\langle u_i(x,t) \rangle$ is a function of space (and time).

Consider for example the reaction $A + B \to \varnothing$. The non-spatial propensity in real space for this reaction is $f(n_A, n_B) = k n_A n_B/\Omega$. However, since this is a Type IV reaction, we replace it according to Table 5.2 by the two reactions $A \to \varnothing$ and $B \to \varnothing$ with propensities $f(n_A, n_B) = k\langle n_B \rangle n_A/\Omega$ and $f(n_A, n_B) = k\langle n_A \rangle n_B/\Omega$, respectively. By replacing $n_i \to u_i(x,t)$ and $\langle n_i \rangle \to \langle u_i(x,t) \rangle$ and omitting Ω terms, we thus obtain the corresponding propensity functions in spatial PR space as $g(u_A(x,t), u_A(x,t)) = k\langle u_B(x,t) \rangle u_A(x,t)$ and $g(u_A(x,t), u_A(x,t)) = k\langle u_A(x,t) \rangle u_B(x,t)$, respectively. A more detailed example of this procedure can be found at the end of this section.

Equation (5.27) looks similar to the spatial chemical Langevin equation, which can be obtained by applying the chemical Langevin equation to the RDME and subsequently taking the continuum limit [49]. However, Eq. (5.27) has a different interpretation since it describes the intensity in PR space. In particular, just as any other PDE or SPDE description in real space, the spatial chemical Langevin equation does not provide a generative model for the actual location of events, and thus would not

allow us to directly model particle locations statistically. Notice that, as a consequence of the mean-field approximation, the mean value of the $u_i(x,t)$ fields is the same as in a deterministic rate equation description; however, the dynamics of the observed variable, i.e., the points in space, remain stochastic even when the intensity field evolves deterministically.

Note that Gardiner has derived equations similar to Eq. (5.27) in [67] by applying the full PR without any approximations to the RDME and taking the continuum limit. For reactions of Type I and II, the result is equivalent to Eq. (5.27). For reactions of Type III and IV, however, Gardiner's equations are complex-valued and differ from Eq. (5.27). To our knowledge, the relation of SRDPs to spatio-temporal point processes that we will establish in the following using Eq. (5.27) has never been reported in the literature.

Example

As an example, consider the following reaction system

$$X \xrightarrow{k_1} X + X, \quad X + X \xrightarrow{k_2} \varnothing. \tag{5.28}$$

The corresponding stoichiometric matrix reads

$$\mathbf{S} = (1, -2). \tag{5.29}$$

The first reaction in (5.28) is of Type II and thus does not need to be approximated. The corresponding non-spatial propensity function in real space is given by $f_1(n) = k_1 n$, where n is the variable denoting the number of X particles. The second reaction in Eq. (5.28) is of Type IV and hence needs to be approximated. According to Table 5.2 we approximate it by the reaction $X \xrightarrow{k_2\langle n\rangle/\Omega} \varnothing$ with propensity $f_2(n) = k_2\langle n\rangle n/\Omega$. The corresponding propensity functions in spatial PR space are obtained by replacing $n \to u(x,t)$ and $\langle n\rangle \to \langle u(x,t)\rangle$, where $u(x,t)$ is the PR field of species X. We thus have

$$\begin{aligned} X \xrightarrow{k_1} X + X, \quad & f_1(n) = k_1 n, \\ \downarrow & \\ X \xrightarrow{k_1} X + X, \quad & g_1(u(x,t)) = k_1 u(x,t), \end{aligned} \tag{5.30}$$

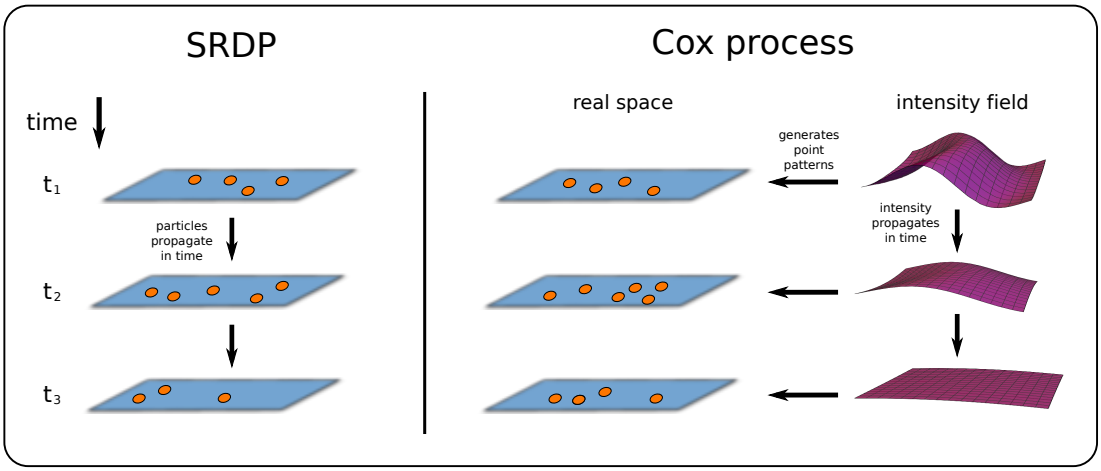

Figure 5.1: Visualisation of Cox process representation of SRDPs. Left: time evolution of the true SRDP in space. Particles diffuse in space, may decay or are created and react with each other. Right: time evolution of a Cox process. Here, the intensity field evolves in time, rather than the points in real space. The latter are merely noisy realisations of the intensity field. In particular, the points patterns at two different time points are independent of each other conditioned on the intensity field.

for the first reaction and

$$\begin{aligned} X+X \xrightarrow{k_2} \varnothing, \quad & f_2(n) = \frac{k_2}{\Omega} n(n-1), \\ \downarrow & \\ X \xrightarrow{k_2} \varnothing, \quad & f_2(n) = \frac{k_2}{\Omega} \langle n \rangle n, \\ \downarrow & \\ X \xrightarrow{k_2} \varnothing, \quad & g_2(u(x,t)) = k_2 \langle u(x,t) \rangle u(x,t), \end{aligned} \tag{5.31}$$

for the second reaction. The corresponding stoichiometric matrix becomes

$$\mathbf{S} = (1, -1). \tag{5.32}$$

Using the general equation in (5.27) we hence obtain the SPDE for the intensity field $u(x,t)$,

$$\begin{aligned} du(x,t) = & \left[D\Delta u(x,t) + k_1 u(x,t) - k_2 \langle u(x,t) \rangle u(x,t) \right] dt \\ & + \sqrt{2k_2 \langle u(x,t) \rangle u(x,t)} dW(x,t). \end{aligned} \tag{5.33}$$

We would like to emphasise once more that $\langle u(x,t) \rangle$ denotes the local expectation of the stochastic intensity field $u(x,t)$ and not a spatial averaging.

5.4 Cox process representation and inference

We will now use the derived results of the previous section to establish a novel relationship between SRDPs and Cox (Poisson) processes. The latter were introduced in Section 2.5. Note that in the classical view of the PR the auxiliary variables u are simply introduced as a mathematical device. Here, however, we will reinterpret the PR variables by considering a *joint* process over the u and the particle number variables. Formally, this is equivalent to what is called *demarginalisation* in statistics: a complex process is replaced by a (simpler) process in an augmented state space, such that the marginals of the augmented process return exactly the initial process.

To this end, we will interpret the PR intensity field $u(x,t)$ as the intensity field of a spatio-temporal Cox (Poisson) process. In this case the intensity field can be thought of as the state variable of the system, with the actual spatial points as noisy realisations of this state (see Figure 5.1 for a graphical visualisation of this concept.). The results of the present section then follow immediately from the derivation of the previous section. Our first observation follows directly from Gardiner's analysis of the continuum limit of the RDME (see next section for a proof)

Remark 5.1. *Consider an SRDP on a spatial domain $\mathcal{D}$ and temporal domain $\mathcal{T}$ with deterministic initial conditions, and let all reactions be of Type I, i.e., involve production or decay of at most one particle. Then, under appropriate initial conditions, the single-time-point spatial probability distribution of the SRDP is* ***exactly the same*** *as of a spatial Poisson process $\forall t \in \mathcal{T}$. The intensity field of the latter fulfils the PDE given in Equation* (5.27).

The corresponding equation in (5.27) if only applied to Type I reactions has been derived by Gardiner before [67]. To our knowledge, however, the connection to spatio-temporal Poisson processes in the statistical sense has not been formulated in the literature to this date.

Next, by using the real-valued PR developed in the previous section, we can generalise Remark 5.1 to general reaction systems, leading to (see next section for a proof)

Result 5.1. *Consider the same setting as in Remark 5.1. Under appropriate initial conditions, if there is at least one reaction with two product particles of the same species, i.e., a reaction of Type II, the system's single-time-point distribution* ***is exactly given*** *by a Cox process, whose intensity fulfils the stochastic PDE (SPDE) given in Equation* (5.27). *If the system involves other types of reactions, i.e., reactions of Type*

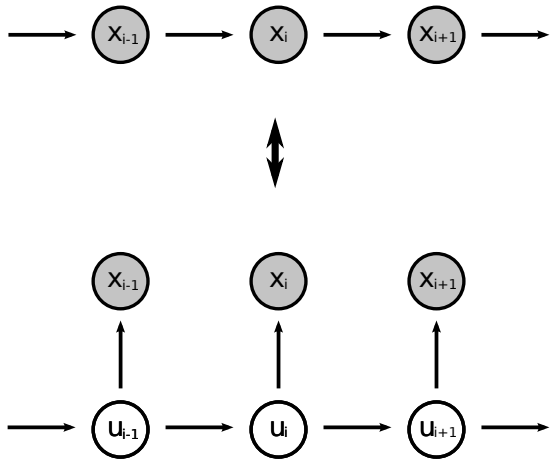

Figure 5.2: Top: Graphical model resembling the true system. The measurements x_i constitute the true state of the system. Bottom: approximation when PR is interpreted as point process. The intensity field is the state of the system, the measured point patterns being noisy realisations thereof. Grey and white circles indicate observed and unobserved variables, respectively.

III or IV, the single-time probability distribution of the SRDP is approximated in a mean-field sense by that of a Poisson (Cox) process whose intensity fulfils Equation (5.27).

Result 5.1 provides an efficient means to calculate statistics such as expected number of agents within a certain volume, without the need to perform extensive Monte Carlo simulations, since it only requires to solve a (S)PDE for which a rich literature of numerical methods exists [78, 102]. Importantly, we can use Result 5.1 to approximate the likelihood function of a configuration of points arising from an SRDP by using the well-known Cox process likelihood, which we introduced in Section 2.5.3. This function can be easily optimised to yield statistical estimates of kinetic parameters from single-time observations.

We consider next the problem of approximating the joint distribution of point patterns arising from an SRDP at different time points. This is important when we have time series observations, i.e., spatial measurements $\mathbf{x} = (\mathbf{x}_0, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i \subset \mathcal{D}$, at discrete time points $t_0, \ldots, t_n$, and we want to compute the likelihood $p(\mathbf{x}|\Theta)$ of the data given

a model Θ. Since we consider Markovian systems, the likelihood can be written as

$$p(\mathbf{x}|\Theta) = p(\mathbf{x}_0|\Theta)\prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{x}_{i-1},\Theta). \tag{5.34}$$

We would like to approximate this likelihood using the relation to Cox processes established in Result 5.1. While this is in principle straightforward, computing the terms $p(\mathbf{x}_i|\mathbf{x}_{i-1},\Theta)$ involves determining the distribution over the associated u_{i-1} variable in PR space. This would involve inverting the PR transformation in Eq. (5.2), which is computationally difficult. Instead, we opt for an approximation strategy: assume that we have determined the PR distribution $p(u_{i-1})$ at time t_{i-1}, where we introduced the shorthand $u_i = u(x,t_i)$. By definition of the intensity of a Poisson process, u_{i-1} represents the expectation of the random configuration of points $\mathbf{x}_{i-1}$ at time t_{i-1}. We then approximate $p(\mathbf{x}_i|\mathbf{x}_{i-1},\Theta)$ in a mean-field way by replacing the explicit dependence on $\mathbf{x}_{i-1}$ with its expectation: $p(\mathbf{x}_i|\mathbf{x}_{i-1},\Theta) \approx \langle p(\mathbf{x}_i|\mathbf{x}_{i-1},\Theta)\rangle_{p(\mathbf{x}_{i-1}|u_{i-1})} = p(\mathbf{x}_i|u_{i-1},\Theta)$. Figure 5.1 visualises this approximation. The left panels shows the time evolution in an SRDP, while the right panel shows the time evolution of a corresponding approximating Cox process. This leads to a new interpretation of the measured points $\mathbf{x} = (\mathbf{x}_0,\ldots,\mathbf{x}_n)$: while they are snapshots of the actual state in the true system, they correspond to noisy realisations of the state $u(x,t)$ in the Cox process picture. This corresponds to replacing the graphical model in the upper panel in Figure 5.2 by the graphical model in the lower panel.

We thus have:

Result 5.2. *The joint n-time-point marginal distribution of an SRDP can be approximated in a mean-field sense by the joint distribution of a Poisson (Cox) process with intensity governed by the (S)PDE in Equation* (5.27)*.*

Result 5.2 is particularly powerful statistically, because it enables us to analytically approximate the exact (intractable) likelihood $p(\mathbf{x}|\Theta)$ in Equation (5.34) by the likelihood of a spatio-temporal Cox process with intensity $u(x,t)$. This allows us to develop efficient algorithms for approximate maximum likelihood estimation in general SRDPs; Section 5.5 is dedicated to illustrating the performance of this method on a range of case studies. Let us first however prove Remark 5.1 and Result 5.1 and briefly review the used inference procedure for Poisson and Cox processes.

5.4.1 Proof of Remark 5.1 and Result 5.1

For the proof of Remark 5.1 and Result 5.1 we only need to consider the behaviour of probability distributions when taking the continuum limit in Section 5.3. For simplicity, we consider a one-dimensional system with one species X in the interval $[0,1]$. Let us go one step back and consider the PR of the discretised version in Eq. (5.26) before taking the continuum limit. Consider first a system involving only reactions of Type I. In that case we do not have to perform any approximations to arrive at (5.26) and the second sum including the noise terms vanishes, i.e., (5.26) reduces to a PDE. For deterministic initial conditions the u_i thus remain deterministic, and the probability distribution of n^l in the l^{th} compartment at time t is given by a Poisson distribution with mean value u^l. The mean number of molecules in an interval $I = [(m_1 - \frac{1}{2})h, (m_2 + \frac{1}{2})h], m_1 < m_2 \in \mathbb{N}$ at time t is hence

$$\langle N(I,t)\rangle = \sum_{i=m_1}^{m_2} \langle n^i\rangle = \sum_{i=m_1}^{m_2} \langle u^i\rangle = \sum_{i=m_1}^{m_2} u^i, \tag{5.35}$$

where $N(I,t) = \sum_{i=m_1}^{m_2} n^i$. Since the n_i are independent Poisson random variables, $N(I,t)$ is also a Poisson random variable with mean $\langle N(I,t)\rangle = \sum_{i=m_1}^{m_2} u^i(t)$.

Defining $u^i/h \to u(x^i)$, where x^l is the centre of the l^{th} compartment, allows us to take the continuum limit $h \to 0$ of Eq. (5.26) at constant $I = [(m_1 - \frac{1}{2})h, (m_2 + \frac{1}{2})h]$ which gives the PDE in Eq. (5.27). The mean value of $N(I,t)$ can be written as $\langle N(I,t)\rangle = \sum_{i=m_1}^{m_2} hu(x^i,t)$, which is a Riemann sum. Taking the limit $h \to 0$ for constant I thus gives

$$\langle N(I,t)\rangle \to \int_I dx\, u(x,t). \tag{5.36}$$

According to the "Countable additivity theorem" [103], the sum of an infinite number of Poisson distributed independent random variables converges with probability 1 if the sum of the mean values converges, and the sum has a Poisson distribution with corresponding mean value. We assume that the mean particle density is bound everywhere, which means that the values $u^i/h = u(x^i)$ are bound for all i and all h. Let B be such an upper bound. Since

$$\left|\sum_{i=m_1}^{m_2} hu(x^i)\right| \leq h \sum_{i=m_1}^{m_2} \left|u(x^i)\right| \leq h \sum_{i=m_1}^{m_2} B = (m_2 - m_1)B, \tag{5.37}$$

the sum converges in the limit $h \to 0$. We thus find that $N(I,t)$ is Poisson distributed in the continuum limit with mean value $\int_I dx\, u(x,t)$, and we can write

$$P(N(I,t) = n) \xrightarrow{h\to 0} \mathcal{P}(n; \int_I dx\, u(x,t)). \tag{5.38}$$

The same can be shown similarly for a countable union of subintervals of $[0,1]$, and $N(U_1,t)$ and $N(U_2,t)$ are obviously independent for disjunct U_1 and U_2. The probability distribution for any fixed t is thus *exactly the same as the one of a spatial Poisson process* with intensity $u(x,t)$.

Suppose now that the system also includes reactions of Type II. In this case the PR becomes stochastic, i.e., Eq. (5.26) and its continuum version in Eq. (5.27) contain non-vanishing noise terms, which means that the field $u(x,t)$ is a random process. Given a realisation of $u(x,t)$, the same considerations from before for the deterministic case apply and the single-time distribution behaves like a spatio-temporal Poisson process. Since $u(x,t)$ is now a random process, the single-time distribution of the system *corresponds exactly* to the one of a *Cox-process* with intensity $u(x,t)$. The same considerations hold in an approximate sense for Type III and IV reactions. These findings can easily be generalised to multiple-species systems and general spatial dimensions and domains. This concludes the derivation of Remark 5.1 and Result 5.1 of the previous section.

5.4.2 Computing the likelihood for Poisson and Cox processes

We briefly review the computation of likelihoods for Poisson and Cox processes and the numerical solution of SPDEs via basis projection which we presented in Sections 2.5.3 and 2.5.2, respectively. For simplicity we consider a one-species system here. The methods can however easily be generalised to multi-species systems.

Numerical solution of (S)PDEs via basis projection

Consider first the numerical solution of an (S)PDE for the intensity field as in Eq. (5.27). As explained in Section 2.5.2, we make the ansatz $u(x,t) = \sum_{i=1}^{n} c_i(t)\phi_i(x)$ of writing the intensity field $u(x,t)$ as a linear combination of a finite set of n spatial basis functions $\phi_i(x)$ and project the (S)PDE onto these basis functions. This leads to (S)ODEs for the coefficient vector $\mathbf{c}(t) = (c_1(t),\ldots,c_n(t))$. In the case of a Poisson process, i.e., vanishing noise terms in Eq. (5.27), $\mathbf{c}(t)$ fulfils a set of ODEs which can be integrated numerically. In the case of a Cox process, $\mathbf{c}(t)$ fulfils a set of SDEs. In the case of linear SDEs, i.e, drift terms depending linearly on $\mathbf{c}(t)$ and noise terms being independent of $\mathbf{c}(t)$, the system of SDEs is solved by a multivariate normal distribution whose mean and covariance can be obtained by direct numerical integration. If the SDEs are non-linear, the solution is no longer a multivariate normal distribution. However, it

can easily be seen that, due to the approximations that we apply to Type III and Type IV reactions, the time evolution equations for the mean and covariance of $\mathbf{c}(t)$ decouple from higher order moments, and can thus be obtained by numerical integration. We then simply approximate the solution of the set of SDEs by a multivariate normal distribution with corresponding mean and covariance. For details on the solution of (S)PDEs via basis projection see Section 2.5.2.

Filtering

Assume now that we have a (Cox) Poisson process whose intensity fulfils the (S)PDE in Eq. (5.27) and that we can solve this (S)PDE via basis projection as described above. The state of the system at time t is thus represented by the vector $\mathbf{c}(t)$ of basis projection coefficients. Assume further that we have spatial measurements $\mathbf{x} = (\mathbf{x}_0, \ldots, \mathbf{x}_n)$ at discrete times $t_0, \ldots, t_n$. For a Poisson process with intensity field $u(x,t)$, the likelihood of one measurement $\mathbf{x}_i$ at time t_i is given by

$$p(\mathbf{x}_i|\Theta) = \prod_{s\in\mathbf{x}_i} u(s,t_i) e^{-\int dx u(x,t_i)}, \tag{5.39}$$

where Θ represents the model. In the case of a Cox process, for which $u(x,t_i)$ is stochastic, Eq. (5.39) has to be averaged accordingly with respect to the distribution of $u(x,t_i)$.

To perform inference we need to compute the likelihood $p(\mathbf{x}|\Theta)$ of the data given our model Θ. We omit in the following the conditioning on Θ for notational clarity. The likelihood factorises as

$$p(\mathbf{x}) = p(\mathbf{x}_0) \prod_{i=1}^{n} p(\mathbf{x}_i|\mathbf{x}_{i-1}, \ldots, \mathbf{x}_0). \tag{5.40}$$

This can be solved iteratively as follows. Assume that we know the *posterior* $p(\mathbf{c}_{i-1}|\mathbf{x}_{i-1}, \ldots \mathbf{x}_0)$ at time step $i-1$ and that we can propagate it forward in time via basis projection to obtain the predictive distribution $p(\mathbf{c}_i|\mathbf{x}_{i-1}, \ldots \mathbf{x}_0)$. To obtain the posterior at time step i, we have to perform the measurement update

$$p(\mathbf{c}_i|\mathbf{x}_i, \ldots \mathbf{x}_0) = \frac{p(\mathbf{x}_i|\mathbf{c}_i) p(\mathbf{c}_i|\mathbf{x}_{i-1}, \ldots \mathbf{x}_0)}{p(\mathbf{x}_i|\mathbf{x}_{i-1}, \ldots \mathbf{x}_0)}. \tag{5.41}$$

We find that the normalisation gives the contribution to the full likelihood in Eq. (5.40) at time step i. We can thus in principle compute the full likelihood in Eq. (5.40) iteratively.

In the case of a Poisson process, $\mathbf{c}(t)$ is deterministic, which means that the measurement updates in Eq. (5.41) become trivial and effectively no measurement updates have to be performed. Rather, one only needs to solve the PDE forward in time over the whole measurement time interval, plug the measurements into the likelihood in Eq. (5.39) for single time points, and subsequently take the product over the time points.

In the case of a Cox process, the measurement updates in Eq. (5.41) have to be performed. Generally, these cannot be computed analytically. We approximate them here by means of the *Laplace approximation* [79], which approximates the posterior by a Gaussian centred at the posterior's maximum and with covariance being the negative Hessian of the posterior in the maximum. For more details on the computation of the likelihood for Poisson and Cox processes we refer the reader to Section 2.5.3.

We perform parameter inference in this work by maximising the likelihood with respect to the parameters. This is probably the most straightforward way for parameter inference. We would like to point out however, that the possibility of (approximately) computing the likelihood in principle allows the development of more advanced inference methods such as Bayesian inference [79].

5.5 Examples

In the following we apply Result 5.1 and Result 5.2 to perform parameter inference for several systems. We solve the corresponding (S)PDEs for the different systems via basis projection as explained in Section 5.4.2 using locally constant, non-overlapping basis functions defined in Section 2.5.2. For details see Appendix C.

5.5.1 Gene expression

To demonstrate the accuracy of our method, we first consider simulated time-series data in this section. To this end, consider a gene expression system in a one-dimensional container as depicted in Figure 5.3. A gene located in the nucleus is transcribed into mRNA molecules. The latter decay and diffuse across the whole cell and are translated into proteins in the cytosol. The protein molecules also diffuse across the whole cell and decay. For simplicity, we do not model the gene explicitly but assume that mRNA becomes transcribed with a certain fixed rate m_1 homogeneously in the nucleus. The

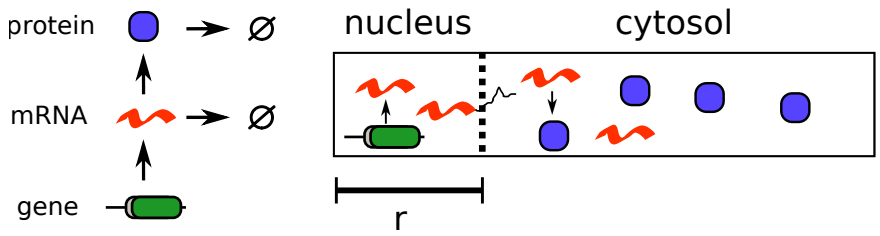

Figure 5.3: Gene expression system. Left: chemical reactions taking place. Right: system in space. The mRNA becomes transcribed in the nucleus, and becomes translated to proteins in the cytosol. mRNA and protein molecules decay stochastically and undergo Brownian diffusion across the whole cell.

Table 5.3: Inference for gene expression system in Figure 5.3 and reactions in Eqs. (5.42) and (5.43). We assume measurements of the protein while the mRNA is unobserved. The inference is carried out by maximising the likelihood of simulated data for thirty measurement points separated by $\Delta t = 0.5$. This procedure is carried out for hundred simulated data sets, and the mean value and standard deviation (in parenthesis) of the inferred results are displayed.

	r	d_m	d_p	m_1	m_2	p_1	p_2
exact	0.3	0.1	0.1	20	0.5	20	0.2
inferred	0.31 (0.06)	0.12 (0.08)	0.14 (0.06)	23 (12)	0.51 (0.4)	26 (18)	0.25 (0.1)

corresponding reactions are

$$\varnothing \xrightarrow{m_1} M, \quad M \xrightarrow{m_2} \varnothing, \tag{5.42}$$

$$M \xrightarrow{p_1} M + P, \quad P \xrightarrow{p_2} \varnothing, \tag{5.43}$$

where M and P denote the mRNA and protein, respectively. For this system, the SPDE of our method in Eq. (5.27) becomes deterministic and thus corresponds to a Poisson process, see Appendix C.1 for details.

In addition to the reaction parameters m_1, m_2, p_1 and p_2, we have to infer the nucleus size r, as well as the diffusion rates d_m and d_p of the mRNA and protein, respectively, summing up to a total number of seven parameters. We assume that the positions of the protein molecules are observed at thirty time points, while the mRNA is unobserved. The inference results for one parameter set are shown in Table 5.3. Considering that we observe the protein at only thirty time points with unobserved

Table 5.4: Inference for the same system as in Table 5.3 but with the additional autocatalytic reaction in Equation (5.44). Since only the difference $p_2 - p_3$ is identifiable we fix $p_3 = 0.01$ and infer the other seven parameters. The table shows the average and standard deviations (in parenthesis) of the inference results for one hundred simulated data sets.

	r	d_m	d_p	m_1	m_2	p_1	p_2
exact	0.3	0.1	0.1	20	0.5	20	0.2
inferred	0.30 (0.05)	0.14 (0.08)	0.088 (0.03)	27 (17)	0.57 (0.3)	24 (21)	0.19 (0.08)

mRNA and that we have seven unknown parameters, the inferred average values are remarkably close to the exact values. Moreover, the standard deviations of the inferred results for single data sets are small, demonstrating the accuracy and precision of our method.

Next, we extend the system in Figure 5.3 by adding an autocatalytic reaction for the protein,

$$P \xrightarrow{p_3} P + P. \tag{5.44}$$

Including this reaction leads to a non-vanishing noise term in Equation (5.27) and the system corresponds to a Cox process. We note that the system has a steady state only if $p_3 < p_2$, with an otherwise exponentially growing mean protein number. On the mean level only the difference $p_2 - p_3$ is identifiable, and we fix $p_3 = 0.01$. We thus infer the same parameters as in the previous case, but this time modelled as a Cox process. Table 5.4 shows the results indicating the accuracy of our method. See Appendix C.1 for details on the used inference method and equations.

5.5.2 SIRS model

We next consider the SIRS system, a popular model for describing the dynamics of an infection spreading through a population. The model contains a susceptible species (S), which can be infected by interacting with an infected (I) species. The latter stochastically recover by making a transition to a recovered species (R), which in turn can become susceptible again. Such systems are frequently modelled as SRDPs [76] or discretised versions thereof [104]. We consider a system in the two-dimensional square

Table 5.5: Inference for the SIRS model with reactions given in Equation (5.45). The inference is carried out by maximising the likelihood of simulated data for forty measurement points. This procedure is carried out for two hundred simulated data sets, and the mean value and standard deviation (in parenthesis) of the inferred results are displayed.

	$10^3 \times d$	$10 \times r$	$10 \times s$	$10^3 \times k^{\mathrm{PR}}$	S_{ini}
exact	1	0.2	2	-	200
inferred	0.8 (0.3)	0.19 (0.09)	1.8 (1.2)	1.5 (0.5)	
exact	1	0.2	2	-	300
inferred	0.9 (0.4)	0.15 (0.06)	1.4 (0.9)	2.4 (0.5)	
exact	1	2	2	-	200
inferred	1.0 (0.6)	1.6 (0.7)	1.5 (1.0)	3.4 (1.1)	
exact	1	0.2	2	-	200
inferred	0.8 (0.4)	0.21 (0.11)	2.2 (1.6)	2.4 (0.5)	
exact	2	0.2	2	-	100
inferred	1.6 (0.8)	0.19 (0.09)	1.9 (1.2)	4.6 (1.1)	

$[0,1] \times [0,1]$. The particles perform Brownian diffusion and interact via the reactions

$$S + I \xrightarrow{k,w} 2I, \quad I \xrightarrow{r} R, \quad R \xrightarrow{s} S, \tag{5.45}$$

where the bimolecular infection reaction is characterised by the microscopic reaction rate k and the reaction range w. We assume that all three species diffuse with the same diffusion rate d. We assume further that initially there are no recovered (R) particles, S_{ini} susceptible (S) particles placed uniformly over the whole area, and one infected (I) particle located at $[0.05, 0.05]$. We consider a fully observed system and perform inference for forty simulated data points using Result 5.2, thereby replacing k and w by an effective bimolecular reaction parameter k^{PR}. The model thus has four parameters that need to be inferred: the diffusion rate d, the recovery rate r, the susceptible rate s and the bimolecular infection rate k^{PR}. Table 5.5 shows the corresponding results, demonstrating the accuracy and precision of our method. The computational efficiency of our method in comparison to stochastic simulations is particularly pronounced here. For the first parameter set in Table 5.5, for example, the Brownian dynamics simulation of *a single realisation* of the system takes about 250 seconds, while *the whole inference procedure* for the four parameters using our method takes only about ten seconds for one simulated data set on a 3.1GHz core. See Appendix C.2 for details on the used inference method.

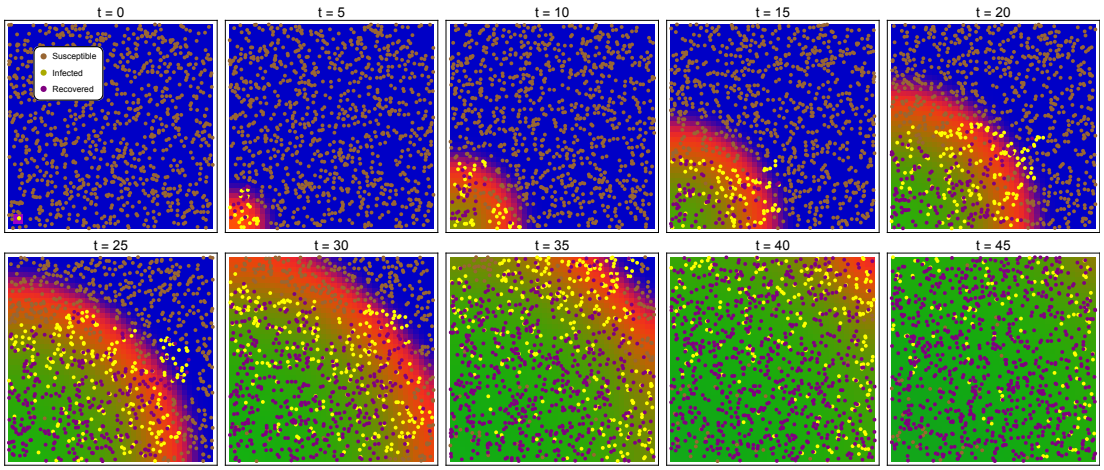

Figure 5.4: Visualisation of a single simulated realisation (points) and prediction of our method (background colours) for the SIRS system with reactions in Equation (5.45). The figures show the time evolution from $t = 0$ to $t = 45$ with steps of $\Delta t = 5$. For the simulation we use the parameters $(S_{\text{ini}}, k, w, d, r, s) = (10^3, 10^4, 0.02, 0.0002, 0.3, 0.01)$ and for the point process prediction the corresponding inferred parameters. The background is an RGB image with the three colour components being proportional to the intensity fields of the three species S (blue), I (red) and R (green). Notice how the mean-field approximation still captures the complex behaviour of a wave of infection spreading through the domain from the bottom left corner.

Figure 5.4 visualises the dynamics of the SIRS system for one parameter set. Individual points from a simulation are shown in different colours (brown for S, yellow for I and purple for R), while the background RGB colours represent the intensity of the respective PR fields with optimised parameters (blue for S, red for I and green for R). Notice how the PR approximation is able to capture the emerging behaviour of a wave of infection sweeping through the domain from bottom left to top right, before the establishment of a dynamic equilibrium between the three population. Such a phenomenon is clearly due to the spatial aspect of the system, and could not have been recovered using an inference method that did not incorporate spatial information. Indeed, the overall number of infected individuals rises rapidly and remains essentially constant between time 20 and time 35 before dropping to steady state, a behaviour which is simply not possible in a non-spatial SIRS model.

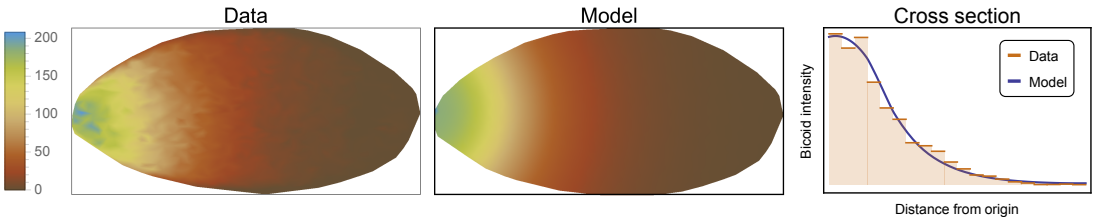

Figure 5.5: Visualisation of results for the Drosophila embryo Bicoid data. Left and middle: Bicoid density across a single embryo for measurement data and point process intensity, respectively. Right: corresponding density along the major axis for one embryo, from experimental data (red histogram) and point process intensity (blue line). In both cases the point process prediction agrees well with the experimental data. The point process prediction is obtained by solving Equation (5.27) numerically for the inferred parameter values.

5.5.3 Drosophila embryo

Finally, we apply our method to real gene expression data for the Bicoid protein at cleavage stage 13 in the Drosophila embryo. The data for seventeen embryos can be obtained from the FlyEx database [105]. The data consists of fluorescence intensity measurements on a spatial grid and is shown for one embryo in the left panel of Figure 5.5. The protein becomes expressed in some region at the left end of the embryo and then diffuses across the embryo and decays. The system is typically modelled by a linear birth-death process [36, 37], and we assume the protein to be expressed within a certain distance r from the left end of the Embryo (see Appendix C.3 for details). At cleavage stage 13 the system is supposed to be in steady state and we can perform inference using Result 5.1 and Equation (5.27). For simplicity, we project the data to one dimension (see Appendix C.3 for details).

The system has four parameters: the creation range r, the diffusion rate d, the production rate c_1 and decay rate c_2 of the Bicoid protein. For steady-state data not all parameters but only certain ratios are identifiable. We thus infer the creation range r, the diffusion rate d and the ratio $c = c_1/c_2$. For the average of the inferred parameters and their standard deviations (shown here in parentheses) across the ensemble of embryos we obtain

$$r = 0.26(0.09), \quad d = 0.023(0.005), \quad c = 1.3(0.2) \times 10^4, \tag{5.46}$$

with standard deviations of about 20% to 30%. We find that these results do not change significantly under variations of the initial parameter values used in the likelihood optimiser.

Figure 5.5 illustrates the inference result for one embryo. The left figure shows the Bicoid density along the major axis of the embryo and the middle and right figures show the corresponding densities across the whole embryo for experimental data and the PR prediction, respectively. In both cases we observe good agreement between the measurement data and the point process approximation. See Appendix C.3 for details on the used inference method.

5.6 Conclusion

In this chapter we studied two popular classes of models for studying stochasticity in spatio-temporal systems; stochastic reaction-diffusion processes (SRDPs) and spatio-temporal point processes. The two classes of models are both commonly used in many disciplines such as epidemiology [21, 77] and social sciences [75], however they are widely perceived as conceptually distinct. SRDPs are microscopic mechanistic descriptions used to predict the dynamics of spatially interacting particles, whereas point processes are typically used empirically to perform inference tasks for systems for which no microscopic description exists. The two approaches therefore seem to be orthogonal to each other.

However, we have shown that the two methods are intimately related. By using the Poisson representation (PR) we established a Cox process representation of SRDPs, which is exact for certain classes of systems and approximate for others. This novel representation enables us to apply statistical inference methods to SRDPs, which has not been possible so far. We applied the developed method to several example systems from systems biology and epidemiology and obtained remarkably accurate results.

It is straightforward to see that our method agrees with a deterministic rate equation description on the mean level. Bimolecular reactions may hence lead to deviations from the true mean, which is known to be the case in some non-spatial scenarios [64]. Moreover, since in our method distributions are given as real Poisson mixtures, sub-Poissonian fluctuations cannot be captured. However, Gardiner showed that fluctuations of SRDPs are dominated by diffusion on small length scales and therefore Poissonian [49], which may explain the accuracy of our method.

Our method approximates certain reaction types. We sometimes call this procedure

"linearisation of reactions". We would like to point out, however, that this approximation introduces modified propensity functions that depend on the mean fields of certain species. The resulting reactions are thus not truly linear in the elementary reaction sense. Importantly, the SPDEs governing the time evolution of the intensity fields may depend non-linearly on the intensity fields and are thus capable of capturing non-linear behaviour.

For the inference from multi-time-point data, the derived method approximates the dependence between different measurement time points by a conditional dependence on the mean fields. Conditioned on these mean fields the measurements at different time points are then independent. We thus expect our method to become less accurate whenever the positions of particles between two consecutive measurements are highly correlated. It is non-trivial to quantify this statement, but one can expect the correlations to be high whenever the time interval between measurements is small in comparison to the time scale of the reactions and, most importantly, small in comparison to the time scale of spatial diffusion. However, since measurements with very short time intervals contain only little information about the system, it is difficult to test our method against this criterion.

One disadvantage of our method that we would like to point out is that it cannot infer kinetic parameters of bimolecular reactions since such non-local interactions get replaced by local interactions with effective parameters. We would like to stress, however, that the same is true for any inference method that relies on the discretisation of space, i.e., on the RDME. Modifications of our method that treat bimolecular reactions more accurately are subject to future research.

Most inference methods in the literature for SRDPs are either based on Brownian dynamics simulations or stochastic simulations of spatially discretised systems using the RDME. Both approaches are computationally extremely expensive and quickly become unfeasible for larger systems and in particular for inference purposes. In contrast, our method relies on the solution of (S)PDEs for which a rich literature of efficient numerical methods exists. For the studied example systems our method turned out to be highly efficient, in particular for systems that correspond to Poisson processes within our framework: the computational time for inferring seven unknown parameters in the gene expression system without the autocatalytic reaction, for example, was found to be about 500 seconds on a 3.1GHz processor. We therefore expect our method to be applicable to significantly larger systems containing more species and unknown parameters. Remarkably, for the SIRS system, for which four unknown parameters had

to be inferred, simulating *a single realisation* of the process from Brownian dynamics simulations took an order of magnitude more (about 250 seconds) than *the whole inference procedure* using our method (about 10 seconds), i.e., optimising the likelihood with respect to the parameters. This indicates the immense computational costs of inference methods based on Brownian dynamics simulations.

We therefore believe that the derived representation of SRDPs in terms of Cox processes has the potential to serve as the basis for a new class of statistical inference methods for SRDPs.

Chapter 6

Conclusion and Outlook

Stochastic effects play a crucial role in many chemical reaction networks in living cells. With a rapidly increasing amount of data becoming available for such systems, the development of mathematical models to describe these effects and to infer information from data constitutes an outstanding research problem. Significant effort has been spent in recent years on the development of such models. Generally, the mathematical description of the stochastic chemical process is intractable, and methods typically approximate it by a tractable model.

In particular in the case of non-spatial descriptions of chemical networks, a large number of different approximation methods have been proposed in the literature. Many of these approximations have no systematic justification, and little is known about their applicability and mathematical limitations. We studied two such classes of approximations in this work; the chemical Langevin equation (CLE) and a certain class of moment closure approximation (MA) methods. We have elucidated fundamental mathematical problems for both methods, which can considerably limit their applicability. We found that wrong usage of these methods may lead to substantially inaccurate results, as well as to qualitatively wrong or unphysical conclusions. We believe that our results significantly advance the understanding of the studied methods.

In comparison to the non-spatial case, the description of spatial processes, i.e., stochastic reaction-diffusion processes (SRDPs), is much less developed. In particular for the task of inference from experimental data, there are barely any methods available. We developed a novel inference method in this thesis that was found to be highly efficient and accurate. We think that the proposed method constitutes an important advancement to the state of the art of inference methods for SRDPs and that it may serve as a basis for future research in this field.

Future perspectives

In the context of the CLE's mathematical limitations, we proposed a novel version of the CLE: the complex CLE which extends the state space of the CLE from real-valued to complex-valued variables. We showed that the complex CLE does not suffer from the problems of the conventional CLE and that it is significantly more accurate than other modifications of the CLE proposed in the literature.

However, the complex CLE has several drawbacks: first of all, it is not clear how to obtain a probability distribution over real variables from its simulations. The exploration of this problem is subject of future research. Another disadvantage of the complex CLE is that it is computationally more expensive than the real-valued CLE. It would thus be useful to derive criteria that validate usage of the real-valued CLE. We would like to mention, however, that for systems for which the real-valued CLE does not break down or whenever the probability of the real-valued CLE to break down in a specified finite time interval is sufficiently small, the complex CLE effectively reduces to the real-valued CLE, and the complex CLE's drawbacks disappear. In cases of doubt, the usage of the complex CLE hence seems favourable.

The analysis of the validity of MAs performed in this work was a purely numerical case study. The question hence arises to what extent our results can be generalised to general reaction networks. Moreover, since we have mainly focused on second-order MAs here, it would be interesting to see how our results generalise to higher order MAs. A more extensive numerical study could shed some light on this issue. Optimally, one would want to find analytic proofs for the properties of MAs that we found numerically in this work for certain example systems. Due to the complicated non-linear form of MAs, however, general analytical proofs seem hardly accessible. Yet, for some properties this might be possible, in particular regarding the unphysical sustained oscillations for large system sizes of certain systems.

The developed inference method for SRDPs incorporates two approximations. The first approximation concerns bimolecular reactions and certain linear reactions. One possibility to improve the proposed method would thus rest upon improving this approximation or to abandon it partially. The latter option would mean to dispense with the representation of SRDPs in terms of spatio-temporal Cox processes, at the cost of more expensive computations. The second approximation concerns the partial decorrelation of measurements at different time points. This could potentially also be improved by extending the state space of our method to complex-valued variables. How-

ever, once again, this would mean to abandon the Cox process picture and would make computations more involved.

Another possibility to improve our method concerns the actual inference procedure. We performed parameter inference in the most straightforward way by maximising the data likelihood. An interesting next step would be to incorporate our method into more elaborate inference methods, such as Bayesian methods.

One drawback of our method is that it cannot be applied directly without further modifications to systems with conservation laws in particle numbers. A gene that switches between an *on* and an *off* state and that transcribes mRNA in the *on* state, for example, would be difficult to model with our method. One possibility to circumvent this problem would be to incorporate our method into a hybrid model, where the gene is modelled explicitly in real space.

We hope that the methods and results developed in this thesis will stimulate corresponding research.

Appendix A

Breakdown of the chemical Langevin equation

A.1 Breakdown analysis for a two enzyme catalysed reaction

Consider the system

$$\begin{aligned} E_A + A &\xrightarrow{c_1} E_A + B, \quad \varnothing \underset{c_4}{\overset{c_3}{\rightleftharpoons}} E_A, \\ E_B + B &\xrightarrow{c_2} E_B + A, \quad \varnothing \underset{c_6}{\overset{c_5}{\rightleftharpoons}} E_B. \end{aligned} \tag{A.1}$$

The enzymes are denoted by E_A and E_B and the substrates by A and B. We do not model the intermediate states of the catalyse process but rather assume they are fast enough such that they can be replaced by one effective reaction. Let N_0, x, x_A and x_B be the total number of substrate molecules (those of A and B), and the number of molecules of A, E_A and E_B, respectively. It can be easily seen that N_0 is constant, which means that the system can be reduced to three variables by using that the number of B molecules is equal to $N_0 - x$.

We rescale time as $\tau = c_6 t$ and define $k_1 = c_1/(\Omega c_6), k_2 = c_2/(\Omega c_6), k_3 = \Omega c_3/c_6, k_4 = c_4/c_6$ and $k_5 = \Omega c_5/c_6$. The CLE in standard form then reads

$$dx = (-k_1 x_A x + k_2(N_0 - x)x_B)d\tau - \sqrt{k_1 x_A x}\, dW_1 + \sqrt{k_2(N_0 - x)x_B}\, dW_2, \tag{A.2}$$

$$dx_A = (k_3 - k_4 x_A)d\tau + \sqrt{k_3}\, dW_3 - \sqrt{k_4 x_A}\, dW_4, \tag{A.3}$$

$$dx_B = (k_5 - x_B)d\tau + \sqrt{k_5}\, dW_5 - \sqrt{x_B}\, dW_6. \tag{A.4}$$

It can be easily seen that this CLE breaks down, since some of the noise terms are non-zero independently of the values of the variables.

An alternative form of the CLE is given by

$$dx = (-k_1 x_A x + k_2(N_0 - x)x_B)d\tau + \sqrt{\lambda_1}\, dW_1, \tag{A.5}$$

$$dx_A = (k_3 - k_4 x_A)d\tau + \sqrt{\lambda_2}\, dW_2, \tag{A.6}$$

$$dx_B = (k_5 - x_B)d\tau + \sqrt{\lambda_3}\, dW_3, \tag{A.7}$$

where

$$\lambda_1 = (k_1 x_A - k_2 x_B)x + N_0 k_2 x_B, \tag{A.8}$$

$$\lambda_2 = k_3 + k_4 x_A, \tag{A.9}$$

$$\lambda_3 = k_5 + x_B. \tag{A.10}$$

Using Ito's formula [49] and the above CLEs for x_A and x_B, we can derive the CLEs for λ_2 and λ_3 leading to

$$d\lambda_2 = k_4(2k_3 - \lambda_2)d\tau + k_4\sqrt{\lambda_2}\, dW_2, \tag{A.11}$$

$$d\lambda_3 = (2k_5 - \lambda_3)d\tau + \sqrt{\lambda_3}\, dW_3. \tag{A.12}$$

Thus the CLEs in the variables λ_2 and λ_3 do not break down because when the noise terms equal zero (for λ_2 and λ_3 equal zero respectively), the drift terms become positive which leads to the eventual increase of the variables. This in fact could be deduced from our previous results as follows. The enzymes E_A and E_B are not influenced by the reactions involving A and B. They simply undergo the simple birth and death process that has been investigated earlier (see Section 3.1.1) and whose alternative form CLE has been shown to not suffer from breakdown.

Similarly, we can deduce the CLE for variable λ_1 using the CLE for variable x above. Under the constraint $k_1 x_A - k_2 x_B = \frac{k_1}{k_4}(\lambda_2 - k_3) - k_2(\lambda_3 - k_5) \neq 0$, the new CLE reads

$$d\lambda_1 = f(\lambda_1, \lambda_2, \lambda_3)d\tau + \sqrt{\lambda_1}\left(\frac{k_1(\lambda_2 - k_3)}{k_4} + k_2(k_5 - \lambda_3)\right)dW_1 + \sqrt{\lambda_2}k_1 \times$$
$$\frac{k_4(k_2 N_0(k_5 - \lambda_3) + \lambda_1)}{k_1(\lambda_2 - k_3) + k_2 k_4(k_5 - \lambda_3)}dW_2 - \sqrt{\lambda_3}k_2\frac{k_4\lambda_1 + k_1 N_0(k_3 - \lambda_2)}{k_1(\lambda_2 - k_3) + k_2 k_4(k_5 - \lambda_3)}dW_3, \tag{A.13}$$

where f is a complicated function of the variables λ_1, λ_2, and λ_3 and whose particular form is not important to our analysis. We find that as λ_1 becomes zero, the first noise term vanishes, however the other two noise terms are generally non-zero which implies that noise can drive λ_1 to negative values and hence the CLE breaks down.

Thus, as for the two variable example in Section 3.1.1, the alternative form of the CLE does not circumvent the problems of the standard form of the CLE. Also similar to the results there, the breakdown is intimately related to the properties of the diffusion matrix of the CFPE. To see this, note that the diffusion matrix for the system (A.1) is given by

$$\mathbf{B} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}. \tag{A.14}$$

Since this matrix is diagonal in the basis x, x_A, x_B, it follows that λ_1, λ_2 and λ_3 are its eigenvalues and that the matrix is positive semi-definite only if all three eigenvalues are positive. However, we have seen that the alternative form of the CLE breaks down at $\lambda_1 = 0$ which indeed corresponds to $\mathbf{B}$ losing its positive semi-definite form and hence to an ill-defined CFPE. Hence the breakdown of all possible CLEs in real variable space for this system is guaranteed.

A.2 Positive semi-definiteness of the diffusion matrix associated with the CLE

Let $\mathbf{C} \in \mathbb{R}^{m\times n}$ be a real matrix, $m, n \in \mathbb{N}$, and $\mathbf{B} = \mathbf{C}\mathbf{C}^T \in \mathbb{R}^{m\times m}$. Let $\mathbf{v} \in \mathbb{R}^m, \mathbf{v} \neq 0$, be an eigenvector of $\mathbf{B}$ with eigenvalue $\lambda \in \mathbb{R}$ such that

$$\mathbf{B}\mathbf{v} = \lambda\mathbf{v}. \tag{A.15}$$

For a vector $\mathbf{w} \in \mathbb{R}^p$, $p \in \mathbb{N}$, let $||\mathbf{w}||_p$ denote the Euclidean norm in $\mathbb{R}^p$, $||\mathbf{w}||_p = (\mathbf{w}^T\mathbf{w})^{1/2}$. Consider

$$\lambda||\mathbf{v}||_m^2 = \lambda\mathbf{v}^T\mathbf{v} = \mathbf{v}^T\mathbf{B}\mathbf{v} = \mathbf{v}^T\mathbf{C}\mathbf{C}^T\mathbf{v} = (\mathbf{C}^T\mathbf{v})^T(\mathbf{C}^T\mathbf{v}) = ||\mathbf{C}^T\mathbf{v}||_n \geq 0. \tag{A.16}$$

Since $\mathbf{v} \neq 0$, we have $||\mathbf{v}||_m^2 > 0$ and thus $\lambda \geq 0$. Since $\mathbf{B} = \mathbf{C}\mathbf{C}^T$ is symmetric, it is diagonalisable. We have shown that all eigenvalues are non-negative and can thus conclude that $\mathbf{B}$ is positive semi-definite. Conversely it follows that if the diffusion matrix $\mathbf{B}$ is not positive semi-definite then the matrix $\mathbf{C}$ cannot be real.

A.3 Exact solution of the CME describing catalysis by a single enzyme molecule

Here we derive an exact solution to the CME for the enzyme reaction system with reactions in Eq. (3.40). To the best of our knowledge this has not been reported to this date; a previous exact derivation led only to an explicit expression for the mean value of the substrate [106].

Define $c = c_1/\Omega$, rescale time as $\tau = ct$ and define $k_2 = c_2/c, k_3 = c_3/c, k_4 = \Omega c_4/c$. Let $P_0(n,\tau)$ and $P_1(n,\tau)$ be the probability of having n substrate molecules at time τ given 0 and 1 free enzyme molecules, respectively. The coupled CMEs describing the time evolution of these two probabilities are then given by

$$\partial_\tau P_0(n,\tau) = k_4 P_0(n-1,\tau) + (n+1)P_1(n+1,\tau) - k_4 P_0(n,\tau) - k_2 P_0(n,\tau) - k_3 P_0(n,\tau), \tag{A.17}$$

$$\partial_\tau P_1(n,\tau) = k_4 P_1(n-1,\tau) + k_2 P_0(n-1,\tau) + k_3 P_0(n,\tau) - k_4 P_1(n,\tau) - n P_1(n,\tau). \tag{A.18}$$

Define the generating functions as

$$G_0(s) = \sum_n s^n P_0(n), \tag{A.19}$$

$$G_1(s) = \sum_n s^n P_1(n). \tag{A.20}$$

Multiplying (A.17) and (A.18) in steady state ($\partial_\tau P_0 = \partial_\tau P_1 = 0$) with s^n and summing over n leads to

$$0 = (k_4(s-1) - k_2 - k_3)G_0(s) + \partial_s G_1(s), \tag{A.21}$$

$$0 = (k_2 s + k_3)G_0(s) + (k_4(s-1) - s\partial_s)G_1(s). \tag{A.22}$$

Solving the second equation for G_0 and inserting into the first gives

$$0 = \frac{(k_4(s-1) - k_2 - k_3)k_4(1-s)}{k_2 s + k_3} G_1(s) + \left(\frac{(k_4(s-1) - k_2 - k_3)s}{k_2 s + k_3} + 1 \right) \partial_s G_1(s), \tag{A.23}$$

which leads to the solution

$$G_0(s) = C \frac{e^{k_4 s}}{(k_3 - k_4 s)^{k_2 + k_4 + 1}} = C' \frac{e^{k_4 s}}{(k_{34} - s)^{k+1}}, \tag{A.24}$$

$$G_1(s) = C \frac{e^{k_4 s}}{(k_3 - k_4 s)^{k_2 + k_4}} = C' \frac{e^{k_4 s}}{(k_{34} - s)^{k}}, \tag{A.25}$$

where $k_{34} = k_3/k_4$, $k = k_2 + k_4$, and C' is a normalisation constant. The latter can be obtained from the normalisation condition

$$\sum_n (P_0(n) + P_1(n)) = G_0(1) + G_1(1) = 1, \tag{A.26}$$

which leads to

$$C' = \frac{e^{-k_4}}{k_{34}}(k_{34} - 1)^{k+1}. \tag{A.27}$$

Hence the generating function solution is given by

$$G_0(s) = \frac{e^{k_4(s-1)}}{k_{34}}\left(\frac{k_{34} - 1}{k_{34} - s}\right)^{k+1}, \tag{A.28}$$

$$G_1(s) = \frac{e^{k_4(s-1)}}{k_{34}}\frac{(k_{34} - 1)^{k+1}}{(k_{34} - s)^k}. \tag{A.29}$$

One can show by induction that

$$\begin{aligned}
\partial_s^n G_0(s) = &= \frac{C'}{k}\frac{\sum_{i=0}^n \binom{n}{i}[k]^{n-i+1}(k_4(k_{34} - s))^i}{(k_{34} - s)^{k+n+1}}e^{k_4 s} \\
&= \frac{(k_{34} - 1)^{k+1}}{k_{34}k}\frac{e^{k_4(s-1)}}{(k_{34} - s)^{k+n+1}}\sum_{i=0}^n \binom{n}{i}[k]^{n-i+1}(k_4(k_{34} - s))^i, \qquad \text{(A.30)} \\
\partial_s^n G_1(s) = &= C'\frac{\sum_{i=0}^n \binom{n}{i}[k]^{n-i}(k_4(k_{34} - s))^i}{(k_{34} - s)^{k+n}}e^{k_4 s} \\
&= \frac{(k_{34} - 1)^{k+1}}{k_{34}}\frac{e^{k_4(s-1)}}{(k_{34} - s)^{k+n}}\sum_{i=0}^n \binom{n}{i}[k]^{n-i}(k_4(k_{34} - s))^i, \qquad \text{(A.31)}
\end{aligned}$$

where we have used the definition for the rising factorial

$$[k]^i = k \cdot (k+1) \ldots (k+i-1), \tag{A.32}$$

$$[k]^0 = 1. \tag{A.33}$$

The probability distributions can now be obtained by using their definition in terms of the generating functions which read

$$P_0(n) = \frac{1}{n!}\partial_s^n G_0(s)|_{s=0}, \tag{A.34}$$

$$P_1(n) = \frac{1}{n!}\partial_s^n G_1(s)|_{s=0}. \tag{A.35}$$

Substituting Eqs. (A.30) and (A.31) in Eqs. (A.34) and (A.35), leads to

$$P_0(n) = \frac{C''}{k}\frac{k_4^{n+1}}{n!}\sum_{i=0}^n \binom{n}{i}[k]^{n-i+1}k_3^{i-n-1}, \tag{A.36}$$

$$P_1(n) = C''\frac{k_4^n}{n!}\sum_{i=0}^n \binom{n}{i}[k]^{n-i}k_3^{i-n}, \tag{A.37}$$

where $C'' = e^{-k_4}(k_{34}-1)^{k+1}/k_{34}^{k+1}$. These can be compactly represented in terms of the confluent hypergeometric function ${}_1F_1$ as

$$P_0(n) = \frac{C''}{kn!}\left(\frac{k_4}{k_3}\right)^{n+1}\frac{\Gamma(n+k+1)}{\Gamma(k)}\,{}_1F_1\left[-n;-(k+n);k_3\right], \tag{A.38}$$

$$P_1(n) = \frac{C''}{n!}\left(\frac{k_4}{k_3}\right)^{n}\frac{\Gamma(k+n)}{\Gamma(k)}\,{}_1F_1\left[-n;-(k+n-1);k_3\right]. \tag{A.39}$$

Analytic expressions for moments of arbitrary order can be directly computed by taking appropriate derivatives of the generating functions in Eqs. (G14) and (G15). Let $\langle n\rangle$ and Σ denote the mean and variance of the substrate, and further $\langle n_E\rangle$ and Σ_E the mean and variance of the free enzyme. We obtain the following expressions for these:

$$\langle n\rangle = \frac{k_4k_3(k_3+k_2)+k_4^2}{k_3(k_3-k_4)}, \tag{A.40}$$

$$\Sigma = \frac{k_4(k_3^2(k_4^2+k_3(-k_4+k_2+k_3))+k_4(k_3(k_4+k_3)-k_4^2))}{k_3^2(k_4-k_3)^2}, \tag{A.41}$$

$$\langle n_E\rangle = 1-\frac{k_4}{k_3}, \tag{A.42}$$

$$\Sigma_E = \frac{k_4(k_3-k_4)}{k_3^2}. \tag{A.43}$$

Appendix B

Non-uniqueness of moment closure approximations

Here, we investigate in detail the non-uniqueness of the Poisson and log-normal MAs for systems with conservation laws. To this end, we consider the simple reversible reaction system

$$A + B \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C. \tag{B.1}$$

We now compute the MA equations by applying the conservation laws of the system once *after*, and once *before* closing the equations.

Closing the equations first

This approach involves deriving the moment equations from the CME and subsequently imposing the conservation laws. The stoichiometric matrix and propensity vector for this system read (c.f. Eq. (2.24))

$$\mathbf{S} = \begin{pmatrix} -1 & 1 \\ -1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \mathbf{f}(n_1, n_2, n_3) = \left(\frac{k_1}{\Omega} n_1 n_2, k_2 n_3\right)^T, \tag{B.2}$$

where n_1, n_2 and n_3 denote the copy numbers of species A, B and C, respectively, and Ω is the volume of the system. The corresponding time evolution equations for the first and second-order moments can be obtained from Eqs. (2.33) and (2.34). For $y_1 = \langle n_1 \rangle$ and $y_{1,1} = \langle n_1^2 \rangle$, for example, they read

$$\partial_t y_1 = -\frac{k_1}{\Omega} y_{1,2} + k_2 y_3, \tag{B.3}$$

$$\partial_t y_{1,1} = -2\frac{k_1}{\Omega} y_{1,1,2} + 2k_2 y_{1,3} + \frac{k_1}{\Omega} y_{1,2} + k_2 y_3. \tag{B.4}$$

Note that due to the term $n_1 n_2$ in the first element of the propensity vector in Eq. (B.2), the equation for $y_{1,1}$ depends on the third-order moment $y_{1,1,2}$, but not on any *diagonal* third-order moment, i.e., not on $y_{1,1,1}, y_{2,2,2}$ or $y_{3,3,3}$. The same is of course true for the equations of the other second-order moments: they do not depend on a diagonal third-order moment. This means that the second-order normal and Poisson MAs are equivalent, since they differ only in their expressions for diagonal moments (c.f. Eqs. (2.39)-(2.41)). The corresponding second-order normal and Poisson MAs for y_1 and $y_{1,1}$ are obtained by setting the corresponding third-order cumulant $c_{1,1,2}$ to zero which leads to $y_{1,1,2} = 2y_1 y_{1,2} + y_2 y_{1,1} - 2y_1^2 y_2$ and thus gives

$$\partial_t y_1 = -\frac{k_1}{\Omega} y_{1,2} + k_2 y_3 \tag{B.5}$$

$$\partial_t y_{1,1} = -4\frac{k_1}{\Omega} y_1 y_{1,2} - 2\frac{k_1}{\Omega} y_2 y_{1,1} + 4\frac{k_1}{\Omega} y_1^2 y_2 + 2k_2 y_{1,3} + \frac{k_1}{\Omega} y_{1,2} + k_2 y_3, \tag{B.6}$$

and similarly for the other first and second-order moments. Note that the system has two conservation laws which can be written as

$$n_1 + n_3 = A^t = const., \tag{B.7}$$

$$n_2 + n_3 = B^t = const. \tag{B.8}$$

To simplify the following equations, let us assume $A^t = B^t$, which implies $n_1 = n_2$. The system of moment equations of three variables can thus be reduced to a system with only one variable, since all moments of first and second order can be expressed in terms of y_1 and $y_{1,1}$ using Eqs. (B.7) and (B.8). For example, we have $y_3 = \langle n_3 \rangle = \langle A^t - n_1 \rangle = A^t - y_1$ and $y_{1,2} = \langle n_1 n_2 \rangle = \langle n_1 n_1 \rangle = y_{1,1}$ and similarly for the other first and second-order moments. The resulting equations for y_1 and $y_{1,1}$ are thus closed and read

$$\partial_t y_1 = -\frac{k_1}{\Omega} y_{1,1} + k_2 (A^t - y_1), \tag{B.9}$$

$$\partial_t y_{1,1} = -6\frac{k_1}{\Omega} y_1 y_{1,1} + 4\frac{k_1}{\Omega} y_1^3 + 2k_2 (A^t y_1 - y_{1,1}) + \frac{k_1}{\Omega} y_{1,1} + k_2 (A^t - y_1). \tag{B.10}$$

Note that these are the resulting second-order MA equations for both the normal *and* the Poisson MA.

Applying the conservation laws first

Alternatively, we can start from the reduced CME with species B and C eliminated, whose stoichiometric matrix and propensity vector are given by

$$\mathbf{S} = \begin{pmatrix} -1 & 1 \end{pmatrix}, \quad \mathbf{f}(n_1) = \left(\frac{k_1}{V} n_1^2, k_2 (A^t - n_1)\right)^T. \tag{B.11}$$

Note that due to the term n_1^2, the time evolution equation for the second-order moment $y_{1,1}$ depends on the *diagonal* third-order moment $y_{1,1,1}$ (all moments are diagonal here of course, since we deal with a system with only one variable). The corresponding equations for the first two moments can be obtained using Eqs. (2.33) and (2.34) and read

$$\partial_t y_1 = -\frac{k_1}{\Omega} y_{1,1} + k_2(A^t - y_1), \tag{B.12}$$

$$\partial_t y_{1,1} = -2\frac{k_1}{\Omega} y_{1,1,1} + 2k_2(A^t y_1 - y_{1,1}) + \frac{k_1}{\Omega} y_{1,1} + k_2(A^t - y_1). \tag{B.13}$$

For closing these equations to second order, we need to express $y_{1,1,1}$ in terms of y_1 and $y_{1,1}$. The corresponding expression is now not the same anymore for the normal and Poisson MAs. For the normal MA we have $y_{1,1,1} = 3y_1 y_{1,1} - 2y_1^3$. Inserting the latter into Eq. (B.13) one obtains the same result as in Eqs. (B.9) and (B.10) which we obtained by applying the conservation laws *after* closing the equations. In contrast, if we apply the Poisson MA, which sets $y_{1,1,1} = 3y_1 y_{1,1} - 2y_1^3 + y_1$, the resulting equation for $y_{1,1}$ is *not* equal to Eq. (B.10). The reason for this is that the Poisson MA does not treat diagonal and non-diagonal moments equivalently. Here, this means that the replacements of $y_{1,1,1}$ and $y_{1,1,2}$ differ from each other if one sets the index 2 to 1 in the expression for $y_{1,1,2}$. Since the same is true for the log-normal MA, the latter also gives differing results depending if the equations are closed before or after the conservation laws are applied. Since the normal and CMN-MA do treat diagonal and non-diagonal moments equivalently (so the expressions for $y_{1,1,1}$ and $y_{1,1,2}$ are the same after setting 2 to 1), these MAs do not suffer from this flaw.

Appendix C

Cox process representation - details on examples

C.1 Gene expression

Equations

Consider the gene expression system in Figure 5.3 in Chapter 5. We do not model the gene explicitly, but rather assume a homogeneous production of mRNA in the nucleus. The corresponding reactions are

$$\text{nucleus}: \quad \varnothing \xrightarrow{m_1} M, \tag{C.1}$$

$$\text{whole cell}: \quad M \xrightarrow{m_2} \varnothing, \tag{C.2}$$

$$\text{cytosol}: \quad M \xrightarrow{p_1} M + P, \tag{C.3}$$

$$\text{whole cell}: \quad P \xrightarrow{p_2} \varnothing, \tag{C.4}$$

and both the mRNA M and protein P diffuse across the whole cell with diffusion constants d_M and d_P, respectively. After linearising the reaction in Eq. (C.3) as explained in Section 5.2.2 the PR for this system is real and deterministic, and we obtain using Eq. (5.27)

$$du_M(x,t) = [d_M \Delta u_M(x,t) + m_1 h_n(x) - m_2 u_M(x,t)]dt, \tag{C.5}$$

$$du_P(x,t) = [d_P \Delta u_P(x,t) + p_1 h_c(x) u_M(x,t) - p_2 u_P(x,t)]dt, \tag{C.6}$$

$$h_n(x) = \frac{1}{r}\Theta(r-x), \tag{C.7}$$

$$h_c(x) = \frac{1}{1-r}\Theta(x-r), \tag{C.8}$$

where r is the size of the nucleus and Θ the Heaviside step function. The functions $h_n(x)$ and $h_c(x)$ arise because M and P only become created in the nucleus and cytosol, respectively. If we additionally include the reaction

$$P \xrightarrow{p_3} P + P, \tag{C.9}$$

the equation for $u_P(x,t)$ becomes a SPDE and reads

$$\begin{aligned} du_P(x,t) = &\left[d_P \Delta u_P(x,t) + p_1 h_c(x) u_M(x,t) + p_3 u_P(x,t) - p_2 u_P(x,t)\right]dt \\ &+ \sqrt{2p_3 u_P(x,t)}\, dW(x,t). \end{aligned} \tag{C.10}$$

Inference

Consider first the system without the reaction in Eq. (C.9). In this case the system corresponds to a Poisson process. After basis function projection of the PDEs in Eqs. (C.5) and (C.6) as explained in Section 2.5.2, we are left with solving a coupled system of ODEs and can compute data likelihoods as explained in Section 2.5.3. We fix the parameters to

$$r = 0.3, \quad d_M = 0.1, \quad m_1 = 20, \quad m_2 = 0.5, \quad d_P = 0.1, \quad p_1 = 20, \quad p_2 = 0.2, \tag{C.11}$$

and assume that initially there are zero mRNA molecules and zero protein molecules in the cell. We further assume that the mRNA is unobserved and consider measurements of the protein at thirty equally separated time points separated by $\Delta t = 0.5$. We solve the PDEs in Eqs. (C.5) and (C.6) by projecting them onto twenty basis functions as explained in Section 2.5.2. We then optimise the likelihood of the data with respect to the parameters to obtain the inferred parameter values. We vary the initial values for the parameters in the likelihood optimiser randomly between 0.5 times and 2 times the exact value. The inference results are shown in Table 5.3 in the main text.

Next, we consider the same system but with the additional reaction in Eq. (C.9), for which the PDE in Eq. (C.6) gets replaced by the SPDE in Eq. (C.10). Now the system corresponds to a Cox-process, and we are left with solving a coupled system of SDEs after basis function projection. Since the drift and mean are both linear in the variables, the mean and covariance can be obtained by direct numerical integration, and we approximate the solution of the SDEs by a corresponding multivariate normal distribution. The corresponding likelihoods can then be computed as explained in Section 2.5.3. We again consider measurements of the protein at thirty equally separated time points separated by $\Delta t = 0.5$ and optimise the corresponding likelihood. The results are shown in Table 5.4 in the main text.

C.2 SIRS model

Equations

The reactions of the SIRS system are

$$S + I \xrightarrow{k,r} 2I, \quad I \xrightarrow{r} R, \quad R \xrightarrow{s} S. \tag{C.12}$$

We consider a system in the two-dimensional square $[0,1] \times [0,1]$. After linearising the first reaction in Eq. (C.12) as explained in Section 5.2.2 the PR for this system is real, and we obtain using Eq. (5.27) for the intensity fields of S, I and R,

$$du_S(x,t) = d\Delta u(x,t) - ku_S(x,t)u_I(x,t) + su_R(x,t), \tag{C.13}$$

$$du_I(x,t) = d\Delta u_I(x,t) + ku_S(x,t)u_I(x,t) - ru_I(x,t), \tag{C.14}$$

$$du_R(x,t) = d\Delta u_R(x,t) + ru_I(x,t) - su_R(x,t), \tag{C.15}$$

where we omitted noise terms in the equation for $u_I(x,t)$ for simplicitly and hence treat the system deterministically.

Inference

Initially at $t = 0$, we distribute S_{ini} particles of species S randomly across the whole area, one I particle at $[0.05, 0.05]$ and assume zero R particles. We simulate data for forty time points equally spaced by $\Delta t = 1$. As a basis we take 100 basis functions equally distributed in both dimensions. The inference results are shown in Table 5.5 in the main text.

C.3 Drosophila embryo

Data and equations

The data of the Bicoid protein in Drosophila embryos we consider here consists of two-dimensional fluorescence data as depicted in the left panel in Figure 5.5 in the main text. Since the relation of measured fluorescence intensity to actual protein numbers is unknown we simply translate them one to one. The Bicoid system is typically modelled by a simple birth-death process with the reactions

$$\varnothing \xrightarrow{k_1} P, \quad P \xrightarrow{k_2} \varnothing. \tag{C.16}$$

For simplicity, since diffusion is radially symmetric, we only consider the data within a certain distance from the major axis of the embryos, which we then project onto this axis, thus effectively obtaining one-dimensional data. We assume further that the protein is produced within a certain range around the left tip of the embryos. Mathematically the system is thus basically equivalent to the mRNA system in Section 5.5.1. The intensity of the protein fulfils the PDE

$$du(x,t) = \left(d\Delta u(x,t) + k_1 f(x) - k_2 u(x,t)\right)dt, \tag{C.17}$$

where x is the distance from the left top of the embryo, d is the diffusion constant, k_1 the production rate, $f(x) = 1/r$ for $x \in [0,r]$ and $f(x) = 0$ for $x \notin [0,1]$, r is the production radius around the origin and k_2 is the decay rate.

Inference

Since we have steady-state data, not all parameters are identifiable. One can easily see that multiplication of k_1 and k_2 with the same factor leads to the same steady state. We thus infer the creation range r, the diffusion rate d, and the ratio $c = k_2/k_1$. We solve the PDE in Eq. (C.17) by projecting it onto twenty basis functions and solve the resulting ODEs for large times to ensure steady state. We optimise the likelihood for each of the embryos independently to obtain the inferred parameter values with mean values and standard deviations (in parenthesis) given in Eq. (5.46).

Bibliography

[1] Bernhard Palsson. *Systems biology*. Cambridge University Press, 2015.

[2] Matthew M Crane, Ivan BN Clark, Elco Bakker, Stewart Smith, and Peter S Swain. A microfluidic system for studying ageing and dynamic single-cell responses in budding yeast. *PloS One*, 9(6):e100042, 2014.

[3] Anders S Hansen and Erin K O'Shea. Promoter decoding of transcription factor dynamics involves a trade-off between noise and control of gene expression. *Molecular systems biology*, 9(1):704, 2013.

[4] Michael B Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[5] Christopher J Paddon, P J? Westfall, DJ Pitera, K Benjamin, K Fisher, D McPhee, MD Leavell, A Tai, A Main, D Eng, et al. High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature*, 496(7446):528–532, 2013.

[6] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[7] David LK Toner and Ramon Grima. Molecular noise induces concentration oscillations in chemical systems with stable node steady states. *The Journal of Chemical Physics*, 138(5):055101, 2013.

[8] David M Raskin and Piet AJ de Boer. Rapid pole-to-pole oscillation of a protein required for directing division to the middle of Escherichia coli. *Proceedings of the National Academy of Sciences*, 96(9):4971–4976, 1999.

[9] Daniel T Gillespie. The chemical Langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[10] Andrew Golightly and Darren J Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005.

[11] Christoph Zechner, Jakob Ruess, Peter Krenn, Serge Pelet, Matthias Peter, John Lygeros, and Heinz Koeppl. Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences*, 109(21):8340–8345, 2012.

[12] Andrew Duncan, Radek Erban, and Konstantinos Zygalakis. Hybrid framework for the simulation of stochastic chemical kinetics. *arXiv preprint arXiv:1512.03988*, 2015.

[13] Andrew Duncan, Shuohao Liao, Tomáš Vejchodskỳ, Radek Erban, and Ramon Grima. Noise-induced multistability in chemical systems: Discrete versus continuum modeling. *Physical Review E*, 91(4):042111, 2015.

[14] Lukasz Szpruch and Desmond J Higham. Comparing hitting time behavior of Markov jump processes and their diffusion approximations. *Multiscale Modeling & Simulation*, 8(2):605–621, 2010.

[15] Desmond J Higham. Stochastic ordinary differential equations in applied and computational mathematics. *IMA Journal of Applied Mathematics*, 76(3):449–474, 2011.

[16] Leo A Goodman. Population growth of the sexes. *Biometrics*, 9(2):212–225, 1953.

[17] Peter Whittle. On the use of the normal approximation in the treatment of stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 268–281, 1957.

[18] Ramon Grima. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 136(15):154105, 2012.

[19] Eszter Lakatos, Angelique Ale, Paul DW Kirk, and Michael PH Stumpf. Multivariate moment closure techniques for stochastic kinetic models. *The Journal of Chemical Physics*, 143(9):094107, 2015.

[20] Geoffrey R Holmes, Sean R Anderson, Giles Dixon, Anne L Robertson, Constantino C Reyes-Aldasoro, Stephen A Billings, Stephen A Renshaw, and Visakan Kadirkamanathan. Repelled from the wound, or randomly dispersed? Reverse migration behaviour of neutrophils characterized by dynamic modelling. *Journal of The Royal Society Interface*, page rsif20120542, 2012.

[21] Toby P Davies, Hannah M Fry, Alan G Wilson, and Steven R Bishop. A mathematical model of the London riots and their policing. *Scientific Reports*, 3, 2013.

[22] Hendrik A Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

[23] José E Moyal. Stochastic processes and statistical physics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):150–210, 1949.

[24] Joshua Wilkie and Yin M Wong. Positivity preserving chemical Langevin equations. *Chemical Physics*, 353(1):132–138, 2008.

[25] Saswati Dana and Soumyendu Raha. Physically consistent simulation of mesoscale chemical kinetics: The non-negative FIS-α method. *Journal of Computational Physics*, 230(24):8813–8834, 2011.

[26] Donald A McQuarrie, CJ Jachimowski, and ME Russell. Kinetics of small systems. II. *The Journal of Chemical Physics*, 40(10):2914–2921, 1964.

[27] Carlos A Gomez-Uribe and George 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.

[28] John Goutsias. Classical versus stochastic kinetics modeling of biochemical reaction systems. *Biophysical Journal*, 92(7):2350–2365, 2007.

[29] Matt J Keeling. Multiplicative moments and measures of persistence in ecology. *Journal of Theoretical Biology*, 205(2):269–281, 2000.

[30] Ingemar Nåsell. An extension of the moment closure method. *Theoretical Population Biology*, 64(2):233–239, 2003.

[31] Joao Hespanha. Moment closure for biochemical networks. In *3rd International Symposium on Communications, Control and Signal Processing*, pages 142–147. IEEE, 2008.

[32] Marian V Smoluchowski. Über Brownsche Molekularbewegung unter Einwirkung äußerer Kräfte und deren Zusammenhang mit der verallgemeinerten Diffusionsgleichung. *Annalen der Physik*, 353(24):1103–1112, 1916.

[33] Marian V Smoluchowski. Versuch einer mathematischen Theorie der Koagulationskinetic kolloider Lösungen. *Zeitschrift für Physikalische Chemie*, 92:129–168, 1917.

[34] Masao Doi. Second quantization representation for classical many-particle system. *Journal of Physics A: Mathematical and General*, 9(9):1465, 1976.

[35] Masao Doi. Stochastic theory of diffusion-controlled reaction. *Journal of Physics A: Mathematical and General*, 9(9):1479, 1976.

[36] Michael A Dewar, Visakan Kadirkamanathan, Manfred Opper, and Guido Sanguinetti. Parameter estimation and inference for stochastic reaction-diffusion systems: application to morphogenesis in D. melanogaster. *BMC Systems Biology*, 4(1):1, 2010.

[37] Andreas Ruttor and Manfred Opper. Approximate parameter inference in a stochastic reaction-diffusion model. In *International Conference on Artificial Intelligence and Statistics*, pages 669–676, 2010.

[38] David Schnoerr, Guido Sanguinetti, and Ramon Grima. The complex chemical Langevin equation. *The Journal of Chemical Physics*, 141(2):024103, 2014.

[39] David Schnoerr, Guido Sanguinetti, and Ramon Grima. Validity conditions for moment closure approximations in stochastic chemical kinetics. *The Journal of Chemical Physics*, 141(8):084103, 2014.

[40] David Schnoerr, Guido Sanguinetti, and Ramon Grima. Comparison of different moment-closure approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 143(18):185101, 2015.

[41] David Schnoerr, Ramon Grima, and Guido Sanguinetti. Cox process representation and inference for stochastic reaction-diffusion processes. *Nature Communications*, 7:11729, 2016.

[42] Ertugrul M Ozbudak, Mukund Thattai, Iren Kurtser, Alan D Grossman, and Alexander van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73, 2002.

[43] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005.

[44] Arren Bar-Even, Johan Paulsson, Narendra Maheshri, Miri Carmi, Erin O'Shea, Yitzhak Pilpel, and Naama Barkai. Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643, 2006.

[45] William J Blake, Mads Kærn, Charles R Cantor, and James J Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

[46] John RS Newman, Sina Ghaemmaghami, Jan Ihmels, David K Breslow, Matthew Noble, Joseph L DeRisi, and Jonathan S Weissman. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

[47] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010.

[48] Gábor Balázsi, Alexander van Oudenaarden, and James J Collins. Cellular decision making and biological noise: from microbes to mammals. *Cell*, 144(6):910–925, 2011.

[49] Crispin W Gardiner. *Handbook of stochastic methods*, volume 3. Springer Berlin, 1985.

[50] Iosif I Gikhman and Anatoli V Skorokhod. *The theory of stochastic processes II*. Springer, 2015.

[51] Daniel T Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1):404–425, 1992.

[52] Donald A McQuarrie. Stochastic approach to chemical kinetics. *Journal of Applied Probability*, 4(3):413–478, 1967.

[53] R Grima, Deena R Schmidt, and Timothy J Newman. Steady-state fluctuations of a genetic feedback loop: An exact solution. *The Journal of Chemical Physics*, 137(3):035104, 2012.

[54] Daniel 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.

[55] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[56] Sean Mauch and Mark Stalzer. Efficient formulations for exact stochastic simulation of chemical systems. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(1):27–35, 2011.

[57] Edwin L Crow and Kunio Shimizu. *Lognormal distributions: Theory and applications*, volume 88. M. Dekker New York, 1988.

[58] Abhyudai Singh and Joao P Hespanha. Lognormal moment closures for biochemical reactions. In *45th IEEE Conference on Decision and Control*, pages 2063–2068, 2006.

[59] Thomas G Kurtz. Limit theorems and diffusion approximations for density dependent Markov chains. In *Stochastic Systems: Modeling, Identification and Optimization, I*, pages 67–78. Springer, 1976.

[60] Bence Melykuti, Kevin Burrage, and Konstantinos C Zygalakis. Fast stochastic simulation of biochemical reaction systems by alternative formulations of the chemical Langevin equation. *The Journal of Chemical Physics*, 132(16):164109, 2010.

[61] Nicolaas G van Kampen. The expansion of the master equation. *Advance in Chemical Physics*, 34:245–309, 1976.

[62] Nicolaas G van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[63] Ramon Grima. An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. *The Journal of Chemical Physics*, 133(3):035101, 2010.

[64] Rajesh Ramaswamy, Nélido González-Segredo, Ivo F Sbalzarini, and Ramon Grima. Discreteness-induced concentration inversion in mesoscopic chemical systems. *Nature Communications*, 3:779, 2012.

[65] Philipp Thomas, Hannes Matuschek, and Ramon Grima. Computation of biochemical pathway fluctuations beyond the linear noise approximation using iNA. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1–5, 2012.

[66] RF Pawula. Approximation of the linear Boltzmann equation by the Fokker-Planck equation. *Physical Review*, 162(1):186, 1967.

[67] Crispin W Gardiner and S Chaturvedi. The Poisson representation. I. A new technique for chemical master equations. *Journal of Statistical Physics*, 17(6):429–468, 1977.

[68] Jeroen S van Zon and Pieter R Ten Wolde. Simulating biochemical networks at the particle level and in time and space: Green's function reaction dynamics. *Physical Review Letters*, 94(12):128103, 2005.

[69] Aleksandar Donev, Vasily V Bulatov, Tomas Oppelstrup, George H Gilmer, Babak Sadigh, and Malvin H Kalos. A first-passage kinetic Monte Carlo algorithm for complex diffusion-reaction systems. *Journal of Computational Physics*, 229(9):3214–3236, 2010.

[70] Daniel T Gillespie, Andreas Hellander, and Linda R Petzold. Perspective: Stochastic algorithms for chemical kinetics. *The Journal of Chemical Physics*, 138(17):170901, 2013.

[71] Samuel A Isaacson. Relationship between the reaction-diffusion master equation and particle tracking models. *Journal of Physics A: Mathematical and Theoretical*, 41(6):065003, 2008.

[72] Samuel A Isaacson and Charles S Peskin. Incorporating diffusion in complex geometries into stochastic chemical kinetics simulations. *SIAM Journal on Scientific Computing*, 28(1):47–74, 2006.

[73] Radek Erban and S Jonathan Chapman. Stochastic modelling of reaction-diffusion processes: algorithms for bimolecular reactions. *Physical Biology*, 6(4):046001, 2009.

[74] Jin Fu, Sheng Wu, Hong Li, and Linda R Petzold. The time dependent propensity function for acceleration of spatial stochastic simulation of reaction-diffusion systems. *Journal of Computational Physics*, 274:524–549, 2014.

[75] Andrew Zammit-Mangion, Michael Dewar, Visakan Kadirkamanathan, and Guido Sanguinetti. Point process modelling of the Afghan War Diary. *Proceedings of the National Academy of Sciences*, 109(31):12414–12419, 2012.

[76] Fernando Peruani and Chiu F Lee. Fluctuations and the role of collision duration in reaction-diffusion systems. *EPL (Europhysics Letters)*, 102(5):58001, 2013.

[77] Kathrine Grell, Peter J Diggle, Kirsten Frederiksen, Joachim Schüz, Elisabeth Cardis, and Per K Andersen. A three-dimensional point process model for the spatial distribution of disease occurrence in relation to an exposure source. *Statistics in Medicine*, 34(23):3170–3180, 2015.

[78] Noel Cressie and Christopher K Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

[79] Christopher M Bishop. Pattern Recognition. *Machine Learning*, 2006.

[80] Peter E Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*, volume 1. Springer Berlin, 1992.

[81] Alan Fersht. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*. Macmillan, 1999.

[82] Arren Bar-Even, Elad Noor, Yonatan Savir, Wolfram Liebermeister, Dan Davidi, Dan S Tawfik, and Ron Milo. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry*, 50(21):4402–4410, 2011.

[83] Long Cai, Nir Friedman, and X Sunney Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362, 2006.

[84] Mukund Thattai and Alexander Van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, 2001.

[85] Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[86] Thomas Wilhelm. The smallest chemical reaction system with bistability. *BMC Systems Biology*, 3(1):1, 2009.

[87] Philipp Thomas, Hannes Matuschek, and Ramon Grima. Intrinsic noise analyzer: a software package for the exploration of stochastic biochemical kinetics using the system size expansion. *PloS One*, 7(6):e38518, 2012.

[88] Albert Goldbeter and Daniel E Koshland. An amplified sensitivity arising from covalent modification in biological systems. *Proceedings of the National Academy of Sciences*, 78(11):6840–6844, 1981.

[89] Ilya Prigogine and Réné Lefever. Symmetry breaking instabilities in dissipative systems. II. *The Journal of Chemical Physics*, 48(4):1695–1700, 1968.

[90] René Lefever, Grégoire Nicolis, and Pierre Borckmans. The Brusselator: it does oscillate all the same. *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 84(4):1013–1023, 1988.

[91] Pål O Westermark, David K Welsh, Hitoshi Okamura, and Hanspeter Herzel. Quantification of circadian rhythms in single cells. *PLoS Computational Biology*, 5(11):e1000580, 2009.

[92] Carl Troein, Florence Corellou, Laura E Dixon, Gerben van Ooijen, John S O'Neill, François-Yves Bouget, and Andrew J Millar. Multiple light inputs to a simple clock circuit allow complex biological rhythms. *The Plant Journal*, 66(2):375–385, 2011.

[93] Gen Kurosawa and Albert Goldbeter. Amplitude of circadian oscillations entrained by 24-h light-dark cycles. *Journal of Theoretical Biology*, 242(2):478–488, 2006.

[94] Margaritis Voliotis, Philipp Thomas, Ramon Grima, and Clive G Bowsher. Stochastic simulation of biomolecular networks in dynamic environments. *arXiv preprint arXiv:1511.01268*, 2015.

[95] Joao P Hespanha. StochDynTools - a MATLAB toolbox to compute moment dynamics for stochastic networks of bio-chemical reactions, 2006.

[96] Colin S Gillespie. Moment-closure approximations for mass-action models. *Systems Biology, IET*, 3(1):52–58, 2009.

[97] Paul Azunre, C Gómez-Uribe, and George Verghese. Mass fluctuation kinetics: analysis and computation of equilibria and local dynamics. *Systems Biology, IET*, 5(6):325–335, 2011.

[98] See http://grimagroup.bio.ed.ac.uk/publications.html for software package MOCA as well as tutorial and example files.

[99] Philipp Thomas, Arthur V Straube, and Ramon Grima. Communication: Limitations of the stochastic quasi-steady-state approximation in open biochemical reaction networks. *The Journal of Chemical Physics*, 135(18):181103, 2011.

[100] Angelique Ale, Paul Kirk, and Michael PH Stumpf. A general moment expansion method for stochastic kinetic models. *The Journal of Chemical Physics*, 138(17):174101, 2013.

[101] Abhyudai Singh and Joao P Hespanha. Approximate moment dynamics for chemically reacting systems. *IEEE Transactions on Automatic Control*, 56(2):414–418, 2011.

[102] Botond Cseke, Andrew Zammit-Mangion, Tom Heskes, and Guido Sanguinetti. Sparse approximate inference for spatio-temporal point process models. *Journal of the American Statistical Association*, (just-accepted):1–52, 2015.

[103] John FC Kingman. *Poisson processes*, volume 3. Clarendon Press, 1992.

[104] Mohammed Abdullah, Colin Cooper, and Moez Draief. Viral processes by random walks on random regular graphs. In *APPROX-RANDOM*, pages 351–364. Springer, 2011.

[105] Andrei Pisarev, Ekaterina Poustelnikova, Maria Samsonova, and John Reinitz. FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Research*, 37:D560–D566, 2009.

[106] Marianne O Stefanini, Alan J McKane, and Timothy J Newman. Single enzyme pathways and substrate fluctuations. *Nonlinearity*, 18(4):1575, 2005.