Chemical Master Equations for Non-linear Stochastic Reaction Networks: Closure Schemes and Implications for Discovery in the Biological Sciences

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

PATRICK SMADBECK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

YIANNIS KAZNESSIS, ADVISER

July, 2014

Acknowledgements

While I would love to defy cliché and declare myself an island, a man whose inspiration springs unfettered from within, that would of course be incredibly disingenuous. I have been fortunate enough to have been surrounded by an invaluable support system during my doctoral studies, and their contribution to this thesis in some way rivals my own.

I wish to extend my utmost gratitude and appreciation to my adviser Yiannis Kaznessis. Playing the role of sage guide, supportive mentor and enthusiastic colleague at just the right moments has made my time at the University of Minnesota incredibly productive and equally enjoyable. Nothing quite lifted me up like the genuine excitement he felt about my ideas and where and how they might be applied to our models. Without Yiannis the mere thought of moment closure would never have crossed my mind. His professional, academic and personal support cannot be overstated, and I extend my deepest thanks for everything he has contributed to my work throughout the years.

My deepest appriciation to my parents, Diane and Arthur, whose unyielding support never fails to amaze me. And to my brothers, Louis, Mark, Jeff, and Jamie, all of which I love beyond comprehension and are all hilarious and amazing people.

To my friends in and out of the department for consistent amusement and diversion through the years. Kevin, Mel, Tim, Garrett and Alex, happy hour will never be the same without you. To my lab mates both past and present. Kat and Kostas guided me through my formative years in the department and were both champions and critics whenever necessary. Giota more recently helped me keep the computational lab lively in our rather large working quarters. And Michael has managed to put up with my nonsense these past few months. I wish you the best in your future work and studies.

Finally, a special thanks needs to be given to my fiancée Caryn. She suffered through three years of a long distance relationship. She moved to Minnesota, she got a dog with

me and has supported me throughout it all with grace. Nothing can really describe how important she has been to my current and continued success. I love you Caryn.

Those named are only a fraction of the people who have passed through my life these past five years. To name everyone to whom I own a debt upon completion of this thesis would be impossible (or at the very least extremely improbable). You all undoubtedly made my work and life at the University of Minnesota a unique and special experience and I thank you.

I will conclude with an expression of my sincerest gratitude to my final examination committee, and for support from the National Institutes of Health (American Recovery and Reinvestment Act Grant GM086865 and Training Grant), the National Science Foundation (CBET-0644792), and the University of Minnesota Doctoral Dissertation Fellowship.

Dedication

This thesis is dedicated to my parents, Diane and Arthur, to my fiancée Caryn, and to my future in-laws Larry and Mary Louise for their incredible support over the past five years.

Abstract

With the development of genome-wide sequencing, DNA synthesis technologies and the continued growth of supercomputing resources, biology has become a new focus of engineering research across the globe. Our ability to analyze gene function and develop novel synthetic biological systems has made engineered biological constructs a reality. Despite the advancement of computational resources and numerical methods biological research, however, remains the domain of experimental scientists. Novel simulation methods and theories for biological simulation are sorely needed in order to bridge the gap between the experimental and computational sides of biological engineering.

One major issue facing biological simulations is that these systems experience random fluctuations that can strongly influence and drive overarching function. The importance of these random fluctuations to the accuracy of the simulation requires the use of stochastic mathematics. Instead of describing and simulating a single deterministic trajectory through time for a chemical system, a probabilistic distribution of possible states must be determined. In such systems the master equation describes, in full detail, the underlying dynamics. In practice, however, such a solution for non-linear systems has been elusive for over 50 years. From a statistical perspective what is missing is a relationship between complex sets of statistics that has remained unresolved for decades called the *moment closure problem*. Solving this problem would allow for a new way to analyze and optimize stochastic simulations using deterministic numerical methods.

The work presented herein focuses on the full development of a numerical solution to the moment closure problem using maximum-entropy distributions. The intentions of my work were: (1) To develop an algorithm to quickly produce moment equations that fully describe the dynamics of the chemical master equation deterministically; (2) Develop a novel moment closure method using maximum-entropy optimization to solve the master equation; (3) Demonstrate the potential of this method for performing non-linear analysis, power spectral density determination and model reduction on stochastic systems. I will demonstrate in this initial study a new method for the simulation of biological systems (and other systems with a random nature) that is entirely separate from the methods that currently dominate stochastic biological simulation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Stochastic Simulation Motivation

Traditionally the mathematical modeling of biologically relevant systems has relied on a single fundamental assumption that such systems can be represented by a set of chemical reactions following mass action kinetics. These reaction-rate equation (RRE) models are typically represented as a set of ordinary differential equations (ODEs) and concern the macroscopic property of concentration, scaled with the system volume[1]. These simplified ODE models are still widely and effectively used in fields such as systems biology[2], where very large sets of reactions need to be modeled and analyzed efficiently. The use of these simple models in biological system simulation dates back to the early work of Monod[3] and Goodwin[4].

The simple ODE models assume that the system is at the thermodynamic limit. The thermodynamic limit concerns an important condition implicit to ODE simulations: the number of molecules of each component is so large that the concentration can be assumed to be continuous[5]. These systems, as a result, are deterministic, given the initial conditions and set of reaction rate laws the concentration of every component can be determined exactly for all future time. This is, at its core, the difference between statistical mechanics and classical mechanics. In statistical mechanics the noise in a system (the amount macroscopic properties can deviate from average) is inversely proportional to the number of particles, N, within the system[6]. In deterministic ODE models the number of molecules for each chemical component can often be on the order of 10^{23}

(Avagadro's number) for a benchtop experiment. The noise over signal ratio in such a system is thus on the order of 10^{-10}, practically undetectable under even the most precise experimentation. When the system is at the thermodynamic limit, statistical mechanics converges onto classical mechanics and ODE models are applicable.

Biological systems, however, are often far from the thermodynamic limit and are not accurately represented by ODE models[7, 8, 9]. Considering molecules such as DNA (a single macromolecule) and messenger RNA (often existing as a low copy number from 0-10 molecules) the condition of having a large number of molecules is obviously not achieved. Fluctuations in the number of messenger RNA can then have a dramatic effect on the underlying dynamics and biological function of the system[10, 11, 12, 13]. The concentration of each component can experience discrete jumps in value and these jumps will occur at seemingly random times. These systems are decidedly non-deterministic and given a set of initial conditions the system can progress in an ensemble of possible trajectories through time.

Biological systems are more accurately simulated using stochastic mathematics[14, 15, 16, 17, 18]. The state of the system at any given time can take on a wide range of values, each with an associated probability. Stochastic simulations describe the states as an underlying probability distribution and typically concern themselves with the statistics, such as the mean and variance, of these distributions. It has been demonstrated that the dynamics of stochastic systems can differ substantially from equivalent deterministic models[19]. Noise can have a profound effect on system dynamics, including being linked to complex behavior such as spontaneous state switching and oscillations[20, 21, 22, 23, 24].

In the following section a comprehensive review of stochastic simulation will be presented as a motivation for the development of novel stochastic simulation methods for chemical reaction networks.

1.2 Stochastic Simulation

1.2.1 Historical Review

Probability theory as a mathematical discipline had been well established by the 1700s by great mathematicians such as Bayes, Laplace and Bernoulli among many others[25].

And while many of the statistical methods utilized by contemporary scientists were established prior to 1900, the mathematical theory itself was not specifically applied to natural processes until the development of statistical mechanics, quantum theory, and chaos theory. The development of the Maxwell-Boltzmann distribution and the groundbreaking work on Brownian motion by Albert Einstein[26] and others[27] firmly established a place for probability theory in the natural sciences. Up until that point the universe itself was widely consider to be deterministic, fully predictable given only the laws of nature and well-defined initial conditions. During the early part of the 20th century, however, descriptions of electrical processes (Nyquist) and a general rigorous description of statistical dynamics (Kolomogorov) soon brought a more probabilistic view of the universe to the natural sciences[28].

That is not to completely invalidate or reject a deterministic description of the universe (although the uncertainty principle does make obtaining sufficiently accurate initial conditions somewhat questionable). Stochastic theory and statistical mechanics merely provides a simplified mathematical construct within which to explore the dynamics of such systems. Consider the simplest of the games of chance often explored by the early pioneers of probability theory: the flipping of a coin. It is possible to describe the entire isolated system in which the coin exists; the initial positions of every air molecule, their initial velocities, the exact position of the perfectly fair coin itself, etc. And perhaps once this incredible simulation is complete it could determine precisely whether a heads or tails will be obtained on the particular flip being studied. But such a simulation is unnecessary because a simplified description is available: the probability of obtaining one of the two states (heads and tails) is for all practical purposes uniform (50-50) for a fair coin. Stochastic simulation does not necessarily defy a deterministic universe, but it does accept our inability to provide a perfect description of the state of a system being modeled. In this way stochastic simulation can provide a mathematically tractable version of systems that would otherwise be too complex (or unknowable) to model deterministically.

From the turn of the century to approximately 1940 stochastic mathematics were developed for applications in a wide range of fields. The primary pioneers of the theoretical aspect of such work were pure mathematicians such as Kolomagorov, Ito and Lagevin[29] (Ito and Kolomagorov's student Stratonovich developed the fundamental

theorems for stochastic calculus). These ideas, however, were not specifically applied to chemical processes until the 1950s and 1960s.

In 1940, Delbrück[30] described the possibility of observable chemical fluctuations in autocatalytic processes. Later, Renyi[31], Bartholomay[32, 33] and Ishida[34, 35] among others[36] would all explore the stochastic description of simple chemical systems. Applications to general chemical networks were established in 1967 by McQuarrie who fully described what is referred to as the chemical master equation (CME)[37]:

$$\frac{\partial P\left(\underline{X};t\right)}{\partial t} = \sum_{\underline{X}'} \left[T\left(\underline{X}|\underline{X}'\right) P\left(\underline{X}';t\right) - T\left(\underline{X}'|\underline{X}\right) P\left(\underline{X};t\right)\right] \tag{1.1}$$

Here $P\left(\underline{X};t\right)$ is the probability of being in state $\underline{X}$ at time t, and $T\left(\underline{X}|\underline{X}'\right)$ is the probability of transitioning from state $\underline{X}'$ to state $\underline{X}$. The state $\underline{X}$ in this context means the vector of chemical concentrations. This is a forward Kolomagorov equation, and theoretically provides an exact description of the underlying probability distribution dynamics for chemical systems. Much later Gillespie would prove that, given a small set of assumptions (a well mixed volume), the CME is, indeed, an exact theoretical description of a chemical system from a probabilistic viewpoint[38]. McQuarrie's two groundbreaking papers[39, 40] described and solved the master equation for a select few very simple chemical reaction networks (both linear and non-linear). Unfortunately, it is also clear that such a solution method was impossible for anything but the simplest networks, and that, from a mathematical perspective analytically solving stochastic chemical reaction simulations via the CME was impractical.

For about ten years the exact stochastic simulation of complex chemical reaction networks remained practically impossible, instead relying on the development of approximate stochastic differential equation methods[41, 42]. In 1976, extending the theoretical work previously done by Doob[43], Gillespie described the stochastic simulation algorithm (SSA) which outlines a Monte Carlo method for sampling the master equation solution to arbitrary chemical reaction networks[44]. In particular, by assuming the propensity for reaction (the probability a reaction event occurring in a small time step Δt) is a Poisson distribution the time to the next reaction is:

$$\tau = -\frac{\ln URN}{\sum_r a_r\left(\underline{X}\right)} \tag{1.2}$$

Where URN is a uniformly distributed random number and $a_r\left(\underline{X}\right)$ is the propensity for the reaction r to occur given the current state of the system $\underline{X}$. The probability a specific reaction, r', occurred in that time period is proportional to the propensity relative to the total propensity of reaction:

$$P\left(r' \text{ is the next reaction}\right) = \frac{a_{r'}\left(\underline{X}\right)}{\sum_r a_r\left(\underline{X}\right)} \tag{1.3}$$

By generating random numbers these two equations determine a chain of reaction events and reaction times describing a single stochastic trajectory. By constructing a multitude of such trajectories the CME probability distribution can be obtained. Much later this incredible achievement would be recognized as the key to efficient stochastic simulation of chemical systems. Later still the necessity of stochastic simulations for chemical networks would be recognized in the cutting edge fields of bioengineering.

1.2.2 Contemporary Review

In the late 1990s, after the incredible leaps made by physics in the 20th century, many scientists declared the next frontier of research to be biology[45]. The advancements in the understanding of biology with the human genome project offered a full scale perspective of the genes that control human physiology[46]. More importantly the real potential of constructing novel and deliberate function within biological systems on a large scale was beginning to become apparent with the advent of cheap DNA synthesis technology[47]. Using these tools fields such as synthetic biology began to look at a bottom-up approach to the engineering of biological function[48].

The advancements concerning the stochastic simulation of chemical reaction networks began on three fronts. From an experimental perspective Elowitz and Liebler were able to show the construction of a synthetic oscillator within a prokaryotic cell (along with a simple stochastic model to help describe its function)[49]. Elsewhere, Gardner et al (among many others) showed experimental evidence that cellular behavior is, indeed, significantly affected by fluctuations and the necessity of stochastic simulation for accurate descriptions of biological systems began to become apparent[50, 51, 52, 53, 54, 55]. And finally, from the theoretical perspective the first major advancement of Gillespie's SSA was made by Gibson and Bruck[56], potentially making the efficient simulation of large biologically relevant chemical reaction networks possible.

For the last twenty years there has been an explosion of research from all three fronts. Many synthetic constructs were developed and introduced into the growing toolbox for synthetic biology (including our own AND-gate and tetON/OFF constructs)[57, 58, 18, 59]. The evidence for the necessity and importance of stochastic simulations has mounted, with the intrinsic and extrinsic noise in a system being credited for the stability of some systems, toggle switch behavior and even oscillations[60, 61, 62, 63]. And a multitude of major and minor advancements in the SSA and other simulation algorithms[64, 65, 66, 67] or approximations[68, 69, 70, 71] have been developed as well (including our own hybrid stochastic simulation algorithm for supercomputers, Hy3S[72]). These methods are summarized nicely by Li et al[73]. All of these advancements have made the construction, simulation and evaluation of synthetic biological constructs both cheap and easy.

Yet, despite this advancement the computational side of synthetic biology, and more generally biology itself, has remained a step behind the experimental work being done in the field. Looking to industrial engineering or the aerospace fields it is obvious that simulation and prediction is a major component of any well developed engineering field. And so, from a theoretical perspective there is a lot left to be desired. Stochastic simulation is very much available for just about as large and complex of systems as one can find using kinetic Monte Carlo simulations. Unfortunately, Monte Carlo sampling, while incredibly efficient for finding dynamic behavior of a system where the kinetic constants are known, cannot obtain steady states or perform analysis on stochastic systems efficiently. The potential for what is called moment closure to provide a faster method for determining steady states and behavior of stochastic chemical networks will be introduced and motivated in the next section.

1.3 Moment Closure

The motivation for moment closure is that for biological systems the computational cost of kinetic Monte Carlo methods becomes prohibitive. This is especially true when the simulations concern either steady states, which result in wasted computation time, or analysis, which often rely on numerical approximation and multiple simulations to achieve results. The desirable approach to analysis would be to take advantage of

the wealth of knowledge available to deterministic modelers. With deterministic systems steady states can be obtained immediately from optimization programs. Using the steady-state solution non-linear analysis, sensitivity analysis and other numerical methods can give insight into the underlying behavior of the system and the important kinetic components. With kinetic Monte Carlo simulations performing such analysis is either impossible or impractical. Moment closure, as will be shown below, has the potential to provide a deterministic viewpoint for stochastic simulations allowing for deterministic analysis to be applied.

1.3.1 What is Moment Closure?

The moments, often referred to as statistics, of a probability distribution describe the shape and characteristic form of the distribution function. Readers may be familiar with the first two lowest-order moments, the mean and variance, but less familiar higher-order moments become necessary when describing more complex distributions. Mathematically, for a simple single component system the mean is:

$$\langle x \rangle = \int_{-\infty}^{\infty} xP\left(x\right) dx \tag{1.4}$$

Where $P\left(x\right)$ is the probability of being in state x. The mean is often described as providing an expectation for the distribution, the average value the random variable described by the distribution will take. Higher-order moments are defined similarly:

$$\left\langle x^M \right\rangle = \int_{-\infty}^{\infty} x^M P\left(x\right) dx \tag{1.5}$$

Here M is called the order of the moment. The variance, for example, is related to the second-order moment (it is the second central moment) and describes the width of a given distribution. The skewness is related to the third-order moment and describes asymmetries. And the kurtosis is related to the fourth-order moment and describes the squatness.

Looking back to Equation 1.1 it is possible (see Chapter 2) to use transforms to obtain dynamic equations for the moments that are equivalent to the CME. For well-behaved distributions (i.e. no discontinuities, this is a good assumption for the chemical systems dealt with throughout this thesis) the infinite set of moments have a one-to-one

relationship with the distribution they represent. Thus solving the moment equations is, in theory, equivalent to solving the CME.

The immediate benefit of the moment viewpoint is that we are now dealing with deterministic equations. Given a set of initial moments (defined by an initial distribution) and the moment equations defined by the reaction network and all of the moments through time can be calculated immediately. The issue, however, is the same as was discovered by McQuarrie in 1967[37]. The set of equations is unclosed. The mean dynamics depend on the variance, the variance dynamics on the skewness, etc. In theory an infinite set of moments is necessary to solve the moment equations, and this is called the moment closure problem. In practice, using the moment viewpoint for all but the simplest networks requires a moment closure solution, a way to relate a higher-order moment to the lower-order moments (either numerically or analytically).

1.3.2 Moment Closure Review

The first moment closure schemes predate even the description of chemical dynamics by stochastic mathematics[74, 75] and dealt exclusively with general stochastic processes. The early closure schemes were of a single type: normal distribution approximations. The simplest analytical closure method, this approximation was used almost exclusively by ecologists through the 1990s [76, 77, 9]. While simple, this approximation is still widely used today as a benchmark example of moment closure. Fundamentally, however, this method was wanting. Using the normal distribution can result in negative values for molecules and is only applicable as a third-order closure scheme (although some researchers have looked into higher-order closure[78]). Theoretically there was a lot of room for improvement.

The interesting thing about moment closure is the almost 40 year gap in significant publications that utilize (let alone advance) the method. After Whittle a novel closure scheme wouldn't be proposed and explored until 2005. At this time an explosion of papers come out in a short 5 years stretch.

In 2005 Krishnarajah[79] proposed a new closure scheme for epidemic models that exhibited significant skewness. In this closure scheme a beta-binomial distribution is used in conjunction with a log-normal distribution in order to represent non-normal distributions. This closure scheme is one of the only examples of mixing multiple distributions

to represent complex behavior and was extended to include zero-modified distributions as well[80]. Krishnarajah's work represents some of the best multi-distribution schemes available, although the method has not been widely applied since.

Hespanha and Singh introduced one of the first numerical approximations for moment closure with the seperable-derivative matching (SMD) method[81]. In this scheme derivatives of the moments can be set equal to a chosen power-law ansatz in order to produce a flexible high-order closure scheme. First proposed in 2005 the method has been subject to continued study and refinement over the last 10 years [82, 83, 84]. Their 2011 paper provides the most complete description of the method which, essentially, assumes that high-order moments are consistent with a log-normal distribution. This method is widely considered to be one of the most promising analytical closure method developed.

Lee introduced a Taylor series expansion inspired method involving setting central-moments to zero[85]. The research is remarkable in two ways. First, as a numerical moment-closure scheme it convincingly exhibits, on a limited number of example systems, increased accuracy on increased closure order without bound, an important and rare characteristic. More importantly, however, outside of the closure method presented herein this is the only publication to claim the ability to immediately obtain steady-state distributions (and correctly deduces that finding such solutions may be substantially faster than similar SSA results). It extends this observation by suggesting (although no results are presented) the ability to assess steady-state results via non-linear eigenvalue analysis. The use of moment-closure to achieve stochastic non-linear analysis is presented in the context of our own closure method in Chapter 4.

Other works of note include Gomez-Uribe[86] who essentially assumed the third-central moments to be zero (symmetry) for a multi-component distribution. Stumpf[87] extends this by allowing any order central moment to be zero as desired. Saddlepoint approximations[88] provide an approach to approximating distributions given a set of moments to provide analytical solutions to higher-order moments, but often results in inaccurate or abberant results. Ruess[89] uses a small number of stochastic simulations to estimate moments. Software for the use of moment-closure is also provided by Gillespie[90]. These twelve papers comprise the majority of novel moment closure work developed and applied to chemical reaction networks over the last 50 years.

Overall the available moment closure methods developed during this period fall into two broad categories: distribution-based methods and numerical methods. With distribution based methods one assumes a distribution and assigns analytically calculated moments based on this assumed information. One could assume the data will follow a normal distribution[75], or a log-normal distribution[79], etc. There are two immediate issues with this approach. First, if the distribution is not normal or log-normal such methods will, obviously, provide bad results with no recourse for increasing accuracy. Second, all of these methods are limited to 3rd-order closure because analytically defined probability distributions are generally limited to one or two variables (the normal, log-normal and beta-binomial distributions are certainly limited in this way). Numerical methods, on the other hand, can generally be defined by a specific single assumption. Perhaps a central moment is set to zero (analytically[87] or numerically[85]) or derivatives are matched to an ansantz as in Hespanha[84]. This neatly avoids the second issue described above: now the closure-order can be increased without bound, while compounding the first issue. Now, not only will there be a risk that the distribution being offered does not describe the underlying data, it is extremely probable that the moments being calculated don't belong to a valid distribution at all!

Ultimately, while moment closure has been researched extensively over the past 10 years, the field itself, when this thesis project was begun, was still somewhat undirected and qualitative. Researchers each had their own proposed closure scheme, few of these schemes went beyond single-component low-order closure and rarely were the assumptions necessary for these schemes to hold true tested for applicability to biological systems in general. There was ample room and motivation for a novel closure scheme that could do several things:

- A closure-scheme that was not based on a specific distribution and thus could be applied universally to any distribution.
- A closure scheme that could be applied to multi-component systems and at high-closure order without restrictions.
- The assumptions concerning what types of reaction networks the closure scheme can be applied to should be well-established and also non-restrictive when considering biologically relevant systems.

Ultimately our zero-information closure (ZI-Closure) scheme (see Chapter 3) fits these criteria nicely. In the following chapters the topics to be discussed are:

- **Chapter 2**: In order to facilitate the development and analysis of moment closure schemes a more efficient open-source method for determining moment equations was developed. This method is described in detail along with some of the issues surrounding moment closure in general.

- **Chapter 3**: The ZI-Closure method is described in detail along with results for three models. This demonstrated the universality and accuracy of the method while also highlighting some of the drawbacks of utilizing numerical schemes for moment closure.

- **Chapter 4**: The ZI-Closure method is extended in order to perform non-linear analysis deterministically on simple stochastic networks. Eigenvalues and eigenvectors are found, and the fundamental differences between deterministic and stochastic models are discussed.

- **Chapter 5**: The non-linear analysis method is extended in order to demonstrate the calculation of exact power spectral densities of stochastic systems without the need for Monte Carlo simulations. Such calculations will have impacts on the analysis of oscillatory systems in the future.

- **Chapter 6**: A brief discussion of an earlier project concerning model reduction is described and then discussed within the context of applications of ZI-Closure in the future.

Chapter 2

Efficient Moment Matrix Generation for Arbitrary Chemical Networks

2.1 Introduction

As outlined in Chapter 1 the use of stochastic simulations in biological modeling and simulation is increasingly common due to the tendency of such systems to lie far from the thermodynamic limit[5, 18]. This condition, common when only a handful of reactants exist in a system, renders deterministic models inaccurate[91]. The primary problem with stochastic simulations is the mathematical intractability of the govering equation, the chemical master equation (CME), that makes analysis particularly difficult[92]. The motivation for the research presented within this chapter is to propose and provide an efficient tool for the analysis of stochastic systems potentially without the need of the costly kinetic Monte Carlo sampling employed by Gillespie's SSA or its derivatives[38, 56, 93, 73]. Indeed, the open-source tool presented here was used extensively for generating the results of Chapters 3-5.

An analogue to the CME proposed initially by McQuarrie, among others, is the use of time differentials of the moments of the probablility distribution[40]. In principle, this deterministic analogue allows for system analysis without the need for Monte Carlo

sampling. Such methods are in use, but predominantly in smaller systems like simple transient gene activation[94]. Complications arise when 2nd-order reactions are included in a chemical network. In such cases lower-order moments explicitly depend on higher-order moments, necessitating an infinitely large matrix[86]. The production and analysis of closure schemes for these open systems is an active area of research across several fields[82, 95, 85, 84, 78].

What is not often considered is the generation of the moment equations necessary once a particular closure scheme is chosen. Such sets of equations can become enormous considering that reproducing probability distributions accurately for complex systems have been shown to require as many as eight moments[96]. Recent publications provide analytical solutions for the moment equations[96, 90], but do not touch on the scaling problems or memory allocation of such systems. The intention of this work is to facilitate moment closure research by making matrix construction a quick and simple step. In particular, because biological systems often contain many components and exhibit complex underying behavior, a method for generating an arbitrarily large set of moments equations has become increasingly desireable.

The current work focuses on the efficient generation of a concise set of moment equations for arbitrary chemical networks. It should be noted that the method, as described, only works with elementary rate laws with integer stoichiometry. The method can generate either factorial moments (indicated by curly brackets $\{\cdot\}$) or traditional polynomial moments (indicated by angled brackets $\langle\cdot\rangle$). Factorial moments can, in principle, conserve memory allocation by forming a banded matrix. Factorial and polynomial moment formulation are entirely equivalent basis sets. The use of the probability generation function, the Z-transform of a probability distribution, also allows for a recursive algorithm as the moment equations are produced by the systematic differentiation of the Z-transformed CME (Z-CME). The results demonstrate the reduced memory load, reduced bandwidth, and computational efficiency in several example systems.

2.2 Theory

The traditional approach to chemical kinetic modeling is deterministic in nature. Stochastic models reframe chemical kinetics to take into account the randomness inherent to

systems far from the thermodynamic limit. This viewpoint requires probability-based mathematics developed by McQuarrie, Van Kampen, Gardiner, and Gillespie, among others[40, 97, 28, 41]. The theory presented here focuses on Markov processes, the CME along with its Z-transform analogue, and the moment viewpoint of chemical reaction dynamics. The primary novel work is an alternative formulation of the moment equations for arbitrary chemical networks, including the unclosed matrices for systems with bimolecular reactions. The algorithm produces both the traditional and factorial moment equation matrix for either numerical or symbolic kinetic reaction rates for any desired order of moments.

2.2.1 Markov Processes and the Chemical Master Equation

For a stochastic process to be a Markov process the current state, x_n at time t_n, must be fully determined by the previous state, x_{n-1} at time t_{n-1}. This condition is often called a "memoryless" condition, where the memory of states prior to the current state have no effect on the future dynamics. In the particular case of chemical kinetic systems the time scale is considered continuous, but the state space can be either continuous or discrete. Herein, the state space is considered to be strictly discrete by viewing the state $\underline{X} = [x_0, x_1, \ldots, x_n]'$ as a set of the number of molecules for each of the n chemical species. The state exists within a volume, Ω, that is well mixed to eliminate diffusive effects.

For stochastic processes the state can be described by a probability distribution $P(\underline{X}; t)$ at a given time t. This distribution is a continuous, bounded, n-dimensional, real-valued function. Given the Markov condition, the change in the probability distribution in time can thus be described generally by:

$$P(\underline{X}; t+\delta t) = \left(1 - \sum_{\underline{X}'} T\left(\underline{X}'|\underline{X}\right)\delta t\right) P(\underline{X}; t) + \sum_{\underline{X}'} T\left(\underline{X}|\underline{X}'\right)\delta t P\left(\underline{X}'; t\right) \qquad (2.1)$$

where $\underline{X}$ is the current state, $\underline{X}'$ is any state that is not $\underline{X}$, and $T(Y|X)$ is the transition probability from state X to Y. By rearrangement one can reach what is known as the

chemical master equation (CME) for a discrete state space:

$$\frac{\partial P\left(\underline{X};t\right)}{\partial t}=\sum_{\underline{X}'}\left[T\left(\underline{X}|\underline{X}'\right)P\left(\underline{X}';t\right)-T\left(\underline{X}'|\underline{X}\right)P\left(\underline{X};t\right)\right] \tag{2.2}$$

The transition probability is a catch-all for any event that may bring the system into or out of a given state. In chemical kinetics these events will be reactions and thus the transition probabilities are equivalent to reaction propensities. For the general reaction $aA+bB\overset{k_r}{\rightarrow}cC+dD$ the propensity for a reaction event to occur is:

$$a_r\left(\underline{X}\right)=k_r\binom{x_A}{a}\binom{x_B}{b} \tag{2.3}$$

Here x_A is the number of molecules of A in the system, and x_B is the number of molecules of B. Thus Equation 2.2 can be reformulated as a sum over the set of reactions in a chemical network:

$$\frac{\partial P\left(\underline{X};t\right)}{\partial t}=\sum_{r}\left[a_r\left(\underline{X}-\underline{\nu}_r\right)P\left(\underline{X}-\underline{\nu}_r;t\right)-a_r\left(\underline{X}\right)P\left(\underline{X};t\right)\right] \tag{2.4}$$

Here $\underline{\nu}_r$ is the stoichiometic vector for reaction r. The CME is an equation which perfectly describes the time dynamics of any stochastic chemical network given an initial condition. The problem is that, except for a select few simple equations, analytically solving such a system is mathematically intractable.

To fully describe a system by probability states Equation 2.4 would form an infinite set of ODEs. Analysis is predominately performed using the Z-CME, the CME in terms of the probability generating function (Z-transform). The Z-transform is formed by a simple change of variables:

$$G\left(\underline{S};t\right)=\sum_{x_0=0}^{\infty}\sum_{x_1=0}^{\infty}\cdots\sum_{x_n=0}^{\infty}s_0^{x_0}s_1^{x_1}\cdots s_n^{x_n}P\left(\underline{X};t\right) \tag{2.5}$$

The vector $\underline{S}=\left[s_0,s_1,\ldots,s_n\right]'$ represents the continuous transform variables for an n-component system. The Z-transform simplifies the system by transforming from a discrete space ($\underline{X}$) to a continuous space ($\underline{S}$).

The Z-CME is thus formed by mathematical manipulation of Equation 2.4:

$$\frac{\partial G\left(\underline{S};t\right)}{\partial t}=\sum_{r}\frac{k_r}{\prod_{i=1}^{n}\nu_{i,r}^{+}!}\left[\prod_{i=1}^{n}s_{i,r}^{\nu_{i,r}^{-}}-\prod_{i=1}^{n}s_{i,r}^{\nu_{i,r}^{+}}\right]\frac{\partial^{\sum_{i=1}^{n}\nu_{i,r}^{+}}G\left(\underline{S};t\right)}{\partial\prod_{i=1}^{n}s_{i,r}^{\nu_{i,r}^{+}}} \tag{2.6}$$

Here k_r is modified since the differential creates permutations instead of combinations. Thus k_r must be divided by the appropriate factorials. Also $\nu_{i,r}^{+}$ refers to product stoichiometries whereas $\nu_{i,r}^{-}$ refers to reactant stoichiometries. Both are strictly positive constants. This formulation may appear daunting, but is extremely simple to implement for any arbitrary chemical system with polynomial reaction rates (not non-linear rates). To make things clearer the Z-CME for a system with one general reaction $aA+bB\overset{k_1}{\rightarrow}cC+dD$ is:

$$\frac{\partial G\left(\underline{S};t\right)}{\partial t}=\frac{k_1}{a!\cdot b!}\left[s_C^c s_D^d-s_A^a s_B^b\right]\frac{\partial^{a+b}G\left(\underline{S};t\right)}{\partial s_A^a\partial s_B^b} \tag{2.7}$$

The advantages of using the Z-CME will become clear after a brief discussion of different types of moments available to form the full moment equation matrix.

2.2.2 Moments and Factorial Moments

Moments are expected values that describe properties of a probability distribution. The common examples are the mean (first moment) and the variance (related to the second moment). For a molecular system the distribution is defined on the positive integer line. The polynomial moments are:

$$\left\langle x_1^{m_1}\cdot x_2^{m_2}\cdots x_n^{m_n}\right\rangle=\sum_{x_1=0}^{\infty}\sum_{x_1=0}^{\infty}\cdots\sum_{x_1=0}^{\infty}x_1^{m_1}\cdot x_2^{m_2}\cdots x_n^{m_n}\cdot P\left(\underline{X};t\right) \tag{2.8}$$

When more than one chemical species has an order greater than zero these are known as joint moments. The order of the moment is defined by $m=m_1+m_2+\ldots m_n$. While many distributions can be described primarily by low-order moments (m equals one or two), more complex systems can require as many as eight or ten moments to adequately reproduce the probability distribution[96].

The literature concerning moment equations and closure schemes has focused on the traditional set of polynomial moments[90, 96, 84], but any moment basis set is valid when describing a probability distribution. In this chapter we utilize factorial moments,

an equivalent basis set. As will be shown shortly this has to do with the relationship between the Z-CME and the moments of a distribution. The factorial moments are defined as:

$$\{x_1^{m_1} \cdot x_2^{m_2} \cdots x_n^{m_n}\} = \sum_{x_1=0}^{\infty} \sum_{x_1=0}^{\infty} \cdots \sum_{x_1=0}^{\infty} (x_1)_{m_1} \cdot (x_2)_{m_2} \cdots (x_n)_{m_n} \cdot P(\underline{X}; t) \tag{2.9}$$

The form $(x_i)_{m_i}$ refers to the m_i-permutation of the variable x_i: $(x_1)_2 = x_1(x_1 - 1)$, $(x_1)_3 = x_1(x_1 - 1)(x_1 - 2)$, etc. The Z-transform of a distribution relates to the factorial moments as follows:

$$G(\underline{S}; t)|_{\underline{S}=\underline{1}} = 1$$

$$\left.\frac{\partial G(\underline{S}; t)}{\partial s_i}\right|_{\underline{S}=\underline{1}} = \{x_i\}$$

$$\left.\frac{\partial^2 G(\underline{S}; t)}{\partial s_i^2}\right|_{\underline{S}=\underline{1}} = \left\{x_i^2\right\} \tag{2.10}$$

Using the factorial moment basis there is a simple way to obtain the full set of factorial moment equations that has not been previously considered. By starting at the Z-CME (Equation 2.6) and then systematically differentiating by the elements of the $\underline{S}$ vector then setting $\underline{S} = \underline{1}$ one can obtain the time derivatives for any factorial moment desired.

2.2.3 Matrix Equation Production

The analytical derivation of polynomial moment equations has been previously developed by several authors[90, 96]. The novelty of the following approach is that the described method lends itself to efficient recursion. Higher-order moments are formed by applying an additional differentiation to a lower-order moment. In this way there is no need to re-derive information previously determined to form lower-order moments, thus saving substantial computational time. What is often not taken into consideration when producing a full analytical expression of moment equations is that it is almost never the case where only a select few moments are required. Typically, a system will require *all* moments up to a specific order (denoted as M throughout this thesis).

Providing an analyical expression does not provide an efficient way to fill a matrix of moment equations when all moments up to order M are necessary.

To expand on this last point a bit further, the available moment equation generation techniques exclusively rely on and provide an analytical equation for general moment equations[90, 96]. Sotiropoulos provided the following equation:

$$\frac{\partial \left\langle X_1^{m_1} X_2^{m_2} \cdots X_N^{m_N} \right\rangle}{\partial t} = \sum_{j_1,j_2,\ldots,j_N=0}^{m_1,m_2,\ldots,m_N} \binom{m_1}{j_1}\binom{m_2}{j_2}\cdots\binom{m_N}{j_N} \cdot \left\langle X_1^{m_1-j_1} X_2^{m_2-j_2} \cdots X_N^{m_N-j_N} a_{j_1+j_2+\cdots+j_N}\left(\underline{X}\right)\right\rangle \quad (2.11)$$

where

$$a_{j_1+j_2+\cdots+j_N}\left(\underline{X}\right) = \sum_r \left[\prod_{i=1}^{N} \nu_{i,k}^{j_i}\right] a_r\left(\underline{X}\right) \quad (2.12)$$

There are three issues with this formulation. First, the summation is of indeterminate size. When there are N components there are N summations. When building a matrix generation program this issue makes implementation difficult. Second, the propensity term, $a^{index}_{j_1+j_2+\cdots+j_N}\left(\underline{X}\right)$, does not have a computationally friendly form. Rather, it is a complicated equation with multiple binomial terms and stoichiometric considerations (Equations 22 and 29 in [96]). Finally, as mentioned above, there is information waste in construction of the moment matrix that becomes costly as the number of components and moments increase. Note that the analytical form is not recursive: there is no intutive way to construct the higher-order moments from the moment equations previously determined. The proposed method may not provide an analytical equation for the moment equations, but by utilizing the Z-CME and factorial moments it does avoid this last problem in particular to produce an efficient recursive algorithm.

Using the Z-CME differentials can produce the moment equations. A quick application to Equation 2.7 illustrates this point. The left side of the equation is simple:

$$\frac{\partial}{\partial s_A}\frac{\partial G\left(\underline{S};t\right)}{\partial t}\bigg|_{\underline{S}=\underline{1}} = \frac{\partial}{\partial t}\frac{\partial G\left(\underline{S};t\right)}{\partial s_A}\bigg|_{\underline{S}=\underline{1}} = \frac{\partial \{A\}}{\partial t}$$

The full transformed equation before simplification is:

$$\frac{\partial\{A\}}{\partial t}=\left[-\frac{k_1}{a!\cdot b!}\left[a\cdot s_A^{a-1}s_B^b\right]\frac{\partial^{a+b}G\left(\underline{S};t\right)}{\partial s_A^a\partial s_B^b}+\ \frac{k_1}{a!\cdot b!}\left[s_C^c s_D^d-s_A^a s_B^b\right]\frac{\partial^{a+b+1}G\left(\underline{S};t\right)}{\partial s_A^{a+1}\partial s_B^b}\right]\Bigg|_{\underline{S}=\underline{1}}$$

The reduced form then contains a single term:

$$\frac{\partial\{A\}}{\partial t}=-\frac{k_1}{(a-1)!\cdot b!}\left\{A^aB^b\right\} \tag{2.13}$$

It is easy to then produce $\partial_t\left\{A^2\right\}$ by applying a second differential. The final expression would be:

$$\frac{\partial\left\{A^2\right\}}{\partial t}=-\frac{k_1}{(a-2)!\cdot b!}\left\{A^aB^b\right\}-2\cdot\frac{k_1}{(a-1)!\cdot b!}\left\{A^{a+1}B^b\right\} \tag{2.14}$$

This method is applicable to any arbirary set of chemical reactions.

A concern when producing moment matrices and moment closure schemes is a standardized and efficient indexing scheme for moments. Here it is proposed that the ideal indexing will achieve one main goal: keep the system as banded as possible. In the case of factorial moments the easiest way to band the matrix is to keep the moments of order-m together. In elementary chemical systems the stoichiometry of a chemical component in a reaction will rarely exceed an absolute value of two. In these cases then the factorial moment equations of order-m will *only* depend on the moments of order-(m-1), order-m, or order-(m+1). This tends to create systems with lower bandwidths. The only other consideration would be to put chemical species which are not involved in any common reactions apart. Within each order the moments are simply indexed in descending order for each of the components. For example, with three components the third-order moments are ordered: $\left[\left\langle x_1^3\right\rangle,\left\langle x_1^2x_2\right\rangle,\left\langle x_1^2x_3\right\rangle,\left\langle x_1x_2^2\right\rangle,\left\langle x_1x_2x_3\right\rangle,\left\langle x_1x_3^2\right\rangle,\right.$ $\left\langle x_2^3\right\rangle,\left\langle x_2^2x_3\right\rangle,\left\langle x_2x_3^2\right\rangle,\left.\left\langle x_3^3\right\rangle\right]$.

The final point to make is how the factorial moment matrix produced here relates to the polynomial moment matrix typically considered in the literature. The factorial moment matrix takes the form:

$$\frac{\partial\underline{\mu}_f}{\partial t}=\left[A_f|A_f'\right]\left[\begin{array}{c}\underline{\mu}_f\\ \underline{\mu}_f'\end{array}\right] \tag{2.15}$$

The vector $\underline{\mu}_f$ represents all of the factorial moments up to order M, and $\underline{\mu}'_f$ represented the higher-order moments necessary to close the system. The number of moments up to order-M for N components is the same as the number of terms in a general polynomial with N variables and up to order-M:

$$\text{Number of Terms} = N_M = \frac{(N+M)!}{N! \cdot M!} \tag{2.16}$$

Therefore, a system with four components, as decribed above, with eight moments would require $N_M = 495$ rows. The vector $\underline{\mu}'_f$ will be of size N'_M representing the number of additional moments necessary to close the system. The size of matrix A_f is $N_M \times N_M$, and the size of the matrix A'_f is $N_M \times N'_M$. This matrix can be quite large, especially when there is a large number of components involved in a system.

The polynomial matrix takes the same form and is the same size:

$$\frac{\partial \underline{\mu}}{\partial t} = \left[A|A'\right] \begin{bmatrix} \underline{\mu} \\ \underline{\mu}' \end{bmatrix} \tag{2.17}$$

The polynomial moments and factorial moments are related through a similarity tranform. This means they have the same eigenvalues, just in a different basis set:

$$\left[A_f|A'_f\right] = T^{-1}\left[A|A'\right] S \tag{2.18}$$

The creation of the $N_M \times N_M$ transform matrices T and the $(N_M + N'_M) \times (N_M + N'_M)$ transform matrix S is rather simple and thus the programs outlined here are able to produce A_f and A, whichever is preferred.

2.3 Results and Examples

In the following section three examples of varying complexity and size are presented to show the primary advantages of the Z-CME approach. The algorithm used to produce the matrices of interest is written in Matlab. The matrices can be output for both numerical and symbolic kinetic constants. The main advantages claimed for the algorithm is speed, especially when producing high-order moments, and a reduced bandwidth and memory load in most cases. The three examples will be a simple reversible dimerization ($2A \leftrightarrow B$), a Michaelis-Menten system ($S + E \leftrightarrow S : E \rightarrow E + P$), and a gene

regulatory network (9 components, 10 reactions). In all cases the matrices provided will utilize symbolic kinetic constants. In order to provide a better estimate of computational efficiencies, generation times are provided for these three matrices for the recursive algorithm, an equivalent non-recursive algorithm, and for the python code described by Gillespie[90]. These represent a non-analytic recursive algorithm, a non-analytic non-recursive algorithm, and an analytic algorithm, respectively. We should stress that a direct comparison could not be rigorously developed, but we present these CPU times only to provide a sense of the computational efficiencies gained.

2.3.1 Reversible Dimerization

In reversible dimerization two monomers (A) combine in a second-order monomolecular reaction to form a dimer (B). This dimer can then split apart to reform the monomer bases. This reaction network is one of the simplest non-linear systems available with only two reactions and two components. The network and Z-CME are as follows:

$$\begin{array}{ccc} 2A & \stackrel{k_1}{\rightarrow} & B \\ B & \stackrel{k_2}{\rightarrow} & 2A \end{array}$$

$$\frac{\partial G\left(s_A, s_B; t\right)}{\partial t} = k_1\left(s_B - s_A^2\right)\frac{\partial^2 G\left(s_A, s_B; t\right)}{\partial s_A^2} + k_2\left(s_A^2 - s_B\right)\frac{\partial G\left(s_A, s_B; t\right)}{\partial s_B} \tag{2.19}$$

For moment equations up to order-2 $\left[A_f | A'_f\right]$ is:

$$\frac{\partial}{\partial t}\begin{bmatrix} \{A\} \\ \{B\} \\ \left\{A^2\right\} \\ \{AB\} \\ \left\{B^2\right\} \end{bmatrix} = \left[\begin{array}{ccccc|cc} 0 & 2k_2 & -2k_1 & 0 & 0 & 0 & 0 \\ 0 & -k_2 & k_1 & 0 & 0 & 0 & 0 \\ 0 & 2k_2 & -2k_1 & 4k_2 & 0 & -4k_1 & 0 \\ 0 & 0 & 0 & -k_2 & 2k_2 & k_1 & -2k_1 \\ 0 & 0 & 0 & 0 & -2k_2 & 0 & 2k_1 \end{array}\right] \begin{bmatrix} \{A\} \\ \{B\} \\ \left\{A^2\right\} \\ \{AB\} \\ \left\{B^2\right\} \\ \left\{A^3\right\} \\ \left\{A^2B\right\} \end{bmatrix} \tag{2.20}$$

and $[A|A']$ is:

$$\frac{\partial}{\partial t}\begin{bmatrix} \langle A\rangle \\ \langle B\rangle \\ \langle A^2\rangle \\ \langle AB\rangle \\ \langle B^2\rangle \end{bmatrix} = \left[\begin{array}{ccccc|cc} 2k_1 & 2k_2 & -2k_1 & 0 & 0 & 0 & 0 \\ -k_1 & -k_2 & k_1 & 0 & 0 & 0 & 0 \\ -4k_1 & 4k_2 & 8k_1 & 4k_2 & 0 & -4k_1 & 0 \\ 2k_1 & -2k_2 & -3k_1 & 2k_1-k_2 & 2k_2 & k_1 & -2k_1 \\ -k_1 & k_2 & k_1 & -2k_1 & -2k_2 & 0 & 2k_1 \end{array}\right] \begin{bmatrix} \langle A\rangle \\ \langle B\rangle \\ \langle A^2\rangle \\ \langle AB\rangle \\ \langle B^2\rangle \\ \langle A^3\rangle \\ \langle A^2B\rangle \end{bmatrix} \tag{2.21}$$

For the recursive algorithm the generation time for eight moments is on average 0.0502 seconds. This is compared to 0.962 seconds for the non-recursive version and 18.3 seconds for the code provided by Gillespie[90]. Figure 2.1 highlights the non-zero entries in the factorial moment matrix compared to the polynomial moment matrix for 8 moments. In all cases the factorial moment matrix will have at most as many non-zero entries as the polynomial moment matrix. Here there are 179 non-zero entres for the factorial moment case and 846 non-zero entries for the polynomial moment case. The factorial moment matrix will also have a smaller bandwidth in all cases. The bandwidth for the factorial moment matrix is 16 (7 left bandwidth, 9 right bandwidth), and for the polynomial matrix is 52 (43 left bandwidth, 9 right bandwidth).

It should be noted that the system as described has a single degree of freedom since $2 \cdot B = 2 \cdot B_0 + A_0 - A$. In later chapters the matrix is reduced to include only the moments of A as a simplification.

2.3.2 Michaelis-Menten Network

The Michaelis-Menten reaction system is comprised of three reactions. First a substrate (S) complexes with an enzyme (E) to form a complex (S:E). Then the complex either degrades back to its original components or a product is formed (P). The network has four components and three reactions:

$$\begin{aligned} S + E &\xrightarrow{k_1} E:S \\ E:S &\xrightarrow{k_2} S + E \\ E:S &\xrightarrow{k_3} E + P \end{aligned}$$

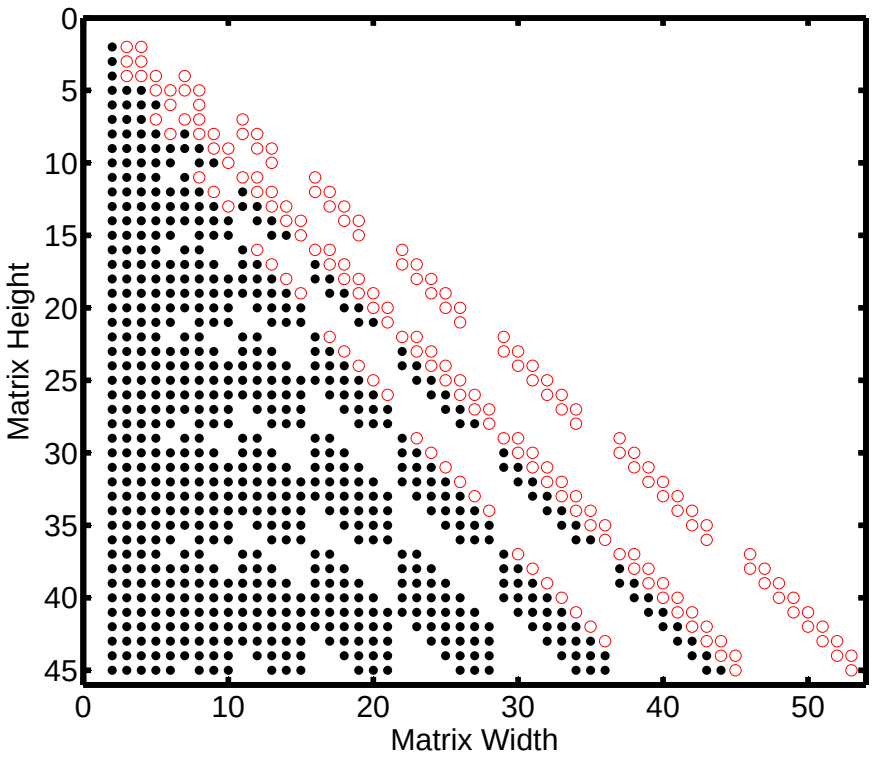

Figure 2.1: Comparison of the matrix form for the factorial moment equations and polynomial moment equations for reversible dimerization up to eight moments. The non-zero entries present in both matrices are represented by red empty circles, while the non-zero entries present only in the polynomial matrix are represented by black dots.

This system has only two degrees of freedom. Using the following relations the system is reduced to two components (S and E): $S+E:S+P=S_T$ and $E+E:S=E_T$. Using these relations the reduced Z-CME can be determined to be:

$$\begin{aligned}\frac{\partial G\left(\underline{S};t\right)}{\partial t} &= k_1\left(1-s_S s_E\right)\frac{\partial^2 G\left(\underline{S};t\right)}{\partial s_S \partial s_E} \\ &+ k_2 E_T\left(s_S s_E-1\right)G\left(\underline{S};t\right)-k_2\left(s_S s_E^2-s_E\right)\frac{\partial G\left(\underline{S};t\right)}{\partial s_E} \\ &+ k_3 E_T\left(s_E-1\right)G\left(\underline{S};t\right)-k_3\left(s_E^2-s_E\right)\frac{\partial G\left(\underline{S};t\right)}{\partial s_E}\end{aligned} \tag{2.22}$$

For moment equations up to order-2 $\left[A_f|A'_f\right]$ is:

$$\frac{\partial}{\partial t}\begin{bmatrix}\{1\}\\ \{S\}\\ \{E\}\\ \{S^2\}\\ \{SE\}\\ \{E^2\}\end{bmatrix} = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 \\ E_T k_2 & 0 & -k_2 & 0 & -k_1 & 0 \\ E_T(k_2+k_3) & 0 & -k_2-k_3 & 0 & -k_1 & 0 \\ 0 & 2E_T k_2 & 0 & 0 & -2k_2 & 0 \\ E_T k_2 & E_T(k_2+k_3) & (E_T-2)k_2 & 0 & -k_1-k_2-k_3 & -k_2 \\ 0 & 0 & 2(E_T-1)(k_2+k_3) & 0 & 0 & -2k_2-2k_3 \end{array}\right]\begin{bmatrix}\{1\}\\ \{S\}\\ \{E\}\\ \{S^2\}\\ \{SE\}\\ \{E^2\}\\ \{S^3\}\\ \{S^2E\}\\ \{SE^2\}\end{bmatrix} \tag{2.23}$$

and $[A|A']$ is:

$$\frac{\partial}{\partial t}\begin{bmatrix}\langle 1\rangle\\ \langle S\rangle\\ \langle E\rangle\\ \langle S^2\rangle\\ \langle SE\rangle\\ \langle E^2\rangle\end{bmatrix} = \left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0 & 0\\ E_T k_2 & 0 & -k_2 & 0 & -k_1 & 0\\ E_T(k_2+k_3) & 0 & -k_2-k_3 & 0 & -k_1 & 0\\ E_T k_2 & 2E_T k_2 & -k_2 & 0 & k_1-2k_2 & 0\\ E_T k_2 & E_T(k_2+k_3) & (E_T-1)k_2 & 0 & k_1-k_2-k_3 & -k_2\\ E_T(k_2+k_3) & 0 & 2(E_T-1)(k_2+k_3) & 0 & k_1 & -2k_2-2k_3 \end{array}\right] \begin{bmatrix}\langle 1\rangle\\ \langle S\rangle\\ \langle E\rangle\\ \langle S^2\rangle\\ \langle SE\rangle\\ \langle E^2\rangle\\ \langle S^3\rangle\\ \langle S^2E\rangle\\ \langle SE^2\rangle\end{bmatrix} \tag{2.24}$$

For the non-reduced set of reactions the recursive algorithm generation time for eight moments is on average 0.126 seconds. This is compared to 21.9 seconds for the non-recursive version and 37.2 seconds for the code provided by Gillespie[90]. Figure 2.2 highlights the non-zero entries in the factorial moment matrix compared to the polynomial moment matrix for 8 moments. There are 244 non-zero entires for the factorial moment case and 762 non-zero entries for the polynomial moment case. The bandwidth for the factorial moment matrix is 26 (16 left bandwidth, 10 right bandwidth), and for the polynomial matrix is 54 (44 left bandwidth, 10 right bandwidth).

2.3.3 Gene Network Model

While the previous two examples effectively show the memory conserving and bandwidth reducing advantages of the method, both were simple enough to require little time to complete. To demonstrate the final advantage, an ability to produce large matrices quickly, a larger network was chosen, that of a single gene model. The network consists of ten reactions involving nine components presented in Table 2.1.

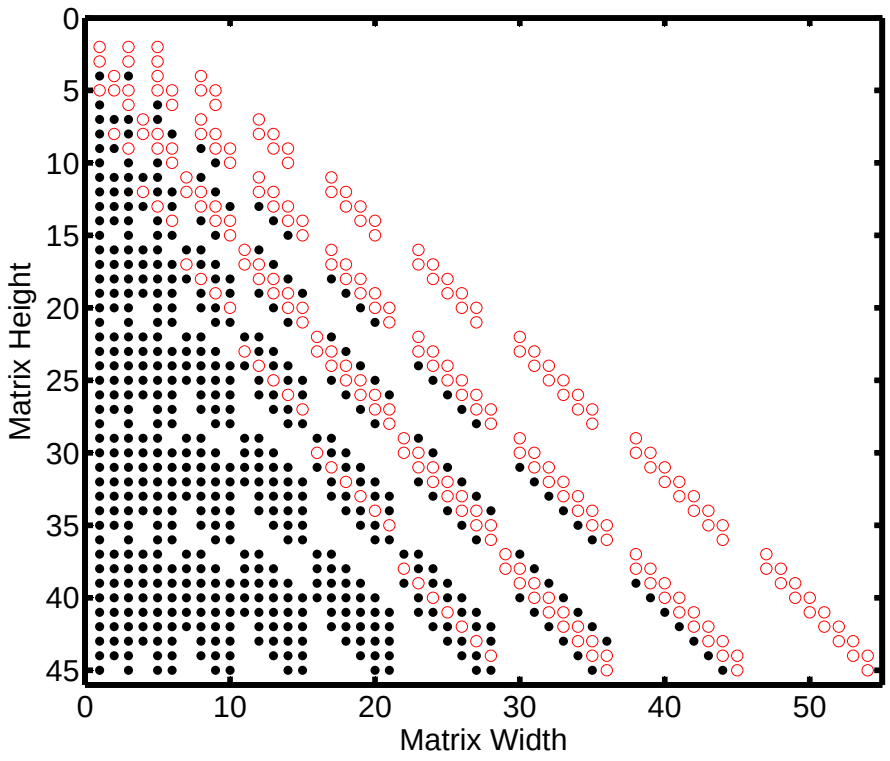

Figure 2.2: Comparison of the matrix form for the factorial moment equations and polynomial moment equations for a Michaelis-Menten network up to eight moments. The non-zero entries present in both matrices are represented by red empty circles, while the non-zero entries present only in the polynomial matrix are represented by black dots.

Table 2.1: Gene Network Model

DNA + RNAp	$\rightarrow$	DNA:RNAp
DNA:RNAp	$\rightarrow$	DNA + RNAp
DNA:RNAp	$\rightarrow$	DNA + RNAp*
RNAp*	$\rightarrow$	RNAp + mRNA
mRNA + rib	$\rightarrow$	mRNA:rib
mRNA:rib	$\rightarrow$	mRNA + rib
mRNA:rib	$\rightarrow$	mRNA + rib*
rib*	$\rightarrow$	rib + Product
mRNA	$\rightarrow$	$\emptyset$
Product	$\rightarrow$	$\emptyset$

A two-moment matrix for nine components requires 55 rows and 194 columns. For the non-reduced set of reactions the recursive algorithm generation time for four moments is on average 0.251 seconds. This is compared to 12.8 seconds for the non-recursive version and 152.6 seconds for the code provided by Gillespie[90]. For eight moments, the non-recursive and Gillespie's code both run out of memory, but the recursive algorithm takes 105 seconds to complete. The Z-CME and two-moment matrix are not shown for size considerations. Figure 2.3 shows the non-zero entries for 8 moments. The factorial moment matrix is a 24310 by 48355 matrix, a large matrix that pushes the memory limits of Matlab when produced. The program takes about two minutes to produce the symbolic matrix on a desktop computer. The transform matrix is larger (48355 by 48355) and takes three minutes to produce. The number of non-zero entires in the factorial moment matrix is 237299, whereas for the polynomial matrix there are 1439085 non-zero entries. This matrix is far larger than is currently feasable for methods utilizing the moments of a stochastic simulation.

2.4 Discussion

The first two reaction networks presented (Dimerization in Figure 2.1 and the Michaelis-Menten system in Figure 2.2) demonstrate advantages in bandwidth and memory allocation for the factorial moment matrix, $\left[A_f | A'_f\right]$, compared to the polynomial moment matrix, $[A|A']$, present in the literature. In the dimerization system the factorial matrix is banded with three main bands corresponding to the (m-1)-order, m-order, and

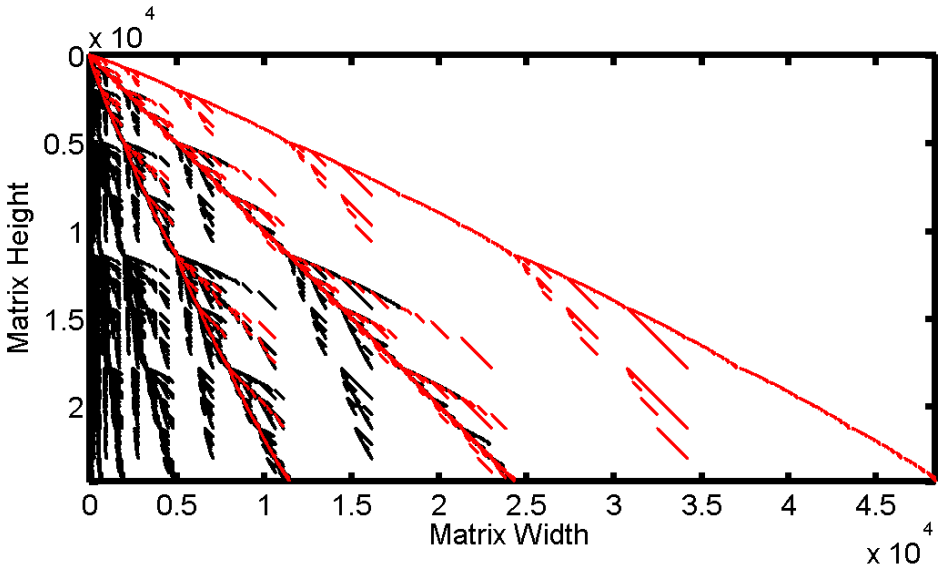

Figure 2.3: Comparison of the matrix form for the factorial moment equations and polynomial moment equations for a nine component gene network up to eight moments. The non-zero entries present in both matrices are represented by red empty circles, while the non-zero entries present only in the polynomial matrix are represented by black dots.

(m+1)-order moments. The polynomial moment matrix is nearly lower-triangular in nature and contains over five times as many non-zero entries. Similarly the Michaelis-Menten system has four main bands in the factorial moment matrix. The fourth band is an (m-2)-order moment resulting from the reduction using the conservation equation $E + E : S = E_T$. The polynomial moment matrix has about three times as many non-zero entries and is nearly lower triangular. Both systems take a negligible amount of time to produce up to order-eight moments. The important point is that the resulting matrix will have at most the same number of non-zero entries and bandwidth as the polynomial matrix, never more. As far as memory allocation and bandwidth reduction are concerned the factorial moment equations are unequivocally better than traditional polynomial moment equations in all elementary chemical networks. Although it would be useful, we do not have a rigorous mathematical means for providing scaling arguments that show precisely how much sparser the factorial moment matrix is compared to the polynomial moment matrix or to what degree these matrices can take advantage of sparse algorithms.

The final system (Gene network, Figure 2.3) demonstrates the most impressive aspect of the method presented, its efficiency. The matrix produced is massive with over 10^9 entries in the system, pushing the limit of memory allocation in many programs. The production of such a set of moment equations takes less than five minutes in total and shows that even in this extreme case the system does not require a prohibitively long period of time to run.

This final network also illustrates one of the drawbacks of utilizing the moments matrix in stochastic simulation for complex systems. Some systems will require four, six, or even eight moments to produce accurate results. For systems with, say, nine components this obviously becomes prohibitive in the analytical sense. The number of equations in deterministic systems scale $O(N)$ where N is the number of components. On the other hand, the number of equations in stochastic systems will scale $O\left(N^M\right)$ where M is the maximum order of moments. In Chapters 3 and 4 our methods are limited to small reaction networks partly because of this scaling issue inherent to moment based methods. Any analysis using moment equations will have to consider the number of components and moments necessary for accurate results. Model reduction techniques (see Chapter 6) will be necessary for the practical application of moment closure.

2.5 Conclusion

The results provided in the previous section illustrate two main points: 1) The factorial moment basis produces a lower bandwidth and less memory intensive set of equations than the polynomial moment basis and in many cases should be preferentially chosen for analysis. Indeed the analysis shown in Chapters 3-5 uses this form of moment equations exclusively; 2) The use of the transforms of the CME (here the Z-CME, but other transforms could also be used) allow for efficient equation production through application within a recursive algorithm. These results show that for simple systems production of the set of equations necessary for analysis is not a time-consuming step.

Moment closure is the primary problem in the way of using such systems for stochastic analysis. The results as presented facilitate exploration of higher-order closure schemes. Such schemes are necessary in the case of complex probability systems, like bimodal systems, that are impossible to duplicate with anything less than six moments. In Chapter 2 a general purpose closure scheme is presented that can be applied at any order and on arbitrary chemical networks. The method presented here was vitally important in that work, and will be crucial in future improvements to the algorithm and moment closure in general.

2.6 Availability and Requirements

https://sourceforge.net/projects/multikin/

The code necessary to produce moment matrices efficiently is open-source and available on the sourceforge website listed above (the MomentsMat.zip file). All code was generated on Matlab R2009b and a version as or more up to date may be necessary. The files provided work together and should be housed in the same directory.

Chapter 3

On a Closure Scheme for Chemical Master Equations

3.1 Introduction

Mathematical models of biological systems are generally founded on determinism and may be excellent approximations of reality when the number of molecules is very large, approaching the limit of an infinitely sized molecular population[37, 98, 99, 100, 101]. As established in Chapter 1, however, the size of biomolecular systems is far from infinite. We know that the behavior of a few molecules fluctuating from the average in unexpected ways may significantly impact the dynamics of a biological system. It has thus been commonly recognized that models of small, evolving molecular populations better account for the noisy, probabilistic nature of outcomes[49, 102, 103].

Chapters 1 and 2 introduced the concept of the chemical master equation (CME)[104]. The "master" in the name reflects the all-encompassing nature of an equation that purports to govern all possible outcomes for all time. Because of its ambitious character, the master equation has remained unsolved for all but the simplest of molecular interaction networks, even though it is now over 50 years since the first master equations were set up for chemical systems[30, 39]. In this chapter we present a numerical solution to master equations of small chemical or biochemical reaction networks.

Generally, for a system of N molecular species, the state of the system is described by an N-dimensional vector $\underline{X} = [X_1, X_2, \ldots, X_N]$, where X_i is the number of molecules

of species i. The master equation governs the evolution of the probability, $P\left(\underline{X};t\right)$, that the system is at state $\underline{X}$ at time t:

$$\frac{\partial P\left(\underline{X};t\right)}{\partial t}=\sum_{\underline{X}'}\left[T\left(\underline{X}|\underline{X}'\right)P\left(\underline{X}';t\right)-T\left(\underline{X}'|\underline{X}\right)P\left(\underline{X};t\right)\right] \tag{3.1}$$

This is a probability conservation equation, where $T\left(\underline{X}|\underline{X}'\right)$ is the transition propensity from any possible state $\underline{X}'$ to state $\underline{X}$ per unit time. In a network of chemical or biochemical reactions, the transition probabilities are defined by the reaction rate laws and dictate how many reaction events take place per unit time.

In Chapter 2 we recast the CME in equivalent terms of probability moments[105]; the probability distribution average, the variance, and so on:

$$\frac{\partial\underline{\mu}}{\partial t}=A\underline{\mu}+A'\underline{\mu}' \tag{3.2}$$

The vector $\underline{\mu}$ is the set of lower-order moments of length $N_M=\binom{N+M}{M}$ and $\underline{\mu}'$ is the set of higher-order moments necessary for closure. If m additional orders of moments are needed for closure then $\underline{\mu}'$ is of size $N'_M=\binom{N+M+m-1}{M+m}$. Typically, m is equal to one if second-order reactions are present, two if third-order reactions are present, etc. Using this notation A is an $N_M\times N_M$ matrix and A' is an $N_M\times N'_M$ matrix.

For linear systems with only 0th or 1st-order reactions, A' is empty. For other systems, A' is not empty and the set of ordinary differential equations (ODEs) becomes infinite, and thus intractable. Note that Equation 3.2 assumes reaction networks are comprised of only polynomial reaction rates laws, as described in Chapter 2[105].

In order to use Equation 3.2, a closure scheme must be defined. A closure scheme approximates the infinite set of moment equations with a finite one that accepts a solution. Typically, a closure scheme is an approximation:

$$\underline{\mu}'=F\left(\underline{\mu}\right) \tag{3.3}$$

where F is a function that uses the lower-order moments to approximate the higher-order moments. There have been numerous attempts to define F, either by assuming an underlying distribution[90, 75, 82] or through numerical approximation[88, 80, 85]. However, closure schemes thus far exhibit limited accuracy and uncertain utility.

The motivation for the development of closure schemes is to establish a deterministic alternative to Gillespie's stochastic simulation algorithm (SSA). The SSA is a

kinetic Monte Carlo method that generates ensembles of stochastic trajectories in state space[97]. Although this algorithm and the numerous improvements have found ample use by an ever-widening community[56, 72], this approach becomes cumbersome when reaction rates span multiple time scales. Furthermore, this approach does not facilitate important analysis methods, such as steady-state and stability analysis, or perturbation and bifurcation analysis, which find use in the study of evolving molecular systems.

Here we establish a closure scheme that is accurate and may be implemented on any chemical reaction network. The proposed method affords the determination of how reaction networks evolve in time, when away from the thermodynamic limit i.e., when the molecular population size is infinite - offering an alternative to kinetic Monte Carlo sampling methods. Perhaps more importantly, this formulation facilitates the calculation of steady-state probability distributions of reaction networks without resorting to dynamic simulations. As such, it may facilitate the type of analysis of dynamic trajectories or steady states that is either impossible or impractical using kinetic Monte Carlo techniques. It should be noted that the present incarnation of the method, used throughout the rest of this thesis, necessitates the determination of probabilities at all relevant states in the state space and thus scales poorly as the reachable state space expands.

3.2 The Theoretical basis of Zero-Information Closure

For the sake of brevity, we limit the discussion in this section to one-dimensional state spaces. In particular, for a single random variable that can attain a discrete set of values, $(x_1, x_2, \dots)$, each with probability $P(x_i)$, the information entropy is defined as[106]:

$$H = -\sum_i P(x_i) \ln P(x_i) \tag{3.4}$$

We conjecture that a finite number of probability moments may capture all the information needed to precisely construct the probability distribution. Consequently, we maximize the information entropy under the constraints of the first M moment definitions:

$$\Lambda = H - \alpha_0 g_0 - \alpha_1 g_1 - \cdots - \alpha_M g_M \tag{3.5}$$

$$g_0 = \left[\sum_{x=0}^{\infty} p\left(x\right)\right] - 1$$

$$g_1 = \left[\sum_{x=0}^{\infty} xp\left(x\right)\right] - \langle x \rangle$$

$$\vdots$$

$$g_M = \left[\sum_{x=0}^{\infty} x^M p\left(x\right)\right] - \left\langle x^M \right\rangle \tag{3.6}$$

where α_j is the Lagrange multiplier of the j_{th} moment constraint. These are readily computed with appropriate root-finding numerical methods, such as the simple Newton-Raphson described in Section 3.2.1.

Taking Equations 3.5 and 3.6, the maximum is found by differentiating by p(x) and setting the result to zero[107]:

$$\frac{\partial \Lambda}{\partial p\left(x\right)} = -\ln p\left(x\right) - 1 - \alpha_0 - \alpha_1 x - \cdots - \alpha_M x^M = 0 \tag{3.7}$$

$$p_H\left(x\right) = exp\left(-1 - \alpha_0 - \alpha_1 x - \cdots - \alpha_M x^M\right) \tag{3.8}$$

An analytical expression for the maximum entropy distribution is determined with the same number of parameters as the number of known lower-order moments. The derivation of Equation 3.8 is easily extended to multi-component systems and continuous state-space systems.

Using the lower-order moments, $\underline{\mu}$, the maximum entropy distribution is found by determining the Lagrange parameters, $\underline{\alpha}$. The maximum-entropy distribution, denoted as $p_H\left(x\right)$, can be determined with the known lower-order moments ($\underline{\mu}$). Take, for example, the determination of $\langle x^m \rangle_H$ for any arbitrarily high m in a single component system:

$$\langle x^m \rangle_H = \sum_{x=0}^{\infty} x^m p_H\left(x\right) \tag{3.9}$$

Once again, this method may be trivially extended to multi-component systems.

Zero-information closure uses maximum-entropy moments as approximations for higher-order moments in simulation:

$$\underline{\mu}' = \underline{\mu}'_H \tag{3.10}$$

Note that the order of closure, M, used throughout the section is applied without drawing any conclusions about accuracy. At present the order of closure is chosen in a trial-and-error fashion. Further research into error analysis may illuminate the necessary or optimal closure order using ZI-Closure.

3.2.1 Algorithm 1: Newton-Raphson Method

The method used within this chapter for the determination of the Lagrange parameters, $\underline{\alpha}$, given a set of known lower-order moments, $\underline{\mu}$, is a simple Newton-Raphson optimization scheme. The pseudo-algorithm is provided below:

1. Initial Lagrange parameters guess, $\underline{\alpha} = \underline{\alpha}_0$, is given. A range for each component is provided along with the values for the known moments that need to be matched, $\underline{\mu}$. Typically, the guess will be a set of zeros, although if a previous step in the simulation determined the parameters $\underline{\alpha}$ with reasonable accuracy those values can be used to speed up the simulation.

2. Calculate $p_H(x)$ using $\underline{\alpha}$ (Equation 3.8)

3. Calculate the lower-order maximum entropy moments, $\underline{\mu}_H$, using $p_H(x)$ (Equation 3.9)

4. Calculate the difference between the known moments and maximum entropy moments:
$$\Delta\underline{\mu} = \underline{\mu} - \underline{\mu}_H \tag{3.11}$$

5. Calculate the 2-norm error:
$$\epsilon = \Delta\underline{\mu}^T \Delta\underline{\mu} \tag{3.12}$$

6. If $\epsilon \leq$ tolerance proceed to (7). The tolerance was assumed to be the machine error ($1 \cdot 10^{-16}$) in this case. Else:

(a) Calculate the Jacobian matrix J:

$$J_{i,j} = \frac{\partial \mu_{H,i}}{\partial \alpha_j} \tag{3.13}$$

For a single component system the result is trivial (substituting Equation 3.8 into Equation 3.9) and is easily extended to multi-component systems:

$$J_{i,j} = \frac{\partial}{\partial \alpha_j}\left[\sum_{x=0}^{\infty} x^i exp\left(-1-\alpha_0-\alpha_1 x-\cdots-\alpha_M x^M\right)\right] = -\left\langle x^{i+j}\right\rangle_H \tag{3.14}$$

(b) As with any Newton-Raphson system the method uses a first-order Taylor expansion:

$$\Delta\underline{\mu} \approx J\Delta\underline{\alpha} \tag{3.15}$$

Thus an approximate parameter step is calculated:

$$\Delta\underline{\alpha} = J^{-1}\Delta\underline{\mu} \tag{3.16}$$

(c) This is used to get a new set of parameters:

$$\underline{\alpha} = \underline{\alpha} + \Delta\underline{\alpha} \tag{3.17}$$

(d) Return to (2)

7. Output $\underline{\alpha}$ for use in calculating the higher-order moments $\underline{\mu}'_H$ using Equation 3.9

This basic Newton-Raphson scheme is used by the ODE time trajectory algorithm in determining the higher-order moments. This method is also modified to produce steady-state distributions in Section 3.3.1.

Throughout this chapter the lower-order moments in the simulation are often described as being known. Please note that while the values of the lower-order moments are not known through time *a priori*, at each time step these values will have been determined. This information is then utilized within the Newton-Raphson optimization scheme. The results of the ODE solving method will be, in fact, these lower order moments.

3.2.2 The ODE Method for Moment Determination

Equations 3.2 are now closed and can be integrated to evolve the probability moments by a time step Δt. Given an initial condition for the value of the probability moments, the system may be propagated in time. With newly calculated moments up to order M, the information entropy is maximized again, generating new values for the Lagrange multipliers, and so on. Herein we use the adaptive time step, 5th-order Runge-Kutta ODE solver (ode15s) built into MATLAB.

The following is a simple ODE solving method provided to illustrate how Algorithm 1 allows for the closure of Equation 3.2 and the determination of moment time trajectories.

1. Define $t = t_0$ and $\underline{\mu} = \underline{\mu}_0$. The moment matrices A and A' are determined via the reaction network. The higher-order moments $\underline{\mu}' = \underline{\mu}'_0$ are initially known as well. Ranges for the system components are necessary for the Newton-Raphson optimization.

2. The first step is calculated with Equation 3.2 using the chosen ODE method obtaining $\Delta\underline{\mu}$ and Δt.

3. Step forward: $\underline{\mu} = \underline{\mu} + \Delta\underline{\mu}$ and $t = t + \Delta t$. The higher-order moments $\underline{\mu}'$ are now unknown at t.

4. If $t \geq t_{max}$ proceed to (5), else:

   (a) Use Algorithm 1 to obtain parameters $\underline{\alpha}$ from $\underline{\mu}$ at time t.

   (b) Use Equation 3.9 to obtain $\underline{\mu}'_H$ from $\underline{\alpha}$

   (c) The next step, $\Delta\underline{\mu}$ and Δt, is calculated with Equation 3.2 with the chosen ODE method. Again, Δt is adaptive to account for system stiffness and particular to the chosen solver.

   (d) Return to (3)

5. Return t, a vector of desired time points, and the corresponding vector through time for each of the lower-order moments $\underline{\mu}$.

Using this ODE-Optimization method it is possible to find dynamic results, equivalent to Gillespie's SSA, using Zero-Information Closure. The next section will explore

the possibility of finding steady-state moment values for chemical reaction networks using ZI-Closure.

3.3 The Theoretical Basis of Steady-State Determination Using ZI-Closure

Steady-state determination is structurally similar to the dynamic results described above with one key difference. In Section 3.2 lower-order moment values, $\underline{\mu}$, are known with given initial conditions (e.g. a Kronecker delta function or a Gaussian distribution). This information can then be used to obtain the desired information, the higher-order moments, $\underline{\mu}'_H$. In the case of steady-state determination the lower-order moments are instead the desired information, $\underline{\mu}_{SS}$, and must be obtained using other available information.

The information that is available is the reaction network structure, Equation 3.2. In particular, take the augmented matrix B:

$$B = \left[A|A'\right] \tag{3.18}$$

This matrix is $N_M \times (N_M + N'_M)$ and is thus under-defined. The rank of B can be at most $N_M - 1$, depending on the reaction network. Given this idealized case the null space will have a rank of $N'_M + 1$. The null space basis can be described by an $(N_M + N'_M) \times (N'_M + 1)$ matrix with $N'_M + 1$ independent moments. The set of basis vectors for the null space can typically be manipulated such that it takes the form:

$$C_{SS} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ C_{1,1} & C_{1,2} & C_{1,3} & \cdots & C_{1,N'_M+1} \\ C_{2,1} & C_{2,2} & C_{2,3} & \cdots & C_{2,N'_M+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{N_M,1} & C_{N_M,2} & C_{N_M,3} & \cdots & C_{N_M,N'_M+1} \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{3.19}$$

The entry $C(m,n)$ indicates how the mth-moment scales with the nth independent moment with respect to the null-space of matrix B. The independent moments can, for the most part, be chosen arbitrarily. We typically choose the independent moments to be the higher-order moments and the 0th moment (which is always independent). Any linear combination of the columns in C_{SS} (i.e. C_{SS} matrix multiplied by any vector) is in the null space of the matrix B.

This basis set is the additional information needed to obtain steady-states: a relationship between the lower-order moments and higher-order moments has now been established that must be satisfied by the steady-state distribution. This relationship results in a modified objective function:

$$\begin{bmatrix} \Delta\underline{\mu}_{SS} \\ \Delta\underline{\mu}'_{SS} \end{bmatrix} = \begin{bmatrix} \underline{\mu}_H \\ \underline{\mu}'_H \end{bmatrix} - \begin{bmatrix} \underline{\mu}_{SS} \\ \underline{\mu}'_{SS} \end{bmatrix} = \begin{bmatrix} \underline{\mu}_H \\ \underline{\mu}'_H \end{bmatrix} - C_{SS} \begin{bmatrix} 1 \\ \underline{\mu}'_H \end{bmatrix} \tag{3.20}$$

Note that the final form takes into account our choice of independent moments as the 0th moment (always equal to one) and the higher order moments, $\underline{\mu}'_H$. Using the modified objective function, $\Delta\underline{\mu}_{SS}$ in place of $\Delta\underline{\mu}$, the same Newton-Raphson algorithm from Section 3.2.1 can be used to obtain steady-state distributions.

When generating the results for this thesis an expanded Jacobian was used exclusively. The expanded Jacobian matrix is an $(N_M + N'_M) \times (N_M + N'_M)$ matrix whereas in Section 3.2.1 it was an $N_M \times N_M$ matrix. The higher-order moments, $\underline{\mu}'$, are now present as both matched variables and fitted parameters. We have found that using an expanded Jacobian improves convergence to the steady-state solution.

3.3.1 Algorithm 2: Steady-State Newton-Raphson Method

1. Initial parameter guess, $\underline{\alpha}_{SS} = \underline{\alpha}_{SS,0}$, is given. Ranges for the components are also provided. Typically a guess is a vector of zeros, but if a previous step gave a reasonably close distribution the output parameters can be utilized as an initial guess.

2. Calculate the null space (the matrix C_{SS}) for Equation 3.18.

3. Calculate $p_H(x)$ using $\underline{\alpha}_{SS}$ (Equation 3.8)

4. Calculate $\underline{\mu}_H$ and $\underline{\mu}'_H$ using $p_H(x)$ (Equation 3.9).

5. The calculated moments, $\underline{\mu}_H$ and $\underline{\mu}'_H$, are then compared to those given by the null space (Equation 3.19 and 3.20):

$$\begin{bmatrix} \Delta\underline{\mu}_{SS} \\ \Delta\underline{\mu}'_{SS} \end{bmatrix} = \begin{bmatrix} \underline{\mu}_H \\ \underline{\mu}'_H \end{bmatrix} - \begin{bmatrix} \underline{\mu}_{SS} \\ \underline{\mu}'_{SS} \end{bmatrix} = \begin{bmatrix} \underline{\mu}_H \\ \underline{\mu}'_H \end{bmatrix} - C_{SS} \begin{bmatrix} 1 \\ \underline{\mu}'_H \end{bmatrix}$$

6. Calculate the 2-norm error:

$$\epsilon_{SS} = \Delta\underline{\mu}_{SS}^T \Delta\underline{\mu}_{SS} \tag{3.21}$$

   Note that due to the C_{SS} form chosen $\mu'_H = \mu'_{SS}$ necessarily, so those terms can be ignored in calculating the error.

7. If $\epsilon_{SS} <$ tolerance proceed to (8). The tolerance was taken to be $1 \cdot 10^{-16}$, the machine error, but with a catch to prevent infinite loops. Else:

   (a) Calculate the expanded Jacobian, J_{SS}. This is similar to the Jacobian in Algorithm 1, but expanded to include the higher order moments, $\underline{\mu}'_H$, as rows and the corresponding null space basis vectors as new columns. Essentially, the information for determining the steady-state lower-order moments are provided by the null space. The Jacobian takes the form:

$$J_{SS} = \left[\begin{array}{c|c} \frac{\partial \underline{\mu}_H}{\partial \underline{\alpha}} & \\ & C_{SS} \\ \frac{\partial \underline{\mu}'_H}{\partial \underline{\alpha}} & \end{array}\right] \tag{3.22}$$

   The Jacobian from Algorithm 1 is present as the top-left corner of the steady-state Jacobian (Equation 4.22 and 4.23). See the example at the end of this section for a clearer view on how this steady-state Jacobian is developed.

   (b) As with any Newton-Raphson scheme the method uses a first order Taylor expansion:

$$\begin{bmatrix} \Delta\underline{\mu}_{SS} \\ \Delta\underline{\mu}'_{SS} \end{bmatrix} \approx J_{SS} \begin{bmatrix} \Delta\underline{\alpha}_{SS} \\ \Delta\underline{\mu}'_{SS} \end{bmatrix} \tag{3.23}$$

   You are not fitting the higher-order moments, but rather they are providing the null-space information.

(c) An approximate parameter step is calculated:

$$\begin{bmatrix} \Delta\underline{\alpha}_{SS} \\ \Delta\underline{\mu}'_{SS} \end{bmatrix} \approx J_{SS}^{-1} \begin{bmatrix} \Delta\underline{\mu}_{SS} \\ \Delta\underline{\mu}'_{SS} \end{bmatrix} \tag{3.24}$$

(d) A new parameter set is determined:

$$\underline{\alpha}_{SS} = \underline{\alpha}_{SS} + \Delta\underline{\alpha}_{SS} \tag{3.25}$$

(e) Return to (3)

8. Output $\underline{\alpha}_{SS}$ for use in calculating $\underline{\mu}_{SS}$ using Equation 3.9.

The pseudo-algorithms presented in the last two sections represent a method for the stochastic simulation of chemical reaction networks that is totally independent from Gillespie's stochastic simulation algorithm and its descendents. The steady-state algorithm presented within this section in particular marks, potentially, the first universal one-step method for stochastic steady-state distribution determination.

3.4 Example Problem: Reversible Dimerization

In this section a full example problem (Reversible Dimerization, see Table 3.1) will be described concerning the theoretical basis for both dynamic and steady-state moment closure using ZI-Closure. The Reversible Dimerization equations are:

$$\begin{aligned} 2A &\xrightarrow{k_1} B \\ B &\xrightarrow{k_2} 2A \end{aligned}$$

If we are looking to apply 2nd-order closure (approximating 3rd-order moments and higher) then the moment vectors are:

$$\underline{\mu} = \begin{bmatrix} \langle A^0 \rangle \\ \langle A \rangle \\ \langle A^2 \rangle \end{bmatrix} \quad \text{and} \quad \underline{\mu}' = \begin{bmatrix} \langle A^3 \rangle \end{bmatrix}$$

Note that since B can be calculated by knowing A and the initial amounts A_0 and B_0 there is only one independent component which we take to be A.

As for bounding the system the total number of monomer (A) in the system is A_0 + 2 B_0. For each reaction two monomers are consumed or produced, so the sets $A \in [0, 2, \ldots, A_0 + 2B_0]$ or $[1, 3, \ldots, A_0 + 2B_0]$ represent the two possible componant ranges for the system. At this point such considerations are important in ensuring the method operates correctly.

With the known lower-order moments, $\underline{\mu}$, the Lagrange parameters, $\underline{\alpha}$, can be determined. Since we know three moments there will be three parameters to determine:

$$\underline{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

The final entropy maximized distribution takes the form:

$$p_H(x) = exp\left(-1 - \alpha_0 - \alpha_1 x - \alpha_2 x^2\right)$$

This is determined using Algorithm 1. The most important part of this is the Jacobian which takes on a simple three by three form:

$$J = \begin{bmatrix} \frac{\partial \langle A^0 \rangle_H}{\partial \alpha_0} & \frac{\partial \langle A^0 \rangle_H}{\partial \alpha_1} & \frac{\partial \langle A^0 \rangle_H}{\partial \alpha_2} \\ \frac{\partial \langle A \rangle_H}{\partial \alpha_0} & \frac{\partial \langle A \rangle_H}{\partial \alpha_1} & \frac{\partial \langle A \rangle_H}{\partial \alpha_2} \\ \frac{\partial \langle A^2 \rangle_H}{\partial \alpha_0} & \frac{\partial \langle A^2 \rangle_H}{\partial \alpha_1} & \frac{\partial \langle A^2 \rangle_H}{\partial \alpha_2} \end{bmatrix}$$

$$J = \begin{bmatrix} -\langle A^0 \rangle & -\langle A \rangle & -\langle A^2 \rangle \\ -\langle A \rangle & -\langle A^2 \rangle & -\langle A^3 \rangle \\ -\langle A^2 \rangle & -\langle A^3 \rangle & -\langle A^4 \rangle \end{bmatrix}$$

Note the symmetry of the matrix and the necessity to determine much higher-order moments. This structure poses problems when the Jacobian is nearly singular (e.g. a Kronecker delta) or when considering systems with nearly independent components.

In order to complete the dynamic results the ODE solver requires the moment equations as well. The method for determining A and A' is outlined in detail in Chapter 2. It can be noted that since $B = (A_0 + 2B_0 A)/2$ there is only one independent component (N = 1). We are looking at 2nd-order closure (M = 2) and m is one because there is a

second-order reaction. Therefore:

$$N_M = \binom{N+M}{M} = \frac{3!}{2!1!} = 3$$

$$N'_M = \binom{N+M}{M+1} = \frac{3!}{3!0!} = 1$$

Therefore A will be a 3 × 3 matrix, and A' will be a 3 × 1 matrix. Let $S_0 = A_0 + 2B_0$, and the two matrices can be determined to be:

$$A = \begin{bmatrix} 0 & 0 & 0 \\ k_2 S_0 & -k_2 & -k_1 \\ k_2 S_0 & 2k_2 S_0 - 3k_2 & -k_1 - 2k_2 \end{bmatrix} \quad \text{and} \quad A' = \begin{bmatrix} 0 \\ 0 \\ -2k_1 \end{bmatrix}$$

The steady-state results are obtained in a slightly different manner. First of all we operate on the augmented matrix B in Equation 3.18:

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 \\ k_2 S_0 & -k_2 & -k_1 & 0 \\ k_2 S_0 & 2k_2 S_0 - 3k_2 & -k_1 - 2k_2 & -2k_1 \end{bmatrix}$$

The null space is thus simple to obtain:

$$C_{SS} = \begin{bmatrix} 1 & 0 \\ \frac{S_0 k_2}{k_2 + k_1(S_0 - 1)} & \frac{k_1^2}{k_2(k_2 + k_1(S_0 - 1))} \\ \frac{S_0 k_2 (S_0 - 1)}{k_2 + k_1(S_0 - 1)} & \frac{-k_1}{k_2 + k_1(S_0 - 1)} \\ 0 & 1 \end{bmatrix}$$

C_{SS} has two columns, one corresponding to the 0th moment (necessarily independent since it is invariant), and one corresponding to the single higher-order moment necessary for closure, $\langle A^3 \rangle$. Only the second column is used in the expanded Jacobian.

The null basis vector corresponding to $\langle A^3 \rangle$ is incorporated into the expanded Jacobian by including $\langle A^3 \rangle$ as both a matched variable and parameter. So the variables to be matched are: $[\langle A^0 \rangle_{SS}, \langle A \rangle_{SS}, \langle A^2 \rangle_{SS}, \langle A^3 \rangle_{SS}]$, and the parameters to be fit

are: $[\alpha_0, \alpha_1, \alpha_2, \langle A^3 \rangle_{SS}]$:

$$J_{SS} = \begin{bmatrix} \frac{\partial \langle A^0 \rangle_{SS}}{\partial \alpha_0} & \frac{\partial \langle A^0 \rangle_{SS}}{\partial \alpha_1} & \frac{\partial \langle A^0 \rangle_{SS}}{\partial \alpha_2} & \frac{\partial \langle A^0 \rangle_{SS}}{\partial \langle A^3 \rangle_{SS}} \\ \frac{\partial \langle A \rangle_{SS}}{\partial \alpha_0} & \frac{\partial \langle A \rangle_{SS}}{\partial \alpha_1} & \frac{\partial \langle A \rangle_{SS}}{\partial \alpha_2} & \frac{\partial \langle A \rangle_{SS}}{\partial \langle A^3 \rangle_{SS}} \\ \frac{\partial \langle A^2 \rangle_{SS}}{\partial \alpha_0} & \frac{\partial \langle A^2 \rangle_{SS}}{\partial \alpha_1} & \frac{\partial \langle A^2 \rangle_{SS}}{\partial \alpha_2} & \frac{\partial \langle A^2 \rangle_{SS}}{\partial \langle A^3 \rangle_{SS}} \\ \frac{\partial \langle A^3 \rangle_{SS}}{\partial \alpha_0} & \frac{\partial \langle A^3 \rangle_{SS}}{\partial \alpha_1} & \frac{\partial \langle A^3 \rangle_{SS}}{\partial \alpha_2} & \frac{\partial \langle A^3 \rangle_{SS}}{\partial \langle A^3 \rangle_{SS}} \end{bmatrix}$$

Because C_{SS} describes a relationship between the lower-order moments and higher-order moments it can be used for the last column in the Jacobian. The rest are found in the same manner as Equation 4.23:

$$J_{SS} = \begin{bmatrix} -\langle A^0 \rangle & -\langle A \rangle & -\langle A^2 \rangle & 0 \\ -\langle A \rangle & -\langle A^2 \rangle & -\langle A^3 \rangle & \frac{k_1^2}{k_2(k_2+k_1(S_0-1))} \\ -\langle A^2 \rangle & -\langle A^3 \rangle & -\langle A^4 \rangle & \frac{-k_1}{k_2+k_1(S_0-1)} \\ -\langle A^3 \rangle & -\langle A^4 \rangle & -\langle A^5 \rangle & 1 \end{bmatrix}$$

This section should provide anyone interested in the specifics of the method an understanding of how this can be applied to any arbitrary reaction network of interest.

3.5 Models

To illustrate the utility of ZI-Closure in generating both dynamic trajectories and steady-state results, we investigate three models, described in Table 3.1. Model 1 represents a simple reversible dimerization reaction network with a second order reaction. There is a single independent component, A, as conservation arguments can be used to eliminate B. Model 2 represents a Michaelis-Menten reaction network[108]. There are two independent components, the substrate S and the enzyme E. Model 3 represents the Schlögl model[24], a four reaction network that can produce bimodal distributions. There is a single free component, X, and two reservoirs, A and B, assumed constant. It may be noted that the values for A and B are incorporated into the first and third reaction rate constants whenever specified. All systems are considered isothermal and comprised of a well mixed volume of 10^{-15} L, a typical size for common bacteria. In

Table 3.1: Model descriptions for ZI-Closure.

Model	(1) Reversible Dimerization	(2) Michaelis-Menten	(3) Schlögl
Reactions	$2A \xrightarrow{k_1} B$ $B \xrightarrow{k_2} 2A$	$S+E \xrightarrow{k_1} S:E$ $S:E \xrightarrow{k_2} S+E$ $S:E \xrightarrow{k_3} P+E$ $P \xrightarrow{k_4} S$	$2X+A \xrightarrow{k_1} 3X$ $3X \xrightarrow{k_2} 2X+A$ $B \xrightarrow{k_3} X$ $X \xrightarrow{k_4} B$
Degrees of freedom	A	S, E	X
Initial condition	$A_0 = 10$	$S_0 = 10$; $E_0 = 10$	$X_0 = 25$

all cases the initial conditions in Table 3.1 define the initial probability distributions as Kronecker delta functions. For example, for Model 1, $P(A; t = 0) = \delta_{A,10}$. Comparisons of dynamic and steady state results are made between the ZI-Closure method and the widely used SSA improvement prescribed by Gibson and Bruck[56].

While it is understood that the chosen models are simple, the presence of second and higher order reactions necessitates moment closure (i.e., matrix A' is not empty). Here, we use these models to show that the proposed closure scheme meets three important goals. First, the method remains accurate regardless of the separation of time scales in the reaction rates. Second, the method remains valid for systems with multiple degrees of freedom. Third, the method is accurately implemented for higher-order closures (e.g., 12th-order closure is successfully implemented for the Schlögl model).

3.6 Results and Discussion

The time trajectory is shown in Figure 3.1A for the average and the variance (inset) of the number of A molecules in Model 1, as calculated with ZI-Closure. The results are also shown for the stochastic simulation algorithm, as improved by Gibson and Bruck[56] and implemented in Hy3S, an open-license software package for simulating stochastic reaction networks[72]. In all comparisons the results of 10^6 SSA trajectories are shown, unless otherwise stated.

Four distinct dynamic trajectories are shown for different equilibrium constant values spanning four orders of magnitude. The comparison is favorable between the ZI-Closure method and the SSA. Interestingly, we find that while at the extremes of equilibrium constant K values, 2nd-order closure is adequate, the match becomes relatively poor when k_1 is equal to k_2. Despite the simplicity of Model 1, 4th-order closure is necessary

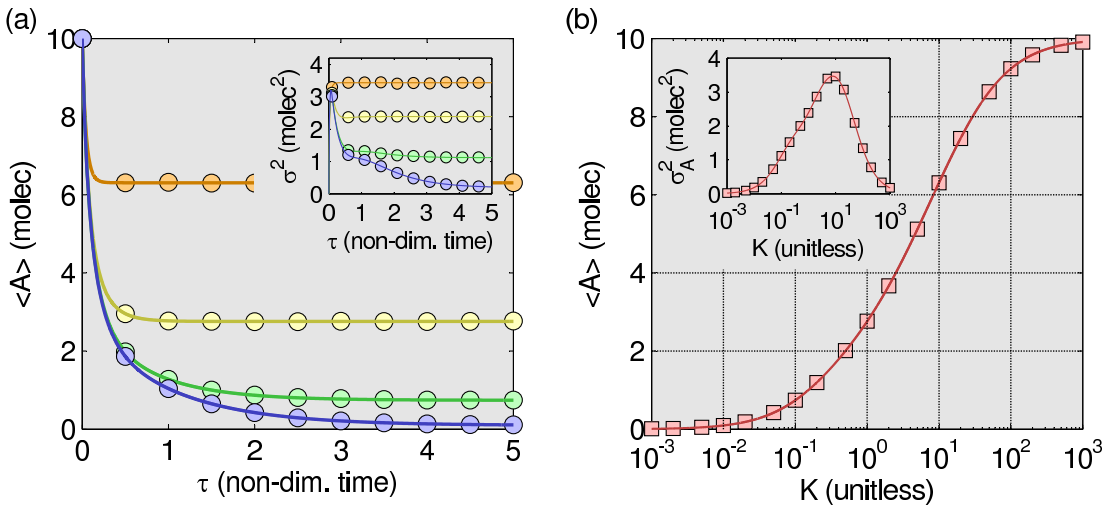

Figure 3.1: Dynamic trajectory and steady-state results for reversible non-linear dimerization. (a) Evolution of the average and variance (inset) of number of A molecules using 4th-order ZI-Closure. Different colors represent various values of equilibrium constant $K = k_2/ k_1$. The forward reaction rate is constant at $k_1 = 1$ (1/molec-s). The initial distribution for the trajectory results is a Kronecker-delta function, $P_0 = \delta_{A,10}$. Solid lines are from 4th-order ZI-Closure and circles are results from 100,000 SSA trajectories. Colors refer to dissociation constants with orange (K = 10), yellow (K = 1), green (K = 0.1), and blue (K = 0.01). (b) Steady-state results for a range of K values using 4th-order ZI-Closure (line), compared to SSA results (squares). Results for 10,000 k_2 values were modeled ranging from 10^{-3} to 10^3 (1/s) ($k_1 = 1$ (1/molec-s)). SSA results for 20 k_2 values, each with 100,000 trajectories. The variance results are shown in the inset.

for accurate results across all studied kinetic values. This presents a major advantage of ZI-Closure over previous closure schemes, most all of which cannot be assumed to remain accurate as closure order is increased to an arbitrarily high order.

We note that the ZI-Closure scheme is not as computationally efficient as the SSA is for simulating the dynamic evolution of reaction networks. This drawback is due to the computationally taxing optimization step present at each time step. There may be benefits in using the ZI-Closure method for stiff reaction networks, but exploring these is beyond the scope of this thesis.

The steady state of Model 1 is shown in Figure 3.1B across seven orders of magnitude for the kinetic constant (10^{-3} to 10^{3}). Again, 4th-order closure accurately describes the mean and variance for all constant values. Steady-state results are produced astonishingly fast, since only a single optimization step is needed. For demonstration purposes, we calculated 10,000 steady states varying the equilibrium constant value. Simply put, this is a staggering amount of data for even a highly efficient SSA algorithm to produce. As such, these results suggest the potential in using ZI-Closure in accurately and efficiently performing steady state and sensitivity analysis of stochastic reaction systems.

We also tested the ZI-Closure scheme for the Michaelis-Menten model. In Figure 3.2, the evolving average and variance for molecules S and E are shown for a range of kinetic constants. Steady-state results for the Michaelis-Menten model are presented in Figure 3.3 for a wide range of the four kinetic parameters over multiple orders of magnitude. Evidently, the ZI-Closure steady-state optimization algorithm is applicable to multi-component systems. The Michaelis-Menten results in particular demonstrate that this type of optimization may be used efficiently to perform sensitivity analysis. The slopes are equivalent to sensitivities of the mean and variance of the steady-state distribution to the four kinetic constants. Here again we observe that while 2nd-order ZI-Closure is for the most part adequate, it diverges slightly in several cases. In all cases 4th-order ZI-Closure is accurate.

Figure 3.4A shows the time trajectories for the Schlögl model with kinetic constants chosen to result in extreme bimodality (Figure 3.4B inset). This complex network was chosen because many moments are necessary to accurately describe the system evolution. The results clearly demonstrate how ZI-Closure can accurately capture even

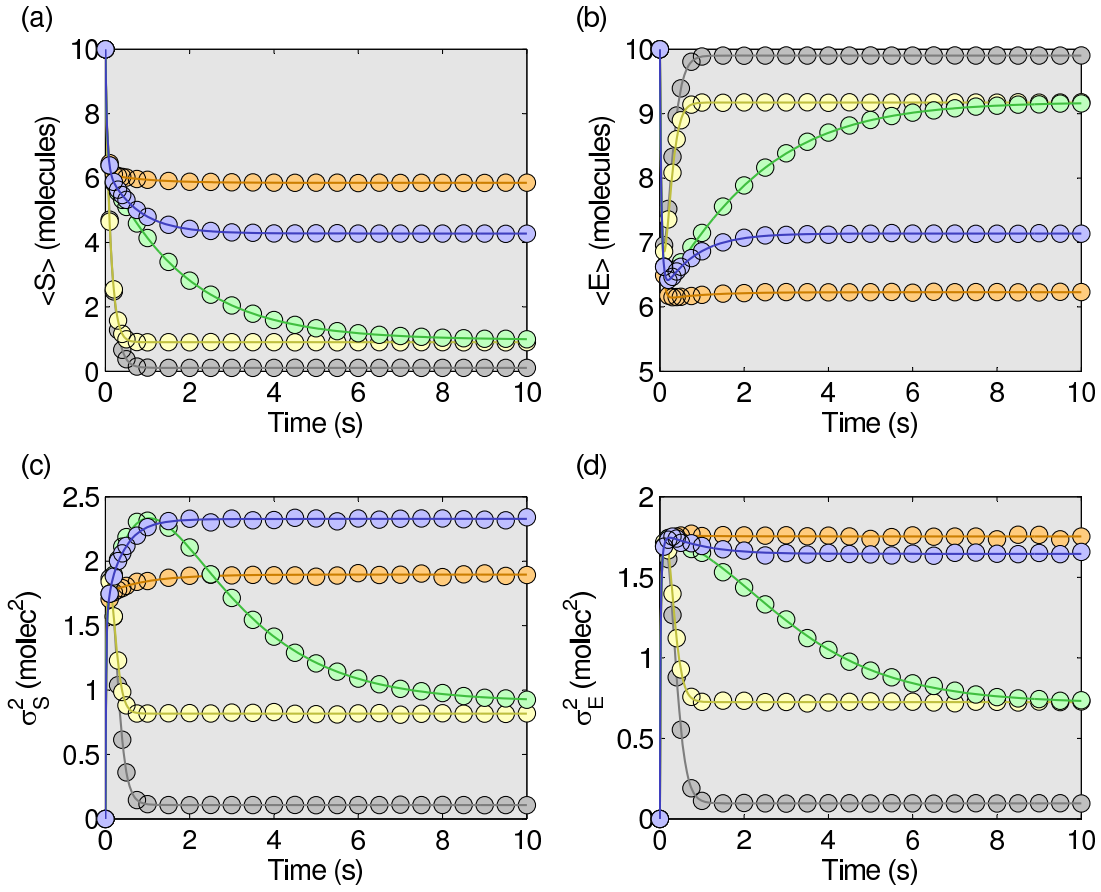

Figure 3.2: Michaelis-Menten trajectory results. Shown here are the average number and the variance of E and S molecules (plots a, b, c, and d, respectively) for five kinetic constant combinations. Results from ZI-Closure in solid lines and from SSA as circles. Colors refer to different kinetic rates with gray ([k_2, k_3, k_4] = [1, 10, 0.1] 1/s), orange ([k_2, k_3, k_4] = [10, 0.1, 1] 1/s), yellow ([k_2, k_3, k_4] = [0.1, 10, 1] 1/s); green ([k_2, k_3, k_4] = [10, 1, 0.1] 1/s); and blue ([k_2, k_3, k_4] = [10, 1, 1] 1/s). The initial distribution is a Kronecker-delta function, $P_0 = \delta_{S,10} \cdot \delta_{E,10}$.

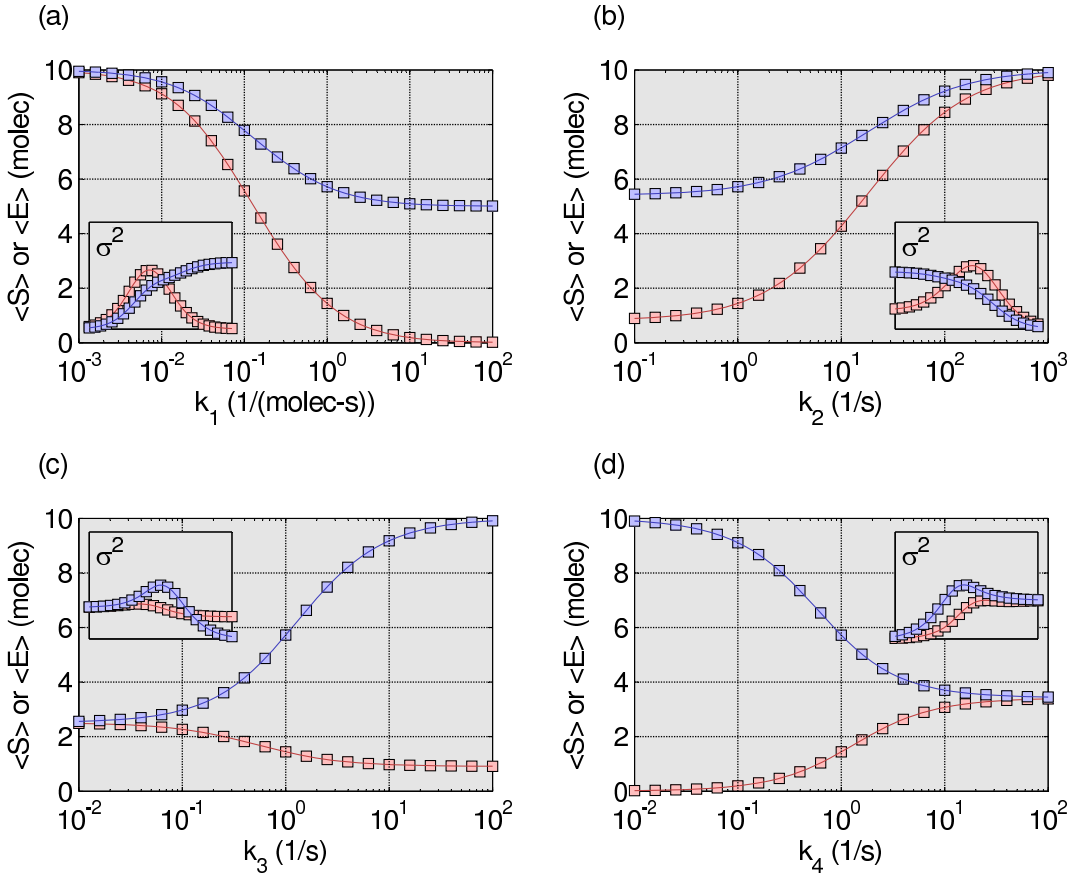

Figure 3.3: Steady-state results for the Michaelis-Menten model. The steady-state results for a wide range of kinetic parameter values (centered around $k_1 = 1$ (1/molec-s), $k_2 = 1$ (1/s), $k_3 = 1$ (1/s), and $k_4 = 1$ (1/s)) for the Michaelis-Menten model ($S_0 = 10$. $E_0 = 10$). (a) Both the mean substrate (S, red) and enzyme (E, blue) count are shown for 4th-order ZI-Closure (solid lines) and compared to SSA simulations (squares) with one million trajectories. Identical conditions are shown for (b), (c) and (d) except applied to k_2, k_3 and k_4 respectively. Note that over each parameter range in (a-d) all other parameters are held constant. The insets show variances for S (red) and E (blue) for both ZI-Closure (line) and SSA (squares) results.

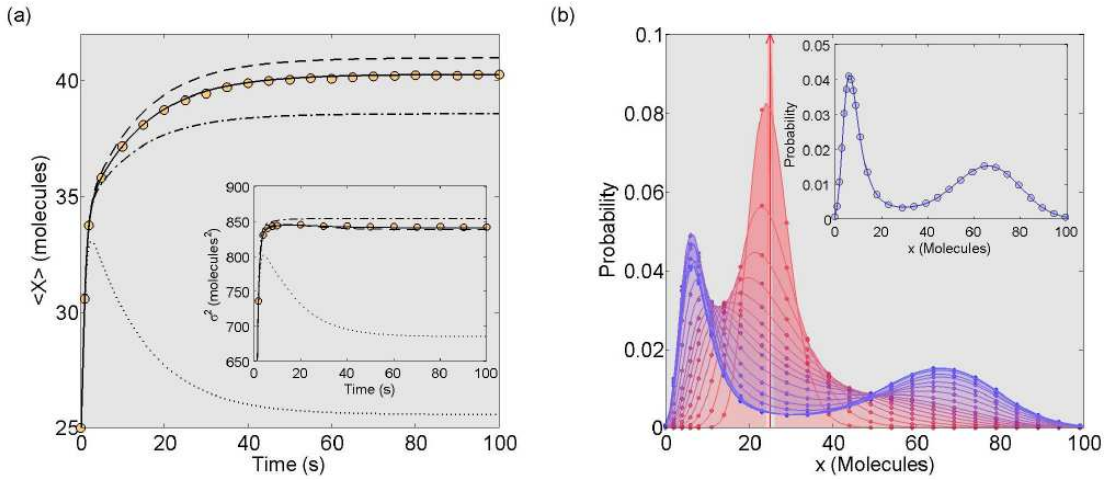

Figure 3.4: Schlögl model trajectory results and steady-state distribution. The ZI-Closure trajectory results for the Schlögl model ($k_1 \cdot A = 0.15$ (1/molec-s), $k_2 = 0.0015$ (1/molec^2-s), $k_3 \cdot B = 20$ (molec/s), and $k_4 = 3.5$ (1/s)). The initial distribution is a Kronecker-delta function, $P_0 = \delta_{X,25}$. (a) The mean output of X through time; for 6th-order (dotted line), 8th-order (dot-dash line), 10th-order (dashed line), and 12th-order ZI-Closure (solid line). The trajectories are compared to one million SSA trajectories (circles). The inset shows corresponding variance results. (b) To demonstrate how this method replicates the actual underlying distribution distribution found using 12th-order ZI-Closure (lines) is plotted with the SSA-simulated distribution (circles). Time flows from red to blue demonstrating good reproduction of the actual underlying distribution throughout time. The inset compares the steady-state 12th-order ZI-Closure (line) to the steady-state SSA distribution (circles).

complex distributions. Figure 3.4A shows 6th, 8th, 10th, and 12th-order ZI-Closure compared to results of one million SSA trajectories. It is demonstrated that 12th-order closure is necessary to accurately match the mean of the SSA trajectories. These results are significant because this is the first time a closure scheme achieves accuracy at such high moment order.

Figure 3.4B shows the probability distributions through time for both 12th-order ZI-Closure and the SSA. The distributions computed with ZI-Closure match the actual SSA distributions remarkably well throughout time. These results provide a most convincing argument for ZI-Closure as a powerful closure scheme. Indeed, the bimodal distribution of the Schlögl model is particularly challenging to simulate.

Steady-state results are presented in Figure 3.5 for the Schlögl model over a wide

range of values for the four kinetic constants. Only the 12th-order closure is shown for the mean and variance results compared to SSA results. A quantitative sensitivity analysis is now possible, investigating the impact of kinetic constants on the behavior of the network. The steady-state optimization method renders a thorough analysis feasible, where none was previously available. Sampling the probability distribution with SSA quickly becomes untenable because of the combinatorial explosion of necessary trajectories.

For all the appeal of the ZI-Closure scheme drawbacks exist. First, in its present form, the algorithm tends to be less efficient in producing trajectories through time than SSA. This is because optimization, the complexity of which scales with the state space size, is needed in every time step. Second, the number of moment equations scales as $(N + M)$ choose M, while ODE or stochastic differential equation models for chemical networks scale with the number of components, N. This challenge inherently limits all moment closure schemes, and it likely limits the utility to relatively small networks until large scale algorithms are developed. Finally, ZI-Closure, in its present form, faces numerical implementation challenges, in particular when delta functions best describe the probability distributions. At this limit the α parameters diverge to infinity, although starting from a multivariate Gaussian distribution addresses this drawback.

3.7 Conclusions

Since the days of Newton and Leibnitz, mathematical models have been at the heart of physical and engineering sciences. Founded on universally accepted physicochemical laws, these models capture the essential aspects of systems, phenomena and processes, all in a way fit for analysis, explanation, understanding and then for design, engineering, optimization and control.

There are many reasons why mathematical models are not presently at the heart of biological sciences. To name a few: we are still discovering the parts that comprise living organisms; there are very many of these components; the kinetic parameters and thermodynamic strength of interactions are not known; there are environmental, exogenous dependencies that dictate biological behaviors; there are evolutionary, historical links that determine the nature of biosystems. They all impose significant epistemological

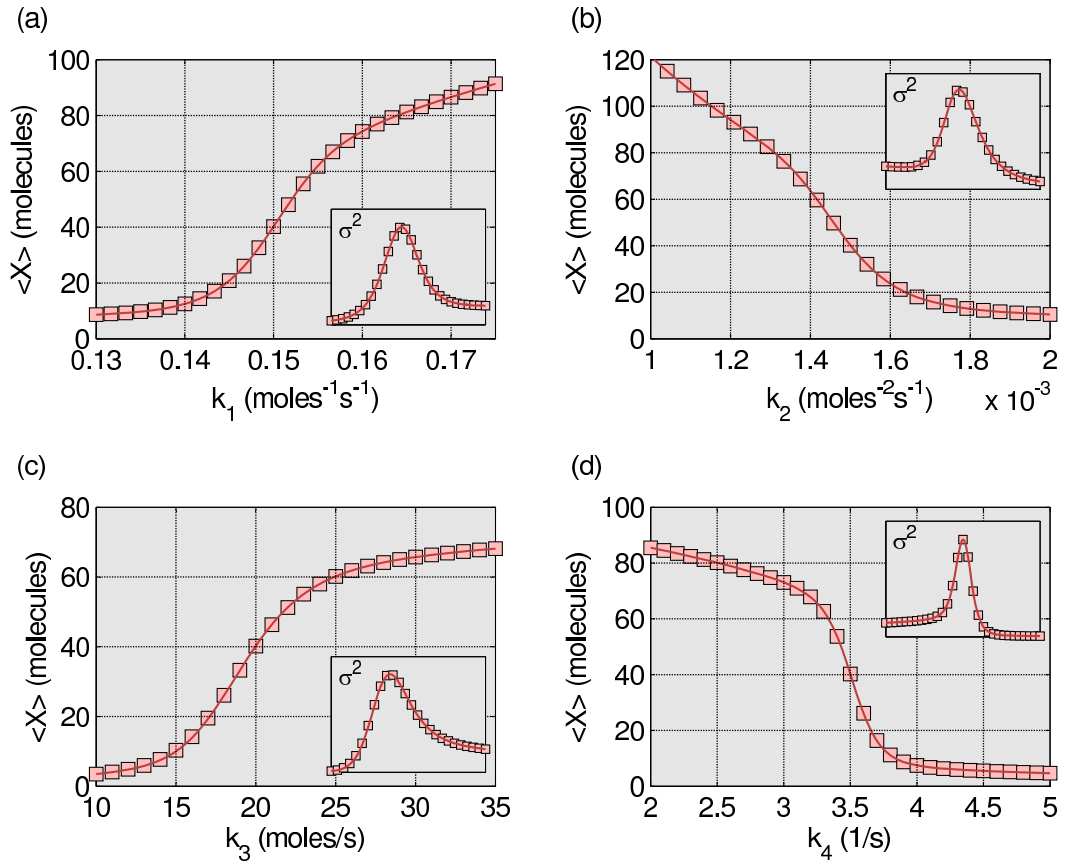

Figure 3.5: Steady-state Schlögl model results. Sensitivity analysis around the dynamic simulation parameter values. (a) The mean steady-state output of X was simulated for 450 $k_1 \cdot A$ values between 0.13 and 0.175 (1/molec-s) using a 12th-order ZI-Closure (line), and compared to 28 points simulated using an SSA (squares). (b) Results for 200 k_2 values between 0.001 and 0.002 (1/molec^2-s) using a 12th-order ZI-Closure (line), compared to 25 points simulated using an SSA (squares). (c) Results for 250 $k_3 \cdot B$ values between 10 and 35 (molec/s) using a 12th-order ZI-Closure (line), compared to 25 points simulated using an SSA (squares) (d) Results for 300 k_4 values between 2 and 5 (1/s) using a 12th-order ZI-Closure (line), compared to 30 points simulated using an SSA (squares). The insets show variance comparisons.

hurdles, let alone the practical ones involved in reliable model development.

Another challenge facing scientists is related to capturing random molecular events that frequently determine the fate of a living organism. It has been argued that probabilistic fluctuations are a defining feature of biomolecular systems, conferring necessary elasticity under environmental stresses. Living organisms can then explore a distribution of states with finite probability. As a result, stochastic outcomes become incredibly important in describing underlying dynamics: they equip a population of organisms with adaptation under evolutionary pressures, but may also doom an individual organism.

The modeling framework for capturing probabilistic outcomes in evolving molecular populations has been cast for over seven decades with the work of Delbrück and McQuarrie on chemical master equations. However, a solution has been elusive thus far, when there are second or higher order reactions. It is inconceivable to describe biomolecular systems without the presence of interacting molecular partners. Consequently, master equations have not inspired biological discovery.

This chapter presented the ZI-Closure scheme for master probability equations that govern the evolution of small molecular populations reacting with reaction rates of order higher than one. We demonstrate progress in three main areas. First, ZI-Closure works on single or multi-component reaction networks. Second, ZI-Closure is applicable across a wide range of kinetic parameters in simple models. Third, ZI-Closure can be applied for high order closure with commensurate accuracy gains.

ZI-Closure represents the first closure scheme that renders the moment viewpoint a viable alternative to kinetic Monte Carlo methods for stochastic chemical simulation. Importantly, ZI-Closure represents a unique method for determining steady-state distributions without having to simulate reaction networks through time. ZI-Closure requires only a single optimization step to determine steady-state distributions, significantly more efficient than kinetic Monte Carlo simulations. As such, ZI-Closure facilitates the use of many well-defined analysis tools such as the exploration of eigenvalue analysis and power spectral densities presented in Chapters 4 and 5.

We believe these results to be of special importance for the biological sciences. For example, with a microscopic definition of irreversible processes, a range of experimental observations of biomolecular interactions may be mathematically conceptualized. Ultimately, general principles that govern biological phenomena may be established.

Chapter 4

On a Theory of Stability for Nonlinear Stochastic Chemical Reaction Networks

4.1 Introduction

In Chapter 3 a novel closure scheme for chemical reaction networks was established using information theory called zero-information closure (ZI-Closure)[116]. While dynamic solutions can be obtained, the method, which involves a computationally costly optimization element, shows greater promise in its ability to quickly determine steady-state probability distributions. In this chapter the well-established deterministic analytical tool of non-linear analysis is developed using ZI-Closure.

Improved understanding of compelling dynamic phenomena, from molecular chaos to relativistic cosmological models, has been possible to large extent only thanks to such mathematical theories as Lyapunov's theory of stability of nonlinear dynamic systems[109]. First developed over 120 years ago, Lyapunov's elegant theories also propelled the development of the disciplines of systems and control engineering[110, 111]. In the area of chemical dynamics, Lyapunov's stability theory laid the foundation for understanding nonlinear chemical systems, e.g., the Belusov-Zhabotinsky reaction pattern formation[112], or the non-isothermal, continuously stirred chemical reactor [113].

As established in Chapter 1, stochasticity is often a significant, impactful feature in dynamic systems[1], and in the recent past stochasticity has become a significant focus of studies of biochemical reaction systems[12]. Powerful mathematical formalisms have been developed to model and analyze stochastic dynamic systems, including models based on master equations[41]. A prominent place in the theory of stochastic dynamics is occupied by the fluctuation-dissipation theorem, that the linear response of a given system to an external perturbation is expressed in terms of fluctuation properties of the system in thermal equilibrium[114, 115]. For chemical reaction systems, Gillespie's stochastic simulation algorithm has sparked the development of a plethora of modeling approaches[44].Yet, despite the significance of stochasticity in nonlinear dynamics, attempts to develop a theory for the stability of stochastic reacting systems have largely remained unfruitful.

Here we present a theory for stability analysis of stochastic nonlinear chemical dynamics. With ZI-Closure, the steady-state probability distribution of small nonlinear chemical reaction networks may be obtained accurately and quickly. In a process that resembles Lyapunov's first method, eigenvalues can be computed directly from the Jacobian matrix of probability moments at the steady state. We show with numerous examples that steady-state solutions are stable with the eigenvalues being negative real numbers. We discuss stochastic oscillations, when eigenvalues appear with non-zero imaginary parts, along with the possibility of observing positive eigenvalues in stochastic reacting systems. Intriguingly, parallels are drawn to the fluctuation-dissipation theorem[115], with correlation functions of probability distribution moments being congruent to transient relaxation dynamics when probability distributions of chemical reaction networks are perturbed from their steady state.

In what follows we present the theory for computing the Jacobian matrix of steady-state probability moments and for computing the Jacobian eigenvalues. We discuss the calculation of correlation functions of probability distribution moment fluctuations. We model and analyze prototypical nonlinear chemical reaction network models with stochastic dynamics. We compute steady-state solutions and analyze their stability. We conclude by stressing the inherent limitations of the proposed methods and by speculating about their significance in the field of stochastic nonlinear dynamics.

4.2 Theory

4.2.1 Zero Information Closure Scheme

A complete model of randomly evolving chemical reactions is one based on the chemical master equation (CME):

$$\frac{\partial P\left(\underline{X};t\right)}{\partial t}=\int\left[T\left(\underline{X}|\underline{X}'\right)P\left(\underline{X}';t\right)-T\left(\underline{X}'|\underline{X}\right)P\left(\underline{X};t\right)\right]d\underline{X}' \tag{4.1}$$

Here $\underline{X}$ is the state of the system, an N-dimensional vector with the concentrations of the N reactants and products in the chemical reaction network. $P\left(\underline{X};t\right)$ is the probability of being in the state $\underline{X}$ at time t. $T\left(\underline{X}|\underline{X}'\right)$ is the transition probability per unit time of reaction events changing the state from $\underline{X}'$ to $\underline{X}$.

If the reaction rate laws for the reaction network are elementary, the CME can be used to generate a set of equations of the form[96, 105]:

$$\frac{\partial\underline{\mu}}{\partial t}=A\underline{\mu}+A'\underline{\mu}' \tag{4.2}$$

Here $\underline{\mu}$ is a vector of probability moments up to order M (length N_{M}) and $\underline{\mu}'$ is a vector of the higher-order moments needed to close the system (moment order M' and vector length N'_{M}). A is thus an $N_M\times N_M$ matrix and A' is an $N_M\times N'_M$ matrix.

The presence of higher-order moments ($\underline{\mu}'$) in the set of equations results in the moment closure challenge. These terms are present if there are second-order reactions or higher; A' is then a non-empty matrix. Our moment closure scheme, ZI-Closure, is outlined in detail in Chapter 3 and rests on the simple notion that higher order moments, although certainly non-negligible in arithmetic value, offer little in the way of information necessary to reconstruct the underlying probability distribution[116]. We have shown that ZI-Closure, based on the maximization of the system's information entropy, offers a closure scheme for the CME of small, nonlinear, stochastic chemical reaction networks.

The entropy of a distribution is defined by Shannon as follows[106]:

$$H=\sum_{x=0}^{\infty}-p\left(x\right)\ln p\left(x\right) \tag{4.3}$$

The entropy defined in Equation 4.3 can be maximized subject to constraints (the assumed known lower-order moments) to define the maximum entropy distribution)[107]:

$$p_H(x) = exp\left(-\alpha_0 - \alpha_1 x - \alpha_2 x^2 - \cdots - \alpha_M x^M\right) \tag{4.4}$$

Note that this derivation is for a single component system with M lower-order moments chosen for the closure scheme; the choice of M depends on the complexity of modeled probability distribution (in Chapter 3 up to twelve moments were necessary to accurately produce the steady-state Schlögl distribution). The parameters α_i are Legrange multipliers. With the now known maximum entropy probability distribution, higher-order moments can be computed as:

$$\mu_i' = \sum_{x=0}^{\infty} f_{\mu_i'}(x)\, p_H(x) \tag{4.5}$$

where $f_{\mu_i'}(x)$ is the function for the i-th higher-order moment (for the above example μ' might be $\left\langle x^{M+1} \right\rangle$ and $f_{\mu'}(x) = x^{M+1}$).

4.2.2 Steady-State Probability Distributions, Jacobian Matrices and Eigenvalues

The steady-state distribution of a stochastic chemical reaction network can be determined by setting the left side of Equations 4.2 to zero. This calculation can be fast for small reaction networks, certainly faster than kinetic Monte Carlo sampling of the probability distribution[116].

A linearized Jacobian (J_{SS}) may be computed when the steady-state distribution is calculated with ZI-Closure, and Equation 4.2 can be approximated around the steady state as:

$$\frac{\partial \underline{\mu}}{\partial t} = J_{SS}\underline{\mu} \tag{4.6}$$

Nonlinear stability analysis can then be performed by calculating the eigenvalues and eigenvectors of the Jacobian. Before we discuss how these eigenvalues and eigenvectors dictate the dynamic behavior and the stability of the reaction network around steady states, we present the calculation of J_{SS}.

4.2.3 Detailed Derivation of $\mathbf{J}_{SS}$

Looking to Equation 4.2 what is needed for this analysis is the Jacobian, denoted as J_{SS}, near the steady-state distribution. Through a closure scheme the unknown higher-order

moments are related to the lower-order moments ($\underline{\mu}' = \underline{F}\left(\underline{\mu}\right)$). Although the analysis presented can be applied to arbitrarily complex networks, for illustrative purposes let's assume a single component (x) system and the lower-order moments are up to order two ($\underline{\mu} = \left[\left\langle x^0\right\rangle, \left\langle x\right\rangle, \left\langle x^2\right\rangle\right]^T$) and the only higher-order moment necessary for closure is the third-moment ($\underline{\mu}' = \left\langle x^3\right\rangle$). In this case Equation 4.2 is more clearly defined as:

$$\frac{\partial}{\partial t}\begin{bmatrix} \left\langle x^0\right\rangle \\ \left\langle x\right\rangle \\ \left\langle x^2\right\rangle \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\begin{bmatrix} \left\langle x^0\right\rangle \\ \left\langle x\right\rangle \\ \left\langle x^2\right\rangle \end{bmatrix} + \begin{bmatrix} a'_{11} \\ a'_{21} \\ a'_{31} \end{bmatrix}\left\langle x^3\right\rangle \tag{4.7}$$

The Jacobian matrix can be computed with the following equation (from Equation 4.2):

$$J_{SS} = \left.\frac{\partial\left[\partial\underline{\mu}/\partial t\right]}{\partial\underline{\mu}}\right|_{SS} = A + A'\left.\frac{\partial\underline{\mu}'}{\partial\underline{\mu}}\right|_{SS} \tag{4.8}$$

Given three variables (the lower-order moments) and three functions the Jacobian is defined as:

$$J_{SS} = \begin{bmatrix} \frac{\partial\left[\partial_t\left\langle x^0\right\rangle\right]}{\partial\left\langle x^0\right\rangle} & \frac{\partial\left[\partial_t\left\langle x^0\right\rangle\right]}{\partial\left\langle x\right\rangle} & \frac{\partial\left[\partial_t\left\langle x^0\right\rangle\right]}{\partial\left\langle x^2\right\rangle} \\ \frac{\partial\left[\partial_t\left\langle x\right\rangle\right]}{\partial\left\langle x^0\right\rangle} & \frac{\partial\left[\partial_t\left\langle x\right\rangle\right]}{\partial\left\langle x\right\rangle} & \frac{\partial\left[\partial_t\left\langle x\right\rangle\right]}{\partial\left\langle x^2\right\rangle} \\ \frac{\partial\left[\partial_t\left\langle x^2\right\rangle\right]}{\partial\left\langle x^0\right\rangle} & \frac{\partial\left[\partial_t\left\langle x^2\right\rangle\right]}{\partial\left\langle x\right\rangle} & \frac{\partial\left[\partial_t\left\langle x^2\right\rangle\right]}{\partial\left\langle x^0\right\rangle} \end{bmatrix}_{SS} =$$
$$\begin{bmatrix} a_{11} + a'_{11}\frac{\partial\left\langle x^3\right\rangle}{\partial\left\langle x^0\right\rangle} & a_{12} + a'_{11}\frac{\partial\left\langle x^3\right\rangle}{\partial\left\langle x\right\rangle} & a_{13} + a'_{11}\frac{\partial\left\langle x^3\right\rangle}{\partial\left\langle x^2\right\rangle} \\ a_{21} + a'_{21}\frac{\partial\left\langle x^3\right\rangle}{\partial\left\langle x^0\right\rangle} & a_{22} + a'_{21}\frac{\partial\left\langle x^3\right\rangle}{\partial\left\langle x\right\rangle} & a_{23} + a'_{21}\frac{\partial\left\langle x^3\right\rangle}{\partial\left\langle x^2\right\rangle} \\ a_{31} + a'_{31}\frac{\partial\left\langle x^3\right\rangle}{\partial\left\langle x^0\right\rangle} & a_{32} + a'_{31}\frac{\partial\left\langle x^3\right\rangle}{\partial\left\langle x\right\rangle} & a_{33} + a'_{31}\frac{\partial\left\langle x^3\right\rangle}{\partial\left\langle x^2\right\rangle} \end{bmatrix}_{SS} \tag{4.9}$$

Three unknowns (sensitivities of the higher-order moments with regard to the lower-order moments) need to be calculated at the steady-state distribution. These relations can be found while determining the steady-state distribution through the Legrange parameters.

The higher-order moments are related to the lower-order moments through the Lagrange parameters $\underline{\alpha}$ in $P_H(x)$. Thus, for μ'_j with $j \geq M+1$, we write:

$$d\mu'_j = \frac{\partial\mu'_j}{\partial\alpha_0}d\alpha_0 + \frac{\partial\mu'_j}{\partial\alpha_1}d\alpha_1 + \cdots + \frac{\partial\mu'_j}{\partial\alpha_M}d\alpha_M \tag{4.10}$$

and taking derivatives at the steady state:

$$\left.\frac{\partial \mu'_j}{\partial \mu_i}\right|_{SS} = \left[\frac{\partial \mu'_j}{\partial \alpha_0}\frac{\partial \alpha_0}{\partial \mu_i} + \frac{\partial \mu'_j}{\partial \alpha_1}\frac{\partial \alpha_1}{\partial \mu_i} + \cdots + \frac{\partial \mu'_j}{\partial \alpha_M}\frac{\partial \alpha_M}{\partial \mu_i}\right]\Bigg|_{SS} \tag{4.11}$$

Note that this is an $N'_M \times N_M$ matrix.

From a computational standpoint it is easier to obtain the following two matrices at steady state:

$$J_{NR} = \left.\frac{\partial \underline{\mu}}{\partial \underline{\alpha}}\right|_{SS} \tag{4.12}$$

$$J'_{NR} = \left.\frac{\partial \underline{\mu}'}{\partial \underline{\alpha}}\right|_{SS} \tag{4.13}$$

J_{NR} is an $N_M \times N_M$ matrix and J'_{NR} is an $N'_M \times N_M$ matrix. In the steady-state distribution Newton-Raphson optimization the expanded Jacobian is determined[116] comprised of J_{NR} and J'_{NR}. Part of this matrix for the above example is as follows:

$$\left[\frac{J_{NR}}{J'_{NR}}\right] = \left[\begin{array}{ccc} \frac{\partial\langle x^0\rangle}{\partial\alpha_0} & \frac{\partial\langle x^0\rangle}{\partial\alpha_1} & \frac{\partial\langle x^0\rangle}{\partial\alpha_2} \\ \frac{\partial\langle x\rangle}{\partial\alpha_0} & \frac{\partial\langle x\rangle}{\partial\alpha_1} & \frac{\partial\langle x\rangle}{\partial\alpha_2} \\ \frac{\partial\langle x^2\rangle}{\partial\alpha_0} & \frac{\partial\langle x^2\rangle}{\partial\alpha_1} & \frac{\partial\langle x^2\rangle}{\partial\alpha_2} \\ \hline \frac{\partial\langle x^3\rangle}{\partial\alpha_0} & \frac{\partial\langle x^3\rangle}{\partial\alpha_1} & \frac{\partial\langle x^3\rangle}{\partial\alpha_2} \end{array}\right] \tag{4.14}$$

Combining Equations 4.11-4.13 results in:

$$\left.\frac{\partial \underline{\mu}'}{\partial \underline{\mu}}\right|_{SS} = J'_{NR}J_{NR}^{-1} \tag{4.15}$$

And more explicitly:

$$J'_{NR}J_{NR}^{-1} = \left[\begin{array}{c} \frac{\partial\langle x^3\rangle}{\partial\alpha_0}\frac{\partial\alpha_0}{\partial\langle x^0\rangle} + \frac{\langle x^3\rangle}{\partial\alpha_1}\frac{\partial\alpha_1}{\partial\langle x^0\rangle} + \frac{\langle x^3\rangle}{\partial\alpha_2}\frac{\partial\alpha_2}{\partial\langle x^0\rangle}, \\ \frac{\partial\langle x^3\rangle}{\partial\alpha_0}\frac{\partial\alpha_0}{\partial\langle x\rangle} + \frac{\langle x^3\rangle}{\partial\alpha_1}\frac{\partial\alpha_1}{\partial\langle x\rangle} + \frac{\langle x^3\rangle}{\partial\alpha_2}\frac{\partial\alpha_2}{\partial\langle x\rangle}, \\ \frac{\partial\langle x^3\rangle}{\partial\alpha_0}\frac{\partial\alpha_0}{\partial\langle x^2\rangle} + \frac{\langle x^3\rangle}{\partial\alpha_1}\frac{\partial\alpha_1}{\partial\langle x^2\rangle} + \frac{\langle x^3\rangle}{\partial\alpha_2}\frac{\partial\alpha_2}{\partial\langle x^2\rangle} \end{array}\right]$$

$$J'_{NR}J_{NR}^{-1} = \left[\begin{array}{ccc} \frac{\partial\langle x^3\rangle}{\langle x^0\rangle} & \frac{\partial\langle x^3\rangle}{\langle x\rangle} & \frac{\partial\langle x^3\rangle}{\langle x^2\rangle} \end{array}\right] \tag{4.16}$$

As can be seen with Equation 4.9 these are all of the unknowns necessary to calculate J_{SS}.

Generally J_{SS} (Equation 5.8) can be written as:

$$J_{SS} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \begin{bmatrix} a'_{11} \\ a'_{21} \\ a'_{31} \end{bmatrix} \begin{bmatrix} \frac{\partial \langle x^3 \rangle}{\partial \langle x^0 \rangle} & \frac{\partial \langle x^3 \rangle}{\partial \langle x \rangle} & \frac{\partial \langle x^3 \rangle}{\partial \langle x^2 \rangle} \end{bmatrix} \tag{4.17}$$

Finally, a very simple and general equation can be defined for J_{SS}:

$$J_{SS} = A + A' J_{NR} J_{NR}^{-1} \tag{4.18}$$

In this chapter, for small chemical reaction networks, we compute the Jacobian matrix, and the eigenvalues/eigenvectors of J_{SS}, which are now trivial to calculate. The significance of eigenvalues and eigenvectors becomes evident with the introduction of time correlation functions and response functions, which we introduce briefly next.

4.2.4 Correlation Equation Dynamics

The next step is to show the connection between the moment equations and the correlation equation dynamics. The moment equations for this particular analysis will take on a slightly different form as to how it has been written previously:

$$\frac{\partial \underline{\mu}}{\partial t} = A\underline{\mu} + \underline{b} \tag{4.19}$$

This is for a linear system (thus A' and $\underline{\mu}'$ are absent), but also in this case the zeroth-moment terms are taken out of the matrix A into a separate vector $\underline{b}$. For an N-component system then the moment vector (for up to order-one) is $\underline{\mu} = [\langle x_1 \rangle, \langle x_2 \rangle, \ldots, \langle x_N \rangle]^T$. The regression theorem states simply that with these moment equations the following is also true[117]:

$$\frac{\partial \left\langle \underline{\mu}(t)\, \underline{\mu}(0)^T \right\rangle}{\partial t} = A \left\langle \underline{\mu}(t)\, \underline{\mu}(0)^T \right\rangle + \underline{b} \left\langle \underline{\mu}(0)^T \right\rangle \tag{4.20}$$

The correlation functions are defined as: $C_{x_i,x_j}(t) = \langle (x_i(t) - \langle x_i(t) \rangle)(x_j(0) - \langle x_j(0) \rangle) \rangle$ and thus the following is also true:

$$\frac{\partial}{\partial t}\begin{bmatrix} C_{x_1,x_1} & C_{x_1,x_2} & \cdots & C_{x_1,x_N} \\ C_{x_2,x_1} & C_{x_2,x_2} & \cdots & C_{x_2,x_N} \\ \vdots & \vdots & \ddots & \vdots \\ C_{x_N,x_1} & C_{x_N,x_2} & \cdots & C_{x_N,x_N} \end{bmatrix} = A \begin{bmatrix} C_{x_1,x_1} & C_{x_1,x_2} & \cdots & C_{x_1,x_N} \\ C_{x_2,x_1} & C_{x_2,x_2} & \cdots & C_{x_2,x_N} \\ \vdots & \vdots & \ddots & \vdots \\ C_{x_N,x_1} & C_{x_N,x_2} & \cdots & C_{x_N,x_N} \end{bmatrix} \tag{4.21}$$

Note that the zeroth-order moment terms ($\underline{b}$) disappears. Ultimately that vector is always multiplied by zero and can be eliminated. It is important to note that this is not done in the prior section with the steady-state Jacobian. The first Legrange parameter (for the zeroth moment) is important when calculating J_{SS}, but then is eliminated with regard to the correlation dynamics since its corresponding eigenvalue is zero and represents a steady-state which is subtracted out.

First, assume there is an N-component linear system, and you are interested in performing analysis on the first random variable (x_1). In this case you need to obtain an analytical solution to $C_{x_1,x_1}(t)$ and thus the correlation matrix can be reduced to a single vector:

$$\frac{\partial \underline{C}_{x_1}}{\partial t} = A\underline{C}_{x_1} \tag{4.22}$$

where $\underline{C}_{x_1} = [C_{x_1,x_1}, C_{x_2,x_1}, \ldots, C_{x_N,x_1}]^T$.

Analysis on matrix A can produce a vector of eigenvalues $\underline{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_N]^T$ and a corresponding matrix of eigenvectors $S = \begin{bmatrix} \underline{\nu}_1 & \underline{\nu}_2 & \ldots & \underline{\nu}_N \end{bmatrix}$. Since the system is linear ZI-Closure is not needed. Spectral resolution dictates that the analytical solution (near the steady-state) for $C_{x_1,x_1}(t)$ can be written as:

$$\underline{C}_{x_1}(t) = S \exp(I\underline{\lambda}t) S^{-1} \underline{C}_{x_1}(0) \tag{4.23}$$

where I is an $N \times N$ identity matrix. Since we assume the probability distribution is stationary the initial condition of the correlation functions will relate to the steady-state covariances:

$$\underline{C}_{x_1}(0) = \begin{bmatrix} \left\langle (x_1 - \langle x_1 \rangle)^2 \right\rangle \\ \langle (x_2 - \langle x_2 \rangle)(x_1 - \langle x_1 \rangle) \rangle \\ \vdots \\ \langle (x_N - \langle x_N \rangle)(x_1 - \langle x_1 \rangle) \rangle \end{bmatrix}_{SS} = \begin{bmatrix} \sigma_{x_1}^2 \\ \sigma_{x_2,x_1} \\ \vdots \\ \sigma_{x_N,x_1} \end{bmatrix}_{SS} \tag{4.24}$$

The key relationship between the correlation functions and the eigenvalues is that the correlation functions can be reduced to:

$$C_{x_1,x_1}(t) = D_1 \exp(\lambda_1 t) + D_2 \exp(\lambda_2 t) + \cdots + D_N \exp(\lambda_N t) \tag{4.25}$$

With ZI-Closure, this theoretical treatment is extended for nonlinear stochastic chemical dynamics. We domonstrate that the correlation functions of probability distribution moments can be computed near the steady state as

$$\frac{\partial C}{\partial t} = J_{SS} C \tag{4.26}$$

Analogously to linear system cases, using J_{SS}, one can obtain eigenvalues (λ) for nonlinear systems at steady state, and determine time correlation functions using the eigenvalues and eigenvectors of J_{SS}, as in Equation 4.25.

This function is calculated exclusively with ZI-Closure and does not require the use of the SSA at all. We use the autocorrelation functions to compare our analytical results to SSA results in order to make a quantitative statement about whether the eigenvalues calculated using moment closure are truly representative of the CME behavior.

Non-linear Correlation Function Note

There is an important but subtle note about the form of the correlation functions in non-linear systems. For illustrative purposes assume that we are dealing with a two component system (like the Michaelis-Menten system) and choose the closure order as two. The moment vector will be $\underline{\mu} = \left[\langle x_1\rangle, \langle x_2\rangle, \langle x_1^2\rangle, \langle x_1 x_2\rangle, \langle x_2^2\rangle\right]^T$. If we are concerned with component one in particular then the corresponding correlation vector for the non-linear system will be:

$$\underline{C}_{x_1}(t) = \begin{bmatrix} \langle (x_1(t) - \langle x_1(t)\rangle)(x_1(0) - \langle x_1(0)\rangle)\rangle \\ \langle (x_2(t) - \langle x_2(t)\rangle)(x_1(0) - \langle x_1(0)\rangle)\rangle \\ \langle (x_1^2(t) - \langle x_1^2(t)\rangle)(x_1(0) - \langle x_1(0)\rangle)\rangle \\ \langle (x_1(t)x_2(t) - \langle x_1(t)x_2(t)\rangle)(x_1(0) - \langle x_1(0)\rangle)\rangle \\ \langle (x_2^2(t) - \langle x_2^2(t)\rangle)(x_1(0) - \langle x_1(0)\rangle)\rangle \end{bmatrix} = \begin{bmatrix} C_{x_1,x_1} \\ C_{x_2,x_1} \\ C_{x_1^2,x_1} \\ C_{x_1x_2,x_1} \\ C_{x_2^2,x_1} \end{bmatrix} \tag{4.27}$$

The reason this introduces a subtle complication to the non-linear method has to do with the spectral resolution which relies on the initial condition of the correlation

vector ($\underline{C}_{x_1}(0)$). For this particular example the initial condition is as follows:

$$\underline{C}_{x_1}(0) = \begin{bmatrix} \left\langle (x_1 - \langle x_1 \rangle)^2 \right\rangle \\ \langle (x_2 - \langle x_2 \rangle)(x_1 - \langle x_1 \rangle) \rangle \\ \left\langle \left(x_1^2 - \left\langle x_1^2 \right\rangle\right)(x_1 - \langle x_1 \rangle) \right\rangle \\ \langle (x_1 x_2 - \langle x_1 x_2 \rangle)(x_1 - \langle x_1 \rangle) \rangle \\ \left\langle \left(x_2^2 - \left\langle x_2^2 \right\rangle\right)(x_1 - \langle x_1 \rangle) \right\rangle \end{bmatrix}_{SS} = \begin{bmatrix} \langle x_1^2 \rangle & - & \langle x_1 \rangle \langle x_1 \rangle \\ \langle x_2 x_1 \rangle & - & \langle x_2 \rangle \langle x_1 \rangle \\ \langle x_1^3 \rangle & - & \langle x_1^2 \rangle \langle x_1 \rangle \\ \langle x_1^2 x_2 \rangle & - & \langle x_1 x_2 \rangle \langle x_1 \rangle \\ \langle x_1 x_2^2 \rangle & - & \langle x_2^2 \rangle \langle x_1 \rangle \end{bmatrix}_{SS} \tag{4.28}$$

As can be seen these are not straightforward covariances as in the linear system, but rather cover multiple moment orders and result in somewhat unexpected form for the initial conditions.

4.2.5 Fluctuation-Dissipation

Furthermore, we demonstrate that the fluctuation-dissipation theorem applies to non-linear stochastic chemical reactions. Chemical reaction networks can be perturbed to new steady states, by changing the value of system parameters. The system's probability distribution relaxes from one steady state to another. These response functions are congruent to steady state correlation functions of probability moment fluctuations, so long as the perturbation results in a linear system response.

The normalized response function for any component x_i is calculated as:

$$F_{x_i}(t) = \frac{\langle x_i \rangle (t) - \langle x_i \rangle_{SS,2}}{\langle x_i \rangle_{SS,1} - \langle x_i \rangle_{SS,2}} \tag{4.29}$$

In this case the system starts in steady state 1. The system parameters are then changed and relaxation occurs towards a new steady-state 2. Typically a single parameter value is increased by 5%. In the reversible dimerization model for example only k_2 is modified.

In the following sections the eigenvalues for several simple models across a range of kinetic values are obtained. We evaluate the accuracy of the eigenvalue analysis by comparing the correlation functions calculated via ZI-Closure (Equation 4.25) to the correlation functions calculated by kinetic Monte Carlo sampling). We then confirm fluctuation-dissipation congruency relations between steady state correlations and non-steady state linear responses.

To assess the accuracy of the methods, results from ZI-Closure are compared to results from kinetic Monte Carlo simulations. The kinetic Monte Carlo results are obtained from 10^7 stochastic simulation trajectories. This ensemble size ensures sampling convergence for probability distributions of small chemical reaction networks[72]. The stochastic simulations were conducted using the Hy3S algorithm [72].

4.3 Example Calculation

The goal of this section is to provide a numerical example to help illustrate some of the calculations previously developed in the chapter. A single simple system was chosen, reversible dimerization, and analysis is performed all the way from generating moment equations to producing eigenvalues.

4.3.1 Moment Matrix Generation

The first step in the analysis is producing the correct moment equations. The kinetic equations are:

$$\begin{aligned} 2A &\xrightarrow{k_1} B \\ B &\xrightarrow{k_2} 2A \end{aligned}$$

There is only one free variable (A) because given an initial monomer count ($S_0 = A+2B$) there is a mass conservation relation ($B = \frac{1}{2}(S_0 - A)$). The CME is as follows:

$$\begin{aligned} \frac{\partial P(A;t)}{\partial t} = k_1 \frac{(A+2)(A+1)}{2} P(A+2;t) \\ + k_2 \frac{S_0 - (A-2)}{2} P(A-2;t) - \left[k_1 \frac{A(A-1)}{2} + k_2 \frac{S_0 - A}{2} \right] P(A;t) \end{aligned}$$

After some manipulation the final moment equations are:

$$\frac{\partial \langle A \rangle}{\partial t} = -k_1 \langle A^2 \rangle + (k_1 - k_2) \langle A \rangle + k_2 S_0$$

$$\frac{\partial \langle A^2 \rangle}{\partial t} = -2k_1 \langle A^3 \rangle + 2(2k_1 - k_2) \langle A^2 \rangle + 2(k_2 S_0 - k_1 - k_2) \langle A \rangle + 2k_2 S_0$$

Ultimately these equations can be put into its final form (Equation 4.7):

$$\frac{\partial}{\partial t} \begin{bmatrix} \langle A^0 \rangle \\ \langle A \rangle \\ \langle A^2 \rangle \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ k_2 S_0 & (k_1 - k_2) & -k_1 \\ 2k_2 S_0 & 2(k_2 S_0 - k_1 - k_2) & 2(2k_1 - k_2) \end{bmatrix} \begin{bmatrix} \langle A^0 \rangle \\ \langle A \rangle \\ \langle A^2 \rangle \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -2k_1 \end{bmatrix} \langle A^3 \rangle$$

For the rest of the example we will use $k_1 = 0.1$ (1/molecules-s) and $k_2 = 1$ (1/s) with an initial monomer count of $S_0 = 20$ molecules. The numerical moment equations are thus:

$$\frac{\partial}{\partial t}\begin{bmatrix} \langle A^0 \rangle \\ \langle A \rangle \\ \langle A^2 \rangle \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 20 & -0.9 & -0.1 \\ 40 & 37.8 & -1.6 \end{bmatrix} \begin{bmatrix} \langle A^0 \rangle \\ \langle A \rangle \\ \langle A^2 \rangle \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -0.2 \end{bmatrix} \langle A^3 \rangle$$

4.3.2 Steady-State Jacobian and Correlation Functions

The first step to determining J_{SS} is to obtain the steady-state distribution, which in this case has parameter values of $\underline{\alpha}_{SS} = [\alpha_0, \alpha_1, \alpha_2] = [8.732, -1.492, 0.074]$. The resulting lower-order moments are $\langle A^0 \rangle = 1$, $\langle A \rangle = 10.1$ molec., and $\langle A^2 \rangle = 109.0$ molec.^{2}.

Given the form of J_{SS} in Equation 4.9 the following is true:

$$J_{SS} = \begin{bmatrix} 0 & 0 & 0 \\ k_2 S_0 & (k_1 - k_2) & -k_1 \\ 2k_2 S_0 - 2k_1 \frac{\partial \langle A^3 \rangle}{\partial \langle A^0 \rangle} & 2\left(k_2 S_0 - k_1 - k_2\right) - 2k_1 \frac{\partial \langle A^3 \rangle}{\partial \langle A \rangle} & 2\left(2k_1 - k_2\right) - 2k_1 \frac{\partial \langle A^3 \rangle}{\partial \langle A^2 \rangle} \end{bmatrix}$$

$$J_{SS} = \begin{bmatrix} 0 & 0 & 0 \\ 20 & -0.9 & -0.1 \\ 40 - 0.2 \frac{\partial \langle A^3 \rangle}{\partial \langle A^0 \rangle} & 37.8 - 0.2 \frac{\partial \langle A^3 \rangle}{\partial \langle A \rangle} & -1.6 - 0.2 \frac{\partial \langle A^3 \rangle}{\partial \langle A^2 \rangle} \end{bmatrix}$$

So we need to determine three unknowns: $\frac{\partial \langle A^3 \rangle}{\partial \langle A^0 \rangle}$, $\frac{\partial \langle A^3 \rangle}{\partial \langle A \rangle}$, and $\frac{\partial \langle A^3 \rangle}{\partial \langle A^2 \rangle}$. The key to obtaining values for these unknowns is the maximum entropy distribution:

$$P_H(A) = exp\left(-\alpha_0 - \alpha_1 A - \alpha_2 A^2\right)$$
$$\langle A^0 \rangle = \textstyle\sum_{A=0}^{S_0} P_H(A)$$
$$\langle A \rangle = \textstyle\sum_{A=0}^{S_0} A P_H(A)$$
$$\langle A^2 \rangle = \textstyle\sum_{A=0}^{S_0} A^2 P_H(A)$$

Thus the Jacobians that are given by the steady-state Newton-Raphson algorithm (Equation 4.14) have particular values (note: $\langle A^3 \rangle = 1239.0$ molec.^{3}, $\langle A^4 \rangle = 14740.2$

molec.^{4}, and $\langle A^5 \rangle = 182569.0$ molec.^{5}):

$$\left[\begin{array}{c} J_{NR} \\ \hline J'_{NR} \end{array}\right] = \left[\begin{array}{ccc} \frac{\partial\langle x^0\rangle}{\partial\alpha_0} & \frac{\partial\langle x^0\rangle}{\partial\alpha_1} & \frac{\partial\langle x^0\rangle}{\partial\alpha_2} \\ \frac{\partial\langle x\rangle}{\partial\alpha_0} & \frac{\partial\langle x\rangle}{\partial\alpha_1} & \frac{\partial\langle x\rangle}{\partial\alpha_2} \\ \frac{\partial\langle x^2\rangle}{\partial\alpha_0} & \frac{\partial\langle x^2\rangle}{\partial\alpha_1} & \frac{\partial\langle x^2\rangle}{\partial\alpha_2} \\ \hline \frac{\partial\langle x^3\rangle}{\partial\alpha_0} & \frac{\partial\langle x^3\rangle}{\partial\alpha_1} & \frac{\partial\langle x^3\rangle}{\partial\alpha_2} \end{array}\right] = -\left[\begin{array}{ccc} 1 & 10.1 & 109.0 \\ 10.1 & 109.0 & 1239.0 \\ 109.0 & 1239.0 & 14740.2 \\ \hline 1239.0 & 14740.2 & 182569.0 \end{array}\right]$$

The required unknowns are determined numerically (Equation 4.16):

$$J'_{NR}J_{NR}^{-1} = \left[\begin{array}{ccc} \frac{\partial\langle A^3\rangle}{\partial\langle A^0\rangle} & \frac{\partial\langle A^3\rangle}{\partial\langle A\rangle} & \frac{\partial\langle A^3\rangle}{\partial\langle A^2\rangle} \end{array}\right] = \left[\begin{array}{ccc} 828.2 & -286.3 & 30.3 \end{array}\right]$$

And a single calculation then produces numerical values for Equations 4.17:

$$J_{SS} = \left[\begin{array}{ccc} 0 & 0 & 0 \\ 20 & -0.9 & -0.1 \\ 40 - 0.2 \cdot 828.2 & 37.8 + 0.2 \cdot 286.3 & -1.6 - 0.2 \cdot 30.3 \end{array}\right] = \left[\begin{array}{ccc} 0 & 0 & 0 \\ 20 & -0.9 & -0.1 \\ -125.6 & 95.1 & -7.7 \end{array}\right]$$

This matrix J_{SS} represents the approximate dynamics of the moments near the steady-state.

The final step is to obtain eigenvalues and eigenvectors and get the analytical solution to the autocorrelation function for A. First, we establish the correlation vector for this system (Equation 4.22):

$$\underline{C}_A(t) = \left[\begin{array}{c} C_{A,A}(t) \\ C_{A^2,A}(t) \end{array}\right] = \left[\begin{array}{c} \langle (A(t) - \langle A(t)\rangle)(A(0) - \langle A(0)\rangle)\rangle \\ \left\langle \left(A(t)^2 - \left\langle A(t)^2 \right\rangle\right)(A(0) - \langle A(0)\rangle)\right\rangle \end{array}\right]$$

The initial conditions for the correlation functions with a stationary distribution is (Equation 4.24):

$$\left[\begin{array}{c} C_{A,A}(0) \\ C_{A^2,A}(0) \end{array}\right] = \left[\begin{array}{c} \langle A^2\rangle - \langle A\rangle^2 \\ \langle A^3\rangle - \langle A^2\rangle\langle A\rangle \end{array}\right]\Bigg|_{SS} = \left[\begin{array}{c} 6.77 \\ 136.9 \end{array}\right]$$

Based on Equation 4.19 the correlation functions are calculated as follows:

$$\frac{\partial}{\partial t}\left[\begin{array}{c} C_{A,A} \\ C_{A^2,A} \end{array}\right] = J_{SS}\left[\begin{array}{c} C_{A,A} \\ C_{A^2,A} \end{array}\right] = \left[\begin{array}{cc} -0.9 & -0.1 \\ 95.1 & -7.7 \end{array}\right]\left[\begin{array}{c} C_{A,A} \\ C_{A^2,A} \end{array}\right]$$

Note that only the bottom right portion of J_{SS} is used since the zeroth-moment disappears from the analysis. But the zeroth Legrange multiplier is necessary when initially determining J_{SS} in the previous section. The eigenvalues and eigenvectors can thus be determined:

$$\lambda_{S,1} = -2.891 \quad \lambda_{S,2} = -5.674 \quad \underline{\nu}_{S,1} = \begin{bmatrix} 0.050 \\ 0.999 \end{bmatrix} \quad \underline{\nu}_{S,2} = \begin{bmatrix} 0.021 \\ 0.999 \end{bmatrix}$$

This makes Equation 4.23:

$$\begin{bmatrix} C_{A,A} \\ C_{A^2,A} \end{bmatrix} = \begin{bmatrix} 0.050 & 0.021 \\ 0.999 & 0.999 \end{bmatrix} \begin{bmatrix} e^{-2.891t} & 0 \\ 0 & e^{-5.674t} \end{bmatrix} \begin{bmatrix} 34.2 & -0.716 \\ 34.2 & 1.72 \end{bmatrix} \begin{bmatrix} 6.77 \\ 136.9 \end{bmatrix}$$

The final analytical equation for the normalized autocorrelation function of component A is thus (Equation 4.25):

$$C_{A,A}(t) = (0.4718)\, exp(-2.891 \cdot t) + (0.5282)\, exp(-5.674 \cdot t)$$

The individual D_i values (pre-normalization) can be back calculated as $D_1 = 0.067$ (molec.^{2}) and $D_2 = 0.075$ (molec.^{2}).

4.4 Results

We examine four small reaction network models: a simple reversible dimerization system (Model 1), a two component Michaelis-Menten system (Model 2)[118], the Schlögl model (Model 3)[119], and an N-stage linear cycle. Each model serves a particular purpose in showing that the proposed analysis of stochastic chemical reacting systems is accurate and generally applicable. The description, initial conditions and parameters for each model are provided in Table 4.1.

Proof of concept that ZI-Closure and eigenvalue analysis capture nonlinear stochastic chemical dynamics can be demonstrated with the reversible dimerization model. The reversible dimerization system requires no more than 4th-order moment closure in order to accurately construct the probability distribution [116]. Consequently, the derivation of the Jacobian is straightforward and was detailed analytically previously. The eigenvalues of the Jacobian are trivial to compute using Equation 4.18, leading to the calculation of autocorrelation functions.

Table 4.1: Model descriptions for Demonstration of Moment-Based Non-Linear Analysis

Model	(1) Reversible Dimerization	(2) Michaelis-Menten	(3) Schlögl	(4) n-Stage Linear Cycle
Reactions	$2A \xrightarrow{k_1} B$ $B \xrightarrow{k_2} 2A$	$S+E \xrightarrow{k_1} S:E$ $S:E \xrightarrow{k_2} S+E$ $S:E \xrightarrow{k_3} P+E$ $P+E \xrightarrow{k_4} S:E$	$2X+A \xrightarrow{k_1} 3X$ $3X \xrightarrow{k_2} 2X+A$ $B \xrightarrow{k_3} X$ $X \xrightarrow{k_4} B$	$X_1 \xrightarrow{k_1} X_2$ $X_2 \xrightarrow{k_2} X_3$ $\vdots$ $X_N \xrightarrow{k_N} X_1$
Parameters	Specified in text	k_1, k_2 specified in text $k_3 = 1$ (1/s) $k_4 = 0.1$ (1/molc.-s)	$k_1 \cdot A = 0.15$ (1/molec-s) $k_2 = 0.0015$ (1/molec^2-s) $k_3 \cdot B = 20$ (molec/s) k_4 specified in text	$k_i = N(1/s)$ i= [1, 2, … N]
Initial Conditions (molecules)	$A_0 = 10$	$S_0 = 100$, $E_0 = 50$ $S:E_0 = 0$, $P_0 = 0$	$X_0 = 25$	$X_1 = 20$ $X_{i\neq 1} = 0$

In Figure 4.1A, autocorrelation functions for component A, $C_{AA}(t)$, in Model 1 are plotted for varying kinetic constant values. These functions are calculated via eigenvalue analysis (solid black lines) and kinetic Monte Carlo sampling (shapes). The results clearly demonstrate the accuracy of eigenvalue analysis based on ZI-Closure.

Not unexpectedly, the decay rate of probability moment correlations decreases with increasing kinetic constant values. The relaxation of autocorrelation functions for this simple system is captured by a single exponential term, whose exponent is the dominant eigenvalue of the Jacobian matrix. The inset to Figure 4.1A shows this dominant eigenvalue calculated across the chosen range of kinetic parameters for situations where $k_2 = k_1$ (blue) and $k_2 = 1$ (1/s) (red). Note that eigenvalues change smoothly with kinetic parameter changes, and linearly when the numerical values of k_2 and k_1 are kept numerically identical. Importantly, the eigenvalues are negative real numbers in all studied ranges of kinetic constants. Because the deterministic model is always stable, it is unsurprising that the stochastic steady states of reversible dimerizations are always stable as well.

In Figure 4.1A, the moment response curve subject to perturbation is also plotted (x's). In accordance with the fluctuation-dissipation theorem, the non-steady state response function is congruent to the steady state correlation function. In other words, fluctuations at steady state dictate the response of a reaction network to small external perturbations. We find this to be true for any tested parameter combination.

Figure 4.1B focuses on the response of the probability distribution to perturbation for the case of $k_1 = 1$ (1/molec.-s) and $k_2 = 1$ (1/s). In this example, the system is

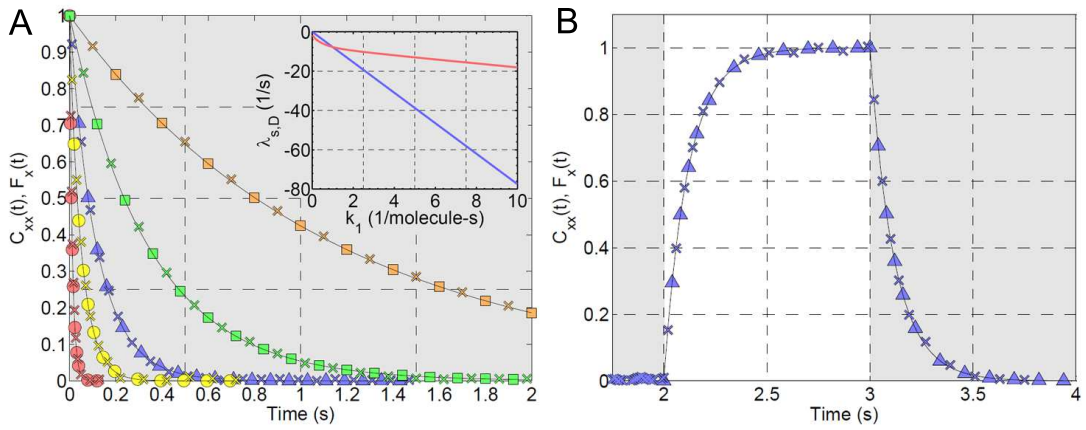

Figure 4.1: (A) Moment correlation functions for the reversible dimerization reaction network. Autocorrelation functions are plotted for five kinetic parameter sets using 4th-order ZI-Closure (solid black lines) and kinetic Monte Carlo results (shapes). The parameters are [k_1 (1/s-molec.), k_2 (1/s)] = [10, 10] (blue circles), [10, 1] (red circles), [1, 1] (purple triangles), [0.1, 1] (red squares), and [0.1, 0.1] (blue squares). Also shown is the response function of the reaction network to an external perturbation (F_x, x's). The inset shows the dominant Jacobian eigenvalue across a range of values for k_1 where $k_2 = k_1$ (blue) or $k_2 = 1$ (1/s). (B) Congruency between response and correalation functions. A closer look at the system's response function to an external perturbation (x symbols). Initially the rate constants are [k_1 (1/s-molec.), k_2 (1/s)] = [1, 1]. At time = 2 (s) the parameters change to [k_1 (1/s-molec.), k_2 (1/s)] = [1, 1.05] (white area). At time = 3 (s) the parameters return to their initial values and the system relaxes back to its initial distribution. Also shown are autocorrelation function results from kinetic Monte Carlo (blue triangles) and ZI-Closure(4th-order, black line).

perturbed away from the first steady state at $t = 2$, allowed to relax to a second steady state, and then perturbed back at $t = 3$. The two responses are congruent, if inverted, confirming that the system exhibits a linear response. All perturbation results presented herein exhibit similar behavior.

Analysis of the Michaelis-Menten reaction network demonstrates that the presented theory is applicable to multi-component systems. This reaction network has two independent components, the substrate S and the enzyme E. This example system exhibits symmetry, such that the eigenvalues and eigenvectors computed when changing k_1 and k_2 are identical to those computed when changing k_4 and k_3, respectively. Because of this, presented results are limited to ranging k_1 and k_2.

Figures 4.2A show the autocorrelation function for the substrate S, $C_{SS}(t)$, calculated with eigenvalue analysis (solid black lines). The same autocorrelation function is calculated with kinetic Monte Carlo sampling (shapes). In addition, the response of the mean substrate $< S >$, $F_S(t)$, is shown when the system is subjected to a small pertubation (x's). Figure 4.2B shows the corresponding results for the enzyme, E. Again, steady-state correlations, computed by either ZI-Closure or kinetic Monte Carlo, and response functions display relaxation congruency.

The inset to Figure 4.2B shows the dominant eigenvalues calculated for a range of k_1 and k_2 parameters. In order to generate the inset plot, nearly 1,000 steady states were calculated varying the kinetic constant values, and the eigenvalues of steady-state Jacobians were computed. This detailed description of the stochastic dynamics of a Michaelis-Menten model is produced in a matter of seconds with ZI-Closure, in contrast to an equivalent calculation using kinetic Monte Carlo simulations, which requires many CPU hours. Again, the eigenvalues of the Jacobian are real negative numbers for physically relevant kinetic constant values. Consequently, the reaction network steady-state probability distribution is stable.

The final nonlinear model is the Schlögl model. As in Chapter 3, the Schlögl model is studied because of its underlying bimodal distribution. Figure 4.3A shows that with slight changes to a single kinetic parameter (here k_4) a range of interesting distributions can be sampled. At high kinetic constant values, $k_4 > 3.8$ (1/s), a single peak is observed at high numbers of molecules of X. There is a transition range from $k_4 = 3.8$ to 3.2 (1/s) where the system exhibits a bimodal distribution. At low values, $k_4 < 3.2$, a single

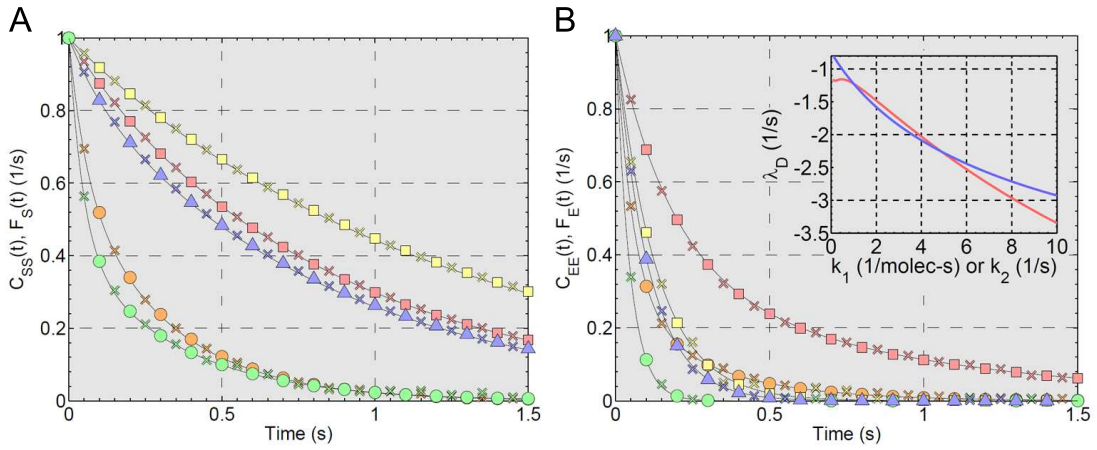

Figure 4.2: Moment correlation functions for the Michaelis-Menten reaction network. Five kinetic parameter sets were chosen and the normalized autocorrelation function plotted for both kinetic Monte Carlo results (shapes) and using 4th-order ZI-Closure (solid black lines). Results are shown for both S (Figure 4.2A, substrate) and E (Figure 4.2B, enzyme) components. The base parameter values are [k_1 (1/molec.-s), k_2 (1/s), k_3 (1/s), k_4 (1/molc.-s)] = [0.1, 1, 1, 0.1] (purple triangles). Because of the network symmetry we only varied two parameters, as follows: $k_1 = 1$ (red circle), $k_1 = 0.01$ (red squares), $k_2 = 10$ (blue circles), $k_2 = 0.1$ (blue squarse). The inset shows the dominant eigenvalue (λ_D) as both k_1 (red line) and k_2 (blue line) are ranged.

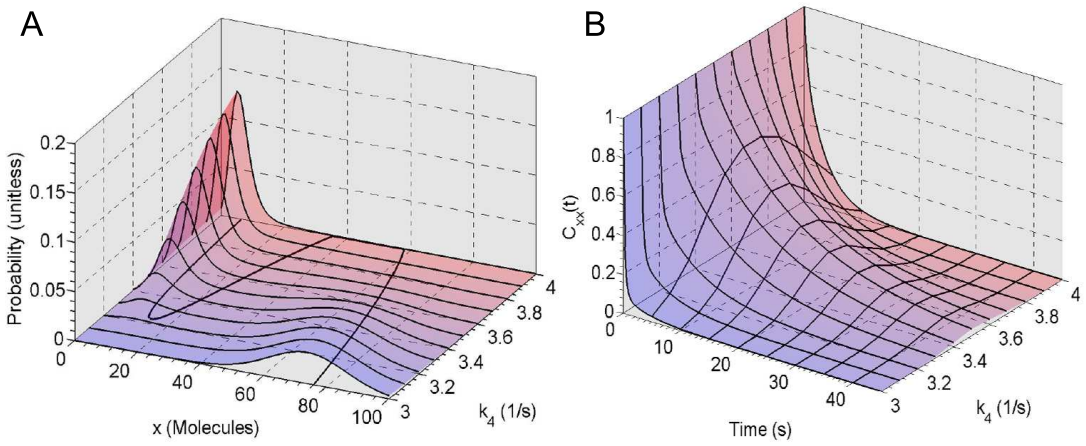

Figure 4.3: Steady-state probability distributions for the Schlögl model across a range of values of k_4 (1/s). The system transitions from a single high-value peak (k_4 = 3.0-3.2), to a bimodal distribution (k_4 = 3.2-3.8), and back to a single low-value peak (k_4 = 3.8-4.0). The solution of the deterministic Schlögl model is also shown on the X-k_4 plane (black solid line). It is apparent that within a range of k_4 values there are three solutions, the middle of which is unstable; (B) Normalized autocorrelation functions calculated for the range of parameter k_4 using 9th-order ZI-Closure. The correlation functions reveal slower dynamics in the bimodal region as compared to the single peaked region.

peak is observed at low numbers of molecules of X. Because of its complex form, the probability distribution is captured accurately with 9th-order ZI-Closure [116]. On the X-k_4 plane the steady-state results for the deterministic model are also plotted as a solid black line.

The autocorrelation functions for the X component, $C_{XX}(t)$, are computed for numerous steady states, varying the value of k_4 and shown in Figure 4.3B. Of note is that at the extremes of parameter values ($k_4 > 3.8$, $k_4 < 3.2$) the autocorrelation function relaxes rapidly as a single exponential term, whereas in the bimodal region (around k_4 = 3.5) the autocorrelation relaxes more slowly.

Looking more closely at three of the kinetic constant values (k_4 = 3.0 (1/s) (blue), k_4 = 3.5 (1/s) (purple), and k_4 = 4.0 (1/s) (red)) in Figure 4.4 we observe the fast drop in the autocorrelation function for single-peaked distributions (red and blue) compared to the bimodal distribution (purple). This behavior is confirmed by the kinetic Monte

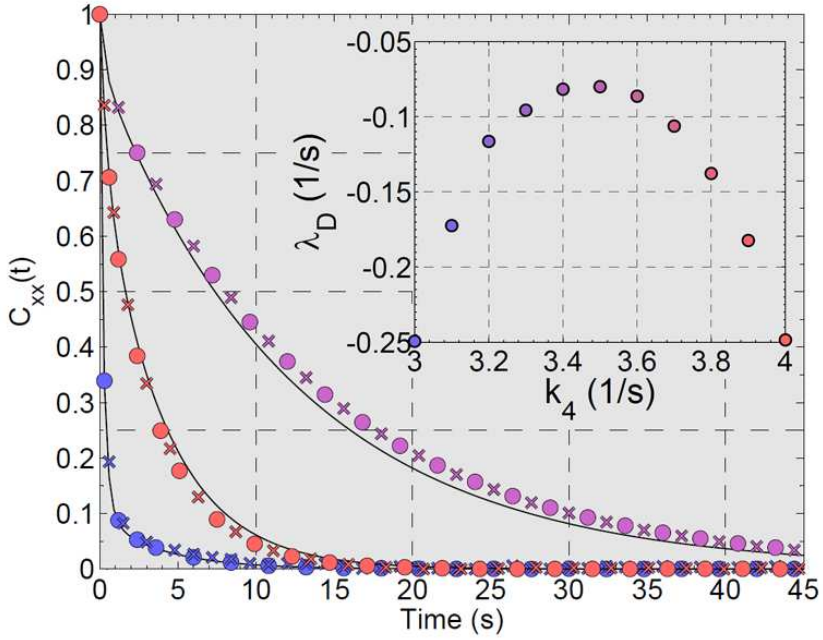

Figure 4.4: Schlögl model correlation and and response functions. Results are shown for the three extreme points of the chosen kinetic range ($k_4 = 3.0$ (1/s) as blue circles, $k_4 = 3.5$ (1/s) as purple circles, and $k_4 = 4.0$ (1/s) as red circles). The 9th-order ZI-Closure results are shown as solid black lines, the kinetic Monte Carlo results are shown as circles and the response functions are shown with x's. The inset shows the dominant eigenvalue determined with 9th-order ZI-Closure. Slower dynamic behavior is observed in the center of the kinetic range (the bimodal range for the probability distribution) than at either end of the range (the unimodal range).

Carlo autocorrelation functions (shapes) and the moment response curves subject to perturbation (x's).

Looking at the largest, dominant eigenvalues (Figure 4.4 inset) a parabolic shape emerges for the transition between unimodal and bimodal distrbutions. The system dynamics are faster when the distribution is unimodal than when it is bimodal. This may be attributed to the time necessary for a system to cross the barrier and transition from one probability mode to another.

In all previous examples, the eigenvalues of the Jacobian matrix are negative real numbers, resulting in monotonic decays for the correlation functions. In order to investigate periodic dynamic behaviors we model a cyclical system, Model 4. In this linear

cycle a large number of first-order reactions are linked to form a cycle (with size N, and $k_i = N$ (1/s) such that the average cycle time is always one second). The moment equations for such a system are analytically solvable and thus autocorrelation functions can be produced without moment closure. In Figure 4.5 the power spectral densities (see [117]) are shown for $N = 10$ (green) and $N = 50$ (purple). The spectral densities are computed as a Fourier transform of autocorrelation functions. The moment equation system (solid black lines) and kinetic Monte Carlo (circles) results show a good match. Note that the imaginary part of the eigenvalues (marked out as dashed lines) match up exactly with the frequency of oscillation for the system.

In particular, for $N = 10$, the peak corresponds well with the calculated eigenvalue. For $N = 50$, two eigenvalues are marked which correspond to the first two power-spectral density peaks. For $N = 50$ there are, in fact, 12 unique imaginary eigenvalue parts (frequency modes) calculated with moment equations. This suggests higher order modes may be discernible with moment equation analysis, even when they are impossible to determine numerically. The inset shows the eigenvalues for the cycle as N is increased from 3 to 60, with $N = 10$ and 50 highlighted. As expected, as the cycle size increases, the imaginary part of the dominant eigenvalue converges onto one cycle.

4.5 Discussion

An important step toward understanding nonlinear stochastic systems is the calculation of steady-state probability distributions. A panoply of tools, first developed for stability analysis of deterministic dynamic systems, can then find use in stochastic dynamic systems.

The steady-state probability distribution is directly obtained with ZI-Closure for small reaction networks. This calculation is challenging with kinetic Monte Carlo algorithms, such as Gillespie's stochastic simulation algorithm, not least because there is no certainty that an ensemble of stochastic trajectories has indeed relaxed to steady-state.

Probability moments are computed up to a finite order with ZI-Closure, with higher-order moments computed by maximizing the probability distribution information entropy. To our knowledge, this is the first time that a master equation closure scheme has be used for stability analysis of nonlinear stochastic chemical reactions.

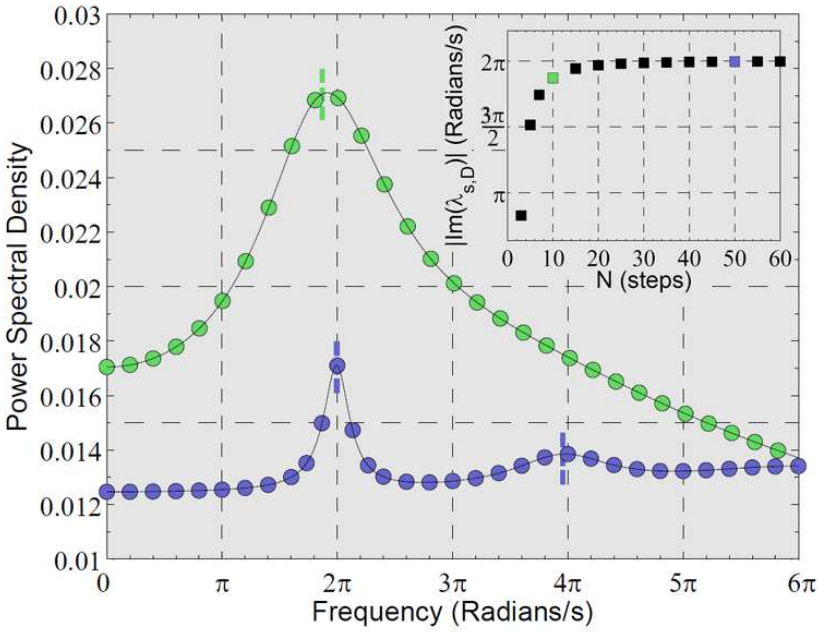

Figure 4.5: Linear cycle reaction network. Results are for cycle sizes of $N = 10$ (green) and $N = 50$ (purple) (all steps having a rate constant of $k = N(1/s)$). Results are obtained for both kinetic Monte Carlo results (black solid line) and calculations via moment equations (dots). The imaginary part of the dominant eigenvalues (dashed colored lines) are also shown. With $N = 10$, the peak corresponds well with the calculated eigenvalue. For $N = 50$, two eigenvalues are marked which correspond to the first two power spectral density peaks. For $N = 50$ there are, in fact, 12 unique imaginary eigenvalue parts (frequency modes) calculated with the moment equations. This suggests higher order modes may be discernible with moment equation analysis, even when they are impossible to determine numerically. The inset shows the imaginary part to the dominant eigenvalue for a range of N values asymtotically approaching 2π as N goes to infinity.

Nonlinear stochastic chemical reaction dynamics can now be linearized around the steady state. With the resulting Jacobian matrix a stability analysis is possible around steady-state probability distributions. The eigenvalues of the Jacobian provide direct insight into the dynamic behavior and stability of chemical reacting systems and into possible oscillatory behaviors.

With Jacobian eigenvalues and eigenvectors at steady state, the autocorrelation functions for the chemical components can be computed. These correlations capture the entirety of dynamic features of nonlinear reaction networks. Usuful information can be obtained about the time scales of stochastic processes at or near steady states.

In the examples shown, the eigenvalues decrease smoothly with increasing kinetic constants. As expected, higher kinetic constant values result in faster reaction kinetics, captured quantitatively by eigenvalues.

All calculated eigenvalues presented herein have real negative parts. Consequently, the computed steady state probability distributions are all stable. The reversible dimerization and Michaelis-Menten models have a single stable steady-state in the deterministic models, so this should come as no suprise.

For the Schlögl model, the stability of stochastic nonlinear dynamics is less apparent. The deterministic Schlögl model has three steady states for a range of parameter values, as shown in Figure 4.3A. Two of the steady states are stable, as determined using Lyapunov methods, but the third is not. In notable contrast, the stochastic Schlögl model always has a single stable, steady-state probability distribution. This probability distribution at steady state may be bimodal but the Schlögl model exhibits ergodicity and visits the two distinct modes with a frequency proportional to their probability. The time scales the system resides within each mode and the time scale for crossing through the probability trough are captured by time correlation functions of probability moments at steady state.

This is an important difference between deterministic and stochastic modeling formalisms with significant implications on the relevance of different modeling formalisms. It appears that chemical reacting systems away from the thermodynamic limit may indeed be observed in states that according to deterministic models are impossible to observe.

We find that all autocorrelation functions determined using eigenvalue analysis

match the autocorrelation functions determined with kinetic Monte Carlo simulations. However kinetic Monte Carlo simulations require simulating up to 10^7 stochastic trajectories in order to obtain equally acurate results. Such simulations require many CPU hours, even for the small reaction networks studied here. For ZI-Closure and the subsequent analysis, only one optimization step is necessary for the information entropy. As long as the size of the discrete state space of reaction network can be enumerated, ZI-Closure may outperform kinetic Monte Carlo simulations. For larger reaction networks or for larger numbers of molecules than the ones studied here, ZI-Closure will be computationally challenging to implement. ZI-Closure may also be computationally challenging to implement in cases of complex probability distributions. Eigenvalue analysis of the Schlögl model using 9th-order ZI-Closure, for example, required the accurate calculation of 18th-order moments. While higher-order closure would have been desirable (12th-order closure was used in Chapter 3), numerical challenges made such analysis impractical. In all these cases when ZI-Closure is difficult to implement, model and state space reduction techniques may prove useful.

Intriguingly, steady state autocorrelation functions are found to be congruent to perturbation responses. This means that steady-state fluctuations provide information not only on the stability of a system around the steady state, but also on how the system will respond dynamically to both external forcing functions and parameter variations. In linear response regimes, the fluctuation-dissipation theorem predicts such congruency, but to our knowledge this has never before been shown for nonlinear stochastic reactions.

We believe that the analysis is in principle applicable to any nonlinear stochastic dynamic system for which the master probability equation can be expressed in terms of probability moments.

4.6 Conclusions

ZI-Closure solves the chemical master equation of small nonlinear reaction networks. The steady state of nonlinear stochastic chemical reaction networks can be calculated with ZI-Closure in a single step maximization of the system's information entropy. The alternative, with a significant computational cost, has been to sample the probability with kinetic Monte Carlo simulations.

Nonlinear dynamics can be linearized around the steady-state probability distribution with ZI-Closure. Stability analysis tools, such as Lyapunov's first method, can then be employed to reveal the stability characteristics of stochastic nonlinear reacting systems. Parameter sensitivity analysis is possible and the design of control systems for nonlinear reactions under the influence of noise becomes conceivable. With these theoretical advances, not possible before the development of ZI-Closure, small chemical and biochemical systems may be rigorously studied. The possibility also opens up for designing stochastic reaction control systems.

Important differences are observed with this analysis between the stability of deterministic and stochastic modeling formalisms. In stochastic systems, steady states may be observed with finite probability, whereas these states may be unstable according to deterministic models, and thus practically impossible to attain. These are important considerations that impact the choice of modeling formalisms used in capturing the dynamic behavior of small reacting systems, for example ones often observed in biological systems.

Chapter 5

Extending Power Spectral Density Generation to Non-Linear Stochastic Chemical Reaction Networks

5.1 Introduction

The two initial chapters (Chapters 2 and 3) of this thesis focused on the efficient generation of a large set of moment equations and the subsequent development of a novel moment closure scheme, ZI-Closure, that holds the potential to make the practical use of moment closure in biological simulation a reality. Chapter 4 began to refocus the thesis on new analytical methods using ZI-Closure by introducing non-linear analysis of stochastic chemical reaction models. These exciting results, showing that producing eigenvalues that represent the underlying probability distribution dyamics is possible for stochastic systems, opened the door to further expansion of moment-closure techniques into other common analytical methods.

As mentioned previously Monte Carlo simulation methods have become an important part of stochastic simulation for chemical reaction systems. However, there are still many aspects of these methods which remain underwhelming. In particular, Monte

Carlo simulations cannot immediately produce steady-state results. This drawback can result in wasting CPU resources on dynamic simulations when steady-state results are of sole interest to researchers. This is where ZI-Closure is able to exploit its deterministic approach to stochastic simulation, by producing steady-state results quickly and performing analysis *a priori*. Chapter 4 focused on the production of eigenvalues for stochastic systems, an exciting development in stochastic simulation of chemical reaction network models.

Here we build off of Chapter 4 by utilizing the stochastic eigenvalue analysis using ZI-Closure to produce analytical power spectral density (PSD) diagrams. Previously, Warren was able to show that for linear systems PSD diagrams can be produced analytically using eigenvalue analysis[117]. In that same publication it was also suggested that by linearizing the moment equations one could theoreticaly produce such diagrams for non-linear systems as well. This result is the subject of this study intended to show that by using ZI-Closure accurate PSD diagrams can be analytically produced, even for complex and especially oscillatory systems.

Power spectral density is one of the only methods in which researchers are able to effectively analyze the oscillations in a noisy system. Unlike in deterministic systems, a stochastic system exhibiting oscillatory behavior may not show clearly delimited peaks and troughs conducive to direct analysis. Indeed, in stochastic systems the underlying distribution can be stationary while the underlying Monte Carlo trajectories experience significant and sustained oscillatory behavior. Power spectral density diagrams transform the time dependant system into the fequency domain and can allow the determination of dominant frequency modes for even very noisy stochastic systems.

Oscillations in real chemical systems was a fairly recent discovery and continues to be an area of intense study from a theoretical perspective[112]. Oscillatory behavior is a desireable trait in synthetic biologyl and was, in fact, one of the first synthetic systems produced *in vivo*[49]. The ability to quickly and analytically determine PSDs could, in theory, lead to the optimization and careful calibration of oscillatory behavior in stochastic chemical models.

5.2 Background

In this section the chemical master equation, moment dynamics and ZI-Closure will be described as a foundation for extending this work to power spectral densities in the theory sections. Much of this section is a restatement of prior background and theory sections from Chapter 3 and 4. The equations are restated for clarity.

As mentioned previously stochastic chemical dynamics are, in theory, governed by a single equation called the chemical master equation (CME) first described by McQuarrie[37]:

$$\frac{\partial P\left(\underline{X};t\right)}{\partial t} = \int_{\underline{X}'} \left[T\left(\underline{X}|\underline{X}'\right) P\left(\underline{X}';t\right) T\left(\underline{X}'|\underline{X}\right) P\left(\underline{X};t\right)\right] \tag{5.1}$$

Where $P\left(\underline{X};t\right)$ is the probability of being in state $\underline{X}$ at time t, and $T\left(\underline{X}|\underline{X}'\right)$ is the transition probability of moving from state $\underline{X}'$ to state $\underline{X}$. As mentioned above this set of equations is often complex and mathematically intracable. By relating these transition probabilities to kinetic rate laws and assuming a well-mixed volume Gillespie was able to show (and later prove[38]) that a Monte Carlo algorithm, often referred to as the stochastic simulation algorithm (SSA), can exactly reproduce the solution to the CME for arbitrary chemical networks[44]. While invaluable to current stochastic chemical simulation the SSA cannot easliy perform analysis and an alternative deterministic simulation would be perferred.

In order to, instead, utilize the moments of the underlying probability distribution in the CME a set of dynamic moment equations needs to be calculated. These equations are deterministic and the generation of the equation set has been done numerically[105] and analytically[96]. Generally the CME can be equivalently represented as a complete set of moment equations, often written as:

$$\frac{\partial \underline{\mu}}{\partial t} = A\underline{\mu} + A'\underline{\mu}' \tag{5.2}$$

Where $\underline{\mu}$ is a moment vector of length N_M representing the moments up to order M in the system and $\underline{\mu}'$ is a moment vector of length N'_M representing the higher-order moments necessary to fully define the moment equations. The matrices A (size $N_M \times N_M$) and A' (size $N_M \times N'_M$) represent the linear and non-linear part of the moment dynamics respectively. If a system is linear then A' is empty and the moment equations

can be solved immediately with an ODE solver of choice. When the system is non-linear, there are second-order reactions or higher, A' is non-empty and a closure scheme is needed.

A closure scheme in its most general sense is simply a relationship between the higher and lower-order moments that will allow Equation 5.2 to be solved:

$$\underline{\mu}' = F\left(\underline{\mu}\right) \tag{5.3}$$

This function, F, can be either an analytical or numerical relationship. In the case of zero-information closure the relationship is numerical and based on Shannon's definition of entropy:

$$H = \sum_{x=0}^{\infty} -p\left(x\right) \ln p\left(x\right) \tag{5.4}$$

This definition is for a single component system on a discrete state space, but can be generalized for multi-component and continuous systems.

With a set of M known lower-order moments ($\underline{\mu}$) this entropy can be constrained and minimized using a Legrange multiplier method. This trivially results in a definition for a probability distribution as a function of these calculated Legrange multipliers:

$$p_H\left(x\right) = exp\left(\alpha_0 + \alpha_1 x + \cdots + \alpha_M x^M\right) \tag{5.5}$$

Where α_i represents the i_{th} Legrange multiplier. Once again, it is trivial to generalize Equation 5.5 to multi-component systems. This distribution satisfies Equation 5.3 by allowing for the determination of higher-order moments:

$$\mu_i' = \sum_{x=0}^{\infty} f_{\mu_i'}\left(x\right) p_H\left(x\right) \tag{5.6}$$

Where $f_{\mu_i'}\left(x\right)$ is the function representing the i_{th} higher-order moment (e.g. for $\left\langle x^M \right\rangle$ the function $f_{x^M}\left(x\right) = x^M$). This method closes Equation 5.2 and is called zero-information closure because it assumes no knowledge of the system beyond its lower-order moments. The accuracy and versatility of this method was established in Chapter 3.

While Equation 5.2 can be solved using an ODE solving method, it is also possible to very quickly determine steady states by instead solving:

$$\underline{0} = A\underline{\mu} + A'\underline{\mu}' \tag{5.7}$$

Instead of utilizing a set of known lower-order moments in order to determine higher-order moments, the null space of the augmented matrix $B = [A|A']$ can be used to instead determine a set of steady-state lower-order moments $\underline{\mu}_{SS}$. These methods, while only currently applicable to small simple reaction networks due to computational concerns, are potentially far faster than SSA methods for determining steady-state distributions. It is this steady-state method that is utilized in the theory section that follows.

5.3 Theory

In the following section the ZI-Closure scheme described in the background section is extended to allow for the exact determination of power spectral densities for a stochastic chemical reaction simulation. We first establish the linearization of the moment equations around the steady-state calculated in Equation 5.7 and the equivalence of these equations for correlation functions. Solving the eigenproblem then allows for the exact determination of power spectral densities for non-linear stochastic chemical reaction simulation.

When the steady-state moments, $\underline{\mu}_{SS}$, are calculated via Equation 5.7 an implicit relation between the higher and lower-order moments are established with the Legrange parameters (α_i). Using this relationship Equation 5.2 can be linearized around the steady-state with a calculated Jacobian (J_{SS}):

$$\frac{\partial \underline{\mu}}{\partial t} = J_{SS}\underline{\mu} \tag{5.8}$$

This solution is only valid near the steady-state moments $\underline{\mu}_{SS}$.

It has been established that for a linear system the correlation function dynamics can be calculated using the same matrix as in Equation 5.2:

$$\frac{\partial C}{\partial t} = AC \tag{5.9}$$

Note that A' is not included since the system is linear. Here C is a correlation matrix where:

$$C_{i,j} = \left\langle \left(f_{\mu_i}(x) - \langle f_{\mu_i}(x)\rangle\right)\left(f_{\mu_j}(x) - \left\langle f_{\mu_j}(x)\right\rangle\right)\right\rangle \tag{5.10}$$

As an example consider a single component system, then $C_{2,1} = \left\langle \left(x^2 - \left\langle x^2 \right\rangle\right)(x - \langle x \rangle) \right\rangle$. This has been suggested, although not demonstrated, that for non-linear systems an equivalent equations holds as well[117]:

$$\frac{\partial C}{\partial t} = J_{SS} C \tag{5.11}$$

The Jacobian, J_{SS}, can be used to generate a set of eigenvalues (λ_i) and eigenvectors (ν_i), and subsequently with spectral decomposition a closed form solution for a correlation function can be calculated that takes the form:

$$C_{ij}(t) = \sum_{k=1}^{N_M} D_k exp(-\lambda_k t) \tag{5.12}$$

Where D_k is related to the initial conditions and eigenvectors.

It is from the closed form solution for the autocorrelation functions (Equation 5.12 where $i = j$) that power spectral densities can be determined using the relation:

$$S_x(\omega) = \hat{C}_{x,x}(\omega) \tag{5.13}$$

Here x refers to a chemical component of concern (although you can look at high-order autocorrelation functions to obtain higher-order power spectral densities if desired), and $\hat{C}(\omega)$ is the Fourier transform of the autocorrelation function. This function is defined on the frequency domain, ω. Ultimately, through Equation 5.13, a closed form solution for the power spectral density (PSD) can be obtained as well:

$$S_x(\omega) = \sum_{k=1}^{N_M} \frac{2 D_k \lambda_k}{\lambda_k^2 + \omega^2} \tag{5.14}$$

Note that for a non-normalized correlation function (units of molecules^2) the power spectral density has units of molecules^2/Hz. For the normalized correlation function these units are instead 1/Hz. Results presented herein tend to use the non-normalized version.

5.4 Example Calculation

In previous Chapters example calculations were presented in order to illustrate how results were generated. Like in Chapter 4 the reversible dimerization was chosen. While the results will be restated for clarity the initial presentation will be abridged. Refer to Section 4.3 for a more complete presentation.

5.4.1 Initial Calculations

Recall that the kinetic equations for the reversible dimerization model are:

$$\begin{aligned} 2A &\xrightarrow{k_1} B \\ B &\xrightarrow{k_2} 2A \end{aligned}$$

The moment equations as generated in Chapter 2 take the form:

$$\frac{\partial}{\partial t}\begin{bmatrix} \langle A^0 \rangle \\ \langle A \rangle \\ \langle A^2 \rangle \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ k_2 S_0 & (k_1 - k_2) & -k_1 \\ 2k_2 S_0 & 2\left(k_2 S_0 - k_1 - k_2\right) & 2\left(2k_1 - k_2\right) \end{bmatrix} \begin{bmatrix} \langle A^0 \rangle \\ \langle A \rangle \\ \langle A^2 \rangle \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -2k_1 \end{bmatrix} \langle A^3 \rangle$$

Take $k_1 = 0.1$ (1/molecules-s) and $k_2 = 1$ (1/s) with an initial monomer count of $S_0 = 20$ molecules. The numerical moment equations are:

$$\frac{\partial}{\partial t}\begin{bmatrix} \langle A^0 \rangle \\ \langle A \rangle \\ \langle A^2 \rangle \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 20 & -0.9 & -0.1 \\ 40 & 37.8 & -1.6 \end{bmatrix} \begin{bmatrix} \langle A^0 \rangle \\ \langle A \rangle \\ \langle A^2 \rangle \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -0.2 \end{bmatrix} \langle A^3 \rangle$$

The first step to determining the power spectral density is the determination of J_{SS}. To obtain this matrix it is necessary to obtain the steady-state distribution. This distribution has parameter values of $\underline{\alpha}_{SS} = [\alpha_0, \alpha_1, \alpha_2] = [8.732, -1.492, 0.074]$. The resulting lower-order moments are $\langle A^0 \rangle = 1$, $\langle A \rangle = 10.1$ molec., and $\langle A^2 \rangle = 109.0$ molec.^{2}.

Using these values the numerical J_{SS} is then calculated to be:

$$J_{SS} = \begin{bmatrix} 0 & 0 & 0 \\ 20 & -0.9 & -0.1 \\ 40 - 0.2 \cdot 828.2 & 37.8 + 0.2 \cdot 286.3 & -1.6 - 0.2 \cdot 30.3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 20 & -0.9 & -0.1 \\ -125.6 & 95.1 & -7.7 \end{bmatrix}$$

This matrix J_{SS} represents the approximate dynamics of the moments near the steady-state.

5.4.2 Autocorrelation Dynamics

In the previous chapter the eigenvalues and eigenvectors was used to obtain the auto-correlation dynamics for component A. Recall that the correlation vector for this system

is:

$$\underline{C}_A(t) = \begin{bmatrix} C_{A,A}(t) \\ C_{A^2,A}(t) \end{bmatrix} = \begin{bmatrix} \langle (A(t) - \langle A(t) \rangle)(A(0) - \langle A(0) \rangle) \rangle \\ \left\langle \left(A(t)^2 - \left\langle A(t)^2 \right\rangle \right)(A(0) - \langle A(0) \rangle) \right\rangle \end{bmatrix}$$

The initial conditions for the correlation functions with a stationary distribution is calculated as:

$$\begin{bmatrix} C_{A,A}(0) \\ C_{A^2,A}(0) \end{bmatrix} = \begin{bmatrix} \langle A^2 \rangle - \langle A \rangle^2 \\ \langle A^3 \rangle - \langle A^2 \rangle \langle A \rangle \end{bmatrix}\Bigg|_{SS} = \begin{bmatrix} 6.77 \\ 136.9 \end{bmatrix}$$

Ultimately the following is true of the correlation functions.

$$\frac{\partial}{\partial t} \begin{bmatrix} C_{A,A} \\ C_{A^2,A} \end{bmatrix} = J_{SS} \begin{bmatrix} C_{A,A} \\ C_{A^2,A} \end{bmatrix} = \begin{bmatrix} -0.9 & -0.1 \\ 95.1 & -7.7 \end{bmatrix} \begin{bmatrix} C_{A,A} \\ C_{A^2,A} \end{bmatrix}$$

The eigenvalues and eigenvectors can thus be determined:

$$\lambda_{S,1} = -2.891 \quad \lambda_{S,2} = -5.674 \quad \underline{\nu}_{S,1} = \begin{bmatrix} 0.050 \\ 0.999 \end{bmatrix} \quad \underline{\nu}_{S,2} = \begin{bmatrix} 0.021 \\ 0.999 \end{bmatrix}$$

The spectral decomposition of the correlation functions are thus:

$$\begin{bmatrix} C_{A,A} \\ C_{A^2,A} \end{bmatrix} = \begin{bmatrix} 0.050 & 0.021 \\ 0.999 & 0.999 \end{bmatrix} \begin{bmatrix} e^{-2.891t} & 0 \\ 0 & e^{-5.674t} \end{bmatrix} \begin{bmatrix} 34.2 & -0.716 \\ 34.2 & 1.72 \end{bmatrix} \begin{bmatrix} 6.77 \\ 136.9 \end{bmatrix}$$

The final analytical equation for the non-normalized autocorrelation function of component A is thus (Equation 4.25):

$$C_{A,A}(t) = (0.067)\, exp(-2.891 \cdot t) + (0.075)\, exp(-5.674 \cdot t)$$

5.4.3 Power Spectral Density Determinzation

After the determination of the correlation function (Equation 5.12) the determination of the power spectral density is a single step process and straightforward through Equation 5.13. Analytically the Fourier transform of every term in Equation 5.12 is a straightforward calculation.

$$P_A(\omega) = \frac{2(0.067)(-2.891)}{(-2.891)^2 + \omega^2} + \frac{2(0.075)(-5.674)}{(-5.674)^2 + \omega^2}$$

This power has units of molecules^2/Hz. The resulting 4th-order plot is shown in the Results section Figure 5.1 as green squares confirming that the ZI-Closure result matches the PSD as calculated via the SSA.

The results section will demonstrate the use of moment closure in determining power spectral densities for simple chemical reaction networks. The ZI-Closure results will be compared to the Monte Carlo SSA results to verify exact matching. Since the power spectral densities are typically used to determine dominate frequencies of oscillation, the results shown herein suggested that exact oscillation frequencies can be determined for non-linear stochastic systems through moment closure.

5.5 Results

In the following results section calculations of power spectral densities for three distinct models (Table 5.1) will be produced. First, simple reversible dimerization is intended to show the applicability of our analytical PSD equation (Equation 5.14) to a simple non-linear reaction network. Second, the Schlögl model (previously used in Chapter 3) is used to show that the accuracy of this method is not limited to simple networks, that even complex underlying distributions can be accurately reproduced using high-order ZI-Closure (in this case, 9th-order closure is needed). Finally, the Brusselator model is presented. This model is the simplest oscillatory chemical model available, and its purpose is two-fold. First, it demonstrates that the power spectral density method can be applied to multi-component systems. But more importantly, as an oscillatory system it is able to demonstrate the analytical determination of complex eigenvalues and dominant frequencies using ZI-Closure. The consequences of this theoretical development will be discussed further in the discussion section.

For the initial study for the reversible dimerization model we chose five parameter sets across three orders of magnitude in order to confirm that accurate results could be obtained across a wide parameter range. Figure 5.1 shows PSD results for these parameter sets for both Monte Carlo SSA simulation and the analytical solution as determined by ZI-Closure analysis (4th-order). As can be seen the results are accurate across the whole parameter range. Additionally, the dominant eigenvalue as determined through eigenvalue analysis can be seen as the “break” in the logarithmically plotted

Table 5.1: Model descriptions for Demonstration of Moment-Based Power Spectral Density Calculation

Model	(1) Reversible Dimerization	(2) Schlögl	(3) Brusselator
Reactions	$2A \xrightarrow{k_1} B$ $B \xrightarrow{k_2} 2A$	$2X + A \xrightarrow{k_1} 3X$ $3X \xrightarrow{k_2} 2X + A$ $B \xrightarrow{k_3} X$ $X \xrightarrow{k_4} B$	$\emptyset \xrightarrow{k_1} X$ $2X + Y \xrightarrow{k_2} 3X$ $X \xrightarrow{k_3} Y$ $X \xrightarrow{k_4} \emptyset$
Parameters	Specified in text	$k_1 \cdot A = 0.15$ (1/molec-s) $k_2 = 0.0015$ (1/molec^2-s) $k_3 \cdot B = 20$ (molec/s) k_4 specified in text	k_1 specified in text $k_2 = 1$ (1/molec^2-s) $k_3 = 1$ (1/s) $k_4 = 1$ (1/s)
Initial Conditions (molecules)	$A_0 = 10$	$X_0 = 25$	$X_0 = 10$ $Y_0 = 0$

PSD diagram. Overall, the results confirm the accuracy of analytical PSD determination using ZI-Closure.

A focus of this study was the concern that the linearization of the moment equations (Equation 5.8) might reder our PSD determination invalid for more complex (more non-linear) systems. In order to test this we looked to the Schlögl model. This one component system, through the use of third-order reaction rate laws, is able to take on a complex bimodal (double peaked) distribution. Such a distribution has been shown on numerous occasions to require high-order closure in order to faithfully reproduce the underlying bimodal shape.This will show definitively whether non-linear effects cause issue with analytical PSD determination for stochastic systems.

In Figure 5.2A PSD results across a range of k_4 values were chosen such that at high values (3.8 to 4.0) and low values (3.0 to 3.2) the distribution exhibits unimodality, whereas in the middle (3.2 to 3.8) the distribution is bimodal. Thus a range of simple and complex behaviors are sampled.

Figure 5.2B focuses in on three of these specific PSD diagrams. In particular k_4 = 3.0 (blue triangles), k_4 = 3.5 (green squares) and k_4 = 4.0 (red circles). As can be seen, accurate PSD diagrams are determined analytically for these three distinct systems, where k_4 = 3.0 exhibits a single high-valued peak, k_4 = 3.5 exhibits two peaks (bimodality), and k_4 = 4.0 exhibits a single low valued peak. Also, note that, especially

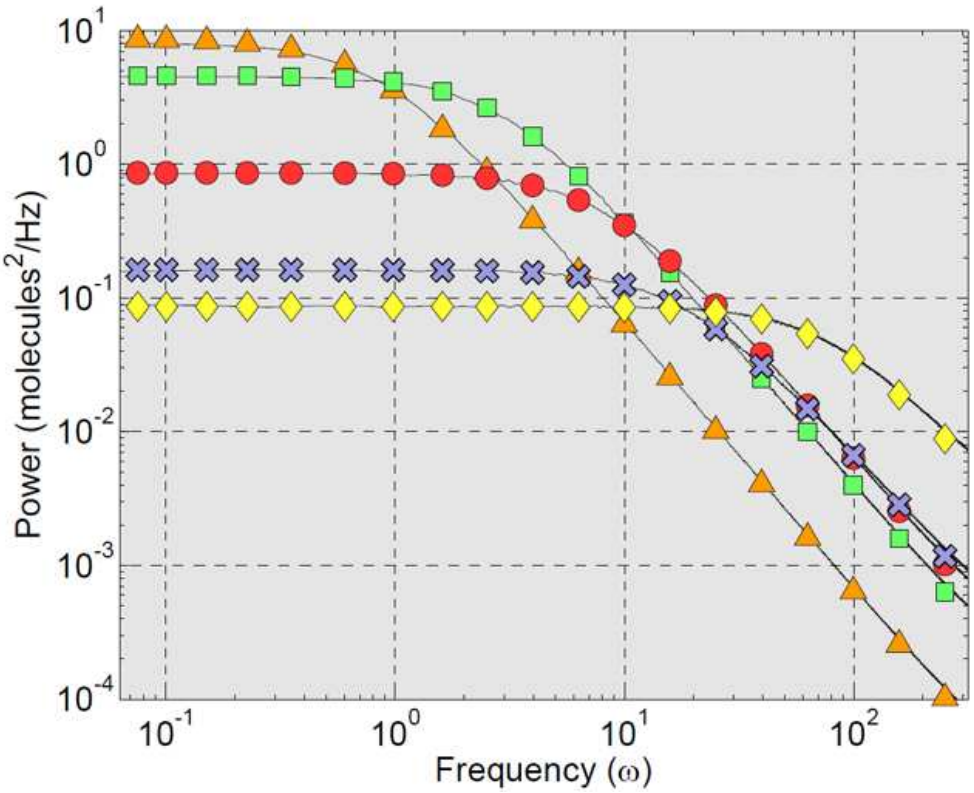

Figure 5.1: Power spectral densities for reversible dimerization model. Power spectral densities are produced using 10^6 trajectories with the SSA (shapes) and compared to the analytical solution obtained through eigenvalue analysis with ZI-Closure (4th-order) (lines). The five lines correspond to different kinetic parameter combinations k_1 (molec.$^{-1}$-s) and k_2 (1/s): [k_1, k_2] = [0.1, 0.1] (orange triangles), [0.1, 1] (green squares), [1, 1] (red circles), [10,1] (blue crosses) and [10,10] (yellow diamonds).

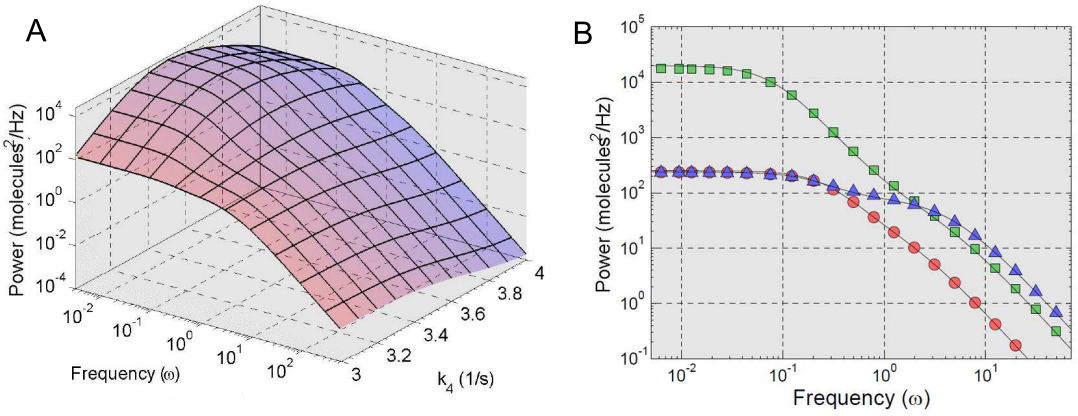

Figure 5.2: Power spectral densities for select parameters of the Schlögl model. (a) Power spectral densities are produced using 9th-order ZI-Closure for 11 different k_4 values from 3 to 4 (1/s). This range is chosen because for k_4 at high values (3.8-4.0) and low values (3.0-3.2) the Schlögl model is unimodal, while in the middle (3.2-3.8) the model exhibits a bimodal distribution (two distinct peaks). (b) A more detailed look is provided for both SSA results with 10^6 trajectories (shapes) and the analytical solution obtained through eigenvalue analysis with ZI-Closure (9th-order) (lines). The three kinetic parameters are $k_4 = 3.0$ (blue triangles), $k_4 = 3.5$ (green squares), and $k_4 = 4.0$ (red circles). Note the two distinct breaks (corresponding to two distinct timescales) exhibited by the Schlögl system, especially prevalent at $k_4 = 3.0$.

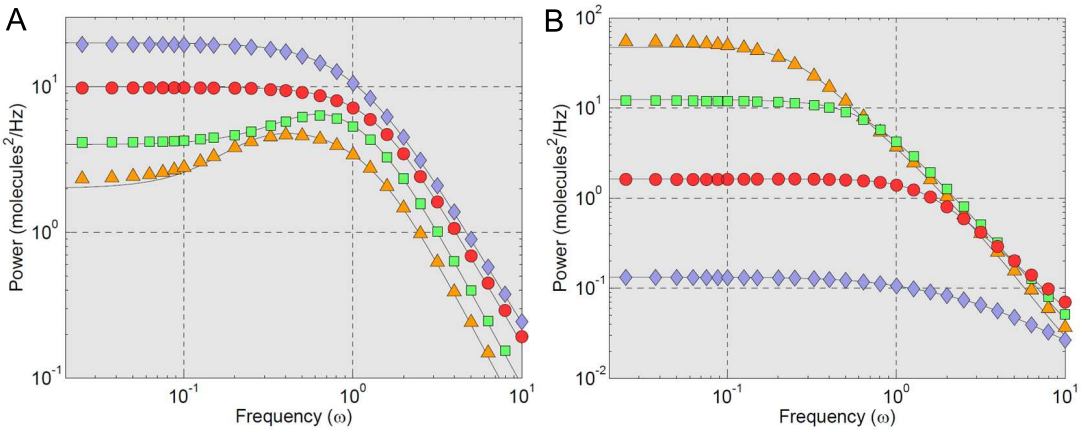

Figure 5.3: Power spectral densities for select parameters of the Brusselator model. Both the (A) X-component and (B) Y-component PSDs for four parameters values. $k_2 = 1$ (1/molec.^{2}-s), $k_3 = 1$ (1/s) and $k_4 = 1$ (1/s) for all models. For k_1 values of 1 (molec./s) (green triangles), 2 (molec./s) (green square), 5 (molec./s) (red circle) and 10 (molec./s) (blue diamond). Corresponding dominant eigenvalues are -0.1721 $\pm$ 0.1162i, -0.5664 $\pm$ 0.3725i, -1.1267 and -1.0621 for $k_1 = 1$, 2, 5, and 10 (molec./s) respectively.

for $k_4 = 3.0$, two distinct breaks are observed in the logarithmically plotted results. Indeed, we hypothesize that these represent two timescales of import for the stochastic system.

For the Brusselator model the PSD diagrams for both the X and Y component were generated and are provided in Figure 5.3. The results of particular interest are that peaks are observed in the X-component PSD diagrams for $k_1 = 1$ (molec./s) and $k_1 = 2$ (molec./s). This coincides nicely with the dominant eigenvalues of -0.1721 $\pm$ 0.1162i, -0.5664 $\pm$ 0.3725i, -1.1267 and -1.0621 (1/s) for $k_1 = 1$, 2, 5 and 10 (molec./s) respectively. Note that these peaks are conspicuously absent in the Y-Component PSD diagrams. These results are extensively discussed in the following Discussion section.

5.6 Discussion

In the previous Results sections we presented results for a simple reversible dimerization model, the complex single component Schlögl model, and the oscillatory Brusselator model. These results confirm our initial goal of extending previous efforts to produce

analytical power spectral density diagrams through eigenvalue analysis of moment equations. By extending this method to non-linear chemical reaction networks the possibility of using moment closure to determine power spectral densities and thus dominant frequencies is now a reality.

While the results for the reversible dimerization model were neither particularly challenging nor surprising several key points can be determined through the analysis. First, this is the first time an analytical power spectrum was determined for a non-linear chemical simulation. Second, as mentioned in the Results section the break in the PSD plots represent increasing dominant eigenvalues as kinetic parameters increase. This clear representation of deterministic eigenvalues in Monte Carlo stochastic simulations (the SSA results) provides further evidence that the moment dynamics have a direct and measurable influence on individual trajectory dynamics. In other words: the dynamics of individual Monte Carlo trajectories and the overall ensemble dynamics are not two distinct phenomena, but rather one in the same.

The fact that non-linear analysis, which is only valid around the steadystate at which is it calculated, provides exact solutions to the PSD for non-linear stochastic systems is an amazing and possibly unexpected result. One could easily imagine a situation where non-linear effects make exact PSD matching impossible. This, however, does not seem to be the case, even for the complex distributions of models 2 and 3.

The Schlögl model results were intended to demonstrate the universality of the method in the face of a complex distribution, in this case a bimodal distribution needing 9th-order ZI-Closure to obtain accurate results. Figure 5.2A shows how a smooth curve can be produced across a range of kinetic parameters, $k_4 = 3.0$ to 4.0 (1/s), and behaviors where the distribution is unimodal at its edges and bimodal in the middle. Figure 5.2B focuses on three of these PSD diagrams. The fact that for very non-linear systems (using third-order reaction rate laws) the PSD diagrams remain faithfully reproduced using linearization and eigenvalue analysis is of particular interest. Additionally, the ability to discern multiple eigenvalues (breaks) in the diagram, and thus definitively discern multiple timescales, is also an interesting consequence. Overall, the Schlögl results definitely show that the method can be used to produce exact power spectral densities for non-linear chemical reaction networks regardless of the complexity of the underlying distribution being studied.

The consequences of the analytical determination of the power spectral density for a set of parameters in the Brusselator can not be overstated. One can immediately recognize an important realization that the systems that exhibit peaks in the PSD also exhibit a complex dominant eigenvalue. Indeed, the analysis provided suggests that a bifurcation point can be found between $k_1 = 2$ and 5. Deterministic eigenvalue analysis of the system shows a bifurcation at $k_1 = 2\sqrt{2}$, and it appears the stochastic system acts similarly.

A more subtle point is that for the stochastic system there is a single set of eigenvalues that govern the entirety of the system, the X and Y components both have complex eigenvalues. Yet, in the PSD disgrams (both Monte Carlo SSA produced and analytically derived) only the X-component PSD exhibits a peak. This suggests that there is an oscillatory component to the underlying Y-component behavior that is not observed in the PSD. Indeed, if SSA results were only produced or analyzed for the Y-component researchers could come to the conclusion that the system doesn't oscillate at all! For this reason we believe that there is a significant place for deterministic eigenvalue analysis in stochastic simulations using moment closure techniques.

Also note that the Brusselator system represents the most complex system thus far studied via ZI-Closure. At only two components the complexity may not be immediately realized, but requiring 8 moments for accurate reproduction means the method meeded to optimize across 45 parameters. The restrictions placed on ZI-Closure are still a major issue facing its widespread use as an analytical tool. The fact that the reachable state space scales poorly and the number of equations blow up quickly with the number of components continues to be a major issue facing the method. But our initial results showing the accuracy and versatility of ZI-Closure and its use in determining both eigenvalues and power spectral densities shows great promise for the future of moment closure techniques in general for stochastic chemical simulation.

5.7 Conclusions

For biological simulations the importance of stochastic simulation has now become firmly established over the past 15 years. Most current simulations utilize Monte Carlo methods like Gillespie's SSA, but these method fall short when it comes to analytical techniques. In particular, the ability to obtain steady-states immediately and perform *a priori* analysis on biological simulations would be incredibly valuable, but impossible with Monte Carlo methods. It would thus be highly desireable to utilize moment closure schemes which can, in theory, produce such results.

In chapter 3 we introducted the zero-information closure scheme that exhibited both accuracy and versatility. Over the past two chapters we have been extending this moment closure method in order to demonstrate its potential utility in applying well-extablished analytical methods to chemical reaction simulation. Non-linear analysis, the determination of power spectral densities and control theory were of particular interest because they are already widely used in stochastic simulation of chemical systems.

It was previously established that eigenvalue analysis of linear chemical reaction networks could yield analytical solutions to power spectral densities. It was postualted that the linearization of the moment equations could produce equivalent results for non-linear systems. The results herein confirm the universal and accurate reproduction of power spectral densities for both simple and complex reaction networks.

The abilty to quickly and immediately determine the dominant frequency of oscillation (a common use of power spectral densities in noisy systems) for biologically relevant simulations is an exciting development. Optimization of frequencies, production of bifurcation diagrams and the continued exploration of stochastic eigenvalues has the potential to revolutionize how we view stochastic simulation for chemical reaction networks. Indeed, the power of ZI-Closure in particular in applications to a wide variety of networks and behaviors hints at the distinct possibility of a more wide-spread use for moment-closure based methods in the future.

Chapter 6

The Importance of Higher-Order Moments in Stochastic Model Reduction

6.1 Introduction

In Chapters 1-5 we motivated the need for moment closure, presented novel programs and algorithms for the development of our moment closure scheme ZI-Closure, and showed promising analytical results related to this work. In this concluding chapter we will tie these results back to previous work on model reduction in order to show more clearly how ZI-Closure could aid (and potentially revolutionize) biological simulation in a real and immediate way.

Models of gene regulatory and cell signaling networks can provide useful insight and quantitative understanding of dynamics for biological systems where experimental data is lacking[2, 17]. These models often find applications in systems and synthetic biology, in an effort to decipher the complexity of naturally occurring gene networks or other cellular pathways. Current research tends to focus on the use of well-studied naturally occuring components, such as the lactose[57, 19], tryptophan[58], tetracylcin[120, 121], and arabinose[122] operons. The hope is that biological simulation may aid in the design and construction of new gene networks, such as repressilators[49, 18], AND-gates[59],

and toggle switches[123, 55].

As addressed throughout this thesis biological models often rely on the use of stochastic simulation techniques such as the stochastic simulation algorithm (SSA) and its derivatives[97, 56, 73, 93]. The challenge with stochastic simulations is that the number of reactions and components necessarily becomes large with system complexity, dramatically increasing computational load. More importantly the number of unknown parameters that need to be fit to experimental data will almost always also increase.

There are two ways to address this prohibitive computational burden. First, develop novel algorithms to speed up simulation for highly complex and stiff reaction networks with accuracy. I have addressed this need already with the development of ZI-Closure allowing for the immediate determination of steady-states and quicker parameter determination. The second way is to reduce the reaction set through the elimination or replacement of subsets of reactions[125] known as model reduction. The simplest method for model reduction is through the use of complex reaction rate laws. Suites of such reduction techniques have been developed to dynamically speed up stochastic simulation[126].

The purpose of this chapter is to demonstrate the importance of higher-order statistics when utilizing complex reaction rate laws, such as Hill-type kinetics, in stochastic simulation. The Hill-type reaction rate law is a natural choice as a transcription regulation rate law because it takes into account multibinding site mechanisms, and is already naïvely used as an approximation for gene network repression in stochastic models[49]. We previously showed the paramount importance of higher-order statistics in accurately representing the dynamics of stochastic chemical networks in Chapters 3-5. The use of complex rate laws requires careful consideration of higher-order moments. This initial study is intended to show how even small miscalculations in how these higher-order moments are affected can lead to inaccurate results when placed into larger gene network simulations.

The use of the related Michaelis-Menten[108] approximation has been explored for stochastic simulations [127] with promising results, but requires complex steady-state assumptions[128]. The use of Hill-type rate laws poses an even greater difficulty as there is no mathematical derivation of this complex reaction rate[129]. Additionally, since exact mathematical analysis through the chemical master equation (CME) is

intractable[92], this particular type of reduction is performed almost exclusively on small reaction networks. Thus, ZI-Closure is a prime candidate for future model reduction studies.

The use of the Hill rate law for stochastic simulation was tested using a small gene regulatory model utilizing complex rate laws. The form and parameter values for this model were developed from a full elementary model previously fit to experimental data. We show computationally that accounting for the variance of the system by splitting the Hill-type reaction allows the incorporation of the complex reaction rate into gene network models without sacrificing accuracy or modularity.

In the results section four sample transcription models are simulated using the next-reaction Gillespie SSA. In simple systems a typical method of transcription repression is the binding of a repression factor onto one or several operator sites in the promoter region of a gene. Additionally, repressor proteins may be active in a monomer form or in a complex (e.g. an active dimer). The four models take into account the combinations of these two variables: 1) One operator with monomer binding (1M), 2) One operator with dimer binding (1D), 3) Two operators with monomer binding (2M), 4) Two operators with dimer binding (2D).

The four example configurations use basal steady-state protein output to determine suitable Hill parameters. In each configuration the repression protein is looped back to form a feedback repression loop in order to compare the accuracy of the Hill and split-Hill reaction rate laws. Finally, the importance of higher-order moments to modularity is demonstrated by linking two of these configurations into a larger network to form a bistable switch.

6.2 Background

The Hill-type rate law is a simple non-linear function with versatile dynamics and is an important building block of transcription regulation models. The full models include all of the important biomolecular interactions, including transcription, translation, degradation and repression, and are comprised of reactions with elementary kinetics. The

reduced models are generally identical to the full models except that transcription repression is described by the single, more complex, Hill-type reaction rate law:

$$a_{Hill} = \frac{a_{max} S^n}{S^n + K_m^n} \quad \text{or} \quad a_{Hill} = \frac{a_{max} K_m^n}{S^n + K_m^n} \tag{6.1}$$

The first equation is a positive form (activation), the second is a negative form (repression). Here a_{max} is the maximum propensity for reaction, S is the amount of substrate (in molecules), K_m is the Michaelis constant describing the number of molecules of S necessary to reduce the propensity to $a_{max}/2$, and n describes the shape of the function. The Hill model was first proposed in the early 1900s as a description of cooperative and multi-site binding in hemoglobin[130]. Many transcription regulation mechanisms involve multiple binding sites in the promoter region. The Hill rate law could potentially replace many of these complex regulation mechanisms in their entirety.

In the Results section for Hill parameter matching the Hill coefficients are found by a simple linear regression. It is first assumed that the steady state reporter protein concentration (P) is proportional to the production rate of the Hill function representing transcription repression (a_{Hill}). When no repressor (S in Equation 6.1) is present in the system the resulting output is called the basal protein concentration and basal production rate (P_b and $a_{Hill,b}$ respectively). By dividing the protein output and production rate by their basal counterparts one obtains the following equation:

$$\frac{a_{Hill}}{a_{Hill,b}} = \frac{P}{P_b} = P^* = \frac{K_m^n}{K_m^n + S} \tag{6.2}$$

Through simple mathematical manipulation the following equation is determined.

$$\log \frac{1 - P^*}{P^*} = n \cdot \log S - n \cdot \log K_m \tag{6.3}$$

Note that P^* ranges from 0 to 1 in repression. A linear regression for log-distributed repression concentrations ($\log S$) will thus produce a slope n and an intercept of $-n \cdot \log K_m$. In this way the parameters n and K_m are determined from the reporter protein concentration data in a simple gene repression network.

6.2.1 Full Models: Gene Network and Kinetics

In the analysis of the simple models that follow refer to Table 6.1 for a list of reactions and kinetic constants. The 1M-model utilizes Reactions 1-12, the 1D-model utilizes Reactions 1-12 and 19-21, the 2M-model utilizes Reactions 1-18, and the 2D-model utilizes

Table 6.1: List of reactions utilized in full and reduced gene network models.

#	Reactions	k	#	Reactions	k
1	DNA + RNAp $\rightarrow$ DNA:RNAp	$10^7 \frac{1}{M \cdot s}$	13	DNA + $X_a \rightarrow DNA_{x2}$	$10^8 \frac{1}{M \cdot s}$
2	DNA:RNAp $\rightarrow$ DNA + RNAp	$0.1 \frac{1}{s}$	14	$DNA_{x2} \rightarrow$ DNA + X_a	$0.01 \frac{1}{s}$
3	DNA:RNAp $\rightarrow$ DNA + RNAp*	$0.01 \frac{1}{s}$	15	DNA_{x1} + $X_a \rightarrow DNA_{x12}$	$10^8 \frac{1}{M \cdot s}$
4	RNAp* $\rightarrow$ RNAp + mRNA	$\Gamma_{(30,600)}$	16	$DNA_{x12} \rightarrow DNA_{x1}$ + X_a	$0.01 \frac{1}{s}$
5	mRNA + rib $\rightarrow$ mRNA:rib	$10^5 \frac{1}{M \cdot s}$	17	DNA_{x2} + $X_a \rightarrow DNA_{x12}$	$10^8 \frac{1}{M \cdot s}$
6	mRNA:rib $\rightarrow$ mRNA + rib	$0 \frac{1}{s}$	18	$DNA_{x12} \rightarrow DNA_{x2}$ + X_a	$0.01 \frac{1}{s}$
7	mRNA:rib $\rightarrow$ mRNA + rib*	$33 \frac{1}{s}$	19	$2 \cdot X \rightarrow X_a$	$10^8 \frac{1}{M \cdot s}$
8	rib* $\rightarrow$ rib + Product	$\Gamma_{(33,200)}$	20	$X_a \rightarrow 2 \cdot X$	$1 \frac{1}{s}$
9	mRNA $\rightarrow$	$0.002 \frac{1}{s}$	21	$X_a \rightarrow X$	$0.00116 \frac{1}{s}$
10	Product $\rightarrow$	$0.000578 \frac{1}{s}$	22	DNA + RNAp $\rightarrow$ DNA + RNAp*	k_{Hill}
11	DNA + $X_a \rightarrow DNA_{x1}$	$10^8 \frac{1}{M \cdot s}$	23	DNA $\rightarrow DNA_x$	k^-
12	$DNA_{x1} \rightarrow$ DNA + X_a	$0.01 \frac{1}{s}$	24	$DNA_x \rightarrow$ DNA	k^+

Abbreviations: RNAp = RNA polymerase, RNAp* = DNA bound RNA polymerase, mRNA = messenger RNA, rib = ribosomes, rib* = RNA bound ribosomes, X_a = active repressor. $\Gamma_{(rate,step)}$ reaction events are described by a rate $\left(\frac{1}{s}\right)$ and a number of steps

Reactions 1-21. The reaction rates presented here are used in producing all elementary model results. The network configuration and kinetic parameters used throughout the study are designed for stochastic simulation and matched to experimental data. The network represents the type of model that would be replaced by a reduced model with complex reaction rates. The network provides a test-of-concept in applying complex reaction rates to what can be considered an experimentally fitted, modular, stochastic model.

In the base model only transcription, translation, and degradation reactions are used to build a straightforward model of protein synthesis. Reactions 1-4 represent transcription. Reactions 1 and 2 are RNA polymerase binding and unbinding events, respectively. Reaction 3 is initiation of transcription, and Reaction 4 is elongation and termination. It should be noted that elongation events are represented as Γ distributed events. The Γ-distribution is equivalent to a string of identical first order reactions (see Table 6.1 footnote), and models a protein or ribosome stepping along DNA or messenger RNA (mRNA)[56]. Translation takes on a similar form, Reactions 5 and 6 involve the binding and unbinding of ribosomes on a length of mRNA. Reaction 7 and Reaction 8 are the initiation and elongation/termination of translation. Two degradation reactions for mRNA (Reaction 9) and Products (Reaction 10) are also included in the base model.

The values for k_1-k_3 are approximated from equilibrium constants for the tetracycline

and lactose operons[131, 132, 133, 134]. The values for k_4 and k_8 (and the number of steps for each) are from elongation approximations[135, 136]. The values for k_5-k_7 and k_9 are designed to produce approximately 20 production events during the lifetime of a single mRNA molecule. The variable k_{10} specifies a 20 minute half-life[121]. Reaction 10 can refer to an output protein or a repressor protein. For multiple products Reaction 10 represents multiple reactions.

To form the four sample models, additional reactions are then added to the base model. To take into account a single operator site Reactions 11-12 are added. To add a second operator site Reactions 13-18 are added. Finally to add dimerization Reactions 19-21 are added. For the sample examples k_{11}, k_{13}, k_{15}, and k_{17} are on the same order as RNA polymerase binding. Then, k_{12}, k_{14}, k_{16}, k_{18} give relatively tight binding affinities. The constants k_{19} and k_{20} are adapted from data for tetracycline repressors (dimer active)[137]. Finally, k_{21} is simply two times the degradation rate of a protein monomer (X).

While the use of these specific kinetic constants limits the generality of the conclusions presented herein, the gene network described in Table 6.1 exhibits sufficiently interesting behavior to more generally suggest that when transcription repression diverges from a simple exponentially distributed reaction, matching higher-order statistics is necessary to guarantee accuracy.

6.2.2 Reduced Models: Hill-type Reaction Rate Law

The purpose of a reduced model is two-fold: it should reduce the number of reactions and/or simplify the reaction network, and it should maintain accuracy. The addition of repressor protein binding to operator sites (Reactions 11-21) creates a multisite binding system in which the more versatile Hill-type kinetics (Equation 6.1) can be applied. Once reduced, all four models (1M, 1D, 2M, 2D) have the same number of reactions, representing a severe reduction in the number of reactions. In all cases the transcription repression mechanism is replaced in its entirety by the Hill rate law. This is a common technique in both deterministic and stochastic simulations already[49], although naïvely applied in the case of stochastic simulation.

The fundamental difference between the elementary and reduced models is the replacement of Reactions 1-3 and Reactions 11-21 by Reaction 22. Here the kinetic rate,

k_{Hill}, takes the form:

$$k_{Hill} = \frac{k_{base} K_m^n}{X_a^n + K_m^n} \tag{6.4}$$

Thus three parameters: k_{base} (units of $\frac{1}{M \cdot s}$ because the reaction is 2_{nd} order), K_m (units of M), and n (unitless) are fit to the mean stochastic simulation results for the full elementary models. The remainder of each model is simply Reactions 4-10 condensing the repression function into strength (K_m) and shape (n) parameters.

6.2.3 Split Hill Model Theory

As will be shown the Hill-type kinetics are not readily applicable to stochastic simulations. Part of the problem is the inability for the Hill-type kinetics to capture the higher-order statistical properties found in the elementary model. The proposed solution is to split the Hill-kinetic model into multiple reactions. The additional degree of freedom will match higher-order statistical properties and allow the model to reclaim some of the loss of accuracy.

Reactions 23 and 24 are added to the Hill model and they function as availability reactions for the DNA molecule described in Reaction 22. To implement these reactions a form for the rate laws (k^- and k^+) need to be determined. Reactions 23 and 24 are both first order and thus the mathematics describing DNA availability can be determined exactly. By setting k_{Hill} to k_{base} the steady-state mean availability of the DNA is determined by:

$$\frac{\partial \langle DNA \rangle}{\partial t} = 0 = -k^- \langle DNA \rangle + k^+ \langle DNA_x \rangle \tag{6.5}$$

$\langle DNA_x \rangle$ is unavailable DNA and $\langle DNA_x \rangle = DNA_{Total} - \langle DNA \rangle$. The following equation is obtained.

$$\frac{\langle DNA \rangle_{ss}}{DNA_{Total}} = \frac{k^+}{k^+ + k^-} = \frac{1}{1 + \left(\frac{k^-}{k^+}\right)} \tag{6.6}$$

Note that what we have found is the percentage of DNA available at a given time.

To get the appropriate form the availability is matched to the Hill function. The following equation needs to be satisfied:

$$k_{base} \langle DNA \rangle_{ss} = k_{Hill} DNA_{Total} \tag{6.7}$$

Thus:

$$\frac{k_{base}}{1+\left(\frac{k^-}{k^+}\right)} = \frac{k_{base}}{1+\left(\frac{X_a}{K_m}\right)^n} \tag{6.8}$$

This provides an additional degree of freedom (k^+, a first-order reaction rate) and a form for k^- (also first-order):

$$k^- = k^+\left(\frac{X_a}{K_m}\right)^n \tag{6.9}$$

This derivation relies on RNA polymerase concentration remaining relatively constant throughout the simulation.

The same k_{base}, K_m, and n fitted for the reduced model are used. Other forms for Equation 6.9 are possible. For example, if k^- was the positive form of Equation 6.1 and k^+ was the negative form of Equation 6.1, then Equation 6.9 would remain valid. The chosen construction was used because if $n = 1$, splitting the Hill rate law should result in a first order reaction for k^- and zeroth order reaction for k^+. This condition is desirable when reducing Michaelis-Menten like systems, and holds for Equation 6.9. In this case the Split-Hill model is still conceptually simple and eliminates 2-11 reactions depending on the form of repression.

The non-Poisson behavior exhibited by adding the availability reactions into the Hill network is due to a state of latency as DNA enters the off-state. In theory, as the k^+ value goes to infinity the network will immediately reach steady state and the Poisson behavior is restored. Lower k^+ values correspond to a higher period of latency and non-Poisson behavior.

It should be noted that additional splits could be applied to capture more information about the higher-order statistics. Such additional splits would increase accuracy in theory, especially for systems with very strong stochastic effects, but also complexity. Eventually, the number of parameters to be matched and size of the model will be equivalent to the elementary model. The give-and-take between complexity and accuracy would be interesting to observe in larger or more complex networks than were chosen for this study.

6.2.4 Bistable Switch Model

The final test of the use of complex reaction rates laws in stochastic gene network simulation is the effect higher-order moments have on modularity. Modularity represents the universality of a model. A large complex model comprised of multiple smaller models should retain accuracy without altering the model parameters determined for the small models when isolated. A simple system for testing modularity is a dual-repressor bistable switch (Figure 6.4, inset). Two 2-operator dimer repressed genes (A and B) with kinetic constants identical to the 2D model are used. Gene A outputs a repressor monomer, X_B, that complexes and represses the function of gene B. Gene B acts similarly, producing a repressor monomer, X_A, the represses the function of gene A.

In bistable switches there is often a signal that causes the network to switch states. In the sample model this is not the case. Instead the network is either dominated by gene A or gene B from the start, and the steady-state behavoir is thus of little interest. Modularity is instead compared by observing the dynamic results early in the simulation.

6.3 Results

In the reduced models it is necessary to fit the three parameters (k_{base}, K_m, and n) to data procured from full stochastic simulations. Only the steady state product output, for a range of repressor levels, is used to obtain the fit. This method is used because, while we are fitting to data from stochastic simulations, it would be desirable to fit the parameters to experimentally realizable data. Typically, the data available experimentally is the level of a ouput fluorescence protein. The parameters are fit using a simple linear regression based on Equation 6.3. In these fits the maximum propensity is approximately equal to 0.00975 $\frac{1}{s}$ for all models. By dividing by the total RNA polymerase in the elementary model a k_{base} of 21800 $\frac{1}{M \cdot s}$ is found. The total number of RNAp molecules on average was chosen as 270 molecules, enough to saturate the system. The fits are provided in Figure 6.1.

In Figure 6.1, simulation results are shown from all four elementary models for various levels of repressor (X_a for monomer repression, X for dimer repression). By eliminating generation and degradation reactions the level of repressor is constant and

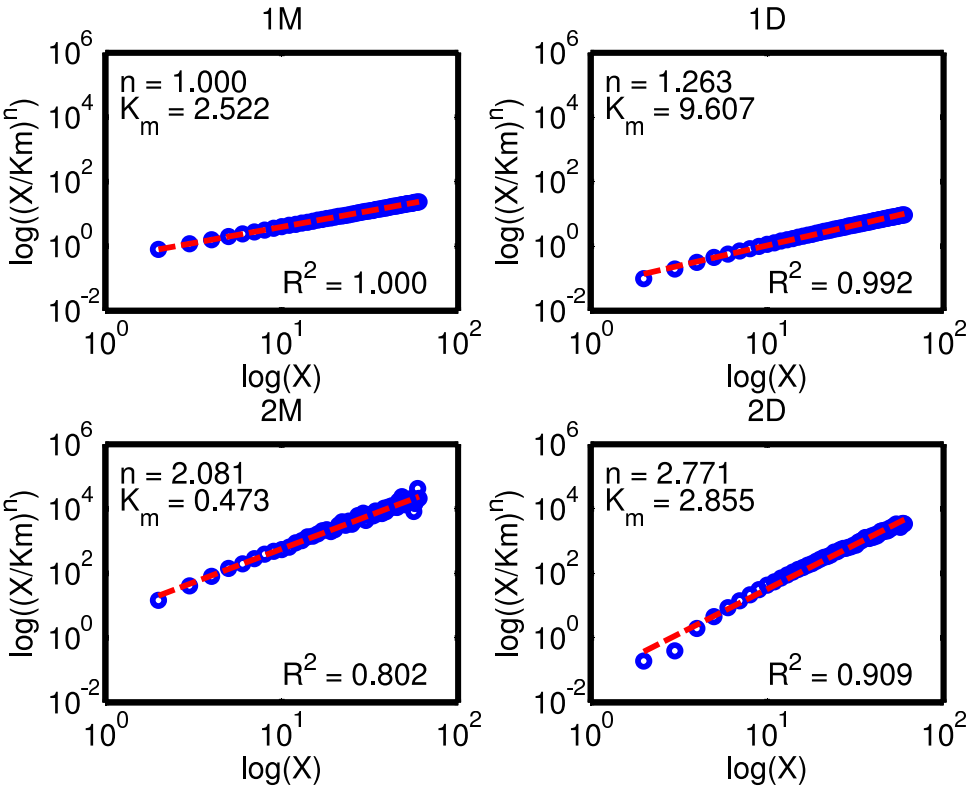

Figure 6.1: The Hill parameter fits for models 1M, 1D, 2M, and 2D. The parameters K_M and n are given in the plot along with R^2 values for quantitative comparison. The steady-state of the product from elementary stochastic simulations (blue dots), and the linear fit from Equation 6.3 (red dashed line). In the figure, X refers to X_a (monomer).

a steady-state product concentration is found. This is approximately equivalent to determining a steady-state output when a known concentration of repressor protein is provided to a sample. The final Hill parameters are provided in the figure insets along with the R^2 values for the fits.

Once the Hill parameters are determined the reduced models are tested in feedback loop repression. Feedback loops are simulated using both elementary and reduced models. The mean product outputs are compared by relative root mean squared differences (rRMS). The variance-to-mean data (VMR) for the mRNA output is also provided for analysis of the higher-order moments. In stochastic systems reactions are assumed to be Poisson distributed events, so the time between reaction events is exponentially distributed. Poisson processes have the distinct feature of a variance-to-mean ratio of one. The VMR data gives a rough metric for determining when repression diverges from a simple Poisson process, and thus when higher-order statistics become significant.

The mean product output for all four models are shown in Figure 6.2. In all four cases the transient behavior, specifically the overshoot of the steady-state early in the simulation, differs significantly. Quantitative comparisons are reported in Figure 6.2. For the 1M, 1D, 2M, 2D models the steady-state VMR for mRNA production for the full models are 0.983, 0.989, 1.247, 1.321. Thus the 2-operator models diverge significantly from simple Poisson processes. For the reduced models the VMR are 0.918, 0.902, 0.992, 0.950, respectively. The reduced models tend to lie close to a VMR of one, and fail to capture the more complex behavior in the 2-operator system. By reducing mRNA production to a single reaction an implicit assumption was made, that repression is an approximately Poisson distributed event. This naïve application of Hill kinetics fails to accurately describe the behavior of the system.

In the case of the Split-Hill model the same four sample models (1M, 1D, 2M, 2D) were simulated, and the product means were compared. In Figure 6.3 the mean output is plotted for both the elementary and Split-Hill models. The constant k^+ is adjusted to create the best quantitative fit to the elementary data. For the 1M model $k^+ = 0.0086s^{-1}$, for the 1D model $k^+ = 0.0075s^{-1}$, for the 2M model $k^+ = 0.000185s^{-1}$, and for the 2D model $k^+ = 0.00028s^{-1}$. Figure 6.3 also provides a quantitative comparison of the fits (as described by the rRMS). Across all models the fitted Split-Hill model provides a more accurate representation of the full-model results.

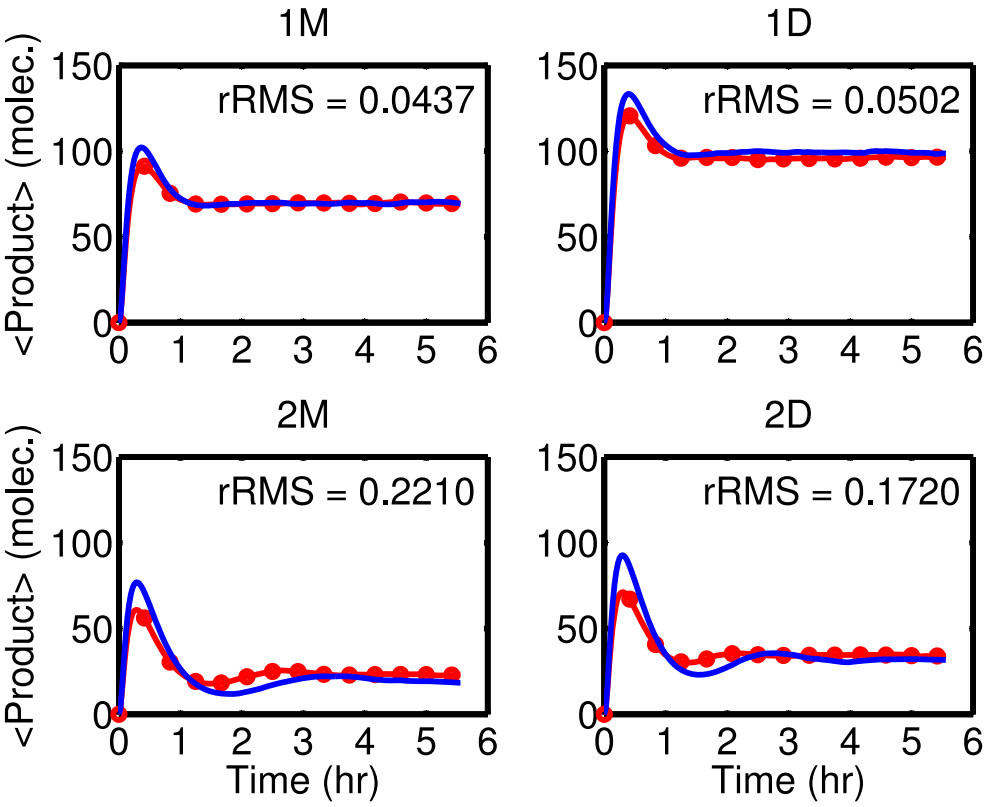

Figure 6.2: Models 1M, 1D, 2M, and 2D, mean Product count versus time. Elementary (blue solid line) and reduced (red circle line) models. Each line represents an average over 10,000 trajectories. Elementary simulations used Hy3S, Reduced simulation used a basic Gillespie next-reaction SSA. Relative root mean squared differences (rRMS) values are provided for comparison.

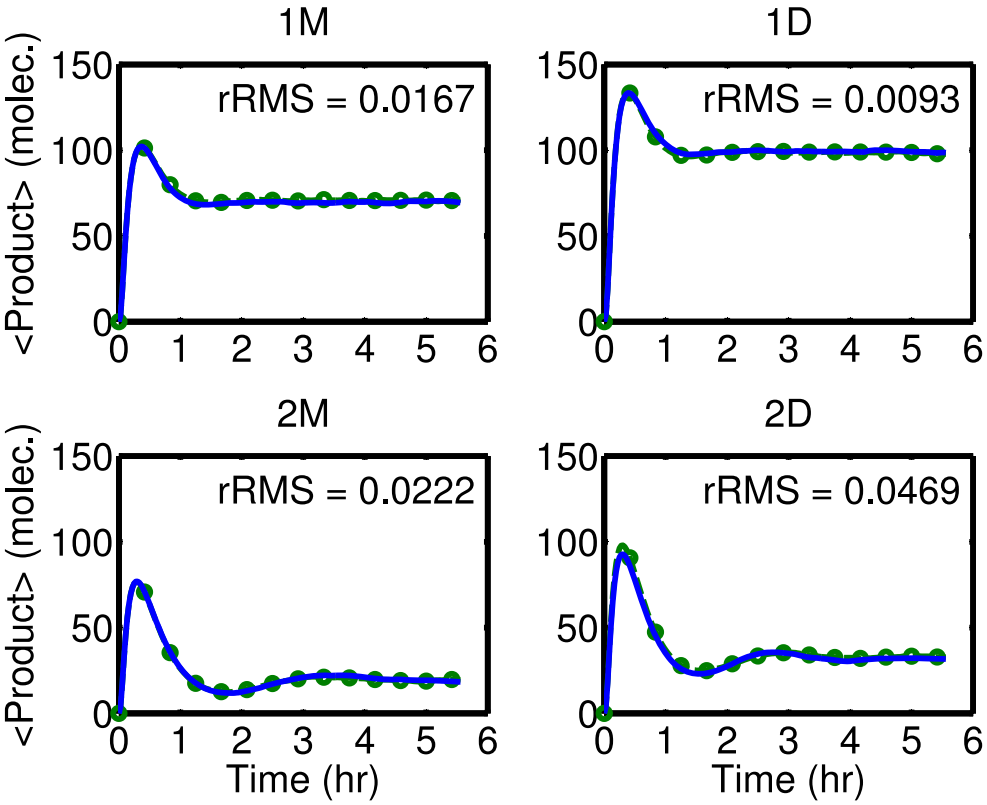

Figure 6.3: Models 1M, 1D, 2M, and 2D, mean Product count versus time. Elementary (blue solid line) and Split-Hill (green circle line) models. For the Split-Hill model $k^+ = 0.0086\ s^{-1}$ (1M), $0.0075\ s^{-1}$ (1D), $0.000185\ s^{-1}$ (2M), and $0.00028\ s^{-1}$ (2D). Each line represents an average over 10,000 trajectories. Relative root mean squared differences (rRMS) values are provided for comparison.

For the 1M, 1D, 2M, 2D models the steady state VMR for mRNA production for the split-Hill models are 0.997, 1.031, 1.380, 1.496, respectively (compared to 0.983, 0.989, 1.247, 1.321, from before). In the case of one-operator models (1M and 1D) the VMR of the Split-Hill model match well with the elementary case. In these cases the higher-order effects are primarily due to the DNA-promoter region going into a period a latency when bound by a repressor. For the one-operator models this effect is restored. In the cases of the two-operator repressors the VMR is still significantly different, but now the reduced model shows a VMR greater than 1, and thus is able to capture non-Poisson behavior.

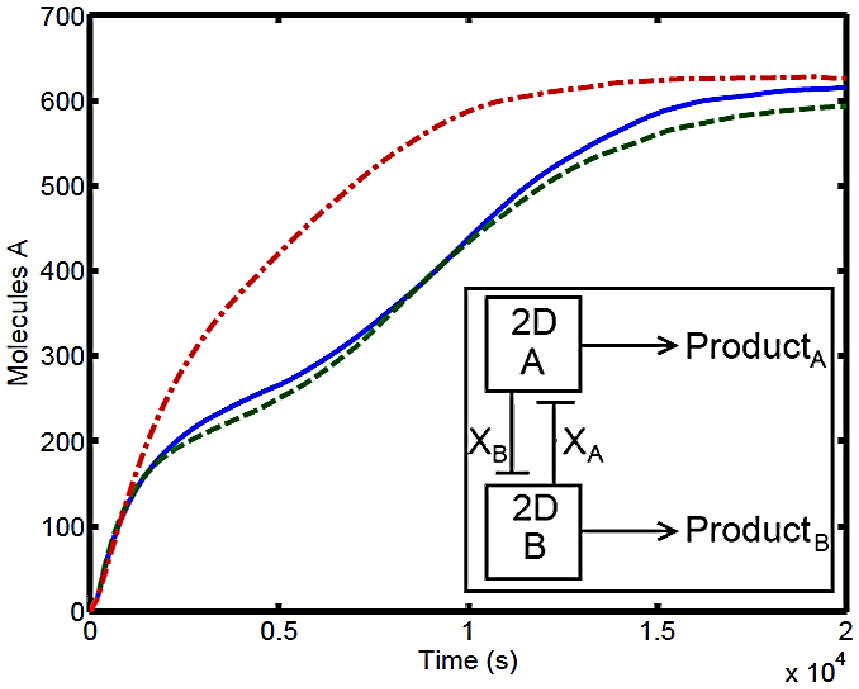

Figure 6.4: Elementary (blue solid line), Hill (red dash-dot line), and Split-Hill (green dashed line) models, mean Product count versus time for gene A. The inset figure describes the reaction network. Each line represents an average over 10,000 trajectories.

Finally, the bistable switch model was simulated for the full model, the reduced Hill model, and the reduced Split-Hill model. The three resulting trajectories for gene A through time are shown in Figure 6.4. Quantitatively, the rRMS for the Hill and Split-Hill models are 0.0363 and 0.2425 respectively. The Split-Hill model tends to conserve modularity more effectively than the Hill-model, presumably due to the more accurate matching of higher-order statistics.

6.4 Discussion

The results provided within this attempt to quantify the affect and importance of higher-order moments in stochastic model reduction using complex reaction rate laws. Hill parameters were fit to mean output protein concentrations and four sample models were created. Simple feedback loop networks were compared to test accuracy and a bistable

switch network was used to test modularity. Relative root mean squred differences analysis was used to quantify the accuracy of the simulation compared to the full elementary simulation results.

The Hill-type rate law was able to match the mean output protein results for the four systems with reasonable accuracy (see Figure 6.1). The single complex rate law was thus used to replace a number of elementary reactions. The four models were then reconfigured to form simple feedback loops. The mean output of the Hill models were compared to elementary reaction network results to assess accuracy of the Hill-models (see Figure 6.2). As was demonstrated the one-operator systems exhibited reasonable accuracy, but the two-operator models diverged significantly from the full-model results.

The variance-to-mean (VMR) ratios indicate that the two-operator systems diverge from simple Poisson processes (VMR=1) exhibited by the Hill models. The failure to match output dynamics using the Hill type kinetic rate law is due to the inability to accurately account for complex higher-order statistics observed in gene transcription.

The split-Hill model has an additional degree of freedom and was used to test the direct effect of higher-order statistics on stochastic model reduction. This additional kinetic constant (k^+) can be fit such that accuracy in the feedback loop is much improved (see Figure 6.3). The VMR, while not an exact match, indicate that the split-Hill model is able to exhibit the highly non-Poissonian behavior that is observed in these more complex gene repression schemes. Note that in Figure 6.2 the models with high rRMS values corresponded to lower k^+ values in Figure 6.3. This intuitively makes sense because the poor fits were likely due to an inability to match higher-order statistics corresponding to a need for a higher latency period in the availability reactions.

The final observation was concerning modularity. The ability for a smaller kinetic model to be plugged into a large network without modification to its parameters is a trait that is highly desirable in gene models. The results from the bistable switch (Figure 6.4) suggest that the effects of higher-order statistics can result in disproportionately inaccurate mean output results when present in a larger gene network. Indeed, the slight improvement in dynamic matching in Figures 6.2 and 6.3 results in a dramatic improvement in accuracy in the larger bistable switch. The ability of ZI-Closure to construct accurate steady-state profiles of small reaction networks could be used to aid in the determination of more accurate complex rate laws for use in biological modeling.

6.5 Conclusion

Model reduction can offer simpler and more intuitively constructed gene network models. In order to maintain accuracy and modularity, however, these models must account for higher-order statistics as well. Working with simple models of small gene networks (1-2 genes) and borrowing kinetic parameters from previous studies we were able to show that higher-order statistics have a dramatic effect on the accuracy of gene network simulations. The inability of naïvely applied Hill-type kinetics to account for the variance in the transcription rate suggest that even for simple cases there is room for improvement in the use of complex rates laws for model reduction. Indeed, by combining several complex rate laws together in the split-Hill model, accuracy and modularity were restored in systems exhibiting non-Poisson behavior.

The construction of the split-Hill models were, at the time, performed using SSA results and a trial-and-error approach. We believe that, given the desire for modularity on the single gene scale, the development of ZI-Closure offers an interesting avenue in the development of novel model reduction techniques. ZI-Closure operates easily on small reaction sets, and single gene constructs cannot practically exhibit complex underlying behaviors like bimodality (in the Schlögl model). These simplifications make the immediate implementation of moment closure possible. While Hill-type and Michaelis-Menten kinetics have historical value in biological simulation, there is room for more abstract complex rate laws in modeling that can account for higher-order statistics for stochastic simulation. Such developments could be the key to expanding synthetic biological modeling beyond small gene networks.

Chapter 7

Concluding Remarks

Biology is at the forefront of scientific exploration as techniques to explore biological systems, both experimentally and computationally, grow at an unprecedented pace. Research into the creation of novel functionality in existing biologically systems, in particular, has become an area of intense focus as the cost of DNA synthesis drops precipitously. As an engineering discipline research will, eventually, rely on computation and simulation to save both time and money in the development of complex synthetic cellular machinery.

Models of biological systems are often considered to be a large set of chemical reactions within a homogenous volume. Such models are traditionally modeled using reaction rate equations as sets of ordinary differential equations. Through the development of sophisticated computer algorithms and the adoption of supercomputing technology these models can be of nearly unlimited size and scope. Indeed, full cellular models of prokaryotic cells have been utilized in systems biology for over a decade now.

These deterministic models, however, are in many cases inadequate for the simulation of biological systems. Determinism in thermodynamic systems is only truly applicable at the thermodynamic limit, when the number of interacting particles is large and the magnitude of fluctuations becomes insignificant. This limit is not, however, applicable in many biological systems where chemical components such as DNA or mRNA can exists in single digit molecule counts. Thermal noise can have a dramatic effect on the dynamics of such systems resulting in an ensemble of possible trajectories through time. Such an ensemble is decidedly non-deterministic and best described as

an evolving probability distribution modeled by stochastic mathematics.

Through the use of Gillespie's stochastic simulation algorithm (SSA) and subsequent improvements the stochastic simulation of chemical and biological systems has experiences an explosion of research endeavors over the past 15 years. Despite these many important improvements stochastic biological simulation still suffers from several drawbacks. First, the inability of the SSA to produce steady-state distributions immediately and perform analysis easily results in major computational difficulties even for simple chemical reaction simulation. Second, the development of novel model reduction techniques remains mathematically difficult due to the intractability of the underlying mathematics. The work presented herein on the development of a novel moment closure scheme, zero-information closure (ZI-Closure), and analytical methods shows promise in addressing both of these drawbacks.

In Chapter 2 of this thesis a program was produced in Matlab and made available open-source that can produce moment equations as an augmented under-defined matrix. This program was invaluable to the production of this thesis, but more importantly will be valuable in future work in the development of new analytical closure techniques, the analysis of existing closure techniques and any other analytical techniques developed for stochastic chemical simulation. We were able to completely eliminate any bottleneck in producing moment equations (both numerically and symbolically) in Matlab. By casting the equations as a matrix, as well, it accelerated the use of moment closure to obtain steady-state distributions by enabling the determination of null spaces for use with ZI-Closure.

In Chapter 3 our novel moment closure scheme, zero-information closure, was presented. This closure scheme postulates that there exists a moment-order M above which no additional information can be gained about the underlying probability distribution being determined. This assumption is intuitively pleasing as it is ultimately what all closure schemes claim: that all information is contained in a finite number of lower-order moments. Additionally, by utilizing a numerical approach to the problem the method can, in theory, be applied to any chemical reaction network utilizing elementary rate laws with arbitrary accuracy. We were able to show accuracy and versatility in determining both dynamic and steady-state results, all the way up to 12th-order closure for the Schlögl model, by far the highest closure order attempted for chemical reaction

systems at present.

These first two chapters completed the zero-information closure method by allowing for the determination of moment equations and using ZI-Closure to solve it. The rest of the thesis focused on analytical methods: non-linear analysis, power spectral analysis, and model reduction.

In Chapter 4 we extended this work to explore the possibility of performing non-linear analysis on stochastic systems using the methods of Lyapunov. By determining the steady-state distribution in Chapter 3 this opened the door to linearizing the moment equation system and determining eigenvalues and eigenvectors of the stochastic system. The possibility of using well-established deterministic analytical methods, like the *a priori* analysis of stochastic system dynamics with eigenvalue analysis, has the potential to completly change how stochastic simulation of chemical systems is performed. We were able to show, again, the versatility of this method along with the accuracy of the eigenvalues in describing stochastic dynamics. Intriguingly, the Schlögl model, which has three steady states in the bimodal regime when deterministically modeled, has only the single ensemble steady state in the stochastic simulation. In the future we hope to further explore the possibility of things like limit cycles and other behaviors found in deterministic systems, but never firmly established in stochastic theory.

In Chapter 5 we further extended our analytical toolbox by also including a method for analytically determining power spectral densities for chemical reaction systems using the eigenvalue analysis demonstrated in Chapter 4. This work was inspired by a 2006 paper by Warren[117] where analytically derived power spectra are determined for linear systems. Here we extended this work to include all linear and non-linear chemical systems by using ZI-Closure. Power spectral analysis is often used in determining dominant oscillatory frequency for stochastic chemical simulation, something that is often difficult to calculate due to noise. Analytically producing these values opens the doors to potentially optimizing the oscillatory frequency of biological systems efficiently, something that is, currently, impractical.

Finally in Chapter 6 we showed some earlier work on model reduction focusing on the importance of higher-order moments concerning the use of complex reaction rate laws. We were able to show that the naïve use Hill-type kinetics in the modeling of transcription can often have a dramatic effect on the accuracy of the simulation when

placed into a feedback look. Through the development of a novel complex rate law we called the split-Hill model we were able to show that by restoring accuracy in the higher-order moments (the variance in particular) that the accuracy in the feedback simulation was restored. Further, the slight inaccuracies present in the feedback simulation becomes catastrophic when placed into more complex systems effectively destroying modularity. The fact that complex rate laws are often used in place of small reaction networks and necessitate accuracy in higher-order kinetics suggests that ZI-Closure could aid in the development and refinement of complex non-linear rate laws for use in biological simulation.

Overall, we have been able to establish a closure method for chemical dynamics that is accurate, versatile and theoretically well-founded. We have also been able to show that using moment closure may allow for the use of deterministic analytical methods in stochastic simulation, a tantalizing possibility when considering the complexity of stochastic mathematics. We hope that in the future these analytical methods will be expanded to include model reduction, and look forward to the improvements to ZI-Closure that may allow closure methods to assist and even supplant Monte Carlo simulation techniques in the future.

References

[1] DT Gillespie. Stochastic simulation of chemical kinetics. Annual review of physical chemistry **58** 35-55 (2007).

[2] H De Jong. Modeling and simulation of genetic regulatory systems: a literature review. Journal of computational biology **9.1** 67-103 (2002).

[3] F Jacob and J Monod. Genetic regulatory mechanisms in the synthesis of proteins. Journal of molecular biology **3.3** 318-356 (1961).

[4] BC Goodwin. *Temporal organization in cells: dynamic theory of cellular control processes*. Academic Press (London and New York) 1963.

[5] YN Kaznessis. Multi-scale models for gene network engineering. Chemical engineering science **61.3** 940-953 (2006).

[6] YN Kaznessis. *Statistical Thermodynamics and Stochastic Kinetics: An Introduction for Engineers*. Cambridge University Press, 2012.

[7] MS Bartlett. Deterministic and stochastic models for recurrent epidemics. Proceedings of the third Berkeley symposium on mathematical statistics and probability. **4.81** 109 (1956).

[8] R Durrett and SA Levin. Stochastic spatial models: a user's guide to ecological applications. Philosophical Transactions of the Royal Society of London (Series B: Biological Sciences) **343.1305** 329-350 (1994).

[9] B Bolker and SW Pacala. Using moment equations to understand stochastically driven spatial pattern formation in ecological systems. Theoretical population biology **52.3** 179-197 (1997).

[10] PJ Halling. Do the laws of chemistry apply to living cells?. Trends in biochemical sciences **14.8** 317-318 (1989).

[11] MSH Ko. Problems and paradigms: Induction mechanism of a single gene molecule: Stochastic or deterministic?. Bioessays **14.5** 341-346 (1992).

[12] HH McAdams and A Arkin. Stochastic mechanisms in gene expression. Proceedings of the National Academy of Sciences **94.3** 814-819 (1997).

[13] HH McAdams and A Arkin. It's a noisy business! Genetic regulation at the nanomolar scale. Trends in genetics **15.2** 65-69 (1999).

[14] CD Cox, GD Peterson, MS Allen, JM Lancaster, JM McCollum, D Austin, L Yan, GS Sayler and ML Simpson. Analysis of noise in quorum sensing. OMICS A Journal of Integrative Biology **7.3** 317-334 (2003).

[15] S Krishna, B Banerjee, TV Ramakrishnan and GV Shivashankar. Stochastic simulations of the origins and implications of long-tailed distributions in gene expression. Proceedings of the National Academy of Sciences **102.13** 4771-4776 (2005).

[16] JM Pedraza and A van Oudenaarden. Noise propagation in gene networks. Science **307.5717** 1965-1969 (2005).

[17] H Salis and YN Kaznessis. Numerical simulation of stochastic gene circuits. Computers and chemical engineering **29.3** 577-588 (2005).

[18] LM Tuttle, H Salis, JR Tomshine and YN Kaznessis. Model-driven designs of an oscillating gene network. Biophysical journal **89.6** 3873-3883 (2005).

[19] M Stamatakis and NV Mantzaris. Comparison of Deterministic and Stochastic Models of the lac Operon Genetic Network. Biophysical journal **96.3** 887-906 (2009).

[20] M Samoilov, S Plyasunov and AP Arkin. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. Proceedings of the National Academy of Sciences **102.7** 2310-2315 (2005).

[21] M Acar, A Becskei and A van Oudenaarden. Enhancement of cellular memory by reducing stochastic transitions. Nature **435.7039** 228-232 (2005).

[22] LS Weinberger, JC Burnett, JE Toettcher, AP Arkin and DV Schaffer. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. Cell **122.2** 169-182 (2005).

[23] H Li, Z Hou and H Xin. Internal noise stochastic resonance for intracellular calcium oscillations in a cell system. Physical Review E **71.6** 061916 2005.

[24] R Gunawan, Y Cao, L Petzold and FJ Doyle III. Sensitivity analysis of discrete stochastic systems. Biophysical Journal **88.4** 2530-2540 (2005).

[25] P Gorroochurn. *Classic problems of probability*. John Wiley & Sons, 2012.

[26] A Einstein. Zur theorie der brownschen bewegung. Annalen der physik **324.2** 371-381 (1906).

[27] M Von Smoluchowski. Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen. Annalen der physik **326.14** 756-780 (1906).

[28] CW Gardiner. *Handbook of stochastic methods. Vol. 3*. Berlin: Springer, 1985.

[29] PE Kloeden and E Platen. *Numerical solution of stochastic differential equations. Vol. 23*. Springer, 1992.

[30] M Delbrück. Statistical fluctuations in autocatalytic reactions. The Journal of Chemical Physics **8.1** 120-124 (2004).

[31] A Renyi. A discussion of chemical reactions using the theory of stochastic processes. MTA Alk Mat Int Kozl **2** 83-101 (1953).

[32] AF Bartholomay. Stochastic models for chemical reactions: I. Theory of the unimolecular reaction process. The bulletin of mathematical biophysics **20.3** 175-190 (1958).

[33] AF Bartholomay. Stochastic models for chemical reactions: II. The unimolecular rate constant. The bulletin of mathematical biophysics **21.4** 363-373 (1959).

[34] K Ishida. The stochastic model for unimolecular gas reaction. Bulletin of the Chemical Society of Japan **33.8** 1030-1036 (1960).

[35] K Ishida. Stochastic model for bimolecular reaction. The Journal of Chemical Physics **41.8** 2472-2478 (1964).

[36] CJ Jachimowski, DA McQuarrie and ME Russell. A Stochastic Approach to Enzyme-Substrate Reactions. Biochemistry **3.11** 1732-1736 (1964).

[37] DA McQuarrie. Stochastic approach to chemical kinetics. Journal of Applied Probability **4.3** 413-478 (1967).

[38] DT Gillespie. A rigorous derivation of the chemical master equation. Physica A: Statistical Mechanics and its Applications **188.1** 404-425 (1992).

[39] DA McQuarrie. Kinetics of Small Systems I. Journal of Chemical Physics **38** 433-436 (1963).

[40] DA McQuarrie, CJ Jachimowski and ME Russell. Kinetics of Small Systems II. Journal of Chemical Physics **40** 2914-2921 (1964).

[41] NG Van Kampen. *Stochastic processes in physics and chemistry. Vol. 1.* Elsevier, 1992.

[42] DT Gillespie. The chemical Langevin equation. The Journal of Chemical Physics **113.1** 297-306 (2000).

[43] JL Doob. *Stochastic processes. Vol. 101.* Wiley: New York, 1953.

[44] DT Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of computational physics **22.4** 403-434 (1976).

[45] JC Venter and Dl Cohen. The century of biology. New Perspectives Quarterly **21.4** 73-77 (2004).

[46] JC Venter, *et al.* The sequence of the human genome. Science **291.5507** 1304-1351 (2001).

[47] R Carlson. The changing economics of DNA synthesis. Nature biotechnology **27.12** 1091 (2009).

[48] PEM Purnick and R Weiss. The second wave of synthetic biology: from modules to systems. Nature reviews: Molecular cell biology **10.6** 410-422 (2009).

[49] MB Elowitz and S Leibler. A synthetic oscillatory network of transcriptional regulators. Nature **403.6767** 335-338 (2000).

[50] DC Bennett. Differentiation in mouse melanoma cells: initial reversibility and an on-off stochastic model. Cell **34.2** 445-453 (1983).

[51] MA Dingemanse, PA de Boer, AF Moorman, R Charles and WH Lamers. The expression of liver-specific genes within rat embryonic hepatocytes is a discontinuous process. Differentiation **56.3** 153-162 (1994).

[52] MC Walters, S Fiering, J Eidemiller, W Magis, M Groudine and DI Martin. Enhancers increase the probability but not the level of gene expression. Proceedings of the National Academy of Sciences **92.15** 7125-7129 (1995).

[53] MS Ko, H Nakauchi and N Takahashi. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. The EMBO journal **9.9** 2835 (1990).

[54] A Becskei, B Séraphin and L Serrano. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. The EMBO journal **20.10** 2528-2535 (2001).

[55] TS Gardner, CR Cantor and JJ Collins. Construction of a genetic toggle switch in Escherichia coli. Nature **403.6767** 339-342 (2000).

[56] MA Gibson and J Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. The journal of physical chemistry A **104.9** 1876-1889 (2000).

[57] P Wong, S Gladney and JD Keasling. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. Biotechnology progress **13.2** 132-143 (1997).

[58] ZL Xiu, ZY Chang and AP Zeng. Nonlinear Dynamics of Regulation of Bacterial trpOperon: Model Analysis of Integrated Effects of Repression, Feedback Inhibition, and Attenuation. Biotechnology progress **18.4** 686-693 (2002).

[59] KI Ramalingam, JR Tomshine, JA Maynard and YN Kaznessis. Forward engineering of synthetic bio-logical AND gates. Biochemical Engineering Journal **47.1** 38-47 (2009).

[60] TA Carrier and JD Keasling. Investigating autocatalytic gene expression systems through mechanistic modeling. Journal of theoretical biology **201.1** 25-36 (1999).

[61] T Tian and K Burrage. Stochastic models for regulatory networks of the genetic toggle switch. Proceedings of the National Academy of Sciences **103.22** 8372-8377 (2006).

[62] H Qian, S Saffarian and EL Elson. Concentration fluctuations in a mesoscopic oscillating chemical reaction system. Proceedings of the National Academy of Sciences **99.16** 10376-10381 (2002).

[63] J Paulsson, OG Berg and M Ehrenberg. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. Proceedings of the National Academy of Sciences **97.13** 7148-7153 (2000).

[64] JM McCollum, GD Peterson, CD Cox, ML Simpson and NF Samatova. The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. Computational biology and chemistry **30.1** 39-49 (2006).

[65] M Rathinam, LR Petzold, Y Cao and DT Gillespie. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. The Journal of Chemical Physics **119.24** 12784-12794 (2003).

[66] T Tian and K Burrage. Binomial leap methods for simulating stochastic chemical kinetics. The Journal of chemical physics **121.21** 10356-10364 (2004).

[67] Y Cao, DT Gillespie and LR Petzold. The slow-scale stochastic simulation algorithm. The Journal of chemical physics **122.1** 014116 (2004).

[68] EL Haseltine and JB Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. The Journal of chemical physics **117.15** 6959-6969 (2002).

[69] H Salis and YN Kaznessis. An equation-free probabilistic steady-state approximation: dynamic application to the stochastic simulation of biochemical reaction networks. The Journal of chemical physics **123.21** 214106 (2005).

[70] Y Cao, DT Gillespie and LR Petzold. Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. Journal of Computational Physics **206.2** 395-411 (2005).

[71] B Munsky and M Khammash. The finite state projection algorithm for the solution of the chemical master equation. The Journal of chemical physics **124.4** 044104 (2006).

[72] H Salis, V Sotiropoulos and YN Kaznessis. Multiscale Hy3S: Hybrid stochastic simulation for supercomputers. BMC bioinformatics **7.1** 93 (2006).

[73] H Li, Y Cao, LR Petzold and DT Gillespie. Algorithms and software for stochastic simulation of biochemical reacting systems. Biotechnology progress **24.1** 56-61 (2008).

[74] LA Goodman. Population growth of the sexes. Biometrics **9.2** 212-225 (1953).

[75] P Whittle. On the use of the normal approximation in the treatment of stochastic processes. Journal of the Royal Statistical Society. Series B (Methodological) 268-218 (1957).

[76] R Durrett and C Neuhauser. Particle systems and reaction-diffusion equations. The Annals of Probability **22.1** 289-333 (1994).

[77] BT Grenfell, K Wilson, VS Isham, HEG Boyd and K Dietz. Modelling patterns of parasite aggregation in natural populations: trichostrongylid nematode-ruminant interactions as a case study. Parasitology **111.S1** S135-S151 (1995).

[78] R Grima. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. The Journal of chemical physics **136.15** 154105 (2012).

[79] I Krishnarajah, A Cook,G Marion and G Gibson. Novel moment closure approximations in stochastic epidemics. Bulletin of mathematical biology **67.4** 855-873 (2005).

[80] I Krishnarajah, G Marion and G Gibson. Novel bivariate moment-closure approximations. Mathematical biosciences **208.2** 621-643 (2007).

[81] JP Hespanha and A Singh. Stochastic models for chemically reacting systems using polynomial stochastic hybrid systems. International Journal of robust and nonlinear control **15.15** 669-689 (2005).

[82] A Singh and JP Hespanha. Lognormal moment closures for biochemical reactions. 45th IEEE Conference on Decision and Control 2063-2068 2006.

[83] A Singh and JP Hespanha. A derivative matching approach to moment closure for the stochastic logistic model. Bulletin of mathematical biology **69.6** 1909-1925 (2007).

[84] A Singh and JP Hespanha. Approximate moment dynamics for chemically reacting systems. IEEE Transactions on Automatic Control **56.2** 414-418 (2011).

[85] CH Lee, KH Kim and P Kim. A moment closure method for stochastic reaction networks. The Journal of chemical physics **130.13** 134107 (2009).

[86] CA Gomez-Uribe and GC 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).

[87] A Ale, P Kirk and MPH Stumpf. A general moment expansion method for stochastic kinetic models. The Journal of chemical physics **138.17** 174101 (2013).

[88] CS Gillespie and E Renshaw. An improved saddlepoint approximation." Mathematical biosciences **208.2** 359-374 (2007).

[89] J Ruess, A Milias-Argeitis, S Summers and J Lygeros. Moment estimation for chemically reacting systems by extended Kalman filtering. The Journal of chemical physics **135.16** 165102 (2011).

[90] CS Gillespie. Moment-closure approximations for mass-action models. IET systems biology **3.1** 52-58 (2009).

[91] DT Gillespie. *Markov Processes: An Introduction for Physical Scientists*. Academic Press, (1992).

[92] T Jahnke and W Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. Journal of Mathematical Biology **54** 1-26 (2007).

[93] H Salis and YN Kaznessis. Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. The Journal of Chemical Physics **122** 054103 (2005).

[94] C Zechner, J Ruess, P Krenn, S Pelet, M Peter, J Lygeros and H Koeppl. Moment-based inference predicts bimodality in transient gene expression. Proceedings of the National Academy of Sciences **109.21** 8340-8345 (2012).

[95] S Engblom. Computing the moments of high dimensional solutions of the master equation. Applied Mathematics and Computation, **180** 498-515 (2006).

[96] V Sotiropoulos and YN Kaznessis. Analytical derivation of moment equations in stochastic chemical kinetics. Chemical Engineering Science **66** 268-277 (2011).

[97] DT Gillespie. Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry **81** 2340-2361 (1977).

[98] I Oppenheim, KE Shuler and GH Weiss. Stochastic and deterministic formulation of chemical rate equations. The Journal of Chemical Physics, **50** 460-466 (1969).

[99] TG Kurtz. The relationship between stochastic and deterministic models for chemical reactions. The Journal of Chemical Physics, **57** 2976-2978 (1972).

[100] I Oppenheim, KE Shuler and GH Weiss. Stochastic theory of nonlinear rate processes with multiple stationary states. Physica A: Statistical Mechanics and its Applications, **88** 191-214 (1977).

[101] H Grabert, P Hänggi and I Oppenheim. Fluctuations in reversible chemical reactions. Physica. A, Statistical mechanics and its applications. **117** 300-316 (1983).

[102] MB Elowitz, AJ Levine, ED Siggia and PS Swain. Stochastic gene expression in a single cell. Science, **297** 1183-1186 (2002).

[103] CV Rao, DM Wolf and AP Arkin. Control, exploitation and tolerance of intracellular noise. Nature, **420** 1183–1186 (2002).

[104] I Matheson, DF Walls and CW Gardiner. Stochastic models of first-order nonequilibrium phase transitions in chemical reactions. Journal of Statistical Physics, **12** 21-34 (1975).

[105] P Smadbeck and YN Kaznessis. Efficient moment matrix generation for arbitrary chemical networks. Chemical Engineering Science, **84** 612-618 (2012).

[106] CE Shannon. A mathematical theory of communication. Bell System Technical Journal, **27** 379-423 (1948).

[107] ET Jaynes. *Probability Theory: The Logic of Science*. G. L. Bretthorst Ed, Cambridge University Press (2003).

[108] DL Nelson and MM Cox. *Lehninger Principles of Biochemistry*. D. L. Nelson, M. M. Cox, Eds., W. H. Freeman (2005).

[109] AM Lyapunov. The general problem of the stability of motion. International Journal of Control, **55** 539-589 (1992).

[110] RE Kalman and JE Bertram. Control system analysis and design via the second method of Lyapunov: I - Continuous-time systems. Journal of Basic Engineering, **82** 371-393 (1960).

[111] IR Epstein and JA Pojman. *An Introduction to Nonlinear Chemical Dynamics*. New York: Oxford University Press, (1998).

[112] G László and RJ Field. A three-variable model of deterministic chaos in the Belousov-Zhabotinsky reaction. Nature **355** 808-810 (1992).

[113] RB Warden, R Aris and NR Amundson. An analysis of chemical reactor stability and control - VIII: The direct method of Lyapunov. Chemical Engineering Science **19** 149-172 (1964).

[114] HB Callen and RF Greene. On a theorem of irreversible thermodynamics. Physical Review **86** 702-710, (1952).

[115] R Kubo. The fluctuation-dissipation theorem. Reports on Progress in Physics **29.1** 255 (1966).

[116] P Smadbeck and YN Kaznessis. A closure scheme for chemical master equations. Proceedings of the National Academy of Sciences **35** 14261-14265 (2013).

[117] PB Warren, S Tănase-Nicola and PR Ten Wolde. Exact results for noise power spectra in linear biochemical reaction networks. The Journal of Chemical Physics **125** 144904:1-10 (2006).

[118] BO Palsson and EN Lightfoot. Mathematical modelling of dynamics and control in metabolic networks. I. On Michaelis-Menten kinetics. Journal of Theoretical Biology **111** 273-302 (1984).

[119] M Vellela and H Qian. Stochastic dynamics and non-equilibrium thermodynamics of a bistable chemical system: the Schlögl model revisited. Journal of The Royal Society Interface **6** 925-940 (2009).

[120] W Hillen and C Berens. Mechanisms underlying expression of Tn10 encoded tetracycline resistance. Annual Review of Microbiology **48.1** 345-369 (1994).

[121] K Biliouris, P Daoutidis and YN Kaznessis. Stochastic simulations of the tetracycline operon. BMC Systems Biology **5.1** 9 (2011).

[122] R Schleif. Regulation of the l-arabinose operon of Escherichia coli. Trends in Genetics **16.12** 559-565 (2000).

[123] V Sotiropoulos and YN Kaznessis. Synthetic tetracycline-inducible regulatory networks: computer-aided design of dynamic phenotypes. BMC systems biology **1.1** 7 (2007).

[124] CS Gillespie. Stochastic simulation of chemically reacting systems using multi-core processors. The Journal of Chemical Physics **136.1** 014101 (2012).

[125] S Kadam and K Vanka. A new approximate method for the stochastic simulation of chemical systems: The representative reaction approach. Journal of Computational Chemistry **33.3** 276-285 (2012).

[126] S Wu, J Fu, H Li and L Petzold. Automatic identification of model reductions for discrete stochastic simulation. Journal of Chemical Physics **137.3** 034106 (2012).

[127] KR Sanft, DT Gillespie and LR Petzold. Legitimacy of the stochastic michaelis-menten approximation. IET Systems Biology **5.1** 58-69 (2011).

[128] C Rao and A Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. The Journal of Chemical Physics **118.11** 4999-5010 (2003).

[129] J Murray. *Mathematical biology*. Springer (2003).

[130] A Hill. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. Journal of physiology **40.4** 4-7 (1910).

[131] A Revzin and PH Von Hippel. Direct measurement of association constants for the binding of Escherichia coli lac repressor to non-operator DNA. Biochemistry **16.22** 4769-4776 (1977).

[132] M Dunaway, J Olson, J Rosenberg, O Kallai, R Dickerson and K Matthews. Kinetic studies of inducer binding to lac repressor. operator complex. Journal of Biological Chemistry **255.21** 10115-10119 (1980).

[133] T Lederer, M Kintrup, M Takahashi, PE Sum, GA Ellestad and W Hillen. Tetracycline analogs affecting binding to Tn 10-encoded Tet repressor trigger the same mechanism of induction. Biochemistry **35.23** 7439-7446 (1996).

[134] E Weeding, J Houle and YN Kaznessis. SynBioSS designer: a web-based tool for the automated generation of kinetic models for synthetic biological constructs. Briefings in Bioinformatics **11.4** 394-402 (2010).

[135] U Vogel and KF Jensen. The RNA chain elongation rate in Escherichia coli depends on the growth rate. Journal Of Bacteriology **176.10** 2807-2813 (1994).

[136] MA Sørensen and S Pedersen. Absolute in vivo translation rates of individual codons in Escherichia coli: The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. Journal of Molecular Biology **222.2** 265-280 (1991).

[137] W Hillen, C Gatz, L Altschmied, K Schollmeier and I Meier. Control of expression of the Tn10-encoded tetracycline resistance genes: Equilibrium and kinetic investigation of the regulatory reactions. Journal of Molecular Biology **169.3** 707-721 (1983).