# PARAMETER ESTIMATION IN DETERMINISTIC AND STOCHASTIC MODELS OF BIOLOGICAL SYSTEMS

by

ANKUR GUPTA

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Chemical Engineering)

at the
UNIVERSITY OF WISCONSIN-MADISON
2013

Date of final oral examination: December 11, 2013

The dissertation is approved by the following members of the Final Oral Committee:

James B. Rawlings, Professor Chemical and Biological Engineering
John Yin, Professor, Chemical and Biological Engineering
Christos T. Maravelias, Associate Professor, Chemical and Biological Engineering
Jennifer L. Reed, Associate Professor, Chemical and Biological Engineering
Ross E. Swaney, Associate Professor, Chemical and Biological Engineering
David F. Anderson, Assistant Professor, Mathematics

To my grandfather, whose belief in me never wavered.

# ACKNOWLEDGMENTS

My time as a graduate student at the University of Wisconsin-Madison has been an adventure. I have met unbelievably nice people in this town, made many friends, and inevitably, a few enemies. I entered this department as a naive student and I leave here still as a student, albeit less naive.

I have had the opportunity to be co-advised by two advisors, Prof. James B. Rawlings and Prof. John Yin. I am immensely thankful to both for providing me the opportunity to work under them. My research experience has been truly unique because of the two different perspectives I received from my advisors. I am thankful for the incredible flexibility they allowed me during my research. I was allowed to make mistakes and learn from them. My interactions with Prof. Rawlings have always been thoroughly enlightening and inspiring. Every discussion with him sharpened my mind and improved my ability to think critically about the problem at hand. I can honestly say that I have grown smarter in these five years because of him. Professor Yin has been incredibly supportive over these five years and allowed me to freely pursue my ideas and offer constructive criticism even when I went astray. It is difficult to imagine the five years of research under any other advisors.

I am grateful to Prof. Christos Maravelias, Prof. Jennifer Reed, Prof. Ross Swaney and Prof. David Anderson for taking the time to be on my thesis committee.

Raj Shekhar, a graduate student in dePablo group, was my mentor both in IIT Kharagpur and UW-Madison. He helped me in many ways during first year in Madison and whose honest advice I later ignored at my own peril. I am very grateful to Rishi Amrit who taught me how to use Linux. I can proudly say that I attempted to fill his shoes as the group's network administrator after he left. He took time out of his busy schedule to teach me the ways of Linux. I also thank him for providing me car rides home when I broke my foot, moral support and invaluable career mentoring which continues to this day. I am grateful to John Eaton, who watched over our group computers from afar. Thanks to him, my annoying questions were answered and mistakes were forgiven and then corrected.

My life here would have been incredibly boring without Kaushik Narasimhan, the God of matrices. He is a trusted friend who provided engaging brainstorming sessions in academics, intellectual discussions in fantasy fiction, and much needed career mentoring during tumultuous times. He is a friend I hope to retain for the rest of my life. I am lucky to have met Brett Stewart who was the senior most member of Rawlings group when I joined. He provided the guidance, leadership and entertaining Rawlings group trivia. I still remember our forays into the works of Aristotle and other philosophers. Fernando Lima also provided much needed guidance and help until my final years.

Luo Ji was always a smiling face, even in the toughest of times, and provided help whenever I required. I wish him and other members of Rawlings group — Cuyler Bates, Megan Zagrobelny and Min Yao — the best of luck. I am very thankful to Eric Haseltine who provided me the invaluable internship experience and mentoring. I obtained a completely new perspective and confidence after my internship. Mary Diaz has been wonderful in providing administrative and computer support as well as delicious cakes and cookies all year round. I am grateful for all her hard work and support. Of course, these acknowledgements would be incomplete without mentioning the enjoyable time I spent with the visiting students and researchers in Rawlings group — Emil Sokoler, Ganzhou Wang, Johannes Philippus Maree, Philip Kunz, Giulio Mancuso, Stefano Riverso, Rolf Stierle and Behzad Sharefi.

I am fortunate enough to be the part of both Rawlings research group and Yin lab. While I spent time in Rawlings group offices, I collaborated regularly with the members of Yin lab. I am very thankful to the entire Yin lab for being so immensely accommodating. They enthusiastically listened to my research talks about “die throw experiments” and “problems in optimization” and asked insightful questions. My experience in this department is truly unique because of them. Collin Timm was a classmate and a collaborator who always had an open mind. He saw me make mistakes and helped me look for them. And a more humble person I have not yet met. Andrea Timm was a project group member during my first year and taught me the basics of cell culture and counting during the few times I did experiments. Later, she performed experiments which formed the inspiration for some of my research. My discussions with Jay Warrick regarding modeling, regression and computer softwares in general was very helpful. Much of my skills in collaboration may be attributed to these discussions. I also want to thank other members of Yin lab — Adam Swick, Fulya Apkinar, Emily Voigt, Bahar Inankur, Ashley Baltes.

I have had the pleasure of being friends with many students outside this department. Devashish Das, a fellow KGPian, is a friend whose company I enjoyed

very much. He was always available for any help I asked him for. I am thankful for the discussions we have had ranging from politics to statistics. He listened to my naive questions about statistical methods and helped me learn a great deal of statistics. I am thankful to have met Jacob Morth, who is a trusted friend and was always there to help me. My life in Madison has been exceptionally enjoyable because of him. I would like to thank Goldy Kumar, Jacob Trowbridge, Emily Leibold, Amanda Keen and Santosh Mutyala for a great time in Madison.

Obviously, none of this would have been possible without the support of my family. I have faced turbulent times during these five years and my parents were always there to provide moral support even when I did not deserve it. My brother has been a trusted advisor who helped me in every aspect of my life, be it moral support or computer science. My grandfather is the one whom I owe the most gratitude. His moral fiber, pragmatism in the face of adversity and unwavering belief in me has been my last bastion of strength throughout my life. Even when others were struck with doubt and my own confidence failed me, he was there as a rock. Truly, I would have given up more times than I could count if it were not for him.

Finally, the accidental omission of a name that rightly deserves a mention should be attributed to the forgetfulness of a tense mind instead of an absence of gratitude.

Ankur Gupta
Madison, WI
December 2013

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

## ABSTRACT

Viruses pose a threat to human health. Understanding how viruses work helps us develop vaccines and antivirals. Experimental techniques are now advanced enough to provide quantitative data regarding viral infection. Using this data, we can develop mathematical models to describe viral infection processes. Two such modeling paradigms are deterministic and stochastic reaction systems. Useful mathematical models require accurate estimates of model parameters from data. In this dissertation, I present parameter estimation methods for deterministic and stochastic reaction models, focusing on the stochastic models. I present two new classes of parameter estimation methods for stochastic chemical kinetic models, namely, importance sampling and approximate direct methods. Using examples from systems biology, I demonstrate the use of these newly developed methods and compare them with literature methods with favorable results. Guidelines on experimental and model design and directions for further research are presented in the end.

# 1

# INTRODUCTION

## 1.1 MOTIVATION

Viruses cause diseases like HIV, HCV and Influenza and pose a threat to human health. For example, chronic Hepatitis C (HCV) affects approximately 170 million people around the world. The number of HIV infected patients is about 35 million. Billions of dollars of money is being spent to develop vaccines and cures for these viral illnesses. In contrast, viruses may also be used to treat cancer (oncolytic virus) and may help in gene therapy (viral gene therapy). Study of viruses and other biological entities has increasingly become important. Advances in experimental techniques has allowed the quantitative measurements in various biological processes. For example, during an *in vitro* viral infection process, the amounts of viral RNA, DNA and proteins may be measured over time. Availability of such quantitative data has allowed us to build mathematical models of biological systems. Instead of isolating each quantity and attempting to study it individually, a *systems biology* approach attempts to study the interactions between the systems. Mathematical modeling of biological systems is an especially effective approach for this purpose.

These biochemical processes are represented as chemical reactions, which can then be modeled mathematically. The model describes the various reacting species and contains unknown model parameters, which are usually rate constants. Using experimental time series data, these unknown model parameters may be estimated. The estimation of model parameters using reaction models is the central contribution of this dissertation.

Depending upon the physical regime and experimental data, two kind of reaction kinetic models may be formulated — (1) the traditional, deterministic mass-action models, and (2) stochastic (or random) reaction models. Deterministic formulation applies when the number of molecules of reacting species is large while the stochastic formulation is valid in the regime of far fewer reacting molecules. The two modeling paradigms are different and they require different mathematical approaches. Nevertheless, parameter estimation is a common problem that applies to both modeling paradigms. Considerable research has been done to develop tools and software that can perform parameter estimation in deterministic models but parameter estimation in discrete stochastic chemical kinetic models is still an active area of research.

In this dissertation, I present parameter estimation methods for both kinds of models. For the deterministic models, I present a brief overview of mathematical techniques and software tools. In the case of stochastic models, I present two new classes of parameter estimation methods — importance sampling based methods and approximate direct methods — that I developed over the course of my doctoral research. Using relevant examples from systems biology, I demonstrate the use of the newly developed methods. I also present a comprehensive overview and comparison of the related literature methods which are based on Markov Chain Monte Carlo methods. Finally, I present some experimental and model de-

sign guidelines that are specific to the stochastic chemical kinetic models. This dissertation contains the following classes of parameter estimation methods for stochastic chemical kinetic models

1. Closed-form expressions for parameter estimates given complete data
2. Exact inference method based on Markov chain representation
3. Simulation methods
   (a) MCMC methods
      - MCMC-Unif: MCMC using endpoint-conditioned simulation
      - MCMC-MH: MCMC using Metropolis-Hastings approach
   (b) CDIS: Importance sampling based methods
4. Approximate direct (AD) methods using limited sampling

## 1.2 NOTATION AND LANGUAGE

A central list of symbols is not provided. There is a finite set of symbols which is easily exhausted even with the use of superscripts, subscripts and ornamentation. This thesis covers many fields of study — biology, chemical engineering, probability and statistics — each with its own set of preferred symbols. Therefore, creating a unique notation for the entire thesis not only places a significant burden on the author, but is also unintuitive for the reader. Each chapter should be treated to have its own notation. In cases where some notation is inherited from other chapters, it is so indicated. Notation, within each chapter, is defined incrementally over sections. For example, the notation is presented once in Section 5.1 and then used by the remaining sections of Chapter 5.

*Gamma distribution.* A gamma distribution is represented by $Ga(a, b)$ in which $a$ represents the shape parameter and $b$ represents the rate parameter. The entire

thesis uses the shape-rate parameterization of the gamma distribution.

*Random variables and distributions.* As an example, the terms "gamma random variable", "gamma-distributed random variable", "random variable from a gamma distribution" all have the same meaning.

*Random variables and their instances.* A random variable is denoted by uppercase (Greek or Roman) letters, for example, $X$ or $\Theta$. The specific instance (or sample) of the random variable is represented by lowercase (Greek or Roman) letters, for example, $x$ or $\theta$.

*Symbols $P$, $f$, $p$, $\pi$.* The uppercase symbol, when in the relevant position, $P$ denotes the probability (which is between 0 and 1). For a discrete random variable $X$, $P(X = x)$ denotes the probability that the random variable $X$ assumes a value $x$. As a shorthand, the name of the random variable is dropped, for example, $P(x)$ represents a shorthand for $P(X = x)$. The name of the random variable is either the capitalized version of the argument of $P(\cdot)$ or is otherwise clear from the context. The notation $\pi(\mathbf{Z} = \mathbf{z})$ or $\pi(\mathbf{z})$ is used to denote the value of the probability density function of a vector-valued random variable $\mathbf{Z}$ (or, in general a set of random variables $\mathbf{Z}$). The probability of an event will be denoted using $P(\cdot)$, for example, the probability of the event $\{\mathbf{Z} \in \mathbb{A}\}$ is written as $P(\mathbf{Z} \in \mathbb{A})$. In the appendix, the probability density function and the cumulative distribution function of a scalar, continuously distributed, random variable $T$ are represented as $f_T(t)$ and $F_T(t)$ respectively. To avoid ambiguity between probability and probability density, the notation $p(\cdot)$, when used, is expressly defined.

*Vectors and matrices.* Usually, the matrices and vectors are denoted by boldface characters, for example, $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^{n \times n}$. Greek letters representing vectors

or matrices, for example, the stoichiometric matrix, $\nu$, are not denoted using boldface. Vectors which are indexed with a subscript, for example, $Y_i \in \mathbb{R}^{n_{\text{meas}}}$ are not denoted in boldface.

*Abbreviations.* Lowercase abbreviations, *rv*, *pdf*, *cdf* mean *random variable*, *probability density function*, *cumulative distribution function*, respectively.

*Estimates $\hat{\theta}$, $\hat{k}_i$.* Unless otherwise stated, the symbols, $\hat{\theta}$ and $\hat{k}_i$ represent the maximum *a posteriori* (MAP) estimates obtained using the method being discussed. The subscript MLE is used specifically to denote the maximum likelihood estimates as $\hat{\theta}_{\text{MLE}}$ and $\hat{k}_{i,\text{MLE}}$. When Bayesian inference methods are being compared, relevant subscript such as CDIS will be used to denote the MAP estimates obtained by the corresponding method, for example, $\hat{\theta}_{\text{CDIS}}$ and $\hat{k}_{i,\text{CDIS}}$.

*Reaction rate and reaction propensity.* The term "reaction rate" is only applicable for the deterministic formulation of a reaction while the term "reaction propensity" is applicable only for the stochastic version. The two expressions, reaction rate and reaction propensity, are similar with some differences as described in (Gillespie, 1976 [32]). Reaction propensity is also called the "hazard rate" (Wilkinson, 2012 [119]). In this thesis, I use the term "reaction rate" to refer to "reaction propensity" as well, with the meaning being clear from context.

*Deterministic and stochastic parameters, $k$.* Deterministic formulation of a chemical reaction uses the parameter, $k$, as the "reaction rate constant" (Rawlings and Ekerdt, 2004 [85], Gillespie, 1976 [32]) while the stochastic formulation uses the parameter, $k$, as the "hazard rate constant" (Wilkinson, 2012 [119]) or "reaction parameter" (Gillespie, 1976 [32]). Both deterministic and stochastic formulations use the same symbol, $k$, to denote their respective rate parameters. Ambiguity

arises only when the two formulations are being compared, in which case, the intended meaning of the symbol, $k$, is expressly specified. Further, I use the terms "stochastic rate constant" to refer to the hazard rate constant and "deterministic rate constant" to refer to the reaction rate constant. The term "rate constant" refers to either stochastic rate constant or deterministic rate constant depending upon context.

*Stochastic reactions and deterministic reactions.* As discussed in Chapter 2, a chemical reaction has two formulations, stochastic and deterministic, which are applicable in different regimes. The phrase "stochastic formulation of the chemical reaction" is abbreviated to "stochastic chemical reactions". The same is applicable for the deterministic case.

## 1.3 AN OVERVIEW OF THE THESIS

This dissertation considers two major mechanistic modeling paradigms — deterministic and stochastic — used to study systems biology phenomena. The focus is on assimilating experimentally observed data by estimating relevant model parameters, specifically in the case of stochastic chemical kinetic models. The rest of the dissertation is organized as follows.

**Chapter 2 – Mathematical Modeling in Systems Biology.** This chapter introduces the concept of using mathematical modeling to describe biological systems. Essentially, a biological system is treated as a set of biochemical reactions which may be studied using either deterministic or stochastic framework depending on the physical regime and nature of experimental data. I begin by describing how to convert a system of reactions into either a set of nonlinear ODEs (deterministic

formulation) or a continuous-time, discrete state space Markov chain (stochastic formulation). Various methods of simulation for both formulations are presented, with a focus on the stochastic formulation. Similarities and differences between the two frameworks are discussed. Some definitions and theoretical necessities are also presented with examples. Finally, the parameter estimation problems for both frameworks are introduced. This chapter serves as a foundation for later chapters of the dissertation.

**Chapter 3 – Parameter estimation in deterministic models.** In this chapter, I present the methods for estimating parameters using deterministic models. The nonlinear optimization problem is formulated and the relevant software tools are discussed. A relevant example from virology is briefly dicussed.

**Chapter 4 – Model reduction using stochastic reaction equilibrium assumption.** Stochastic and deterministic versions of the reaction equilibrium assumption (REA) is presented. Two examples, with linear and nonlinear kinetics, respectively, are presented. For the linear kinetics example, a reduced kinetics is obtained for the stochastic REA which was previously unknown. Using, the nonlinear kinetics example, I demonstrate that the deterministically reduced kinetics disagree with the stochastically reduced kinetics, which was not previously proved.

**Chapter 5 – Overview: Parameter estimation in stochastic chemical kinetic models.** In this chapter, I present various methods that have been proposed in the literature to estimate parameters in stochastic chemical kinetics models. Using examples, I demonstrate the strengths and weaknesses of these methods. The problem of parameter inference in the sense of maximum likelihood estimation and Bayesian estimation is formulated incrementally through this chapter. Specific insights regarding the informativeness of data are also presented. This chapter

forms a basis and motivation for the next chapter.

**Chapter 6 – New methods for parameter estimation in stochastic chemical kinetic models.** This chapter is the main contribution of this dissertation. Two new proposed classes of parameter estimation methods are presented, namely importance sampling based methods and approximate methods. Drawing upon the notation in Chapter 5, this chapter provides a detailed development of the two new classes of methods. Using examples, all applicable methods are compared in terms of estimator accuracy and computational expense.

**Chapter 7 – Parameter estimation in systems biology.** Two examples in systems biology are presented. The first example is a new model of early viral gene expression developed to describe experiments performed by members of Yin lab. The second example is taken from the literature. Using simulated data, the newly developed parameter estimation methods are compared against literature methods. Some experimental guidelines, specific to stochastic kinetic modeling, are presented.

**Chapter 8 – Conclusions and future directions.** A summary of the major contributions of this dissertation is presented. Specific areas in the fields of experimental measurement techniques, experimental design, modeling, numerical methods, computational software, probability and statistics that require further research are identified.

# 2

## MATHEMATICAL MODELING IN SYSTEMS BIOLOGY

A system of chemical reactions may be formulated using the traditional deterministic paradigm or using the stochastic paradigm. The [1] classical method to simulate the time evolution of reacting molecules is based on the continuum assumption . When the number of molecules of all reacting species in a set of chemical reactions is of the order of Avagadro's number, the concentration can be assumed to be a continuous real variable. In such cases, classic mass action kinetics can be used to describe the rates of reaction. When the number of molecules of one or more species is on the order of hundreds or thousands, however, we can no longer use the continuum assumption. As a result, instead of real-valued concentrations we need to consider the integer-valued number of molecules. Another effect of such low number of molecules is that the classical mass action kinetics is no longer valid. The reaction rates are no longer deterministic, and a probabilistic approach is required. Instead of accounting for amount of reactants consumed (and products produced) in a time interval, we need to account for the probability that a reaction occurs in a time interval. This approach of modeling chemical reactions has come to be known as stochastic chemical kinetics [70, 32, 121], [112,

[1] Most of this paragraph appears in Gupta and Rawlings, 2013 [43]

p. 166].

Another difference between the deterministic and stochastic regimes of chemical kinetics is seen in the simulation methods. While the classical, deterministic regime involves solving a system of coupled ordinary differential equations (ODEs) , the stochastic regime requires a probabilistic method involving repeated generation of random numbers. Various algorithms have been developed to simulate stochastic chemical kinetics starting with the stochastic simulation algorithm (SSA) , also known as Gillespie's algorithm [32, 33]. Note that unlike the solution of a system of ODEs, the SSA produces a different (random) trajectory for every simulation. Considerable research has been devoted to developing alternative algorithms [31, 4] and approximations with applications to various problems [103, 67, 16, 48, 47].

Cellular processes in systems biology frequently demonstrate intrinsic noise or randomness [25, 10, 83], which is caused by the presence of low number of reacting molecules. Arkin and others [69, 6] used stochastic chemical kinetic models to explain how a homogeneous population of $\lambda$-infected *E. coli* cells can partition into sub-populations that follow different pathways. Weinberger et al. [118] demonstrated that the otherwise unexplained phenotypic bifurcation observed in HIV-1 infected T-cells could be accounted for by the presence of low molecular concentrations and a stochastic transcriptional feedback loop. Hensel et al. [52] developed a detailed model of intracellular growth of *Vesicular stomatitis virus* (VSV) to demonstrate that stochastic gene expression contributes to the variation in viral yields. Neuert et al. [75] developed a stochastic dynamic model to predict the behavior of mRNA expression in *Saccharomyces cerevisiae*.

This chapter is organized as follows. In Section 2.1, I describe a general system of chemical reactions and detail the basic relationship between the deterministic and stochastic formulations. In Section 2.2 and Section 2.3, I introduce the mod-

eling and simulation of deterministic and stochastic formulations, respectively. Finally, in Section 2.4, I define the terms cyclical and non-cyclical kinetics.

## 2.1 CHEMICAL REACTION KINETICS

Consider a system of chemical reactions with $n_r$ reactions, denoted by $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_{n_r}$ and $n_s$ species, denoted by $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_{n_s}$. This $n_r \times n_s$ system may be represented as the following system of chemical reactions

$$\begin{aligned}
\mathcal{R}_1 &: \quad \nu'_{1,1}\mathcal{X}_1 + \nu'_{1,2}\mathcal{X}_1 + \cdots + \nu'_{1,n_s}\mathcal{X}_{n_s} \quad \xrightarrow{k_1} \quad \nu''_{1,1}\mathcal{X}_1 + \nu''_{1,2}\mathcal{X}_1 + \cdots + \nu''_{1,n_s}\mathcal{X}_{n_s} \\
\mathcal{R}_2 &: \quad \nu'_{2,1}\mathcal{X}_1 + \nu'_{2,2}\mathcal{X}_1 + \cdots + \nu'_{2,n_s}\mathcal{X}_{n_s} \quad \xrightarrow{k_2} \quad \nu''_{2,1}\mathcal{X}_1 + \nu''_{2,2}\mathcal{X}_1 + \cdots + \nu''_{2,n_s}\mathcal{X}_{n_s} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
\mathcal{R}_{n_r} &: \quad \nu'_{n_r,1}\mathcal{X}_1 + \nu'_{n_r,2}\mathcal{X}_1 + \cdots + \nu'_{n_r,n_s}\mathcal{X}_{n_s} \quad \xrightarrow{k_{n_r}} \quad \nu''_{n_r,1}\mathcal{X}_1 + \nu''_{n_r,2}\mathcal{X}_1 + \cdots + \nu''_{n_r,n_s}\mathcal{X}_{n_s}
\end{aligned} \tag{2.1}$$

The corresponding, $n_r \times n_s$ sized, stoichiometric matrix, $\nu$, may be written as

$$\nu = \begin{bmatrix} \nu_{1,1} & \nu_{1,2} & \cdots & \nu_{1,n_s} \\ \nu_{2,1} & \nu_{2,2} & \cdots & \nu_{2,n_s} \\ \vdots & \vdots & \vdots & \vdots \\ \nu_{n_r,1} & \nu_{n_r,2} & \cdots & \nu_{n_r,n_s} \end{bmatrix} \tag{2.2}$$

in which,

$$\nu_{i,j} = \nu''_{i,j} - \nu'_{i,j} \qquad\qquad i = 1, 2, \ldots, n_r,\ j = 1, 2, \ldots, n_s$$

The above notation of a system of chemical reactions and the corresponding stoichiometric matrix is common to deterministic and stochastic formulations. In spite of this similarity, both formulations have completely different governing equations

which involve widely different physical and mathematical quantities and definitions. The two important quantities are the parameter $k_i$ associated with reaction $\mathcal{R}_i$ and the rate or propensity at which the reaction proceeds.

For the deterministic formulation, the parameter $k_i$, $i = 1, 2, \ldots, n_r$, represents the traditional *reaction rate constant* while for the stochastic formulation, $k_i$ is the *hazard rate constant* or the *reaction parameter* (Gillespie, 1976 [32]) . Note that the deterministic reaction rate constant and the stochastic hazard rate constants are not necessarily equal. To illustrate the difference between the deterministic these two quantities, I expressly define the deterministic reaction rate constant as $k_i^{\text{det}}$ and the stochastic hazard reaction parameter as $k_i^{\text{stoc}}$. The definition of the stochastic reaction parameter was first provided by Gillespie, 1976 [32]. This definition is re-stated below.

**Assumption 2.1** (Fundamental hypothesis of stochastic chemical kinetics [32])**.** For a chemical reaction $\mathcal{R}_i$, the reaction parameter $k_i^{\text{stoc}}$ may be defined as follows

$$k_i^{\text{stoc}} dt \stackrel{d}{=} \text{average probability, to first order in } dt, \text{ that a specific combination of the reactant molecules of reaction } \mathcal{R}_i \text{ will react in the next } dt \text{ time interval} \tag{2.3}$$

Using the above assumption and other assumptions (see Gillespie, 1976 [32]), the probability that reaction $\mathcal{R}_i$ will occur in the next $dt$ time interval.

$$\begin{aligned} &P(\text{Reaction } \mathcal{R}_i \text{ will occur in the next time interval } dt) \\ &= n_i^{\text{comb}} k_i^{\text{stoc}} dt \; + \; o(dt) \end{aligned} \tag{2.4}$$

in which, $n_i^{\text{comb}}$ represents number of distinct combinations of reactant molecules of $\mathcal{R}_i$. Using Eq. (2.4), it is clear that the stochastic reaction parameter has the units of $\texttt{time}^{-1}$. Table 2.1 shows the values of $n_i^{\text{comb}}$ for various types of reactions.

Table 2.1: Number of reactant combinations for various types of reactions

| Reaction Type | Example | $n^{\text{comb}}$ |
|---|---|---|
| Monomolecular | $\text{A} \xrightarrow{k} \text{B}$ | $A$ |
| Bimolecular | $\text{A} + \text{B} \xrightarrow{k} \text{C}$ | $AB$ |
| Bimolecular | $\text{A} + \text{A} \xrightarrow{k} \text{B}$ | $\frac{A(A-1)}{2}$ |
| Trimolecular | $\text{A} + \text{B} + \text{C} \xrightarrow{k} \text{D}$ | $ABC$ |
| Trimolecular | $2\text{A} + \text{B} \xrightarrow{k} \text{C}$ | $\frac{A(A-1)}{2}B$ |

Note: The variables $A$, $B$ and $C$ denote the number of molecules of the species A, B and C, respectively.

A detailed development of the expressions provided in Table 2.1 may be found in Gillespie, 1976 [32]. Generalizing Eq. (2.4), the term *reaction propensity* may be defined as follows.

**Definition 2.1** (Reaction propensity)**.** The *propensity* of a reaction $\mathcal{R}_i$, denoted by $h_i$, is defined by the following equation.

$$\begin{aligned} &P(\text{Reaction } \mathcal{R}_i \text{ will occur in the next time interval } dt) \\ &= h_i dt \; + \; o(dt) \end{aligned} \tag{2.5}$$

Under Assumption 2.1 and others [32], the *reaction propensity* for reaction $\mathcal{R}_i$ is given by

$$h_i = k_i^{\text{stoc}} n_i^{\text{comb}} \tag{2.6}$$

Since stochastic formulation considers a probabilistic view of chemical reactions, the number of molecules of all species are random variables. The reaction propensity, $h_i$, of a reaction $\mathcal{R}_i$, is a function of the number of reactant molecules. Thus, $h_i$ is also a random variable.

The deterministic formulation, unlike the stochastic version, defines the term *reaction rate*, $r_i^{\text{det}}$ in the familiar manner. The expressions for reaction rate, $r_i^{\text{det}}$ may be found in Rawlings and Ekerdt [85, p. 193]. The relationship between $k_i^{\text{det}}$ and $k_i^{\text{stoc}}$ is explained by (Gillespie, 1976 [32]) using statistical and physical arguments. In its essence the relationship between $k_i^{\text{det}}$ and $k_i^{\text{stoc}}$ is derived from the following basic argument [32]

$$\text{average reaction rate (in } \frac{\texttt{moles}}{\texttt{volume} \cdot \texttt{time}}) = \frac{\mathbb{E}[h_i]}{N_a V} = r_i \tag{2.7}$$

in which, $N_a \approx 6.022 \times 10^{23}\ \texttt{moles}^{-1}$ is the Avagadro's constant, $\mathbb{E}[h_i]$ represents the expectation of the random variable $h_i$, and $V$ represents the volume of the reactor $V$ [2]. The following example of a monomolecular reaction explains the use of this argument to derive the relationship between $k_i^{\text{det}}$ and $k_i^{\text{stoc}}$.

*Example: Monomolecular reaction*

Consider the following monomolecular reaction

$$\text{A} \xrightarrow{k} \text{B} \tag{2.8}$$

in which, $k$ may be interpreted to mean $k_i^{\text{det}}$ or $k_i^{\text{stoc}}$ depending upon the corresponding formulation. For the deterministic formulation the rate of change *rate of reaction*, $r$, is defined as

$$r = k^{\text{det}} c_A \tag{2.9}$$

in which, $c_A$ denoted the concentration of the species A in a reactor of volume $V$. The units of $k^{\text{det}}$ are $\texttt{time}^{-1}$, which are also the units for $k^{\text{stoc}}$. The reaction

[2] For a biological system, for example a cell, the reactor volume $V$, could be the volume of the cell.

propensity for this monomolecular reaction is given by

$$h = k^{\text{stoc}} A \tag{2.10}$$

in which, $A$ is a random variable denoting the the number of molecules of species A in a reactor of volume $V$. Using Eq. (2.7),

$$\begin{aligned} \frac{\mathbb{E}[h]}{N_a V} &= r \\ k^{\text{stoc}} \mathbb{E}[A] &= k^{\text{det}} c_A (N_a V) \end{aligned} \tag{2.11}$$

in which $\mathbb{E}[A]$ denotes the average number of molecules of species A. The concentration is related to the number of molecules as

$$\text{concentration} = \frac{\text{average number of molecules}}{N_a V} \tag{2.12}$$

Using the above relation, Eq. (2.11) can be re-written as

$$\begin{aligned} k^{\text{stoc}} \mathbb{E}[A] &= k^{\text{det}} \mathbb{E}[A] \\ k^{\text{det}} &= k^{\text{stoc}} \end{aligned} \tag{2.13}$$

The same argument may be used (with an additional assumption [32]) to determine the relationship between the deterministic and stochastic parameters for other types of reactions. Table 2.2 provides a few examples. Since the meaning of the parameter is clear from the context, I use the same symbol $k$ to denote the both deterministic and stochastic parameters instead of explicitly denoting them by $k^{\text{stoc}}$ and $k^{\text{det}}$. For the system of chemical reactions in Eq. (2.1), the set of parameters, $k_i$, $i = 1, 2, \ldots, n_r$, are collectively denoted by $\theta = \begin{bmatrix} k_1 & k_2 & \cdots & k_{n_r} \end{bmatrix}^T$.

Table 2.2: Relationship between $k^{\text{stoc}}$ and $k^{\text{det}}$ various types of reactions

| Reaction Type | Example | Relationship |
|---|---|---|
| Monomolecular | $A \xrightarrow{k} B$ | $k^{\text{det}} = k^{\text{stoc}}$ |
| Bimolecular | $A + B \xrightarrow{k} C$ | $k^{\text{det}} = (N_a V)k^{\text{stoc}}$ |
| Bimolecular | $A + A \xrightarrow{k} B$ | $k^{\text{det}} = \frac{(N_a V)}{2}k^{\text{stoc}}$ |
| Trimolecular | $A + B + C \xrightarrow{k} D$ | $k^{\text{det}} = (N_a V)^2 k^{\text{stoc}}$ |
| Trimolecular | $2A + B \xrightarrow{k} C$ | $k^{\text{det}} = \frac{(N_a V)^2}{2}k^{\text{stoc}}$ |

## 2.2 DETERMINISTIC MODELING AND SIMULATION

### 2.2.1 *Nonlinear ODEs*

Given an $n_r \times n_s$ system in Eq. (2.1), the state of the system at time $t$, when modeled as deterministic chemical reactions is denoted by the vector-valued variable $\mathbf{x}(t) \in \mathbb{R}^{n_s}$. The $i^{\text{th}}$ element of $\mathbf{x}(t)$, $x_i$, $i = 1, 2, \ldots, n_s$, represents the concentration of $i^{\text{th}}$ species $\mathcal{X}_i$, $i = 1, 2, \ldots, n_s$. The variable $\mathbf{r} \in \mathbb{R}^{n_r}$ denotes the vector of rate of reactions. The rate of change of the state of the system may now be written as

$$\frac{d}{dt}\mathbf{x} = \nu^T \mathbf{r} \tag{2.14}$$

with the initial conditions, $\mathbf{x}(0) = \mathbf{x}_0$. This system of nonlinear equations forms the deterministic formulation of the chemical reaction system in Eq. (2.1). These nonlinear ODEs may be simulated using a numerical ODE solver (see Section 3.3). Examples of deterministic formulations of various reaction systems are available in Chapter 4 and Section 5.3. Experiments provide us with measurements of states of the system at discrete time points $\{t_i,\ i = 1, 2, \ldots, n\}$. These measurements represented by $\{\mathbf{y}_i,\ i = 1, 2, \ldots, n\}$ may then be used to estimate unknown parameters $\theta$ (see Chapter 3).

## 2.3 STOCHASTIC MODELING AND SIMULATION

### 2.3.1 *The Chemical Master Equation*

Unlike the deterministic formulation, the stochastic formulation involves the integer-valued number of molecules instead of concentration. The state of the system at time $t$ is a vector-valued random variable denoted by $\mathbf{X}(t) \in \mathbb{R}^{n_s}$. Each element $\mathbf{X}_i(t)$ is a scalar random variable denoting the number of molecules of species $\mathcal{X}_i$. A sample of $\mathbf{X}$ is represented by $\mathbf{x}$. Instead of reaction rates, each reaction has an associated *reaction propensity*, denoted by $h_j(\mathbf{x}(t), \theta)$. The expression for reaction propensity is very similar to the the mass-action expression for reaction rate [32] and contains the stochastic rate parameters.

The stochastic chemical reaction system is governed by a Chemical Master Equation [112] which describes the time-evolution of the probability distribution of $\mathbf{X}$. There are many forms of the chemical master equation (CME). One such form is [48],

$$\frac{d}{dt}P(\mathbf{X}(t) = \mathbf{x}) = \sum_{j=1}^{n_r} h_j(\mathbf{x} - \nu_j, \theta)\ P(\mathbf{X}(t) = \mathbf{x} - \nu_j) - \sum_{j=1}^{n_r} h_j(\mathbf{x}, \theta)\ P(\mathbf{X}(t) = \mathbf{x}) \tag{2.15}$$

in which,

1. $P(\mathbf{X}(t) = \mathbf{x})$ is the probability that the system is in a state $\mathbf{x}$ at time $t$.
2. $\nu_j$, $j = 1, 2, \ldots, n_r$, is the $j^{\text{th}}$ row of the stoichiometric matrix $\nu$

Another form of the above CME is called Kolmogorov's forward equation (KFE)

which is represented as

$$\frac{d}{dt}\mathbf{P}(t) = \mathbf{Q}\mathbf{P}(t) \tag{2.16}$$

$$\mathbf{P}(0) = \mathbf{I} \tag{2.17}$$

Kolmogorov's forward equation, as written above, denotes *continuous time, discrete state space* Markov chain . The CME formulation of the stochastic reaction system is used in Chapter 4 while the KFE formulation is used in Chapter 5.

A stochastic chemical reaction system may be simulated in two ways — discrete time simulation and discrete event simulation. These are described next.

### 2.3.2 *Discrete time simulation using CME*

An obvious (but not tractable) method to simulate the time-evolution of a stochastic reaction system is to solve the above system of linear equations (*i. e.*, KFE) as follows:

$$\mathbf{P}(t) = e^{t\mathbf{Q}} \tag{2.18}$$

Clearly, we can now obtain the probability vector at time $t$ as $\mathbf{P}(t)$. Using

$$\mathbf{P}(t) = \begin{bmatrix} P(\mathbf{X}(t) = \mathbf{x}_1) \\ P(\mathbf{X}(t) = \mathbf{x}_2) \\ \vdots \\ P(\mathbf{X}(t) = \mathbf{x}_{|\mathrm{S}|}) \end{bmatrix} \tag{2.19}$$

we can simulate the random variable, $\mathbf{X}(t)$ by sampling a discrete random variable distributed as $\mathbf{P}(t)$ [119, p. 140] using a lookup method. This approach is similar to the Euler's method of solution of ODEs in which we (approximately)

compute the concentration of the system at discrete times. However, instead of concentrations, we compute the probability vector which is then used to sample the state of the system. Clearly, the linear KFE ODEs need only be solved once to generate $\mathbf{P}(t)$. But, unlike, the deterministic formulation, the state of the system in the stochastic formulation is random, *i. e.*, every time the lookup method is used a different state may be sampled. Since in this approach, the state of the system is sampled at discrete times, this method of simulation is called *discrete time simulation*.

### 2.3.3 *Discrete event simulation using SSA*

Discrete event simulation does not use the CME or KFE equations. Instead, in this approach, the system of stochastic chemical reactions is interpreted as a system of independent Poisson processes [41, p. 477] . As a result, without requiring to solve the CME/KFE, the system may be directly simulated using stochastic simulation algorithms (SSAs) [32, 33, 34]. Gillespie [32] presented two algorithms — *direct method* and *first reaction method* to simulate the stochastic reaction system. Other SSAs were developed — *next reaction method* by Gibson and Bruck, 2000 [31] and *modified next reaction method* by Anderson, 2007 [4]. All these four SSA algorithm are equivalent as they exactly simulate the same system of Poisson processes.

A SSA does not simulate the state of the system at discrete time points. Instead it simulates the *reaction events* directly. Essentially, two (random) quantities are simulated starting from an initial condition — the time at which a reaction occurs and the identity of the reaction. This is called *discrete event simulation* because the discrete reaction events are simulated.

Faster but approximate methods to simulate the reaction system have been developed, for example, $\tau$-leaping [84, 35, 17, 5], slow-scale SSA [16], and various

methods based on timescale separation [47, 48, 67, 103] .

## 2.4 CYCLICAL VS NON-CYCLICAL KINETICS

Given the framework described above, the terms *cyclical kinetics* and *non-cyclical kinetics* may be defined as follows.

**Definition 2.2** (Non-cyclical kinetics)**.** *An $n_r \times n_s$ reaction system with a stoichiometric matrix, $\nu$, is said to be non-cyclical if the transposed stoichiometric matrix , $\nu^T$ has full column rank. In other words, all $n_r$ columns of $\nu^T$ are independent.*

*A reaction kinetics is said to be cyclical if the transpose of its stoichiometric matrix, $\nu^T$ has less than full column rank. In other words, $\nu^T$ contains dependent columns.*

***Corollary**: Let $\mathbf{r} \in \mathbb{R}^{n_r}$ denote the number of times each reaction occurred during the time interval $[0, t]$. If the reaction kinetics is non-cyclical, then, given the initial state $\mathbf{x}(0)$ and final state $\mathbf{x}(t)$, $\mathbf{r}$ may be uniquely obtained by solving the linear system*

$$\mathbf{x}(1) - \mathbf{x}(0) = \nu^T \mathbf{r} \tag{2.20}$$

# 3

# PARAMETER ESTIMATION IN DETERMINISTIC MODELS

## 3.1 FORMULATION OF THE OPTIMIZATION PROBLEM

As we have seen in Chapter 2, the deterministic formulation of a chemical reaction system is a system of nonlinear ODEs

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, \theta) \tag{3.1}$$

$$\mathbf{x}(t = 0) = \mathbf{x}_0 \tag{3.2}$$

in which $\mathbf{x}(t) \in \mathbb{R}^{n_s}$ represents the state of the system at time $t$ and $\theta \in \mathbb{R}^{n_p}$. Each element of the vector represents the real-valued concentration of a species. Experimental measurements are denoted by the vector $\mathbf{y} \in \mathbb{R}^{n_{\text{meas}}}$.

### 3.1.1 *Traditional least-squares problem*

In the traditional version of the parameter estimation problem, the measurements are considered functions of the state

$$\mathbf{y}_i = \mathbf{h}(\mathbf{x}(t_i)) \quad i = 1, 2, \ldots, n \tag{3.3}$$

in which $\mathbf{y}_i \in \mathbb{R}^{n_{\text{meas}}}$. Given enough measurements, the number of (independent) equations described above equals the number of unknown parameters, $n_p$, and the parameters may be uniquely determined by solving the above set of non-linear equations. Since the measurements suffer from *experimental error*, such an approach is unreliable because it provides different sets of parameter values for different sets of measurements. Therefore, more than the required number of measurements (which involve experimental error) are obtained, which makes the above system of equations overdetermined [1]. The traditional *least-squares* approach (invented by Gauss) is used to obtain the parameter values which minimize the sum of squared errors, $\Phi(\theta)$

$$\hat{\theta} = \arg\min_{\theta} \Phi(\theta) \tag{3.4}$$

$$\Phi(\theta) = \sum_{i=1}^{n} [\mathbf{y}_i - \mathbf{h}(\mathbf{x}(t_i))]^T [\mathbf{y}_i - \mathbf{h}(\mathbf{x}(t_i))] \tag{3.5}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n_{\text{meas}}} (y_{ij} - h_j(\mathbf{x}(t_i)))^2 \tag{3.6}$$

The above minimization may be performed by a numerical optimizer (see Section 3.3).

### 3.1.2 *Measurement error as a random variable*

The measurement error is interpreted as a random variable, $\mathbf{V} \in \mathbb{R}^{n_{\text{meas}}}$. A sample of the random variable $\mathbf{V}$ is denoted by $\mathbf{v}$. The measurements are therefore modeled as random variables themselves

$$\mathbf{Y} = \mathbf{h}(\mathbf{X}) + \mathbf{V} \tag{3.7}$$

[1] It is assumed that the model structure in Eq. (3.1) is correct

and the obtained experimental measurements, $\mathbf{y}$ are interpreted as the samples of $\mathbf{Y}$. Note that the vector-valued *measurement function*, $\mathbf{h}(\cdot)$ is fixed and not considered random. The measurement errors, $V$, are assumed to be independently and identically distributed (*iid*). Usually, the measurement error, $\mathbf{Y}$ is assumed to have normal distribution and this assumption is justified using central limit theory arguments [13, p. 77-79].

$$\mathbf{V} \sim \mathcal{N}(\mu,\ \sigma^2\mathbf{I}) \tag{3.8}$$

in which the mean $\mu$ and variance $\sigma^2$ may be known or unknown depending upon how well characterized the experimental error is. Usually, the mean is assumed to be $\mu = \mathbf{0}$. Since $V$ are *iid*, $Y$ are also *iid*. If $v$ is normal, then $Y$ are also normals [41].

In this framework, the parameters $\theta$ may be obtained in a *maximum likelihood framework*. The likelihood of the data $\mathbf{y}$ is defined as

$$\begin{aligned}
\pi(\mathbf{Y} = \mathbf{y} \mid \Theta = \theta) &= \pi(\mathbf{y} \mid \theta) \\
&= \pi(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \mid \theta) \\
&= \prod_{i=1}^{n} \pi(\mathbf{y}_i \mid \theta) \\
&= \prod_{i=1}^{n} \pi(\mathbf{V}_i = \mathbf{y}_i - \mathbf{h}(\mathbf{x}(t_i)) \mid \theta)
\end{aligned} \tag{3.9}$$

The maximum likelihood estimation (MLE) problem then becomes

$$\hat{\theta} = \arg\max_{\theta} \pi(\mathbf{y} \mid \theta) \tag{3.10}$$

which is equivalent to the following minimization problem

$$\hat{\theta} = \arg\min_{\theta} \ -log(\pi(\mathbf{y} \mid \theta)) \tag{3.11}$$

in which $-log(\pi(\mathbf{y} \mid \theta))$ is the *negative log-likelihood* . Using the definition of the multivariate normal, the above minimization problem may be re-written as

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta} \sum_{i=1}^{n} [\mathbf{y}_i - \mathbf{h}(\mathbf{x}(t_i))]^T [\mathbf{y}_i - \mathbf{h}(\mathbf{x}(t_i))] \\ &= \arg\min_{\theta} \Phi(\theta) \end{aligned} \tag{3.12}$$

Using the definition of the least-squared error, we can see that the MLE problem is equivalent to the least-squared problem (under the assumptions mentioned).

## 3.2 GRADIENTS, HESSIAN AND SENSITIVITY EQUATIONS

The nonlinear minimization problem in Eq. (3.12) may be solved with less computation if the gradients of the objective function, $\Phi(\theta)$ with respect to $\theta$ are available. Before we can obtain the gradients we need to obtain the *sensitivity* of the state with respect to the parameters

$$S_{j,k}(t_i) = \frac{\partial x_j(t_i, \theta)}{\partial \theta_k} \tag{3.13}$$
$$j = 1, 2, \ldots, n_s, \quad k = 1, 2, \ldots, n_p, \quad i = 1, 2, \ldots, n$$

The elements $S_{j,k}(t_i)$, $j = 1, 2, \ldots, n_s$, $k = 1, 2, \ldots, n_p$, form the sensitivity matrix $\mathbf{S}_i$. The gradient of the objective function, $\bigtriangledown\Phi$ may now be defined as

$$\bigtriangledown\Phi = -2\sum_{i=1}^{n} \mathbf{S}_i^T \frac{\partial \mathbf{h}(\mathbf{x}(t_i))}{\partial \mathbf{x}(t_i)}^T (\mathbf{y}_i - \mathbf{h}(\mathbf{x}(t_i))) \tag{3.14}$$

Using a *Gauss-Newton* approximation [85, p. 541], the approximate Hessian may be obtained as

$$\mathbf{H} = 2\sum_{i=1}^{n} \mathbf{S}_i^T \frac{\partial \mathbf{h}(\mathbf{x}(t_i))}{\partial \mathbf{x}(t_i)}^T \frac{\partial \mathbf{h}(\mathbf{x}(t_i))}{\partial \mathbf{x}(t_i)^T} \mathbf{S} \tag{3.15}$$

In order to obtain the sensitivities at time $t_i$, we need to solve the the following *augmented* system of ODEs [85, p. 583]

$$\frac{d}{dt}\begin{bmatrix}\mathbf{x}(t)\\ \mathbf{S}(t)\end{bmatrix} = \begin{bmatrix}\mathbf{f}(\mathbf{x}(t),\theta)\\ \bigtriangledown_{\mathbf{x}}\mathbf{f}\,\mathbf{S}(t) + \bigtriangledown_{\theta}\mathbf{f}\end{bmatrix} \tag{3.16}$$

$$\mathbf{x}(0) = \mathbf{x}_0 \tag{3.17}$$

$$\mathbf{S}(0) = \bigtriangledown_{\theta}\mathbf{x}_0 \tag{3.18}$$

in which, $\bigtriangledown_{\mathbf{x}}\mathbf{f}$ and $\bigtriangledown_{\theta}\mathbf{f}$ represent the gradients of $\mathbf{f}(\mathbf{x},\theta)$ with respect to $\mathbf{x}$ and $\theta$ respectively.

## 3.3 SOFTWARE TOOLS

The minimization problem in Eq. (3.12) is solved by using a nonlinear optimizer which in turn calls the an ODE solver to solve the system of ODEs in Eq. (3.16). The entire parameter estimation procedure may be performed using the software `parest` , developed in the Rawlings research group.

### *ODE Solvers*

The `cvodes` ODE solver in the SUNDIALS [54] software package is versatile and was found to perform very well. It also allows for the solution of the augmented system of ODEs Eq. (3.16). Other notable ODE solvers include the legacy `lsode` [78, 53], which though not as versatile as was found to be extremely robust.

### *Automatic differentiation*

Note that the solution of the augmented system of ODEs in Eq. (3.16) requires the computation of gradients $\bigtriangledown_{\mathbf{x}}\mathbf{f}$ and $\bigtriangledown_{\theta}\mathbf{f}$. While it is possible to do these numerically, it is much better to perform the differentiation analytically and provide the

relevant expressions to the ODE solver. The SUNDIALS solver, `cvodes` allows the user to provide these expressions. During the course of research, many candidate models are explored before a final model is selected. This iterative procedure of model selection would require the repeated computation of $\bigtriangledown_{\mathbf{x}}\mathbf{f}$ and $\bigtriangledown_{\theta}\mathbf{f}$ by hand. Not only is this process tedious, it is also subject to human error. Automatic differentiation methods [15, 8, 1] provide a convenient alternative. These methods are basically softwares that automatically compute the required gradients analytically using only the function $\mathbf{f}(\cdot,\cdot)$. In particular the ADOL-C package [113] was found to be very convenient as it allowed the combination with `cvodes`.

*Optimizer*

Nonlinear optimizers, especially, sequential quadratic programming (SQP) [77] based methods were found to work well. Octave [78] provides one such implementation called `sqp()`.

*Issues with parameter estimation*

These issues are frequently encountered during the parameter estimation as described above. Firstly, the optimizer requires an initial guess of the parameter values $\theta_0$. In many cases, it is required that the initial guess be close enough to a local minima of the optimization problem in Eq. (3.12) to allow the optimizer to converge to the local minima. Estimation of an initial guess usually require trial and error. Second, during the course of the optimization, a solution to the ODE is required at various mathematically feasible but physically unreasonable parameter values. Such parameter values are generally associated with a stiff ODE system causing the ODE to fail. This in turn stops the optimization procedure.


Figure 3.1: Diagram of Vesicular Stomatitis virus (VSV) genome

## 3.4 EXAMPLES

In this section, I describe two relevant examples that I worked on during my doctoral research.

### 3.4.1 *Intracellular VSV mRNA kinetics*

Vesicular Stomatitis virus (VSV) is a negative strand RNA virus of the Rhabdoviridae family. VSV's 11-kilobase negative-sense RNA genome encodes 5 genes: the nucleocapsid (N), polymerase protein (P), glycoprotein (G), matrix protein (M) and large protein (L). VSV is a weak human pathogen that does not undergo genetic reassortment (like Influenza A) or recombination and does not integrate into host cell DNA. Nevertheless, it evokes a strong innate immune response from competent cells, allows incorporation of large foreign genes and grows in high titers in laboratory settings. These features make VSV a preferred choice for experimental studies. Figure 3.1 shows a diagram of VSV genome.

*VSV Biology*

According to current literature, VSV transcription follows a *start-and-stop mechanism* [28]. Transcription begins at the 3′ end of the genome when the polymerase (L-P3) first synthesizes a small 47-nucleotide leader sequence. Subsequently, the polymerase pauses at every gene junction for 1-2 minutes and re-initiates mRNA synthesis on the downstream gene with a 70-80% probability. This mechanism causes *transcriptional attenuation* such that N mRNA is present in the most amount

followed by P mRNA until the least abundant L mRNA. It is known that once sufficient amount of N and P proteins have accumulated in the cell, the RNA synthesis switches from transcription mode to replication mode. Literature provides more than one hypothesis to explain this behavior. One of most common hypothesis is that encapsidation of the newly formed leader RNA by N proteins sends an antitermination signal to the polymerase which then begins replication of the genome length RNA to produce positive-sense genomes (or anti-genomes).

*Experimental data and modeling*

A possible modeling approach based on the above VSV biology the following reactions may be written down.
Transcription:

$$
\begin{aligned}
RNP + Pol &\xrightarrow{k_1^*} RNP - Pol_N \\
RNP - Pol_N &\xrightarrow{k} N_{RNA} + RNP - Pol_P \\
RNP - Pol_P &\xrightarrow{k} P_{RNA} + RNP - Pol_M \\
RNP - Pol_M &\xrightarrow{k} M_{RNA} + RNP - Pol_G \\
RNP - Pol_G &\xrightarrow{k} G_{RNA} + RNP - Pol_L \\
RNP - Pol_L &\xrightarrow{k} L_{RNA} + n_1 RNP + n_2 Pol
\end{aligned}
$$

Detachment:

$$RNP - Pol_N \xrightarrow{k_d} RNP + Pol$$
$$RNP - Pol_P \xrightarrow{k_d} RNP + Pol$$
$$RNP - Pol_M \xrightarrow{k_d} RNP + Pol$$
$$RNP - Pol_G \xrightarrow{k_d} RNP + Pol$$
$$RNP - Pol_L \xrightarrow{k_d} RNP + Pol$$

Transcription starts when polymerase (Pol) attaches to ribonucleoprotein (RNP, which is genome encapsidated with N proteins) to create a ribonucleoprotein-polymerase complex at the first gene N. This RNP-Pol complex then synthesizes N mRNA and moves on to RNP-Pol complex at P gene. Due to transcriptional attenuation the RNP-Pol complex may disintegrate into the ribonucleoprotein and polymerase. At this point, the polymerase has to start at the first gene N again in order to synthesize P, M, G, L mRNA.

The experimental data was obtained by a member of the Yin lab as the measurements of VSV mRNA and genomes in terms of numbers of molecules per cell. While the model above was explored, a simpler model was developed by Timm et al., 2013 [108].

# 4

# MODEL REDUCTION USING STOCHASTIC REACTION EQUILIBRIUM ASSUMPTION

Chemical reaction systems usually have a characteristic complicating feature in that the different chemical reactions occur at vastly different time scales. This timescale separation results from difference in reaction rates (or propensities) that span orders of magnitude. In some cases this large difference in reaction rates may be due to a large difference in reactant amounts (number of molecules or concentrations). In other cases, the difference in reaction rates is due to a stark difference in reaction rate (or propensity) constant. Differences in both reactant amounts and reaction rate constants may produce an exacerbated difference in reaction rates. The reaction with larger rates are called *fast reactions* and the those with smaller rates are called *slow reactions*. The system of reactions may be divided into two timescales – *fast timescale* and *slow timescale* .

Timescale separation is present in both stochastic and deterministic chemical reaction systems. In the case of deterministic reactions, existence of fast and slow timescales produces a *stiff* system of ordinary differential equations (ODEs). Though modern ODE solvers have solved this problem to a large extent, there

are chemical reaction models which still pose a challenge. The issue of *stiffness* is especially encountered while estimating parameters in which the optimizer attempts to solve the ODE at an unreasonable parameter value causing the ODE solver to fail. The ODE solver failure then halts the optimizer algorithm causing an abrupt failure of the entire computation. These issues have been discussed in Chapter 3. In the case of stochastic chemical reactions, timescale separation causes *stochastic stiffness* in which a computer simulation spends most of the computation time simulating the fast reactions which advance in extremely small time steps, thus, resulting in an extremely slow simulation (see [51] for an example). Therefore, these stiff stochastic systems require prohibitively large computation effort. When estimating parameters, timescale separation presents a much more challenging problem in the case of stochastic reactions when compared to deterministic reactions. Parameter estimation in stochastic chemical reaction systems (discussed in Chapters 5 and 6) is an inherently computationally expensive procedure even without the presence of disparate timescales. Some parameter estimation methods [105, 81, 63, 110, 99] simulate the stochastic reaction system in order to estimate parameters. The existence of timescale separation causes the overall estimation problem to be magnified by a large factor of 100 or more.

While the availability of faster computers, modern ODE solvers [54, 53, 78], faster (but approximate) stochastic simulation methods [16, 117, 40, 97, 98, 67, 103, 82, 46, 68, 23, 60] has helped reduce the problems associated with timescale separation, this issue is far from being solved. Usually, we do not require the information about the fast timescale at all, thus making the computational effort spent on their simulation rather wasteful. Model reduction methods have been developed which approximate the true (or *full* model) to produce a *reduced* model that does not have a timescale separation but still retains the essential features of the original full model. The process of model reduction also provides valuable insights into

the reactions mechanisms. Two widely used model reduction techniques are the quasi steady state assumption (QSSA) and the reaction equilibrium assumption (REA) . These assumptions arise from the chemistry of the reactions and are applicable in both deterministic and stochastic reaction paradigms. The traditional, deterministic assumptions, denoted by dQSSA and dREA have been developed over a hundred years ago [11, 19, 85] and have been researched well (see Chapter 5 in Rawlings and Ekerdt, 2004 [85] and the references therein). An authoritative discussion of the QSSA may be found in Turanyi et al., 1993 [111]. Pantea et al., 2013 [79] present an alternative method to the traditional use of dQSSA, in which the low concentration species are rescaled instead of being eliminated. The stochastic versions of these assumptions, denoted by sQSSA and sREA, have been developed recently [47, 48, 67, 103, 16]. In this chapter, I present the deterministic and stochastic versions of only the reaction equilibrium assumption.

Application of the REA in the case of deterministic reactions [85, p. 207] not only provides a set of equations that describe the reduced model but also a reduced set of chemical reactions that describe the reduced model exactly. The reduced set of chemical reactions has fewer parameters than the full model, has no timescale separation (and are therefore easy to solve) and can be simulated by solving the corresponding system of ODEs (the reduced model equations) just like any other chemical kinetics. At first, the availability of reduced kinetics may seem unimportant but on a closer look, these reduced kinetics provide many desirable benefits. Firstly, reduced kinetics provide a much more direct understanding of the reduced model than a set of equations. Secondly, the traditional software tools may be used without modifications. Finally, the reaction rate constants associated with the reduced kinetics indicate the combination of the true parameters which can be easily estimated. Similar reduced set of chemical reactions may be obtained by the application of dQSSA to applicable systems [67]. Reaction equilibrium as-

sumption has been used in various ways for stochastic reactions [16, 40, 47, 48]. Haseltine and Rawlings [48] derive a set of reduced model equations which does not have a timescale separation and is computationally inexpensive to simulate. They, however, do not provide a reduced set of chemical reactions describing the reduced model. The reduced model equations contain all of the parameters of the full model and, therefore, provide neither an insight into the behavior of the reduced system nor a reduction in number of parameters to be estimated. Lastly, the reduced model equations developed by Haseltine and Rawlings [48] have a complicated structure which does not conform to the traditional stochastic simulation algorithms (SSAs) [32]. In fact, an auxiliary simulation is required to be run along the standard SSA, thus making the overall simulation of the reduced model difficult to understand and implement. The application of sQSSA, however, does provide reduced kinetics as shown by Mastny et al. [67]. Mastny et al. [67] also show that the reduced kinetics obtained by the application of dQSSA is not necessarily the same as the sQSSA-reduced kinetics . In this chapter, I investigate the existence of sREA-reduced kinetics for two examples.

This chapter is organized as follows. I begin with a typical example which admits an REA-based model reduction and demonstrate the separation of timescales in both deterministic and stochastic modeling regimes. I use the dREA procedure to obtain the reduced model and kinetics. Extending the work done by Haseltine and Rawlings [48] for this example, I show the equivalence of the reduced kinetics from both deterministic and stochastic versions of this assumption. I also present another example in which this equivalence does not hold.

## 4.1 LINEAR KINETICS: AN EXAMPLE

Consider the following typical example

$$\mathrm{B} \xrightarrow{k_1^f} \mathrm{C} \tag{4.1}$$

$$\mathrm{A} \underset{k_2^r}{\overset{k_2^f}{\rightleftharpoons}} \mathrm{B} \tag{4.2}$$

Note that the reactions above have linear reaction rates and propensities. Reaction equilibrium assumption may be used when the following condition holds

$$k_2^f, k_2^f \gg k_1^f \tag{4.3}$$

The equilibrium constant [1] denoted by $K_2$ is defined as the ratio of forward and reverse rate constants

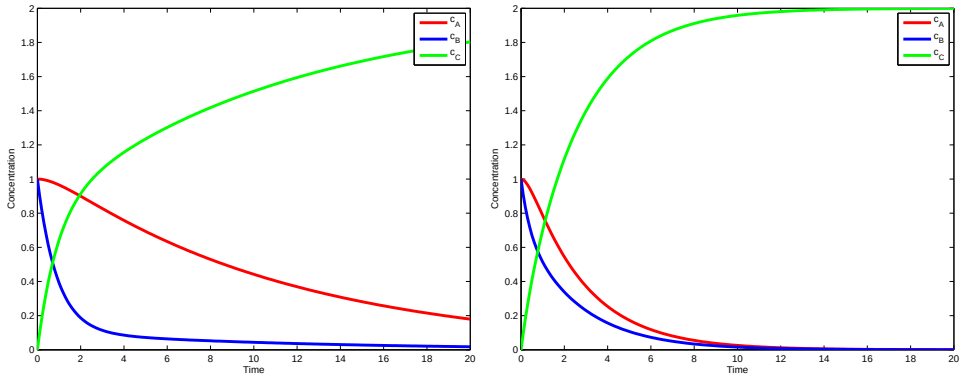
$$K_2 = \frac{k_2^f}{k_2^r} \tag{4.4}$$

### 4.1.1 *Deterministic Reaction Equilibrium Assumption*

The reactions in Eqs. (4.1)-(4.2) are described by the following deterministic formulation

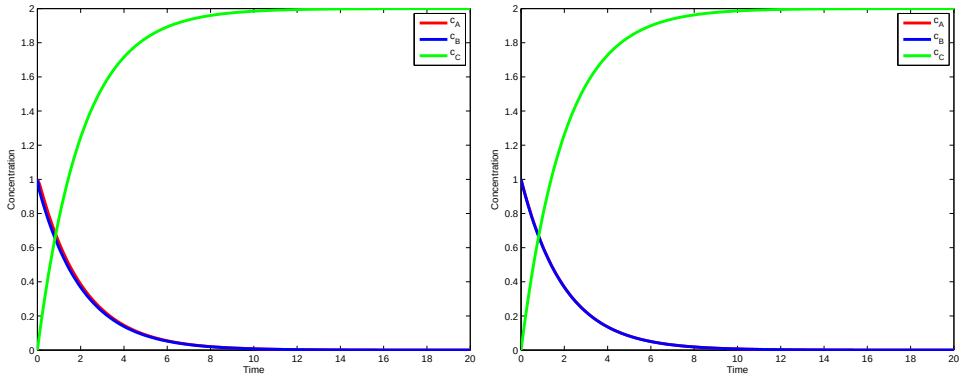
$$\begin{aligned}
\frac{d}{dt}c_A &= -k_2^f c_A + k_r^2 c_B \\
\frac{d}{dt}c_B &= k_2^f c_A - k_r^2 c_B - k_1^f c_B \\
\frac{d}{dt}c_C &= k_1^f c_B
\end{aligned} \tag{4.5}$$

[1] In this chapter, the uppercase symbol $K_2$ is used to describe the equilibrium constant instead of a random variable.




(a) $k_2^f = 0.1, k_2^r = 0.1, k_1^f = 1$

(b) $k_2^f = 1, k_2^r = 1, k_1^f = 1$




(c) $k_2^f = 10, k_2^r = 10, k_1^f = 1$

(d) $k_2^f = 100, k_2^r = 100, k_1^f = 1$

Figure 4.1: Simulation of linear kinetics using deterministic formulation for four different parameter sets showing timescale separation.

with the initial conditions $c_A(0) = c_{A0}$, $c_B(0) = c_{B0}$, $c_C(0) = c_{C0}$.

Figure 4.1 shows the simulation of the system of nonlinear ODEs in Eq. (4.5) for four different sets of parameter values. The rate constant for the first reaction, $k_1^f$ is set to 1 for all Figures 4.1a-4.1d. For the second reaction, the forward rate constant $k_2^f$ is set to equal the reverse rate constant $k_2^r$. As the value of $(k_2^f,\ k_2^r)$ is increased from 0.1 (in Figure 4.1a) to 100 (in Figure 4.1d), the concentrations of species A and B start to follow the equilibrium equation.

Under the reaction equilibrium assumption of Eq. (4.3), the kinetics can di-

vided into fast and slow time scales [85, p. 211] described by Eqns 4.6-4.7.

$$\text{Fast Time Scale } (\tau = k_2^r t): \quad \begin{aligned} \frac{dc_A}{d\tau} &= -K_2 c_A + c_B & c_A(\tau = 0) &= c_{A0} \\ \frac{dc_B}{d\tau} &= K_2 c_A - c_B & c_B(\tau = 0) &= c_{B0} \\ \frac{dc_C}{d\tau} &= 0 & c_C(\tau = 0) &= c_{C0} \end{aligned} \tag{4.6}$$

$$\text{Slow Time Scale } (t): \quad \begin{aligned} \frac{dc_A}{dt} &= -k_1^f \frac{K_2}{(1+K_2)} c_A & c_A(t = 0) &= \frac{(c_{A0}+c_{B0})}{(1+K_2)} \\ \frac{dc_B}{dt} &= -k_1^f \frac{K_2}{(1+K_2)} c_B & c_B(t = 0) &= \frac{K_2(c_{A0}+c_{B0})}{(1+K_2)} \\ \frac{dc_C}{dt} &= k_1^f \frac{K_2}{(1+K_2)} c_A + k_1^f \frac{K_2}{(1+K_2)} c_B & c_C(t = 0) &= c_{C0} \end{aligned} \tag{4.7}$$

On inspection of Eq. (4.7), we find that the slow time scale can be written as

$$\mathrm{B} \xrightarrow{k'} \mathrm{C} \tag{4.8}$$

$$\mathrm{A} \xrightarrow{k'} \mathrm{C} \tag{4.9}$$

in which, the *effective rate constant* , $k'$ is given by

$$k' = k_1^f \frac{K_2}{(1+K_2)} \tag{4.10}$$

The system of chemical reactions Eqs. (4.8)-(4.9) is called the *dREA-reduced kinetics*.

### 4.1.2 *Stochastic Reaction Equilibrium Assumption*

A stochastic description of reactions 4.1-4.2 is required when the number of reacting species is low (1-100). Instead of concentrations, the system is described in terms of probability of being in a particular state. The time evolution of these

probabilities is described by a Chemical Master Equation [34]. The chemical master equation in terms of reaction extents can be written as [48],

$$\frac{dP(\mathbf{x};t)}{dt} = \sum_{k=1}^{m} a_k^f(\mathbf{x}-\mathbf{I}_k)P(\mathbf{x}-\mathbf{I}_k;t) + a_k^r(\mathbf{x}+\mathbf{I}_k)P(\mathbf{x}+\mathbf{I}_k;t) - (a_k^f(\mathbf{x})+a_k^r(\mathbf{x}))P(\mathbf{x};t) \tag{4.11}$$

For a reaction network with $p$ species and $m$ reactions, $\mathbf{x} \in \mathbb{I}^m$ is the vector of reaction extents described by

$$\mathbf{n} = \mathbf{n}_0 + \nu^T \mathbf{x} \tag{4.12}$$

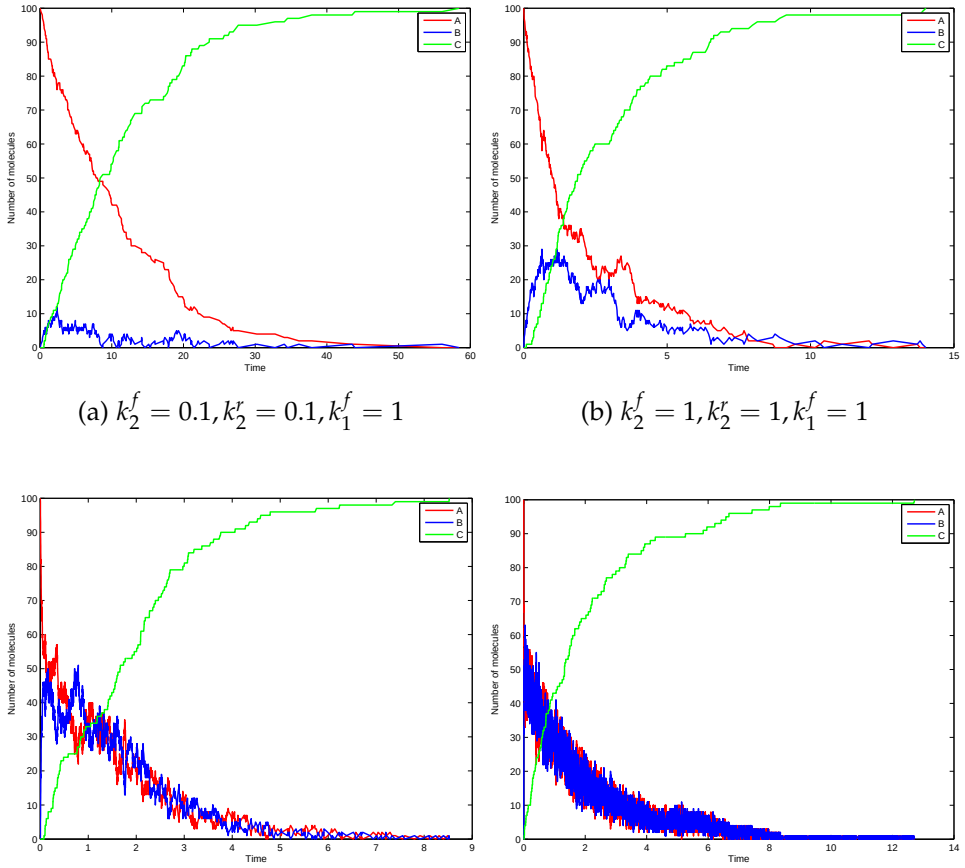
in which, $\mathbf{n} \in \mathbb{I}_{\geq 0}^p$ represents the number of molecules of each species, $\mathbf{n}_0$ is the initial number of molecules and $\nu \in \mathbb{I}_{\geq 0}^{m\times p}$ is the stoichiometric matrix. $P(\mathbf{x};t)$ is the probability that the system is in state $\mathbf{x}$ at time $t$. In the particular case of reactions 4.1-4.2, we have $m = 2$, $p = 3$ with the following initial condition:

$$a(t=0) = a_0,\ b(t=0) = b_0,\ c(t=0) = c_0$$
$$x_1(t=0) = 0,\ x_2(t=0) = 0$$

in which, $x_1$, $x_2$ are the extents of reactions 4.1 and 4.2 respectively, such that $x_2 \in [-b_0, a_0]$, $x_1 \in [0, x_2 + b_0]$. The chemical master equation becomes,

$$\begin{aligned}\frac{dP(x_1,x_2;t)}{dt} =& k_1^f(b_0 - x_1 + x_2)P(x_1-1,x_2) - k_1^f(b_0 - x_1 + x_2)P(x_1,x_2) \\ &+ k_2^f(a_0 - x_2)P(x_1,x_2-1) + k_2^r(b_0 - x_1 + x_2 + 1)P(x_1,x_2+1) \\ &- \big(k_2^f(a_0 - x_2) + k_2^r(b_0 - x_1 + x_2)\big)P(x_1,x_2)\end{aligned} \tag{4.13}$$

Figure 4.2 shows a sample simulation of the reaction system for the same four sets of parameter values as in Figure 4.1. The rate constant for the first reaction,




(a) $k_2^f = 0.1, k_2^r = 0.1, k_1^f = 1$

(b) $k_2^f = 1, k_2^r = 1, k_1^f = 1$

(c) $k_2^f = 10, k_2^r = 10, k_1^f = 1$

(d) $k_2^f = 100, k_2^r = 100, k_1^f = 1$

Figure 4.2: Simulation of linear kinetics using stochastic formulation for four different parameter sets showing timescale separation.

$k_1^f$ is set to 1 for all Figures 4.2a-4.2d. For the second reaction, the forward rate constant $k_2^f$ is set to equal the reverse rate constant $k_2^r$. As the value of $(k_2^f,\ k_2^r)$ is increased from 0.1 (in Figure 4.2a) to 100 (in Figure 4.2d), the concentrations of species A and B start to follow the equilibrium probability distribution given approximately in Eq. (4.15).

When reaction 4.2 is much faster than reaction 4.1, Eq. (4.13) may be approximated [48] such that the reduced model is represented by a *reduced evolution*

equation and an *algebraic constraint*. The *reduced evolution* equation is given by

$$\frac{dP(x_1;t)}{dt} = \sum_{x_2=x_1-b_0}^{x_2=a_0} \big[k_1^f(b_0 - x_1 + 1 + x_2)P_A(x_2 \mid x_1 - 1)P(x_1 - 1) \\ - k_1^f(b_0 - x_1 + x_2)P_A(x_2 \mid x_1)P(x_1)\big] \tag{4.14}$$

in which, $P_A(x_2 \mid x_1)$ is the approximate conditional probability described by the following *algebraic constraint*

$$k_2^f(a_0 - x_2 + 1)P_A(x_2 - 1 \mid x_1) + k_2^r(b_0 - x_1 + x_2 + 1)P_A(x_2 + 1 \mid x_1) \\ - \big(k_2^f(a_0 - x_2) + k_2^r(b_0 - x_1 + x_2)\big)P_A(x_2 \mid x_1) = 0 \tag{4.15}$$

The first three moments of $P_A(x_2 \mid x_1)$ may be obtained as

$$\sum_{x_2=x_1-b_0}^{x_2=a_0} P_A(x_2 \mid x_1) = 1 \tag{4.16}$$

$$\sum_{x_2=x_1-b_0}^{x_2=a_0} x_2 P_A(x_2 \mid x_1) =< x_2 \mid x_1 > \tag{4.17}$$

$$\sum_{x_2=x_1-b_0}^{x_2=a_0} x_2^2 P_A(x_2 \mid x_1) =< x_2^2 \mid x_1 > \tag{4.18}$$

Substituting Eqs. (4.16), (4.17), (4.18) in Eq. (4.14), we obtain

$$\frac{dP(x_1;t)}{dt} = k_1^f P(x_1 - 1)\big(\ (b_0 - x_1 + 1) + < x_2 \mid x_1 - 1 >\ \big) \\ - k_1^f P(x_1)\big(\ (b_0 - x_1) + < x_2 \mid x_1 >\ \big) \tag{4.19}$$

In order to eliminate $< x_2 \mid x_1 >$ and $< x_2 \mid x_1 - 1 >$, we compute the first moment of Eq. (4.15):

$$k_2^f a_0 + k_2^r(x_1 - b_0) - \big(k_2^f + k_2^f\big) < x_2 \mid x_1 >= 0 \tag{4.20}$$

Substituting Eq. (4.20) in Eq. (4.19), we obtain the following *reduced evolution* equation in terms of only the slow extent ($x_1$),

$$\frac{dP(x_1)}{dt} = \frac{k_1^f K_2}{(1+K_2)}\big(a_0 + b_0 - (x_1 - 1)\big)P(x_1 - 1) - \frac{k_1^f K_2}{(1+K_2)}\big(a_0 + b_0 - x_1\big)P(x_1) \tag{4.21}$$

Using the definition of $k'$, the above equation may be re-written as

$$\frac{dP(x_1)}{dt} = k'\big(a_0 + b_0 - (x_1 - 1)\big)P(x_1 - 1) - k'\big(a_0 + b_0 - x_1\big)P(x_1) \tag{4.22}$$

In the next section, I demonstrate an equivalence between dREA-reduced kinetics and sREA-reduction in general.

### 4.1.3 *Equivalence of the deterministic and stochastic reductions*

First, consider a simple case with the initial conditions $a_0 = b_0 = c_0 = 1$. Using the reduced evolution equation in Eq. (4.22) we can obtain

$$\begin{aligned} P(x_1 = 0) &= P(a = 2, b = 0, c = 1) + P(a = 1, b = 1, c = 1) + P(a = 0, b = 2, c = 1) \\ &= e^{-2t} \end{aligned} \tag{4.23}$$

$$\begin{aligned} P(x_1 = 1) &= P(a = 1, b = 0, c = 2) + P(a = 0, b = 1, c = 2) \\ &= 2e^{-t} - 2e^{-2t} \end{aligned} \tag{4.24}$$

$$\begin{aligned} P(x_1 = 2) &= P(a = 0, b = 0, c = 3) \\ &= e^{-2t} - 2e^{-t} + 1 \end{aligned} \tag{4.25}$$

I, now, compare the above result with the dREA-reduced reaction kinetics in Eqs. (4.8)-(4.9). For the same initial conditions, $a_0 = b_0 = c_0 = 1$, the stochastic formulation of the dREA-reduced reaction kinetics in Eqs. (4.8)-(4.9) yields the

following probabilities

$$
\begin{aligned}
&P(a=1,b=1,c=1) = e^{-2t} \\
&P(a=1,b=0,c=2) = P(a=0,b=1,c=2) = e^{-t} - e^{-2t} \\
&P(a=0,b=0,c=3) = e^{-2t} - 2e^{-t} + 1
\end{aligned} \tag{4.26}
$$

These equations in Eq. (4.26) predict exactly the same probabilities as the stochastically reduced equations 4.25. This indicates that in this case a stochastic reduction yields the same reduced reaction set 4.8-4.9 as the deterministic one. This may not always be the case, as shown by Mastny et al. [67] for a QSSA reduction. It may be noted that Eq. (4.22) does not give any information about the fast extent, *i.e.*, it tells us the probability $P(a+b)$ but not $P(a)$ or $P(b)$ individually. Such a reduction may be represented as

$$
Z \xrightarrow{k'} C \tag{4.27}
$$

in which Z is a species representing both A and B with the initial condition $z(t=0) = z_0 = a_0 + b_0$.

Next, I show that, in general, the dREA-reduced kinetics in Eqs. (4.8)-(4.9), when modeled as a stochastic reaction system, is equivalent to the sREA reduction described by Eqs. (4.14)-(4.15). The stochastic formulation of the dREA-reduced kinetics in Eqs. (4.8)-(4.9) is given as

$$
\begin{aligned}
&\frac{d}{dt}P(A=a,B=b,C=c;t) = \\
&\quad k'(a+1)P(A=a+1,B=b,C=c-1;t) - k'(a)P(A=a,B=b,C=c;t) \\
&\quad k'(b+1)P(A=a,B=b+1,C=c-1;t) - k'(b)P(A=a,B=b,C=c;t).
\end{aligned} \tag{4.28}
$$

Converting $(A, B, C)$ (population space of reduced model) to $(X_1, X_2)$ (extent space of full model)

$$\begin{aligned}\frac{d}{dt}P(X_1 = x_1, X_2 = x_2; t) = \\ & k'(A_0 - x_2 + 1)P(X_1 = x_1 - 1, X_2 = x_2 - 1; t) \\ & + k'(B_0 - x_1 + x_2 + 1)P(X_1 = x_1 - 1, X_2 = x_2; t) \\ & - (A_0 + B_0 - x_1)P(X_1 = x_1, X_2 = x_2; t)\end{aligned}$$

Note the the extents $(X_1, X_2)$ are that of the full model in Eqs. (4.1)-(4.2) and not of the reduced kinetics in Eqs. (4.8)-(4.9). Summing over the slow extent, $X_2$, to eliminate it

$$\begin{aligned}&\frac{d}{dt}\sum_{X_2} P(X_1 = x_1, X_2 = x_2; t) = \\ &\sum_{X_2} k'(A_0 - x_2 + 1)P(X_1 = x_1 - 1, X_2 = x_2 - 1; t) \\ &+ \sum_{X_2} k'(B_0 - x_1 + x_2 + 1)P(X_1 = x_1 - 1, X_2 = x_2; t) \\ &- \sum_{X_2} (A_0 + B_0 - x_1)P(X_1 = x_1, X_2 = x_2; t)\end{aligned}$$

which results in the following equation in fast extent, $X_1$,

$$\begin{aligned}\frac{d}{dt}P(X_1 = x_1; t) = k'\big(A_0 + B_0 - (x_1 - 1)\big)P(X_1 = x_1 - 1; t) \\ - k'\big(A_0 + B_0 - x_1\big)P(X_1 = x_1; t)\end{aligned} \tag{4.29}$$

Note that the above equation in Eq. (4.29) is equivalent to the reduced evolution equation Eq. (4.22). This proves, that the two reductions are equivalent. This result provides a promising indication that the same may be true for any linear kinetics in general and therefore warrants further research.

## 4.2 NONLINEAR KINETICS: COUNTER EXAMPLE

Let's consider another example which admits the use of reaction equilibrium assumption but with nonlinear reaction rates and propensities.

$$2\mathrm{B} \xrightarrow{k_1^f} \mathrm{C} \tag{4.30}$$

$$2\mathrm{A} \underset{k_2^r}{\overset{k_2^f}{\rightleftharpoons}} 2\mathrm{B} \tag{4.31}$$

As before, the REA may be used when Eq. (4.3) holds. The definition of the equilibrium constant, $K_2$, is still given by Eq. (4.4). In this section, I will show that while dREA and sREA yield the same reduced kinetics for the linear kinetics example in Eqs. (4.1)-(4.2), the same is not true for this example with nonlinear kinetics. Thus, this example serves as a counter example.

### 4.2.1 *Deterministic Reaction Equilibrium Assumption*

The reactions in Eqs. (4.30)-(4.31) are described by the following deterministic formulation

$$\begin{aligned}
\frac{d}{dt}c_A &= -2\left(k_2^f c_A^2 + k_r^2 c_B^2\right) \\
\frac{d}{dt}c_B &= -2k_1^f c_B^2 + 2\left(k_2^f c_A^2 + k_r^2 c_B^2\right) \\
\frac{d}{dt}c_C &= k_1^f c_B^2
\end{aligned} \tag{4.32}$$

with the initial conditions $c_A(0) = c_{A0}$, $c_B(0) = c_{B0}$, $c_C(0) = c_{C0}$. Note that the mass balance hold according to the following equation

$$\frac{d}{dt}(c_A + c_B + 2c_C) = 0 \tag{4.33}$$

Under the reaction equilibrium assumption of Eq. (4.3), the following *slow time scale* equation may be derived

$$\begin{aligned}
\frac{d}{dt}c_A &= -2\frac{k_1^f K_2}{1+\sqrt{K_2}}c_A^2 \\
\frac{d}{dt}c_B &= -2\frac{k_1^f \sqrt{K_2}}{1+\sqrt{K_2}}c_B^2 \\
\frac{d}{dt}c_C &= \frac{k_1^f K_2}{1+\sqrt{K_2}}c_A^2 + \frac{k_1^f \sqrt{K_2}}{1+\sqrt{K_2}}c_B^2
\end{aligned} \tag{4.34}$$

with the *adjusted initial conditions* for the slow time scale

$$\begin{aligned}
c_A(t=0) &= \frac{1}{(1+\sqrt{K_2})}(c_{A0}+c_{B0}) \\
c_B(t=0) &= \frac{\sqrt{K_2}}{(1+\sqrt{K_2})}(c_{A0}+c_{B0}) \\
c_C(t=0) &= c_{C0}
\end{aligned} \tag{4.35}$$

In terms of the *effective rate constants*, $k_1'$ and $k_2'$,

$$k_1' = \frac{k_1^f K_2}{1+\sqrt{K_2}} \tag{4.36}$$

$$k_2' = \frac{k_1^f \sqrt{K_2}}{1+\sqrt{K_2}} \tag{4.37}$$

the nonlinear ODEs for slow time scale system may be re-written as

$$\begin{aligned}
\frac{d}{dt}c_A &= -2k_1' c_A^2 \\
\frac{d}{dt}c_B &= -2k_2' c_B^2 \\
\frac{d}{dt}c_C &= k_1' c_A^2 + k_2' c_B^2
\end{aligned} \tag{4.38}$$

The above system of ODEs also represent the following system of reactions

$$2\mathrm{B} \xrightarrow{k_1'} \mathrm{C} \tag{4.39}$$

$$2\mathrm{A} \xrightarrow{k_2'} \mathrm{C} \tag{4.40}$$

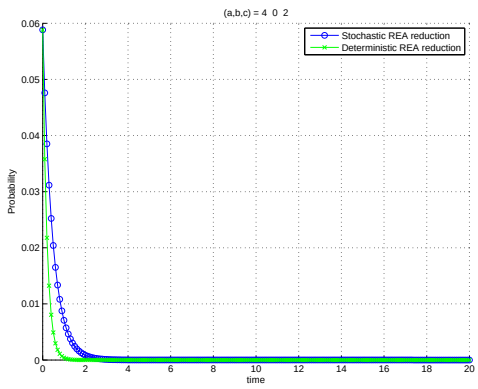
which represents the *dREA-reduced kinetics*.

### 4.2.2 *Stochastic Reaction Equilibrium Assumption*

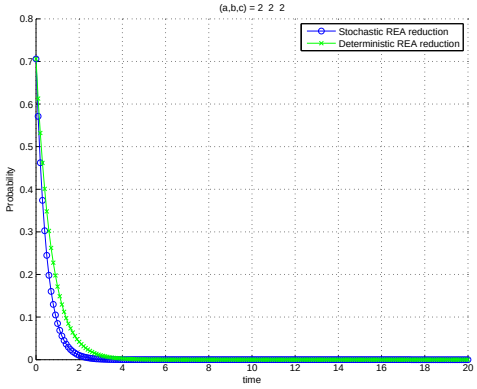
The evolution equation and the algebraic constraint describing the sREA-reduced system of equations may be derived using the same procedure as described in Section 4.1 and Haseltine and Rawlings [48]. The derivation is skipped for brevity.

### 4.2.3 *Deterministic and stochastic reductions mismatch*

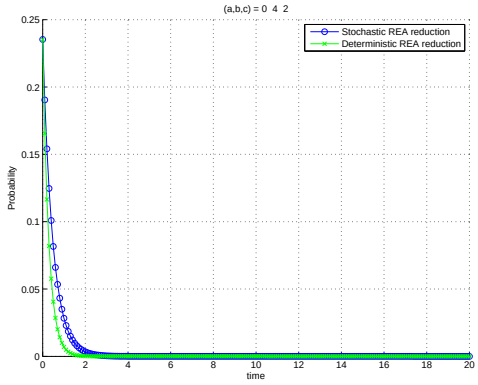
Before an attempt is made to reconcile the dREA and sREA reductions, it is informative to simulate the two reductions for a specific initial condition, $(a_0,\ b_0,\ c_0) = (2,\ 2,\ 2)$. Figure 4.3 shows the comparison of the two reductions for this example. Clearly, the two reductions produce different probabilities, which implies that dREA-reduced kinetics is not equivalent to the sREA reduction.



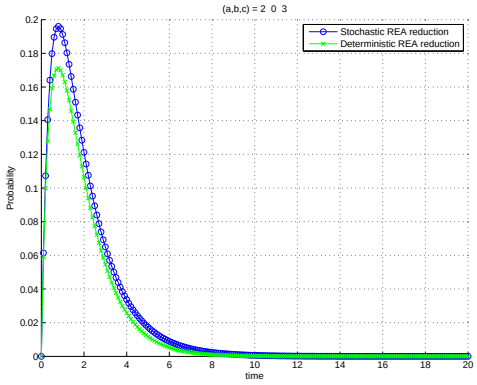

(a) Probability of state $P(4,0,2;t)$



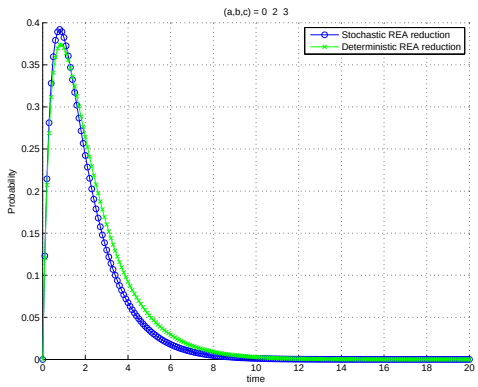

(b) Probability of state $P(2,2,2;t)$



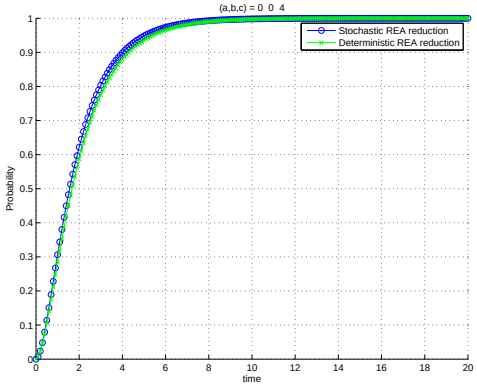

(c) Probability of state $P(0,4,2;t)$




(d) Probability of state $P(2,0,3;t)$




(e) Probability of state $P(0,2,3;t)$




(f) Probability of state $P(0,0,4;t)$

Figure 4.3: Comparison of dREA and sREA reductions of Eqs. (4.30)-(4.31). Initial Condition: $(a_0, b_0, c_0) = (2,2,2)$.

# 5

# OVERVIEW: PARAMETER ESTIMATION IN STOCHASTIC CHEMICAL KINETIC MODELS

The [1] acceptance of stochastic reaction kinetics as a standard approach to explain intrinsic noise and the advances in experimental measurement techniques, especially fluorescence microscopy, led to the inverse problem of estimating parameters from data. Golding et al. [36] used fluorescence microscopy to measure the number of mRNA molecules in a single cell at a sampling rate of 2 measurements per minute. They showed that the average number of mRNA transcripts per cell was between 0 and 10 and proposed a stochastic gene expression model to explain the observed behavior. Initial attempts to estimate model parameters (usually the reaction rate constants) were rather ad hoc – the parameters were set to biologically plausible values and then tuned by eye so that the model simulation resembled the experimental data [6]. Another approach is to use the continuum assumption and treat the reaction kinetics as deterministic instead of stochastic. In this case, parameter estimates can be easily obtained by least-squares fitting or maximum likelihood estimation. Tian et al. [105] demonstrated that this approach does not produce good parameter estimates when the number of molecules is

[1] Part of this paragraph appears in Gupta and Rawlings, 2013 [43]

small. Later approaches maintained the stochasticity of the model and utilized various statistical Monte Carlo methods to estimate parameters. These estimation methods may be classified into two familiar categories – maximum likelihood methods and Bayesian inference methods. Maximum likelihood methods include simulated maximum likelihood (SML) [105], density function distance (DFD) [81], approximate maximum likelihood [81, 87] and accelerated maximum likelihood methods [22]. methods. All of these methods maximize an approximation of the true likelihood to obtain parameter estimates. Stochastic gradient descent (SGD), proposed by Wang et al. [114], estimates the gradients of the likelihood function with respect to the parameters using a reversible jump Markov chain Monte Carlo (RJMCMC) method. By contrast, Bayesian inference methods attempt to obtain the posterior distribution of the parameters, which can then be maximized to obtain maximum *a posteriori* (MAP) estimates. Most Bayesian inference methods rely on Markov chain Monte Carlo (MCMC) techniques. Rempala et al. [88] utilize MCMC-enabled Bayesian inference to estimate parameters in a specific model of gene transcription. Wilkinson and others [50, 38, 37, 39, 119] have developed a range of Bayesian inference algorithms, all of which employ MCMC methods. In many cases, they replace the stochastic chemical kinetic model with another statistical model [50] or approximate it using the diffusion approximation [38, 37, 39]. In other cases, they use the true stochastic model [14, 119]. Similar to the MCMC method of Wilkinson [119], Choi and Rempala [21] proposed a Bayesian inference algorithm based on Gibbs sampling. Approximate Bayesian computation (ABC) [63, 110, 99, 120] is another class of Bayesian methods in which the parameters are estimated by measuring the "distance" between the experimental data and SSA simulations for a candidate parameter value and then accepting or rejecting based on a tolerance. A rather new class of parameter estimation methods is Bayesian regression using polynomial chaos representation [3].

All of these estimation methods suffer from their own particular shortcomings

and have restricted applicability. These issues are discussed in this chapter.

Many of the examples in the references above come from systems biology. Starting with the Prokaryotic gene-autoregulation [6, 114, 39], gene transcription models [87, 88, 88], RNA expression [36, 81, 87] and single cell mtDNA dynamics [50]. Other examples include the Lotka-Volterra model [14, 39, 119] and epidemic models [21]. These models consist of 2–8 reactions involving 2-5 species.

Throughout this thesis, I use the following *gamma prior*

$$\begin{aligned} \pi(\theta) &= \prod_{i=1}^{n_r} \pi(k_i) \\ \pi(k_i) &= Ga(a_i, b_i) = \frac{b_i^{a_i}}{\Gamma(a_i)} k_i^{a_i - 1} e^{-b_i k_i} \\ K_i &\sim Ga(a_i, b_i) \quad i = 1, 2, \ldots, n_r \end{aligned} \tag{5.1}$$

in which $a_i$ and $b_i$ are the shape and rate parameters of the corresponding gamma distribution. The motivation for this choice is algebraic convenience resulting from the prior-likelihood conjugacy. Details are provided in Section 5.1. Analysis of results throughout this chapter (especially in Section 5.2) and the next chapter show that the choice of gamma prior does not affect the posterior and parameter estimates significantly.

I use the following system of reactions as a common example [2] throughout this chapter and the next chapter.

$$\mathrm{A} \xrightarrow{k_1} \mathrm{B} \tag{5.2}$$

$$\mathrm{B} \xrightarrow{k_2} \mathrm{C} \tag{5.3}$$

This system of reactions is chosen because it allows for parameter estimation using

[2] This example also appears in Gupta and Rawlings, 2013 [42]

every parameter method described in this thesis and therefor allows a thorough comparison. Specific reasons, listed below, are also explained wherever relevant.

1. The system of reactions is simple to understand and simulate.
2. The reaction rates are linear in the reactant amounts (assuming that the reactions are elementary [85, p. 188]).
3. The system of reactions in not cyclical (the definition of cyclical kinetics is presented in Section 2.4).
4. Mass balance is not violated which allows the use of exact parameter estimation.
5. Two parameter system which allows the plotting of joint posterior on a three-dimensional graph.

This chapter is organized as follows. In Section 5.1, I explain the concept of complete data and provide the analytical expressions of complete-data likelihood, posterior, and marginal likelihood. Next, I restate the parameter inference procedure under the Bayesian estimation (BE) framework given complete data as described by Wilkinson, 2012 [119], Boys et al., 2008 [14]. I also present parameter inference procedure under the maximum likelihood estimation (MLE) framework. Closed-form expressions for the parameter estimates using both BE and MLE frameworks are presented and contrasted. I also discuss the conditions under which a given complete-data trajectory is informative or non-informative. This section aids in the development of approximate methods in Section 6.2 and the experimental design guidelines in Section 7.3. In Section 5.2, I explain the concept of measurement data and its relationship with complete data. Next, I present an exact method to estimate parameters using measurement data and demonstrate the limited applicability of this method. In Section 5.3, I demonstrate that the deterministic formulation of the chemical reaction system does not necessarily provide

reliable parameter estimates. Sections 5.4 and 5.5 describe two simulation-based parameter estimation methods proposed by Wilkinson, 2012 [119], Boys et al., 2008 [14] and Choi and Rempala, 2012 [21]. Another simulation-based method is described in Section 6.1. Relevant examples have been used throughout this chapter to explain and compare the parameter estimation methods. A common notational scheme is built incrementally over Sections 5.1–5.5. Chapter 6 builds upon this chapter to develop two new classes of parameter inference methods.

## 5.1 ESTIMATION USING COMPLETE DATA

Much of the notation required to describe a system of stochastic chemical reactions has been defined in Chapter 2. The required notation is restated here. Consider a system of chemical reactions with $n_r$ reactions, denoted by $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_{n_r}$ and $n_s$ species, denoted by $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_{n_s}$. This $n_r \times n_s$ system may be represented as the following system of chemical reactions

$$\begin{aligned}
\mathcal{R}_1 &: \quad \nu'_{1,1}\mathcal{X}_1 + \nu'_{1,2}\mathcal{X}_1 + \cdots + \nu'_{1,n_s}\mathcal{X}_{n_s} \quad \xrightarrow{k_1} \quad \nu''_{1,1}\mathcal{X}_1 + \nu''_{1,2}\mathcal{X}_1 + \cdots + \nu''_{1,n_s}\mathcal{X}_{n_s} \\
\mathcal{R}_2 &: \quad \nu'_{2,1}\mathcal{X}_1 + \nu'_{2,2}\mathcal{X}_1 + \cdots + \nu'_{2,n_s}\mathcal{X}_{n_s} \quad \xrightarrow{k_2} \quad \nu''_{2,1}\mathcal{X}_1 + \nu''_{2,2}\mathcal{X}_1 + \cdots + \nu''_{2,n_s}\mathcal{X}_{n_s} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\quad \vdots \\
\mathcal{R}_{n_r} &: \quad \nu'_{n_r,1}\mathcal{X}_1 + \nu'_{n_r,2}\mathcal{X}_1 + \cdots + \nu'_{n_r,n_s}\mathcal{X}_{n_s} \quad \xrightarrow{k_{n_r}} \quad \nu''_{n_r,1}\mathcal{X}_1 + \nu''_{n_r,2}\mathcal{X}_1 + \cdots + \nu''_{n_r,n_s}\mathcal{X}_{n_s}
\end{aligned}$$

The corresponding, $n_r \times n_s$ sized, stoichiometric matrix, $\nu$, may be written as

$$\nu = \begin{bmatrix} \nu_{1,1} & \nu_{1,2} & \cdots & \nu_{1,n_s} \\ \nu_{2,1} & \nu_{2,2} & \cdots & \nu_{2,n_s} \\ \vdots & \vdots & \vdots & \vdots \\ \nu_{n_r,1} & \nu_{n_r,2} & \cdots & \nu_{n_r,n_s} \end{bmatrix} \tag{5.4}$$

in which,

$$\nu_{i,j} = \nu''_{i,j} - \nu'_{i,j} \qquad i = 1, 2, \ldots, n_r,\ j = 1, 2, \ldots, n_s$$

In the stochastic formulation, each reaction $\mathcal{R}_i,\ i = 1, 2, \ldots, n_r$, has an associated rate constant (or reaction parameter or hazard rate constant) $k_i$. The set of reaction rate constants is denoted by, $\theta = \begin{bmatrix} k_1 & k_2 & \cdots & k_{n_r} \end{bmatrix}^T$.

### 5.1.1 *Complete data*

The random variable, $\mathbf{X}(t) \in \mathbb{R}^{n_s}$ denotes the state of the system at time $t$. Each component, $X_i(t)$, is the scalar random variable denoting the $i^{\text{th}}$ species. A sample of $\mathbf{X}(t)$ is denoted by $\mathbf{x}(t)$. Similarly, a sample of $X_i(t)$ is denoted by $x_i(t)$. A *complete* trajectory of the system over the time interval $t \in [0, T]$ is defined as

$$X = \{\mathbf{X}(t) : t \in [0, T]\} \tag{5.5}$$

Note that the trajectory $X$ is an infinite-dimensional collection of random vectors $\mathbf{X}(t)$. A sample of the complete trajectory is denoted by

$$x = \{\mathbf{x}(t) : t \in [0, T]\} \tag{5.6}$$

A sample trajectory is nothing but a simulation of the reaction system using one of the simulation methods presented in Chapter 2. As shown in Chapter 2, a simulated trajectory follows a *staircase* plot in which the state of the system remains the same until a reaction event occurs. Thus, given the initial state of the system and suitable information about the reaction events, it is possible to create the reaction events. Specifically, let $\mathbf{x}_0 = \mathbf{x}(0)$ be the initial state of the system. Let the random variable $N$ denote the total number of reaction events occurring in the

time interval $t \in [0, T]$. Following the notational scheme, $n$ denotes a sample of the random variable $N$. Let $R_i$, $i = 1, 2, \ldots, n_r$ be the random variable denoting the number of times reaction $\mathcal{R}_i$ occurs during the time interval $t \in [0, T]$. A sample of $R_i$ is denoted by $r_i$. Using this notation,

$$N = \sum_{i=1}^{n_r} R_i, \qquad n = \sum_{i=1}^{n_r} r_i \tag{5.7}$$

Further, let $\mathbf{R} \in \mathbb{R}^{n_r}$ be the random vector formed by the random variables $R_i$, $i = 1, 2, \ldots, n_r$. A sample of $\mathbf{R}$ is denoted by $\mathbf{r}$. Note that the elements of $\mathbf{r}$ are non-negative integers. Given $\mathbf{r}$ and $\mathbf{x}(0)$, the state of the system at time $T$ may be obtained as

$$\mathbf{x}(T) = \mathbf{x}(0) + \nu^T \mathbf{r} \tag{5.8}$$

However, given $\mathbf{x}(T)$ and $\mathbf{x}(0)$, $\mathbf{r}$ may only be recovered if the transposed stoichiometric matrix, $\nu^T$ has full column rank.

Let $N_j \in \{1, 2, \ldots, n_r\}, j = 1, 2, \ldots, n$ be the index of the $j^{\text{th}}$ reaction event. The corresponding sample of $N_j$ is denoted by $n_j$. Let $T_j \in [0, T], j = 1, 2, \ldots, n$ be the random time at which $j^{\text{th}}$ reaction occurs. Note that $T_i \leq T_j$, if $i \leq j$. A sample of $T_j$ is denoted by $t_j$. As suggested by Wilkinson, 2012 [119], I define $t_0 = 0$ and $t_{n+1} = T$ for notational convenience.

Given $\mathbf{x}_0$, $n$, $(n_j, t_j)$, $j = 1, 2, \ldots, n$, it is possible to construct the corresponding complete trajectory $x$ over the time interval $t \in [0, t_n]$ as follows

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{x}(t_{j-1}) + (\nu^T)_{n_j} \quad t \in [t_{j-1}, t_j),\ j = 1, 2, \ldots, n \\ \mathbf{x}(t_n) &= \mathbf{x}(t_{n-1}) + (\nu^T)_{n_n} \end{aligned} \tag{5.9}$$

in which, $(\nu^T)_{n_j}$ is the $n_j{}^{\text{th}}$ column of the $n_s \times n_r$ matrix $\nu^T$. If additional information is given that no reaction event occurred in the time interval $[t_n, T]$, then we

can define $x$ over $t \in [0, T]$ as follows

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{x}(t_{j-1}) + (\nu^T)_{n_j} \quad t \in [t_{j-1}, t_j),\ j = 1, 2, \ldots, n+1 \\ \mathbf{x}(t_{n+1}) &= \mathbf{x}(t_n) \end{aligned} \tag{5.10}$$

Similarly, given the complete trajectory $x = \{\mathbf{x}(t) : t \in [0, T]\}$, it is possible to extract the information $\mathbf{x}_0,\ n,\ (n_j, t_j),\ j = 1, 2, \ldots, n$ over the time interval $[0, T]$. This proves the following equivalence between complete trajectory and complete reaction event data ,

$$\begin{aligned} X = \{\mathbf{X}(t) : t \in [0, T]\} \equiv \\ \mathbf{X}_0,\ n,\ (N_j, T_j),\ j = 1, 2, \ldots, n \text{ and no reaction occurred in } (t_n, T] \end{aligned} \tag{5.11}$$

The benefit of this equivalence is that instead of attempting the impossible task of storing the infinite dimensional complete trajectory $x$ on a computer, we can store the complete reaction event data.

Each reaction $\mathcal{R}_i,\ i = 1, 2, \ldots, n_r$ has an associated reaction propensity (or hazard rate) denoted by $h_i(\mathbf{X}(t), \theta)$. The total reaction propensity (or combined hazard) is defined as

$$h_0(\mathbf{x}(t), \theta) = \sum_{i=1}^{n_r} h_i(\mathbf{X}(t), \theta) \tag{5.12}$$

Note that the reaction propensities are merely functions of the state $\mathbf{X}(t)$ and the model parameters $\theta$ and therefore are random variables as well.

### 5.1.2 *Complete-data distributions*

In this section, I explain how the parameters may be estimation when one (or more) samples of the complete trajectory $X$ is given. A detailed explanation is provided in Wilkinson [119, p. 278-281]. Given a sample complete trajectory $x$ (or the equivalent complete reaction event data), the *complete data-likelihood* is given by Wilkinson [119, p. 279]

$$\pi(x \mid \theta) = \prod_{j=1}^{n} h_{n_j}(\mathbf{x}(t_{j-1}), \theta) \exp\left\{ - \int_0^T h_0(\mathbf{x}(t), \theta) dt \right\} \tag{5.13}$$

**Definition 5.1** (Separable propensity)**.** *Reaction propensity $h_j(\mathbf{X}(t), \theta)$ is called separable if the model parameters $\theta$ and the state $\mathbf{X}(t)$ can be separated as follows:*

$$h_j(\mathbf{X}(t), \theta) = k_j^{c_j} g_j(\mathbf{X}(t)) \tag{5.14}$$

*in which $k_j$ denotes the associated reaction rate constant and $c_j$ is some known constant. The quantity $g_j(\mathbf{X}(t))$ is called rateless propensity.*

I assume that the reaction propensities of all reactions in the system are separable. This assumption is completely reasonable because all elementary reactions, as described in Gillespie, 1976 [32], have separable reaction propensities. Further, the constant $c_j$ is also known to be equal to 1. I will further assume that $c_j = 1$ for all reactions in the system. It is possible that reactions obtained as a result of model reduction may not follow this assumption (see Mastny et al., 2007 [67]). The parameter estimation methods described in this thesis do not seem to require that a model reduction be performed. Thus, the full model which follows the separable propensities assumption , may be used. Under these assumptions, the

likelihood expression in Eq. (5.13) may be re-written as

$$\pi(x \mid \theta) = \prod_{j=1}^{n} g_{n_j}(\mathbf{x}(t_{j-1})) \prod_{i=1}^{n_r} k_i^{r_i} \exp\left\{-k_i \int_0^T g_i(\mathbf{x}(t))dt\right\} \tag{5.15}$$

Renaming the following quantities for convenience,

$$g_{\text{prod}} = \prod_{j=1}^{n} g_{n_j}(\mathbf{x}(t_{j-1})) \tag{5.16}$$

$$G_i = \int_0^T g_i(\mathbf{x}(t))dt \tag{5.17}$$

I obtain the following simplified expression for complete-data likelihood

$$\pi(x \mid \theta) = g_{\text{prod}} \prod_{i=1}^{n_r} k_i^{r_i} e^{-k_i G_i} \tag{5.18}$$

Note that the rateless propensities $g_i(\mathbf{x}(t))$, $i = 1, 2, \ldots, nr$ are non-negative, which implies that $G_i \geq 0$, $i = 1, 2, \ldots, nr$. The likelihood expression above resembles a gamma distribution and therefore motivates the use of the gamma prior described in Eq. (5.1). In other words, the prior in Eq. (5.1) is a *conjugate prior* for the likelihood in Eq. (5.18) (Wilkinson, 2012 [119]). Using the following Bayes' rule for complete data

$$\pi(\theta \mid x) = \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(x)} \propto \pi(x \mid \theta)\pi(\theta) \tag{5.19}$$

the *complete-data posterior*, $\pi(\theta \mid x)$, is proportional to the following expression (Wilkinson, 2012 [119])

$$\pi(\theta \mid x) \propto \prod_{i=1}^{n_r} k_i^{r_i + a_i} e^{-k_i(b_i + G_i)} \tag{5.20}$$

Since $\pi(x)$ does not depend upon $\theta$ and the probability density $\pi(\theta \mid x)$ must

integrate to one, we can obtain the following closed-form expression

$$\pi(\theta \mid x) = \prod_{i=1}^{n_r} \frac{(b_i + G_i)^{a_i + r_i - 1}}{\Gamma(a_i + r_i)} k_i^{r_i + a_i} e^{-k_i(b_i + G_i)} \tag{5.21}$$

The complete-data posterior may also be written as

$$\pi(\theta \mid x) = \prod_{i=1}^{n_r} \pi(k_i \mid x) \tag{5.22}$$

$$\pi(k_i \mid x) = Ga(a_i + r_i, b_i + G_i) \tag{5.23}$$

in which $\pi(k_i \mid x)$ is called the *complete-data marginal posterior* . Since we know three out of the four terms in Eq. (5.19), we can easily obtain the *complete-data marginal likelihood* , $\pi(x)$, as follows:

$$\begin{aligned} \pi(x) &= \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(\theta \mid x)} \\ &= g_{\text{prod}} \prod_{i=1}^{n_r} \frac{\Gamma(a_i + r_i)}{\Gamma(a_i)} \frac{b_i^{a_i}}{(b_i + G_i)^{a_i + r_i}} \end{aligned} \tag{5.24}$$

Note that Eq. (5.24) may also be obtained by using the fact that the probability density $\pi(\theta \mid x)$ must integrate to one:

$$\begin{aligned} \int_\theta \pi(\theta \mid x) d\theta &= 1 \\ \frac{\int_\theta \pi(x \mid \theta)\pi(\theta) d\theta}{\pi(x)} &= 1 \\ \pi(x) &= \int_\theta \pi(x \mid \theta)\pi(\theta) d\theta \end{aligned}$$

The *maximum a posteriori* (MAP) estimate, $\hat{\theta}$, may be obtained by maximizing the posterior

$$\hat{\theta} = \max_\theta \ \pi(\theta \mid x) \tag{5.25}$$

Looking at Eq. (5.22), MAP estimate for each rate constant, $\hat{k}_i$, $i = 1, 2, \ldots, n_r$ may be independently obtained by maximizing the corresponding complete-data marginal posterior, $\pi(k_i \mid x)$.

$$\hat{k}_i = \max_{k_i} \ \pi(k_i \mid x) \tag{5.26}$$

$$\pi(k_i \mid x) = Ga(a_i + r_i, b_i + G_i) \quad i = 1, 2, \ldots, n_r \tag{5.27}$$

Note that since $b_i > 0$, $b_i + G_i > 0$. The closed form solution for $\hat{k}_i$ may be written as the mode of the gamma distribution

$$\hat{k}_i = \begin{cases} \frac{a_i + r_i - 1}{b_i + G_i} & \text{if } (a_i + r_i) > 1 \\ 0 & \text{if } (a_i + r_i) \leq 1 \end{cases} \tag{5.28}$$
$$i = 1, 2, \ldots, n_r$$

Since $r_i$ represents the number of times reaction $\mathcal{R}_i$ occurred, $r_i \geq 0$, $i = 1, 2, \ldots, n_r$. Therefore, if the shape parameter $a_i$ of the gamma (marginal) prior is chosen so that $a_i > 1$, then a non-zero (but not necessarily informative) estimate of the rate constant is obtained. In other words, if the gamma (marginal) prior has a non-zero maximum, then so does the marginal posterior. Also note that if $r_i \geq 1$, *i.e.*, if reaction $\mathcal{R}_i$ is observed to occur at least once then an informative MAP estimate is obtained for the corresponding rate constant $k_i$.

Note that the *maximum likelihood estimation* (MLE) estimates may also be obtained by maximizing the complete-data likelihood in Eq. (5.18)

$$\hat{\theta}_{\text{MLE}} = \max_{\theta} \ \pi(x \mid \theta) \tag{5.29}$$

Since the likelihood expression in Eq. (5.18) is also separable into individual rate constants, estimate of each rate constant may be obtained by the following maxi-

mization

$$\hat{k}_{i,\text{MLE}} = \max_{k_i} \; k_i^{r_i} e^{-k_i G_i} \quad i = 1, 2, \ldots, n_r \tag{5.30}$$

which has the following closed form solution

$$\hat{k}_{i,\text{MLE}} = \begin{cases} \frac{r_i}{G_i} & \text{if } r_i > 0,\ G_i > 0 \\ \text{no estimate} & \text{if } r_i = 0,\ G_i = 0 \\ 0 & \text{if } r_i = 0,\ G_i > 0 \\ \infty & \text{if } r_i > 0,\ G_i = 0 \end{cases} \tag{5.31}$$
$$i = 1, 2, \ldots, n_r$$

under the feasibility conditions that $r_i \geq 0$ and $G_i \geq 0$. Note that an informative estimate is obtained only when both $r_i$ and $G_i$ are positive. If both $r_i$ and $G_i$ are zero, then the likelihood expression in Eq. (5.18) does not depend on $k_i$ at all and, therefore, no information can be obtained about this rate constant from the given trajectory $x$. In such a case, more data (in the form of more trajectories) is required which provide both positive values of both $r_i$ and $G_i$. If $r_i > 0$ but $G_i = 0$, then the likelihood does not have a maximum in $k_i$ and results in an infinitely large MLE estimate of $k_i$. A closer look at the definition of $G_i$ in Eq. (5.17) reveals that for $i = 1, 2, \ldots, n_r$,

$$\begin{aligned} G_i = \int_0^T g_i(\mathbf{x}(t))dt = 0 \\ \implies g_i(\mathbf{x}(t)) = 0 \;\forall\; t \in [0, T] \quad (\because\ g_i(\cdot) \geq 0 \;\forall\; t \in [0, T]) \\ \implies r_i = 0 \quad (\because \text{ reaction cannot occur with zero reaction propensity}) \end{aligned} \tag{5.32}$$

Thus, the case of $r_i > 0$ but $G_i = 0$ is not feasible and therefore an infinitely large MLE estimate of $k_i$ can never be obtained. Conversely, if $r_i = 0$ but $G_i > 0$ (which

is feasible), then the likelihood is maximized when $k_i$ achieves the minimum possible value of 0. Compare the Eq. (5.31) with Eq. (5.28). The use of a gamma prior with parameters $(a_i, b_i)$, ensures the MAP estimate, $\hat{k}_i$ does not become indeterminate. If $r_i = 0$ and $G_i = 0$, the complete-data marginal posterior $\pi(k_i \mid x)$ is the same as marginal prior $\pi(k_i)$, thus indicating that no information was provided by the data $x$. In other words, the trajectory was non-informative with respect to $k_i$. If $r_i = 0$ but $G_i > 0$ then the rate parameter changes from the marginal prior to the marginal posterior, thus, providing some information. Finally, the trajectory $x$ provides the most information, when both $r_i$ and $G_i$ are positive because both shape and rate parameters change from $(a_i, b_i)$ to $(a_i + r_i, b_i + G_i)$. Chapters 6 and 7 deals with these issues.

Another important point to note is that the variables, $r_i = r_i(x)$ and $G_i = G_i(x)$, $i = 1, 2, \ldots, n_r$, form the *sufficient statistics* for the posterior, $\pi(\theta \mid x)$.

## 5.2 ESTIMATION USING EXACT METHOD

Parameters may be easily estimated using both maximum likelihood estimation and Bayesian inference when a complete-data trajectory $x$ is available. However, as discussed in Chapter 2, experimental techniques are not advanced enough (yet) to provide us with reaction event data. Instead, we have to rely on *measurement data* which provides us with the number of molecules of species at different time points. This measurement data or *discrete time data* is denoted by $y$. In this section, I describe an exact method to estimate parameters given only the measurement data.

### 5.2.1 *Measurement data*

In this subsection, I define the structure of measurement data. Some of this notation has been defined previously in Chapter 2 but the entire notation has been restated here. Let $\{s_0, s_1, \ldots, s_m\}$ be a strictly increasing sequence of $m+1$ time points for which the measurements are performed. Let $Y_i$, $j = 0, 1, \ldots, m$ be the random variable denoting the measurement of the state of the system at time $s_i$. The measurement $Y_i$ is said to have the following characteristics

$$Y_i = \mathbf{C}\mathbf{X}(s_i) + V_i \tag{5.33}$$

$$Y_i \in \mathbb{R}^n_{\text{meas}},\ \mathbf{X}(s_i) \in \mathbb{R}^{n_s},\ V_i \in \mathbb{R}^n_{\text{meas}} \tag{5.34}$$

$$\mathbf{C} \in \mathbb{R}^{n_{\text{meas}} \times n_s} \tag{5.35}$$

$$0 < n_{\text{meas}} \leq n_s,\ i = 0, 1, \ldots, m$$

in which,

1. the rows of $\mathbf{C}$ form a unique subset of the rows of the identity matrix, $\mathbf{I}_{n_s \times n_s} \in \mathbb{R}^{n_s \times n_s}$
2. $\mathbf{X}(s_i)$ is the (random) state of the system
3. $V_i$ is the random variable denoting measurement error
4. $n_{\text{meas}}$ is the number measured states

Note that if all $n_s$ species are measured, then the measurements are called *full measurements* and $\mathbf{C} = \mathbf{I}_{n_s \times n_s}$. If only a subset of the $n_s$ species are measured, then the measurements are called *partial measurements*. The set

$$Y = \{Y_0, Y_1, \ldots, Y_m\} \tag{5.36}$$

represents the measurements available at times $\{s_0, s_1, \ldots, s_m\}$. A sample of $Y$ is denoted by

$$y = \{y_0, y_1, \ldots, y_m\} \tag{5.37}$$

I will only consider the cases in which measurement error is absent or negligible *i. e.*

$$V_i = 0 \qquad i = 0, 1, \ldots, m$$

**Proposition 5.1.** *If measurement error is zero,* i. e. $V_i = 0$, $i = 0, 1, \ldots, m$, *then* $\{X = x\} \subseteq \{Y = y\}$. *Equivalently* $\{X = x\} \implies \{Y = y\}$. *In shorthand,* $X \subseteq Y$ *and* $X \implies Y$.

*Proof.* Let $Y_{j,i}$ be the $j^{\text{th}}$ element of the $i^{\text{th}}$ measurement, for $j = 1, 2, \ldots, n_{\text{meas}}$, $i = 0, 1, \ldots, m$. Then, $Y_{j,i}$ is equal to one of the elements of $\mathbf{X}(s_i)$. Thus, the set of random variables $Y_i = (Y_{1,i}, Y_{2,i}, \ldots, Y_{n_{\text{meas}},i})$ is a subset of the set of random variables $\mathbf{X}(s_i) = (X_1(s_i), X_2(s_i), \ldots, X_{n_s}(s_i))$. Noting that the use of "comma" indicates an "intersection", we can rewrite the sets of random variables as events as follows

$$Y_i \text{ corresponds to the event } \{Y_i = y_i\} \tag{5.38}$$

$$\mathbf{X}(s_i) \text{ corresponds to the event } \{\mathbf{X}(s_i) = \mathbf{x}(s_i)\} \tag{5.39}$$

in which,

$$\{Y_i = y_i\} = \{Y_{1,i} = y_{1,i}\} \cap \{Y_{2,i} = y_{2,i}\} \cap \cdots \cap \{Y_{n_{\text{meas}},i} = y_{n_{\text{meas}},i}\} \tag{5.40}$$

$$\{\mathbf{X}(s_i) = \mathbf{x}(s_i)\} = \{X_1(s_i) = x_1(s_i)\} \cap \{X_2(s_i) = x_2(s_i)\} \cap \cdots \cdots \cap \{X_{n_s}(s_i) = x_{n_s}(s_i)\} \tag{5.41}$$

Every event in Eq. (5.40) is also present in Eq. (5.41). However, the set in Eq. (5.41) may be more restricted due to intersection with other events that are not present in Eq. (5.40). Hence, $\{\mathbf{X}(s_i) = \mathbf{x}(s_i)\} \subseteq \{Y_i = y_i\}$. Consequently,

$$\begin{aligned}\{\mathbf{X}(s_0) = \mathbf{x}(s_0)\} \cap \{\mathbf{X}(s_1) = \mathbf{x}(s_1)\} \cap \cdots \cap \{\mathbf{X}(s_m) = \mathbf{x}(s_m)\} \\ \subseteq \{Y_0 = y_0\} \cap \{Y_1 = y_1\} \cap \cdots \{Y_m = y_m\}\end{aligned} \tag{5.42}$$

Further,

$$\bigcap_{t \in [s_0, s_m]} \{\mathbf{X}(t) = \mathbf{x}(t)\} \subseteq \{\mathbf{X}(s_0) = \mathbf{x}(s_0)\} \cap \cdots \cap \{\mathbf{X}(s_m) = \mathbf{x}(s_m)\} \tag{5.43}$$

Using Eq. (5.42) and Eq. (5.43),

$$\bigcap_{t \in [s_0, s_m]} \{\mathbf{X}(t) = \mathbf{x}(t)\} \subseteq \{Y_0 = y_0\} \cap \{Y_1 = y_1\} \cap \cdots \{Y_m = y_m\} \tag{5.44}$$

Finally, re-writing the terms in familiar "comma" notation,

$$\{X = x\} = \{\mathbf{X}(t) : t \in [s_0, s_m]\} = \bigcap_{t \in [s_0, s_m]} \{\mathbf{X}(t) = \mathbf{x}(t)\} \tag{5.45}$$

$$\{Y = y\} = \{Y_0 = y_0\} \cap \{Y_1 = y_1\} \cap \cdots \{Y_m = y_m\} \tag{5.46}$$

and substituting these expressions into Eq. (5.44),

$$\{X = x\} \subseteq \{Y = y\} \tag{5.47}$$

$$\text{Or, } X \subseteq Y \tag{5.48}$$

□

### 5.2.2 *Measurement-data distributions*

Similar to the complete-data distributions describes in Section 5.1.2, the *measurement-data distributions* are related to each other by the following Bayes' rule

$$\pi(\theta \mid y) = \frac{\pi(y \mid \theta)\pi(\theta)}{\pi(y)} \propto \pi(y \mid \theta)\pi(\theta) \tag{5.49}$$

in which,

1. $\pi(y \mid \theta)$ is the *measurement-data likelihood*
2. $\pi(\theta \mid y)$ is the *measurement-data posterior*
3. $\pi(y)$ is the *measurement-data marginal likelihood*
4. $\pi(\theta)$ is the prior

As discussed in Chapter 2, a system of stochastic chemical reactions is essentially a continuous time, discrete space Markov chain. Let $\mathbf{X}(t)$ denote the state of this Markov chain at time $t$. Let $\mathbb{S}$ be the set of all possible states of the Markov chain, also known as the *state space*. Since the state space is discrete, $\mathbb{S}$ can be at most countable. The defining property of the continuous-time Markov chain is stated as

$$P(\mathbf{X}(t+dt) = \mathbf{x} \mid \mathbf{X}(s) = \mathbf{x}(s),\ s \in [0,t]) = P(\mathbf{X}(t+dt) = \mathbf{x} \mid \mathbf{X}(t) = \mathbf{x}(t)) \quad \forall t \geq 0,\ x \in \mathbb{S} \tag{5.50}$$

The behavior of the Markov chain is described by the *transition kernel*

$$p(\mathbf{x}, t, \mathbf{x}', s) = P(\mathbf{X}(t+s) = \mathbf{x}' \mid \mathbf{X}(t) = \mathbf{x}) \tag{5.51}$$

The equation describing this Markov chain may be derived from the basic hypothesis (see Section 2.1) [32, Eq. (13)]

$$h_i(\mathbf{x}(t),\theta)dt = \text{probability to first order in } dt, \text{ that reaction } \mathcal{R}_i$$
$$\text{will occur in the next time interval } dt$$

If the reaction propensities, $h_i(\mathbf{x}(t),\theta)$, $i = 1,2,\ldots,n_r$, do not depend directly upon the time $t$, then the transition kernel $p(\mathbf{x},t,\mathbf{x}',s)$ does not depend on $t$, thus, making the Markov chain *time-homogeneous* or just *homogeneous*. For a homogeneous Markov chain, the transition kernel may be represented by the following reduced notation

$$\begin{aligned} p(\mathbf{x},\mathbf{x}',s) &= p(\mathbf{x},t,\mathbf{x}',s) \\ &= P(\mathbf{X}(t+s) = \mathbf{x}' \mid \mathbf{X}(t) = \mathbf{x}) \end{aligned} \tag{5.52}$$

Since $\mathbb{S}$ is at most countable, we can index all the states in the state space by the set (or subset) of natural numbers, $\mathbb{N}$. Thus, a particular state $\mathbf{x}$ may be represented only by an index $i$, $i = 1,2,\ldots,|\mathbb{S}|$. Further, the transition probabilities defined in Eq. (5.52) may be arranged into a *transition matrix* , $\mathbf{P}(s)$, as follows

$$\begin{aligned} &\mathbf{P}(s) \in \mathbb{R}^{|\mathbb{S}|\times|\mathbb{S}|} \\ &P_{i,j}(s) = p(i,j,s) = P(\mathbf{X}(t+s) = j \mid \mathbf{X}(t) = i) \quad \forall t \\ &\qquad i = 1,2,\ldots,|\mathbb{S}| \quad j = 1,2,\ldots,|\mathbb{S}| \end{aligned} \tag{5.53}$$

Note that $\mathbf{P}(0) = \mathbf{I}$. Chapman-Kolmogorov equations may be derived as follows

$$\begin{aligned}
P_{i,j}(t+s) &= P(\mathbf{X}(t+s) = j \mid \mathbf{X}(0) = i) \\
&= \sum_{k=1}^{|\mathbb{S}|} P(\mathbf{X}(t+s) = j, \mathbf{X}(s) = k \mid \mathbf{X}(0) = i) \\
&= \sum_{k=1}^{|\mathbb{S}|} P(\mathbf{X}(t+s) = j \mid \mathbf{X}(s) = k, \mathbf{X}(0) = i)\ P(\mathbf{X}(s) = k \mid \mathbf{X}(0) = i) \\
&= \sum_{k=1}^{|\mathbb{S}|} P(\mathbf{X}(t+s) = j \mid \mathbf{X}(s) = k)\ P(\mathbf{X}(s) = k \mid \mathbf{X}(0) = i) \\
&= \sum_{k=1}^{|\mathbb{S}|} P_{i,k}(s)\ P_{k,j}(t)
\end{aligned}$$

The matrix version of Chapman-Kolmogorov equations is given as

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s) = \mathbf{P}(s)\mathbf{P}(t) \tag{5.54}$$

The *transition rate matrix*, $\mathbf{Q}$, is defined as the derivative of $\mathbf{P}(t)$ at $t = 0$

$$\begin{aligned}
\mathbf{Q} &= \frac{d}{dt}\mathbf{P}(t)\Big|_{t=0} \\
&= \lim_{dt \to 0} \frac{\mathbf{P}(dt) - \mathbf{P}(0)}{dt} \\
&= \lim_{dt \to 0} \frac{\mathbf{P}(dt) - \mathbf{I}}{dt}
\end{aligned} \tag{5.55}$$

Note that $\mathbf{Q} \in \mathbb{R}^{|\mathbb{S}| \times |\mathbb{S}|}$. A differential equation may be derived using the above expressions Wilkinson [119, p. 138]

$$\begin{aligned}
\frac{d}{dt}\mathbf{P}(t) &= \lim_{dt \to 0} \frac{\mathbf{P}(t+dt) - \mathbf{P}(t)}{dt} \\
&= \lim_{dt \to 0} \frac{\mathbf{P}(dt)\mathbf{P}(t) - \mathbf{P}(t)}{dt} \\
&= \lim_{dt \to 0} \frac{\mathbf{P}(dt) - \mathbf{I}}{dt}\mathbf{P}(t) \\
&= \mathbf{Q}\mathbf{P}(t)
\end{aligned} \tag{5.56}$$

Equation (5.56) is the matrix version of Kolmogorov's forward equations (KFEs). The matrix $\mathbf{P}(t)$ may be obtained as the solution to the following matrix differential equation

$$\begin{aligned} \frac{d}{dt}\mathbf{P}(t) &= \mathbf{Q}\mathbf{P}(t) \\ \mathbf{P}(0) &= \mathbf{I} \end{aligned} \tag{5.57}$$

The solution to Eq. (5.57) is simply

$$\mathbf{P}(t) = e^{\mathbf{Q}t} \tag{5.58}$$

in which $e^{\mathbf{Q}t}$ is the matrix exponential.

Assuming that full measurements are available, *i. e.*, $\mathbf{C} = \mathbf{I}_{n_s \times n_s}$, the measurement-data likelihood, $\pi(y \mid \theta)$, may be obtained as follows

$$\begin{aligned} \pi(y \mid \theta) &= P(Y(s_m) = y_m, Y(s_{m-1}) = y_{m-1}, \ldots, Y(s_0) = y_0 \mid \theta) \\ &= P(\mathbf{X}(s_m) = y_m, \mathbf{X}(s_{m-1}) = y_{m-1}, \ldots, \mathbf{X}(s_0) = y_0 \mid \theta) \\ &= P(\mathbf{X}(s_m) = y_m \mid \mathbf{X}(s_{m-1}) = y_{m-1}, \ldots, \mathbf{X}(s_0) = y_0, \ \theta) \\ &\quad \times P(\mathbf{X}(s_{m-1}) = y_{m-1}, \ldots, \mathbf{X}(s_0) = y_0 \mid \theta) \\ &= P(\mathbf{X}(s_m) = y_m \mid \mathbf{X}(s_{m-1}) = y_{m-1}, \ \theta) \\ &\quad \times P(\mathbf{X}(s_{m-1}) = y_{m-1} \mid \mathbf{X}(s_{m-2}) = y_{m-2}, \ \theta) \\ &\quad \vdots \\ &\quad \times P(\mathbf{X}(s_1) = y_1 \mid \mathbf{X}(s_0) = y_0, \ \theta) \\ &\quad \times P(\mathbf{X}(s_0) = y_0 \mid \theta) \end{aligned} \tag{5.59}$$

Equation (5.59) may be written more compactly as

$$\pi(y \mid \theta) = P(\mathbf{X}(s_0) = y_0 \mid \theta) \prod_{i=1}^{m} P(\mathbf{X}(s_i) = y_i \mid \mathbf{X}(s_{i-1}) = y_{i-1},\ \theta) \tag{5.60}$$

Assuming that the initial condition $y_0$ is fixed and known,

$$P(\mathbf{X}(s_0) = y_0 \mid \theta) = 1$$

The term, $P(\mathbf{X}(s_i) = y_i \mid \mathbf{X}(s_{i-1}) = y_{i-1},\ \theta)$, is an element of the transition matrix $\mathbf{P}(s_i - s_{i-1})$ which is given as the following matrix exponential (Moler and Van Loan, 1978 [72], Moler and Van Loan, 2003 [73])

$$\mathbf{P}(s_i - s_{i-1}) = e^{\mathbf{Q}(s_i - s_{i-1})}$$

The elements of $\mathbf{Q}$ are comprised of reaction propensities making $\mathbf{Q}$ a function of $\theta$. Thus, the measurement-data likelihood, $\pi(y \mid \theta)$, may be computed for every value of $\theta$. Evaluation of $\pi(y \mid \theta)$ for one value of $\theta$ requires evaluation of as many as $m$ matrix exponentiations. If the time points $\{s_0, s_1, \ldots, s_m\}$ are equally spaced then $\Delta s = s_i - s_{i-1}$, $\forall\ i = 1, 2, \ldots, m$ and only one matrix exponentiation $e^{\mathbf{Q}\Delta s}$ is required.

### 5.2.3 *Examples*

The exact method described in Section 5.2, requires full measurements *i.e.*, all species must be measured, which limits the applicability of the method. In the case of partial measurements , the unmeasured species may have to be sampled using a suitable distribution and then averaged over appropriately. However, as I will show through the following series of examples, even in the case of full measurements, the exact method is intractable for many systems of interest.

*Example 1*

Consider the simplest possible example

$$\mathrm{A} \xrightarrow{k_1} \mathrm{B} \tag{5.61}$$

Note that the above reaction conserves the total number of molecules in the system, *i. e.*,

$$A(t) + B(t) = A(0) + B(0) = \text{ constant } \quad \forall\, t \tag{5.62}$$

in which, the $A(t)$ and $B(t)$ represent the number of molecules of species A and B respectively. The state of the system is represented by $\mathbf{X}(t) = \begin{bmatrix} A(t) & B(t) \end{bmatrix}^T$. For an initial condition of $\mathbf{x}(0) = \begin{bmatrix} a(0) & b(0) \end{bmatrix}^T = \begin{bmatrix} 100 & 0 \end{bmatrix}^T$, the total number of molecules in the system is 100.

A chemical reaction system involving $n_s$ species that conserves the number of molecules can only assume a finite number of states. The number of states in the state space, $|\mathbb{S}|$, is bounded by the following combinatorial formula

$$|\mathbb{S}| \leq \binom{N_0 + n_s - 1}{n_s - 1} \tag{5.63}$$

in which, $N_0$ represents the total (conserved) number of molecules and the expression $\binom{n}{k}$ has its usual meaning $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

For the example at hand, the state space attains the maximum, *i.e.*,

$$\begin{aligned} |\mathbb{S}| &= \binom{a(0) + b(0) + n_s - 1}{n_s - 1} \\ &= \binom{101}{1} = 101 \end{aligned} \tag{5.64}$$

Thus, the transition matrix, $\mathbf{Q}$, is of size $101 \times 101$. In order to compute the likelihood in Eq. (5.60), the exponentiation of matrix, $\mathbf{Q}(s_i - s_{i-1})$ needs to be computed. For every value of $\theta$, at least one and at most $m$ matrix exponentiations are required. For, $\mathbf{Q} \in \mathbb{R}^{101 \times 101}$, this computation is not expensive.

*Example 2*

Consider another example with only four reactions

$$\mathrm{A} \xrightarrow{k_1} \mathrm{B} \tag{5.65}$$

$$\mathrm{B} \xrightarrow{k_2} \mathrm{C} \tag{5.66}$$

$$\mathrm{C} \xrightarrow{k_3} \mathrm{D} \tag{5.67}$$

$$\mathrm{D} \xrightarrow{k_4} \mathrm{A} \tag{5.68}$$

which also conserves the number of molecules. The initial condition, $\mathbf{x}(0) = \begin{bmatrix} a(0) & b(0) & c(0) & d(0) \end{bmatrix}^T = \begin{bmatrix} 100 & 0 & 0 & 0 \end{bmatrix}^T$ corresponds to the same number of total molecules, $N_0 = 100$. The number of states the system of reactions in Eqs. (5.65)-(5.66) can take is given by

$$\begin{aligned} |\mathbb{S}| &= \binom{a(0) + b(0) + c(0) + d(0) + n_s - 1}{n_s - 1} \\ &= \binom{103}{3} = 176851 \end{aligned} \tag{5.69}$$

Thus, the transition matrix, $\mathbf{Q}$, is of size $176851 \times 176851$. The computation of likelihood now becomes prohibitively expensive even for one matrix exponentiation.

As a result, parameter estimation for this reaction system using the exact method is intractable.

*Example 3*

Consider the *viral RNA genome replication* reaction (in lumped form)

$$\mathrm{G} + nA \xrightarrow{k_1'} 2\mathrm{G} \tag{5.70}$$

in which G represents the viral genome and A represents nucleic acids. The stoichiometric coefficient $n$ represents the number of nucleic acid molecules required to create a new genome. Usually, nucleic acids exist in large quantities (millions) while the viral genomes are very few (as few as one). Further, the time-evolution of nucleic acids is unimportant. In such a case, the reaction in Eq. (5.70) may be approximated as

$$\mathrm{G} \xrightarrow{k_1} 2\mathrm{G} \tag{5.71}$$

Such an approximation is ubiquitous in systems biology [36, 87, 81, 114, 39, 88, 14, 50, 105, 6, 119, 21] and has served well in various modeling studies. Note that the reaction in Eq. (5.71), unlike the reaction in Eq. (5.70), does not conserve the total number of molecules. In fact, Eq. (5.71) corresponds to a possibly infinite number of molecules. In other words the state space is infinite. This causes the transition matrix, $\mathbf{Q}$ to become infinite-dimensional, thus making it almost impossible exponentiate. While some research has been done to solve the chemical master equation (and equivalently Kolmogorov's forward equation) analytically under special cases [58, 74], it is usually not possible to do so. As a result, if no method is available to solve the CME or KFE, exact method as described in this section may not be used. Note that even if the reaction in Eq. (5.70) is used, the

state space, $\mathbb{S}$, would be very close to being infinite-dimensional.

*Example 4*

While exact method as described in this section is intractable even for many systems of interest, I demonstrate the use of the exact method using the simple following example

$$\mathrm{A} \xrightarrow{k_1} \mathrm{B} \tag{5.72}$$

$$\mathrm{B} \xrightarrow{k_2} \mathrm{C} \tag{5.73}$$

Since the system of reactions above has a finite state space*state space!finite*, the exact method may be used. Also note that the number of molecules is conserved. Thus, if the total number molecules at the initial condition is chosen to be small, the exact method may be used to estimate parameters.

Figure 5.1 shows a typical dataset generated using the above reactions with initial conditions $\mathbf{X}(0) = [A(0)\ B(0)\ C(0)]^T = [7\ 8\ 0]^T$ and true parameter values, $\theta_0 = [0.04\ 0.11]^T$. The gamma prior is described by the parameters described in Table 5.1. Note that the mode of the prior distribution is chosen to be orders of magnitude different from the true parameter values. Using the exact method, the joint posterior, $\pi(\theta \mid y)$, may be obtained as shown in Figure 5.2. The joint gamma prior, $\pi(\theta)$ is also shown. Note that the joint prior when compared is almost uniform compared to the joint posterior. This indicates that the prior does not bias the parameter estimation procedure. The corresponding marginal priors and posteriors are shown in Figure 5.3.

The MAP estimates, $\hat{\theta}_{\text{exact}}$, are shown in Table 5.1. Note that while the MAP estimates are close to the true values, $\theta_0$, and far away from the mode of the

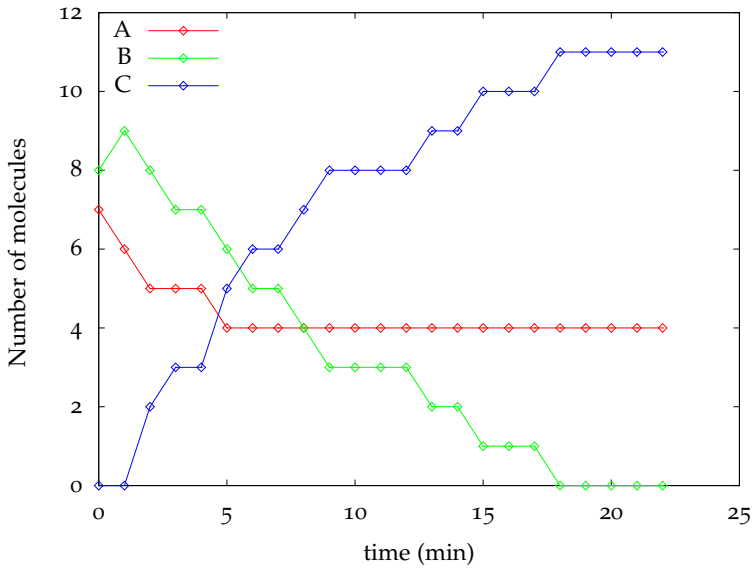



Figure 5.1: Measurement data ($y$) simulated using true parameters, $\theta_0 = [k_{1,0}\ k_{2,0}]^T = [0.04\ 0.11]^T$ and the initial conditions $\mathbf{X}(0) = [A(0)\ B(0)\ C(0)]^T = [7\ 8\ 0]^T$.

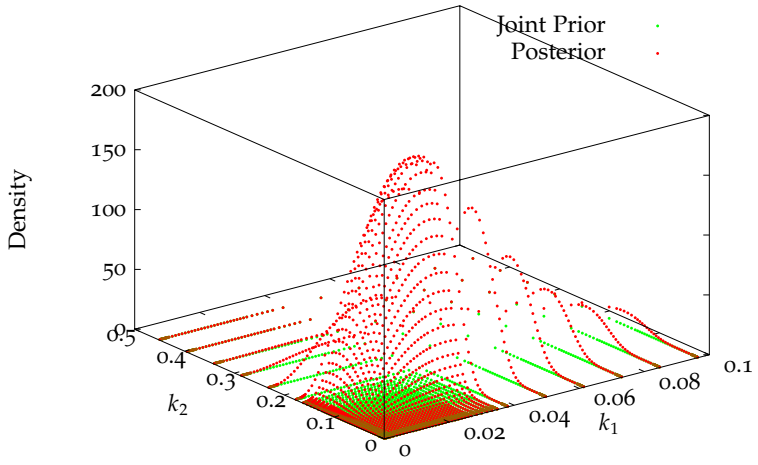



Figure 5.2: Joint prior ($\pi(\theta)$) and posterior ($\pi(\theta \mid y)$) obtained using exact method for measurement data in Figure 5.1

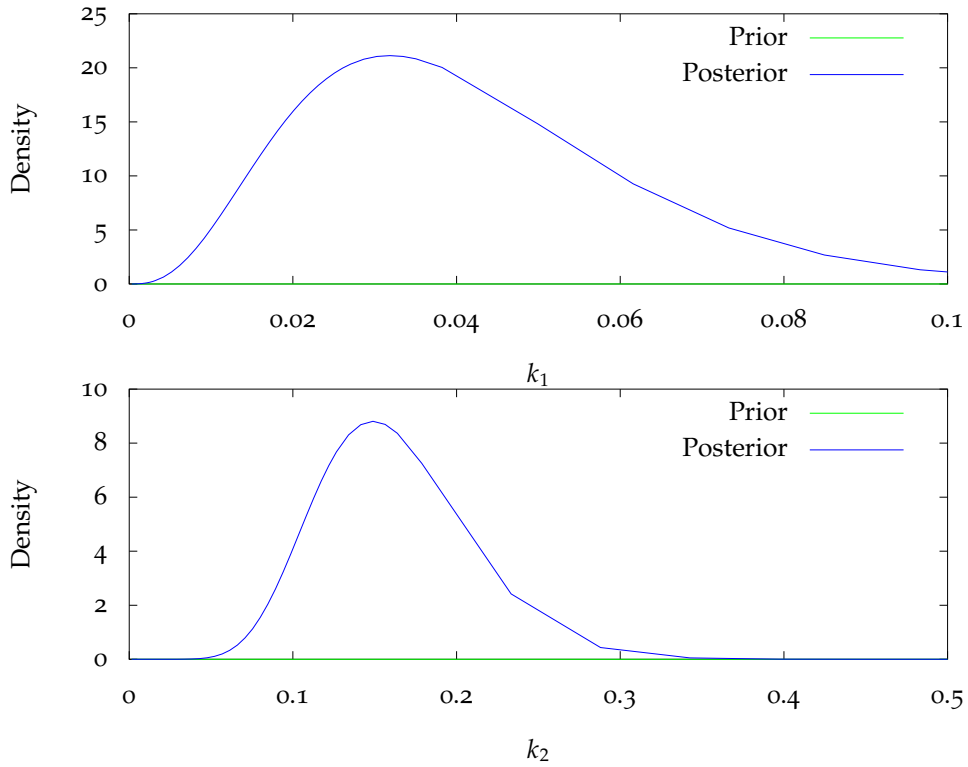



Figure 5.3: Marginal prior ($\pi(k_i)$) and posterior($\pi(k_i \mid y)$) obtained using exact method for measurement data in Figure 5.1

Table 5.1: Parameter true values, Exact MAP estimates and prior parameters

| Reactions | Parameters | True values | MAP estimates | Prior Parameters | | |
|---|---|---|---|---|---|---|
| | | $\theta_0$ | $\hat{\theta}_{\text{exact}}$ | $\hat{\theta}_{\text{prior}}$ | $a$ | $b$ |
| Reaction 1 | $k_1$ | 0.04 | 0.0318 | 15 | 1.01 | 0.00067 |
| Reaction 2 | $k_2$ | 0.11 | 0.148 | 8 | 1.01 | 0.00125 |

prior distribution. However, the MAP estimates are not exactly equal to the true values. This difference between parameter estimates and the true values is called *estimator bias due to finite data*. As the term indicates, in the limit of infinite data, the difference between parameter estimates and the true values should converge to zero. Since this parameter estimation is exact, the bias $\theta_0 - \hat{\theta}_{\text{exact}}$ is expected to converge to $\mathbf{0}$.

## 5.3 ESTIMATION USING DETERMINISTIC FORMULATION

As discussed in Chapter 2, a system of chemical reactions may be represented as a system of (possibly nonlinear) ODEs or a continuous-time, discrete state space Markov chain corresponding to the deterministic and stochastic formulations, respectively. These two formulations are applicable in different regimes. The stochastic formulation is more suited when the number of molecules of the reacting species is small (usually 1-100) while the deterministic formulation is valid when the number of molecules is large (usually 1000 or more). As we have seen in Chapter 2, the stochastic formulation, though extremely slow, may be used even when the number of molecules is large. In fact, the stochastic formulation converges to the deterministic formulation in the limit of infinite number of molecules. Even though, deterministic formulation is not applicable in the regime of small number of molecules, it is tempting to use a deterministic formulation for the purposes of parameter estimation. Obviously, if a deterministic model is to be used, the parameter estimation methods described in Chapter 3 may be used. The reasoning behind such an attempt is two-fold. Firstly, the deterministic formulation arises from the stochastic formulation as a limit. Thus, the underlying molecular mechanisms for both formulations are the same. Consequently, it may be expected that the deterministic model, even when inapplicable, can provide the same information about the reaction rate constants as the stochastic model. Sec-

ondly, extensive research has been done on nonlinear ODE solvers (Hindmarsh et al., 2005 [54]) and nonlinear optimizers (Nocedal and Wright, 2006 [77]) which allows relatively easier estimation of parameters from deterministic models (Rawlings and Ekerdt, 2004 [85]). Indeed, if the deterministic formulation provides reasonable parameter estimates, there would be no need to develop other methods that estimate use the stochastic formulation. In such a case, the parameter estimates may be easily obtained through the deterministic formulation and plugged back into the stochastic model for the purposes of prediction. However, as I show in this section, the deterministic formulation does not necessarily provide reliable parameter estimates. I also state the conditions under which the deterministic formulation may or may not be able to provide accurate parameter estimates. In many cases, biological phenomena that are described using a stochastic model do not meet the requirements to allow the use of deterministic formulation for parameter estimation.

### 5.3.1 *A Simple Example*

Consider the following system of reactions

$$\mathrm{A} \xrightarrow{k_1} \mathrm{B} \tag{5.74}$$

$$\mathrm{B} \xrightarrow{k_2} \mathrm{C} \tag{5.75}$$

As discussed in Chapter 2, a system of chemical reactions described by a system of ODEs in concentrations or a Markov chain in the number of molecules depending upon the regime. The rate constants for both regimes, however, may be interpreted differently. In Eq. (5.75), the variables $k_1$, $k_2$ represent the stochastic rate constants (or hazard rate constants; see Section 1.2). Since the reactions in Eq. (5.75), are monomolecular, the deterministic reaction rate constants equal the

stochastic rate constants (see Section 2.1). Thus, in the deterministic regime, the system of reactions may be described using the following system of ODEs

$$\begin{aligned}\frac{d}{dt}c_A &= -k_1 c_A \\ \frac{d}{dt}c_B &= k_1 c_A - k_2 c_B \\ \frac{d}{dt}c_C &= k_2 c_B\end{aligned} \tag{5.76}$$

with some initial conditions $c_A(0) = c_{A0}$, $c_B(0) = c_{B0}$, $c_C(0) = c_{C0}$. Here, the variables $c_A(t)$, $c_B(t)$, $c_C(t)$ denote the concentrations of species A, B and C respectively, measured in moles/volume. Concentration is related to the number of molecules as

$$\text{concentration} = \frac{\text{number of molecules}}{N_a V} \tag{5.77}$$

in which, $N_a \approx 6.022 \times 10^{23}$ $\texttt{moles}^{-1}$ is the Avagadro's constant and $V$ represents the (fixed) volume of the reaction vessel (or cell volume; see Section 2.1). Let $A(t)$, $B(t)$, $C(t)$ denote the number of molecules of species A, B, C respectively. The system of ODEs, Eq. (5.76), involving concentrations may be converted to use number of molecules as follows [3]

$$\begin{aligned}\frac{1}{N_a V}\frac{d}{dt}A &= -\frac{1}{N_a V}k_1 A \\ \frac{1}{N_a V}\frac{d}{dt}B &= \frac{1}{N_a V}\left(k_1 A - k_2 B\right) \\ \frac{1}{N_a V}\frac{d}{dt}C &= \frac{1}{N_a V}k_2 B\end{aligned}$$

[3] In this discussion, average number of molecules, $\mathbb{E}[A(t)]$, is obtained from the data. Hence $\mathbb{E}[A(t)] = A(t)$. Same argument applies to other species.

Table 5.2: Parameter true values, exact and deterministic/least-squares estimates

| Reactions | Parameters | True values $\theta_0$ | Exact estimates $\hat{\theta}_{\text{exact}}$ | Least-squares estimates $\hat{\theta}_{\text{LS}}$ |
|---|---|---|---|---|
| Reaction 1 | $k_1$ | 0.04 | 0.0318 | 0.0452 |
| Reaction 2 | $k_2$ | 0.11 | 0.148 | 0.155 |

or,

$$\begin{aligned}
\frac{d}{dt}A &= -k_1 A \\
\frac{d}{dt}B &= k_1 A - k_2 B \\
\frac{d}{dt}C &= k_2 B \\
A(0) = A_0,\ B(0) &= B_0,\ C(0) = C_0
\end{aligned} \tag{5.78}$$

The system of ODEs in Eq. (5.78) describes the time-evolution of the number of molecules of each species. Also note that these equations involve the stochastic rate constants (which happen to be equal to the deterministic rate constants in this example). Therefore, given measurement data, the deterministic formulation in Eq. (5.78) may be used to estimate parameters $\theta = \begin{bmatrix} k_1 & k_2 \end{bmatrix}$ using the parameters estimation methods described in Section 3.1. Note that given a set of initial conditions $A_0$, $B_0$, $C_0$ and the (stochastic) rate constants in $\theta$, Eq. (5.78) produces only one solution trajectory. But the stochastic formulation (based on a Markov chain) of the reactions in Eq. (5.75) generates different random trajectories.

I estimate the parameters for the measurement data shown in Figure 5.1 and the deterministic formulation in Eq. (5.78) using least-squares minimization (see Section 3.1). Figure 5.4 shows the model fit at the optimal parameter values, $\hat{\theta}_{\text{LS}}$. The deterministic formulation in Eq. (5.78) fits the measurement data reasonably well. The optimal parameters $\hat{\theta}_{\text{LS}}$ are shown in Table 5.2. The estimates are quite close to the true values, $\theta_0$.

The results of this example indicate that the deterministic formulation with

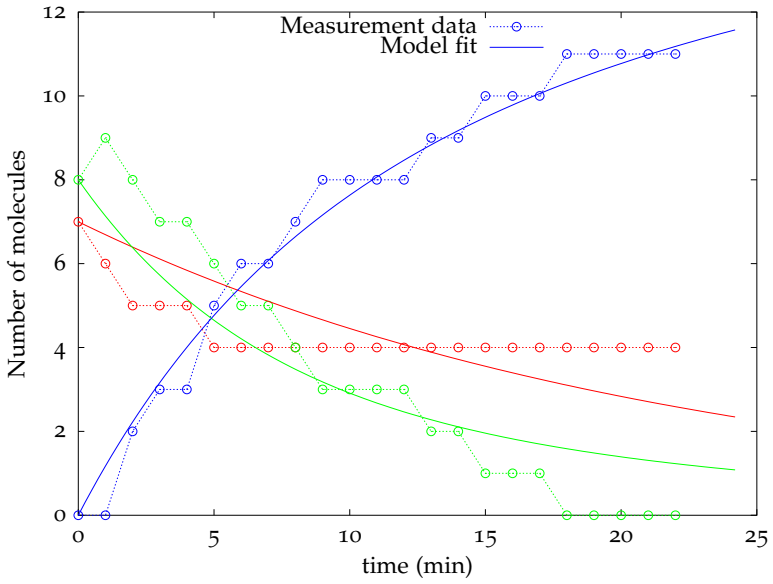



Figure 5.4: Model fit for the measurement data in Figure 5.1

least-squares minimization provides reasonable parameter estimates, $\hat{\theta}_{\text{LS}}$. These estimates are close to the true values and the exact estimates. It may be expected that with more data, the deterministic formulation may provide better estimates.

### 5.3.2 *A Counter Example*

Consider the following system of reactions

$$\text{A} \xrightarrow{k_1} \text{B} \qquad \text{Excitation} \tag{5.79}$$

$$\text{B} + \text{D} \xrightarrow{k_2} 2\text{B} \qquad \text{Replication} \tag{5.80}$$

$$\text{A} \xrightarrow{k_3} \text{C} \qquad \text{Extinction} \tag{5.81}$$

As stated in Section 5.3.1, the variables $k_1$, $k_2$, and $k_3$ denote the stochastic rate constants. The deterministic rate constants for the monomolecular excitation (Eq. (5.79)) and extinction (Eq. (5.81)) reactions are the same as the stochastic rate constants, $k_1$ and $k_3$ respectively. Using Table 2.2, the deterministic rate constant

for the replication reaction (Eq. (5.80)), is given by

$$k_2^{\text{det}} = (N_a V) k_2 \tag{5.82}$$

in which $N_a$ and $V$ have the usual meanings described in the previous section (Section 5.3.1). The deterministic formulation of the reactions in Eqs. (5.79)-(5.81) is given by the following system of ODEs

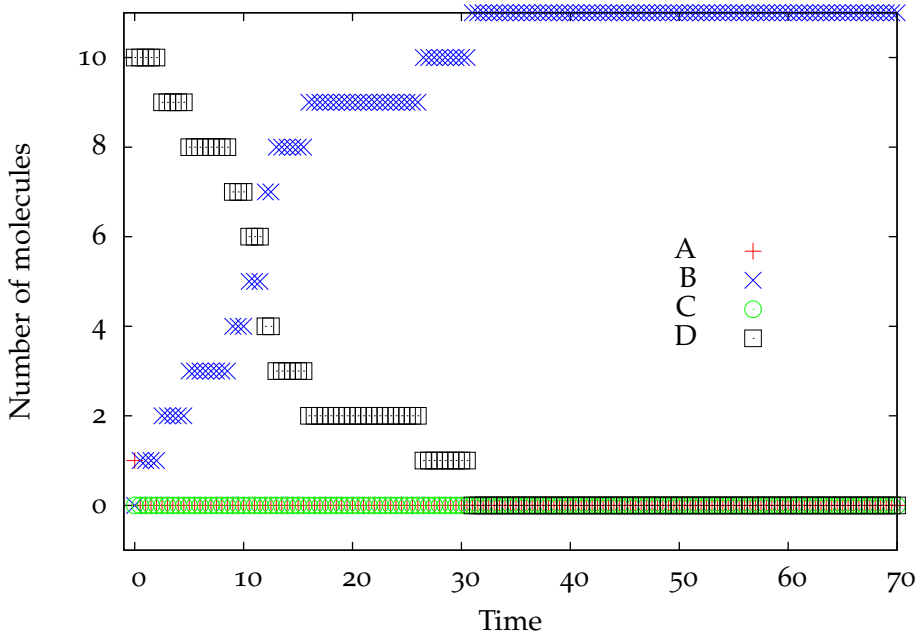
$$\begin{aligned}
\frac{d}{dt} c_A &= -\left(k_1 + k_3\right) c_A \\
\frac{d}{dt} c_B &= k_1 c_A - k_2^{\text{det}} c_B c_D \\
\frac{d}{dt} c_C &= k_3 c_A \\
\frac{d}{dt} c_D &= -k_2^{\text{det}} c_B c_D
\end{aligned} \tag{5.83}$$

with some initial conditions $c_A(0) = c_{A0}$, $c_B(0) = c_{B0}$, $c_C(0) = c_{C0}$, $c_D(0) = c_{D0}$. Using the now familiar conversion of concentrations to number of molecules, we obtain the following system of ODEs

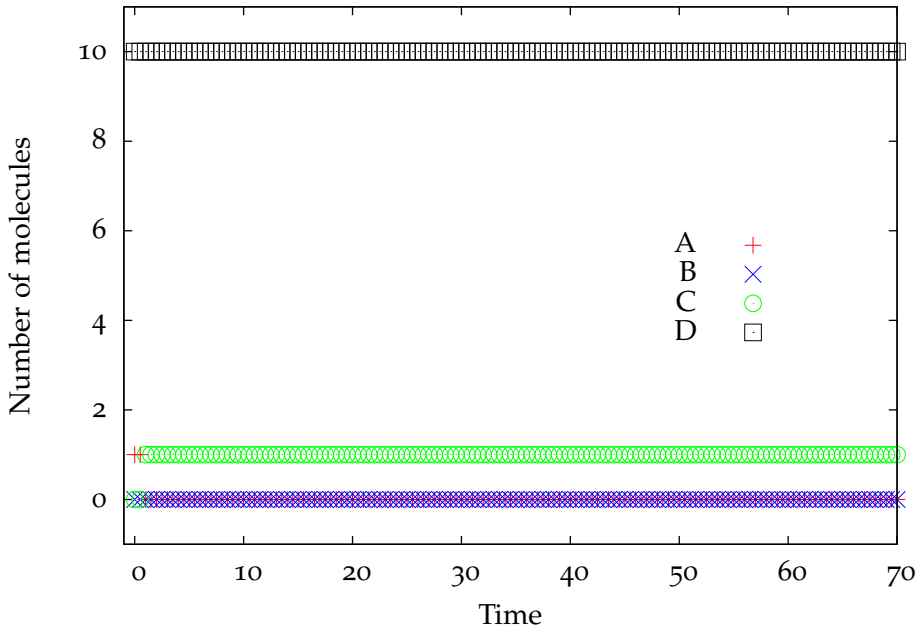
$$\begin{aligned}
\frac{d}{dt} A &= -\left(k_1 + k_3\right) A \\
\frac{d}{dt} B &= k_1 A - k_2 B D \\
\frac{d}{dt} C &= k_3 A \\
\frac{d}{dt} D &= -k_2 B D \\
A(0) = A_0,\ B(0) &= B_0,\ C(0) = C_0,\ D(0) = D_0
\end{aligned} \tag{5.84}$$

Note that Eq. (5.84) does not contain the deterministic rate constant $k_2^{\text{det}}$ at all. Thus, the nonlinear system of ODEs in Eq. (5.84) may be used to estimate parameters given measurement data.

Figures 5.5a-5.5b show the simulation of the stochastic system of reactions in

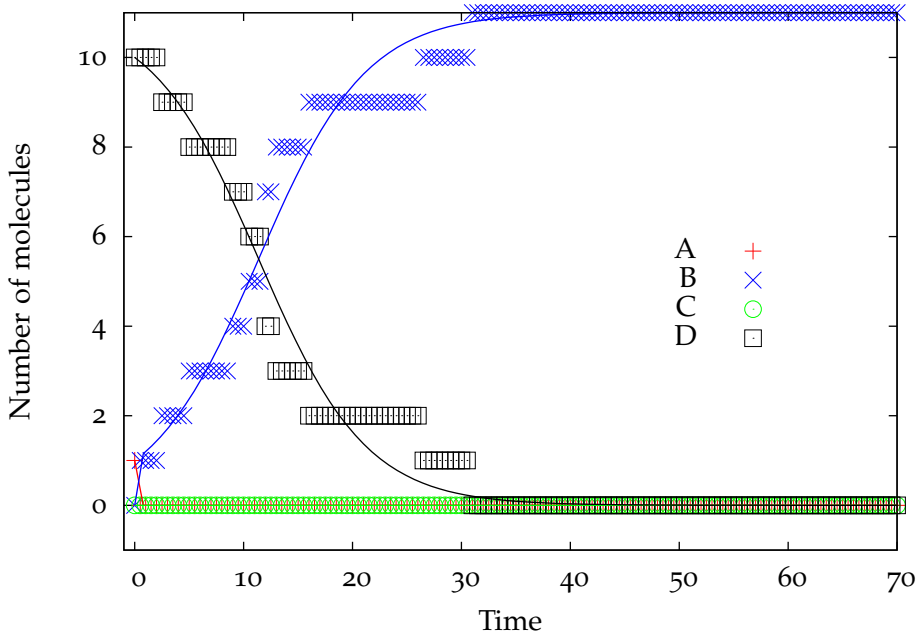



(a) Excitation branch: Measurement data

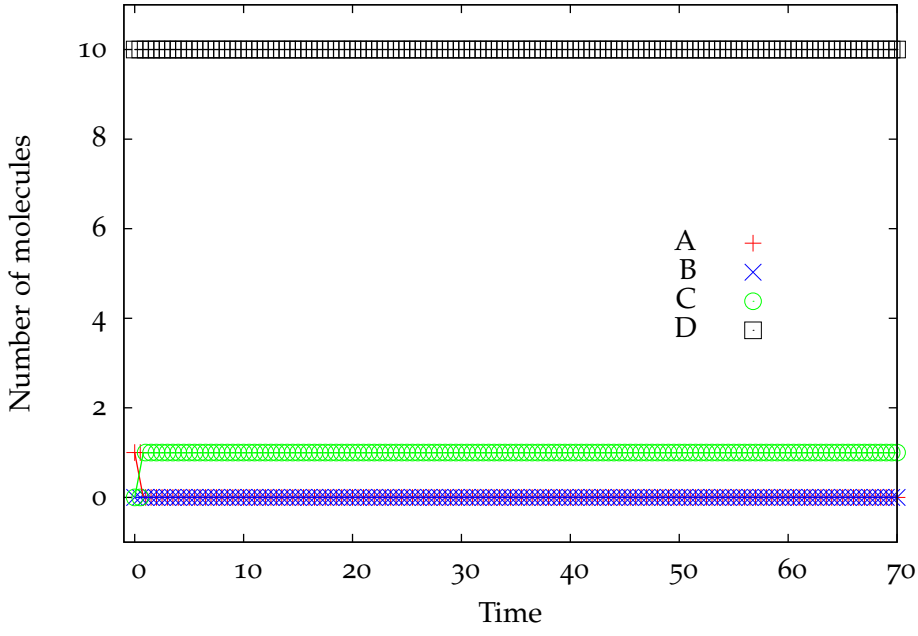



(b) Extinction branch: Measurement data

Figure 5.5: Two typical (random) samples of measurement data (y) showing excitation and extinction branches, obtained using true parameter values, $\theta_0 = [1\ 0.01\ 1]^T$ and initial conditions, $\mathbf{x}(0) = [A_0\ B_0\ C_0\ D_0]^T = [1\ 0\ 0\ 10]^T$




(a) Excitation branch: Model fit




(b) Extinction branch: Model fit

Figure 5.6: Model fit using the same deterministic formulation in Eq. (5.84) and initial conditions. Estimates obtained using least-squares minimization.

Table 5.3: Excitation branch: True values and deterministic/least-squares estimates

| Reactions | Parameters | True values $\theta_0$ | Least-squares estimates $\hat{\theta}_{\text{LS}}$ |
|---|---|---|---|
| Extinction | $k_1$ | 1 | $10^6$ |
| Replication | $k_2$ | 0.01 | 0.018 |
| Extinction | $k_2$ | 1 | $3.5 \times 10^{-4}$ |

Eqs. (5.79)-(5.81) using the common initial condition of $\mathbf{x}(0) = [A_0\ B_0\ C_0\ D_0]^T = [1\ 0\ 0\ 10]^T$ and true parameters, $\theta_0 = [k_{1,0}\ k_{2,0}\ k_{3,0}]^T = [1\ 0.01\ 1]^T$. Clearly, the two datasets are very different. In Figure 5.5a, the single molecules of A converts to B via the excitation reaction which results in replication of B. In Figure 5.5b, the extinction branch is taken. These two trajectories demonstrate the random behavior called *phenotypic bifurcation* which cannot be explained by a deterministic formulation. A deterministic system of equations, no matter how complex or nonlinear, result in the same time-evolution trajectories when started with the same initial condition. Individually fitting the data in Figures 5.5a-5.5b using the deterministic formulation in Eq. (5.84) and least-squares minimization results in two completely different model fits as shown in Figures 5.6a-5.6b. Note that unlike the example in Section 5.3.1, the datasets in this example have very high measurement frequency. The results of these individual fits is provided in Tables 5.4-5.3. Firstly, the estimates for both the datasets are very different from the true values, unlike the previous example in Section 5.3.1. Secondly, the two datasets predict parameters which are very different from one other. Excitation dataset predicts that the excitation rate constant is high ($10^6$) while the extinction dataset predicts a low value (1.173). The same is true for the extinction rate constant.

This example shows that deterministic formulation does not provide reliable estimates. However, it may be considered that the unreliable estimation is an artifact of the using only one dataset. It may be expected that the estimation using a deterministic formulation may perform better if more data is available. To

Table 5.4: Extinction branch: True values and deterministic/least-squares estimates

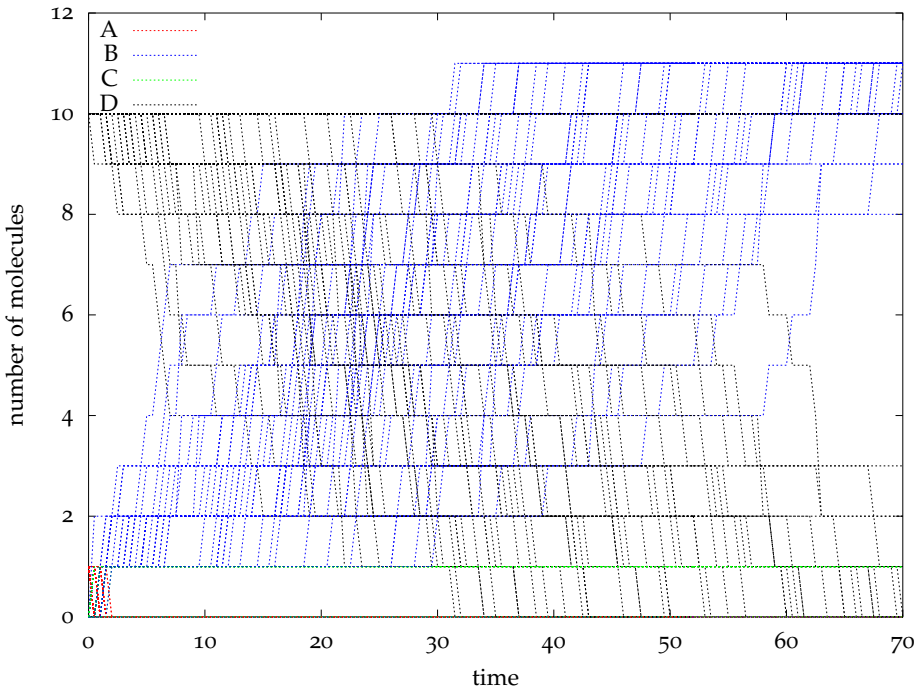
| Reactions | Parameters | True values $\theta_0$ | Least-squares estimates $\hat{\theta}_{\text{LS}}$ |
|---|---|---|---|
| Extinction | $k_1$ | 1 | 1.173 |
| Replication | $k_2$ | 0.01 | $10^{-6}$ |
| Extinction | $k_2$ | 1 | 327.29 |

Table 5.5: Averaged data: True values and deterministic/least-squares estimates

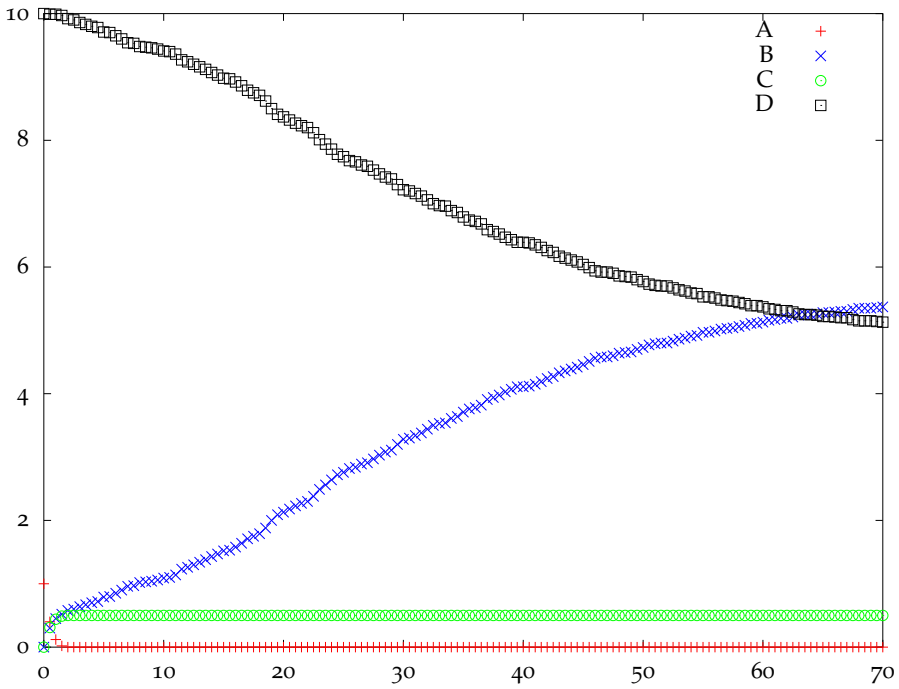
| Reactions | Parameters | True values $\theta_0$ | LS estimates $\hat{\theta}_{\text{LS}}$ | CDIS estimates $\hat{\theta}_{\text{CDIS}}$ |
|---|---|---|---|---|
| Extinction | $k_1$ | 1 | 46936 | 0.906 |
| Replication | $k_2$ | 0.01 | 0.004 | 0.0093 |
| Extinction | $k_2$ | 1 | 7029 | 0.906 |

address this reasoning, I estimate parameters using 100 datasets in the following exercise. Figure 5.7a shows 100 datasets generated using the same initial condition and true parameters as in the examples above. Figure 5.7b shows the averaged dataset obtained by averaging over the 100 datasets in Figure 5.7a. Note that there is a lot of variance in the 100 different trajectories. It turns out that 50 trajectories follow the excitation branch and the other 50 go extinct. Nevertheless, the average trajectory in Figure 5.7b appears to be smooth with very little "stochasticity".

Parameter estimates obtained using the averaged data and the deterministic formulation are shown in Table 5.5. The corresponding model fit is shown in Figure 5.8. Clearly, the model does not capture the behavior of the averaged dataset. As a result of least-squared minimization, the estimated parameters, $\hat{\theta}_{\text{LS}}$ are very different from the true parameters, $\theta_0$. As a comparison, the parameter estimates obtained using the 100 datasets in Figure 5.7a and the CDIS method (Section 6.1), $\hat{\theta}_{\text{CDIS}}$, are also provided in Table 5.5. Since the CDIS method uses the exact stochastic chemical kinetic model, the parameters obtained are close to the true values.




(a) 100 measurement datasets using the same initial condition




(b) Averaged measurement data over 100 datasets

Figure 5.7: Measurement datasets obtained using $\theta_0 = [1\ 0.01\ 1]^T$ and initial conditions, $\mathbf{x}(0) = [A_0\ B_0\ C_0\ D_0]^T = [1\ 0\ 0\ 10]^T$

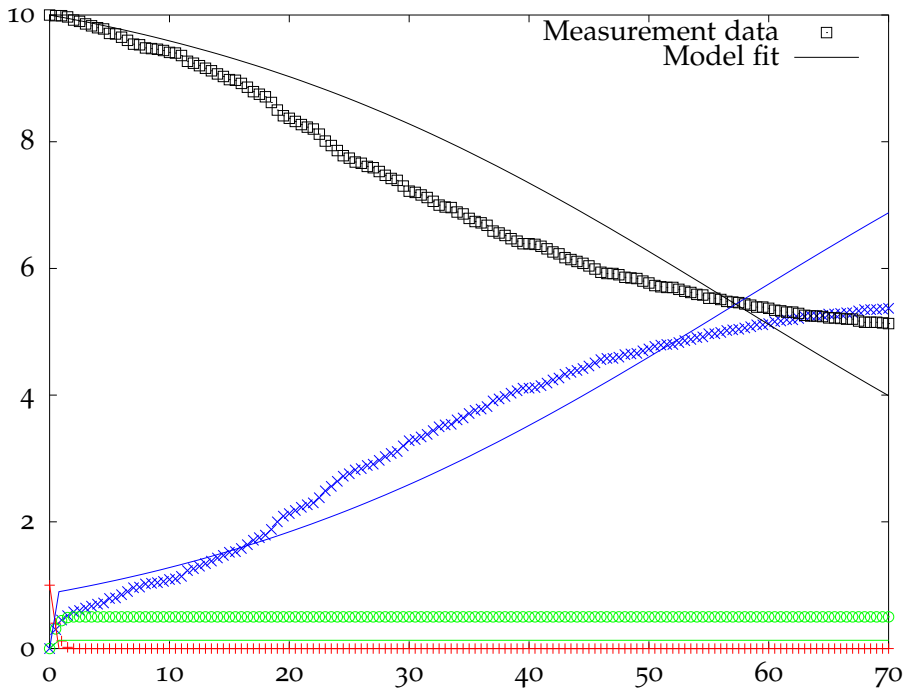



Figure 5.8: Model fit for the measurement data in Figure 5.7b

This example shows that estimation using deterministic formulation does not yield reliable parameter estimates but the CDIS method does.

### 5.3.3 *Necessary conditions to use deterministic formulation*

As demonstrated in Section 5.3.2, the deterministic formulation does not necessarily provide the correct parameter estimates when the measurement data is obtained from a stochastic kinetic model. There a few things to note here. Firstly, the (simulated) measurement data used Section 5.3.2 was free of measurement errors. The experimental methods have not yet reached a level of accuracy to provide such error-free measurements. Secondly, 100 datasets with full measurements were used. Thirdly, there was no model-experiment mismatch, *i. e.*, exactly same reaction kinetics were used as the model and for generating simulated measurement data. In spite of these advantages, the deterministic formulation was not

able to provide (point) estimates that were even remotely close to the true values.

It is informative to discover the necessary conditions under which the deterministic formulation may yield the correct (point) estimates. The example shown in Section 5.3.2 displays *stochastic bifurcation* [85, p. 168] while the example in Section 5.3.1 does not. It so happens that for the example shown in Section 5.3.2, the average behavior of the stochastic model does not equal to the deterministic formulation of the same reactions. As a result, when the number of molecules is small, then even with infinite data, the parameter estimates obtained from the deterministic model do not equal the parameter estimates using the stochastic formulation. Thus, the necessary condition to be able to use the deterministic formulation is that $\mathbb{E}[\text{stochastic formulation}] = \text{deterministic formulation}$. In other words, the averaged behavior of stochastic formulation should be the same as the deterministic behavior.

## 5.4 ESTIMATION USING MCMC AND UNIFORMIZATION

This section begins the use of *simulation methods* for parameter inference using a Bayesian framework. Sections 5.4, 5.5, and 6.1 are the three simulation methods presented in this dissertation. Section 5.4 describes a method which I call MCMC-Unif (abbreviation for Markov Chain Monte Carlo with uniformization) and Section 5.5 describes the MCMC-MH method (abbreviation for Markov Chain Monte Carlo with Metropolis-Hastings). As their titles suggest, Sections 5.4 and 5.5 have a common underlying theme of employing *Markov Chain Monte Carlo* (MCMC) procedure with the difference only in the use of an embedded simulation method. MCMC-Unif (Section 5.4) uses an *endpoint-conditioned* simulation method called *uniformization* while MCMC-MH (Section 5.5) uses an approximate simulation method within a *Metropolis-Hastings* framework. In fact, the MCMC-MH method was developed first by Wilkinson and others (Boys et al. [14], Henderson et al.

[50], Golightly and Wilkinson [38, 37, 39], Wilkinson [119]). Later, Choi and Rempala [21] replaced the Metropolis-Hastings step used by Boys et al. [14] with the uniformization technique of Hobolth and Stone [56] to produce MCMC-Unif method which is described in this section. Choi and Rempala [21] call their algorithm "MCMC–Gibbs sampler" and present it rather informally. A Gibbs sampling algorithm is a specific MCMC-type algorithm and the prominent feature that differentiates the "MCMC–Gibbs sampler" from the MCMC-MH algorithm is the use of uniformization technique. Thus, in this dissertation, I choose to refer to their "MCMC–Gibbs sampler" as the MCMC-Unif method and I present it, rather formally, in a notation that is consistent with other parameter estimation methods (for stochastic chemical kinetic models) in this dissertation. In spite of the chronological development of these methods, MCMC-Unif is easier to understand when compared to MCMC-MH, and is therefore presented first. The two methods — MCMC-Unif (Section 5.4) and MCMC-MH (Section 5.5) — belong to the class of *MCMC methods*. A short history of MCMC methods may be found in Robert and Casella, 2011 [92], Richey, 2010 [91].

### 5.4.1 *Gibbs Sampling Algorithm*

A description of Gibbs sampling is required in order to understand the MCMC-Unif method. Gibbs sampling allows us to sample from a joint multivariate distribution using only the conditional distributions (see Wilkinson [119, p. 255] for a simple example). Gibbs sampling was first proposed by Geman and Geman, 1984 [30] who used it to study image-processing models. Since then, Gibbs sampling has become a very popular MCMC method with many variations. It has been used very successfully for multiple (DNA and other) sequence alignment and weak motif detection (Lawrence et al., 1993 [62], Liu et al., 1995 [65]) with over 500 relevant research articles on PubMed [2]. A primer on Gibbs sampling

containing theory and examples may be found in Casella and George, 1992 [18]. Smith and Roberts, 1993 [101] provide a review of Gibbs sampling and other MCMC methods. Computer software specially dealing with MCMC techniques is also available [66]. An extremely introductory exposition with examples and a discussion of practical implementation issues may be found in Resnik and Hardisty, 2010 [89].

Instead of describing a general Gibbs sampling algorithm, I only describe it as it applies to the problem of parameter estimation in the given framework. For both MCMC methods (MCMC-Unif and MCMC-MH) discussed in this dissertation have a common basic premise — instead of attempting to estimate the measurement-data posterior, $\pi(\theta \mid y)$ these MCMC methods attempt to obtain the *full posterior* distribution, $\pi(\theta, x \mid y)$. The full posterior may be written in terms of the conditionals as

$$\pi(\theta, x \mid y) = \pi(\theta \mid x, y)\pi(x \mid y) \tag{5.85}$$

$$\pi(\theta, x \mid y) = \pi(x \mid \theta, y)\pi(\theta \mid y) \tag{5.86}$$

Note that since $X \subseteq Y$ (from Proposition 5.1), the event $\{X,\ Y\} = \{X \cap Y\}$ is the same as the event $\{X\}$. Hence,

$$\pi(\theta \mid x, y) = \pi(\theta \mid x) \tag{5.87}$$

in which, $\pi(\theta \mid x)$ is the complete-data posterior which is known analytically in Eqs. (5.21)-(5.23). A basic Gibbs sampling algorithm is provided in Algorithm 5.1.

Algorithm 5.1 iteratively generates samples of $\theta$ and $x$ from the conditional distributions $\pi(\theta \mid x)$ and $\pi(x \mid \theta, y)$, respectively. These samples of $\theta$ and $x$ comes from the true full posterior, $\pi(\theta, x \mid y)$ and therefore, provide a way to indirectly sample from the full posterior using only the conditionals. The samples of $\theta \in \mathbb{R}^{n_r}$

**Algorithm 5.1** Basic Gibbs Sampling for $\pi(\theta, x \mid y)$

1: Given measurement-data $y$
2: Sample $x \sim \pi(x \mid y)$ ▷ Initialize $x$
3: **repeat**
4: Sample $\theta \sim \pi(\theta \mid x)$ using current $x$ ▷ $\pi(\theta \mid x)$ known analytically
5: Sample $x \sim \pi(x \mid \theta, y)$ using current $\theta$
6: Store $\theta$ as a sample
7: **until** $N_s$ samples of $\theta$ are obtained
8: Generate histogram of $\pi(\theta \mid y)$ by binning over $N_s$ samples of $\theta$

may be binned to create an $n_r$-dimensional histogram representing the posterior, $\pi(\theta \mid y)$. Further, every component of $\theta \in \mathbb{R}^{n_r}$, $k_i$, $i = 1, 2, \ldots, n_r$, may also be binned to create a one-dimensional histogram representing the marginal posterior, $\pi(k_i \mid y)$. We also obtain samples of $x$ from the marginal posterior, $\pi(x \mid y)$. The samples $(\theta, x)$ form a Markov chain (in discrete time) whose stationary distribution is the full posterior $\pi(\theta, x \mid y)$ and hence the name Markov chain Monte Carlo. Note that this Markov chain is already in it's stationary distribution, *i.e.*, even when only $N_s = 1$ sample is generated using Algorithm 5.1, that sample of $(\theta, x)$ comes from the true distribution $\pi(\theta, x \mid y)$. In other words, if the samples of $(\theta, x)$ are collected by running Algorithm 5.1 (with $N_s = 1$) 100 times, then the histogram obtained by binning over these 100 samples would truly represent the full posterior. The discrete-time, continuous-state Markov chain in $(\theta, x)$ should not be confused with the continuous-time, discrete-state Markov chain in $\mathbf{x}(t)$ (Section 5.2.2) formed by stochastic chemical reactions.

This basic algorithm has two major issues. Firstly, note that the initialization step requires a sample $x$ from the marginal posterior $\pi(x \mid y)$, which is not available. Secondly, even if initialized, the algorithm samples $x$ iteratively from the conditional $\pi(x \mid \theta, y)$. This conditional is very difficult to sample from.

The first issue may be resolved by initializing the Markov chain using any

arbitrary but feasible $x$. In other words, initialize the Markov chain using any $x = x_0$ in the support of $\pi(x \mid y)$, *i.e.*, $\pi(x_0 \mid y) > 0$. In terms of the stochastic chemical kinetic model, the algorithm may be initialized using any valid complete trajectory, $x_0$, which is consistent with the measurement data $y$. The initial sample $x_0$ need not come from any statistical distribution, it can be set manually (while maintaining consistency with $y$). Obviously, when the initialization is not performed using the true marginal $\pi(x \mid y)$, the resulting samples $(\theta, x)$ obtained using Algorithm 5.1 are no longer distributed as $\pi(\theta, x \mid y)$. However, these samples, $(\theta, x)$, still form a discrete-time, continuous-state Markov Chain whose transition kernel is given by

$$\begin{aligned} p(\theta, x, \theta', x') &= \pi(\theta', x' \mid \theta, x, y) \\ &= \pi(x' \mid \theta', \theta, x, y)\ \pi(\theta' \mid \theta, x, y) \\ &= \pi(x' \mid \theta', y)\ \pi(\theta' \mid x) \end{aligned} \tag{5.88}$$

It may be shown that this Markov chain in $(\theta, x)$ has a stationary distribution equal to the full posterior, $\pi(\theta, x \mid y)$ (Wilkinson [119, p. 256]). Further, under mild conditions, it may also be shown that the Markov chain converges to its stationary distribution (Tierney, 1994 [106], Smith and Roberts, 1993 [101], Roberts and Smith, 1994 [93]) . These convergence criterion and proofs ([101, Theorem 1], [106, Theorem 1]) are not discussed in this dissertation and the required convergence is assumed. Thus, when not initialized from the true distribution, the Markov chain may be run to equilibrium to achieve (approximate) stationarity and samples of $(\theta, x)$ may be obtained. These samples will effectively be distributed as the full posterior and may be binned as described before to obtain a histogram of the required posterior, $\pi(\theta \mid y)$. Ensuring convergence of the Markov chain is not trivial and the chain must be monitored for convergence during simulation by looking at the trace and autocorrelation plots. The first $N_b$ samples generated by

the Markov chain are discarded as *burn-in* required to achieve convergence. After the burn-in period, if the the Markov chain "appears" converged, more samples are simulated which are then used to generate histogram posterior(s).

The second issue may be solved in two ways. In specific cases, it may be possible to exactly sample from the conditional, $\pi(x \mid \theta, y)$, using *endpoint-conditioned methods*. The use of a specific endpoint-conditioned simulation method, called *uniformization*, constitutes the MCMC-Unif method. The other way to effectively sample from the conditional, $\pi(x \mid \theta, y)$, is by using a *Metropolis-Hastings* step, which produces the MCMC-MH method. I discuss the endpoint-conditioned simulation methods (specifically uniformization) in Section 5.4.3. In order to use endpoint-conditioned simulation methods, the entire complete-data trajectory $x$ has to broken down into intervals, which is allowed because of the Markov property (discussed next in Section 5.4.2) of the stochastic chemical kinetic models.

### 5.4.2 *Markov property*

The conditional, $\pi(x \mid \theta, y)$ contains the entire complete-data trajectory, $x$, and the entire measurement data, $y$. Recall that $y$ is composed of measurements, $\{y_0, y_1, \ldots, y_m\}$ taken at discrete times, $\{s_0, s_1, \ldots, s_m\}$.

**Definition 5.2** (Complete-data interval)**.** *Complete-data interval, $X_{[s_i,s_j]}$, $s_i \leq s_j$, is a set of random variables, defined as*

$$X_{[s_i,s_j]} = \{\mathbf{X}(t) : t \in [s_i, s_j]\}$$

*A sample of $X_{[s_i,s_j]}$ is denoted by $x_{[s_i,s_j]}$.*

A complete-data trajectory, $x$, defined over the time interval $[s_0, s_m]$ is the same as $x_{[s_0,s_m]}$. The set of random variables $X_{[s_0,s_m]}$ may be divided into smaller intervals

as

$$X_{[s_0,s_m]} = X_{[s_0,s_{m-1}]} \cap X_{[s_{m-1},s_m]} \tag{5.89}$$

and the conditional, $\pi(x \mid \theta, y)$, may be manipulated as

$$\begin{aligned}
\pi(x \mid \theta, y) &= \pi(x_{[s_0,s_m]} \mid \theta, y) \\
&= \pi(x_{[s_0,s_{m-1}]}, x_{[s_{m-1},s_m]} \mid \theta, y) \\
&= \pi(x_{[s_{m-1},s_m]} \mid x_{[s_0,s_{m-1}]}, \theta, y)\ \pi(x_{[s_0,s_{m-1}]} \mid \theta, y)
\end{aligned} \tag{5.90}$$

Using the Markov property (Eq. (5.50)) of the continuous-time, discrete-state Markov chain [4],

$$\pi(x_{[s_{m-1},s_m]} \mid x_{[s_0,s_{m-1}]}, \theta, y) = \pi(x_{[s_{m-1},s_m]} \mid \mathbf{x}(s_{m-1}), \theta, y_m) \tag{5.91}$$

Substituting Eq. (5.91) into Eq. (5.90) and repeating the same procedure for other intervals,

$$\pi(x \mid \theta, y) = \prod_{i=1}^{m} \pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \theta, y_i) \tag{5.92}$$

Thus, given the initial condition, $\mathbf{x}_0 = \mathbf{x}(s_0)$, a sample of the complete-data trajectory, $x \sim \pi(x \mid \theta, y)$, may be generated by sequentially sampling the complete-data intervals, $x_{[s_{i-1},s_i]}$, $i = 1, 2, \ldots, m$, starting from $i = 1$. The Markov property, therefore, allows us to break the entire complete-data trajectory into intervals. The algorithm to generate samples of the complete-data trajectory, $x$, from the conditional, $\pi(x \mid \theta, y)$ is presented in Algorithm 5.2.

When full measurements (see Section 5.2.1) are available, then, Eq. (5.92) may

[4] For a continuous-time Markov chain, this is a non-trivial calculation which is not presented in this dissertation

**Algorithm 5.2** Sample $x \sim \pi(x \mid \theta, y)$

1: Given $\mathbf{x}_0$, $y$, and $\theta$
2: Initialize $x$ as an empty variable
3: **for** $i = 1$ to $m$ **do**
4: Sample $x_{[s_{i-1},s_i]} \sim \pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \theta, y_i)$
5: $x \leftarrow x \cap x_{[s_{i-1},s_i]}$ ▷ Stitch intervals together
6: **end for**

be re-written as

$$\pi(x \mid \theta, y) = \prod_{i=1}^{m} \pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \mathbf{x}(s_i), \theta) \tag{5.93}$$

### 5.4.3 *Endpoint-conditioned simulation methods*

For a continuous-time, discrete-space Markov chains, *endpoint-conditioned simulation methods* (Hobolth and Stone, 2009 [56]) generate samples of $x_{[s_{i-1},s_i]}$ from the conditional distribution $\pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \mathbf{x}(s_i), \theta)$. As the name indicates, the conditional distribution. $\pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \mathbf{x}(s_i), \theta)$, is conditioned not only on the initial condition $\mathbf{x}(s_{i-1})$ and parameter vector $\theta$ but also on the *endpoint* $\mathbf{x}(s_i)$. Contrast this simulation with the previously discussed *forward simulation* of a complete-data trajectory $x \sim \pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \theta)$ using the stochastic simulation algorithms (SSAs), in Section 2.3. Both endpoint-conditioned and forward simulations generate complete-data trajectory, $x = x_{[s_{i-1},s_i]}$, using the initial condition and parameters. While numerous methods (both exact and approximate) are available for the forward simulation, endpoint-conditioned simulation methods are fewer and less advanced. Further, the probability density, $\pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \theta)$, is essentially the complete-data likelihood in Eq. (5.13) and Eqs. (5.15)-(5.18). Thus, in general for any stochastic chemical kinetics, the expression, $\pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \theta)$, may be evaluated for any value of $x_{[s_{i-1},s_i]}$. Such a luxury is not available for the endpoint-conditioned probability density, $\pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \mathbf{x}(s_i), \theta)$. At the time of writing this thesis, the endpoint-conditioned density may only be evaluated for

specific cases, as discussed next.

Sophisticated endpoint-simulation methods that work for any continuous-time, discrete-state space Markov chain are not available. I will discuss a set of endpoint-conditioned simulation methods which are applicable if the following conditions are met

1. Homogeneous Markov chain in continuous time and discrete state
2. (Discrete) State space is finite
3. Markov chain is irreducible and positive recurrent (*i. e.* a stationary distribution exists)

Many methods for generating samples from the endpoint-conditioned density have been proposed. The most intuitive method is the *naive rejective sampling* method (Blackwell, 2003 [9]) , in which $x$ is simulated using forward simulation (conditioned only on initial condition and parameters). This (tentative) sample is then rejected if it does not agree with the endpoint $\mathbf{x}(s_i)$ and the forward simulation is repeated. The iteration stops when a sample $x$ is obtained that agrees with the endpoint. Note that the naive rejection sampling method is applicable to any general stochastic kinetics which do not meet the conditions specified above. However, naive rejection has an extremely low probability of acceptance and therefore, is computationally prohibitive (Hobolth and Stone, 2009 [56]). Better methods than the naive method, available under the conditions listed above, are *modified rejection sampling* (Nielsen, 2002 [76]) , *direct sampling* (Hobolth, 2008 [55]) , *uniformization* (Jensen [59], Fearnhead and Sherlock [27], Hobolth and Stone [56], Ross [95, p. 282]) , and *bisection sampling* (Asmussen and Hobolth, 2008 [7]) . An excellent review and in-depth comparison of the uniformization and rejection, modified rejection and direct sampling methods is provided by (Hobolth and Stone, 2009 [56]). Hobolth and Stone [56] demonstrate that, in general, no

one method dominates the others.

The conditions for application of these methods are similar to those required for the application of the exact method. This is because both endpoint-conditioned simulation methods and the exact method uses the transition rate matrix, $\mathbf{Q}$. In fact, the direct sampling method also requires matrix exponentiation of $\mathbf{Q}$. The limitations mentioned for the exact method in Section 5.2 are also true for the endpoint-conditioned simulation methods which limits the application of MCMC-Unif method. Another important point to note is that the endpoint-conditioned methods described here are developed for a different type of Markov chain that are used to study mutations in DNA sequences (Nielsen [76], Hobolth [55], Fearnhead and Sherlock [27], Hobolth and Stone [56]). Since these methods are not developed with the stochastic chemical kinetic model in mind, they fail to use the inherent structure of this type of Markov chain. This indicates that there is a scope for very useful research in this field.

In this dissertation, I only use the uniformization method to simulate $x$ from the endpoint-conditioned density $\pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \mathbf{x}(s_i), \theta)$. See Hobolth and Stone [56, Algorithm 5] for details. Other endpoint-conditioned algorithms may be used to produce similar methods, for example, MCMC-Direct and MCMC-Bisection.

### 5.4.4 *MCMC-Unif Algorithm*

The MCMC-Unif algorithm may be easily constructed using the Gibbs sampling algorithm with appropriate initialization method, as described in Section 5.4.1, and the uniformization algorithm in Section 5.4.3. Algorithm 5.3 presents the MCMC-Unif method.

**Algorithm 5.3** MCMC with Uniformization (MCMC-Unif)

1: Given measurement-data $y$
2: Initialize $x$ using Algorithm 5.4
3: **repeat**
4: Sample $\theta \sim \pi(\theta \mid x)$ using current $x$ ▷ $\pi(\theta \mid x)$ known analytically
5: Sample $x \sim \pi(x \mid \theta, y)$ using Algorithm 5.2 with Uniformization
6: Store $\theta$ as a sample
7: **until** $N_s$ samples of $\theta$ are obtained
8: Remove $N_b \leq N_s$ samples as burn-in
9: Generate histogram of $\pi(\theta \mid y)$ by binning over $(N_s - N_b)$ samples of $\theta$

**Algorithm 5.4** Initialize $x$

1: Given measurement-data $y$
2: Set $\theta$ as the mode of prior distribution in Eq. (5.1)
3: Sample $x \sim \pi(x \mid \theta, y)$ using Algorithm 5.2 using Uniformization

## 5.5 ESTIMATION USING MCMC AND METROPOLIS-HASTINGS

This chapter describes a parameter estimation method called the Markov Chain Monte Carlo with Metropolis-Hastings (MCMC-MH). As mentioned in the introduction to this dissertation and in Section 5.4, MCMC-MH method belongs to the class of *simulation methods*, specifically *MCMC methods* (Robert and Casella, 2011 [92], Richey, 2010 [91]). MCMC-MH method was developed by Wilkinson and others (Boys et al. [14], Henderson et al. [50], Golightly and Wilkinson [38, 37, 39], Wilkinson [119]). The development of this section continues from the previous sections, specifically from Section 5.4.1 on Gibbs sampling. As mentioned in Section 5.4.1, samples of $x$ need to be generated from the conditional distribution, $\pi(x \mid \theta, y)$. In limited cases, endpoint-conditioned simulation methods may be used. However, in most examples of interest, the current endpoint-conditioned methods are not applicable. Another method to generate samples from of $x$ from $\pi(x \mid \theta, y)$ is using *Metropolis-Hastings* sampling which is discussed next.

### 5.5.1 *Metropolis-Hastings Algorithm*

Metropolis-Hastings is a Markov chain Monte Carlo sampling method first proposed by Metropolis et al., 1953 [71] for performing equations of state calculations and later generalized by Hastings, 1970 [49]. Since it's introduction, and especially after the convincing paper by Gelfand and Smith, 1990 [29], Metropolis-Hastings (and other MCMC methods) have been used extensively in many fields. Among these fields, a special mention is required for statistical physics literature in which the use Metropolis-Hastings (and MCMC) sampling is ubiquitous (Krauth [61, p.21], Hammersley and Handscomb [44, p. 117], Handscomb [45], Honorkamp [57, p. 413]). An introductory exposition to the Metropolis-Hastings algorithm may be found in Chib and Greenberg, 1995 [20] and Wilkinson [119, p. 264].

Instead of describing a general Metropolis-Hastings sampling method, I only describe the sampling method as it applies to the problem at hand. As mentioned in Section 5.4.1, the MCMC-MH method attempts to obtain the full posterior distribution, $\pi(\theta, x \mid y)$. Further, the underlying algorithm to obtain $\pi(\theta, x \mid y)$ is the same Gibbs sampling algorithm that has already been described in Algorithm 5.1. The main difference between the MCMC-Unif algorithm and MCMC-MH method is in the generation of samples $x$ from the conditional distribution, $\pi(x \mid \theta, y)$. As discussed in Section 5.4.1, MCMC-Unif employs the uniformization method to generate samples of $x$ from the "true" conditional distribution, $\pi(x \mid \theta, y)$ while the MCMC-MH method generates $x$ "approximately" using Metropolis-Hastings (MH) sampling.

It is usually not possible to sample from the true conditional distribution, $\pi(x \mid \theta, y)$ (see Section 5.4.3). Such a condition is aptly suited to the application of Metropolis-Hastings algorithm in which samples of $x$ may be generated

iteratively from a *proposal function* and then accepted or rejected based on an acceptance probability. The samples of $x$, so generated, form a Markov chain with a stationary distribution equal to the target distribution, $\pi(x \mid \theta, y)$. Thus, MH sampling, by itself, is a MCMC method.

However, it is difficult to create a proposal function, that can directly sample the complete-data trajectory, $x$, in its entirety. Instead, a proposal function has been developed (by Wilkinson and others [119, 14]) to sample only the complete-data intervals, $x_{[s_{i-1},s_i]}$, $i = 1, 2, \ldots, m$. The Markov property (Section 5.4.2) of the stochastic chemical kinetic model, then, allows us to construct the entire path, $x$, by "stitching" the complete-data intervals together. Thus, the MH method is essentially applied to generate samples of $x_{[s_{i-1},s_i]}$, $i = 1, 2, \ldots, m$. The proposal function, $q(x^{\star}_{[s_{i-1},s_i]} \mid x_{[s_{i-1},s_i]})$, generates "new" samples, $x^{\star}_{[s_{i-1},s_i]}$, using the "old" samples, $x_{[s_{i-1},s_i]}$.

The details of the MH step is available in Wilkinson, 2012 [119], Boys et al., 2008 [14]. In Algorithm 5.5, I provide a general algorithm to generate samples of $x_{[s_{i-1},s_i]}$ for a given $i \in \{1, 2, \ldots, m\}$. Algorithm 5.6 provides the algorithm to generate sample of the entire complete-data path $x$.

### 5.5.2 *MCMC-MH Algorithm*

The MCMC-MH method is a *Metropolis-within-Gibbs* style overall algorithm. The details may be found in Wilkinson, 2012 [119], Boys et al., 2008 [14]. I provide the entire MCMC-MH algorithm in Algorithm 5.7.

**Algorithm 5.5** MH step to sample $x^{\star}_{[s_{i-1},s_i]} \sim \pi(x^{\star}_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \theta, y_i)$

1: Given $i$, $\mathbf{x}(s_{i-1})$, $\theta$, $y_i$ and $x_{[s_{i-1},s_i]}$
2: **repeat**
3: Sample $x^{\star}_{[s_{i-1},s_i]} \sim q(x^{\star}_{[s_{i-1},s_i]} \mid x_{[s_{i-1},s_i]})$
4: Compute acceptance ratio, A

$$A = \frac{\pi(x^{\star}_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \theta, y_i)\; q(x_{[s_{i-1},s_i]} \mid x^{\star}_{[s_{i-1},s_i]})}{\pi(x_{[s_{i-1},s_i]} \mid \mathbf{x}(s_{i-1}), \theta, y_i)\; q(x^{\star}_{[s_{i-1},s_i]} \mid x_{[s_{i-1},s_i]})} \tag{5.94}$$

5: Calculate acceptance probability, $\alpha(x_{[s_{i-1},s_i]}, x^{\star}_{[s_{i-1},s_i]})$

$$\alpha(x_{[s_{i-1},s_i]}, x^{\star}_{[s_{i-1},s_i]}) = \min\{1, A\} \tag{5.95}$$

6: Accept $x^{\star}_{[s_{i-1},s_i]}$ with probability $\alpha(x_{[s_{i-1},s_i]}, x^{\star}_{[s_{i-1},s_i]})$; else reject
7: **until** Acceptance

**Algorithm 5.6** Overall MH step to sample $x^{\star} \sim \pi(x^{\star} \mid \theta, y)$

1: Given $\mathbf{x}_0$, $\theta$, $y$ and $x$
2: Initialize $x^{\star}$ as an empty variable
3: **for** $i = 1$ to $m$ **do**
4: Sample $x^{\star}_{[s_{i-1},s_i]}$ using Algorithm 5.5
5: $x^{\star} \leftarrow x^{\star} \cap x^{\star}_{[s_{i-1},s_i]}$ ▷ Stitch intervals together
6: **end for**

**Algorithm 5.7** MCMC with Metropolis-Hastings (MCMC-MH)

1: Given measurement-data $y$
2: Initialize $x$ using Algorithm 5.8 or Algorithm 5.4
3: **repeat**
4: Sample $\theta \sim \pi(\theta \mid x)$ using current $x$ ▷ $\pi(\theta \mid x)$ known analytically
5: Sample $x \sim \pi(x \mid \theta, y)$ using Algorithm 5.6
6: Store $\theta$ as a sample
7: **until** $N_s$ samples of $\theta$ are obtained
8: Remove $N_b \leq N_s$ samples as burn-in
9: Generate histogram of $\pi(\theta \mid y)$ by binning over $(N_s - N_b)$ samples of $\theta$

**Algorithm 5.8** Initialize $x$

```
Given measurement-data y
for i = 1 to m do
    Either set x_[s_{i-1}, s_i] arbitrarily, or
    procedure
        Set θ as the mode of prior distribution in Eq. (5.1)
        Solve a linear integer program [119, p. 289], and/or
        Simulate the approximate process [119, p. 289]
    end procedure
end for
```

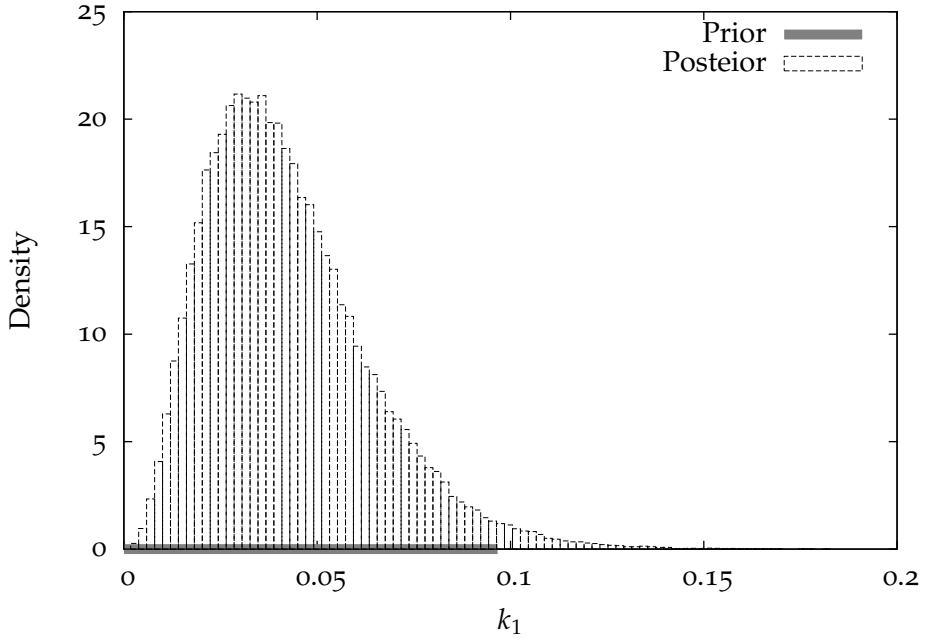
Table 5.6: Parameter true values, MCMC-MH estimates and prior parameters

| Reactions | Parameters | True values | MAP estimates | Prior Parameters | | |
|---|---|---|---|---|---|---|
| | | $\theta_0$ | $\hat{\theta}_{\text{MCMC-MH}}$ | $\hat{\theta}_{\text{prior}}$ | $a$ | $b$ |
| Reaction 1 | $k_1$ | 0.04 | 0.0295 | 15 | 1.01 | 0.00067 |
| Reaction 2 | $k_2$ | 0.11 | 0.150 | 8 | 1.01 | 0.00125 |
| Time taken ($N_s = 10^5$) = 11,616 seconds | | | | | | |

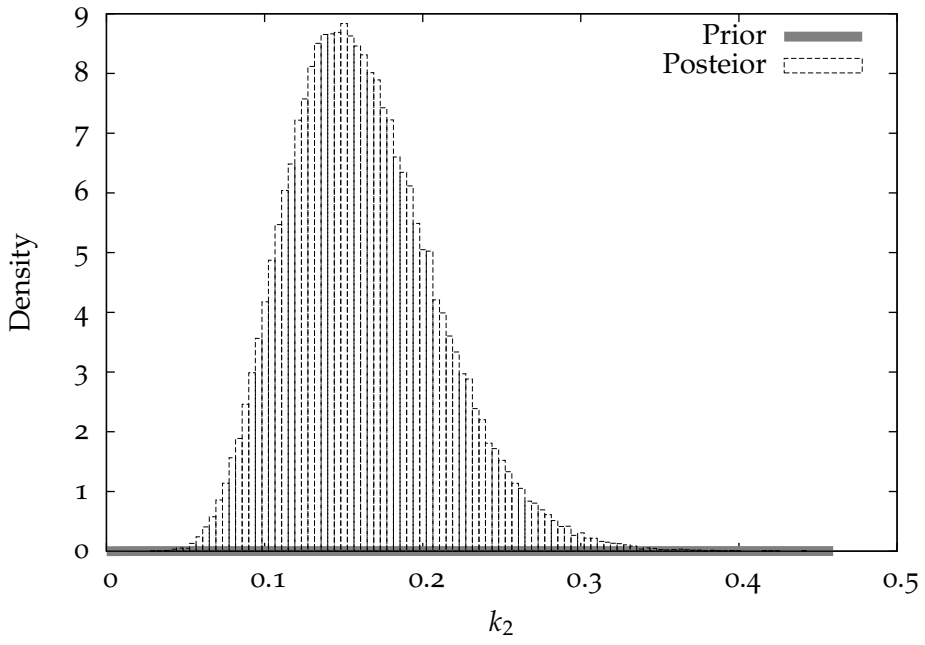
### 5.5.3 *A Simple Example*

I demonstrate the application of MCMC-MH method using the same example and measurement data in Figure 5.1. Figures 5.9a-5.9b show the marginal priors and posteriors for both parameters. The MCMC-MH procedure was used with $N_s = 10^5$ samples. A burn-in of 10%, *i.e.*, $10^4$ samples was used. The remaining 90,000 samples were used to generate the histograms in Figures 5.9a-5.9b. The parameter estimates obtained using histogram estimation are presented in Table 5.6. The same gamma prior was used as in Section 5.2.3.

Comparing the parameter estimates using the exact method (in Table 5.1) and the MCMC-MH ($N_s = 10^5$) method, it appears that even with $10^5$ samples, the MCMC-MH estimate has still not converged to the exact estimate. Table 5.6 also shows the computational, in seconds, time taken for sampling. Figure 5.10 shows the joint posterior obtained by binning the samples of $\theta$ in two dimensions.




(a) Marginal prior, $\pi(k_1)$ and marginal posterior, $\pi(k_1 \mid y)$




(b) Marginal prior, $\pi(k_2)$ and marginal posterior, $\pi(k_2 \mid y)$

Figure 5.9: Marginal priors and posteriors obtained using MCMC-MH method with $N_s = 10^5$.

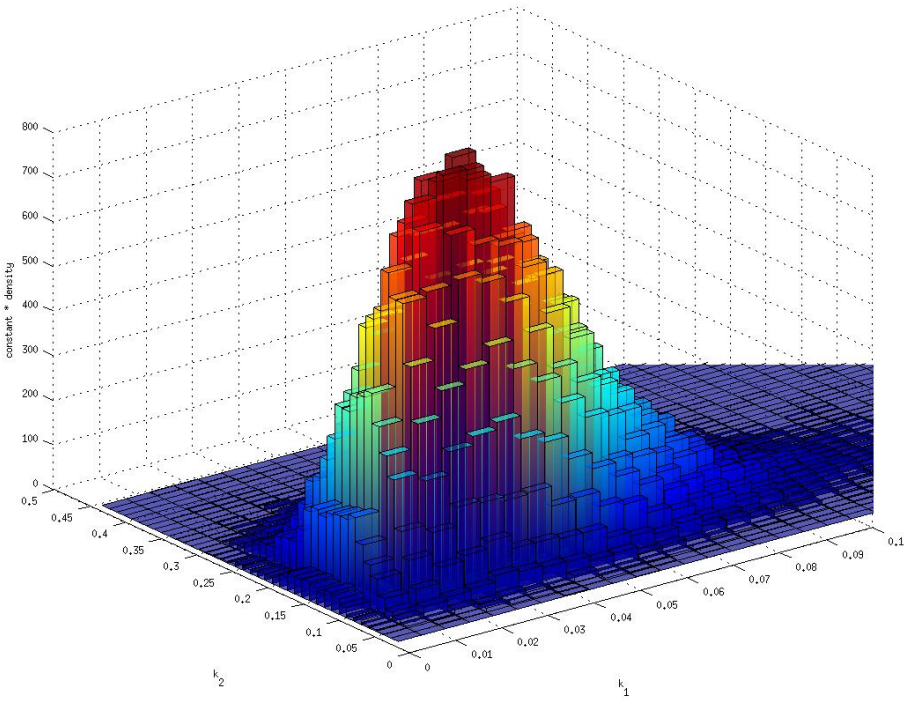



Figure 5.10: Joint posterior obtained using MCMC-MH method with $N_s = 10^5$.

# 6

# NEW METHODS FOR PARAMETER ESTIMATION IN STOCHASTIC CHEMICAL KINETIC MODELS

In this chapter, I present two new classes of methods for estimating parameters in stochastic chemical kinetic models. The first class of methods, described in Section 6.1, is based on importance sampling, which I call the *conditional density importance sampling* (CDIS) method. The CDIS method of parameter estimation belongs to the class of simulation methods. Similar to the two other simulation methods (MCMC-Unif and MCMC-MH) discussed in Sections 5.4 and 5.5 respectively, the CDIS method of parameter estimation also generates samples of the complete-data trajectory, $x$. However, the CDIS method, as shown in this chapter, uses a different conditional distribution than the MCMC methods. Importance sampling is used to generate a *semi-analytical* (approximate) expression for the coveted posterior distribution, $\pi(\theta \mid y)$. All the three simulation methods, MCMC-Unif, MCMC-Mh and CDIS, have the property that the posteriors generated by these methods converge to the true exact posterior in the limit of infinite samples of $x$. The second class of methods, described in Section 6.2, is called *approximate direct methods*. This class of method is developed based on the insights obtained from the expressions of complete-data posterior, $\pi(\theta \mid x)$ and measurement-data (CDIS) posterior,

$\hat{\pi}_{\text{CDIS}}(\theta \mid x)$. Specifically, approximate direct methods exploit the fact that the variables $r_i = r_i(x)$ and $G_i = G_i(x)$, $i = 1, 2, \ldots, n_r$ form *sufficient statistics* for these posteriors. As the name suggests, approximate direct methods do not generate samples of $x$; instead, the sufficient statistics are either computed directly from the data (without any sampling whatsoever) or the sufficient statistics are directly sampled (instead of through $x$). As a result, the approximate direct methods have significant advantage over the simulation methods. However, as of this dissertation, the approximate direct methods do not guarantee convergence to the true exact posterior.

This chapter continues with the notation in Chapter 5.

## 6.1 ESTIMATION USING IMPORTANCE SAMPLING

Importance sampling is a very common sampling method to compute integrals when direct Monte Carlo sampling of the intended *target distribution* is not possible. Importance sampling is a tool as ubiquitous as MCMC methods. In fact, importance sampling finds a place in almost every problem that has an MCMC based solution. A review of importance sampling methods may be found in Smith et al., 1997 [102], Tokdar and Kass, 2010 [109]. A primer on importance sampling with examples may be found in Rawlings and Mayne [86, p. 316].

The importance sampling based parameter estimation methods described in this section are similar in spirit to the *data augmentation*(DA) algorithms by Tanner and Wong, 1987 [104], *sampling-importance-resampling* (SIR) algorithm by Rubin, 1987 [96] and especially, the *poor man's data augmentation* (PMDA) algorithms by Wei and Tanner, 1990 [116]. Gelfand and Smith, 1990 [29] provide an excellent comparison of the three sampling methods – Gibbs sampling, data augmentation and sampling-importance-resampling algorithms. A more recent review of some

of these ideas may be found in Fearnhead, 2008 [26].

These previous methods (DA, SIR, PMDA) may not be applied as trivially as it appears. The DA algorithm by Tanner and Wong, 1987 [104] is an iterative procedure to obtain posteriors which requires sampling $x$ from the conditional distribution, $\pi(x \mid \theta, y)$. As discussed in Chapter 5, sampling from $x \sim \pi(x \mid \theta, y)$ is indeed the major obstacle in the entire estimation procedure. The SIR algorithm by Rubin, 1987 [96] eliminates the iterative step of the DA algorithm but still requires one of following two scenarios — sampling $x$ from $\pi(x \mid \theta, y)$, or sampling $(\theta, x)$ from an importance function, $h(\theta, x \mid y)$. Thus, an appropriate importance function, $h(\theta, x \mid y)$, is required. In my opinion, effort required to discover such an appropriate importance function may be used otherwise to develop a simpler importance sampling method, which is presented as the one of the main contributions of this dissertation. Further, even if the importance function, $h(\theta, x \mid y)$, is readily available, the SIR method of Rubin, 1987 [96] generates only samples of $\theta$, which then have to be binned into a histogram. Thus, the SIR method still suffers from the same histogram estimation problems as the MCMC methods in Chapter 5. The PMDA algorithms by Wei and Tanner, 1990 [116], similar to the DA algorithm, do provide an analytical expression for the desired distribution, $\pi(\theta \mid y)$. Further, PMDA algorithms are non-interative in nature, unlike the DA algorithm. However, the PMDA algorithms have their own requirements, some of which are similar to the requirements of DA and SIR algorithms. Firstly, both PMDA algorithms (PMDA 1 and PMDA 2) require us to know, $\hat{\theta}$, the mode of the posterior, $\pi(\theta \mid y)$. Obviously, if $\hat{\theta}$ is already known, there would be no need to perform parameter estimation in the first place. In fact, Wei and Tanner, 1990 [116] present a Monte Carlo version of the Expectation Minimization algorithm (called MCEM) that provides the $\hat{\theta}$. The MCEM algorithm is an iterative procedure which requires sampling from $\pi(x \mid \theta, y)$ and provides only the MAP estimate, $\hat{\theta}$ and not

the entire posterior, $\pi(\theta \mid y)$. The PMDA algorithms were then provided by Wei and Tanner, 1990 [116] to obtain an analytical expression for the entire posterior, which justifies the name "data augmentation" in PMDA. Thus, PMDA algorithms are not parameter estimation methods at all. The aim of this discussion is to illustrate that while PMDA algorithms appear similar to the CDIS algorithm, PMDA algorithms cannot be used to estimate parameters. Secondly, even if $\hat{\theta}$ is available, both PMDA algorithms still require sampling from $\pi(x \mid \theta, y)$. Even further, even if it is possible to sample from $\pi(x \mid \theta, y)$, PMDA 1 still provides only a first-order approximation to the true posterior.

As the discussion subtly indicates, the main obstacle lies in the sampling of the complete-data trajectory $x$, conditioned on the given measurement data, $y$ through various conditional distributions. While different conditional distributions of the form $\pi(x, \cdot \mid \cdot, y)$ may be adjusted by using a specific method, samples of $x$ conditioned on $y$ **must** be generated. Further, since exact sampling from these conditionals is not possible in general, approximations to these conditionals have to be used. For example, the MCMC-MH method uses a proposal function and the Metropolis-Hastings sampling algorithm while the SIR method uses an importance function. While the choice of the approximate distribution may also be adjusted by choosing an appropriate method, some approximate distribution **must** be developed. And therein lies the practical challenge.

In this section, I develop and describe an importance sampling based parameter inference method, named *conditional sampling importance sampling* (CDIS), that is specific to the estimation framework described in this dissertation. While the analytical estimation of the posterior using an importance sampling method is certainly not novel, the importance function is. This importance function helps provide the following benefits

1. a non-iterative procedure
2. a (semi-)analytical expression of the entire posterior, as opposed to only the mode
3. convergence towards the true exact posterior
4. smaller computational expense than other methods

The importance sampling described in this section may also be categorized as *sequential importance sampling* (SIS) (Tokdar and Kass, 2010 [109], Richard and Zhang, 2007 [90] and Liu [64, p. 46]) due to the "sequential" nature of generating samples of a high-dimensional random variable.

### 6.1.1 *Importance sampling*

The MCMC methods in Sections 5.4 and 5.5 aim to estimate the *full* measurement-data posterior, $\pi(\theta, x \mid y)$. The CDIS method presented in this section attempts to obtain only the required posterior, $\pi(\theta \mid y)$. The procedure begins as follows

$$\begin{aligned}\pi(\theta \mid y) &= \int_x \pi(\theta, x \mid y) dx \\ &= \int_x \pi(\theta \mid x, y)\ \pi(x \mid y)\ dx \\ &= \int_x \pi(\theta \mid x)\ \pi(x \mid y)\ dx \quad (\because \text{ Eq. (5.87)})\end{aligned} \tag{6.1}$$

Obviously, if it was possible to sample from $\pi(x \mid y)$, then the above integral may be approximated by Monte Carlo sampling as

$$\begin{aligned}\pi(\theta \mid y) &= \int_x \pi(\theta \mid x)\ \pi(x \mid y)\ dx \\ &\approx \frac{1}{N_s} \sum_{k=1}^{N_s} \pi(\theta \mid x_k)\end{aligned} \tag{6.2}$$

in which, $x_k$, $k = 1, 2, \ldots, N_s$, are samples from $\pi(x \mid y)$. Since we cannot sample from $\pi(x \mid y)$, importance sampling may be used to perform the integral in Eq. (6.1)

$$\begin{aligned} \pi(\theta \mid y) &= \int_x \pi(\theta \mid x)\ \pi(x \mid y)\ dx \\ &\approx \sum_{k=1}^{N_s} \pi(\theta \mid x_k) \frac{\pi(x_k \mid y)}{q(x_k)} \end{aligned} \tag{6.3}$$

in which, $q(\cdot)$ is the *importance function (or distribution)* and $x_k$, $k = 1, 2, \ldots, N_s$, are samples from $q(x)$ instead. The ratios of the *target distribution* and the importance function, are called (non-normalized) *weights*, $w'_k$

$$w'_k = \frac{\pi(x_k \mid y)}{q(x_k)} \qquad k = 1, 2, \ldots, N_s \tag{6.4}$$

Re-writing Eq. (6.3) compactly,

$$\begin{aligned} \pi(\theta \mid y) &= \int_x \pi(\theta \mid x)\ \pi(x \mid y)\ dx \\ &\approx \sum_{k=1}^{N_s} w'_k\ \pi(\theta \mid x_k) \end{aligned} \tag{6.5}$$

Note that if the samples, $x_k$, can be generated from $q(x)$ and the conditional, $\pi(x \mid y)$, can be evaluated, then the weights, $w'_k$, can be computed numerically. Further, the complete-data posteriors, $\pi(\theta \mid x_k)$ are known analytically as product of gamma distributions in Eqs. (5.22)-(5.23). Investigating the conditional, $\pi(x \mid y)$,

$$\begin{aligned} \pi(x \mid y) &= \frac{\pi(x, y)}{\pi(y)} \\ &= \frac{\pi(y \mid x)\pi(x)}{\pi(y)} \end{aligned} \tag{6.6}$$

in which, $\pi(x)$ is the complete-data marginal likelihood known analytically in Eq. (5.24). The *reverse conditional*, $\pi(y \mid x)$, is very simply

$$\pi(y \mid x) = \begin{cases} 1 & \text{if } y \text{ is consistent with } x \\ 0 & \text{otherwise} \end{cases} \tag{6.7}$$

because $Y \subseteq X$ (Proposition 5.1). The measurement-data marginal likelihood, $\pi(y)$, is still unknown. However, $\pi(y)$ does not depend on $x_k$ and may be eliminated by normalizing the weights as follows. Enforcing the condition that the posterior density, $\pi(\theta \mid y)$, approximately integrates to one,

$$\begin{aligned} \int_\theta \pi(x \mid y)\, d\theta &\approx \int_\theta \sum_{k=1}^{N_s} w'_k \pi(\theta \mid x_k)\, d\theta \\ 1 &\approx \sum_{k=1}^{N_s} w'_k \int_\theta \pi(\theta \mid x_k)\, d\theta \\ 1 &\approx \sum_{k=1}^{N_s} w'_k \\ 1 &\approx \frac{1}{\pi(y)} \sum_{k=1}^{N_s} \frac{\pi(y \mid x_k)\pi(x_k)}{q(x_k)} \end{aligned}$$

results in the following approximate expression for the measurement-data marginal likelihood

$$\pi(y) \approx \hat{\pi}(y) = \sum_{k=1}^{N_s} \frac{\pi(y \mid x_k)\pi(x_k)}{q(x_k)} \tag{6.8}$$

in which, $\hat{\pi}(y)$ is the approximate measurement-data marginal likelihood. An approximate posterior may now be defined as

$$\begin{aligned}\pi(\theta \mid y) &\approx \sum_{k=1}^{N_s} w'_k \, \pi(\theta \mid x_k) \\ &= \frac{1}{\pi(y)} \sum_{k=1}^{N_s} \frac{\pi(y \mid x_k)\pi(x_k)}{q(x_k)} \pi(\theta \mid x_k) \\ &\approx \frac{1}{\hat{\pi}(y)} \sum_{k=1}^{N_s} \frac{\pi(y \mid x_k)\pi(x_k)}{q(x_k)} \pi(\theta \mid x_k) \\ &= \hat{\pi}(\theta \mid y) \end{aligned} \tag{6.9}$$

Re-arranging,

$$\hat{\pi}(\theta \mid y) = \sum_{k=1}^{N_s} \frac{\pi(y \mid x_k)\pi(x_k)}{q(x_k)\hat{\pi}(y)} \pi(\theta \mid x_k)$$

in which,

$$\frac{\pi(y \mid x_k)\pi(x_k)}{q(x_k)\hat{\pi}(y)} = w_k = \frac{\frac{\pi(y|x_k)\pi(x_k)}{q(x_k)}}{\sum_{j=1}^{N_s} \frac{\pi(y|x_j)\pi(x_j)}{q(x_j)}} \tag{6.10}$$

Here, $w_k$, $k = 1, 2, \ldots, N_s$, are the *normalized weights* which may also be written as

$$w_k = \frac{w'_k}{\sum_{j=1}^{N_s} w'_j} = \frac{\frac{\pi(y|x_k)\pi(x_k)}{q(x_k)}}{\sum_{j=1}^{N_s} \frac{\pi(y|x_k)\pi(x_k)}{q(x_j)}} \qquad k = 1, 2, \ldots, N_s \tag{6.11}$$

Note that the normalized weights sum up to one,

$$\sum_{k=1}^{N_s} w_k = 1 \tag{6.12}$$

Also, note that when the sampled complete-data trajectory $x_k$ is inconsistent with the measurement data $y$, *i. e.*, when $\pi(y \mid x_k) = 0$,

$$w_k = 0, \qquad k = 1, 2, \ldots, N_s \tag{6.13}$$

Re-writing the approximate posterior, $\hat{\pi}(\theta \mid y)$, compactly

$$\hat{\pi}(\theta \mid y) = \sum_{k=1}^{N_s} w_k \, \pi(\theta \mid x_k) \tag{6.14}$$

Note that in the above expression of the approximate posterior, all terms are either analytically known or may be numerically computed if an appropriate importance function $q(x)$ is available. The only restrictions on the importance function are

1. $q(x)$ must have the same support as the target distribution $\pi(x \mid y)$
2. (independent) sampling of $x \sim q(x)$ must be possible
3. evaluation of probability density $q(x_k)$ must be possible for any sample $x_k$

Any importance function that satisfies these three mandatory restrictions provides us with an importance-sampling-based parameter inference method which has the first three properties promised in the introduction to Section 6.1. Firstly, the method to obtain the approximate posterior, $\hat{\pi}(\theta \mid y)$ is non-iterative. Secondly, for any number of samples, $N_s$, including when $N_s = 1$, Eq. (6.14) provides a semi-analytical posterior. The complete-data posteriors are available analytically (see Eqs. (5.22)-(5.23)) and the weights, $w_k$, $k = 1, 2, \ldots, N_s$, may be numerically computed (see Eq. (6.11)). This expression, $\hat{\pi}(\theta \mid y)$, in Eq. (6.14), is called *semi-analytical* to indicate that its evaluation requires a detailed sampling procedure and is not as straightforward as the evaluation of the posterior (Eqs. (5.22)-(5.23)) when complete-data is given. Thirdly, the convergence to the true exact posterior, $\pi(\theta \mid y)$ is guaranteed in the limit of infinite number of samples, *i. e.*, as $N_s \to \infty$.

The restrictions above are rather mild and allow the use of an importance function that may not even be related to the stochastic chemical kinetic model but satisfies the restrictions. In such a case, where the importance function, $q(x)$ does not resemble the target distribution, $\pi(x \mid y)$, the convergence of $\hat{\pi}(\theta \mid y)$ towards the $\pi(\theta \mid y)$ is extremely slow. Since $q(x)$ does not resemble $\pi(x \mid y)$, the ratio $\frac{\pi(x_k|y)}{q(x)}$ has a large variance over $k = 1, 2, \ldots, N_s$ which results in a large variance in the weights $w_k$. Consequently, while the convergence is still guaranteed, a prohibitively large number of samples is required to obtain a good estimate of the true posterior. In contrast, when the importance function is similar to the target distribution, the weights, $w_k$, $k = 1, 2, \ldots, N_s$ have low variance and better convergence is achieved using fewer samples. Thus, the choice of the importance function is the key to practical use of this parameter estimation method. In the next subsection, I describe a few importance functions which perform well. The class of parameter inference methods described in this section are collectively called *importance sampling based methods*, which belong to the category of simulation methods.

### 6.1.2 *Importance functions*

The importance function, $q(x)$, described in the previous subsection, aims to mimic the target distribution, $\pi(x \mid y)$. As mentioned before, importance sampling would not even be required if it was possible to sample from the target distribution itself. In that case, $q(x) = \pi(x \mid y)$, and the weights would all be equal, $w_k = \frac{1}{N_s}$, $k = 1, 2, \ldots, N_s$, which corresponds to an ideal condition of *zero-variance weights* and the approximate posterior is given by Eq. (6.2). Since it is not possible to sample from $\pi(x \mid y)$, I attempt to develop an importance function, $q(x)$, as an approximation to the target distribution, that allows sampling and is as close to the the target distribution, $\pi(x \mid y)$, as possible. I begin with the target

distribution itself,

$$\begin{aligned}\pi(x \mid y) &= \int_{\theta} \pi(\theta, x \mid y)\, d\theta \\ &= \int_{\theta} \pi(x \mid \theta, y)\pi(\theta \mid y)\, d\theta \end{aligned} \tag{6.15}$$

The above expression requires the use of the posterior, $\pi(\theta \mid y)$, which is obviously not available. However, since only an approximation to $\pi(x \mid y)$ is required I can approximate $\pi(\theta \mid y)$ with another known distribution, namely the prior, $\pi(\theta)$. This approximation is not that far-fetched, given that in the case of complete-data measurements (Section 5.1.2), both the prior and posterior belong to the same family of distributions, albeit with different parameters.

$$\begin{aligned}\pi(x \mid y) &= \int_{\theta} \pi(x \mid \theta, y)\pi(\theta \mid y)\, d\theta \\ &\approx \int_{\theta} \pi(x \mid \theta, y)\pi(\theta)\, d\theta \end{aligned} \tag{6.16}$$

in which, the coveted conditional distribution, $\pi(x \mid \theta, y)$ is still unknown. Using Eq. (5.92), we know that,

$$\pi(x \mid \theta, y) = \prod_{i=1}^{m} \pi(x_{[s_{i-1}, s_i]} \mid \mathbf{x}(s_{i-1}), \theta, y_i)$$

As discussed in Sections 5.4 and 5.5, a sample of $x$ may be generated by sampling the complete-data intervals, $x_{[s_{i-1}, s_i]}$, $i = 1, 2, \ldots, m$. The aim for the rest of this subsection is to develop an importance function that generate samples of complete-data intervals. Essentially, if $q(x_{[s_{i-1}, s_i]} \mid \theta, y)$ represents the (approximate) distribution that can generate the corresponding complete-data interval, then $q\,(x \mid \theta, y)$ may be represented by

$$q\,(x \mid \theta, y) = \prod_{i=1}^{m} q(x_{[s_{i-1}, s_i]} \mid \theta, y) \tag{6.17}$$

Substituting Eq. (6.17) into Eq. (6.15), and continuing the approximation

$$\begin{aligned}\pi(x \mid y) &\approx \int_\theta \pi(x \mid \theta, y)\pi(\theta)\, d\theta \\ &\approx \int_\theta q\left(x \mid \theta, y\right) \pi(\theta)\, d\theta \\ &= \int_\theta \prod_{i=1}^{m} q(x_{[s_{i-1},s_i]} \mid \theta, y)\ \pi(\theta)\, d\theta \end{aligned} \tag{6.18}$$

Using the above expression, and the identity, $\int_\theta \pi(\theta)\, d\theta = 1$, and approximating further,

$$\begin{aligned}\pi(x \mid y) &\approx \int_\theta \prod_{i=1}^{m} q(x_{[s_{i-1},s_i]} \mid \theta, y)\ \pi(\theta)\, d\theta \\ &= \int_\theta \prod_{i=1}^{m} q(x_{[s_{i-1},s_i]} \mid \theta, y)\ \pi(\theta)\, d\theta \times \prod_{i=2}^{m} \int_\theta \pi(\theta)\, d\theta \\ &\approx \prod_{i=1}^{m} \int_\theta q(x_{[s_{i-1},s_i]} \mid \theta, y)\ \pi(\theta)\, d\theta \end{aligned} \tag{6.19}$$

Using the equivalence of complete-data trajectory, $x_{[s_{i-1},s_i]} = \{\mathbf{x}(t) : t \in [s_{i-1}, s_i]\}$, and the complete event data in Eq. (5.11),

$$q\left(x_{[s_{i-1},s_i]} \mid \theta, y\right) = q\left(\begin{array}{c|c} \mathbf{x}(s_{i-1}),\ n, & \\ (n_j, t_j),\ j = 1, 2, \ldots, n, & \theta, y \\ \text{no reaction in } (t_n, s_i] & \end{array}\right) \tag{6.20}$$

in which, $n$ and $(n_j,\ t_j)$, $j = 1, 2, \ldots, n$, have their usual meanings from Section 5.1. The random variable, $N$ (and its sample $n$), represents the total number of reactions occurred during the time interval $[s_{i-1}, s_i]$. The vector-valued random variable, $\mathbf{R} \in \mathbb{R}^{n_r}$ (and its sample $\mathbf{r}$), represents the number of times each reaction $\mathcal{R}_i$, $i = 1, 2, \ldots, n_r$, occurred in time interval $[s_{i-1}, s_i]$ (see Section 5.1.1). Thus, given $\mathbf{R}$, $N$ is completely given by the Eq. (5.7). Hence, the event $\{N = n,\ \mathbf{R} = \mathbf{r}\}$ is

equivalent to the event $\{\mathbf{R} = \mathbf{r}\}$.

$$\{N = n,\ \mathbf{R} = \mathbf{r}\} \equiv \{\mathbf{R} = \mathbf{r}\} \tag{6.21}$$

Using the above equivalence with the understanding that the complete event data contains the entire information about the path $x_{[s_{i-1},s_i]}$, we can establish the equivalence of the following events

$$\begin{aligned} x_{[s_{i-1},s_i]} &\equiv \left\{ \begin{array}{c} \mathbf{x}(s_{i-1}),\ n, \\ (n_j, t_j),\ j = 1, 2, \ldots, n, \\ \text{no reaction in } (t_n, s_i] \end{array} \right\} \equiv \left\{ \begin{array}{c} \mathbf{x}(s_{i-1}),\ n,\ \mathbf{r} \\ (n_j, t_j),\ j = 1, 2, \ldots, n, \\ \text{no reaction in } (t_n, s_i] \end{array} \right\} \\ &\equiv \left\{ \begin{array}{c} \mathbf{x}(s_{i-1}),\ \mathbf{r} \\ (n_j, t_j),\ j = 1, 2, \ldots, n, \\ \text{no reaction in } (t_n, s_i] \end{array} \right\} \end{aligned} \tag{6.22}$$

Substituting from Eq. (6.22) into Eq. (6.20),

$$q\left(x_{[s_{i-1},s_i]} \mid \theta, y\right) = q\left( \begin{array}{c} \mathbf{x}(s_{i-1}),\ \mathbf{r}, \\ (n_j, t_j),\ j = 1, 2, \ldots, n, \\ \text{no reaction in } (t_n, s_i] \end{array} \middle| \ \theta, y \right) \tag{6.23}$$

By repeated conditioning,

$$
\begin{aligned}
q\left(x_{[s_{i-1},s_i]} \mid \theta, y\right) &= q\left(\begin{array}{c|c} \mathbf{r}, & \\ (n_j, t_j),\ j=1,2,\ldots,n, & \mathbf{x}(s_{i-1}), \theta, y \\ \text{no reaction in } (t_n, s_i] & \end{array}\right) \\
&\quad \times\ q\left(\mathbf{x}(s_{i-1}) \mid \theta, y\right) \\
&= q\left(\begin{array}{c|c} (n_j, t_j),\ j=1,2,\ldots,n, & \mathbf{r},\ \mathbf{x}(s_{i-1}), \theta, y \\ \text{no reaction in } (t_n, s_i] & \end{array}\right) \\
&\quad \times\ q\left(\mathbf{r} \mid \mathbf{x}(s_{i-1}), \theta, y\right) \\
&\quad \times\ q\left(\mathbf{x}(s_{i-1}) \mid \theta, y\right) \\
&= q\left(\{t_j,\ j=1,2,\ldots,n\} \mid \{n_j,\ j=1,2,\ldots,n\},\ \mathbf{r},\ \mathbf{x}(s_{i-1}), \theta, y\right) \\
&\quad \times q\left(\{n_j,\ j=1,2,\ldots,n\} \mid \mathbf{r},\ \mathbf{x}(s_{i-1}), \theta, y\right) \\
&\quad \times q\left(\mathbf{r} \mid \mathbf{x}(s_i), \theta, y\right) \\
&\quad \times q\left(\mathbf{x}(s_{i-1}) \mid \theta, y\right)
\end{aligned} \tag{6.24}
$$

Renaming the approximate distributions for convenience,

$$
\begin{aligned}
q^{(i)}_{t|\theta} &= q\left(\{t_j,\ j=1,2,\ldots,n\} \mid \{n_j,\ j=1,2,\ldots,n\},\ \mathbf{r},\ \mathbf{x}(s_{i-1}), \theta, y\right) \\
q^{(i)}_{\text{seq}|\theta} &= q\left(\{n_j,\ j=1,2,\ldots,n\} \mid \mathbf{r},\ \mathbf{x}(s_{i-1}), \theta, y\right) \\
q^{(i)}_{\mathbf{r}|\theta} &= q\left(\mathbf{r} \mid \mathbf{x}(s_{i-1}), \theta, y\right) \\
q^{(i)}_{\text{ic}|\theta} &= q\left(\mathbf{x}(s_{i-1}) \mid \theta, y\right)
\end{aligned} \tag{6.25}
$$

Eq. (6.24) may be written compactly as

$$
q\left(x_{[s_{i-1},s_i]} \mid \theta, y\right) = q^{(i)}_{t|\theta}\, q^{(i)}_{\text{seq}|\theta}\, q^{(i)}_{\mathbf{r}|\theta}\, q^{(i)}_{\text{ic}|\theta} \tag{6.26}
$$

It is important to note that we could choose the distribution functions, $q^{(i)}_{t|\theta}$, $q^{(i)}_{\text{seq}|\theta}$, $q^{(i)}_{\mathbf{r}|\theta}$ and $q^{(i)}_{\text{ic}|\theta}$ arbitrarily as long as the final importance distribution $q(x)$ follows the stated restrictions in Section 6.1.1. Howvere, the aim of this exercise is to obtain only an approximation of $\pi(x \mid y)$.

Substituting the above expression in $i^{\text{th}}$ product term of the right-hand side of Eq. (6.19),

$$\int_\theta q\left(x_{[s_{i-1},s_i]} \mid \theta, y\right)\ \pi(\theta)\ d\theta = \int_\theta q^{(i)}_{t|\theta}\ q^{(i)}_{\text{seq}|\theta}\ q^{(i)}_{\mathbf{r}|\theta}\ q^{(i)}_{\text{ic}|\theta}\ d\theta \tag{6.27}$$

Since the initial condition, $\mathbf{x}(s_0)$ is assumed to be known (*i. e.*, fixed), and $\mathbf{x}(s_{i-1})$, $i = 2, \ldots, m$ is known using the previous complete-data interval, $x_{[s_{i-2},s_{i-1}]}$,

$$q^{(i)}_{\text{ic}|\theta} = q\left(\mathbf{x}(s_{i-1}) \mid \theta, y\right) = 1 \tag{6.28}$$

Further, assuming that $q^{(i)}_{\mathbf{r}|\theta}$ does not depend on $\theta$,

$$\begin{aligned} q^{(i)}_{\mathbf{r}} &= \int_\theta q^{(i)}_{\mathbf{r}|\theta}\ \pi(\theta)\ d\theta \\ &= q^{(i)}_{\mathbf{r}|\theta} \int_\theta \pi(\theta)\ d\theta \\ &= q^{(i)}_{\mathbf{r}|\theta} \end{aligned} \tag{6.29}$$

Eq. (6.27) may now be reduced to

$$\begin{aligned} \int_\theta q\left(x_{[s_{i-1},s_i]} \mid \theta, y\right)\ \pi(\theta)\ d\theta &= \int_\theta q^{(i)}_{t|\theta}\ q^{(i)}_{\text{seq}|\theta}\ q^{(i)}_{\mathbf{r}|\theta}\ q^{(i)}_{\text{ic}|\theta}\ d\theta \\ &= q^{(i)}_{\mathbf{r}|\theta} \int_\theta q^{(i)}_{t|\theta}\ q^{(i)}_{\text{seq}|\theta}\ \pi(\theta)\ d\theta \\ &= q^{(i)}_{\mathbf{r}|\theta} \int_\theta q^{(i)}_{t|\theta}\ q^{(i)}_{\text{seq}|\theta}\ \pi(\theta)\ d\theta \int_\theta \pi(\theta)\ d\theta \\ &\approx q^{(i)}_{\mathbf{r}|\theta} \int_\theta q^{(i)}_{t|\theta}\ \pi(\theta)\ d\theta \int_\theta q^{(i)}_{\text{seq}|\theta}\ \pi(\theta)\ d\theta \end{aligned} \tag{6.30}$$

Further renaming the two distributions,

$$q_t^{(i)} \approx \int_\theta q_{t|\theta}^{(i)}\ \pi(\theta)\ d\theta \tag{6.31}$$

$$q_{\text{seq}}^{(i)} \approx \int_\theta q_{\text{seq}|\theta}^{(i)}\ \pi(\theta)\ d\theta \tag{6.32}$$

Rewriting Eq. (6.30) in new variables,

$$\int_\theta q\left(x_{[s_{i-1},s_i]} \mid \theta, y\right)\ \pi(\theta)\ d\theta \approx q_{\mathbf{r}}^{(i)}\ q_t^{(i)}\ q_{\text{seq}}^{(i)} \tag{6.33}$$

Using Eq. (6.19) and Eq. (6.33), the overall importance function may be written as the following crude approximation of $\pi(x \mid y)$,

$$q(x) = \prod_{i=1}^{m} q_{\mathbf{r}}^{(i)}\ q_{\text{seq}}^{(i)}\ q_t^{(i)} \tag{6.34}$$

A sample $x$ may be generated using $q(x)$ by sequentially sampling each complete-data interval, $x_{[s_{i-1},s_i]}$, $i = 1, 2, \ldots, m$. Note that this sequential algorithm is similar to the Algorithms 5.2 and 5.6. The $i^{\text{th}}$ complete-data interval may be sampled in three stages — (1) sample $\mathbf{r} \sim q_{\mathbf{r}}^{(i)}$, (2) given $\mathbf{r}$, sample $\{n_j,\ j = 1, 2, \ldots, n\}$ using $q_{\text{seq}}^{(i)}$, and (3) given $\mathbf{r}$, $\{n_j,\ j = 1, 2, \ldots, n\}$, sample $\{t_j,\ j = 1, 2, \ldots, n\}$ using $q_t^{(i)}$. These steps are explained next.

*(1) Sample* $\mathbf{r} \sim q_{\mathbf{r}}^{(i)}$

**Case 1**: *Non-cyclical kinetics, full measurements.* Since the kinetics in non-cyclical, $\nu^T$ has full column rank. Since full measurements are available, $\mathbf{r}$ for the time interval $[s_{i-1}, s_i]$ may be directly computed by solving the system of linear equations

$$\mathbf{x}(s_i) - \mathbf{x}(s_{i-1}) = \nu^T \mathbf{r} \tag{6.35}$$

which has a unique solution. In this case, the random variable $\mathbf{R}$ is actually known deterministically.

**Case 2**: *Cyclical kinetics, full measurements.* Since full measurements are available, $\mathbf{r}$ for the time interval $[s_{i-1}, s_i]$ is still constrained according to the Eq. (6.35). However, since the kinetics is cyclical, $\nu^T$ does not have full column rank, and the solution to Eq. (6.35) is no longer unique. In such a case, a *tuning distribution* is required to generate samples of $\mathbf{r}$. This issue is also encountered during the implementation of MCMC-MH algorithm. Wilkinson, 2012 [119], Boys et al., 2008 [14] used a modified Bessel function with a rather ad hoc tuning parameter which is specific to their Lotka-Volterra model. Since the tuning distribution provides only the proposal values which are later corrected by an acceptance/rejection step, the exact choice of the tuning distribution is not very restrictive. The only consideration is to achieve "good" convergence. The tuning distribution provided by Wilkinson, 2012 [119], Boys et al., 2008 [14] induces "good" mixing of the chain thereby resulting in "good" convergence. Further, a general "tuning distribution" is difficult to specify. Therefore, I describe only a general framework of choosing a "tuning distribution" and provide examples. Also note that, as in the case of MCMC-MH method, the choice of the importance function is also not very restrictive.

Let $d$, $d < n_r$ be the degrees of freedom in the solution space of Eq. (6.35). Partition the solution space, $\mathbf{r} \in \mathbb{R}^{n_r}$, into $\mathbf{r}^{\mathrm{f}} \in \mathbb{R}^{n_r - d}$ and $\mathbf{r}^{\mathrm{v}} \in \mathbb{R}^{d}$. Here $\mathbf{r}^{\mathrm{v}}$ denotes the independent variables and $\mathbf{r}^{\mathrm{f}}$ denotes the rest of the solution space. Note that, $\mathbf{r}$ only assumes non-negative integer values. As a result of this partition,

$$q^{(i)}(\mathbf{r}^{\mathrm{f}} \mid \mathbf{r}^{\mathrm{v}}) = \begin{cases} 1 & \text{if } \mathbf{r}^{\mathrm{f}} \text{ is consistent with } \mathbf{r}^{\mathrm{v}} \\ 0 & \text{otherwise} \end{cases} \tag{6.36}$$

and,

$$\begin{aligned} q^{(i)}(\mathbf{r}) &= q^{(i)}(\mathbf{r}^{\mathrm{f}}, \mathbf{r}^{\mathrm{v}}) \\ &= q^{(i)}(\mathbf{r}^{\mathrm{f}} \mid \mathbf{r}^{\mathrm{v}})\; q^{(i)}(\mathbf{r}^{\mathrm{v}}) \end{aligned} \tag{6.37}$$

Thus, a distribution, $q^{(i)}(\mathbf{r}^{\mathrm{v}})$, that is defined only on the independent variables is required. When $d = 1$ (*e.g.*, Lotka-Volterra model [14]), the random variable $\mathbf{r}^{\mathrm{v}}$ is a scalar and a conditional Poisson distribution may be used to generate the required samples. The gene on-off example presented in Section 7.2 also uses such a Poisson distribution.

**Case 3**: *Partial measurements.* Since partial measurements are available, instead of Eq. (6.35), the following system of linear equations must be solved to obtain $\mathbf{r}$ (see Section 5.2.1),

$$y_i - \mathbf{C}\mathbf{x}(s_{i-1}) = \mathbf{C}\; \nu^T \mathbf{r} \tag{6.38}$$

This case is similar to either one of the above cases and may be handled similarly.

Given $\mathbf{r}$, $n$ is known according to the Eq. (5.7). Next, a sequence of reaction indices, $n_j$, $j = 1, 2, \ldots, n$ needs to be sampled. And then, a sequence of reaction times, $t_j$, $j = 1, 2, \ldots, n$ needs to sampled. Note that the corresponding importance distributions, $q^{(i)}_{\mathrm{seq}}$ and $q^{(i)}_t$, do not involve the parameters, $\theta$. In fact, $\theta$ has been deliberately integrated out. At a first look, especially from a physical standpoint, sampling a complete-data path $x_{[s_{i-1}, s_i]}$ or any of its components (*e.g.*, reaction index $n_j$, reaction time $t_j$) without conditioning on $\theta$ seems absurd. However, a look at Eqs. (6.31)-(6.32) would suggest an intuitive explanation. Consider the following Monte Carlo thought experiment — (a) sample $\theta_k \sim \pi(\theta)$, (b) conditioned on $\theta_k$, generate the required random variable, for ex-

ample, the sequence of reaction indices, $n_j,\ j = 1, 2, \ldots, n$ from the conditional distribution, $\pi(n_j,\ j = 1, 2, \ldots, n \mid \theta, \cdot)$. The samples of $n_j$, thus obtained, come from the joint distribution $\pi(\theta,\ n_j,\ j = 1, 2, \ldots, n)$ (see Gibbs sampling in Section 5.4.1). The samples of $n_j$ can now be binned to generate a histogram of the marginal $\pi(n_j,\ j = 1, 2, \ldots, n)$. Thus, sampling relevant random variables without conditioning on $\theta$ *does* make sense. However, I do not use such a sampling scheme. Instead, the importance distributions, $q^{(i)}_{\text{seq}}$ and $q^{(i)}_{t}$, are obtained by analytically integrating out $\theta$ (with approximation).

*(2) Sample* $\{n_j,\ j = 1, 2, \ldots, n\}$ *using* $q^{(i)}_{seq}$

**Case 1**: *Small n.* Given **r**, it is known how many times each reaction occurred in the time interval $[s_{i-1}, s_i]$. All that is required to sample an arrangement of the corresponding reaction indices. The total number of reactions, $n$ is given as,

$$n = \sum_{u=1}^{n_r} r_u$$

while the number of times each reaction, $\mathcal{R}_u,\ u = 1, 2, \ldots, n_r$, occurred in the $i^{\text{th}}$ interval, is given by $r_u$. The maximum number of sequences possible is given by the following multinomial coefficient

$$N_{\text{seq, max}} = \binom{n}{r_1\ r_2\ \cdots\ r_{n_r}} = \frac{n!}{r_1!\ r_2!\ \cdots\ r_{n_r}!}$$

Since $n$ is small enough, $N_{\text{seq, max}}$ is small as well. Note that $N_{\text{seq, max}}$ represents the maximum possible sequences. The reaction kinetics may allow only a fraction of these $N_{\text{seq, max}}$ sequences to actually occur. For example, some sequences may correspond to negative species which is not feasible. When $N_{\text{seq, max}}$ is small or when it is expected (by examining the reaction kinetics) that the actual number of feasible sequences, $N_{\text{seq}}$ is small then all of these feasible sequences may be enumerated. Assuming that the reaction propensities do not directly depend on

time, the following probability may be exactly and analytically specified

$$\begin{aligned}\pi\left(\{n_j,\ j=1,2,\ldots,n\} \mid \theta, \mathbf{x}(s_{i-1})\right) &= \prod_{j=1}^{n} \frac{h_{n_j}(\mathbf{x}(t_{j-1}),\theta)}{h_0(\mathbf{x}(t_{j-1}),\theta)} \\ &= \prod_{j=1}^{n} \frac{k_{n_j} g_{n_j}(\mathbf{x}(t_{j-1}))}{h_0(\mathbf{x}(t_{j-1}),\theta)}\end{aligned} \tag{6.39}$$

in which, $t_0 = s_{i-1}$ and $t_{n+1} = s_i$ are defined for notational convenience and $h_0(\cdot)$ is the total reaction propensity as defined in Eq. (5.12) (see Section 5.1.1). Using the definition of $g_{\text{prod}}$ in Eq. (5.16) for the $i^{\text{th}}$ interval,

$$\pi\left(\{n_j,\ j=1,2,\ldots,n\} \mid \theta, \mathbf{x}(s_{i-1})\right) = g_{\text{prod}}^{(i)} \prod_{v=1}^{n_r} {k_v}^{r_v} \frac{1}{\prod_{j=1}^{n} h_0(\mathbf{x}(t_{j-1}),\theta)} \tag{6.40}$$

in which

$$g_{\text{prod}}^{(i)} = \prod_{j=1}^{n} g_{n_j}(\mathbf{x}(t_{j-1})) \quad t_j \in [s_{i-1}, s_i] \tag{6.41}$$

Conditioning on $y_i$ may be performed by only admitting the feasible sequences, *i.e.*,

$$\begin{aligned}&\pi\left(\{n_j,\ j=1,2,\ldots,n\} \mid \theta, \mathbf{x}(s_{i-1}), y_i\right) \\ &\begin{cases} \propto g_{\text{prod}}^{(i)} \prod_{v=1}^{n_r} {k_v}^{r_v} \frac{1}{\prod_{j=1}^{n} h_0(\mathbf{x}(t_{j-1}),\theta)} & \text{if sequence is feasible} \\ = 0 & \text{otherwise} \end{cases}\end{aligned} \tag{6.42}$$

Choosing $q_{\text{seq}|\theta}^{(i)}$ as the above distribution and integrating as described in Eq. (6.32), the required importance distribution, $q_{\text{seq}}^{(i)}$, may be written as

$$q_{\text{seq}}^{(i)}\left(\{n_j,\ j=1,2,\ldots,n\}\right) \propto \int_{\theta} g_{\text{prod}}^{(i)} \prod_{v=1}^{n_r} {k_v}^{r_v} \frac{1}{\prod_{j=1}^{n} h_0(\mathbf{x}(t_{j-1}),\theta)} \pi(\theta) d\theta \tag{6.43}$$

in which $\{n_j,\ j = 1, 2, \ldots, n\}$ is a feasible sequence and $W$ is just a proportionality constant. Approximating further,

$$q_{\text{seq}}^{(i)}\left(\{n_j,\ j = 1, 2, \ldots, n\}\right) \propto g_{\text{prod}}^{(i)} \tag{6.44}$$

$$q_{\text{seq}}^{(i)}\left(\{n_j,\ j = 1, 2, \ldots, n\}\right) = \frac{g_{\text{prod}}^{(i)}}{\sum_{\text{all feasible sequences}} g_{\text{prod}}^{(i)}} \tag{6.45}$$

Note that the expression $g$prod also appears in the expression of $\pi(x)$ in Eq. (5.24). In fact, if the above importance function, $q_{\text{seq}}^{(i)}$, is used, then the term $g_{\text{prod}}$ cancels out leading to a lower variance in weights. Thus, if all sequences may be enumerated, then the above expression provides an excellent importance function.

**Case 2**: *Large $n$.* Given **r**, it is known how many times each reaction occurred in the time interval $[s_{i-1}, s_i]$. All that is required to sample an arrangement of the corresponding reaction indices. The total number of reactions, $n$ is given as,

$$n = \sum_{u=1}^{n_r} r_u$$

while the number of times each reaction, $\mathcal{R}_u$, $u = 1, 2, \ldots, n_r$, occurred in the $i^{\text{th}}$ interval, is given by $r_u$. The maximum number of sequences possible is given by the following multinomial coefficient

$$N_{\text{seq, max}} = \binom{n}{r_1\ r_2\ \cdots\ r_{n_r}} = \frac{n!}{r_1!\ r_2!\ \cdots\ r_{n_r}!} \tag{6.46}$$

For large $n$, the number of possible sequences is much larger, and a faster method to sample sequence is needed. One such method is to choose each sequence with the same probability

$$q_{\text{seq}}^{(i)} = \frac{1}{N_{\text{seq, max}}} \quad \text{for every sequence } \{n_j,\ j = 1, 2, \ldots, n\} \tag{6.47}$$

Note that this importance function, $q_{\text{seq}}^{(i)}$ may generate a sample sequence which is not feasible, *i.e.*, at least one species becomes negative. Such a sequence corresponds to $\pi(x_k) = 0$, which implies that the corresponding weight, $w_k = 0$. As a result, this infeasible sample does not contribute to the final estimation. Further, the computer program may be written in such a way that the infeasible sequences are discarded without further computations, thus saving computational effort.

*(3) Sample* $\{t_j,\ j = 1, 2, \ldots, n\}$ *using* $q_t^{(i)}$

Having sampled $\mathbf{r}$ and the reaction index sequence, the last quantity required is the reaction times, $t_j$, $j = 1, 2, \ldots, n$. In other words, now that we know which reactions occur, all we need to do is assign times to those reactions. Note that, conditioned on $\theta$, $\mathbf{x}(s_{i-1})$ and reaction index sequence $\{n_j,\ j = 1, 2, \ldots, n\}$, the reaction times, $t_j$, $j = 1, 2, \ldots, n$, are distributed with known exponential distributions. In fact, simulation of reaction times conditions on parameters, initial condition and reaction index sequences, is just a part of the *forward simulation* described in Section 2.3. Note that since we have conditioned on the reaction index sequence, we do not need to condition on $y_i$ again. However, there is one last conditioning remaining — we need to ensure that last reaction time, $t_n \leq s_i$. This can be accomplished as follows. Let the variables, $V_j$ represent the "inter-event" times as follows

$$V_j = T_j - T_{j-1} \quad j = 1, 2, \ldots, n+1 \tag{6.48}$$

in which, $T_0 = s_{i-1}$ and $T_{n+1} = s_i$ are defined for notational convenience. Given $\theta$, $\mathbf{x}(s_{i-1})$, and $\{n_j,\ j = 1, 2, \ldots, n\}$, each inter-event time, $V_j$, $j = 1, 2, \ldots, n$, are exponentially distributed, independent random variables with known parameters.

$$V_j \sim \text{Exp}(h_0(\mathbf{x}(T_{j-1}), \theta)) \quad j = 1, 2, \ldots, n+1 \tag{6.49}$$

Note that, the reaction propensities $h_{n_j}(\mathbf{x}(T_{j-1}), \theta)$, $j = 1, 2, \ldots, n+1$, even though functions of random variables, $T_{j-1}$, $j = 1, 2, \ldots, n+1$, are known. This is because the (1) reaction propensities do not depend directly on time and, (2) the states, $\mathbf{x}(t_{j-1})$, $j = 1, 2, \ldots, n+1$, are known because the reaction index sequence is known.

The sum of independent exponential distributions form a *Hypoexponential distribution* (Smaili et al., 2013 [100], Bolch et al., 2001 [12]). See Section A.2 in Appendix A. Renaming the propensities,

$$\lambda_j = h_0(\mathbf{x}(T_{j-1}), \theta) \quad j = 1, 2, \ldots, n+1 \tag{6.50}$$

the required Hypoexponential distributions is given as

$$T_n - T_0 = \sum_{j=1}^{n} V_j \sim \text{HypoExp}(\lambda_1, \lambda_2, \ldots, \lambda_n) \tag{6.51}$$

$$T_{n+1} - T_0 = \sum_{j=1}^{n+1} V_j \sim \text{HypoExp}(\lambda_1, \lambda_2, \ldots, \lambda_{n+1}) \tag{6.52}$$

Conditioning on the event

$$\begin{aligned} A &\stackrel{d}{=} \{T_n \leq s_i\} \cap \{\ T_{n+1} \geq s_i\} \\ &\equiv \{T_n - T_0 \leq \Delta s\} \cap \{\ T_{n+1} - T_0 \geq \Delta s\} \end{aligned} \tag{6.53}$$

the random variable, $T_n - T_0 \mid A$, is nothing but a *conditioned Hypoexponential random variable*. See Section A.3 for details. Using similar integration and approximation procedure, the following importance function may be obtained,

$$q_t^{(i)}(\{t_j,\ j = 1, 2, \ldots, n\}) = \prod_{j=1}^{n} q_{t_j}^{(i)}(t_j) \tag{6.54}$$

in which,

$$q_{t_j}^{(i)}(t_j) \propto \frac{t_j^{j-1}}{\prod_{v=1}^{n_r}(1+\alpha_v t_j)^{a_v}} \quad j = 1, 2, \ldots, n \tag{6.55}$$

$$\alpha_v = \frac{\sum_{u=1}^{j} g_v(\mathbf{x}(t_u)) - g_v(\mathbf{x}(t_j))}{j b_v} \tag{6.56}$$

$T_j$, $j = 1, 2, \ldots, n$, may be sampled using the above distribution via acceptance/rejection sampling.

*Large n.* In this case, $T_j$, $j = 1, 2, \ldots, n$, may be sampled using a *truncated gamma* distribution (see Appendix A).

### 6.1.3 *CDIS Algorithm*

The procedure described in the previous section is named *conditional density importance sampling* (CDIS) method. Algorithm 6.1 provides the overall CDIS algorithm.

**Algorithm 6.1** Overall CDIS

1: Given measurement-data $y$
2: Separate $y$ into intervals $[y_{i-1}, y_i]$, $i = 1, 2, \ldots, m$
3: **for** $i = 1$ to $m$ **do**
4:  **repeat**
5:  Sample $\mathbf{r} \sim q_{\mathbf{r}}^{(i)}$
6:  Compute $n = \sum_{v=1}^{n_r} r_v$
7:  Sample $\{n_j,\ j = 1, 2, \ldots, n\} \sim q_{\text{seq}}^{(i)}$
8:  Sample $\{t_j,\ j = 1, 2, \ldots, n\} \sim q_t^{(i)}$
9:  **until** $N_s$ samples of $x_k$, $k = 1, 2, \ldots, N_s$, are obtained
10: **end for**
11: Compute weights, $w_k$, $k = 1, 2, \ldots, N_s$
12: Approximate posterior, $\hat{\pi}(\theta \mid y)$ is given by Eq. (6.14) analytically

Table 6.1: Parameter true values, CDIS estimates and prior parameters

| Reactions | Parameters | True values | MAP estimates | Prior Parameters | | |
|---|---|---|---|---|---|---|
| | | $\theta_0$ | $\hat{\theta}_{\text{CDIS}}$ | $\hat{\theta}_{\text{prior}}$ | $a$ | $b$ |
| Reaction 1 | $k_1$ | 0.04 | 0.0321 | 15 | 1.01 | 0.00067 |
| Reaction 2 | $k_2$ | 0.11 | 0.148 | 8 | 1.01 | 0.00125 |
| Time taken ($N_s = 10^4$) = 3,266 seconds | | | | | | |

Table 6.2: Parameter true values, CDIS estimates for different $N_s$

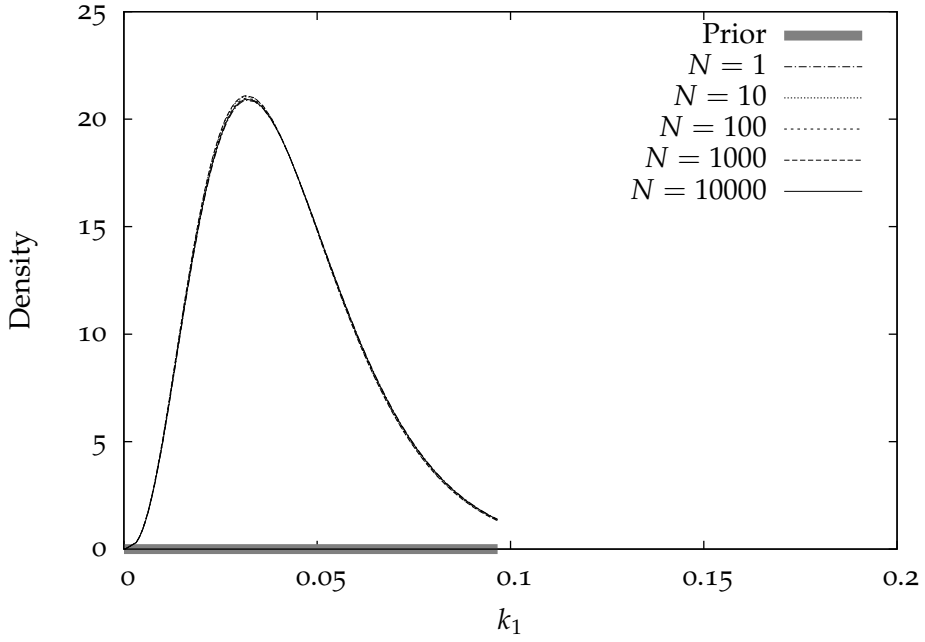
| Reactions | Parameters | True values | MAP estimates, $\hat{\theta}_{\text{CDIS}}$ | | | |
|---|---|---|---|---|---|---|
| | | $\theta_0$ | $N_s = 1$ | $N_s = 10^2$ | $N_s = 10^3$ | $N_s = 10^4$ |
| Reaction 1 | $k_1$ | 0.04 | 0.0322 | 0.0320 | 0.0319 | 0.0321 |
| Reaction 2 | $k_2$ | 0.11 | 0.151 | 0.150 | 0.150 | 0.148 |

### 6.1.4 *A Simple Example*

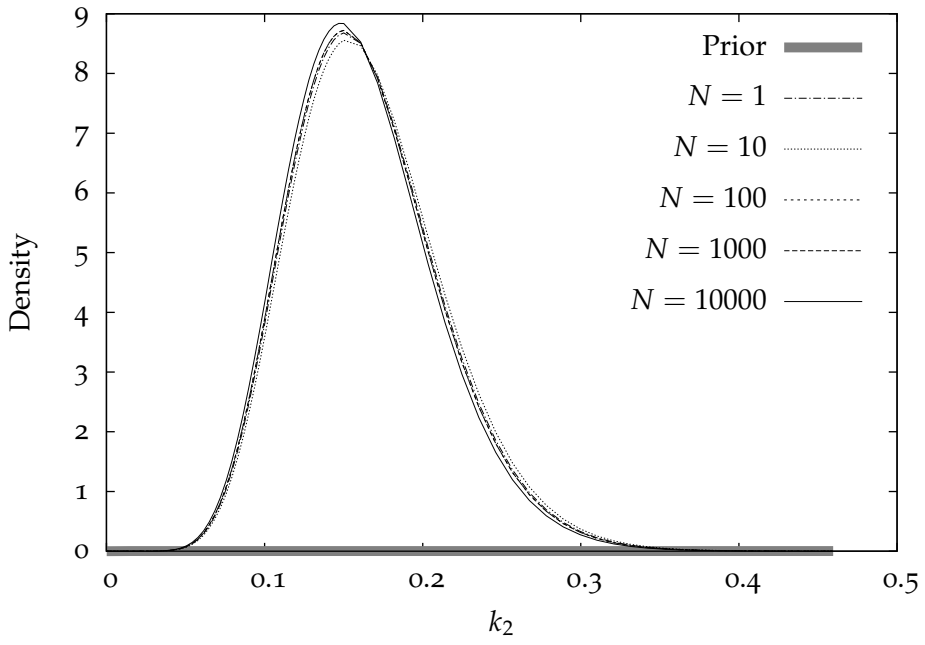
I demonstrate the application of CDIS method using the same example and measurement data in Figure 5.1. Again, the same gamma prior was used as in the previous examples to allow a comparison. Various values of $N_s$ were chosen to demonstrate the convergence of the posterior, as shown in Figures 6.1a-6.1b and Table 6.2. Note that even with just one sample, the parameter estimates are close to the exact estimates in Table 5.1. However, even if a large number of samples are used, the computational time taken for sampling is about three time less than that required by the MCMC-MH method (see Table 5.6). The convergence of the posteriors is very good as can be seen in Figures 6.1a-6.1b.

## 6.2 ESTIMATION USING APPROXIMATE DIRECT METHODS

This is a new class of methods that was developed during my research and at the of writing this dissertation had no close related literature methods. The approximate maximum likelihood (AML) method of Reinker et al., 2006 [87] has some limited similarity with an argument presented in this section. As mentioned in




(a) Marginal prior, $\pi(k_1)$ and marginal posterior, $\pi(k_1 \mid y)$




(b) Marginal prior, $\pi(k_2)$ and marginal posterior, $\pi(k_2 \mid y)$

Figure 6.1: Marginal priors and posteriors obtained using CDIS method with $N_s = 1$, $N_s = 10$, $N_s = 100$, $N_s = 1000$, $N_s = 10^4$

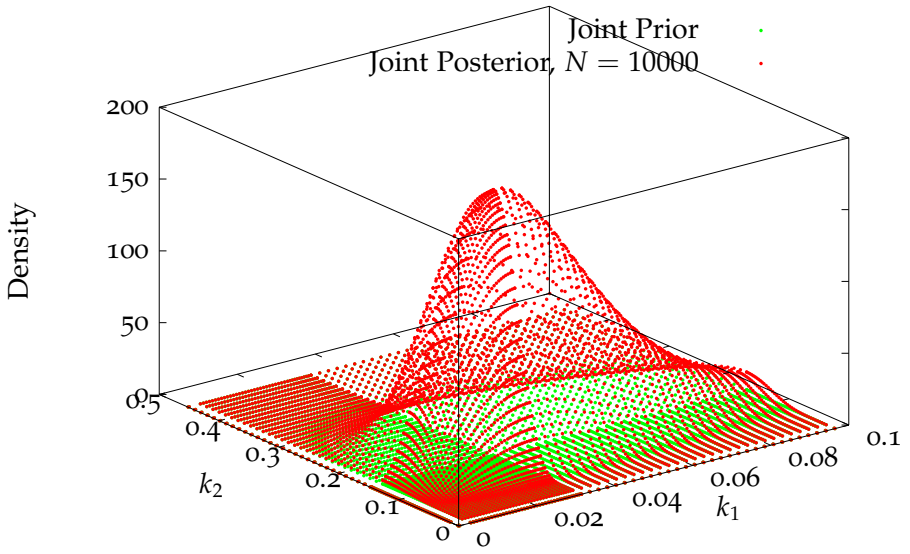



Figure 6.2: Joint and posterior obtained using CDIS method with $N_s = 10^4$.

the introduction to this chapter, the class of *approximate direct* (AD) methods is developed based on the insights obtained from the analytical expression of the CDIS approximate posterior, $\hat{\pi}(\theta \mid y)$ (in Eq. (6.14)), which in turn depends on the complete-data posterior, $\pi(\theta \mid x)$ (in Eq. (5.23)).

Considering the case of the complete-data posterior, $\pi(\theta \mid x)$, it is easy to see that $r_i$ and $G_i$, $i = 1, 2, \ldots, n_r$, form the set of *sufficient statistics* for the posterior $\pi(\theta \mid x)$. In other words, the posterior depends on $x$ through $r_i$ and $G_i$. Thus, if instead of the entire complete-data trajectory, $x$, only $r_i$ and $G_i$, $i = 1, 2, \ldots, n_r$ are provided, it is possible to obtain the posterior without any loss of information.

The CDIS posterior, $\hat{\pi}(\theta \mid y)$, is just a weighted sum of the complete-data posteriors. Thus, the only information needed to generate $\hat{\pi}(\theta \mid y)$, is $w_k$, $k = 1, 2, \ldots, N_s$, and $r_i$ and $G_i$, $i = 1, 2, \ldots, n_r$. However, the weights do contain other pieces of information. Assuming that it is possible to approximate the CDIS pos-

terior, $\hat{\pi}(\theta \mid y)$, further in the following manner

$$\begin{aligned}\hat{\pi}(\theta \mid y) &= \sum_{k=1}^{N_s} w_k \; \pi(\theta \mid x_k) \\ &\approx \pi(\theta \mid \bar{x}) \end{aligned} \tag{6.57}$$

in which, $\bar{x}$ is some "approximation" of the weighted sum of the samples, $x_k$, $k = 1, 2, \ldots, N_s$. Under some conditions, a possible approximation could be

$$\bar{x} = w_1 x_1 \oplus w_2 x_2 \oplus \cdots w_{N_s} x_{N_s} \tag{6.58}$$

in which $\oplus$ denotes an "appropriate" combination of two "weighted" complete-data trajectories. Another possibility is to choose only the sample with the highest weight

$$\bar{x} = x_j : w_j > w_k \;\; \forall \; k = 1, 2, \ldots, N_s \tag{6.59}$$

Thus, if an approximation in Eq. (6.57) is possible, then we obtain a single term which is a product of gamma distributions. Let

$$\bar{r}_i = r_i(\bar{x}) \quad i = 1, 2, \ldots, n_r \tag{6.60}$$

$$\bar{G}_i = G_i(\bar{x}) \quad i = 1, 2, \ldots, n_r \tag{6.61}$$

The approximate posterior in Eq. (6.57) may then be given as

$$\pi(\theta \mid \bar{x}) = \prod_{i=1}^{n_r} \pi(k_i \mid \bar{x}) \tag{6.62}$$

$$\pi(k_i \mid \bar{x}) = Ga(a_i + \bar{r}_i, b_i + \bar{G}_i) \quad i = 1, 2, \ldots, n_r \tag{6.63}$$

in which $\pi(\theta \mid \bar{x})$ is named the approximate direct (AD) posterior. The above expression (which is the result of the sufficient statistics argument) indicates that

a "suitable" weighted sum of $x_k$, $k = 1, 2, \ldots, N_s$, in Eq. (6.58), may be defined only over the values of the sufficient statistics, *i.e.*, $r_i$ and $G_i$, $i = 1, 2, \ldots, n_r$.

$$\bar{r}_i = \sum_{k=1}^{N_s} w_k \, r_{i,k} \tag{6.64}$$

$$\bar{G}_i = \sum_{k=1}^{N_s} w_k \, G_{i,k} \tag{6.65}$$

However, it appears that the weights still need to be computed. Equation (6.63) suggests another method for approximating the posterior. it appears that the form of the posterior is approximately "gamma-like" with two unknown parameters, $\bar{r}_i$ and $\bar{G}_i$. Note that $\bar{r}_i$ denotes the average number of times reaction $\mathcal{R}_i$, $i = 1, 2, \ldots, n_r$ occurred over the time interval $[s_0, s_m]$. Consider the following thought experiment — the true parameter values, $\theta_0$ is provided; how can we compute $\bar{r}_i$ and $\bar{G}_i$. The obvious (and naive) solution is to perform many SSA simulations to generate samples of $x$, then reject the simulations which do not agree with the measurement data, $y$, and using only the accepted samples of $x$, compute a simple average of $r_i$. Similar arguments hold for $\bar{G}_i$.

Let us consider the following series of examples. First, instead of measurement data $y$, complete-data trajectory $x$ is available. Then $r_i$ and $G_i$ are known and substituting these values in the expression of AD posterior (in Eq. (6.63)), we indeed obtain the true complete-data posterior, $\pi(\theta \mid x)$ (in Eqs. (5.22)-(5.23)). Thus, the AD posterior agrees with the true posterior when complete data is provided.

Second, let us assume we do not have complete-data $x$. Instead, measurement data $y$, with very high sampling frequency is available, *i.e.*, $\Delta s \ll \epsilon$. In such a case, it may be assumed that at most one reaction occurs per time interval. As a result, we can obtain an almost exact estimate of $r_i$. Since $\Delta s \ll \epsilon$, $G_i$ may be assumed fixed at the left-endpoint, with very little error. These values of $r_i$ and $G_i$

may now be substituted back into the expression of AD posterior to obtain very good estimates. Note that, as $\Delta s \to 0$, $y \to x \implies \pi(\theta \mid \bar{x}) \to \pi(\theta \mid x)$.

Third, when the reaction kinetics is non-cyclical and full measurements are available, then, irrespective of the value of $\Delta s$, $r_i$ is known exactly (*i. e.* deterministically) by solving the stoichiometric equation in Eq. (5.8). In such a case, the shape parameter of the AD posterior is known with absolute certainty. The rate parameter which depends on $G_i$ may now be approximated by assuming linearly varying propensities, *i. e.*, for the $j^{\text{th}}$ complete-data interval, $x_{[s_{j-1},s_j]}$, $j = 1, 2, \ldots, m$,

$$\begin{aligned} G_i &= \int_{s_{j-1}}^{s_j} g_i(\mathbf{x}(t))dt \\ &\approx \frac{g_i(\mathbf{x}(s_{j-1})) + g_i(\mathbf{x}(s_j))}{2}\,(s_j - s_{j-1}) \end{aligned} \tag{6.66}$$

Such an assumption is used by Wilkinson, 2012 [119] in a completely different framework to generate an approximate proposal process (see Section 5.5). Obviously, other approximations for $G_i$ may be obtained, for example, by assuming a maximum number of reactions that can take place during a complete-data time interval Reinker et al., 2006 [87] and integrating over reaction times using a truncated gamma-distribution (see Section 6.1.2 and Appendix A).

Fourth, consider the remaining cases in which $r_i$ has to be sampled. A biased estimate of the true posterior may be obtained by considering the minimum number of reactions $\mathcal{R}_i$, $i = 1, 2, \ldots, n_r$. This assumption is similar to the parsimony assumption of Reinker et al., 2006 [87]. Alternatively, if the available data has some intervals with a small $\Delta s$ value, then an iterative procedure may be employed to generate good estimates of $(r_i,\ G_i)$. Beginning with the *most informative* interval (possibly with the smallest $\Delta s$), estimate $r_i = r_{i,\min}$ and $G_i$ using Eq. (6.66). Substituting these $(r_i, G_i)$ values into the expression of AD posterior, obtain the mode of

Table 6.3: Parameter true values, AD estimates and prior parameters

| Reactions | Parameters | True values | MAP estimates | Prior Parameters | | |
|---|---|---|---|---|---|---|
| | | $\theta_0$ | $\hat{\theta}_{\text{AD}}$ | $\hat{\theta}_{\text{prior}}$ | $a$ | $b$ |
| Reaction 1 | $k_1$ | 0.04 | 0.0318 | 15 | 1.01 | 0.00067 |
| Reaction 2 | $k_2$ | 0.11 | 0.148 | 8 | 1.01 | 0.00125 |
| Time taken= 0.008 seconds | | | | | | |

AD posterior as $\hat{\theta}$. Given, $\hat{\theta}$, obtain better estimates of $(r_i, G_i)$.

In practice, it is observed that the above AD method produces reasonable estimates when the following conditions hold

1. Non-cyclical kinetics
2. Full-measurements
3. High quality measurement data (small $\Delta s$, multiple datasets)

Obviously, these restrictions make the AD method rather unusable. Before we discuss this issue, I demonstrate the use of AD method for the same example in Figure 5.1.

Table 6.3 shows the results of AD method. The marginal posteriors are presented in Figure 6.3 and the joint posterior in Figure 6.4. These results show an excellent agreement with the exact method results. For these results, Eq. (6.66) was used to compute $G_i$, $i = 1, 2, \ldots, n_r$. The most important feature, however, is the amount of sampling time taken — only 0.008 seconds. Since no sampling was required, the AD method provides a result instantaneously. Comparing these results with MCMC-MH and CDIS method, we see a tremendous difference in computational cost.

Coming back to the issue of usability, it is expected that if high quality measurement data is available, then the average estimates of $r_i$ and $G_i$ obtained using the methods described above would provide reasonable parameter estimates, even

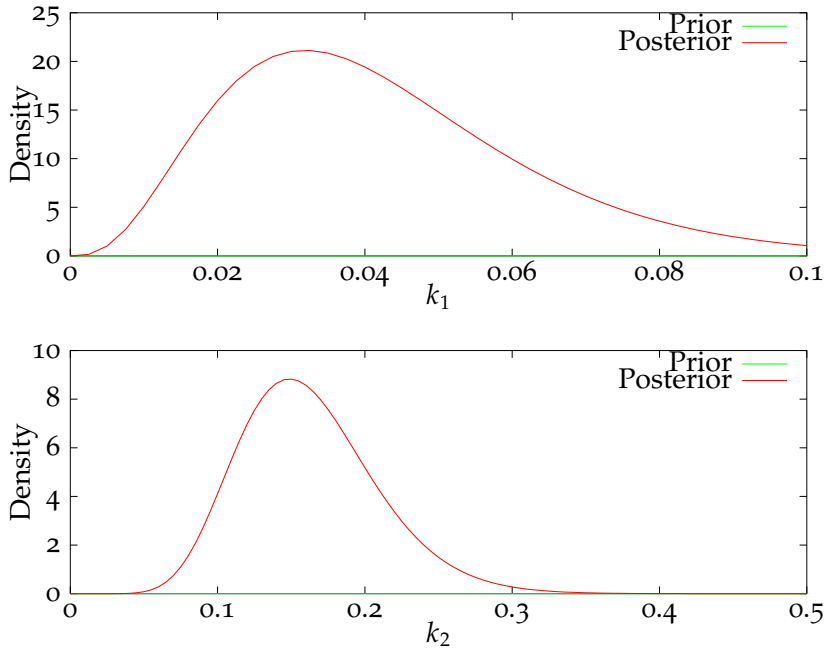



Figure 6.3: Marginal priors and posteriors obtained using approximate direct (AD) method

in the case of partial measurements and reversible kinetics. Further, in the presence of high-throughput data, all simulation methods, would require prohibitively larger computational effort. In such cases, estimates obtained using AD methods, though approximate, would provide tremendously quick results. An important point to note is that simulating the entire complete-data trajectory $x$ is the most computationally expensive component of any simulation method. But, $x$ is not actually required. Instead only $r_i$ and $G_i$ are required, Thus, instead of simulating from $\pi(x \mid \theta, y)$, if we can sample from $\pi(r_i, G_i \mid \theta, y)$ instead, it would be possible to obtain the posterior without any loss of information. Finally, this class of methods is new and as more research effort is directed into understanding the endpoint-conditioned stochastic processes, better estimates of $r_i$ and $G_i$, and in turn of $\theta$, will become available.

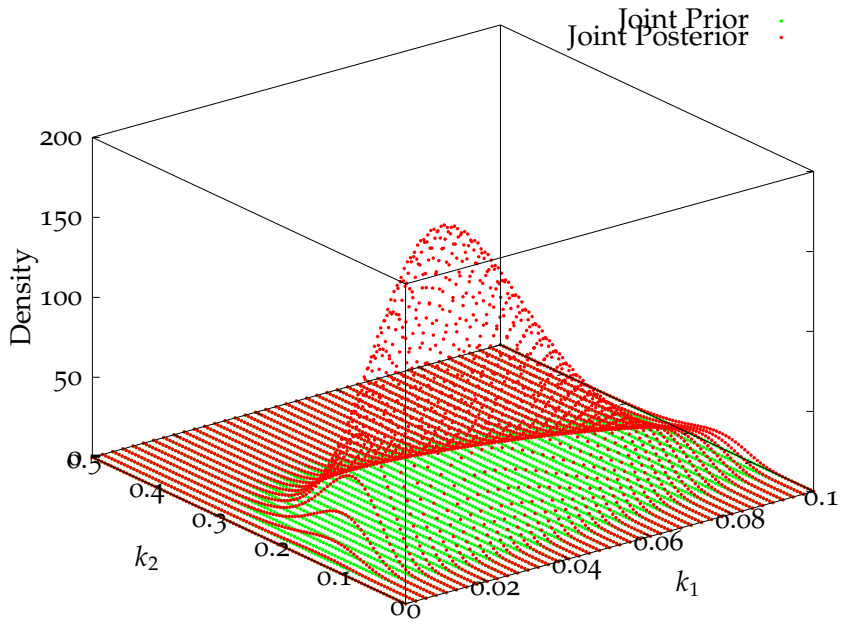



Figure 6.4: Joint and posterior obtained using approximate direct (AD) method with $N_s = 10^4$.

## 6.3 A SIMPLE EXAMPLE: FINAL COMPARISON

In this section, I compare the all the methods used to estimate the parameters from the measurement data in Figure 6.5. Table 6.4 presents a summary of the parameter estimates obtained by each method along with the computational (sampling) time taken by each method. Not surprisingly, the approximate direct (AD) method takes almost no time because there is no sampling involved. In this example, the reaction kinetics are not cyclical and full measurements are available, which means that AD method requires no sampling at all. However, surprisingly, the parameter estimates and the marginal posteriors obtained using the AD method are an excellent match to the exact method. The marginal posteriors produced by each method is shown in Figures 6.6a-6.6b. The CDIS method (with $N_s = 10^4$ samples) requires 10 times fewer samples and three times less computational time than the

Table 6.4: Parameter true values and estimates from all methods

| Reactions | Parameters | True values | Exact | MCMC-MH $N_s = 10^5$ | CDIS $N_s = 10^4$ | AD |
|---|---|---|---|---|---|---|
| | | $\theta_0$ | $\hat{\theta}_{\text{exact}}$ | $\hat{\theta}_{\text{MCMC-MH}}$ | $\hat{\theta}_{\text{CDIS}}$ | $\hat{\theta}_{\text{AD}}$ |
| Reaction 1 | $k_1$ | 0.04 | 0.0318 | 0.0295 | 0.0321 | 0.0318 |
| Reaction 2 | $k_2$ | 0.11 | 0.148 | 0.150 | 0.148 | 0.148 |
| Time taken (in seconds) | | - | - | 11,616 | 3,266 | 0.008 |

MCMC-MH method for a similar parameter estimate.

Also note that while the MCMC methods provide the posterior as a histogram, the CDIS and AD methods provide the posterior as analytical expressions. These expressions may be used to analytically evaluate the mean and variance of the posterior and may be used for further analysis. The histogram provided by MCMC methods has to be fitted with a function to allow such use. Further, as discussed in Chapter 7, in many cases the dataset provides no information about the parameters. Consequently, the marginal prior is returned as the marginal posterior. Since the MCMC methods provide a histogram, far too many samples are required to identify that the posterior is the same as the prior. The CDIS method on the hand provides this information analytically.

This example allowed us to compare all the methods against the exact method, thus providing an assurance of reasonable accuracy. In Chapter 7, I apply the CDIS and MCMC-MH methods on two models from systems biology, which yield further conclusions.

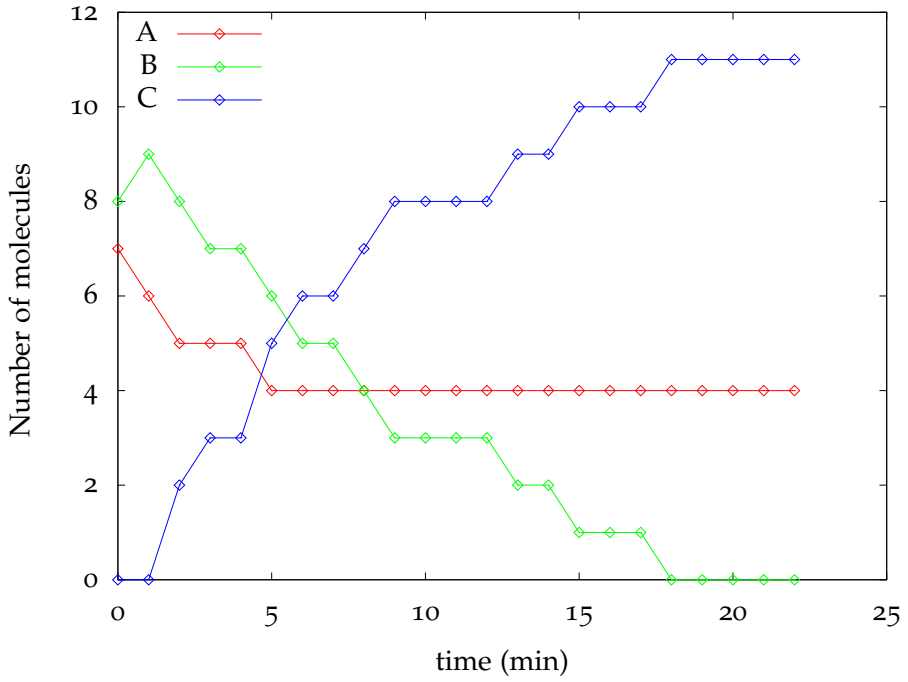
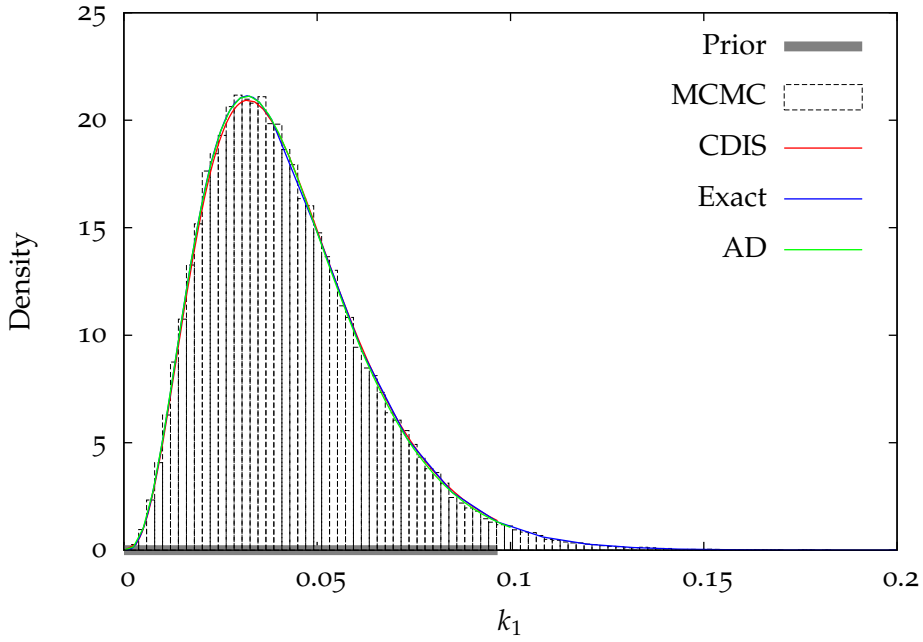
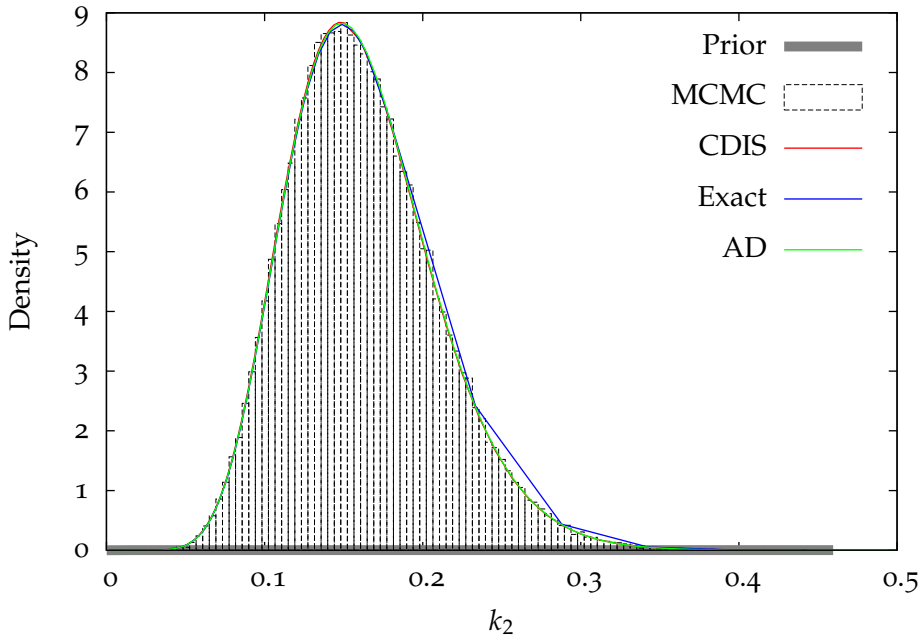



Figure 6.5: Measurement data ($y$) simulated using true parameters, $\theta_0 = [0.04\ 0.11]^T$ and the initial conditions $\mathbf{X}(0) = [7\ 8\ 0]^T$.




(a) Marginal prior, $\pi(k_1)$ and marginal posterior, $\pi(k_1 \mid y)$




(b) Marginal prior, $\pi(k_2)$ and marginal posterior, $\pi(k_2 \mid y)$

Figure 6.6: Marginal priors and posteriors using all estimation methods

# 7

# PARAMETER ESTIMATION IN SYSTEMS BIOLOGY

**Note**: Parts of this chapter appear in Gupta and Rawlings, 2013 [43].

The reaction kinetics in systems biology (whether deterministic or stochastic) differ from the reaction kinetics in chemistry in one large respect – systems biology models usually violate mass balance [36, 87, 81, 114, 39, 88, 14, 50, 105, 6, 119, 21]. Such reaction kinetics can result in an unbounded number of molecules (usually as a result of a *synthesis* reaction) causing the state space $\mathbb{S}$ to become unbounded as well. Violation of mass balance is an ease-of-use approximation in which we do not explicitly model species (for example, amino acids and nucleic acids) that (1) are present in large amounts, (2) are not expected to not vary appreciably, and (3) are not critical to the reaction system we intend to model. Balancing the reactions by explicitly modeling "missing" species does not improve the situation. The state space, $\mathbb{S}$, remains prohibitively large (though bounded) due to the species present in large amounts and the increased number of species causes an unnecessary increase in model complexity and computation. Unboundedness of state space, $\mathbb{S}$, usually rules out the application of exact and MCMC-Unif methods, except when an analytical solution to Eq. (5.56) is available. Our first example, early viral gene

expression, in Section 7.1 has this feature.

Another common feature that complicates parameter estimation is cyclical kinetics. Cyclical kinetics includes reversible reactions as well as oscillatory kinetics such as the Lotka-Voltera model. We define cyclical kinetics as the set of reactions whose transposed stoichiometric matrix $\nu^T$ has less than full column rank (see Section 2.4). As discussed in Section 6.1.2 when $\nu^T$ does not have full column rank, Eq. (5.8) cannot be solved uniquely for $\mathbf{r}$, thus resulting in at least one degree of freedom. In such a case, a large, possibly infinite, number of (hidden) reaction events may occur between two measurements. As a result, the amount of hidden data to be "averaged" over is much larger, which demands more computational effort. The gene on-off example in Section 7.1 has this feature.

## 7.1 EXAMPLE 1: EARLY VIRAL GENE EXPRESSION

The motivation for this example comes from the virus infection experiments performed in [107]. The following model describes the viral infection of a single *baby hamster kidney* (BHK-21) cell by *Vesicular stomatitis virus* (VSV) . The rationale behind this reaction mechanism comes from both the VSV-BHK biology and the characteristics of the data.

$$\begin{aligned} \mathrm{V} &\xrightarrow{k_a} \mathrm{G_1} & \mathrm{G_n} &\xrightarrow{k_a} \mathrm{G} & \mathrm{G} &\xrightarrow{k_G} \varnothing \\ \mathrm{G_1} &\xrightarrow{k_a} \mathrm{G_2} & \mathrm{G} &\xrightarrow{k_t} \mathrm{G+M} & \mathrm{M} &\xrightarrow{k_M} \varnothing \\ &\cdots & \mathrm{G} &\xrightarrow{k_r} 2\mathrm{G} & \mathrm{P} &\xrightarrow{k_P} \varnothing \\ \mathrm{G_{n-1}} &\xrightarrow{k_a} \mathrm{G_n} & \mathrm{M} &\xrightarrow{k_f} \mathrm{M+P} \end{aligned} \tag{7.1}$$

The viral infection process takes place in five distinct stages [94]: adsorption to the host cell and entry, uncoating of the genome, transcription and translation, genome replication, assembly and release of virus progeny. The first two stages, namely, adsorption/entry and uncoating must occur before the viral genome un-

dergoes transcription or replication. A delay is observed before viral protein begins to appear.

In Eq. (7.1), V represents *unactivated* viral genome and G represents *activated* viral genome. The species $G_1, G_2, \ldots, G_n$ represent various intermediate (or delay) states of the viral genome. Unactivated viral genome, V, has to undergo adsorption, entry and uncoating before it is ready for transcription and replication. The reactions in Eq. (7.1) converting V to G represent these initial stages of adsorption, entry and uncoating. The species M and P represent RFP mRNA and red fluorescent protein, respectively. The transcription reaction creates more mRNA, M from activated genome G. Translation creates more red fluorescent protein, P from mRNA, M. Observed RFP signal is proportional to the amount of red fluorescent protein P present in the cell. Here any delays associated with maturation of the RFP signal are embedded in the translation rate.

During the early stages of infection, the degradation of G, M, P is not important. Therefore, we further reduce the model in (7.1) to represent only the early stages of infection and remove the delay states. This leads to the following reduced model:

$$\begin{aligned} \mathrm{V} &\xrightarrow{k_a} \mathrm{G} \\ \mathrm{G} &\xrightarrow{k_t} \mathrm{G} + \mathrm{M} \\ \mathrm{G} &\xrightarrow{k_r} 2\mathrm{G} \\ \mathrm{M} &\xrightarrow{k_f} \mathrm{M} + \mathrm{P} \end{aligned} \tag{7.2}$$

Using true parameter values $\theta_0$ as shown in Table 7.1 and an initial condition of $\mathbf{X}(0) = \begin{bmatrix} \mathrm{V}_0 & \mathrm{G}_0 & \mathrm{M}_0 & \mathrm{P}_0 \end{bmatrix}^T = \begin{bmatrix} 10 & 0 & 0 & 0 \end{bmatrix}^T$ (corresponding to a multiplicity of infection or MOI of 10), we generate six simulated trajectories with Gillespie's direct method [32]. These trajectories, shown in Figure 7.1, use the same random

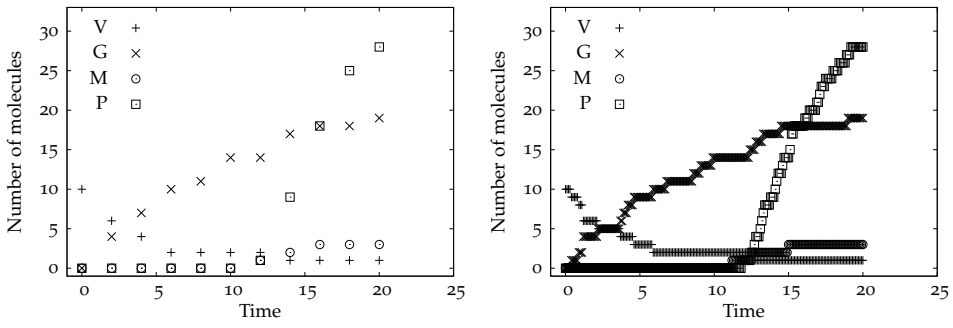
Table 7.1: True rate constants and gamma prior parameters

| Reactions | Parameters | True values | Prior Parameters | | |
|---|---|---|---|---|---|
| | | $\theta_0$ | $\hat{\theta}_{\text{prior}}$ | $a$ | $b$ |
| Activation | $k_a$ | 0.15 | 10 | 1.01 | 0.0010 |
| Transcription | $k_t$ | 0.02 | 20 | 1.01 | 0.0005 |
| Replication | $k_r$ | 0.05 | 1 | 1.01 | 0.0100 |
| Translation | $k_f$ | 1.00 | 30 | 1.01 | 0.0003 |

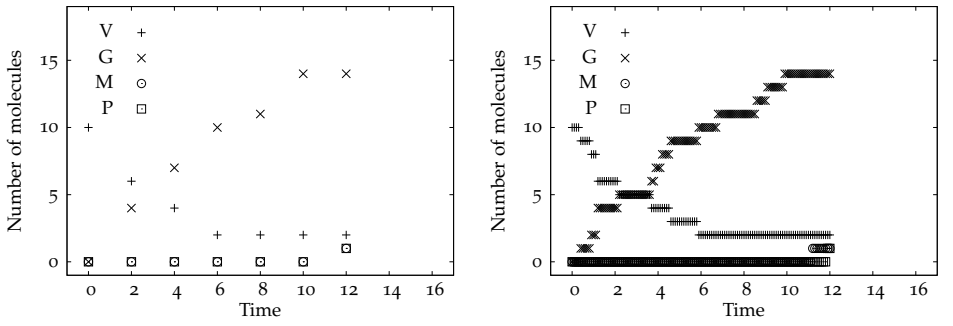
numbers but have different sampling frequency ($\Delta s$) and different number of time points ($m$). Note that Figures 7.1a and 7.1b span the same time interval of $[0, 20]$ but Figure 7.1b has 20 times larger measurement frequency. The same is true for the next two rows of Figure 7.1.

As we can see in (7.2), all reactions except the activation reaction violate mass balance. This rules out the application of exact and MCMC-Unif methods and we only compare the MCMC-MH and CDIS methods. We use a gamma prior as described in Equation (5.1) with shape ($a$) and rate ($b$) parameters shown in Table 7.1. The mode (or peak) of the gamma prior, denoted by $\hat{\theta}_{\text{prior}}$ is also shown in Table 7.1. Note that $\hat{\theta}_{\text{prior}}$ is chosen to be very different than $\theta_0$ so as to not bias the posterior towards the true values. The shape parameter is chosen so that the prior is essentially uniform over a large range of parameter values.

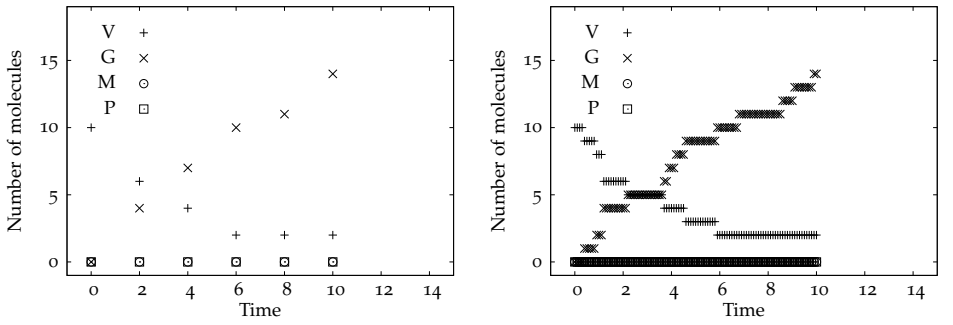
We estimates the parameters for each of the six trajectories using both CDIS (with $N = 1000$ samples) and MCMC-MH methods (with $N = 10000$ steps). The resulting parameter estimates (MAP estimates) are shown in Table 7.2. The marginal posteriors $\pi(k_i \mid y)$ are shown in Figure 7.2. As discussed before, CDIS produces a semi-analytical expression for the marginal posterior while the MCMC-MH method produces samples of $\theta$, which are then binned to produce the histogram. Parameter estimates for the CDIS method are obtained by numerical optimization. The estimates for MCMC-MH method are obtained by simply finding the bin with the largest probability (or height). Histogram estimation suffers from the problem of bias-variance tradeoff [115, p. 303]. As a result,




(a) $m = 10$, $\Delta s = 2$ (b) $m = 200$, $\Delta s = 0.1$




(c) $m = 6$, $\Delta s = 2$ (d) $m = 120$, $\Delta s = 0.1$




(e) $m = 5$, $\Delta s = 2$ (f) $m = 100$, $\Delta s = 0.1$

Figure 7.1: VSV early gene expression. True parameter values of $\theta_0 = \begin{bmatrix} k_a & k_t & k_r & k_f \end{bmatrix}^T = \begin{bmatrix} 0.15 & 0.02 & 0.05 & 1 \end{bmatrix}^T$. Initial condition of $\mathbf{X}(0) = \begin{bmatrix} \mathrm{V}_0 & \mathrm{G}_0 & \mathrm{M}_0 & \mathrm{P}_0 \end{bmatrix}^T = \begin{bmatrix} 10 & 0 & 0 & 0 \end{bmatrix}^T$ corresponding to an MOI of 10. $m$ = number of points, $\Delta s$ = sampling time.

Table 7.2: Parameter estimates from CDIS and MCMC-MH. True values $(k_{a0}, k_{t0}, k_{r0}, k_{f0}) = (0.15, 0.02, 0.05, 1)$.

| Dataset | CDIS ($N = 1000$) | MCMC-MH ($N = 10000$) |
|---|---|---|
| (7.1a) | (0.1674, 0.0124, 0.0411, 1.2767) | (0.1457, 0.0117, 0.0380, 1.3568) |
| (7.1b) | (0.1716, 0.0122, 0.0407, 1.3163) | (0.1768, 0.0112, 0.0387, 1.3871) |
| (7.1c) | (0.1825, 0.0096, 0.0572, 0.8019) | (0.1907, 0.0082, 0.0597, 6.4541) |
| (7.1d) | (0.1832, 0.0095, 0.0567, 1.2069) | (0.1752, 0.0121, 0.0549, 1.1623) |
| (7.1e) | (0.1994, 0.0001, 0.0778, 30.0000) | (0.2012, 0.0006, 0.0700, 155.6734) |
| (7.1f) | (0.2018, 0.0001, 0.0770, 30.0000) | (0.2010, 0.0008, 0.0759, 166.1711) |

MCMC-MH estimates have an inherent error corresponding to the width of the bin ($\sim 10^{-2} - 10^{-5}$ for this example). CDIS estimates in contrast are accurate up to any desired level of accuracy ($\sim 10^{-10}$ due to machine precision).

We begin by comparing the results for trajectories in Figures 7.1a and 7.1b. Figures 7.2a and 7.2b show the corresponding marginal posteriors and priors. In comparison to the posteriors the priors are essentially uniform, indicating that the priors have little effect on the MAP estimates. The CDIS and MCMC-MH methods agree quite well. The first two rows of Table 7.2 provide the parameter estimates. The parameter estimates $\hat{\theta}$ are close to the true values $\theta_0$ indicating that the estimation methods are reasonably accurate. For any statistical estimation method, even the exact method, there is some estimator error (or bias) due to the finite amount of data. Thus, we do not expect the estimates in Table 7.2 to track the true values. However, we expect the parameter estimates, $\hat{\theta}$ to converge as the number of samples ($N$) increases.

Note that increasing the number of measurements by a factor of 20 (from 7.1a to 7.1b) does appreciably change the parameter estimates. Looking at the next trajectory in Figure 7.1c we see that the CDIS and MCMC-MH posteriors for translation rate constant, $k_f$, are different (Figure 7.2c). In this case, increasing the number of measurements causes the two posteriors to agree with each other (Figure 7.2d). Looking at the corresponding parameter estimates in Table 7.2, we

can see that increasing the measurement frequency changes the estimates for $k_f$ significantly but the other parameter estimates do not change. Finally, for the last two trajectories in Figures 7.1e and 7.1f, increased measurement frequency does not change the parameter estimates as well. However, for these trajectories, there are two interesting outcomes – (1) the parameter estimates for $k_t$ and $k_f$ are very different from their corresponding true values and (2) the posteriors for $k_f$ are the same as the prior.

MCMC-MH and CDIS methods use the complete-data posterior $\pi(\theta \mid x)$ (in Eqs. (5.22)-(5.23)) to generate an approximation of the required posterior $\pi(\theta \mid y)$. Eq. (5.23) offers some explanation about behavior of estimates. The marginal complete data posterior, $\pi(k_i \mid x)$ uses the information in $x$ only through $r_i$ and $G_i$. Here, $r_i$ is the number of times reaction $\mathcal{R}_i$ occurs in time interval $[s_0, s_m]$ and can be obtained uniquely by solving Equation (5.8) because $\nu^T$ has full column rank. Changing the measurement frequency while keeping the same end points $y_0$ and $y_m$ does not change $r_i$. But increasing the measurement frequency does provide a better (lower variance) estimate of $G_i$ (defined in Eq. (5.17)). This observation explains why the increasing measurement frequency has a weak effect on the posterior and estimates.

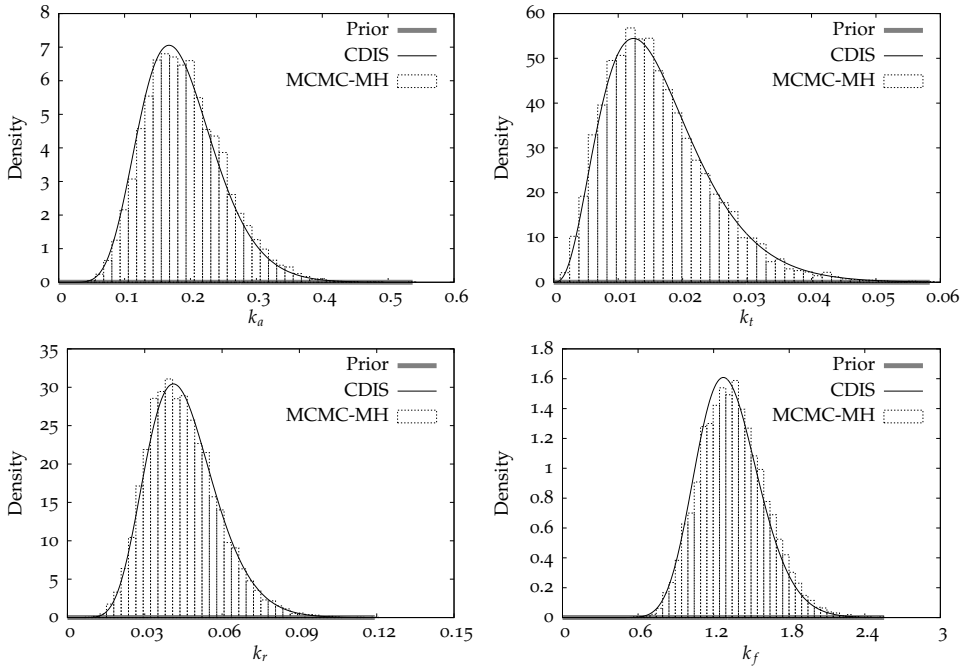
We now compare the data in Figures 7.1e, 7.1c, 7.1a. The trajectory in Figure 7.1e shows that the mRNA (M) and protein (P) measurements remain at zero and neither transcription nor translation reaction events occur during the given $[0, 5]$ time interval. As a result, every (event data) sample $x$ generated by the CDIS or MCMC-MH method has $r_4 = 0$ and $G_4 = 0$. Looking back at Eq. (5.23), we can see that this dataset provides no information at all for the translation reaction. Consequently, the posterior of $k_f$ is exactly equal to the posterior. While the CDIS method provides this information analytically, the MCMC-MH method provides this information through samples of $k_f$. In such a case, the model in (7.2) should be further reduced by removing the translation reaction. While $r_2$ is also equal

Table 7.3: Sampling time (seconds per sample) for CDIS and MCMC-MH.

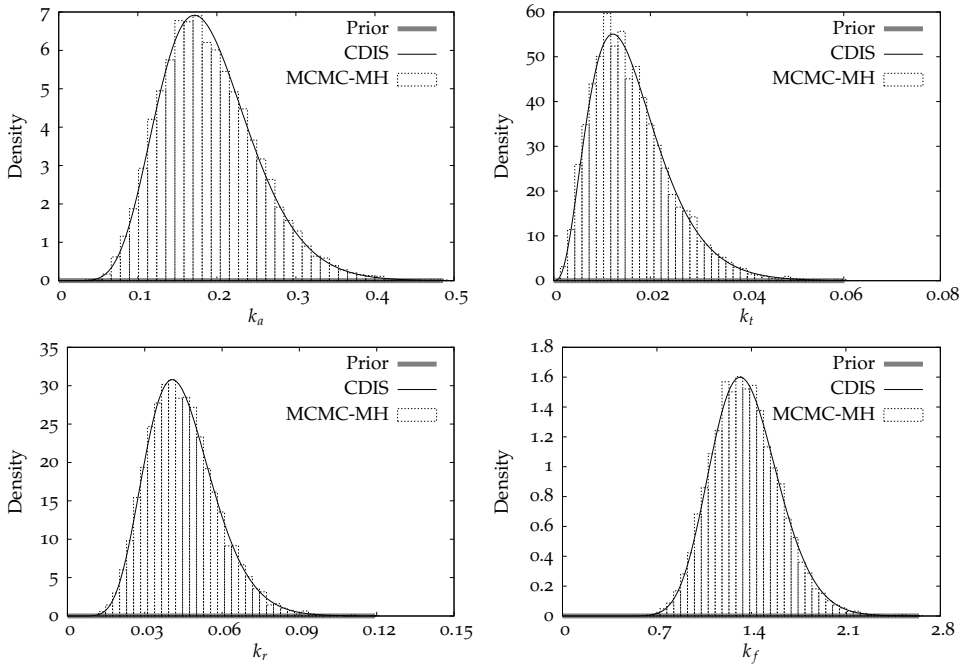
| Dataset | CDIS ($N = 1000$) | MCMC-MH ($N = 10000$) |
|---|---|---|
| (7.1a) | 0.54 | 0.12 |
| (7.1b) | 0.54 | 1.35 |
| (7.1c) | 0.30 | 0.06 |
| (7.1d) | 0.37 | 0.83 |
| (7.1e) | 0.30 | 0.05 |
| (7.1f) | 0.35 | 0.66 |

to zero, $G_2 > 0$, which means that there is some information available regarding the transcription rate constant $k_t$. Consequently, the estimate $\hat{k}_t$ in Table 7.2, while being very different than $k_{t,0}$, is not equal to $\hat{k}_{t,\text{prior}}$. Figures 7.1c and 7.1a have larger (non-zero) $(r_i, G_i)$ values and provide better estimates.

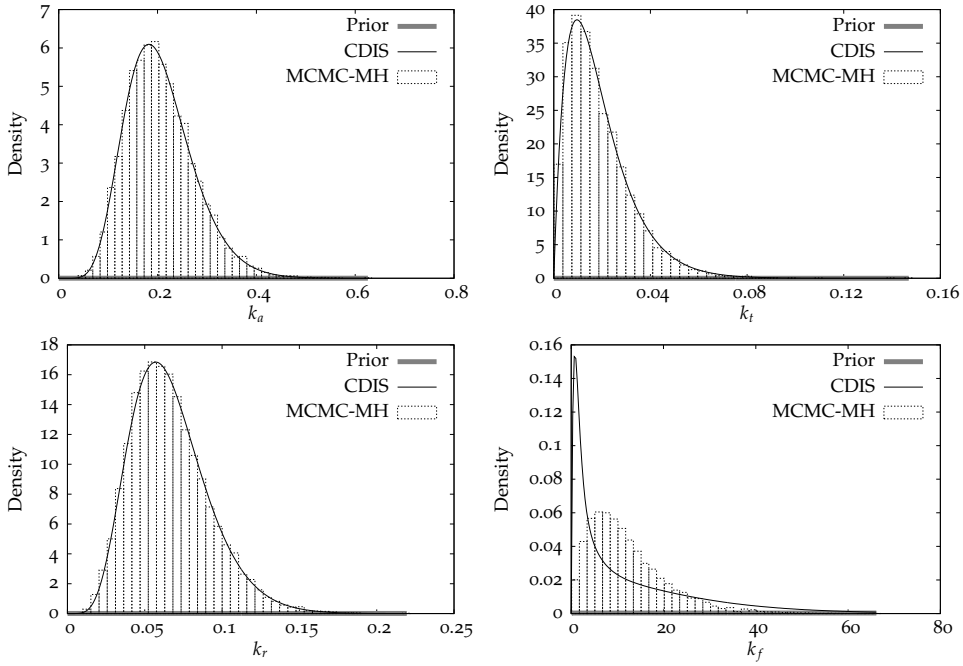
In Table 7.3, we present the computational time spent in terms of seconds taken to generate one sample of $x$ for both methods. All simulations were performed using Octave on an Intel(R) Core(TM) 2 Quad 8400 MHz, 8GB RAM machine running Lubuntu 13.04. For both methods, the computational time per sample is directly proportional to the number of time points, $m$ and is (roughly) inversely proportional to the sampling interval, $\Delta s$. Table 7.3 shows that when $m$ is low (in 7.1a, 7.1c, 7.1e), the CDIS method takes much longer to generate one sample of $x$ than MCMC-MH. However, when $m$ is high (in 7.1b, 7.1d, 7.1f), MCMC-MH is slower than CDIS. Irrespective of the per sample computational time, CDIS requires at least 10 times fewer samples than MCMC-MH. In fact, CDIS method produces a semi-analytical posterior with just one sample while a large ($\sim 1000$) samples are required to obtain a histogram using MCMC methods. Even further, a significant number of samples ($\sim 10 - 50\%$) have to be *burned-in* while using MCMC methods to ensure convergence of the Markov chain to stationarity. The CDIS method, by contrast, may have problems with convergence if the importance distribution $q(x)$ is not "close enough" to the target distribution $\pi(x \mid y)$. We have not encountered convergence issues for this example.



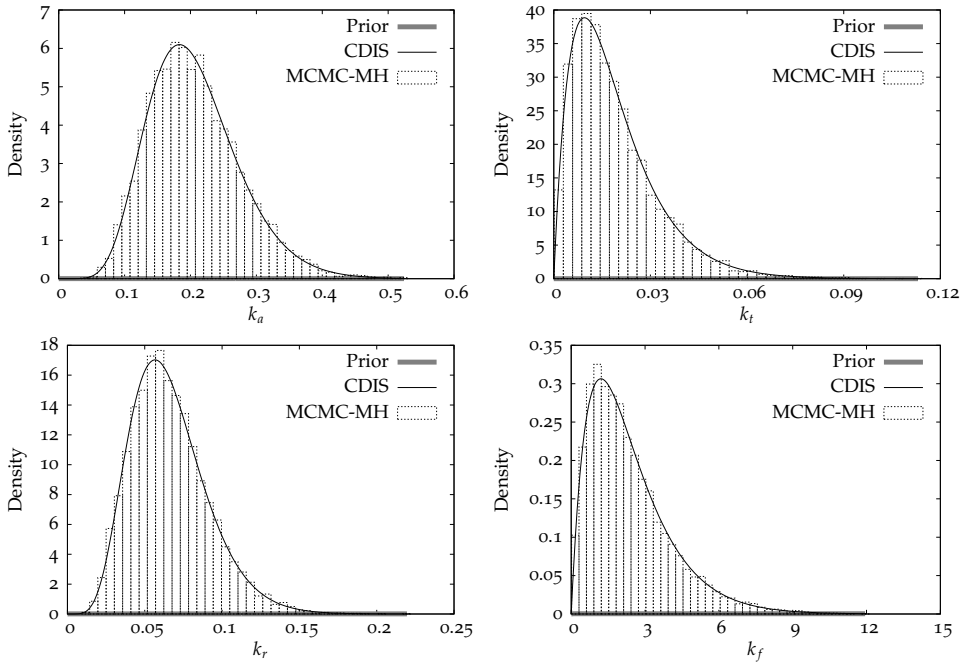

(a) $m = 10$, $\Delta s = 2$



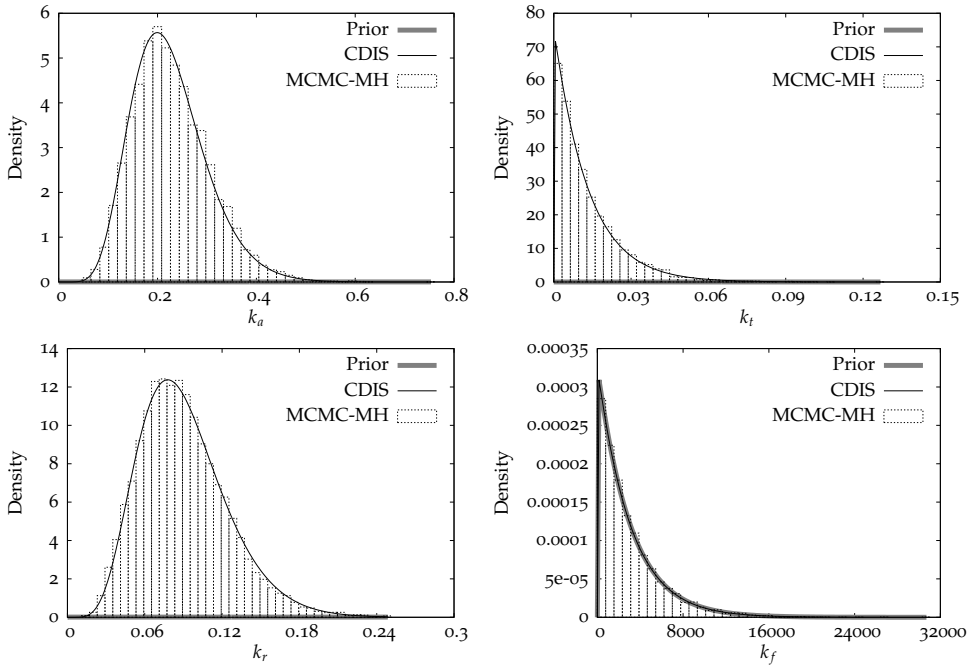

(b) $m = 200$, $\Delta s = 0.1$



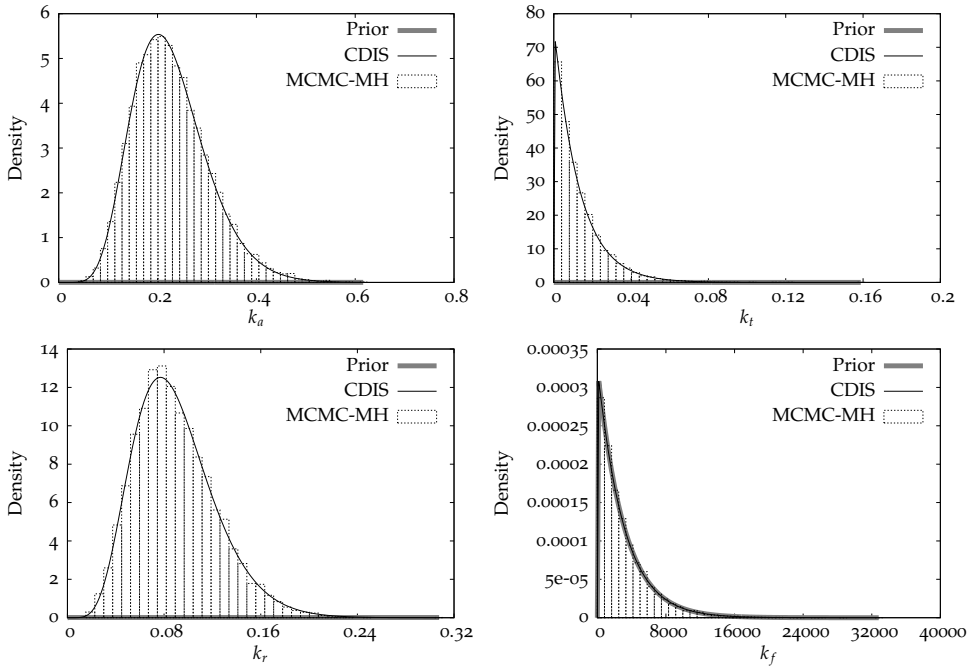

(c) $m = 10$, $\Delta s = 2$




(d) $m = 200$, $\Delta s = 0.1$




(e) $m = 10$, $\Delta s = 2$




(f) $m = 200$, $\Delta s = 0.1$

Figure 7.2: Marginal priors and posteriors obtained using CDIS and MCMC-MH.

The RNA expression model was first proposed by Golding et al. [36] and later studied by Reinker et al. [87] and Poovathingal and Gunawan [81]. This model, as shown in reactions (7.3), has three species and three reactions.

$$\begin{aligned} \mathrm{DNA_{ON}} &\xrightarrow{k_1} \mathrm{DNA_{OFF}} \\ \mathrm{DNA_{OFF}} &\xrightarrow{k_{-1}} \mathrm{DNA_{ON}} \\ \mathrm{DNA_{ON}} &\xrightarrow{k_t} \mathrm{DNA_{ON}} + \mathrm{mRNA} \end{aligned} \tag{7.3}$$

As discovered by Golding et al. [36], the number of RNA molecules in a single cell was around 0–10. This simple model also violates mass balance, which makes the exact and MCMC-Unif methods inapplicable. In order to demonstrate the exact and MCMC-Unif methods and compare them to MCMC-MH and CDIS methods, we consider only the first two reactions in the model. This simplification creates the reduced, gene on-off model presented in reactions (7.4).

$$\begin{aligned} \mathrm{DNA_{ON}} &\xrightarrow{k_1} \mathrm{DNA_{OFF}} \\ \mathrm{DNA_{OFF}} &\xrightarrow{k_{-1}} \mathrm{DNA_{ON}} \end{aligned} \tag{7.4}$$

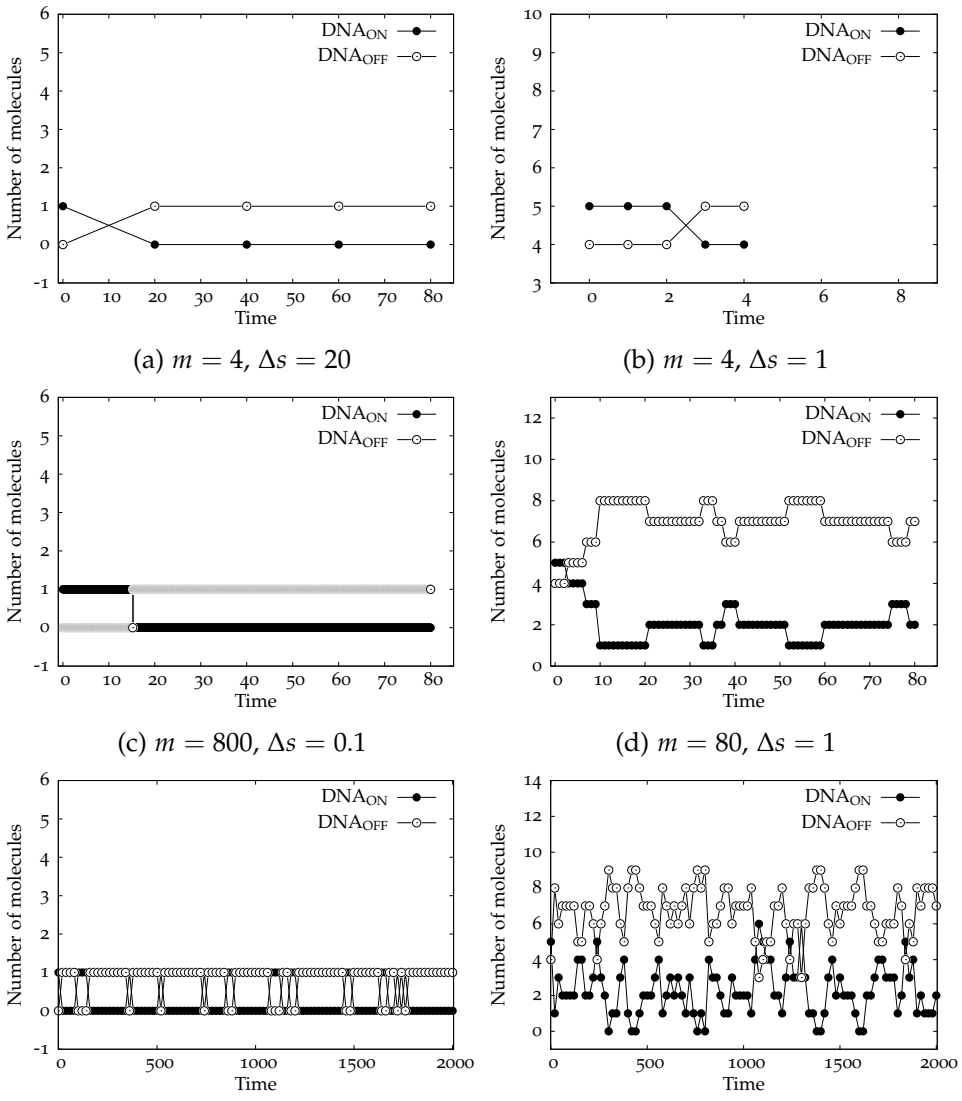
While this model is no longer suited to study mRNA expression, it can be used to study the switching – activation (ON) and deactivation (OFF) – of a gene. Since this model now has only two parameters to be estimated, we can plot likelihood and joint posterior values, which is another motivation for choosing this reduced model.

We generate six simulated trajectories using true parameter values $\theta_0 = \begin{bmatrix} k_1 & k_{-1} \end{bmatrix}^T = \begin{bmatrix} 0.03 & 0.01 \end{bmatrix}^T$ with Gillespie's direct method [32], as shown in Figure 7.3. Trajectories in the first column (Figures 7.3a, 7.3c, 7.3e) are simulated using an initial

condition of $\mathbf{X}(0) = \begin{bmatrix} \text{DNA}_{\text{ON},0} & \text{DNA}_{\text{OFF},0} \end{bmatrix}^T = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$, which corresponds to only one copy of the gene. The second column (Figures 7.3b, 7.3d, 7.3f) is simulated using $\mathbf{X}(0) = \begin{bmatrix} \text{DNA}_{\text{ON},0} & \text{DNA}_{\text{OFF},0} \end{bmatrix}^T = \begin{bmatrix} 5 & 4 \end{bmatrix}^T$, which corresponds to nine copies of the gene.

We begin by computing the likelihood $\pi(y \mid \theta)$ using the exact method as described in Section 5.2 for each trajectory. The resulting three-dimensional likelihood plots, corresponding to each of the trajectory, are presented in Figure 7.4. We can see that for the data in Figure 7.3a, the likelihood is essentially flat when $k_{-1}$ is small and $k_1$ is large enough. The likelihood plots for Figure 7.3b and 7.3c are flat when $k_{-1}$ is small but have a peak in $k_1$. The data in Figures 7.3d and 7.3e produce likelihood curves, which are peaked in both $k_1$ and $k_{-1}$ while the data in Figure 7.3f produces the most sharply peaked likelihood surface.

Intuitively, the likelihood in Figure 7.4a seems reasonable. The corresponding trajectory in Figure 7.3a may be recreated, with probability 1, if we set $k_1 = \infty$ and $k_{-1} = 0$, *i.e.*, $\pi\left(y \mid \theta = \begin{bmatrix} \infty & 0 \end{bmatrix}^T\right) = 1$. Therefore, the only information we can obtain is that $k_1 >> k_{-1}$; we cannot obtain independent estimates of both parameters. Figure 7.3b has the same number of measurements as Figure 7.3a but $k_1 = \infty$ is no longer possible because we observe that $\mathcal{R}_1$ does not occur until $t = 2$. This observation provides some information about $k_1$. Since $\mathcal{R}_2$ does not occur in Figure 7.3b, it is still possible for $k_{-1}$ to be zero. Data in Figure 7.3c has a large number of finely sampled measurements ($m = 800, \Delta s = 0.1$) but it still does not provide more information than Figure 7.3b, which has far fewer measurements. Figure 7.3c indicates that the gene switched off (*i.e.*, $\mathcal{R}_1$ occurred) once around $t = 15$. While the increased measurement frequency provides an accurate time of the switch (compared to Figures 7.3a and 7.3b), the trajectory can still be simulated when $k_{-1} = 0$. Thus, 7.3c provides no information about $\mathcal{R}_2$. In fact, a reduced model consisting of $\mathcal{R}_1$ alone may be used to explain this data. Figure 7.3d has 10 times fewer measurements than Figure 7.3c and spans




(a) $m = 4$, $\Delta s = 20$

(b) $m = 4$, $\Delta s = 1$

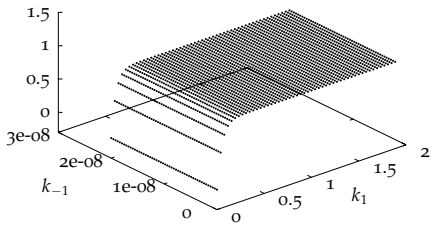
(c) $m = 800$, $\Delta s = 0.1$

(d) $m = 80$, $\Delta s = 1$

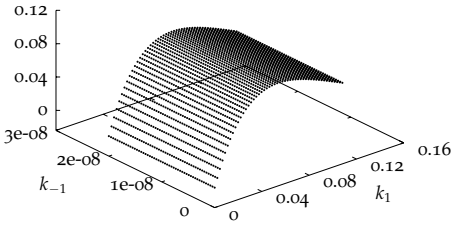
(e) $m = 100$, $\Delta s = 20$

(f) $m = 100$, $\Delta s = 20$

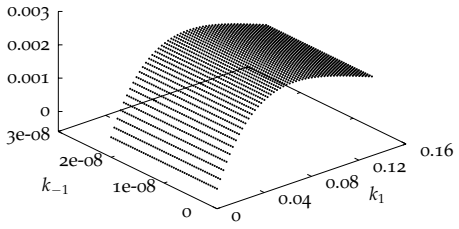
Figure 7.3: Gene on-off model. True parameter values of $\theta_0 = [k_1\ k_{-1}]^T = [0.03\ 0.01]^T$. First column: $\mathbf{X}(0) = [\text{DNA}_{\text{ON},0}\ \text{DNA}_{\text{OFF},0}]^T = [1\ 0]^T$. Second column: $\mathbf{X}(0) = [\text{DNA}_{\text{ON},0}\ \text{DNA}_{\text{OFF},0}]^T = [5\ 4]^T$. $m$ = number of points, $\Delta s$ = sampling time.



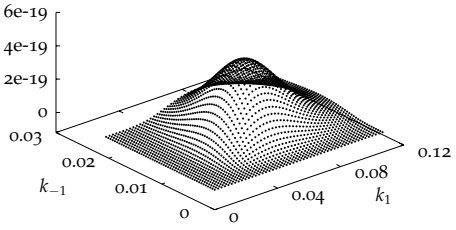

(a) $m = 4$, $\Delta s = 20$



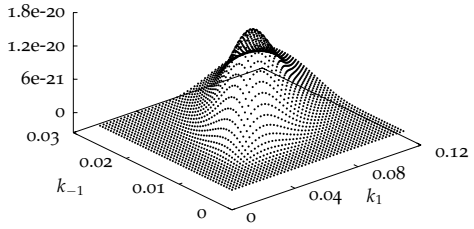

(b) $m = 4$, $\Delta s = 1$



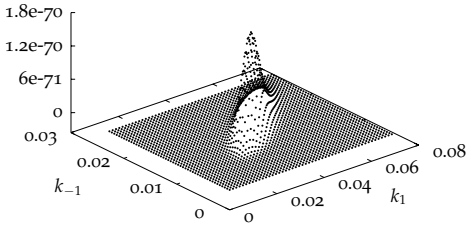

(c) $m = 800$, $\Delta s = 0.1$




(d) $m = 80$, $\Delta s = 1$




(e) $m = 100$, $\Delta s = 20$




(f) $m = 100$, $\Delta s = 20$

Figure 7.4: Gene of-off model. Exact likelihood plots for data in Figure 7.3

the same time interval of $[0, 80]$, but it provides a better estimate of $k_{-1}$. This is because both $\mathcal{R}_1$ and $\mathcal{R}_2$ occur often. Figure 7.3e also provides better estimates for the same reason. Note that the data in Figures 7.3d and 7.3e do not allow the possibility of $k_{-1} = 0$. Figure 7.3f provides the sharpest likelihood surface (and the best estimates) among all trajectories in spite of having a sampling time of $\Delta s = 20$.

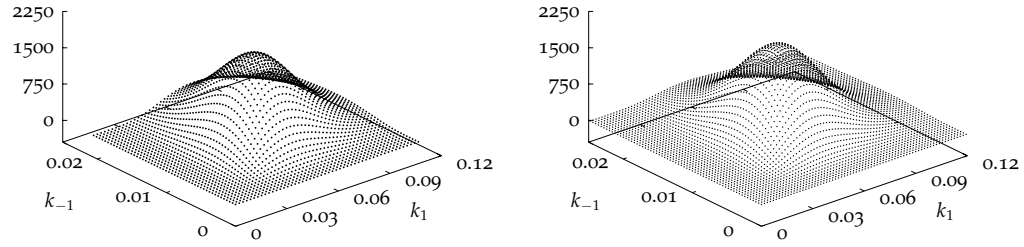
Unlike the previous example, $\nu^T$ for this example does not have full column rank. This means, as we intuitively expect, that we cannot uniquely obtain the number of times each reaction occurred (*i. e.*, $r_i$) between two measurements. For example, in Figure 7.3a, the number of times $\mathcal{R}_1$ and $\mathcal{R}_2$ occurred between $t = 0$ and $t = 20$ is given by $(r_1, r_2) = (n+1, n), n = 0, 1, 2 \ldots$. Therefore, we define $r_{i,\min}$ as the observed (or minimum) number of times a reaction $\mathcal{R}_i$ must occur. As we can see, for Figures 7.3a, 7.3c and 7.3b, $(r_{1,\min}, r_{2,\min}) = (1, 0)$, which indicates that these datasets do not provide information about $k_{-1}$. Figures 7.3d, 7.3e and 7.3f have larger (non-zero) values of $(r_{1,\min}, r_{2,\min})$ and provide better estimates.

We compare other estimation methods with the exact method for the data in Figure 7.3d. The joint posterior for all four estimation methods is presented in Figure 7.5. Parameter estimates obtained by (exact) maximum likelihood estimation and other methods are provided in Table 7.4. The estimates obtained by exact Bayesian estimation are close to the exact MLE estimates, which indicates that the prior does not bias the estimation. The sampling times are also presented in Table 7.4. MCMC-Unif method performs the worst due to slow mixing of the Markov chain (see supplementary document). Further, the uniformization step is extremely slow, which limits the number of Markov chain steps that can be used. Detailed analysis of computation time required by uniformization and other similar methods may be found in [56]. As we can see from Figure 7.5d, 1000 Markov chain steps were not enough to produce a three-dimensional histogram. Consequently, the MCMC-Unif estimates are not accurate. MCMC-Unif must be run

Table 7.4: Parameter estimates and sampling time (seconds per sample) for Figure 7.3d. True values $\theta_0 = (k_{1,0}, k_{-1,0}) = (0.03, 0.01)$.

| Parameter | Exact (MLE) | Exact (BE) | CDIS ($N = 100$) | MCMC-MH ($N = 1000$) | GibbsUnif ($N = 1000$) |
|---|---|---|---|---|---|
| $k_1$ | 0.0533 | 0.0534 | 0.0557 | 0.0556 | 11.9865 |
| $k_{-1}$ | 0.0103 | 0.0103 | 0.0111 | 0.0115 | 3.5485 |
| Sampling time (sec/sample) | – | – | 0.76 | 0.53 | 9.01 |

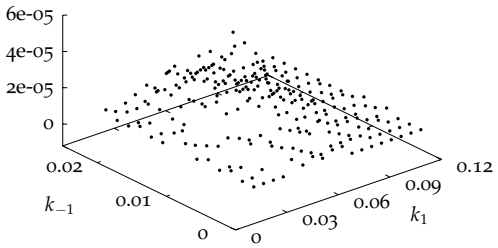
for a much larger number of steps to obtain reliable estimates. By contrast, CDIS ($N = 100$ samples) and MCMC-MH ($N = 1000$ steps) estimates are close to the exact (Bayesian) estimates. A burn-in of 100 steps was used for MCMC-MH method. Variance in CDIS estimates (using $N = 100$ samples) was lower than the variance in MCMC-MH estimates (using $N = 1000$ steps). Reducing the number of Markov chain steps used by the MCMC-MH method increases the variance in the resulting histogram and parameter estimates. Thus, the CDIS method requires ten times fewer number of samples compared to the MCMC-MH method to obtain the same accuracy in estimates. The computational expense in terms of seconds/sample is higher for CDIS compared to the MCMC-MH method but the overall computation time is still $\sim$ 85% lower for CDIS method than MCMC-MH method. In this paper, we implemented the block update version of the Metropolis-within-Gibbs style estimation methods (see Boys et al., 2008 [14] and Wilkinson, 2012 [119]). It is possible that the reversible jump MCMC (RJMCMC) method [14] may perform better. For cyclical kinetics, both the CDIS and the Metropolis-within-Gibbs style methods require a tuning parameter to generate samples of $r_i$. We followed the recommended tuning parameter as suggested by Boys et al. [14], Wilkinson [119]. More details are provided in the supplementary document. The marginal posteriors for both parameters are presented in Figure 7.6. We used $a = \begin{bmatrix} 1.01 & 1.01 \end{bmatrix}^T$, $b = \begin{bmatrix} 0.00067 & 0.00125 \end{bmatrix}^T$ as the parameters for gamma prior.



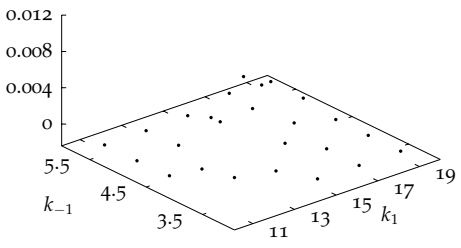

(a) Exact




(b) CDIS

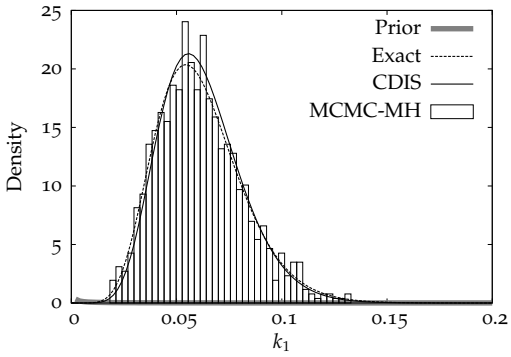



(c) MCMC-MH

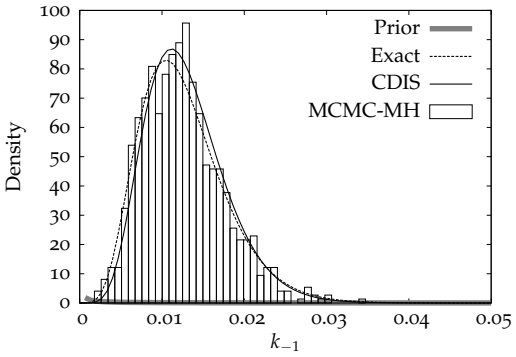



(d) MCMC-Unif

Figure 7.5: Joint posteriors for Figure 7.3d




(a) $k_1$




(b) $k_{-1}$

Figure 7.6: Marginal posteriors and priors for Figure 7.3d

Only a limited number of measurements can be performed in the laboratory due to resource limitations. The analysis presented in the previous section may be used to develop general guidelines on both experimental and model design that aid in parameter estimation. As we have shown, performing more measurements at a higher frequency does not necessarily improve estimates of some parameters. In both examples, we obtained inaccurate estimates of a rate constant $k_i$ when the corresponding reaction $\mathcal{R}_i$ was not observed to occur, *i.e.*, $r_{i,\min}$ was zero. In cases where both $r_i$ and $G_i$ are zero, we obtained no information about the rate constant at all (see data in Figure 7.1c). Note that $G_i$ is zero indicating that the reaction cannot occur during the corresponding time interval and therefore implies that $r_i = 0$, but the converse does not hold.

In order to obtain good estimates, our aim is to ensure large values of both $r_{i,\min}$ and $G_i$ by designing the experiment (*i.e.*, manipulating the times at which measurements are taken). The guidelines we provide here are specific to stochastic chemical kinetic models and are not applicable during the exploratory phase of research when no reliable information is available about the system being measured. These guidelines are useful in the later stages of experimentation when we have a set of candidate reaction models in mind.

(1) Since the parameter estimation depends on $G_i$ (Equation (5.17)), it is more important to measure the reactants than the products. If the reactants are not measured, then the numbers of reactant molecules have to be inferred using Equation (5.8) and statistical sampling, thereby increasing the computational burden. (2) Measuring a reactant species that remains at zero does not provide information about the corresponding reaction(s). As we have seen in Section 7.1, the mRNA (M) remains at zero in Figures 7.1e and 7.1f. The corresponding estimates of translation reaction rate constant $k_f$ were close to zero. The measurement effort

spent during the interval $[0, 10]$ when M was at zero, could have been better spent later during the experiment when the M was large enough. In some experiments, the measurements are left-censored, *i. e.*, we cannot measure them accurately until the measurement is above a certain threshold value. Obviously, measuring left-censored values provides little information, which is especially true for a stochastic chemical kinetic model. (3) Estimation of a rate constant is difficult when we do not observe the occurrence of the corresponding reaction event. We have shown in Section 7.2 that estimates of $k_{-1}$ were close to zero until we were able to observe $\mathcal{R}_2$ occurring, *i. e.*, until $r_{2,\min} > 0$. The same is true for the data in Figure 7.1e. For the early viral gene expression example (Section 7.1), we know from VSV-BHK biology that the translation occurs later in the experiment, which allows us to reserve more resources for later measurements. For systems like the gene on-off model (Section 7.2), the "gene switching on reaction" ($\mathcal{R}_2$) does not have such time dependence. Thus, the only way to observe the reaction events occur is by increasing the number of measurements. From the data in Figures 7.3d and 7.3f, we can see that it is better to perform the same number measurements spread sparsely over a longer period of time instead of finely spread over a shorter period of time. (4) Increased measurement frequency does provide some benefit. Likelihood plots in Figures 7.4a and 7.4c show that increased measurement frequency does improve the estimate of $k_1$. This increased measurement frequency does not change the observed number of events, $r_{1,\min}$, but it does improve the value of $G_1$, thus providing the improvement.

It is also important to choose the correct set of models for which to estimate parameters. If the model has a $\nu^T$ with full column rank, then we can obtain the number of observed reaction events uniquely (*i. e.*, $r_i = r_{i,\min}$) using the data and Equation (5.8). If we can further identify that $G_i$ is zero for a particular reaction (for example, $k_f$ in Figures 7.1e and 7.1f) then the data provides no information about the corresponding reaction, and we should simplify the model by removing

the reaction. Even if $G_i > 0$ but $r_i = 0$, then the estimate of the corresponding reaction rate constant is likely to be zero (for example, $k_r$ in Figures 7.1e and 7.1f) and this reaction may be removed as well. If the model's $\nu^T$ does not have full column rank, the observed number of reaction events $r_{i,\min}$ provides similar guidelines. Using the exact likelihood method, we observed that estimates of $k_{-1}$ were zero when $r_{2,\min} = 0$. In such cases, the "gene switching on" reaction $\mathcal{R}_2$ may be removed to simplify the model.

# 8

# CONCLUSIONS AND FUTURE DIRECTIONS

## 8.1 CONTRIBUTIONS

**Reduced kinetics using sREA reduction.** In Chapter 4, I demonstrate the use of both deterministic and stochastic versions of the reaction equilibrium assumption. It is known that the application of dREA produces a reduced reaction kinetics which is highly desirable for understanding and parameter estimation. But a corresponding reduced kinetics was not available using sREA. I show that for a linear example, the sREA-reduced equations correspond to the dREA-reduced kinetics. Further, I show via a counter example that this result does not hold for nonlinear kinetics.

**Overview of parameter estimation methods in stochastic kinetic models.** In Chapter 5, I present a range of known parameter estimation methods in a consistent, comprehensive manner with common notation. Such a monolithic text is not present in the literature. I present a common example to compare all methods and present counter examples to demonstrate the weaknesses of these methods. Valuable insight into these methods that is not mentioned elsewhere, is presented

and is then fruitfully used in Chapter 6.

**New methods for parameter estimation in stochastic kinetic models.** Chapter 6 is the main contribution of this dissertation. Two new classes of parameter estimation methods are developed. The CDIS method, based on importance sampling, provides a very accurate semi-analytical posterior distribution with desirable convergence properties, at a reduced computational cost when compared to literature methods. The AD method, though approximate, is extremely fast compared to almost any other method, and still provides the entire posterior. The AD method is shown to converge to the true (complete-data) posterior in the limit of continuously sampled data. Comparison of all methods using a common example is provided.

**Examples in systems biology.** Parameter estimation methods are applied to two examples in systems biology. A new model of viral RFP gene expression is presented. The other example is adapted from literature. Guidelines on experimental and model design that are specific to the stochastic chemical kinetic model are presented.

## 8.2 FUTURE RESEARCH DIRECTIONS

**Experimental design.** This dissertation provides some guidelines on experimental design. More research is required in order to obtain the maximum information from a limited number of experiments. Much of the literature is focused on experiments that permit factorial design. Only scarce literature is available for designing experiments that measure continuously-valued variables over continuously-valued time. Once preliminary experiments have been performed and a candidate set of models is identified, systematic methods (for example, entropy minimiza-

tion) must be used to fully utilize the limited resources.

**Experimental techniques.** In order to understand biological phenomena at low number of molecules, better measurement techniques need to be developed. Specifically, systematic characterization of measurement noise and single-molecule resolution live-cell imaging [80, see discussion] would provide an enviable wealth of data.

**Endpoint-conditioned methods.** It was repeatedly discussed in Chapters 5 and 6 that the major bottleneck in parameter estimation is the endpoint-conditioned distribution, $\pi(x \mid \theta, y)$. Unlike the research on forward simulation, surprisingly scant effort has been spent on simulating from and understanding the endpoint-conditioned distribution. $\pi(x \mid \theta, y)$ appears to be an unwieldy distribution which does not yet have an analytical expression. The few methods available for exact simulation from $\pi(x \mid \theta, y)$ suffer from severe limitations and are not suited for stochastic chemical kinetic models. An analytical expression for $\pi(x \mid \theta, y)$ or a fast method to sample from $\pi(x \mid \theta, y)$ would make the parameter estimation problem much easier.

**Approximate parameter estimation methods.** Approximate parameter estimation methods, like the AD method in Chapter 6, are extremely valuable. It is expected that more high throughput but incomplete data will be available in the future. All simulation methods, though accurate, would require enormous amount of computation to "fill-in" the missing data and therefore would be inapplicable, even with the availability of a faster endpoint-conditioned simulation methods. Much faster approximate methods that converge to the true posterior in the limit of infinite (but finitely sampled) data are poised to become the most attractive methods.

**Proofs of convergence rates, error/variance bounds.** The two new methods presented in this dissertation, being new, do not yet have the guarantees of some literature methods. For example, the convergence rate of the CDIS posterior, an upper bound on the variance of the CDIS estimator are desirable to obtain. Error bounds on the estimates of $r_i$ and $G_i$, $i = 1, 2, \ldots, n_r$, when using the AD method would help us determine the range of its applicability.

**Measurement noise.** The methods describe in this dissertation are not directly applicable for measurements with measurement noise. Modifications of these methods are required to deal with measurement error. Advances in measurement techniques have allowed single-molecule resolution [36, 75], and it is expected that continued improvements would provide measurements at low-copy numbers with negligible noise. In any case, the methods presented in this dissertation would serve as a foundation for future methods.

**Generalized parameter estimation methods.** The methods described in this dissertation derive their power by deeply exploiting the structure of stochastic chemical kinetic model. These methods are not applicable when the key assumptions change (for example, separable propensity, homogeneity of the Markov chain). While, these assumptions are much closer to the norm than being far from it, generalized methods that do not require these assumptions, will also find place in the literature.

**Better software tools for parameter estimation in deterministic models.** The tools available for parameter estimation in deterministic reaction models are basically a mixture of individual tools — the ODE solver and the optimizer. Integration of these tools would help immensely in automating the process of parameter estimation. Further, these tools are generic in nature, *i.e.*, they are not designed

specifically to estimate parameters from reaction models. Software tools are required that handle timescale separation, provide good methods for obtaining initial guesses, with integrated optimizer, ODE solver and automatic differentiation.

**Deterministic models to explain phenotypic bifurcation.** It is now accepted that the inherent randomness of a biological system at low number of molecules is the reason behind the phenotypic bifurcation. However, it is possible that a deterministic model with unmodeled (unmeasured, unknown) reactive species may result in a similar behavior. Thus, it is worth exploring if a deterministic model exists which has a partially modeled state and unknown (but deterministic) initial conditions and can explain phenotypic bifurcation. If such deterministic models are found reasonable, further experiments would be required to identify the true model of biological process.

# A

## DISTRIBUTIONS COMPOSED OF EXPONENTIALLY DISTRIBUTED RANDOM VARIABLES

In a system of stochastic chemical reactions, the time to the next reaction is distributed exponentially with the total reaction propensity as the distribution parameter. Let a random variable $V$ denote the time to next reaction and $h_0(\mathbf{x}(t), \theta)$ denote the total reaction propensity. Then,

$$V \sim \text{Exp}\left(h_0(\mathbf{x}(t), \theta)\right) \tag{A.1}$$

Note that the mean of an exponentially distributed random variable is the inverse of the distribution parameter. Also note that an exponentially distributed random variable is also gamma-distributed as given by

$$V \sim \text{Exp}\left(\lambda\right) = \text{Ga}\left(1, \lambda\right). \tag{A.2}$$

As mentioned in the Section 1.2, I use the shape-rate representation of gamma distribution .

The description of the stochastic parameter estimation method in Chapters 5 and 6, it is required to evaluate and sample from joint distributions of exponen-

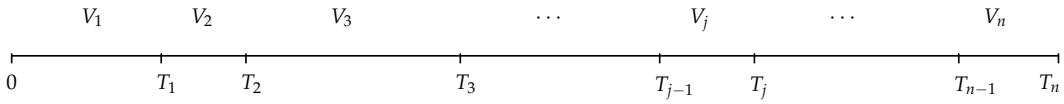



Figure A.1: Schematic of combination of exponential random variables

tially distributed random variables. In this appendix, I will describe many such cases.

## A.1 TRUNCATED GAMMA DISTRIBUTION

I define a *truncated gamma random variable* as a gamma-distributed random variable that is conditioned as follows

$$T \mid T \leq s \tag{A.3}$$

$$T \sim \text{Ga}(\cdot, \cdot), \; s > 0 \tag{A.4}$$

Below, I develop the probability density and cumulative distribution functions of the truncated gamma random variable using exponential random variables. A method to sample from the truncated gamma distributions is also presented.

Let $V_j$, $j = 1, 2, \ldots, n$, $n \geq 1$, be independently and identically distributed exponential distributions with the same distribution parameter $\lambda > 0$.

$$V_j \sim \text{Exp}(\lambda) \quad j = 1, 2, \ldots, n \tag{A.5}$$

$$V_i \perp\!\!\!\perp V_j \quad \text{if } i \neq j \tag{A.6}$$

Let $T_j$, $j = 1, 2, \ldots, n$ be the cumulative (see Figure A.1), defined as

$$T_j = V_1 + V_2 + \cdots + V_j = \sum_{i=1}^{j} V_i \tag{A.7}$$

$$T_n = V_1 + V_2 + \cdots + V_n = \sum_{j=1}^{n} V_j \tag{A.8}$$

Then, the distribution of $T_n$ is given as [119, p. 95]

$$T_n \sim \text{Ga}(n, \lambda) \quad \lambda > 0 \tag{A.9}$$

$$f_{T_n}(t_n) = \begin{cases} \frac{\lambda^n}{\Gamma(n)} t_n^{n-1} e^{-\lambda t_n} & t_n > 0 \\ 0 & t_n \leq 0 \end{cases} \tag{A.10}$$

Renaming $T_n$ as $T$ for convenience and assuming $s > 0$, truncated gamma random variable, $T \mid \{T \leq s\}$, has the following probability density function (pdf)

$$f_{T|\{T\leq s\}}(t) = \frac{P\{T \leq s \mid T = t\} f_T(t)}{P(T \leq s)} \tag{A.11}$$

If $t \leq s$ then $P\{T \leq s \mid T = t\} = 1$, otherwise $P\{T \leq s \mid T = t\} = 0$. Therefore,

$$f_{T|\{T\leq s\}}(t) = \begin{cases} \frac{f_T(t)}{P(T\leq s)} & t \leq s \\ 0 & t > s \end{cases} \tag{A.12}$$

Since $V_j > 0$, $j = 1, 2, \ldots, n$, $T > 0$, hence $f_T(t) = 0, \forall t < 0$. Substituting Eq. (A.10), we obtain

$$f_{T|\{T\leq s\}}(t) = \begin{cases} \frac{\lambda^n}{\Gamma(n)} \frac{t^{n-1} e^{-\lambda t}}{F_T(s)} & 0 \leq t \leq s \\ 0 & \text{otherwise} \end{cases} \tag{A.13}$$

in which, the cumulative distribution function (cdf) of $T$ is

$$F_T(s) = P(T \leq s) = \int_0^s \frac{\lambda^n}{\Gamma(n)} t^{n-1} e^{-\lambda t} dt \tag{A.14}$$

Equivalently,

$$f_{T|\{T\leq s\}}(t) = \begin{cases} \frac{t^{n-1}e^{-\lambda t}}{\int_0^s z^{n-1}e^{-\lambda z}} & 0 \leq t \leq s \\ 0 & \text{otherwise} \end{cases} \tag{A.15}$$

Finally, the conditional cdf of *truncated gamma distribution*, $T \mid \{T \leq s\}$ is given as

$$F_{T|\{T\leq s\}}(t) = \begin{cases} \frac{\int_0^t z^{n-1}e^{-\lambda z}}{\int_0^s z^{n-1}e^{-\lambda z}} & 0 \leq t \leq s \\ 1 & t \geq s \\ 0 & t < 0 \end{cases} \tag{A.16}$$

We can see that when $t \geq s$,

$$\begin{aligned} F_{T|\{T\leq s\}}(t) &= P\{T \leq t \mid T \leq s\} \\ &= \frac{P(\{T \leq t\} \cap \{T \leq s\})}{P(T \leq s)} = \frac{P(T \leq s)}{P(T \leq s)} \\ &= 1 \end{aligned} \tag{A.17}$$

### A.1.1 *Lower incomplete gamma function*

The lower incomplete gamma function, $\gamma(n, t)$, is defined as

$$\gamma(n, t) = \int_0^t z^{n-1}e^{-z}dz \tag{A.18}$$

Consider the numerator in Eq. (A.16),

$$I_{\text{num}} = \int_0^t z^{n-1}e^{-\lambda z}dz$$

Using a change in variable $y = \lambda z$ ($\lambda > 0$)

$$\begin{aligned} I_{\text{num}} &= \frac{1}{\lambda^n} \int_0^{\lambda t} y^{n-1} e^{-y} dy \\ &= \frac{\gamma(n, \lambda t)}{\lambda^n} \end{aligned}$$

Similarly, the denominator in Eq. (A.16),

$$I_{\text{den}} = \frac{\gamma(n, \lambda t)}{\lambda^n}$$

Substituting the values of $I_{\text{num}}$ and $I_{\text{den}}$ from the above two equations into Eq. (A.16) we obtain the following formulation of the cdf

$$F_{T|\{T \leq s\}}(t) = \begin{cases} \frac{\gamma(n,\lambda t)}{\gamma(n,\lambda s)} & 0 \leq t \leq s \\ 1 & t \geq s \\ 0 & t < 0 \end{cases} \tag{A.19}$$

Another way to represent the pdf and cdf is

$$f_{T|\{T \leq s\}}(t) = \begin{cases} 0 & t < 0 \\ \frac{f_T(t)}{F_T(s)} & 0 \leq t \leq s \\ 1 & t \geq s \end{cases} \tag{A.20}$$

$$F_{T|\{T \leq s\}}(t) = \begin{cases} 0 & t < 0 \\ \frac{F_T(t)}{F_T(s)} & 0 \leq t \leq s \\ 1 & t \geq s \end{cases} \tag{A.21}$$

### A.1.2 *Sampling from truncated gamma distribution*

Let $U \sim U[0,1]$ be a uniformally distributed random variable which may be easily generated using the `rand` function in Octave. Then, a sample of the truncated gamma random variable may be obtained by *inverting the cdf* [119, p. 101]

$$T = F^{-1}_{T|\{T \leq s\}}(U) \tag{A.22}$$

If the form in Eq. (A.21) is used, the inversion may be done as

$$t = F_T^{-1}(uF_T(s)) \tag{A.23}$$

in which $T \sim \text{Ga}(n, \lambda)$ and $u$ is a sample of $U$. The Octave code for this inversion is very simply

```
t = gaminv( rand * gamcdf(s, n, 1/lambda), n, 1/lambda )
```

But the code above fails when $uF_T(s)$ is smaller than machine precision (about $10^{-16}$) which happens frequently. In such a case, the formulation based on lower incomplete gamma functions, in Eq. (A.19) may be used. See [24] for details on lower incomplete gamma functions. Alternatively, the nonlinear equation

$$F_T(t) = uF_T(s) \tag{A.24}$$

has a unique solution for $t \in [0,1]$ and may be solved for $t$ numerically using a bijection algorithm.

## A.2 HYPOEXPONENTIAL DISTRIBUTION

In this section, I discuss a generalized version of Section A.1. Let $V_j$, $j = 1, 2, \ldots, n$, $n \geq 1$, be independently and identically distributed exponential distributions with

parameters $\lambda_j > 0$, $j = 1, 2, \ldots, n$.

$$V_j \sim \text{Exp}(\lambda_j) \quad j = 1, 2, \ldots, n \tag{A.25}$$

$$V_i \perp\!\!\!\perp V_j \quad \text{if } i \neq j \tag{A.26}$$

Let $T_j$, $j = 1, 2, \ldots, n$ be the cumulative (see Figure A.1), defined as

$$T_j = V_1 + V_2 + \cdots + V_j = \sum_{i=1}^{j} V_i \tag{A.27}$$

$$T_n = V_1 + V_2 + \cdots + V_n = \sum_{j=1}^{n} V_j \tag{A.28}$$

Then, the distribution of $T_n$ is given as (Smaili et al., 2013 [100], Bolch et al., 2001 [12])

$$T_n \sim \text{HypoExp}(\lambda_1, \lambda_2, \ldots, \lambda_n) \qquad \lambda_j > 0,\ j = 1, 2, \ldots, n \tag{A.29}$$

$$f_{T_n}(t_n) = \begin{cases} -\alpha_n e^{t_n \Lambda_n} \Lambda_n \mathbf{1}_n & t_n > 0 \\ 0 & t_n \leq 0 \end{cases} \tag{A.30}$$

in which,

$$\alpha_n = \begin{bmatrix} 1 & 0 & 0 \cdots 0 \end{bmatrix} \in \mathbb{R}^n \tag{A.31}$$

$$\Lambda_n = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & \cdots & 0 & 0 \\ 0 & 0 & -\lambda_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_{n-1} & \lambda_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & -\lambda_n \end{bmatrix} \in \mathbb{R}^{n \times n} \tag{A.32}$$

$$\mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n \tag{A.33}$$

### A.2.1 *Hypoexponential Shift identity*

In this section, I present an identity involving an integral of the Hypoexponential density function. This identity allows the analytical computation of the cdf of a Hypoexponential distribution.

**Proposition A.1** (Shift Identity)**.** Let $T_j$ and $T_{j+1}$ be random variables distributed as

$$T_j \sim \text{HypoExp}(\lambda_1, \lambda_2, \ldots, \lambda_j) \tag{A.34}$$

$$T_{j+1} \sim \text{HypoExp}(\lambda_1, \lambda_2, \ldots, \lambda_j, \lambda_{j+1}) \tag{A.35}$$

$$\lambda_j > 0 \quad \forall j$$

with the pdfs denoted by $f_{T_j}$ and $f_{T_{j+1}}$, respectively. Then,

$$\int_0^s e^{\lambda_{j+1} t} \, f_{T_j}(t) dt = \frac{1}{\lambda_{j+1}} e^{\lambda_{j+1} s} \, f_{T_{j+1}}(s) \tag{A.36}$$

*Proof.* Let $V_j$, $j = 1, 2, \ldots, n+1$, $n \geq 1$, be independently and identically distributed exponential distributions with parameters $\lambda_j > 0$, $j = 1, 2, \ldots, n+1$.

$$V_j \sim \text{Exp}(\lambda_j) \quad j = 1, 2, \ldots, n+1 \tag{A.37}$$

$$V_i \perp\!\!\!\perp V_j \quad \text{if } i \neq j \tag{A.38}$$

Let $T_j$, $j = 1, 2, \ldots, n+1$ be the cumulative, defined as

$$T_j = V_1 + V_2 + \cdots + V_j = \sum_{i=1}^{j} V_i \tag{A.39}$$

$$T_n = V_1 + V_2 + \cdots + V_n = \sum_{j=1}^{n} V_j \tag{A.40}$$

Let $A$ denote the event

$$A = \{T_n \leq u\} \cap \{T_{n+1} \geq u\} \tag{A.41}$$

Then, for any $j = n-1, n-2, \ldots, 1$,

$$\begin{aligned}
f_{T_j|\{T_{j+1},\ldots,T_n,A\}} &= \frac{f_{T_j,T_{j+1},\ldots,T_n,A}}{f_{T_{j+1},T_{j+2},\ldots,T_n,A}} \\
&= \frac{P\{A \mid T_j = t_j, \ldots, T_n = t_n\}}{P\{A \mid T_{j+1} = t_{j+1}, \ldots, T_n = t_n\}} \; \frac{f_{T_j,T_{j+1},\ldots,T_n}}{f_{T_{j+1},T_{j+2},\ldots,T_n}} \\
&= \frac{f_{T_j,T_{j+1},\ldots,T_n}}{f_{T_{j+1},T_{j+2},\ldots,T_n}} \\
&= \frac{f_{T_n|\{T_j,T_{j+1},\ldots,T_{n-1}\}}}{f_{T_n|\{T_{j+1},T_{j+2},\ldots,T_{n-1}\}}} \; \frac{f_{T_{n-1}|\{T_j,T_{j+1},\ldots,T_{n-2}\}}}{f_{T_{n-1}|\{T_{j+1},T_{j+2},\ldots,T_{n-2}\}}} \cdots \\
&\qquad \cdots \frac{f_{T_{j+1}|\{T_j\}}}{f_{T_{j+1}}} \; f_{T_j} \\
&= \frac{f_{T_{j+1}|\{T_j\}}}{f_{T_{j+1}}} \; f_{T_j} \\
&= \frac{f_{V_{j+1}}(t_{j+1} - t_j) f_{T_j}(t_j)}{f_{T_{j+1}}(t_{j+1})}
\end{aligned} \tag{A.42}$$

Since the $f_{T_j|\{T_{j+1},\ldots,T_n,A\}}$ is a density in $T_j$,

$$1 = \int_{t_j=-\infty}^{t_j=\infty} f_{T_j|\{T_{j+1},\ldots,T_n,A\}} \, dt_j \tag{A.43}$$

Substituting the above simplified expression,

$$\begin{aligned} 1 &= \int_{t_j=-\infty}^{t_j=\infty} \frac{f_{V_{j+1}}(t_{j+1}-t_j) f_{T_j}(t_j)}{f_{T_{j+1}}(t_{j+1})}\, dt_j \\ &= \int_{t_j=0}^{t_j=t_{j+1}} \frac{\lambda_{j+1} e^{\lambda_{j+1}(t_{j+1}-t_j)} f_{T_j}(t_j)}{f_{T_{j+1}}(t_{j+1})}\, dt_j \\ f_{T_{j+1}}(t_{j+1}) e^{\lambda_{j+1} t_{j+1}} &= \int_0^{t_{j+1}} \lambda_{j+1} e^{\lambda_{j+1} t_j} f_{T_j}(t_j)\, dt_j \\ \int_0^{t_{j+1}} e^{\lambda_{j+1} t_j} f_{T_j}(t_j)\, dt_j &= \frac{1}{\lambda_{j+1}} f_{T_{j+1}}(t_{j+1}) e^{\lambda_{j+1} t_{j+1}} \end{aligned} \tag{A.44}$$

Replacing $t_{j+1}$ by $s$ and $t_j$ by $t$ in the above equation,

$$\int_0^s e^{\lambda_{j+1} t} f_{T_j}(t)\, dt = \frac{1}{\lambda_{j+1}} f_{T_{j+1}}(s) e^{\lambda_{j+1} s} \tag{A.45}$$

□

**Corollary**: Note that, when $\lambda_{j+1} = 0$, the right-hand side of the shift identity may be defined as the following limit

$$\begin{aligned} \lim_{\lambda_{j+1}\to 0} \frac{1}{\lambda_{j+1}} f_{T_{j+1}}(s) e^{\lambda_{j+1} s} &= \lim_{\lambda_{j+1}\to 0} \frac{f_{T_{j+1}}(s)}{\lambda_{j+1}} \qquad \text{(if the limit exists)} \\ &= \lim_{\lambda_{j+1}\to 0} \frac{-\alpha_{j+1} e^{s\Lambda_{j+1}} \Lambda_{j+1} \mathbf{1}_{j+1}}{\lambda_{j+1}} \end{aligned} \tag{A.46}$$

Note that

$$\Lambda_{j+1} \mathbf{1}_{j+1} = -\lambda_{j+1} \beta_{j+1} \tag{A.47}$$

$$\beta_{j+1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^{j+1} \tag{A.48}$$

Substituting back into Eq. (A.46),

$$\lim_{\lambda_{j+1}\to 0} \frac{1}{\lambda_{j+1}} f_{T_{j+1}}(s) e^{\lambda_{j+1} s} = \alpha_{j+1} e^{s\Lambda_{j+1}} \beta_{j+1} \tag{A.49}$$

Thus, for $\lambda_{j+1} = 0$, the shift identity may be written as

$$\int_0^s f_{T_j}(t)\, dt = \alpha_{j+1} e^{s\Lambda_{j+1}} \beta_{j+1} \tag{A.50}$$

## A.3 CONDITIONED HYPOEXPONENTIAL DISTRIBUTION

Let $V_j$, $j = 1, 2, \ldots, n+1$, $n \geq 1$, be independently and identically distributed exponential distributions with parameters $\lambda_j > 0$, $j = 1, 2, \ldots, n+1$.

$$V_j \sim \text{Exp}(\lambda_j) \quad j = 1, 2, \ldots, n+1 \tag{A.51}$$

$$V_i \perp\!\!\!\perp V_j \quad \text{if } i \neq j \tag{A.52}$$

Let $T_j$, $j = 1, 2, \ldots, n+1$ be the cumulative, defined as

$$T_j = V_1 + V_2 + \cdots + V_j = \sum_{i=1}^{j} V_i \tag{A.53}$$

$$T_n = V_1 + V_2 + \cdots + V_n = \sum_{j=1}^{n} V_j \tag{A.54}$$

Let $A$ denote the event

$$A = \{T_n \leq s\} \cap \{T_{n+1} \geq s\} \tag{A.55}$$

Then, the distribution of $T_n \mid A$ is given by

$$f_{T_n|A}(t_n) = \begin{cases} \frac{e^{\lambda_{n+1} t_n} f_{T_n}(t_n)}{\int_0^s e^{\lambda_{n+1} t_n} f_{T_n}(t_n)} & 0 \leq t_n \leq s \\ 0 & \text{otherwise} \end{cases} \tag{A.56}$$

Using the shift identity from Section A.2.1, the distribution of $T_n \mid A$ may be simplified to

$$f_{T_n|A}(t_n) = \begin{cases} \lambda_{n+1} \frac{e^{\lambda_{n+1} t_n} f_{T_n}(t_n)}{e^{\lambda_{n+1} s} f_{T_{n+1}}(s)} & 0 \leq t_n \leq s \\ 0 & \text{otherwise} \end{cases} \tag{A.57}$$

This expression may now be used to sample $T_n \mid A$.

# BIBLIOGRAPHY

[1] www.Autodiff.org - Community portal for Automatic Differentiation. URL http://www.autodiff.org/. [cited 2013 Dec 6].

[2] PubMed Health [Internet]. Bethesda (MD). National Library of Medicine (US). URL http://www.ncbi.nlm.nih.gov/pubmedhealth/. [updated 2011 Jan 1; cited 2013 Dec 1].

[3] Alen Alexanderian, Francesco Rizzi, Muruhan Rathinam, Olivier P. Le Maître, and Omar M. Knio. Preconditioned Bayesian regression for stochastic chemical kinetics. *Journal on Scientific Computing*, pages 1–35, 2013. ISSN 0885-7474. doi: 10.1007/s10915-013-9745-5. URL http://dx.doi.org/10.1007/s10915-013-9745-5.

[4] David F Anderson. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *Journal of Chemical Physics*, 127:214107, 2007.

[5] David F. Anderson. Incorporating postleap checks in tau-leaping. *Journal of Chemical Physics*, 128:054103, 2008.

[6] Adam Arkin, John Ross, and Harley H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage $\lambda$-infected escherichia coli cells. *Genetics*, 149(4):1633–1648, 1998. URL http://www.genetics.org/content/149/4/1633.abstract.

[7] S. Asmussen and A. Hobolth. Bisection ideas in end-point conditioned Markov process simulation. In *Proceedings of the 7th International Workshop on Rare Event Simulation*, volume 69, pages 7499–7506, 2008.

[8] C. H. Bischof, H. M. Bücker, and B. Lang. Automatic differentiation for computational finance. In E. J. Kontoghiorghes, B. Rustem, and S. Siokos, editors, *Computational Methods in Decision-Making, Economics and Finance*, volume 74 of *Applied Optimization*, chapter 15, pages 297–310. Kluwer Academic Publishers, Dordrecht, 2002.

[9] P. G. Blackwell. Bayesian inference for Markov processes with diffusion and discrete components. *Biometrika*, 90(3):613–627, 2003. doi: 10.1093/biomet/90.3.613. URL http://biomet.oxfordjournals.org/content/90/3/613.abstract.

[10] William J. Blake, Mads A. Kaern, Charles R. Cantor, and James J. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

[11] Max Bodenstein. Eine Theorie der photochemischen Reaktionsgeschwindigkeiten. *Zeitschrift fu"r phsyikalische Chemie*, 85:329–397, 1913.

[12] Gunter Bolch, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi. *Introduction*, pages 1–34. John Wiley & Sons, Inc., 2001. ISBN 9780471200581. doi: 10.1002/0471200581.ch1. URL http://dx.doi.org/10.1002/0471200581.ch1.

[13] George E. P. Box and George C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison–Wesley, Reading, Massachusetts, first edition, 1973.

[14] R.J. Boys, D.J. Wilkinson, and T.B.L. Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18 (2):125–135, 2008. ISSN 0960-3174. doi: 10.1007/s11222-007-9043-x. URL http://dx.doi.org/10.1007/s11222-007-9043-x.

[15] H.M. Bücker, G. Corliss, P. Hovland, U. Naumann, and B. Norris. *Automatic Differentiation: Applications, Theory, and Implementations*, volume 50 of *Lecture Notes in Computational Science and Engineering*. Springer, 2006.

[16] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. The slow-scale stochastic simulation algorithm. *Journal of Chemical Physics*, 122(1):014116, 2005. doi: http://dx.doi.org/10.1063/1.1824902. URL http://scitation.aip.org/content/aip/journal/jcp/122/1/10.1063/1.1824902.

[17] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. Efficient step size selection for tau-leaping simulation method. *Journal of Chemical Physics*, 124: 044109, 2006.

[18] George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):pp. 167–174, 1992. ISSN 00031305. URL http://www.jstor.org/stable/2685208.

[19] D. L. Chapman and L. K. Underhill. The interaction of chlorine and hydrogen. The influence of mass. *J. Chem. Soc. Trans.*, 103:496–508, 1913.

[20] Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995. doi: 10.1080/00031305.1995.10476177. URL http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1995.10476177.

[21] Boseung Choi and Grzegorz A Rempala. Inference for discretely observed stochastic kinetic networks with applications to epidemic modeling. *Biostatistics*, 13(1):153–165, 2012.

[22] Bernie Daigle, Min Roh, Linda Petzold, and Jarad Niemi. Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. *BMC bioinformatics*, 13(1):68, 2012.

[23] C. Dittamo and D. Cangelosi. Optimized parallel implementation of gillespie's first reaction method on graphics processing units. In *Computer Modeling and Simulation, 2009. ICCMS '09. International Conference on*, pages 156–161, 2009. doi: 10.1109/ICCMS.2009.42.

[24] M. Dohler and M. Arndt. Inverse incomplete gamma function and its application. *Electronics Letters*, 42(1):35–36, 2006. ISSN 0013-5194. doi: 10.1049/el:20063446.

[25] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002. doi: 10.1126/science.1070919. URL http://www.sciencemag.org/content/297/5584/1183.abstract.

[26] Paul Fearnhead. Computational methods for complex stochastic systems: a review of some alternatives to mcmc. *Statistics and Computing*, 18(2):151–171, 2008. ISSN 0960-3174. doi: 10.1007/s11222-007-9045-8. URL http://dx.doi.org/10.1007/s11222-007-9045-8.

[27] Paul Fearnhead and Chris Sherlock. An exact Gibbs sampler for the Markov-modulated poisson process. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):767–784, 2006. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2006.00566.x. URL http://dx.doi.org/10.1111/j.1467-9868.2006.00566.x.

[28] S. J. Flint, L. W. Enquist, R. M. Krug, V. R. Racaniello, and A. M. Skalka. *Principles of Virology: Molecular Biology, Pathogenesis, and Control*. American Society for Microbiology, ASM Press, Washington, D.C., 2000.

[29] Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*,

85(410):398–409, 1990. doi: 10.1080/01621459.1990.10476213. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10476213.

[30] Stuart Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984. ISSN 0162-8828. doi: 10.1109/TPAMI.1984.4767596.

[31] Michael A Gibson and Jehoshua Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *Journal of Physical Chemistry A*, 104(9):1876–1889, 2000.

[32] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434, 1976.

[33] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340–2361, 1977. doi: 10.1021/j100540a008. URL http://pubs.acs.org/doi/abs/10.1021/j100540a008.

[34] Daniel T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A Statistical Mechanics and its Applications*, 188(1-3):404–425, 1992. ISSN 0378-4371. doi: http://dx.doi.org/10.1016/0378-4371(92)90283-V. URL http://www.sciencedirect.com/science/article/pii/037843719290283V.

[35] Daniel T. Gillespie and Linda R. Petzold. Improved leap-size selection for accelerated stochastic simulation. *Journal of Chemical Physics*, 119(16):8229–8234, October 2003.

[36] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 2005. doi: 10.1016/j.cell.2005.09.031.

[37] A. Golightly and D.J. Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics and Data Analysis*, 52(3):1674 – 1693, 2008. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2007.05.019. URL http://www.sciencedirect.com/science/article/pii/S0167947307002198.

[38] Andrew Golightly and Darren J Wilkinson. Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology*, 13(3):838–851, 2006.

[39] Andrew Golightly and Darren J. Wilkinson. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1(6):807–820, 2011. doi: 10.1098/rsfs.2011.0047. URL http://rsfs.royalsocietypublishing.org/content/1/6/807.abstract.

[40] John Goutsias. Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems. *The Journal of Chemical Physics*, 122(18):184102, 2005. doi: http://dx.doi.org/10.1063/1.1889434. URL http://scitation.aip.org/content/aip/journal/jcp/122/18/10.1063/1.1889434.

[41] Michael D. Graham and James B. Rawlings. *Modeling and Analysis Principles for Chemical and Biological Engineers*. Nob Hill Publishing, Madison, WI, 2013. 552 pages, 978-0-9759377-1-6.

[42] Ankur Gupta and James B. Rawlings. Importance-sampling based Bayesian inference for stochastic chemical kinetics. In preparation, 2013.

[43] Ankur Gupta and James B. Rawlings. Comparison of parameter estimation methods in stochastic chemical kinetic models: examples in systems biology. Submitted to AIChE, 2013.

[44] JM Hammersley and DC Handscomb. *Monte Carlo Methods*. Chapman and Hall, New York, 1964.

[45] DC Handscomb. The Monte Carlo method in quantum statistical mechanics. In *Proceedings of the Cambridge Philosophical Society*, volume 58, pages 594–598. Cambridge Univ Press, 1962.

[46] Eric L. Haseltine. *Systems Analysis of Stochastic and Population Balance Models for Chemically Reacting Systems*. PhD thesis, University of Wisconsin–Madison, 2005.

[47] Eric L. Haseltine and James B. Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *Journal of Chemical Physics*, 117(15):6959–6969, 2002. doi: http://dx.doi.org/10.1063/1.1505860. URL http://scitation.aip.org/content/aip/journal/jcp/117/15/10.1063/1.1505860.

[48] Eric L. Haseltine and James B. Rawlings. On the origins of approximations for stochastic chemical kinetics. *Journal of Chemical Physics*, 123(16):164115, 2005. doi: http://dx.doi.org/10.1063/1.2062048. URL http://scitation.aip.org/content/aip/journal/jcp/123/16/10.1063/1.2062048.

[49] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. doi: 10.1093/biomet/57.1.97. URL http://biomet.oxfordjournals.org/content/57/1/97.abstract.

[50] Daniel A. Henderson, Richard J. Boys, Kim J. Krishnan, Conor Lawless, and Darren J. Wilkinson. Bayesian emulation and calibration of a stochastic computer model of mitochondrial DNA deletions in substantia nigra neurons. *Journal of the American Statistical Association*, 104(485):76–87, 2009. doi: 10.1198/jasa.2009.0005. URL http://www.tandfonline.com/doi/abs/10.1198/jasa.2009.0005.

[51] Sebastian Hensel. Stochastic kinetic modeling of the Vesicular stomatitis virus (vsv). Master's thesis, University of Wisconsin–Madison, October 2007. URL http://jbrwww.che.wisc.edu/theses/hensel.pdf.

[52] Sebastian C. Hensel, JamesB. Rawlings, and John Yin. Stochastic kinetic modeling of Vesicular stomatitis virus intracellular growth. *Bulletin of Mathematical Biology*, 71(7):1671–1692, 2009. doi: 10.1007/s11538-009-9419-5. URL http://dx.doi.org/10.1007/s11538-009-9419-5.

[53] Alan C. Hindmarsh. Lsode and lsodi, two new initial value ordinary differential equation solvers. *SIGNUM Newsl.*, 15(4):10–11, December 1980. ISSN 0163-5778. doi: 10.1145/1218052.1218054. URL http://doi.acm.org/10.1145/1218052.1218054.

[54] Alan C. Hindmarsh, Peter N. Brown, Keith E. Grant, Steven L. Lee, Radu Serban, Dan E. Shumaker, and Carol S. Woodward. Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software*, 31(3):363–396, September 2005. ISSN 0098-3500. doi: 10.1145/1089014.1089020. URL http://doi.acm.org/10.1145/1089014.1089020.

[55] Asger Hobolth. A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. *Journal of Computational and Graphical Statistics*, 17(1):138–162, 2008. doi: 10.1198/106186008X289010. URL http://www.tandfonline.com/doi/abs/10.1198/106186008X289010.

[56] Asger Hobolth and Eric A Stone. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *The Annals of Applied Statistics*, 3(3):1204, 2009.

[57] Josef Honorkamp. *Statistical physics: an advanced approach with applications*.

Springer Berlin Heidelberg, 2012. ISBN 978-3-642-28683-4. doi: http://dx.doi.org/10.1007/978-3-642-28684-1.

[58] Tobias Jahnke and Wilhelm Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54(1):1–26, 2007. ISSN 0303-6812. doi: 10.1007/s00285-006-0034-x. URL http://dx.doi.org/10.1007/s00285-006-0034-x.

[59] Arne Jensen. Markoff chains as an aid in the study of markoff processes. *Scandinavian Actuarial Journal*, 1953(sup1):87–91, 1953. doi: 10.1080/03461238.1953.10419459. URL http://www.tandfonline.com/doi/abs/10.1080/03461238.1953.10419459.

[60] Ivan Komarov, Roshan M D'Souza, and Jose-Juan Tapia. Accelerating the gillespie $\tau$-leaping method using graphics processing units. *PloS one*, 7(6): e37370, 2012.

[61] Werner Krauth. *Statistical mechanics: algorithms and computations*, volume 13. Oxford University Press, 2006.

[62] CE Lawrence, SF Altschul, MS Boguski, JS Liu, AF Neuwald, and JC Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, 1993. doi: 10.1126/science.8211139. URL http://www.sciencemag.org/content/262/5131/208.abstract.

[63] Gabriele Lillacci and Mustafa Khammash. The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics*, 29(18):2311–2319, 2013. doi: 10.1093/bioinformatics/btt380. URL http://bioinformatics.oxfordjournals.org/content/29/18/2311.abstract.

[64] Jun S Liu. *Monte Carlo strategies in scientific computing*. springer, 2008.

[65] Jun S. Liu, Andrew F. Neuwald, and Charles E. Lawrence. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American Statistical Association*, 90(432):pp. 1156–1170, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291508.

[66] DavidJ. Lunn, Andrew Thomas, Nicky Best, and David Spiegelhalter. Winbugs - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10(4):325–337, 2000. ISSN 0960-3174. doi: 10.1023/A:1008929526011. URL http://dx.doi.org/10.1023/A%3A1008929526011.

[67] Ethan A. Mastny, Eric L. Haseltine, and James B. Rawlings. Two classes of quasi-steady-state model reductions for stochastic kinetics. *Journal of Chemical Physics*, 127(9):094106, 2007. doi: http://dx.doi.org/10.1063/1.2764480. URL http://scitation.aip.org/content/aip/journal/jcp/127/9/10.1063/1.2764480.

[68] Ethan Allen Mastny. *Multiple time scale order reduction for stochastic kinetics and molecular simulation of crystallization*. PhD thesis, University of Wisconsin, Madison, 2007. URL http://jbrwww.che.wisc.edu/theses/mastny.pdf.

[69] Harley H McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 94(3):814–819, 1997.

[70] Donald A McQuarrie. Stochastic approach to chemical kinetics. *Journal of Applied Probability*, 4(3):413–478, 1967.

[71] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

doi: http://dx.doi.org/10.1063/1.1699114. URL http://scitation.aip.org/content/aip/journal/jcp/21/6/10.1063/1.1699114.

[72] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20(4):801–836, 1978. doi: 10.1137/1020098. URL http://epubs.siam.org/doi/abs/10.1137/1020098.

[73] C. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003. doi: 10.1137/S00361445024180. URL http://epubs.siam.org/doi/abs/10.1137/S00361445024180.

[74] Brian Munsky and Mustafa Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124(4):044104, 2006. doi: http://dx.doi.org/10.1063/1.2145882. URL http://scitation.aip.org/content/aip/journal/jcp/124/4/10.1063/1.2145882.

[75] Gregor Neuert, Brian Munsky, Rui Zhen Tan, Leonid Teytelman, Mustafa Khammash, and Alexander van Oudenaarden. Systematic identification of signal-activated stochastic gene regulation. *Science*, 339(6119):584–587, 2013. doi: 10.1126/science.1231456. URL http://www.sciencemag.org/content/339/6119/584.abstract.

[76] Rasmus Nielsen. Mapping mutations on phylogenies. *Systematic Biology*, 51 (5):729–739, 2002. doi: 10.1080/10635150290102393. URL http://sysbio.oxfordjournals.org/content/51/5/729.abstract.

[77] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, second edition, 2006.

[78] Octave community. GNU/Octave, 2012. URL www.gnu.org/software/octave/.

[79] Casian Pantea, Ankur Gupta, B. Rawlings, James, and Gheorghe Craciun. The QSSA in chemical kinetics: As taught and as practiced. In N Jonoska and M Saito, editors, *Discrete and Topological Models in Molecular Biology*. Springer, 2013.

[80] Juan M. Pedraza and Johan Paulsson. Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, 319(5861):339–343, 2008. doi: 10.1126/science.1144331. URL http://www.sciencemag.org/content/319/5861/339.abstract.

[81] Suresh K Poovathingal and Rudiyanto Gunawan. Global parameter estimation methods for stochastic biochemical systems. *BMC Bioinformatics*, 11(1): 414, 2010.

[82] Christopher V. Rao and Adam P. Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *Journal of Chemical Physics*, 118(11):4999–5010, March 2003.

[83] Jonathan M. Raser and Erin K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–1814, 2004. doi: 10.1126/science.1098641. URL http://www.sciencemag.org/content/304/5678/1811.abstract.

[84] Muruhan Rathinam, Linda R. Petzold, Yang Cao, and Daniel T. Gillespie. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *Journal of Chemical Physics*, 119(24):12784–12794, December 2003.

[85] James B. Rawlings and John G. Ekerdt. *Chemical Reactor Analysis and Design Fundamentals*. Nob Hill Publishing, Madison, WI, 2004. 640 pages, ISBN 0–615–11884–4.

[86] James B. Rawlings and David Q. Mayne. *Model Predictive Control: Theory and*

*Design*. Nob Hill Publishing, Madison, WI, 2009. 576 pages, ISBN 978-0-9759377-0-9.

[87] S. Reinker, R. M. Altman, and J. Timmer. Parameter estimation in stochastic biochemical reactions. *IEE Proceedings Systems Biology*, 153(4):168–178, 2006. ISSN 1741-2471. doi: 10.1049/ip-syb:20050105.

[88] Grzegorz A. Rempala, Kenneth S. Ramos, and Ted Kalbfleisch. A stochastic model of gene transcription: An application to {L1} retrotransposition events. *Journal of Theoretical Biology*, 242(1):101 – 116, 2006. ISSN 0022-5193. doi: http://dx.doi.org/10.1016/j.jtbi.2006.02.010. URL http://www.sciencedirect.com/science/article/pii/S0022519306000610.

[89] Philip Resnik and Eric Hardisty. Gibbs sampling for the uninitiated. Technical report, DTIC Document, 2010. URL http://hdl.handle.net/1903/10058.

[90] Jean-Francois Richard and Wei Zhang. Efficient high-dimensional importance sampling. *Journal of Econometrics*, 141(2):1385 – 1411, 2007. ISSN 0304-4076. doi: http://dx.doi.org/10.1016/j.jeconom.2007.02.007. URL http://www.sciencedirect.com/science/article/pii/S0304407607000486.

[91] Matthew Richey. The evolution of Markov chain Monte Carlo methods. *The American Mathematical Monthly*, 117(5):383–413, 2010. doi: doi:10.4169/000298910X485923. URL http://www.ingentaconnect.com/content/maa/amm/2010/00000117/00000005/art00002.

[92] Christian Robert and George Casella. A short history of Markov chain Monte Carlo: subjective recollections from incomplete data. *Statistical Science*, 26(1):102–115, 2011.

[93] G.O. Roberts and A.F.M. Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and*

*their Applications*, 49(2):207 – 216, 1994. ISSN 0304-4149. doi: http://dx.doi.org/10.1016/0304-4149(94)90134-1. URL http://www.sciencedirect.com/science/article/pii/0304414994901341.

[94] J. K. Rose and M. A. Whitt. Rhabdoviridae: The viruses and their replication. In D. M. Knipe and P. M. Howley, editors, *Fields Virology*, volume 1, pages 1221–1244. Lippincot Williams & Wilkins, Philadelphia, fourth edition, 2001.

[95] Sheldon M Ross. *Stochastic processes*. Wiley Series in Probability and Statistics. New York: Wiley, 2 edition, 1996. ISBN 0471120626.

[96] Donald B. Rubin. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82(398):543–546, 1987. ISSN 01621459. URL http://www.jstor.org/stable/2289460.

[97] Howard Salis and Yannis Kaznessis. Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *Journal of Chemical Physics*, 122(5):054103, February 2005.

[98] A. Samant and D. G. Vlachos. Overcoming stiffness in stochastic simulation stemming from partial equilibrium: A multiscale Monte Carlo algorithm. *Journal of Chemical Physics*, 123:144114, 2005.

[99] S. A. Sisson, Y. Fan, and Mark M. Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007. doi: 10.1073/pnas.0607208104. URL http://www.pnas.org/content/104/6/1760.abstract.

[100] Khaled Smaili, Therrar Kadri, and Seifedine Kadry. Hypoexponential distribution with different parameters. *Applied Mathematics*, 4:624–631, 2013.

[101] A. F. M. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B*, 55(1):3–23, 1993. ISSN 00359246. URL http://www.jstor.org/stable/2346063.

[102] Peter J. Smith, Mansoor Shafi, and Hongsheng Gao. Quick simulation: A review of importance sampling techniques in communications systems. *Selected Areas in Communications, IEEE Journal on*, 15(4):597–613, 1997.

[103] Rishi Srivastava, Eric L. Haseltine, Ethan Mastny, and James B. Rawlings. The stochastic quasi-steady-state assumption: Reducing the model but not the noise. *Journal of Chemical Physics*, 134(15):154109, 2011. doi: http://dx.doi.org/10.1063/1.3580292. URL http://scitation.aip.org/content/aip/journal/jcp/134/15/10.1063/1.3580292.

[104] Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):pp. 528–540, 1987. ISSN 01621459. URL http://www.jstor.org/stable/2289457.

[105] Tianhai Tian, Songlin Xu, Junbin Gao, and Kevin Burrage. Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Bioinformatics*, 23(1):84–91, 2007. doi: 10.1093/bioinformatics/btl552. URL http://bioinformatics.oxfordjournals.org/content/23/1/84.abstract.

[106] Luke Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22(4):1701–1728, 1994. ISSN 00905364. URL http://www.jstor.org/stable/2242477.

[107] Andrea Timm and John Yin. Kinetics of virus production from single cells. *Virology*, 424(1):11 – 17, 2012. ISSN 0042-6822. doi: http://dx.doi.org/

10.1016/j.virol.2011.12.005. URL http://www.sciencedirect.com/science/article/pii/S0042682211005630.

[108] Collin Timm, Ankur Gupta, and John Yin. Kinetics of vesicular stomatitis virus RNA during infection shows mRNA production is predicted by genomes. In preparation, 2013.

[109] Surya T. Tokdar and Robert E. Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010. ISSN 1939-0068. doi: 10.1002/wics.56. URL http://dx.doi.org/10.1002/wics.56.

[110] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P.H Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31):187–202, 2009. doi: 10.1098/rsif.2008.0172. URL http://rsif.royalsocietypublishing.org/content/6/31/187.abstract.

[111] T. Turanyi, A. S. Tomlin, and M. J. Pilling. On the error of the quasi-steady-state approximation. *The Journal of Physical Chemistry*, 97(1):163–172, 1993. doi: 10.1021/j100103a028. URL http://pubs.acs.org/doi/abs/10.1021/j100103a028.

[112] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier Science Publishers, Amsterdam, The Netherlands, second edition, 1992.

[113] A. Walther and A. Griewank. Getting started with adol-c. In U. Naumann and O. Schenk, editors, *Combinatorial Scientific Computing*, pages 181–202. Chapman-Hall CRC Computational Science, 2012.

[114] Yuanfeng Wang, Scott Christley, Eric Mjolsness, and Xiaohui Xie. Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent. *BMC Systems Biology*, 4(1):99, 2010.

[115] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2004.

[116] Greg C. G. Wei and Martin A. Tanner. A Monte Carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):pp. 699–704, 1990. ISSN 01621459. URL http://www.jstor.org/stable/2290005.

[117] E Weinan, Di Liu, and Eric Vanden-Eijnden. Nested stochastic simulation algorithm for chemical kinetic systems with disparate rates. *The Journal of chemical physics*, 123:194107, 2005.

[118] Leor S. Weinberger, John C. Burnett, Jared E. Toettcher, Adam P. Arkin, and David V. Schaffer. Stochastic gene expression in a lentiviral positive-feedback loop: Hiv-1 tat fluctuations drive phenotypic diversity. *Cell*, 122(2):169 – 182, 2005. doi: http://dx.doi.org/10.1016/j.cell.2005.06.006. URL http://www.sciencedirect.com/science/article/pii/S0092867405005490.

[119] Darren James Wilkinson. *Stochastic modelling for systems biology*, volume 44. CRC press, 2012.

[120] Richard David Wilkinson. Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, 12(2):129–141, 2008.

[121] Qiang Zheng and John Ross. Comparison of deterministic and stochastic kinetics for nonlinear systems. *Journal of Chemical Physics*, 94:3644, 1991.

# INDEX

# VITA

Ankur Gupta was born in the city of Jaipur, Rajasthan, India to Drs. Avdhesh and Karuna Gupta. In April 2003, he graduated from Maharaja Sawai Man Singh Vidyalaya, Jaipur. In the August of the same year, he joined the Indian Institute of Technology, Kharagpur, India, better known as IIT Kharagpur, in the Chemical Engineering program. During his undergraduate years, he spent one winter and four summer months at the National Chemical Laboratory, Pune, India working on machine learning methods under the supervision of Dr. V.K. Jayaraman. He spent the summer of 2007 at National Centre for Biological Sciences, Bangalore, India doing experiments in molecular biology. He studied pattern formation in tubular chemical reactors for his Bachelors and Masters thesis under the supervision of Prof. Saikat Chakraborty in the Department of Chemical Engineering at IIT Kharagpur. He graduated with B.Tech (Honours) and M.Tech in Chemical Engineering and an Institute Silver Medal in July 2008.

In September 2008, he joined the Department of Chemical and Biological Engineering at the University of Wisconsin-Madison to pursue a PhD under the supervision of Prof. James B. Rawlings and Prof. John Yin. He spent the summer of 2012 at Vertex Pharamaceuticals, Cambridge, MA as a Modeling & Simulation intern. After graduation, Ankur will begin work as a Quantitative Researcher at The Climate Corporation in San Francisco, California.

Permanent Address: Jaipur, Rajasthan, India

This dissertation was prepared by the author with LaTeX $2_\varepsilon$ using his own template [1]

[1] This University of Wisconsin-Madison compliant thesis style was created by Ankur Gupta using some ideas and code from three different thesis styles, `classicthesis` and `sb-wi-thesis` and `ecsthesis`.