Efficient Finite-difference Methods for Sensitivity Analysis of Stiff Stochastic Discrete Models of Biochemical Systems

by

Monjur Morshed

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Applied Mathematics

Waterloo, Ontario, Canada, 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dr. Yang Cao Associate Professor, Dept. of Computer Science, Virgina Tech
Supervisors	Dr. Brian Ingalls Associate Professor, Dept. Applied Mathematics, University of Waterloo
	Dr. Silvana Ilie Associate Professor, Dept. of Mathematics, Ryerson University
Internal Members:	Dr. Sivabal Sivaloganathan Professor, Dept. of Applied Mathematics, University of Waterloo
	Dr. Matt Scott Associate Professor, Dept. Applied Mathematics, University of Waterloo
Internal-External Member:	Dr. Marc Aucoin Associate Professor, Dept. Chemical Engineering, University of Waterloo

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

I would like to acknowledge the names of my co-authors who collaborated in writing the papers based on my research. The materials from those papers form part of Chapters 4 and 5 of this dissertation: Dr. Silvana Ilie and Dr. Brian Ingalls.

Abstract

In the study of Systems Biology it is necessary to simulate cellular processes and chemical reactions that comprise biochemical systems. This is achieved through a range of mathematical modeling approaches. Standard methods use deterministic differential equations, but because many biological processes are inherently probabilistic, stochastic models must be used to capture the random fluctuations observed in these systems. The presence of noise in a system can be a significant factor in determining its behavior. The Chemical Master Equation is a valuable stochastic model of biochemical kinetics. Models based on this formalism rely on physically motivated parameters, but often these parameters are not well constrained by experiments. One important tool in the study of biochemical systems is sensitivity analysis, which aims to quantify the dependence of a system's dynamics on model parameters. Several approaches to sensitivity analysis of these models have been developed. We proposed novel methods for estimating sensitivities of discrete stochastic models of biochemical reaction systems. We used finite-difference approximations and adaptive tau-leaping strategies to estimate the sensitivities for stiff stochastic biochemical kinetics models, resulting in significant speed-up in comparison with previously published approaches for a similar accuracy. We also developed an approach for estimating sensitivity coefficients involving adaptive implicit tau-leaping strategies. We provide a comparison of these methodologies in order to identify which approach is most efficient depending of the features of the model. These results can facilitate efficient sensitivity analysis, which can serve as a foundation for the formulation, characterization, verification and reduction of models as well as further applications to identifiability analysis.

Acknowledgements

I would like to express my sincere gratitude to my advisors, Dr. Brian Ingalls and Dr. Silvana Ilie for the continuous support of my PhD study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better advisors and mentors for my whole study period. I have infinite gratitude towards both of my supervisors for tolerating my infinite limitations.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Sivabal Sivaloganathan and Dr. Matt Scott for their insightful comments and encouragement, but also for their valuable suggestions and inspiration.

My special thanks go to Dr. Yang Cao and Dr. Marc Aucoin for agreeing to serve as external examiners of my PhD thesis.

Furthermore, I would like to express my sincere appreciation to the Department of Applied Mathematics at the University of Waterloo for the opportunity to study here and its esteemed professors for their constant encouragement. Also, I acknowledge the help and cooperation of all office staffs from the University of Waterloo and Ryerson University: Laura Frazee, Cyntia Bratan, Steve Kanellis, Luisa Chan and Kathy Peter.

I would like to acknowledge the financial support provided by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and by an Ontario Graduate Scholarship (OGS). Computations were performed using the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET) and Compute/Calcul Canada, Ryerson RAMLAB and eight-node Linux Rocks cluster managed by the Math Faculty Computing Facility at the University of Waterloo.

I would like to thank my friends: Sammi, Ripon, Midhun, Keegan for their continued support of my journey.

Last but not least, I would like to thank my family. The foundation for my education and success started at home. I would like to thank my parents, Md Sirajul Islam and Maleka Begum for shaping me into the person I am today. I would like to thank my sisters (Kamrun Nahar, Lutfun Nahar, Najmun Nahar and Kumkum Nahar), my wife (Sanjida Eftakher) and my little princesses (Rufaida Morshed and Nusaiba Morshed) for supporting me spiritually throughout my study and my life in general.

Dedication

This is dedicated to my family, and my supervisors: Dr. Brian Ingalls and Dr. Silvana Ilie for their continuous support and for always being with me by constantly tolerating my limitations during my uneven and long journey.

Finally, I would like to dedicate all of my academic work to the father of the field of my research, who is no longer with us, Dr. Daniel T. Gillespie. May he listen to his favourite jazz records while proving himself right wherever he is now.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

At the intersection of the fields of molecular biology and chemistry is biochemistry: the study of chemical processes of living organisms. Due to the complex behaviour in the dynamics of biochemical systems, rigorous mathematical models and powerful simulation techniques are necessary to better understand them. Various mathematical approaches can be used to model chemical systems to varying degrees of accuracy. Traditional approaches use deterministic models for the time evolution of chemical systems. Although this method is appropriate in many cases, stochastic models must be used to better capture the random fluctuations observed in chemical systems [61].

Biochemical systems are generally modelled using systems of differential equations which describe the evolution of the systems through its various parameters. When appropriate, a system of ordinary differential equations may be used to describe a deterministic model. However, in the presence of random fluctuations in the system, a probabilistic approach is

better suited. The presence of noise in a system can be a significant factor in determining the system's behaviour. This cellular noise is due to the stochastic nature of chemical processes in biochemical systems [28, 78]. Thus, stochastic models and simulation techniques are important mathematical tools in the analysis of biochemical systems.

There are many instances where deterministic approaches fail and stochastic models are necessary for modelling biochemical systems. In living cells, random fluctuations become prevalent where the molecular populations and volumes can be small and readily subject to noise. For example, in some genetic switching a system may switch between two steady states, but a deterministic model can only converge to its single steady state. Therefore, stochastic models can be used to capture the noise responsible for genetic switching. From these examples and many more, it is evident that stochastic methods are essential for the study of biochemical systems.

The Chemical Master Equation (CME) [38] is a discrete stochastic model of biochemical kinetics describing the time evolution of the probability that the system will be in any given state. The state of a biochemical system can be described as the number of molecules of each biochemical species present in the system. In this way the change in the system state can be modelled probabilistically through a Markov process. Solutions to the CME can be probabilistically simulated using the stochastic simulation algorithm (SSA). The SSA, also known as Gillespie's algorithm [41], is a Monte Carlo method and is often used for simulating the dynamics of well-stirred biochemical systems [41, 43]. Although the SSA is an exact way of generating solutions to the CME, the computational resources needed for the SSA become impractical when applied to many biochemical systems encountered in practice.

The SSA requires the simulation of every reaction event that occurs in the system over some time. Therefore, when the population numbers are large and there are many reactions occurring over some time interval, the SSA may become computationally expensive. To overcome this, Gillespie proposed the tau-leaping method in order to more efficiently simulate biochemical systems [42]. In the tau-leaping method, the step-size τ is chosen small enough such that the reaction rates are almost constant. Then, if many reactions occur during these time steps, a more efficient simulation can be achieved.

In 2000, Gibson and Bruck [35] offered an improvement to Gillespie's algorithm known as the Next Reaction Method, which is exact for the CME. By utilizing an appropriate data structure for storing the propensities and avoiding unnecessary updates, the time complexity of the algorithm is improved, leading to a reduction of the computational cost.

The CME and the strategies used to simulate its solution such as the SSA and tau-leaping method are discrete stochastic models. In connection to the continuous and deterministic models of biochemical systems, the Chemical Langevin Equation (CLE) provides a natural transition between the two models. The CLE [37] is a continuous, stochastic model which yields the deterministic model as a limiting case, and is justified in its use when the population numbers are large.

Stiff biochemical systems have two well separated time scales: a slow time scale and a fast time scale, where the fastest modes are stable [93]. Since the explicit tau-leaping method is limited to the fastest time step, it is not suitable for simulations of stiff systems. The explicit tau-leaping strategy for discrete stochastic systems is similar to the explicit Euler method for ordinary differential equations. As such, the explicit tau-leaping method shows

similar instability as the Euler method when taking large time steps in a stiff biochemical system.

To overcome this instability at large time steps, the implicit tau-leaping technique was proposed by Rathinam et al. [93]. This method utilizes time steps that are larger than those of the explicit tau-leaping strategy.

The implicit tau-leaping strategy produces an accurate approximate solution for the slow manifold and for the mean of the fast variable on the slow manifold. This solution is comparable to the solution produced by the explicit tau leaping method [93]. Another property of the implicit tau-leaping strategy is that it may dampen the noise for some systems, as they reach a steady state.

Characteristics exhibited by systems can be described by mathematical methods. System behaviour depends on parametrization. These behaviours or states are outputs affected by many input parameters. This in turn brings the concept of parametric sensitivity which is an important tool in the study of biochemical systems. Using *parametric sensitivity*, we can study the effect of variations in input parameters on system behaviour [116] and measure the parametric sensitivity coefficient of systems. Moreover, when some reactant amounts in a system are small (as is the case for typical biochemical systems) and noise is present in the system, stochastic models must be used for the sensitivity analysis.

If large changes in a system's outcomes occur when there is a small change in a certain input value of a parameter, then the system is highly sensitive to the value of the parameter. In chemical reaction models, relevant input parameters are subject to uncertainties. Reaction kinetics, thermodynamic equilibria and transport properties are measured exper-

imentally or estimated theoretically [116]. Furthermore, initial and operating conditions such as initial amounts for each species can change in response to uncertain environmental reactions. It is therefore important to identify and understand the sensitivity to different parameters affecting a system's behaviour. Useful insight can then be gained about the model regarding dynamics and even help inform the model's own development and accuracy.

A number of approaches to sensitivity analysis of stochastic discrete models of biochemical kinetics have been developed [2, 46, 96]. Most of the techniques developed for approximating parametric sensitivities for these models involve a finite-difference estimator, such as $[E(f(X^{c+h}(t))) - E(f(X^c(t)))]/h$, where h represents a perturbation, c is the parameter of interest, X is the state of the biochemical reaction system, f is the output function of interest and E the expectation value. If the mean abundance of species is the quantity of interest, then $f(X) = X$. Once the mean of the abundance is estimated accurately, the variance of the abundance of the species can then be evaluated by the choice of $f(X) = X^2$. This finite-difference estimator (which is discussed further in Section 3.2.2) approximates the local sensitivity of the expected value of the quantity $f(X^c(t))$ with respect to a parameter c, given a polynomial function f. (Note that higher-order moments can be determined by appropriate combinations of expected sensitivities). Explicitly, the finite difference estimator takes the difference between a nominal system (with parameter value c) and the perturbed system (with parameter value $c + h$) at each time, providing a temporal profile of system sensitivity by the end of the simulation.

The Common Random Number (CRN) method (further described in Section 3.5.1) introduced by Rathinam et al. [96] employs Gillespie's SSA with a shared stream of random

numbers to generate the nominal and perturbed trajectories. A result of sharing a common random number by both nominal and perturbed trajectories is the reduction of variance for the estimator. Consequently, better accuracy of the estimator is achieved for the same computational cost.

The Common Reaction Path (CRP) [96] strategy (which is described in Section 3.5.2) applies the Random Time Change (RTC) algorithm (described in Section 2.7.2) with common random number stream to simulate the sample paths. Like the CRN, the consequence of sharing a common random number stream reduces the variance for the estimator.

The Coupled Finite Difference (CFD) method, proposed by Anderson in [2], simulates the coupled trajectories with a version of the next reaction method [35]. This sensitivity estimator is based on the tight coupling between the nominal process, $X^c(t)$, and the perturbed process, $X^{c+h}(t)$. Numerical examples demonstrate that the CFD produces the smallest variance among these finite-difference estimators (see [2] for more details) which is shown in Section 3.6.

Another sensitivity analysis method for discrete stochastic processes was developed by Gunawan et al. [46]. Their method is based on the density function sensitivity. In their work, the authors used an analogue of classical sensitivity and the Fisher Information Matrix. They compared the deterministic and discrete stochastic analysis when applied to two different models. The importance of applying an appropriate sensitivity analysis was demonstrated in their work in relation to the dynamics of the given models.

In Section 3.6, we provide a comparison for the accuracy of the CRN, CRP and CFD methods on a range of model types. These techniques are based on exact stochastic simulation

algorithms to generate the trajectories.

We developed a novel finite-difference estimator that utilizes an approximate stochastic simulation strategy to generate coupled paths, which we called the Coupled Tau-Leaping scheme (CTL) [83], further discussed in Chapter 4 . To estimate the local sensitivities for stiff biochemical systems, our novel method [83] is computationally efficient for moderately stiff systems. Our strategy couples the nominal and perturbed processes in a manner similar to the CFD method. The CFD scheme couples paths which are in exact agreement with the Chemical Master Equation (CME), whereas our approach couples paths that are obtained using the (approximate) explicit tau-leaping method. Our CTL algorithm makes use of the widely used step-size selection strategy developed by Cao et al. [14] for the explicit tau-leaping method. Our method applies this efficient tau-selection procedure to both the nominal and the perturbed trajectories.

We developed another novel algorithm which we called the Coupled Implicit Tau-leaping (CIT) [84] for estimating local sensitivities that is computationally efficient when applied to moderately stiff to stiff stochastic biochemical systems. This novel strategy (CIT) is described further in Chapter 5. In the CIT sensitivity method, the coupling of the nominal and perturbed processes is similar to that employed by the CFD method [2]. However, our approach couples paths that are obtained with the (approximate) implicit tau-leaping strategy, whereas the CFD method couples paths that are in exact agreement with the Chemical Master Equation (CME). For an efficient implementation of the implicit tau-leaping scheme on the nominal and perturbed trajectories, the CIT method makes use of the state-of-the-art step-size selection strategy introduced by Cao et al. [15].

An effective model of a physical system can be used to predict how it will behave. The mathematical model of a physical system should be constructed such that a unique set of parameters can be found to parameterize the model in a way that is consistent with observable data [4]. It is important that the model can be used to simulate results comparable to the actual observations. Identifiability analysis can be utilized for assessing the confidence of the estimated parameter values. A model is identifiable if we can theoretically discover the true values of the model's parameters by taking an infinite number of observations from it [99]. Generally, identifiability (analysis) consists of two types of analysis [4]. First, structural identifiability analysis will be employed to investigate the theoretical possibility of finding a unique (globally or locally) set of parameter values that are most similar to the observations. Second, practical identifiability investigates the practical possibility of finding a unique (globally or locally) set of parameter values that are most similar to the observations.

The practical idenitifiability analysis is very important in real life situations due to limited amount and quality of experimental data [99]. In some systems which are very complex (highly non-linear), obtaining the true values of the parameters may not be possible. However, in these situations we may obtain quality information about the system, even if the parameter values are not true. Finally, we present an identifiability approach for stochastic models to approximate the Fisher Information Matrix (FIM) by constructing the sensitivity matrix and using it as an identifiability tool to assess the quality of the estimated parameter values and finding the confidence intervals for true values of the model parameters. This is described further in Section 6.5 of Chapter 6.

This thesis is organized as follows. Chapters 2 presents the background on stochastic

modelling and simulation of well-stirred biochemical kinetic systems. In Chapter 3, we describe the established finite-difference approaches to estimating parametric sensitivities for the Chemical Master Equation (CME), with numerical experiments to illustrate their performance, by application to some models of simple biochemical networks. In Chapters 4 and 5 we propose new finite-difference strategies for estimating the sensitivity coefficients, based on the adaptive explicit and implicit tau-leaping methods, respectively. In Chapters 4 and 5, we also present the advantages of our sensitivity estimation methods compared to previously published finite-difference based sensitivity analysis techniques on systems which are mildly stiff to stiff. In Chapter 6, we describe identifiability analysis and illustrate it with applications to some simple model systems. Lastly, in chapter 7 we summarize our results and discuss several future research projects.

Chapter 2

Stochastic Models of Biochemical Kinetics

2.1 Introduction

Many important biological processes have been successfully studied using the techniques of stochastic modelling and simulations. Stochastic models are useful for accurately describing the biochemical system dynamics, in particular for systems with low molecular amount of some species. A widly used stochastic model of well-stirred biochemical systems is the Chemical Master Equation [38]. The CME is a system of ordinary differential equations. One ODE represents the evolution of each possible state of the system. In this model, the continuous time evolution of a system state changes probabilistically through a Markov process. It is a discrete stochastic model of biochemical kinetics. The molecular numbers

of each chemical species present in the system describes the state of the system.

The dimension of the CME depends on the total number of possible states of the system, which inturn depends on its molecular count. Therefore, the CME is often of very high dimension and is analytically solvable for just a few simple systems. The stochastic simulation algorithm (SSA), known as Gillespie's algorithm is a Monte Carlo simulation technique which generates trajactories in exact agrement with the CME. The exact Random Time Change (RTC) algorithm [1, 25] can also be used to simulate the sample paths of the CME model. If the molecular amounts of the species in the systems are large, the silmulations becomes computationally expensive. Gillespie introduced an approximate method called tau-leaping [42], in which time steps are selected dynamically to skip over many reactions when accuracy allows.

The numerical strategies presented above, such as the SSA and the tau-leaping method, apply to the discrete stochastic model of well-stirred biochemical systems, the CME. The Chemical Langevin Equation (CLE) [37] provides a reasonable transition between the discrete stochastic and the deterministic continuous models of biochemical systems. The stochastic continuous CLE model is valid to use for larger molecular populations in each species.

2.2 Deterministic vs Stochastic Approaches

Biological processes may be modelled deterministically or stochastically, depending on the dynamics of the system. A deterministic model describes the evolution of a system in a

predictable manner, whereas a probabilistic description is provided in a stochastic model, which take into account the inherent randomness of the biochemical system. That is, given some input data and parameters for a system, a deterministic model provides a unique evaluation in time of the system. By contrast, a stochastic model yields a variable system output.

One important class of problems where a deterministic model is suitable consists of biochemically reacting systems with very large molecular numbers of each species. For these systems, the average behaviour of the system is considered. Thus, the evolution can be described in a predictable manner. However, when only small numbers of some the molecular species exist in the system, the fluctuations may be significant and a stochastic model is needed to account for the randomness. Consequently, a deterministic behaviour can be understood as a limiting case of a stochastic behaviour, when the number of molecules of each species is large. In Figure 2.1, we illustrate the difference between the stochastic simulation versus the deterministic simulation of the evolution in time to a steady state for the Michaelis-Menton model, described in Section 2.11.1. The SSA algorithm described in Section 2.7.1 has been used for this stochastic simulation.

2.3 Noise and Robustness

The random variability in quantities that arise in cellular biology is referred to as noise [62]. This noise is defined in two ways. Intrinsic noise refers to the inherent stochasticity of biochemical processes within a single cell, such as binding, transcription, and translation [28, 78]. Extrinsic noise refers to the variations in the states of components between

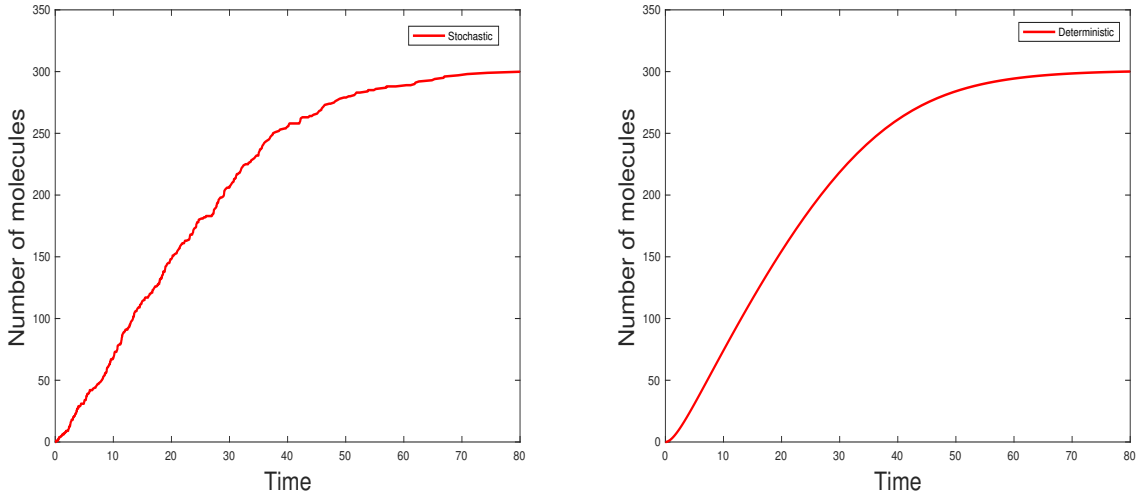


Figure 2.1: Stochastic plot vs. deterministic plot. Evolution in time for steady state.

different cells, such as the variations between cells in the expression of a specific gene. Environmental parameters like temperature, pH, and other kinetic parameters [90] are also responsible for extrinsic noise. Cellular noise may play an important role when some species have a small number of molecules. The probabilistic behaviour of a system is due to the presence of intrinsic and/or extrinsic noise [20, 86, 87, 120]. Both types of noise can lead to fluctuations at the single cell level and thus result in cell-to-cell variability. Identifying the sources of noise is often difficult to do in practice.

The robustness of a system is its ability to maintain its functions in the presence of noise [68]. Robustness is necessary for a system to function with unreliable components in the presence of noise [71]. In this regard, it is believed that evolution can be responsible for selecting and preserving robust traits in a system. This is an example of a system-level phenomena that is not easily explained in terms of only a system's individual components.

This approach of attempting to understand biological systems as a whole is the focus of System Biology [54].

2.4 Markov Processes

A stochastic process represents a random variable which evolves in time, either discretely or continuously. A biochemical system can be modelled stochastically when its system state can be regarded as a random variable. In the special case when the time evolution of the state at any particular time only depends on the current state of the system, a Markov process is used to model the systems behaviour [25, 39]. In a Markov process, provided that the present state of the system is known, the future system state of the system can be determined independent of any past states.

More precisely, let S be the state space of a system and let $X(t) \in S$ represent the state of the system for discrete time steps $t = 0, 1, 2, \ldots$. For some subset $A \subseteq S$, let $P(X(t+1) \in A | X(t) = x)$ denote the probability that $X(t+1) \in A$ provided that $X(t) = x$. Then a Markov process can be defined as one that satisfies

$$P(X(t+1) \in A | X(t) = x, X(t-1) = x_{t-1}, ...X(0) = x_0) = P(X(t+1) \in A | X(t) = x),$$

for all $x, x_{t-1}, ..., x_0 \in S$. Thus, knowing any of the past states, $X(t-1), X(t-2), \ldots$, does not yield any additional information than only knowing the current state $X(t)$, in determining the future state $X(t+1)$.

In the continuous setting, with $t \in [0, \infty)$, a Markov process is defined as one that obeys

$$P(X(t+dt) = y | X(\tau) = x(\tau), \forall \tau \in [0,t]) = P(X(t+dt) = y | X(t) = x(t)),$$

for states $x(\tau) \in S$ where $\tau \in [0,t]$. Continuous time Markov processes are also called diffusion processes [48].

In the time-independent case where $P(X(t+1) \in A | X(t) = x)$ does not depend on time, the probability will simply be written as

$$P(X(t+1) \in A | X(t) = x) = p(x, A), \forall t.$$

2.5 Stochastic Chemical Kinetics

In what follows, a presentation of the concepts used in mathematically modelling of homogeneous biochemical systems will be given, without providing the complete rigorous mathematical details. Thus, consider a chemical process in which N different chemical species, or types of molecules, can interact in M different kinds of chemical reactions. A simple example of one such kind of reaction could be the process in which a molecule of species A and a molecule of species B react to create a new molecule of some other chemical species C. In this way, the state of a system can be expressed by specifying the number of molecules there are of each species at a particular moment in time. This approach benefits from being computationally cheaper in trying to calculate the systems dynamics when compared to the more general molecular dynamics approach. In the molecular dynamics

approach, the individual positions and velocities are tracked in order to account for the various collisions and interactions that may results between molecules. However, the spatial information necessary to specify particle trajectories makes computations involving them computationally expensive. Being concerned with the amounts of the chemical species, without regard to this spatial information, is justified as a simplification when considering *well-stirred* biochemical systems. These are systems with the property that molecules of any type are uniformly spread throughout the reaction volume. These considerations are further justified by also assuming that the system is in thermal equilibrium and the volume of the domain is constant.

Let t be some moment in time, and let $X_i(t) \geq 0$ be an integer denoting the number of molecules of species type i present in the system at some time t. Then, the state of the entire system will be represented as the state vector $X(t) = [X_1(t), X_2(t), ..., X_N(t)]^T$.

The evolution of the state vector $X(t)$ depends on the reactions that take place from one moment to another. Since this evolution depends on the probability that a certain reaction will take place, the resulting state of the system is also probabilistic and usually modelled by a random variable. The Chemical Master Equation, a set of ordinary differential equations (ODEs), describes the system's evolution in this way. In the CME, the k^{th} ODE gives the probability that the system will be in the k^{th} possible system state at time t. The dimension of this system of ODEs is given by the number of all possible states that the system under consideration can be in, given the initial state. Thus the dimension of this ODE system is generally very large in practice and infinite for open systems. The dimension in turn depends on the total number of molecules present in the system and on the nature of the relevant chemical reactions.

Since the CME is a system of ODEs of very large dimension for many systems considered in applications, it is difficult to analyze this model either analytically or computationally. One way of computing solutions to the CME indirectly is by using the stochastic simulation algorithm (SSA), also known as Gillespie's algorithm [41] which is discussed further in Section 2.7.1. Instead of computing complete probability distributions over the state space as the CME is designed to do, the SSA avoids this and merely computes some state trajectory that is sampled from these distributions in such a way that the realized state is computed with probability in accordance with the CME's distributions. Despite being easily implemented, the SSA is inefficient when reactions occur frequently. However, by choosing some time steps τ which leap over many reactions and then updating the state according to what reactions took place over this step, the SSA can be speed-up with only a minor loss in accuracy. This approximation scheme is referred to as the tau-leaping strategy.

A simplified model is the Chemical Langevin Equation (CLE) a system of N stochastic differential equations, where N is the number of different chemical species. The CLE differs from the CME model in several respects. As just stated, the dimension of the CLE is N as opposed to the number of all possible states in the case of the CME. Moreover, instead of the molecule numbers $X_i(t)$ only taking integer values, they now take on real number values in the CLE. Thus, the CLE describes the evolution of a Markov process continuous in space as opposed to a Markov process discrete in space used in the CME. The CLE is easier to analyze and cheaper to solve numerically than the CME. The CLE model is valid in the regime of large molecular amounts.

Even further approximations of the CME can be made by considering only the deterministic

parts of the CLE—obtaining an N-dimensional system of ODEs called the reaction rate equations (RRE). In doing so, this approach ignores certain fluctuations that may be present in the CLE. When compared to the CME and CLE, the RRE are readily solved using numerical integration techniques. The RRE model is valid when the thermodynamic limit (which is discussed further in Section 2.8) applies.

2.6 Chemical Master Equation (CME)

A derivation of the Chemical Master Equation will be presented here. Consider a system with N chemical species labelled as $S_1, S_2, \ldots, S_N$ that can take part in M different of chemical reactions. In what follows, only two kinds of reactions will be considered: unimolecular reactions which involve a single reactant molecule, and bimolecular reactions that involve two reactant molecules.

Let $X(t_1) \in \mathbb{R}^N$ be the state vector of a system at some time t_1 with N chemical species and M permissible reactions. Suppose only a single reaction takes place e.g. R_j in the time interval $[t_1, t_2]$. Then the resulting state $X(t_2)$ can be described by introducing the state-change, or stoichiometric, vector $\nu_j \in \mathbb{R}^N$ that accounts for the change in state. There is a corresponding vector ν_j for each reaction R_j for $j = 1, 2, ..., M$. Note that the matrix formed by taking its columns to be the state change vectors ν_j is called the stoichiometric matrix. In this way, if the j^{th} reaction occurred during the time interval $[t_1, t_2]$, then the resulting state is given by $X(t_2) = X(t_1) + \nu_j$.

Now returning to the general setting, to describe the dynamics of the system provided the

initial state $X(0) = x_0$ is known, it is necessary to compute the probability $P(x,t)$ of the system to be in the particular state $X(t) = x$ at some time t. This probability will depend on the likelihood of a certain reaction occurring. For each reaction R_j, let $a_j(X(t))$ be its propensity function, defined as $a_j(X(t))dt$ is the probability of the j^{th} reaction taking place during the time interval $[t, t+dt)$. Depending on the type of reaction, the propensity for the j^{th} reaction is given by

First Order: $X_m \rightarrow$ products, $\quad a_j(X(t)) = c_j X_m(t)$

Second Order: $X_m + X_n \rightarrow$ products, $\quad a_j(X(t)) = c_j X_m(t) X_n(t)$

Dimerization: $X_m + X_m \rightarrow$ products, $\quad a_j(X(t)) = c_j \frac{1}{2} X_m(t)(X_m(t) - 1)$.

In each case, it is seen here that the propensity is proportional to some combinatorial factor involving chemical species numbers $X_i(t)$ since the likelihood of a certain reaction taking place depends on the number of available reactants at that time. The proportionality constant c_j characterizes the particular reaction R_j by functioning as a scaling factor.

Let us define $P(y,t)$ to be the probability that $X(t) = y$ if $X(t_0) = x_0$. Suppose the probability $P(y,t)$ is known for any state y at time t. Consider a time interval $[t, t+dt)$ where dt is taken to be sufficiently small so that at most only a single reaction occurs during the time interval $[t, t+dt)$. Now, suppose that at time $t+dt$ the system happens to be in the state x, that is $X(t+dt) = x$. This could occur in the following ways. Trivially, one way is if the system at time t was already in the state $X(t) = x$ and no reaction occurred in the interval $[t, t+dt)$ so that the system remains in the state x at $t+dt$. If the j^{th} reaction took place during the time interval $[t, t+dt)$, then the only way the state $X(t+dt) = x$ could result at time $t+dt$ is if the system was originally in the state $X(t) = x - \nu_j$.

Let A be the event where the state of the system at time $t + dt$ is x. To formally derive an equation for the probability $P(A)$ of the event A occurring also consider the events $B_0, B_1, \ldots, B_{M+1}$. Namely, let B_0 be the event that the state of the system is x at time t. For $1 \leq j \leq M$, let B_j correspond to the event that the system is in the state $x - \nu_j$ at time t. Lastly, let B_{M+1} be the event where the system is in any other state at time t. These events are disjoint so only one event can happen, as well as exhaustive, so at least one of them must occur. Thus, the probability of event A occurring is given by the law of total probability:

$$P(A) = \sum_{j=0}^{M+1} P(A|B_j)P(B_j), \tag{2.1}$$

where $P(A|B_j)$ is the conditional probability of event A occurring given that B_j happens. Observe that, by definition of the propensity functions,

$$P(A|B_j) = a_j(x - \nu_j)dt, \qquad 1 \leq j \leq M. \tag{2.2}$$

Now, since either a reaction will happen or not happen, it must be the case that the sum of these probabilities is 1. This implies that

$$P(A|B_0) = 1 - \sum_{j=1}^{M} a_j(x)dt. \tag{2.3}$$

Recall that B_{M+1} is the event where the system is in some state that cannot lead to the desired state x in a single reaction. Therefore, the conditional probability of the event A

happening at time $t + dt$ provided that B_{M+1} was the case at time t is just

$$P(A|B_{M+1}) = 0. \tag{2.4}$$

Using equations (2.2), (2.3) and (2.4), together with the definition of $P(x,t)$, in (2.1) allows the probability $P(A) = P(x, t + dt)$ to be expressed as

$$P(x, t + dt) = \left(1 - \sum_{j=1}^{M} a_j(x)dt\right)P(x,t) + \sum_{j=1}^{M} a_j(x - \nu_j)dtP(x - \nu_j, t).$$

Then upon rearranging this expression, it follows that

$$\frac{P(x, t + dt) - P(x,t)}{dt} = \sum_{j=1}^{M}[a_j(x - \nu_j)P(x - \nu_j, t) - a_j(x)P(x,t)].$$

In the limiting case where $dt \to 0$ the left-hand side of this equation is precisely the time derivative of $P(x,t)$ so that

$$\frac{dP(x,t)}{dt} = \sum_{j=1}^{M}[a_j(x - \nu_j)P(x - \nu_j, t) - a_j(x)P(x,t)]. \tag{2.5}$$

For each state x that could have been considered for the system, there is an equation of the form (2.5). The complete set of such equations gives a system of linear ODEs known as the Chemical Master Equation. As previously mentioned, the dimension of this system of equations is often very large in practice and typically infinite for open systems.

The function $P(x,t)$ can be determined using the CME. Numerical solutions to the CME are generally difficult to compute, and the CME can be solved analytically for only a few

simple closed systems.

2.7 Simulation Algorithms

2.7.1 Stochastic Simulation Algorithm (SSA)

The stochastic simulation algorithm (SSA) [41] offers a means to overcome the computational barrier present in the CME due to its large dimension. The SSA accomplishes this by only computing single realizations of the state in accordance to the underlying probability distribution, as opposed to computing the entire distribution.

For purposes of deriving the SSA, define the quantity $P_0(\tau|x,t)$ to be the probability that no reaction takes place in the time interval $[t, t+\tau)$ provided that the state at time t is $X(t) = x$. Now consider partitioning the infinitesimally extended time interval $[t, t+\tau+d\tau)$ into the two intervals $[t, t+\tau)$ and $[t+\tau, t+\tau+d\tau)$. Moreover, assume that what happens over the first interval, $[t, t+\tau)$, is independent of what happens over the later interval, $[t+\tau, t+\tau+d\tau)$. In this way, the probability $P_0(\tau+d\tau|x,t)$ that no reaction takes place over the extended interval is determined by the probabilities $P_0(\tau|x,t)$ and $P_0(d\tau|x,t+\tau)$ by taking their product since the two events are independent.

Now, since the probability of no reaction happening over a particular interval is complement to the probability of any reaction happening during that interval,

$$P_0(d\tau|x,t+\tau) = 1 - \sum_{k=1}^{M} a_k(x)d\tau.$$

Therefore,

$$P_0(\tau + d\tau|x,t) = P_0(\tau|x,t)\left(1 - \sum_{k=1}^{M} a_k(x)d\tau\right),$$

which can be alternatively expressed as

$$\frac{P_0(\tau + d\tau|x,t) - P_0(\tau|x,t)}{d\tau} = -a_{sum}(x)P_0(\tau|x,t)$$

where $a_{sum}(x) := \sum_{k=1}^{M} a_k(x)$ has been introduced for notational convenience.

Then by taking the limit $d\tau \to 0$ in the preceding expression a linear ODE is derived, which with the initial condition of $P_0(0|x,t) = 1$, has a particular solution of the form

$$P_0(\tau|x,t) = e^{-a_{sum}(x)\tau}. \tag{2.6}$$

With this in mind, consider the probability $P(\tau, j|x,t)$ of the conditional event where no reaction occurs in the first interval $[t, t+\tau)$ and only the j^{th} reaction occurs in the later interval $[t+\tau, t+\tau+d\tau)$:

$$P(\tau, j|x,t)d\tau = P_0(\tau|x,t)a_j(x)d\tau. \tag{2.7}$$

We assumed that $d\tau$ is small enough that at most one reaction may happen during $d\tau$. Upon substituting (2.6) into (2.7) this probability is then given by $P(\tau, j|x,t) = a_j(x)e^{-a_{sum}(x)\tau}$, which can equivalently be expressed as

$$P(\tau, j|x,t) = (\frac{a_j(x)}{a_{sum}(x)})(a_{sum}(x)e^{-a_{sum}(x)\tau}). \tag{2.8}$$

The probability $P(\tau, j|x, t)$ can be interpreted as the joint density function of two random variables, one corresponding to the time to the next reaction, τ, and the other for the index of that reaction.

Here j is called the *next reaction index*. The probability of choosing the j^{th} reaction is proportional to its propensity $a_j(x)$. The other random variable represents the *time index* representing the time until the next reaction; it is an exponential random variable with mean $\frac{1}{a_{sum}(x)}$. It has the density function $a_{sum}(x)e^{-a_{sum}(x)\tau}$.

The SSA is constructed so that it simulates the time and reaction index by sampling a uniform distribution over $(0, 1)$. Assuming state $X(0) = x_0$, the steps of the SSA are outlined below:

1. At time t, compute the propensities $\{a_k(X(t))\}_{k=1}^{M}$ and their sum $a_{sum}(X(t)) := \sum_{k=1}^{M} a_k(X(t))$.
2. Simulate two independent uniform $(0, 1)$ random numbers, ξ_1 and ξ_2.
3. Select j to be the smallest integer which obeys the condition $\sum_{l=1}^{j} a_l(X(t)) > \xi_1 a_{sum}(X(t))$.
4. Take $\tau = \ln(1/\xi_2)/a_{sum}(X(t))$.
5. Set $X(t + \tau) = X(t) + \nu_j$ and update from t to $t + \tau$.
6. Return to step 1 or else stop the simulation.

The simulation will terminate either when t reaches the final time or when some chemical species is larger than a specified upper or lower bound.

For comparing the SSA and the direct solution of the CME, recall that the CME is generally intractable computationally in practice. Similarly, even though the SSA is readily implemented it too lacks computational efficiency. To understand where the computational difficulty may arise in the SSA, observe that the time step τ used in the algorithm depends inversely on the quantity $a_{sum}(x)$ which increases as the population numbers grow. Generally speaking however, perhaps this is to be expected since any algorithmic procedure that requires simulating the events individually will usually be inefficient.

2.7.2 Random Time Change (RTC)

The Random Time Change (RTC) algorithm [1, 25, 96] is another stochastic simulation method which has exact agreement with the CME. Hence, it can be used as an alternative to simulate the sample paths of the CME models that were discussed in Section 2.7.1. It provides each reaction in the system with its own internal time. The internal time $\Gamma_j(t)$ of the reaction channel R_j is defined by

$$\Gamma_j(t) = \int_0^t a_j(X(s))ds. \tag{2.9}$$

The number of firings of the j-th reaction in the interval $[0, t]$ may be represented by $Y_j(\Gamma_j(t))$, where Y_j are independent unit rate Poisson processes, for $j = 1, \ldots, M$. Consequently, the system state X at time t is given by

$$X(t) = X(0) + \sum_{j=1}^{M} \nu_j Y_j \left(\Gamma_j(t)\right). \tag{2.10}$$

The RTC method can be implemented as follows [96]. Let the internal times at which reaction R_j occurs, i.e. the jump times of the Poisson process Y_j, be denoted by I_l^j (such that $I_1^j < I_2^j < I_3^j < \ldots$). For each reaction channel, define $I_+^j(t)$ as the time at which the reaction will occur next:

$$I_+^j(t) = \min\{I_l^j | \Gamma_j(t) < I_l^j, l = 1, 2, \ldots\}.$$

Considering all reactions in the network, let T_i indicate the physical time at which the ith reaction event occurs; let J_i be the index of the corresponding reaction channel. Note from (2.9) that $\Gamma_j(t)$ is piecewise linear, with linear growth at rate $a_j(X)$ between firing events. Thus, at time T_i, the time until the next firing event is given by

$$\Delta T = T_{i+1} - T_i = \min\left\{\frac{I_+^j(T_i) - \Gamma_j(T_i)}{a_j(X(T_i))}, j = 1, \ldots, M\right\}.$$

The corresponding reaction index J_{i+1} is the index j for which the minimum is achieved.

The random time change algorithm can thus be implemented as follows

1. Initialize $I_+^j = E^j$, unit exponential random numbers for $j = 1, \ldots, M$, and set $T_0 = 0$.

2. At each reaction time T_i, compute the propensity function, $a_j(X(T_i))$, for each reaction.

3. Evaluate ΔT. Set j^* to be the index of the minimum.

4. Update the state $X(T_{i+1}) = X(T_i) + \nu_{j^*}$ and the time $T_{i+1} = T_i + \Delta T$, and increment

each internal time Γ_j by $a_j(X(T_i))(\Delta T)$.

5. Increment $I_+^{j^*}$ by a unit exponential random number E^{j^*}.

6. Increment the index i. Return to step 2 or else stop the simulation.

The benefits of the RTC algorithm are three-fold. First, many algorithms and statistical techniques can be formulated by its explicit representation. Second, the RTC can be used in the context of a multi-level Monte Carlo simulation for biochemical kinetic systems to produce an unbiased estimator to couple different versions of processes. This can reduce the computational cost. Third, the RTC can be used to approximate parameter sensitivities by finite-difference methods, which is discussed in Section 3.5.3.

2.7.3 Improvements: Tau-Leaping Method

Since the SSA algorithm is often computationally expensive in solving the CME, a more efficient method is desirable. The tau-leaping method, which was introduced by Gillespie in 2001 [42], offers a more practical algorithm that produces results without compromising too much of the accuracy. The SSA has to deal with simulating every reaction of the system, and is more costly if the system under consideration involves many molecules and first reactions. If the propensity $a_j(x)$ is large, then the time step τ to the next reaction R_j will be small, since this quantity is inversely proportional to the propensity $a_j(X(t))$.

In the tau-leaping method, a fixed time τ is considered with the assumption that during the time interval $[t, t+\tau)$ the propensities $a_j(X(t))$ remain approximately constant ($a_j(X(s)) \approx a_j(X(t))$ for any $t \leq s \leq t+\tau$). As a consequence of this assumption, the number of times

the reaction R_j occurs in the time interval $[t, t+\tau)$ may be approximated by a Poisson random variable, $P_j(a_j(X(t)), \tau)$, with mean and variance, $a_j(X(t))\tau$ (see [42]). Then the state at the later time, $X(t+\tau)$, can be approximated as

$$X(t+\tau) = X(t) + \sum_{j=1}^{M} \nu_j P_j(a_j(X(t)), \tau). \tag{2.11}$$

Hence the tau-leaping method, with stepsize τ, requires the following steps for the simulation:

1. Using the independent Poisson random variables $\{P_j(a_j(X(t)), \tau)\}_{j=1}^{M}$, obtain samples $\{k_j\}_{j=1}^{M}$.

2. Update the state as $X(t+\tau) = X(t) + \sum_{j=1}^{M} \nu_j k_j$ and the time t to $t+\tau$.

3. Return to step 1 or else stop the integration.

Note that the tau-leaping algorithm coincides with the SSA in the limit where $\tau \to 0$. The tau-leaping algorithm is more efficient than the SSA for some systems if many reactions occur during the step τ [42]. However, it is important to make sure appropriate assumptions are in place when using the tau-leaping method.

In the context of stiff systems, which is often the setting for cellular chemical systems, the time step τ with the explicit tau-leaping method will be small as it is bounded by the fastest-time scales relevant to the system.

2.7.4 The Efficient Tau Selection Procedure

Recall the tau-leaping method presented in Section 2.7.3. In this scenario it is necessary that the time step τ is chosen to satisfy the leap condition. Two issues need to be addressed. One pertains to the problem of choosing the largest value of τ satisfying the leap condition, and the other is concerned with being able ensure that the generated samples $\{k_j\}_{j=1}^{M}$ do not cause any of the population numbers to become unrealistic negative values.

In what follows, a method for choosing an appropriate τ satisfying the leap condition for the tau-leaping method will be presented. This stepsize selection scheme is originally described in [42], with further improvements being made in [14, 44]. Consider the relative change in the propensity functions during the time interval $[t, t+\tau)$ given by $\Delta_\tau a_j/a_j$, where

$$\Delta_\tau a_j = a_j(X(t+\tau)) - a_j(X(t)).$$

The objective now is to choose τ so that the $\Delta_\tau a_j/a_j$ is bounded by some small tolerance $0 < \varepsilon << 1$, which is prespecified. This in turn results in choosing other bounds $\varepsilon_i = \varepsilon_i(\varepsilon, X_i)$ so that the relative change $\Delta_\tau X_i/X_i$ of each species population is bounded by ε_i. The individual bounds $\varepsilon_i = \varepsilon_i(\varepsilon, X_i)$ can be computed according to [14].

However, the above bound for the relative changes in the propensity functions may lead to inaccurate numerical results in practice. A more accurate and efficient numerical solution is obtained if instead of bounding the relative change in the propensities, the bound is

applied to the relative change in molecular populations [14].

Consider the following bound obtained from the tau-leaping formula (2.11):

$$|\Delta_\tau X_i| \leq max\{\varepsilon_i X_i, 1\} \quad \text{for any} \quad 1 \leq i \leq N. \tag{2.12}$$

This then gives

$$\Delta_\tau X_i = \sum_{j=1}^{M} P_j(a_j(X(t)), \tau)\nu_{ij}. \tag{2.13}$$

Denote by $\langle . \rangle$ and var $\{.\}$ the mean and the variance of a random variable, respectively. Then since the Poisson random variables $P_j(a_j(X(t)), \tau)$ have mean and variance equal to $a_j(X(t))\tau$, the mean and variance of $\Delta_\tau X_i$ are given by

$$\langle \Delta_\tau X_i \rangle = \sum_j \nu_{ij}(a_j\tau), \quad var\{\Delta_\tau X_i\} = \sum_j \nu_{ij}^2(a_j\tau). \tag{2.14}$$

Thus, both of the quantities $\langle \Delta_\tau X_i \rangle$ and $var\{\Delta_\tau X_i\}$ satisfy the bound expressed in (2.12). This leads to an estimate of the appropriate τ which is explicitly discussed in Section 4.2.

In regards to the second issue of ensuring the sampled numbers $\{k_j\}_{j=1}^M$ don't allow the population numbers to be negative valued, various methods have been proposed. Tian & Burrage [112] and Chatterjee et al. [17] suggest replacing the unbounded Poisson random numbers with bounded binomial random numbers. However, this strategy may give

inaccurate results. On the other hand, Cao et al. [13] offer a different method. In this approach, an integer n_c is chosen and then reactions are identified as being either critical or non-critical in terms of n_c. A critical reaction having non-zero propensity is defined as one that is within n_c reactions away from completely depleting a reactant. Otherwise, a reaction is considered non-critical. Having distinguished between critical and non-critical reactions, then either of the following two methods are applied. Non-critical reactions are simulated by computing a time step τ as just described, and then using the standard tau-leaping method. SSA step is applied to the system of critical reactions. Take,

$$\sum_{1\leq k\leq j,\quad \text{k critical}} a_k(X(t)) > \xi_1 a_{sum,cr}(X(t))$$

and

$$\tau'' = \ln(1/\xi_2)/a_{sum,cr}(X(t))$$

to estimate the time τ'' to the next critical reaction and its index j_c. The time step τ is then chosen to be the minimum of τ' and τ''. If the minimum is τ', no critical reaction fires, and if it is τ'', only one critical reaction R_{jc} fires. Since the number of critical reactions firing during τ is at most one, critical reactions do not drive populations negative.

The tau selection strategy described above was introduced in Cao et al. [14]. Provided that n_c is sufficiently large such that every reaction becomes critical, then this procedure

reduces to the SSA. Tests show that for many biochemical systems, the explicit tau leaping method leads to significantly faster simulations than the SSA with only a minimum loss of accuracy.

2.8 Chemical Langevin Equation (CLE)

If, in addition to the assumption that the propensities $a_j(X(t))$ remain constant over a time interval $[t, t+\tau)$ the following assumption is made: τ is large enough such that $a_j(x)\cdot\tau >> 1 \quad \text{for all} \quad j = 1, ..., M$, then the tau-leap method may be further approximated. Since the mean $a_j(X(t))\cdot\tau$ of the Poisson random variable $P_j\left(a_j(x)\tau\right)$ is large, then this Poisson random variable can be approximated by a normal random variable with the same mean and variance. Making this transition in the state used in the tau-leaping method yields another less refined model, known as the Chemical Langevin Equation (CLE) [37].

To derive the CLE the Poisson random variables $P_j(a_j(X(t)), \tau)$ appearing in (2.11) are substituted with $a_j(X(t))\tau + \sqrt{a_j(X(t))\tau}N_j(0,1)$, where the $N_j(0,1)$ are independent normal random variables having a mean of 0 and a variance of 1. This gives what is knows as the Langevin leaping formula [37] for the state:

$$X(t+\tau) = X(t) + \tau\sum_{j=1}^{M}\nu_j a_j(X(t)) + \sqrt{\tau}\sum_{j=1}^{M}\nu_j\sqrt{a_j(X(t))}N_j(0,1). \tag{2.15}$$

Now, the state $X(t)$ becomes a continuous random variable as opposed to a discrete one in the previous setting. Computationally speaking, since normal random numbers can be

generated more efficiently than Poisson random numbers, the Langevin formula offers an improvement in comparison to the standard tau-leaping method. The algorithm for solving the CLE numerically is analogous to the tau-leaping algorithm and is accomplished as follows:

1. Choosing independent samples $\{N_j\}_{j=1}^{M}$ using the normal (0, 1) distribution.
2. Updating the state $X(t+\tau) = X(t) + \tau \sum_{j=1}^{M} \nu_j a_j(X(t)) + \sqrt{\tau} \sum_{j=1}^{M} \nu_j \sqrt{a_j(X(t))} N_j$ and time t to $t+\tau$.
3. Returning to step 1 or else stopping the integration.

Seperating $X(t+\tau) - X(t)$ in (2.15) with the left hand side, dividing through by τ and taking $\tau \to dt$ leads to the following stochastic differential equation known as the Chemical Langevin Equation (CLE) [37]:

$$dX(t) = \sum_{j=1}^{M} \nu_j a_j(X(t))dt + \sum_{j=1}^{M} \nu_j \sqrt{a_j(X(t))} dW_j(t), \tag{2.16}$$

where $W_j(t)$ are independent scalar Brownian motions for all $1 \leq j \leq M$. In this context, the solution provided by (2.15) is commonly referred to as the Euler-Maruyama method [55, 56] for the SDE given by (2.16). Suffice it to say that SDE models are generally derived through the addition of stochastic terms to an existing deterministic model.

In the thermodynamic limit where species populations X_i and the system's volume Ω all approach ∞, while the concentrations X_i/Ω remain constant, the propensity functions

$a_j(X(t))$ also grow linearly in the size of the system. By examining (2.15) and (2.16), it is seen that the deterministic terms grow linearly in the system's size terms, whereas the stochastic terms grow as the square root of the system's size. This implies that fluctuations in systems of interest generally scale as the inverse of the square root of the system's size.

Recall that one assumption made in this method required the propensities $a_j(X(t))$ to remain approximately constant over the time interval $[t, t+\tau)$, which suggests that τ be kept small. On the contrary, the additional assumption that was made that the means $a_j(X(t))\tau$ also be large requires τ to be sufficiently large. However, these assumptions can all hold if the molecular numbers of the system are large. Thus the CLE model applies when all species have large populations. Also, we note that in the thermodynamic limit the stochastic terms in (2.16) become negligible in comparison to the deterministic terms, so that the CLE reduces to the RRE (2.17). In this way, the tau-leaping method can be thought of as a transition from the discrete, stochastic model of the CME to the continuous, deterministic model provided by the RRE.

2.9 Reaction Rate Equation (RRE)

Modeling of biochemical systems in terms of concentrations and instantaneous rates of change requires that the molecule count is very large for each species. In the presence of thermodynamic limit, species populations and the system's volume all approach ∞, while the concentrations of the species remain constant [47]. The noise present in the stochastic model of the CLE becomes negligible in the thermodynamic limit. Therefore, in such a limit the CLE (2.16) (as was discussed in Section 2.8) can be reduced to the Reaction Rate

Equation

$$\frac{dX(t)}{dt} = \sum_{j=1}^{M} \nu_j a_j(X(t)), \tag{2.17}$$

which is a system of ODEs with dimension given by the number of species in the system. The method of calculating chemical reaction kinetics involves a state vector $X(t) \in \mathbb{R}^N$, where the ith component of the state vector, $X_i(t)$, is a non negative real number which represents the molar concentration of a species, denoted by S_i at a given time t. Molar concentration is $M = \frac{\text{moles of species}}{\text{volume in litres}}$, where 1 mole $= n_A \approx 6.023 \times 10^{23}$ units. As such, $x_i(t) \times n_A$ volume is the number of molecules of a species in a given volume. In such a setting, the concentrations of each species is assumed to vary continuously in time in accordance with the Reaction Rate Equation (RRE) [47]. To determine the RRE, the *law of mass action* is used. This law states that the rate of change of any chemical reaction is proportional to the product of the concentrations of the reacting species.

Despite being able to model systems satisfying certain assumptions, it is important to note that noise may still play a significant role in the system if some of the species populations have small molecular numbers. In such a scenario stochastic models are still required for an accurate description of the system dynamics.

2.10 Simulating Stiff Systems

When the fast and slow time-scales of a system of ODEs are well separated [93], and the fastest mode is stable, the system is considered "stiff" [15]. The solution of the deterministic

problem belongs to a slow manifold [7]. Outside this manifold the state moves rapidly towards it. Stiff problems arise in many practical applications of interest and much work has been invested to overcome the computational difficulties in finding numerical solutions of stiff ODEs [7]. It is worth mentioning that many RREs are stiff.

For simulating the SSA for stiff biochemical systems, small time steps are used. However, the common reactions occurring in the system are generally the fast ones. In practice, this implies that the simulation is often too slow to be efficient [15, 93]. The explicit tau-leaping method is similar to explicit solvers in the context of stiff deterministic systems, in that it is very slow. Recall that in the tau-leaping algorithm as discussed in Section 2.7.3, the time step τ is chosen in order to ensure that certain requirements are met, which also restricts τ to time-scales corresponding to the fasted modes of the system.

2.11 Numerical Simulations

The simulation algorithms for stochastic and deterministic models are previously discussed in Section 2.7. In this chapter we show how under certain assumptions, the algorithms are related to each other. In this section, we show the numerical simulations of each type of algorithm (SSA, Euler-Maruyama for CLE and ODE solver for RRE) on a simple system, known as the Michaelis-Menton model [47, 119]. We also show how to derive the Chemical Langevin Equations (CLE) and Reaction Rate Equations (RRE) for this model.

Table 2.1: Michaelis Menten model

R_j	Reaction	Propensities	Reaction rate
R_1	$S_1 + S_2 \xrightarrow{C_1} S_3$	$a_1 = C_1 X_1 X_2$	$C_1 = 1.661 \times 10^{-3}$
R_2	$S_3 \xrightarrow{C_2} S_1 + S_2$	$a_2 = C_2 X_3$	$C_2 = 10^{-4}$
R_3	$S_3 \xrightarrow{C_3} S_4 + S_2$	$a_3 = C_3 X_3$	$C_3 = 0.1000$

2.11.1 Michaelis-Menten Model

We consider a simple model, known as the Michaelis-Menton model (Figure 2.2) for enzyme kinetics [47, 119]. The model describes the rate an enzyme transforms a substrate into a product. In the Michaelis-Menten model, there are four molecular species involved in three reactions:

$$\begin{aligned} S_1 + S_2 &\xrightarrow{c_1} S_3 \\ S_3 &\xrightarrow{c_2} S_1 + S_2 \\ S_3 &\xrightarrow{c_3} S_4 + S_2. \end{aligned} \tag{2.18}$$

Here, species S_1 is a substrate and species S_2 is an enzyme. Species S_3 represents an enzyme-substrate complex and species S_4 is a product. The values for the Reaction Rate

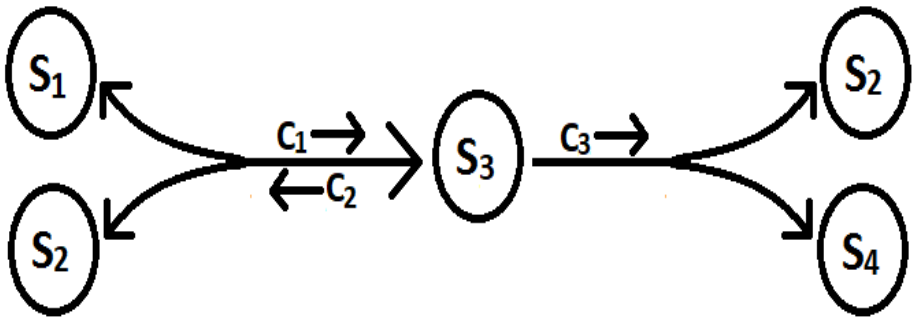


Figure 2.2: Michaelis Menten model reaction chain.

Parameters are $C_1 = 1.661 \times 10^{-3}$, $C_2 = 10^{-4}$ and $C_3 = 0.1$. The propensity functions for the reactions (2.18) are as follows:

$$\begin{aligned} a_1(X) &= C_1 X_1 X_2 \\ a_2(X) &= C_2 X_3 \\ a_3(X) &= C_3 X_3. \end{aligned} \tag{2.19}$$

We suppose the solution of the system (2.18) has the following initial conditions: $X_1(0) = 301$, $X_2(0) = 120$, $X_3(0) = 0$ and $X_4(0) = 0$. We perform the simulation on the time-interval $[0, 50]$ seconds. Note that the values of the rate constants and the initial conditions are from [47, 119]. The state-change vectors are given by the columns of the following stoichiometric matrix

$$V = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 1 & 1 \\ 1 & -1 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, the Chemical Langevin Equations (CLE) for the biochemical system are as follows:

$$dX_1 = [-a_1(x) + a_2(x)]dt - \sqrt{a_1(x)}dW_1 + \sqrt{a_2(x)}dW_2$$

$$dX_2 = [-a_1(x) + a_2(x) + a_3(x)]dt - \sqrt{a_1(x)}dW_1 + \sqrt{a_2(x)}dW_2 + \sqrt{a_3(x)}dW_3$$

$$dX_3 = [a_1(x) - a_2(x) - a_3(x)]dt + \sqrt{a_1(x)}dW_1 - \sqrt{a_2(x)}dW_2 - \sqrt{a_3(x)}dW_3$$

$$dX_4 = a_3(x)dt + \sqrt{a_3(x)}dW_3.$$

Finally, the Chemical Langevin Equation (CLE) for the Michaelis-Menten model can be written as follows:

$$dX_1(t) = [-C_1X_1(t)X_2(t) + C_2X_3(t)]dt - \sqrt{C_1X_1(t)X_2(t)}dW_1 + \sqrt{C_2X_3(t)}dW_2$$

$$dX_2(t) = [-C_1X_1(t)X_2(t) + C_2X_3(t) + C_3X_3(t)]dt - \sqrt{C_1X_1(t)X_2(t)}dW_1 + \sqrt{C_2X_3(t)}dW_2 + \sqrt{C_3X_3(t)}dW_3$$

$$dX_3(t) = [C_1X_1(t)X_2(t) - C_2X_3(t) - C_3X_3(t)]dt + \sqrt{C_1X_1(t)X_2(t)}dW_1 - \sqrt{C_2X_3(t)}dW_2 - \sqrt{C_3X_3(t)}dW_3$$

$$dX_4(t) = C_3X_3(t)dt + \sqrt{C_3X_3(t)}dW_3.$$

There are four species associated with this model and one equation for each different molecular species, as such, the dimension of this system is four for the CLE and RRE models. By neglecting the stochastic terms, we can derive the the deterministic model of the RRE as follows:

$$dX_1(t) = [-C_1X_1(t)X_2(t) + C_2X_3(t)]dt$$

$$dX_2(t) = [-C_1X_1(t)X_2(t) + C_2X_3(t) + C_3X_3(t)]dt$$

$$dX_3(t) = [C_1X_1(t)X_2(t) - C_2X_3(t) - C_3X_3(t)]dt$$

$$dX_4(t) = C_3X_3(t)dt.$$

These equations can also be expressed as follows:

$$\frac{dX_1(t)}{dt} = -C_1X_1(t)X_2(t) + C_2X_3(t)$$

$$\frac{dX_2(t)}{dt} = -C_1X_1(t)X_2(t) + C_2X_3(t) + C_3X_3(t)$$

$$\frac{dX_3(t)}{dt} = C_1X_1(t)X_2(t) - C_2X_3(t) - C_3X_3(t)$$

$$\frac{dX_4(t)}{dt} = C_3X_3(t).$$

The system was simulated with initial conditions $X(0) = [301, 120, 0, 0]$ and the parameters shown in Table 2.1, on the time interval $t = [0, 50]$. Figure 2.3 shows the trajectories of species S_1, S_2, S_3 and S_4 as functions of time simulated using Gillespies algorithm (SSA).

Figure 2.4 shows the trajectories of species S_1, S_2, S_3 and S_4 were simulated with the Euler-Maruyama method for CLE with respect to time. And finally simulations of RRE are presented in Figure 2.5.

The acceptability of the RRE as a model depends on the initial data, system parameters and the purpose for which the RRE model is used [47]. We found that the RRE was a fair match to the single paths that we drew from the Euler-Maruyama method for CLE and the SSA, although the number of molecules were in the hundreds.

The law of mass action, which states that the rate of a chemical reaction is directly proportional to the product of the activities or concentrations of the reactants is in many cases not appropriate for biochemistry within a cell, because when population size is extremely small, describing systems in terms of concentration is inappropriate [47]. Measuring the system responses of many cellular processes depend on precise quantitative values. Often, these processes involve population sizes that are very small and the system may switch between two distinct states that are driven by the inherent noise of the system. As a consequence, stochastic processes often play a vital role in cellular processes [47].

Plot: Michaelis-Menten model

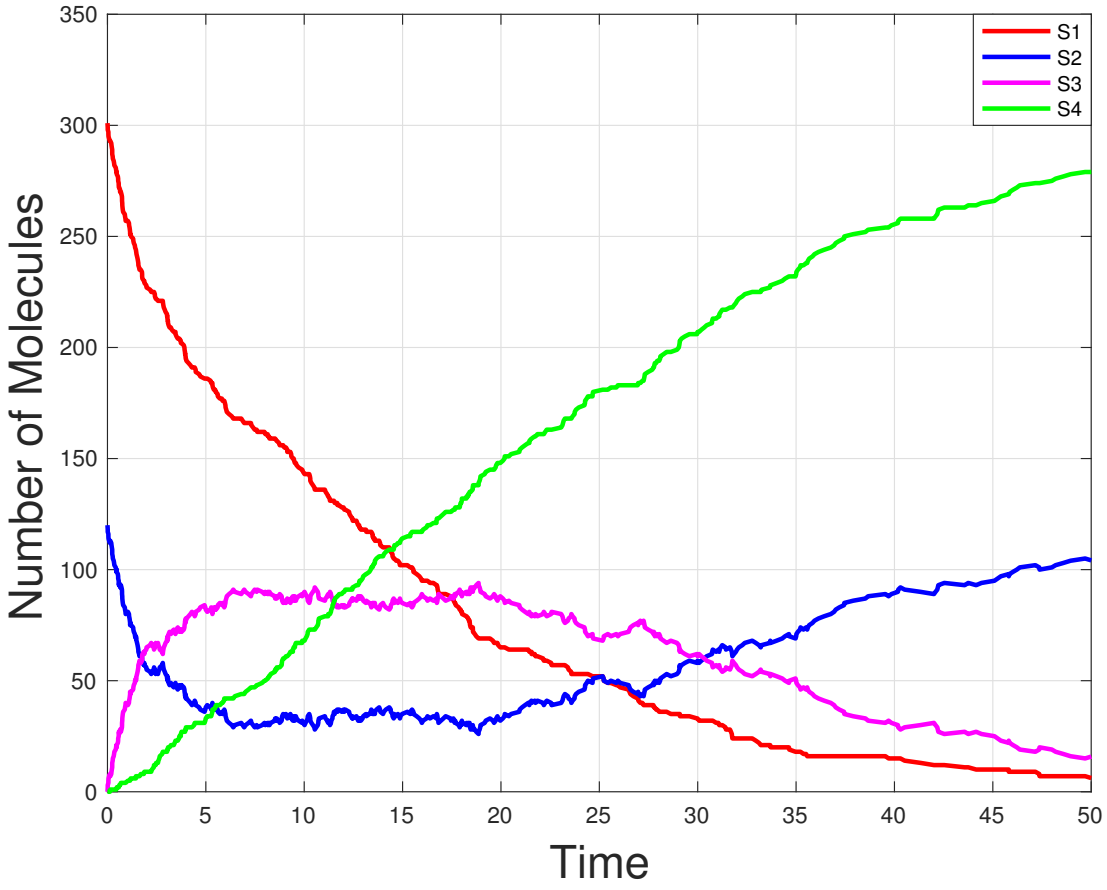


Figure 2.3: Plot for Michaelis-Menten model: Evolution in time of the species S_1 (red), S_2 (blue), S_3 (magenta), S_4 (green) with SSA simulated with time.

Plot: Michaelis-Menten model

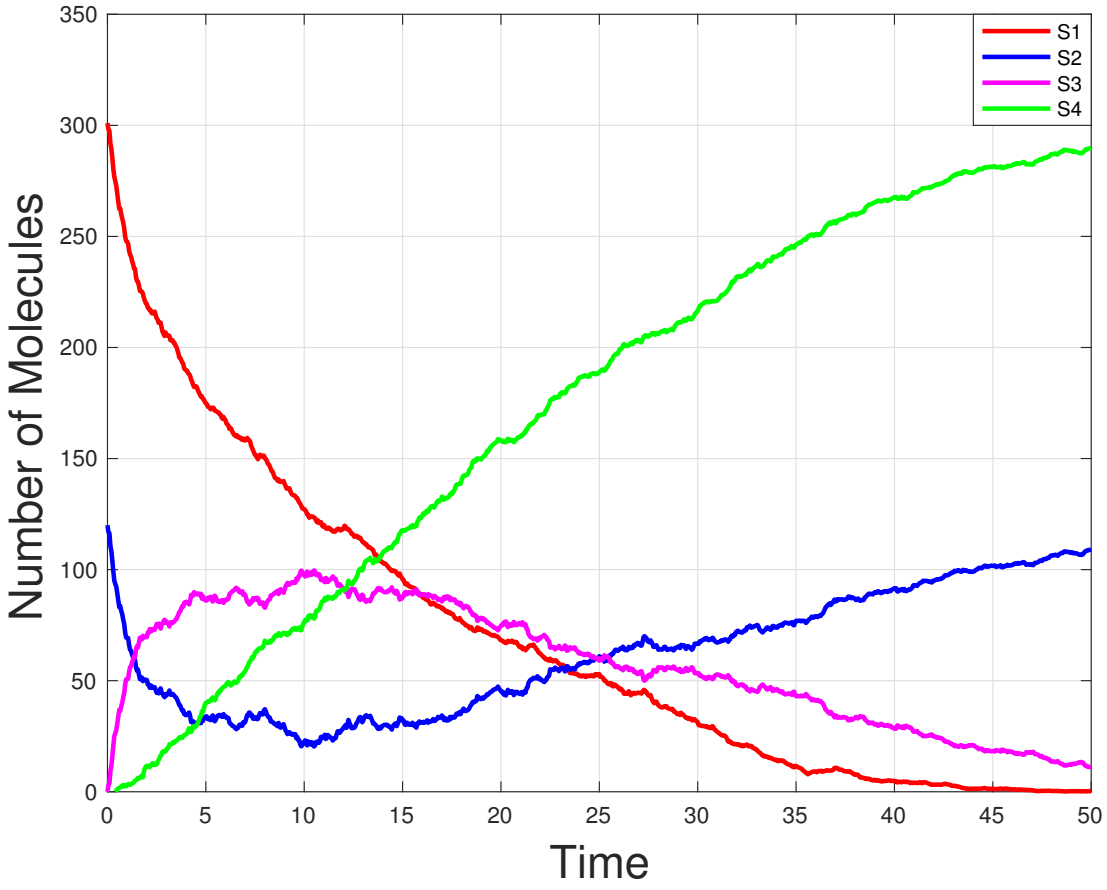


Figure 2.4: Plot for Michaelis-Menten model: Evolution in time of the species S_1 (red), S_2 (blue), S_3 (magenta), S_4 (green); CLE simulated with time.

Plot: Michaelis-Menten model

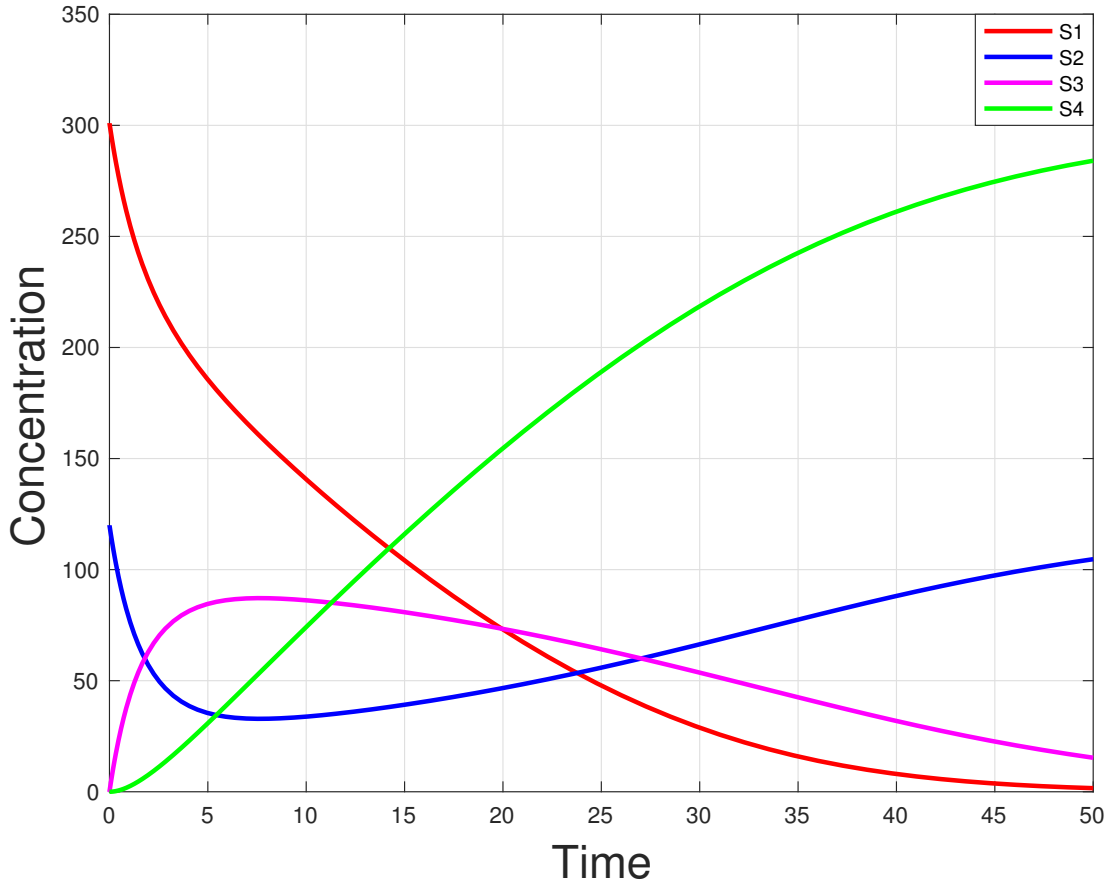


Figure 2.5: Plot for Michaelis-Menten model: Evolution in time of the species S_1 (red), S_2 (blue), S_3 (magenta), S_4 (green); RRE simulated with time.

Chapter 3

Sensitivity Analysis

3.1 Introduction

To study various subjects in the sciences, many computational techniques have been developed in order to simulate and experiment with such systems. One particular tool used to characterize the model of a system is sensitivity analysis, which attempts to understand how certain properties of the model change when variations are introduced into the model's parameter values [116]. A model output's sensitivity to a parameter is a measure of how much change in the system output results as a consequence of varying the parameter values. Model behaviour is regarded as highly sensitive to a parameter's value when a small change in the parameter values results in a large change in the model's outcome. Sensitivity analysis is particularly valuable when addressing biochemical systems for which some parameter values are poorly estimated.

Sensitivity analysis offers a great utility in being able to describe a system in terms of perturbations of the system's parameters. Since there may exist parameters in which only slight variations result in significant changes in the system's output, it is therefore important to identify and understand the effects of such change. In chemical reaction models the relevant parameters include kinetic parameters and initial conditions such as initial amounts for each species. Other undetermined parameters may exist that affect the system due to uncertain environmental interactions. By understanding the sensitivity of a system's model to different parameters, useful insights are revealed about the model which can address issues regarding the model's dynamics, and even help inform the model's own development and accuracy.

Sensitivity analyses can be classified as local (in which only small perturbations around a nominal set of parameter values are considered) or global (in which values over a wide region in parameter space can be addressed). Global analyses are computationally expensive, as they typically demand sampling of a high-dimensional parameter space [107]. In contrast, for deterministic systems, the computation of local parametric sensitivity coefficients poses no challenges. The computational cost to perform a local sensitivity analysis is high for a stochastic model. This is due to the need to simulate a large number of sample paths to generate accurate statistics [5].

3.2 Mathematical Theory of Sensitivity Analysis

To illustrate sensitivity analysis [116] in the context of models of biochemical systems, we can use a variety of chemical systems that exhibit different characteristics. In general,

we use their description in mathematical terms, which is given by models that shows an explicit or implicit relationship between the system behavior and the input parameters. The system behaviour can then be described in terms of dependent output variables or state that change in time and/or space. An important part of these models is input parameters which include the physiochemical parameters [116]. Some examples of the input parameters are transport properties, related reaction kinetics as well as initial conditions and operating conditions. However, these parameters are subject to uncertainties because they are either measured experimentally or estimated theoretically. Sensitivities can be quantified using the partial derivative with respect to a certain parameter. The mathematical methods for estimating parameter sensitivities use finite-difference approximations for derivatives.

3.2.1 Local Sensitivity

Recall the Reaction Rate Equation that was described in Section 2.9. Consider a chemical system that can be described by the following differential equation of variable of X with respect to changes in time t [116],

$$\frac{dX}{dt} = f(X, c, t) \tag{3.1}$$

where the function f is dependent on X, the variable t is the time and c is a vector representing m parameter inputs of the system. To ensure that the above equation has a unique solution, the function f is assumed to be continuous as well as continuously differentiable everywhere in its arguments. It should be noted that this statement holds true for virtually all chemical systems.

The unique solution, called the *nominal solution* [116] is represented by

$$X = X(t, c) \tag{3.2}$$

which is continuous in t and c.

If the jth parameter in the parameter vector c, is changed from c_j to $c_j + \Delta c_j$, then the corresponding nominal solution becomes

$$X = X(t, c_j + \Delta c_j) \tag{3.3}$$

and is called the *current solution*.

If Δc_j is sufficiently small, *i.e.*, $\Delta c_j \ll c_j$ the current solution can be expanded into the following truncated Taylor series:

$$X(t, c_j + \Delta c_j) = X(t, c_j) + \frac{\partial X(t, c_j)}{\partial c_j} \cdot \Delta c_j. \tag{3.4}$$

It follows from this equation that the local sensitivity of the dependent variable, X, with respect to the input parameter c can be written as follows:

$$s(X; c_j) = \frac{\partial X(t, c_j)}{\partial c_j} = \lim_{\Delta c_j \to 0} \frac{X(t, c_j + \Delta c_j) - X(t, c_j)}{\Delta c_j}. \tag{3.5}$$

The local sensitivity, $s(X; c_j)$, is also known as *absolute sensitivity*.

Normalized magnitudes are often used in sensitivity analysis. This *normalized sensitivity* of X with respect to c_j is defined as [116]:

$$S(X; c_j) = \frac{c_j}{X} \cdot \frac{\partial X}{\partial c_j} = \frac{c_j}{X} \cdot s(X; c_j). \tag{3.6}$$

The normalized sensitivity is also referred to as *relative sensitivity* and serves to normalize the magnitudes of the input parameter c_j and the variable X.

Consider the sensitivity of X with respect to each one of the parameters in the m vector c. The *row sensitivity vector* of m indices is now defined as below [116],

$$s^T(X; c) = \frac{\partial X}{\partial c} = \left[\frac{\partial X}{\partial c_1} \frac{\partial X}{\partial c_2} \cdots \frac{\partial X}{\partial \phi_m}\right] = \left[s(X; c_1) s(X; c_2) \cdots s(X; c_m)\right]. \tag{3.7}$$

3.2.2 Global Sensitivity

Local sensitivities, $s(X_i; c_j)$, describe the effect of a small variation in each parameter, c_j, around a fixed nominal value, on each dependent variable, X_i. On the other hand, global sensitivities provide information on the effect of simultaneous large variations of all parameters, c, on the dependent variables [107, 116].

As previously shown, when the perturbation size, Δc_j of input parameter c_j, is small, the Taylor series expansion is truncated after the linear term, and local sensitivities can be well approximated with partial derivatives. As such, the local sensitivity $s(X_i; c_j)$ for a given c_j, is considered as a function of independent variable, t. When considering global sensitivities, all parameters are simultaneously varied over a wide range of values and

the corresponding variations of the dependent variables are then functions of the varied range of the parameters [45, 107, 116]. As such, global sensitivities cannot be defined by mathematical formulae as easily as local sensitivities and can only be evaluated via numerical calculations.

3.3 Sensitivity Methods

3.3.1 Direct Differential Method (DDM)

A natural method for computing sensitivities is the direct differential method (DDM) [116]. Consider equation (3.1) for a single variable system. Now, differentiate both sides with respect to C to compute the local sensitivity of X with respect to the jth parameter, c_j. Applying the definition of (3.5), then the local sensitivity equation can be defined as follows [116]:

$$\frac{d}{dt}\left(\frac{\partial X}{\partial c_j}\right) = \frac{ds(X;c_j)}{dt} = \frac{\partial f}{\partial X}\cdot\frac{\partial X}{\partial c_j} + \frac{\partial f}{\partial c_j}\cdot\frac{\partial c_j}{\partial c_j} = \frac{\partial f}{\partial X}\cdot s(X;c_j) + \frac{\partial f}{\partial c_j} \tag{3.8}$$

which represents the local *sensitivities equation* whose initial conditions can be obtained via a similar differentiation of the initial condition. Furthermore, depending on the chosen input parameter vector c,

$$s(X;c_j)\mid_{t=0} = \begin{cases} 0 & c_j \neq X^i \\ 1 & c_j = X^i \end{cases}$$

When simultaneously solving the model (3.1) and its sensitivity equation (3.8) along with their initial conditions, the dependent variable X along with its corresponding local sensitivity $s(X; c_j)$ can be obtained, both as functions of time. Together, this method is the *direct differential method* (DDM) [116].

In order to compute the sensitivity of the ith output variable, X_i among n output variables with respect to the jth input parameter, c_j, we need to find the sensitivity of all n output variables with respect to c_j due to possible interactions with each other. Therefore we need to solve n sensitivity equations and the n model equations simultaneously.

The n sensitivity equations can be written as follows [116],

$$\frac{d\mathbf{s}(\mathbf{X}; c_j)}{dt} = \mathbf{J}(t) \cdot \mathbf{s}(\mathbf{X}; c_j) + \frac{\partial \mathbf{f}(t)}{\partial c_j}$$

$$\mathbf{J}(t) = \frac{\partial \mathbf{f}}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f_1}{\partial X_1} & \frac{\partial f_1}{\partial X_2} & \cdots & \frac{\partial f_1}{\partial X_n} \\ \frac{\partial f_2}{\partial X_1} & \frac{\partial f_2}{\partial X_2} & \cdots & \frac{\partial f_2}{\partial X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial X_1} & \frac{\partial f_n}{\partial X_2} & \cdots & \frac{\partial f_n}{\partial X_n} \end{bmatrix}, \quad \frac{\partial f(t)}{\partial c_j} = \begin{bmatrix} \frac{\partial f_1}{\partial X_1} \\ \frac{\partial f_2}{\partial X_1} \\ \vdots \\ \frac{\partial f_n}{\partial X_1} \end{bmatrix}$$

where the latter J is the $n \times n$ *Jacobian Matrix* and the former as the $n \times 1$ *nonhomogeneous term*, respectively.

3.3.2 Finite Difference Approximations (FDM)

The finite difference approximation can be used to avoid simultaneously solving model and its sensitivity equation for computing local sensitivities [96, 109, 116]. In trying to estimate the sensitivities of certain parameters in sensitivity analysis, a function representing some property of the system is described in terms of input parameters. In the finite difference method, the derivative of this function is approximated by the difference in the function's values at different given values of the variable parameter. Thus, consider some function $X(t, c_j)$ and the Taylor series expansion of $X(t, c_j + \Delta c_j)$:

$$X(t, c_j + \Delta c_j) = X(t, c_j) + \Delta c_j \frac{\partial X(t, c_j)}{\partial c_j} + \frac{\Delta c_j^2}{2!} \frac{\partial^2 X(t, c_j)}{\partial c_j^2} + \dots . \tag{3.9}$$

Then solving for $\frac{\partial X(t,c_j)}{\partial c_j}$ gives an expression of the form

$$\frac{\partial X(t, c_j)}{\partial c_j} = \frac{X(t, c_j + \Delta c_j) - X(t, c_j)}{\Delta c_j} + O(\Delta c_j)$$

where $O(\Delta c_j)$ denotes all higher order terms in Δc_j. This first-order approximation for a given step Δc_j is called the finite-difference interval. A second order estimate can be obtained by also considering the Taylor expansion of $X(t, c_j - \Delta c_j)$:

$$X(t, c_j - \Delta c_j) = X(t, c_j) - \Delta c_j \frac{\partial X(t, c_j)}{\partial c_j} + \frac{\Delta c_j^2}{2!} \frac{\partial^2 X(t, c_j)}{\partial c_j^2} - \dots . \tag{3.10}$$

Then by subtracting (3.10) from equation (3.9), the resulting equation can be solved for $\frac{\partial^2 X(t,c_j)}{\partial c_j}$ to obtain the central difference formula

$$f^{'}(X) = \frac{X(t, c_j + \Delta c_j) - X(t, c_j - \Delta c_j)}{2\Delta c_j} + O(\Delta c_j^2),$$

where now $O(\Delta c_j^2)$ represents all terms of order Δc_j^2 or greater.

When determining the local sensitivities of n output variables with respect to one among m input parameters, the Direct Difference Method (DDM) requires to solve $m + 1 \times n$ model and sensitivity equations simultaneously [116]. To avoid this, the Finite Difference Method (FDM) is utilized which requires only to solve the n model equations twice for $c_j = c_j$ to get $X(t, c_j)$ and $c_j = c_j + \Delta c_j$ to compute $X(t, c_j + \Delta c_j)$. This method is known as the Finite Difference Method (FDM) [116].

To compute the sensitivities of one among n output variables with respect to one among m input parameters at a given point, it is simple to determine the variation Δc_j for different input parameters. Because of this, the Finite Difference Method (FDM) is very useful [116]. This method can be used in order to find the sensitivity of ith output variable, X_i among n output variables with respect to the jth input parameter, c_j.

In practical uses, there may be cases when DDM cannot be used, since the sensitivity equations cannot be found by directly differentiating the equations of the model. If the sensitivity of the output of interest is implicitly given by a complex form or does not have a mathematical representation, then the FDM is the only method which can be used [116].

3.4 Monte Carlo Approach to Sensitivity Analysis

In the previous section, the sensitivity methods that we discussed can be used for ODE models. These methods are applicable when molecular counts of each species are large. However, biochemical systems of interest show inherently probabilistic behaviour when the molecular count is very small. As such stochastic models are needed to capture the random fluctuations observed in these systems. Stochastic models of biochemical systems and their simulation techniques were discussed in depth in Chapter 2.

The Monte Carlo sensitivity approaches with finite perturbation [96, 109] includes the independent sample method (independent random numbers (IRN) with SSA) and correlated sample method. Among the existing correlated sample methods we have: Common random numbers (CRN) with SSA, Common reaction Path (CRP) with RTC and Coupled finite difference (CFD) method which are discussed in depth in the following sections.

3.5 Established Finite-Difference Methods for Stochastic Systems

In sensitivity analysis, we wish to compare two systems. One system is defined as nominal, whereas the other system's parameter value is finitely perturbed. The same stream of random numbers is used for both systems [96]. This can be achieved by using a common seed if the scripting language allows for it, or by calling a large array of random numbers to be used for both systems. The common seed method is more efficient, if system resources

are sparse or the simulation is sufficiently large.

To approximate the sensitivity of a system output $f(X^c(t))$ with respect to a parameter c via a finite difference [2, 96, 109], (as described in depth in Section 3.2.2) we consider a pair of sample paths, one generated at a nominal parameter value c, the other generated from a perturbed value $c + h$. At a given time t, the nominal system state is $X^c(t)$, while the perturbed state is $X^{c+h}(t)$. we use the notation $a_j^c(x)$ to indicate the propensity of reaction j at state x when the parameter of interest takes the value c, i.e. $a_j^c(x) = a_j(x, c)$.

3.5.1 Common Random Numbers (CRN)

In the CRN method [96], Gillespie's algorithm (as described in depth in Section 2.7.1) is used to simulate sample paths. To test sensitivity with respect to a parameter, c, of the system, we consider a pair of systems: the nominal system $X^c(t)$ and the perturbed system, $X^{c+h}(t)$. The common random number approach is applied to the SSA with a shared stream of random numbers, so that both systems experience the same random input.

Algorithm

1. begin loop over number of trajectories, N, for each i
2. generate large array of random numbers, r^* (if common seed not used)
3. choose system parameter, c, and execute Gillespie's algorithm for the nominal system, $X^c(t)$, using array of random numbers, r_j^*

4. set parameter to $c + h$, and execute Gillespie's algorithm, calculating for perturbed system, $X^{c+h}(t)$, using the same array of random numbers, r_j^* (or common seed)

5. find sensitivity by $Z_i = (f(X^{c+h}(t)) - f(X^c(t)))/h$

6. end loop over i

7. find mean and standard deviation of $\{Z_i\}_{1 \leq i \leq N}$.

3.5.2 Common Reaction Path (CRP)

In the CRP method [96], the RTC algorithm (as described in depth in Section 2.7.2) is used to simulate sample paths, with common random number streams, as above. In this case, the trajectory of the whole system is determined by the collective of the independent trajectories from each reaction. For each reaction in the system, the trajectory is referred to as the *reaction path*. Each reaction path evolves independently as a series of random exponential numbers with unit rate.

Algorithm

1. loop over number of trajectories, N, for each i

2. generate large array of unit exponential random numbers, $E_1^j, E_2^j, \ldots$, for each reaction (or array of common seeds)

3. choose system parameter, c, and execute RTC algorithm for the nominal system, $X^c(t)$, using array of random numbers created

4. set parameter to $c+h$, and execute RTC algorithm, calculating for perturbed system, $X^{c+h}(t)$, using the same array of unit exponential random numbers,$E_1^j, E_2^j, \ldots$ (or common seeds)

5. find sensitivity by $Z_i = (f(X^{c+h}(t)) - f(X^c(t)))/h$

6. end loop over i

7. find mean and standard deviation of $\{Z_i\}_{1 \le i \le N}$.

3.5.3 Coupled Finite Difference (CFD)

The CFD method presented by Anderson [2] imposes tight coupling between the random processes generating the nominal and perturbed sample paths, thus achieving a reduced variance in the estimator. Using the random time change representation (2.10), the coupling of the nominal process $X^c(t)$ and the perturbed process $X^{c+h}(t)$ is obtained as follows:

$$\begin{aligned} X^c(t) &= X^c(0) + \sum_{j=1}^{M} \left\{ \nu_j Y_{j,1} \left(\int_0^t m_{j,c,h}(s)\, ds \right) + \nu_j Y_{j,2} \left(\int_0^t a_j^c(X^c(s)) - m_{j,c,h}(s)\, ds \right) \right\} \\ X^{c+h}(t) &= X^{c+h}(0) + \sum_{j=1}^{M} \left\{ \nu_j Y_{j,1} \left(\int_0^t m_{j,c,h}(s)\, ds \right) + \nu_k Y_{j,3} \left(\int_0^t a_j^{c+h}(X^{c+h}(s)) - m_{j,c,h}(s)\, ds \right) \right\} \end{aligned} \tag{3.11}$$

where $m_{j,c,h}(t) = \min\left\{a_j^c(X^c(t)), a_j^{c+h}(X^{c+h}(t))\right\}$ and $Y_{j,1}$, $Y_{j,2}$ and $Y_{j,3}$ are independent unit rate Poisson processes. The CFD generates the coupled paths with the next reaction method:

1. Begin loop over number of trajectories, N, for each i

2. Initialize the nominal and perturbed system states, $X(0) = \mathbf{x}_0$, at $t = 0$.

3. For each $j = 1, \ldots, M$ and for $k = 1,\ 2,\ 3$, initialize $\mu_{j,k} = rand(0,1)$, $P_{j,k} = \ln(1/\mu_{j,k})$, $T_{j,k} = 0$

4. while $t < T$

   (a) At each time t, for each j:

      i. compute the propensities $a_j^c(X^c(t))$ and $a_j^{c+h}(X^{c+h}(t))$,

      ii. set $A_{j,1} = \min\{a_j^c(X^c(t)), a_j^{c+h}(X^{c+h}(t))\}$,

      $A_{j,2} = a_j^c(X^c(t)) - A_{j,1}$,

      $A_{j,3} = a_j^{c+h}(X^{c+h}(t)) - A_{j,1}$.

      iii. for $k = 1:3$, if $A_{j,k} > 0$, set $\Delta t_{j,k} = (P_{j,k} - T_{j,k})/A_{j,k}$,

      otherwise set $\Delta t_{j,k} = \infty$.

   (b) Find minimum: $\Delta T = \min_{1 \le j \le M, 1 \le k \le 3} \{\Delta t_{j,k}\}$ and indices for min, $\mu = \{j^*, k^*\}$.

   (c) Increment the time: $t = t + \Delta T$

   (d) If $k^* = 1$, increment both $X^c(t)$ and $X^{c+h}(t)$:

   $X^c(t) = X^c(t) + \nu_{j^*}$ and $X^{c+h}(t) = X^{c+h}(t) + \nu_{j^*}$,

   if $k^* = 2$, increment $X^c(t)$: $X^c(t) = X^c(t) + \nu_{j^*}$,

   if $k^* = 3$, increment $X^{c+h}(t)$: $X^{c+h}(t) = X^{c+h}(t) + \nu_{j^*}$.

   (e) For each j and each $k = 1:3$, update $T_{j,k} = T_{j,k} + (A_{j,k})\Delta T$.

   (f) Compute $P_\mu = P_\mu + \ln(1/\mu)$, with $\mu = rand(0,1)$

5. Compute sensitivity by $Z_i = (f(X^{c+h}(t)) - f(X^c(t)))/h$

6. End while

7. End loop over i

8. Find mean and standard deviation of $\{Z_i\}_{1 \leq i \leq N}$.

The mean of the finite differences $Z = [f(X^{c+h}(t)) - f(X^c(t))]/h$ over an ensemble of coupled sample trajectories is an estimator for the sensitivity of $E(f(X^c(t)))$ to parameter c.

3.6 Numerical Results

We carried out an analysis of previously published finite difference approaches to local parametric sensitivity analysis of chemical master equation models. We provide a comparison of the performance of the Common Reaction Path (CRP), Common Random Number (CRN) method by Rathinam et al. [96] and Coupled Finite Deference (CFD) method by Anderson [2] on a range of model types. We tested the method's performance on a rich set of model dynamics. The interpretation of the following three models' results are presented at the end of this section.

3.6.1 Birth-death Model

The Birth-death model is a simple example of a one species reaction network (Figure 3.1). The reactions set and propensities for the Birth-death reaction network are given in Table

Table 3.1: Birth death model

R_j	Reaction	Propensities	Reaction rate
R_1	$\oslash \xrightarrow{C_1} X$	$a_1(x) = C_1$	$C_1 = 100$
R_2	$X \xrightarrow{C_2} \oslash$	$a_2(x) = C_2 X$	$C_2 = 5$

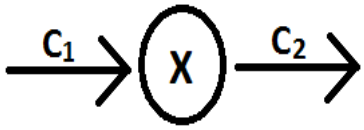


Figure 3.1: Birth death model reaction chain.

3.1, along with specific values for the reaction rate parameters. The initial condition was taken as $X(0) = 0$. The state-change vectors for the reactions are given by the stoichiometric matrix $V = [1, -1]$. For this model, we used the CFD, CRN and CRP methods to determine the sensitivity of the molecular count of species X over the time $t \in [0, 2]$ with respect to parameter C_2. The size of the perturbation was taken as $h = 10^{-1}$ (i.e. a 2% change); the estimates, shown in Figure 3.2, were each calculated from 80,000 sample paths. Figure 3.3 shows how the standard deviation of the sensitivity estimator varies with the perturbation size h for each method.

3.6.2 Schlögl Model

The Schlögl model is a simple example of a bistable reaction network (Figure 3.4). The model is well known for its bistable steady-state distribution. The reaction set and propen-

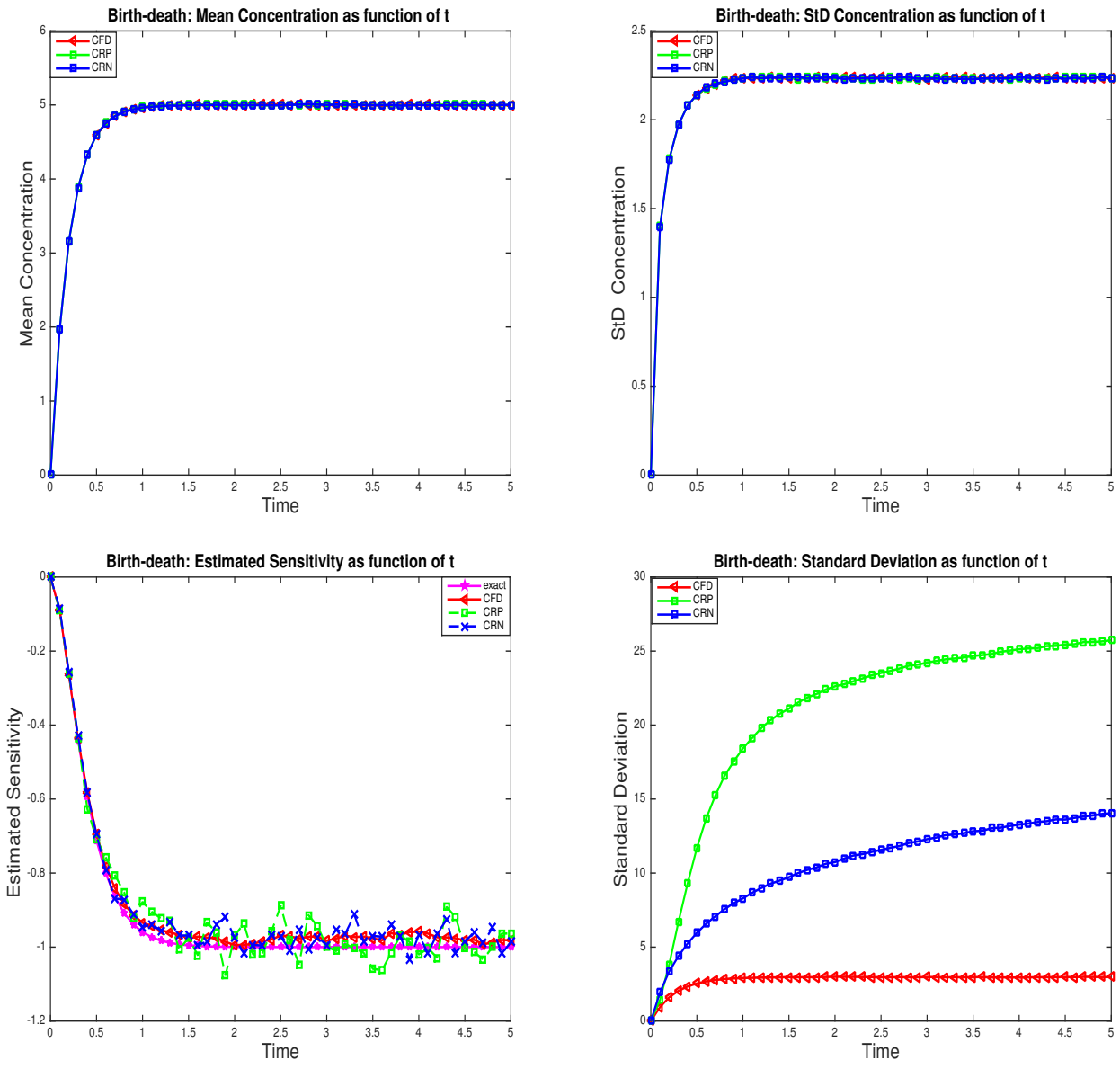


Figure 3.2: Birth death model: comparison of sensitivity methods with perturbation parameter $h = 10^{-1}$, using $80,000$ trajectories, on the interval $[0, 5]$ (species X). Left: estimated sensitivity; right: standard deviation of the estimated sensitivity.

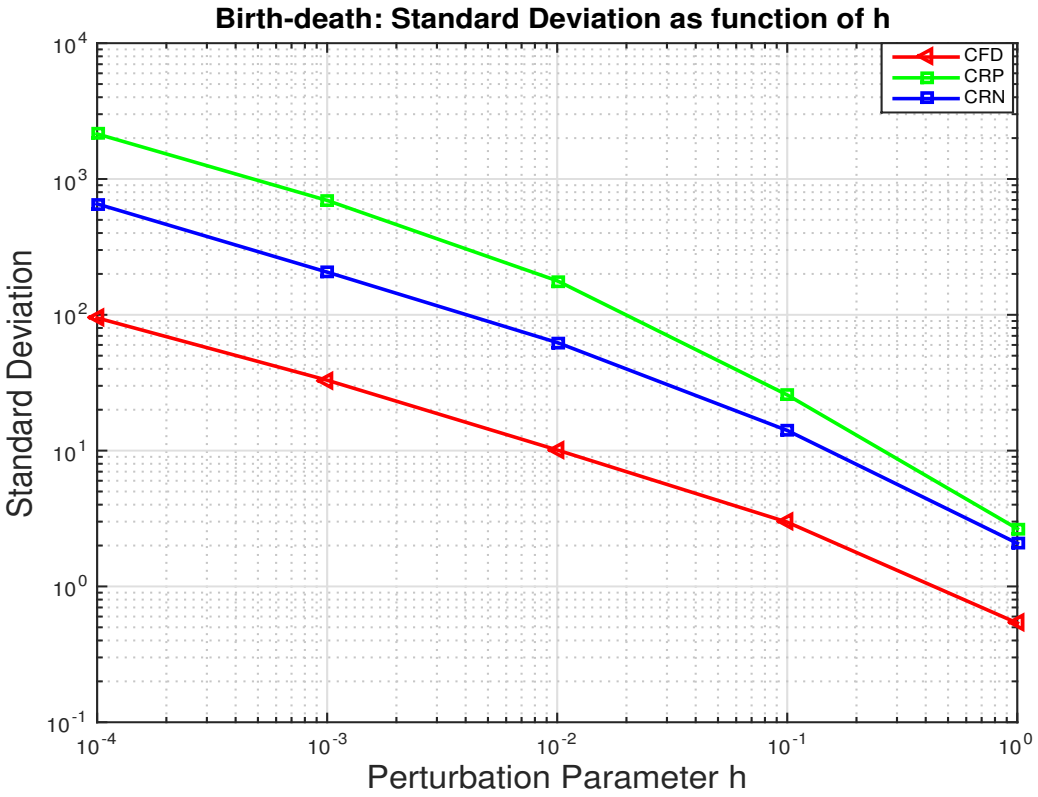


Figure 3.3: Birth death model: the log-log plot of the standard deviation as function of the perturbation parameter h, using $80,000$ trajectories, over $t \in [0, 5]$ (species X).

sities for this model are presented in Table 3.2, along with values for the reaction rate constants. The molecular counts for species A and B, which are denoted as buffered species, are held constant at $A = 10^5$ and $B = 2 \times 10^5$ over the time interval of interest. Species X's molecular population is modelled accurately as a homogeneous jump Markov process X(t) on the non-negative integers [39].

The initial count for species X is taken as $X(0) = 250$. The stoichiometric matrix for this system is $V = [1, -1, 1, -1]$. For this model, we used the CFD, CRN and CRP methods to

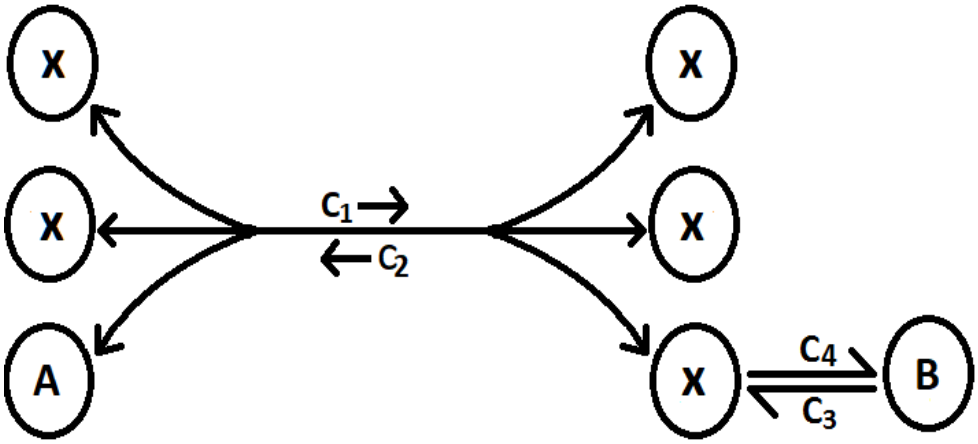


Figure 3.4: Schlögl model reaction network.

determine the sensitivity of the molecular count of species X over the time $t \in [0, 10]$ with respect to parameter C_1. The size of the perturbation was taken as $h = 5 \times 10^{-8}$ (i.e. a 17% change); the estimates, shown in Figure 3.5, were each calculated from 10,000 sample paths.

Table 3.2: Schlögl model

R_j	Reaction	Propensities	Reaction rate
R_1	$A + 2X \xrightarrow{C_1} 3X$	$a_1 = C_1 AX(X-1)/2$	$C_1 = 3 \times 10^{-7}$
R_2	$3X \xrightarrow{C_2} A + 2X$	$a_2 = C_2 X(X-1)(X-2)/6$	$C_2 = 10^{-4}$
R_3	$B \xrightarrow{C_3} X$	$a_3 = C_3 B$	$C_3 = 10^{-3}$
R_4	$X \xrightarrow{C_4} B$	$a_4 = C_4 X$	$C_4 = 3.5$

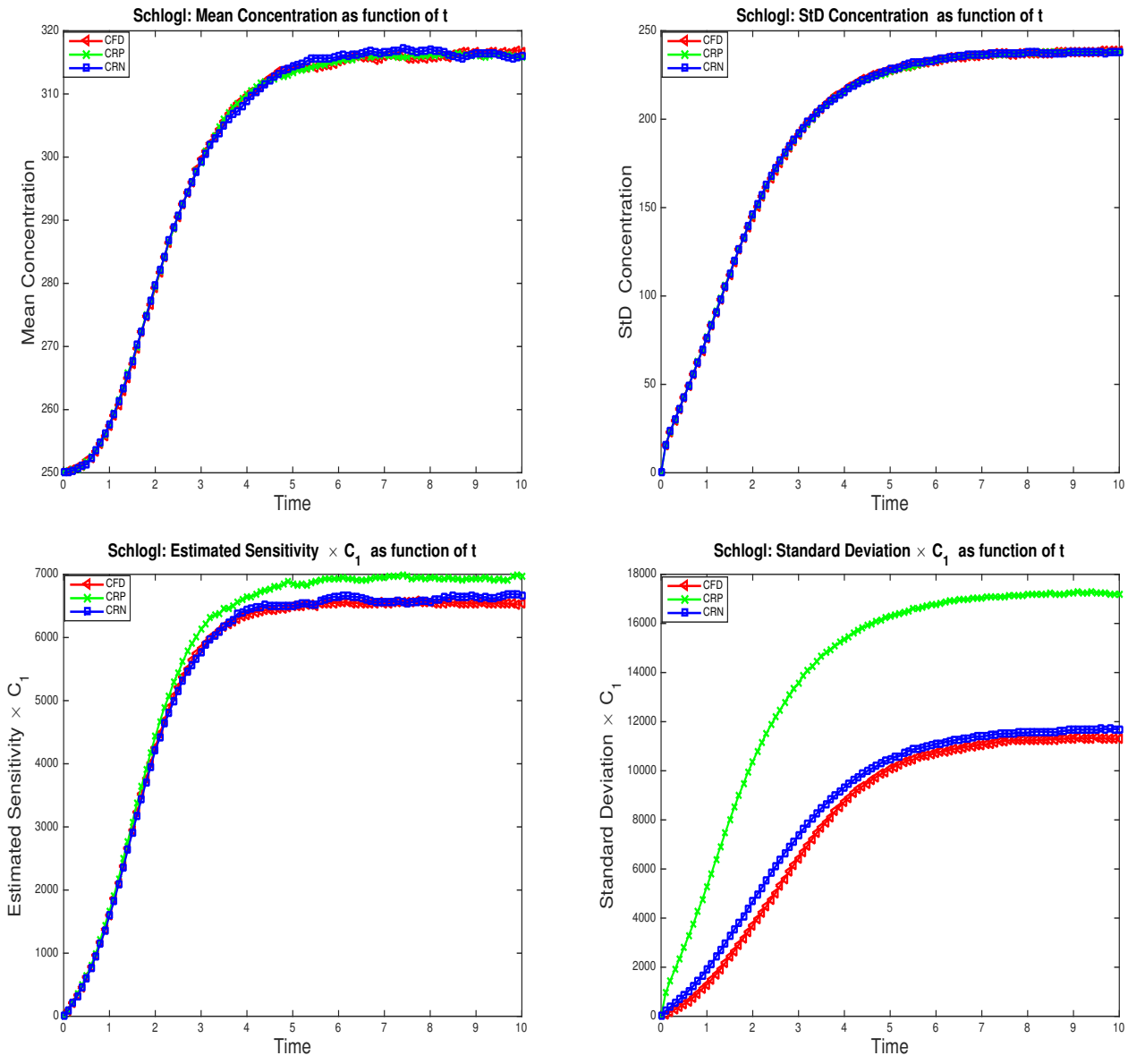


Figure 3.5: Schlögl model: comparison of sensitivity methods with perturbation parameter $h = 5 \times 10^{-8}$, using $10,000$ trajectories, over $t \in [0, 10]$ (species X) Left: estimated sensitivity; right: standard deviation of the estimated sensitivity.

3.6.3 Brusselator Model

The Brusselator model (Figure 3.6) exhibits stable oscillations [41]. The set of reactions for this model, their propensities and reaction rate parameter values are given in Table 3.3. The stoichiometric matrix is

$$V = \begin{bmatrix} 1 & -1 & 1 & -1 \\ 0 & 1 & -1 & 0 \end{bmatrix}.$$

The initial conditions were taken as $[X(0), Y(0)] = [1000, 2000]$. For this model, we used the CFD, CRN and CRP methods to determine the sensitivity of the molecular count of species X over the time $t \in [0, 5]$ with respect to parameter C_4. The size of the perturbation was taken as $h = 1$ (i.e. a 20% change); the estimates, shown in Figure 3.7, were each calculated from 200 sample paths.

Table 3.3: Brusselator model

R_j	Reaction	Propensities	Reaction rate
R_1	$\oslash \xrightarrow{C_1} X$	$a_1 = C_1$	$C_1 = 5000$
R_2	$X \xrightarrow{C_2} Y$	$a_2 = C_2 X$	$C_2 = 50$
R_3	$2X + Y \xrightarrow{C_3} 3X$	$a_3 = C_3 Y X(X-1)/2$	$C_3 = 0.00005$
R_4	$X \xrightarrow{C_4} \oslash$	$a_4 = C_4 X$	$C_4 = 5$

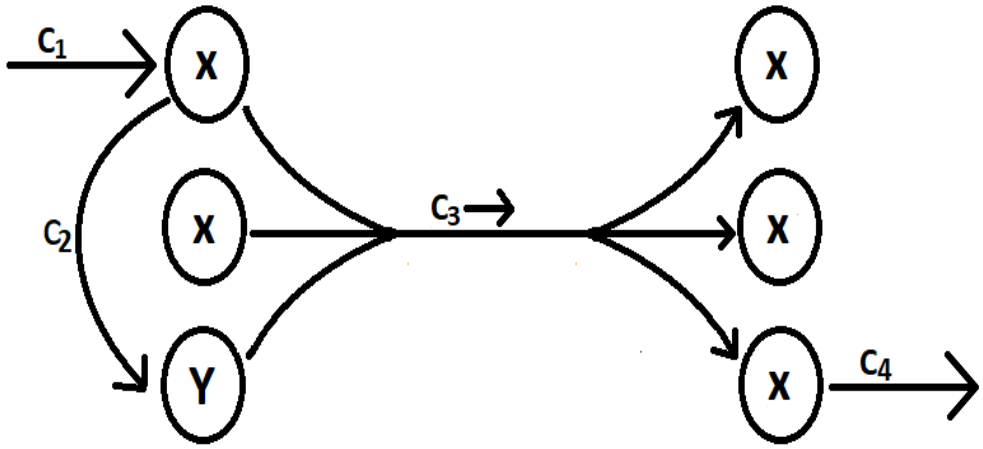


Figure 3.6: Brusselator reaction scheme diagram.

Discussion:

In this comparison we chose three different types of model to verify the accuracy of the existing sensitivity analysis techniques. We consider a simple model (the Birth-death model), a bistable reaction network model (the Schlögl model) and also a model that exhibits stable oscillations (the Brusselator model) for a given set of reaction parameters. For the first model we had the exact solution for comparison. It is to be determined how the behaviors exhibited by the other two models (particularly oscillations and bistability) can be handled by the existing methods.

Regarding the estimator variance, our results confirm that for the birth-death and Brusselator models, the CFD has the lowest variance, followed by the CRN and then the CRP method. This is consistent with the results in [109]. However, in the case of the Schlögl model, the CRN outperforms the CRP, and provides results that are comparable to the CFD.

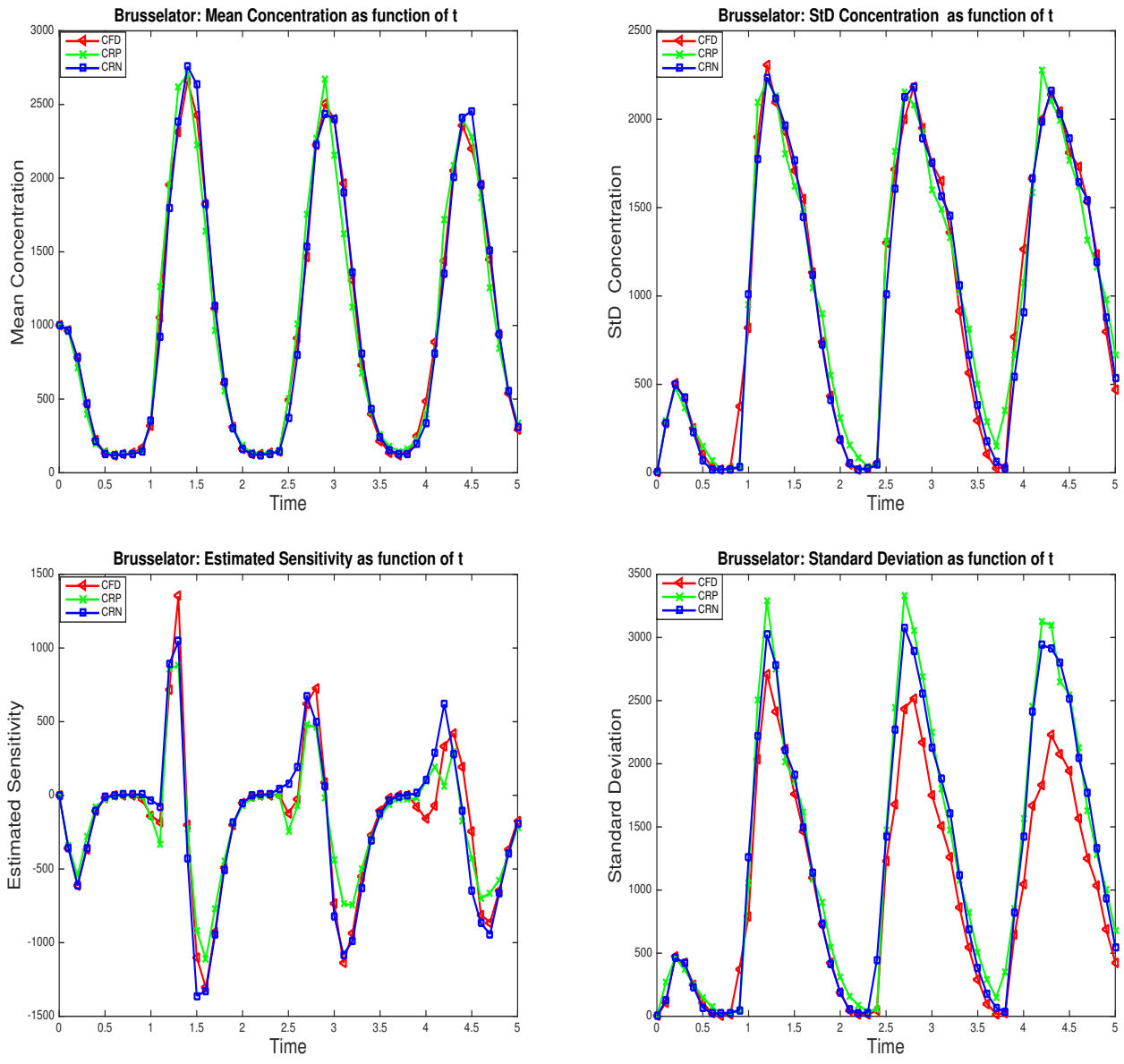


Figure 3.7: Brusselator model: comparison of sensitivity methods with perturbation parameter $h = 1$, using 200 trajectories, on the interval $t \in [0, 5]$ (species X). Left: estimated sensitivity; right: standard deviation of the estimated sensitivity.

As illustrated in Figure 3.3 for the Birth-death process, when the perturbation parameter h is increased, the variance of the sensitivity is reduced. A reduced variance means that a smaller number of trajectories are required to maintain the desired accuracy, thus leading to an improved efficiency of the simulation.

Chapter 4

Adaptive Coupled Tau-Leaping Method

4.1 Introduction

The materials in this Chapter are reproduced directly from the jointly-authored publication by Morshed, Ingalls and Ilie [83]. In this Chapter, we present the new Coupled Tau Leaping (CTL) algorithm [83], that is computationally efficient for approximating parametric sensitivities in moderately stiff stochastic biochemical systems.

Historically, simulations of stochastic models of well-stirred biochemical systems relied on the exact methods such as the SSA. While the tau-leaping algorithm has reduced computational time, large time steps can lead to over-consumption of species, which results in negative numbers. To address this issue, a tau-selection method is used to ensure that

the time step will be sufficiently small when reactions are close to exhausting their species. The state-of-the-art adaptive tau-leaping algorithm due to Cao et al. [14] finds τ such that, for any critical reaction, only one firing can occur during the leap. For non-critical reactions the approximate explicit tau-leaping method [42] is used. Because critical reactions are only allowed to fire once, the chance of a negative result in the reactant population becomes much smaller.

In the original tau-leaping algorithm, the Poisson distribution is unbounded which can lead to overfiring in a reaction channel. When a reaction channel overfires, it may lead to negative numbers in the species populations, which is physically unrealistic and therefore undesirable. When a reactant species has a small population, overfiring is more likely to occur. One solution to this problem was to use a bounded binomial distribution [Tian and Burrage [112] and independently Chatterjee et al. [17]] in place of the Poisson distribution. The binomial leaping method is less accurate than the (Poisson) tau-leaping strategy. However, more recent work [Cao et al. [13, 14]] revised the Poisson tau-leaping strategy to account for negativity.

4.2 Stepsize Selection for Explicit Tau-Leaping

As was discussed in Section 2.7.1, the exact stochastic simulation algorithms are computationally intensive when applied to stiff biochemical systems. One approach to reducing the computational burden is to employ the tau-leaping method, proposed by Gillespie [42] which was discussed further in Section 2.7.3.

In practice, it is more accurate and easier to implement a leap condition bounding the relative change in molecular amounts rather than the relative change in propensities [14]. The most widely used version of the leap condition [14] demands that τ is small enough such that the abundance of each reactant population X_i satisfies (approximately)

$$|X_i(t+\tau) - x_i| \leq \max\{\varepsilon x_i / g_i, 1\} \tag{4.1}$$

where $X_i(t) = x_i$. Here ε is a user-selected tolerance, and the factor g_i is the highest order at which species S_i appears as a reactant (with some modification for reactions in which multiple molecules of S_i occur as reactants, see [14] for details).

As detailed by Cao et al. [14], an efficient implementation of this leap condition begins with a user-specified control parameter, n_c, which characterizes a threshold below which reactant populations are in danger of dropping below zero. It is recommended that $n_c \in [2, 20]$. A reaction is called *critical* if n_c firings of the reaction would result in the population of one of its reactants dropping to zero. Critical reactions are then constrained to fire at most once during a leap. The condition (4.1) is then applied to the non-critical reactions, but is implemented in terms of the mean and standard deviation of the population changes. Specifically, with J_{ncr} as the set of indices of non-critical reaction channels, for each reactant species X_i, we define the auxiliary quantities

$$\hat{\mu}_i(\mathbf{x}) = \sum_{j \in J_{ncr}} \nu_{ij} a_j(\mathbf{x}) \ , \qquad \hat{\sigma}_i^2(\mathbf{x}) = \sum_{j \in J_{ncr}} \nu_{ij}^2 a_j(\mathbf{x}). \tag{4.2}$$

The bound on the step then takes the form (for details see [14] and the CTL algorithm

below):

$$\tau = \min_i \left\{ \frac{\max\{\varepsilon x_i/g_i, 1\}}{|\hat{\mu}_i(\mathbf{x})|}, \frac{\max\{\varepsilon x_i/g_i, 1\}^2}{\hat{\sigma}_i^2(\mathbf{x})} \right\}. \tag{4.3}$$

We make use of this tau-leaping strategy in our algorithm for estimating sensitivities.

We present below a new algorithm for approximating the sensitivity utilizing an adaptive tau-leaping scheme to generate tightly coupled perturbed and nominal trajectories. Finite difference approximations are applied.

4.3 Coupled Tau-Leaping (CTL)

In this section, we present our new algorithm (Morshed et al. [83]) for estimating local sensitivities that is computationally efficient when applied to moderately stiff stochastic biochemical systems. In our sensitivity method, the coupling of the nominal and perturbed processes is similar to that employed by the CFD method [2]. However, our approach couples paths that are obtained with the approximate tau-leaping strategy, whereas the CFD method couples paths that are in exact agreement with the Chemical Master Equation. A similar coupling was first introduced in [69]; it was later used in [3] in the context of multi-level Monte Carlo simulations for biochemical kinetic systems.

In the CTL method [83] the perturbed and nominal sample paths are coupled according

to (compare with (3.11))

$$X^c(t+\tau) = \mathbf{x}^c + \sum_{j=1}^{M} \nu_j \left[P_{1,j}(m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h})\tau) + P_{2,j}((a_j^c(\mathbf{x}^c) - m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h}))\tau)\right]$$
$$X^{c+h}(t+\tau) = \mathbf{x}^{c+h} + \sum_{j=1}^{M} \nu_j \left[P_{1,j}(m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h})\tau) + P_{3,j}((a_j^{c+h}(\mathbf{x}^{c+h}) - m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h}))\tau)\right] \tag{4.4}$$

where $X^{c+h}(t) = \mathbf{x}^{c+h}$ and $X^c(t) = \mathbf{x}^c$. Here $m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h}) = \min\left\{a_j^c(\mathbf{x}^c), a_j^{c+h}(\mathbf{x}^{c+h})\right\}$ and $P_{1,j}$, $P_{2,j}$ and $P_{3,j}$ are independent Poisson random variables.

In the CTL algorithm [83] we make use of the stepsize selection strategy developed by Cao et al. [14] for the tau-leaping method, which proceeds in two steps: candidate tau-leaps are determined separately for the critical and non-critical reactions, and the minimum is selected. We apply the tau-selection procedure to both the nominal and the perturbed trajectories.

CTL Algorithm

1. **Specify simulation parameters:** set the values for the tolerance ε, the critical threshold n_c and the final time T.

2. **Initialize sample paths:** for each trial, initialize the time $t \leftarrow 0$ and the states $X^{c+h} \leftarrow \mathbf{x}$ and $X^c \leftarrow \mathbf{x}$.

3. **Loop:** While $t < T$ do (a)–(g)

   (a) **Compute the propensity functions:** $a_j^{c+h}(X^{c+h})$ and $a_j^c(X^c)$ for each $j = 1, \ldots, M$.

(b) **Determine the set of critical reactions for the nominal and perturbed trajectories:** on the nominal trajectory, for each reaction R_j with propensity $a_j^c(X^c) > 0$, determine

$$L_j^c = \min_{i=1,\ldots,N;v_{ij}<0} \left\lfloor \frac{X_i^c}{|\nu_{ij}|} \right\rfloor \qquad \lfloor\cdot\rfloor \text{ is the floor function (greatest integer less than).}$$

On the perturbed trajectory, for each reaction R_j with propensity $a_j^{c+h}(X^{c+h}) > 0$, determine

$$L_j^{c+h} = \min_{i=1,\ldots,N;v_{ij}<0} \left\lfloor \frac{X_i^{c+h}}{|\nu_{ij}|} \right\rfloor$$

and set $J_{ncr} = \{j : L_j^c \geq n_c \text{ and } L_j^{c+h} \geq n_c\}$, the set of non-critical reaction indexes. That is, $\min(L_j^c, L_j^{c+h})$ is the maximum number of times that R_j can occur without exhausting one of its reactants on either the perturbed or the nominal trajectory.

(c) **Determine candidates leap size, τ_1^c and $\tau_1^{(c+h)}$, for the non-critical reactions:** Compute first τ candidates, $\tau_1^{(c)}$ and $\tau_1^{(c+h)}$ (one candidate for the nominal and one for the perturbed trajectory). If there are no non-critical reactions ($J_{ncr} = \emptyset$), set $\tau_1 = \infty$. Otherwise, determine the set of indices I_{ncr} of species that are reactants of non-critical reactions. For each $i \in I_{ncr}$ and on each of the nominal and perturbed trajectories:

i. Let ψ_i be the highest order at which S_i appears as reactant in a non-critical reaction.

ii. Determine the factors g_i as follows

A. If $\psi_i = 1$, then set $g_i = 1$

B. If $\psi_i = 2$, then set $g_i = 2$, unless the left hand side of the reaction equation is $S_i + S_i$, in which case set $g_i = \left(2 + \frac{1}{x_1 - 1}\right)$.

C. If $\psi_i = 3$, then set $g_i = 3$, unless the left hand side of the reaction equation is of the form $S_i + S_i + S_j$, in which case set $g_i = \frac{3}{2}\left(2 + \frac{1}{x_i - 1}\right)$, or, alternatively, has the form $S_i + S_i + S_i$, in which case set $g_i = \left(3 + \frac{1}{x_i - 1} + \frac{2}{x_i - 2}\right)$.

iii. Evaluate the auxiliary quantities $\hat{\mu}_i(\mathbf{x})$ and $\hat{\sigma}_i^2(\mathbf{x})$ according to (4.2) and $\tau_1^{(c)}$ and $\tau_1^{(c+h)}$ using (4.3).

(d) **Determine candidate leap sizes for the nominal and perturbed trajectories, $\tau_2^{(c)}$ and $\tau_2^{(c+h)}$, for the critical reactions:** set $a_0^{cr,(c)}(X^c)$ and $a_0^{cr,(c+h)}(X^{c+h})$ to be the sum of the critical reaction propensities for the nominal and perturbed trajectories, take $\xi_1^{(c)}$ and $\xi_1^{(c+h)}$ samples from the uniform distribution on $[0, 1]$, and compute $\tau_2^{(c)}$ and $\tau_2^{(c+h)}$ as

$$\tau_2^{(c)} = (1/a_0^{cr,(c)}(X^c)) \ln(1/\xi_1^{(c)}),$$

$$\tau_2^{(c+h)} = (1/a_0^{cr,(c+h)}(X^{c+h})) \ln(1/\xi_1^{(c+h)}).$$

(e) **Select leap size and determine reaction extents k_j:** $\tau_1 = \min\{\tau_1^{(c)}, \tau_1^{(c+h)}\}$, $\tau_2 = \min\{\tau_2^{(c)}, \tau_2^{(c+h)}\}$.

i. If $\tau_1^{(c)} < \tau_2^{(c)}$ and $\tau_1^{(c+h)} < \tau_2^{(c+h)}$, no critical reaction occurs. Set $\tau = \tau_1$ and $k_j^c = k_j^{c+h} = 0$ for all critical reactions. For $j \in J_{ncr}$, compute $m_j =$

$\min(a_j^c(X^c), a_j^{c+h}(X^{c+h}))$.

A. Generate the samples from Poisson distributions

$$\begin{aligned} P_{1,j} &= \text{Poisson}(m_j\tau), \\ P_{2,j} &= \text{Poisson}((a_j^c(X^c) - m_j)\tau), \\ P_{3,j} &= \text{Poisson}((a_j^{c+h}(X^{c+h}) - m_j)\tau). \end{aligned} \tag{4.5}$$

B. Set the reaction extents

$$k_j^c = P_{1,j} + P_{2,j}, \quad k_j^{c+h} = P_{1,j} + P_{3,j}. \tag{4.6}$$

ii. else if $\tau_2^{(c)} < \tau_2^{(c+h)}$, a single critical reaction occurs on the nominal trajectory. Generate a sample, ξ_2, from the uniform distribution over $[0, 1]$. Let j_{cr} be the smallest integer for which $\sum_{k=1}^{j} a_j^c(X^c) > \xi_2 a_0^{cr,(c)}$. Set $\tau = \tau_2$, $k_{j_{cr}}^c = 1$, $k_{j_{cr}}^{c+h} = 0$. For all other critical reactions set $k_j^c = k_j^{c+h} = 0$ and for non-critical reactions compute (4.6) with (4.5).

iii. else if $\tau_2^{(c+h)} < \tau_2^{(c)}$, a single critical reaction occurs on the perturbed trajectory. Generate a sample, ξ_2, from the uniform distribution over $[0, 1]$. Let j_{cr} be the smallest integer for which $\sum_{k=1}^{j} a_j^{c+h}(X^{(c+h)}) > \xi_2 a_0^{cr,(c+h)}$. Set $\tau = \tau_2$, $k_{j_{cr}}^c = 0$, $k_{j_{cr}}^{c+h} = 1$. For all other critical reactions set $k_j^c = k_j^{c+h} = 0$ and for non-critical reactions compute (4.6) with (4.5).

iv. else a single critical reaction occurs on both the nominal and the perturbed trajectories. Set $\tau = \tau_2$. Generate a sample, ξ_2 from the uniform distribution over $[0, 1]$. Let j_{cr} be the smallest integer for which

$\sum_{k=1}^{j} a_j^c(X^{(c)}) > \xi_2 a_0^{cr,(c)}$. Set $\tau = \tau_2$, $k_{j_{cr}}^c = k_{j_{cr}}^{c+h} = 1$. For all other critical reactions set $k_j^c = k_j^{c+h} = 0$ and for non-critical reactions compute (4.6) with (4.5).

(f) **Implement the step:** update the time $t \leftarrow t + \tau$ and the state values

$$X^c \leftarrow X^c + \sum_{j=1}^{M} k_j^c \nu_j,$$
$$X^{c+h} \leftarrow X^{c+h} + \sum_{j=1}^{M} k_j^{c+h} \nu_j.$$

(g) Estimate the sensitivity with respect to c, on the sample path, as $Z = (f(X^{c+h}) - f(X^c))/h$ at time t.

4.4 Numerical Results

In this section, we illustrate the efficiency of the coupled tau-leaping (CTL) method [83], over existing finite-difference methods. We benchmark against three biochemical reaction models with mass action kinetics and simple dynamics. Applications to complex dynamics are of definite interest, but because their interpretation is the subject on ongoing research (e.g. bistability [24], sustained oscillations [53], quasi-steady state approximations [110]), they do not present ideal subjects for unambiguous comparisons.

For each analysis, an ensemble of 10000 pairs of sample paths was simulated. To demonstrate the proposed method's performance, for each model we completed analyses over a range of tolerance values ε. We present the values of the sensitivity estimators in each case, and report the relative computational times. We find that the accuracy of the pro-

posed CTL method is comparable to the CFD method (especially for small ε), and that for systems that are at least moderately stiff, the CTL method is considerably more computationally efficient.

4.4.1 Two-step Closed Reaction Chain Model

Our preliminary analysis is of a simple two-step closed reaction chain (Figure 4.1). The reactions are listed in Table 4.1, along with the reaction propensities and a nominal set of parameter values. Here X_i is the molecular abundance of species S_i. In this model, reactions R_1 and R_2 are fast, while R_3 and R_4 are slow. The propensities of the fast and slow reactions are separated by four orders of magnitude, resulting in significant stiffness.

Table 4.1: Two-step closed reaction chain

R_j	Reaction	Propensity	Nominal rate constant
R_1	$S_1 \xrightarrow{C_1} S_2$	$a_1 = C_1 X_1$	$C_1 = 800$
R_2	$S_2 \xrightarrow{C_2} S_1$	$a_2 = C_2 X_2$	$C_2 = 3200$
R_3	$S_2 \xrightarrow{C_3} S_3$	$a_3 = C_3 X_2$	$C_3 = 0.1$
R_4	$S_3 \xrightarrow{C_4} S_2$	$a_4 = C_4 X_3$	$C_4 = 1$

We simulated the system with initial conditions $(X_1(0), X_2(0), X_3(0)) = (2000, 1000, 100)$ and the kinetic parameters in Table 4.1, on the time-interval $[0, 0.1]$. The Figure 4.2(a) shows the mean abundance of species S_1 for the proposed tau-leaping algorithm (over a

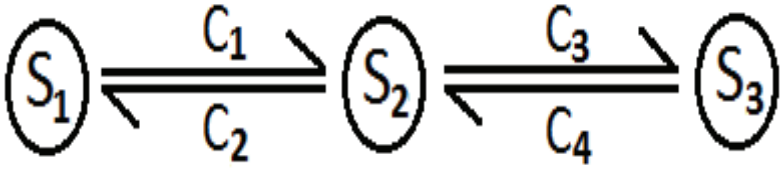


Figure 4.1: Two-step closed reaction chain.

range of tolerances ε) and for the CFD method. The standard deviation of the molecular count of $S_1(t)$ is shown in Figure 4.2(b). (The plots in Figure 4.2(a-b) are equivalent to ensembles generated from independent tau-leaping [14] and next-reaction method [35] simulations, respectively.) As expected, the performance of the tau-leap method deteriorates as the tolerance takes higher values. Figure 4.2(c-d) show the sensitivity of the molecular count of species S_1 with respect to the parameter C_1, with perturbation $h = 1$ (i.e. 0.125% of the nominal parameter value). The insets show the mean (panel (c)) and variance (panel (d)) of the estimator of the sensitivity from the CRN, CRP, CFD, and the CTL ($\varepsilon = 0.03$) methods. As demonstrated in [109], the CFD method provides an estimator with considerably lower variance than the CRN and CRP methods; the proposed CTL method is comparable. The main panels show a comparison of the performance of the CFD and the CTL methods over a range of tolerances ε.

In [2] it was shown that for the CFD method, the standard deviation of the estimator for sensitivity depends on the perturbation size h as $O(h^{-1})$. We investigated this dependence for the simulation of the two-step reaction chain described above. Figure 4.3 shows numerical observations for the CRN, CRP, CFD, and CTL (for a range of tolerances ε). As in Figure 4.2(d), the estimator calculated from the proposed tau-leaping method has variance very close to that of the CFD method, whereas the CRP and the CRN estimators

exhibit much higher variability. We observe that in this case the variance of our adaptive tau-leaping sensitivity estimator appears to be $O(h^{-1})$. The increase in estimator variance with decreased perturbation size h leads to a trade-off: small perturbation size results in large estimator variance, but large perturbation size leads to finite differences that poorly approximate the derivative. We manually explored this trade-off in selecting a value for the parameter h in the examples we considered, but we did not complete an analysis of optimal stepsize choice.

Table 4.2 indicates the relative timing of sample path generation for the simulations in Figure 4.2 (labelled as Simulation I). (Absolute timings will, of course, vary by machine specifications. As an example, in this case, computation of the CTL over 100 trajectory pairs with $\varepsilon = 0.06$ took 25.55 seconds on a Macbook Pro with a single 1.3GHz Intel Core i5 processor the equivalent calculation of the CFD took 2173.74 seconds.) We compared timing for the CFD and CTL methods. We did not carry out a comparison with the CRN or CRP methods, as they produce estimators with considerably higher variability. As expected, the computational efficiency of the CTL depends on the choice of tolerance ε, but, for this analysis, the CTL provides a significant savings in computational effort compared with the CFD (while providing comparable results; Figure 4.2(d)). In Simulation I, the CTL algorithm encountered very few critical reactions. To assess the algorithm's performance when molecule counts are frequently small, we ran another ensemble, with initial condition $X(0) = (3000, 100, 12)$ and rate parameters $(C_1, C_2, C_3, C_4) = (800, 3200, 0.1, 5)$, labelled as Simulation II in Table 4.2. In this case, about 10% of steps involved molecular counts below the critical threshold of $n_c = 10$. The efficiency of the method was thus reduced. However, as shown, the CTL method is still considerably more efficient than CFD

for this simulation.

Table 4.2: Closed reaction chain model: efficiency gain of CTL over the CFD, for approximating the sensitivity of the the abundance of species S_1 with respect to C_1, for $h = 1$. The time interval is $[0, 0.1]$.

Method	Tolerance	Efficiency gain Simulation I	Efficiency gain Simulation II
CTL	$\varepsilon = 0.06$	81.87	24.57
CTL	$\varepsilon = 0.05$	65.89	18.79
CTL	$\varepsilon = 0.04$	44.31	13.04
CTL	$\varepsilon = 0.03$	25.12	7.72
CFD	–	1	1

4.4.2 Oregonator Model

The Oregonator model [31] (Figure 4.4) describes a chemical reaction network capable of exhibiting sustained oscillations. Local sensitivity analysis of periodic behaviour is generally confounded by the fact that phase shifts cause sensitivity coefficients to diverge [53, 97]. Rather than address this issue here, we instead chose nominal rate constants for which the model exhibits damped oscillations. The reactions, the propensities and a nominal set of values for the rate constants are given in Table 4.3. With these parameters, the propensities of the fast reactions (R_2, R_3 and R_5) are two orders of magnitude faster than those of the slow reactions (R_1 and R_4).

Table 4.3: Oregonator model

R_j	Reaction	Propensity	Nominal rate constant
R_1	$S_2 \xrightarrow{C_1} S_1$	$a_1 = C_1 X_2$	$C_1 = 5$
R_2	$S_1 + S_2 \xrightarrow{C_2} \oslash$	$a_2 = C_2 X_1 X_2$	$C_2 = 0.0250$
R_3	$S_1 \xrightarrow{C_3} 2S_1 + S_3$	$a_3 = C_3 X_1$	$C_3 = 130$
R_4	$2S_1 \xrightarrow{C_4} \oslash$	$a_4 = C_4 X_1 (X_1 - 1)/2$	$C_4 = 1.6 \times 10^{-4}$
R_5	$S_3 \xrightarrow{C_5} S_2$	$a_5 = C_5 X_3$	$C_5 = 130$

We ran simulations from initial condition $(X_1(0), X_2(0), X_3(0)) = (5000, 400, 800)$ over the time interval $[0, 2]$, and addressed the sensitivity of the molecular count of species S_1 with respect to parameter C_1, with a perturbation size of $h = 0.01$ (i.e. 0.2% of the nominal value).

Figure 4.5 (a-b) show the mean and standard deviation of the molecular count of species S_1 as simulated by the next reaction method and adaptive tau-leaping scheme (over a range of tolerances ε). The behaviours are similar, with the tau-leaping approach showing less accuracy as higher tolerance values are chosen. The mean and standard deviation of the sensitivity estimators generated by the CFD and CTL methods are shown in Figure 4.5(c-d). Again, the results are similar, with accuracy dependent on tolerance threshold ε. The relative timings for this analysis, shown in Table 4.4, show considerable efficiencies (up to 216-fold) for the tau-leap approach.

Table 4.4: Oregonator model: efficiency gain of CTL over CFD, for approximating the sensitivity of the abundance of species S_1 with respect to C_1, for $h = 0.01$. The time interval is $[0, 2]$.

Method	Tolerance	Efficiency gain
CTL	$\varepsilon = 0.05$	216.17
CTL	$\varepsilon = 0.04$	164.38
CTL	$\varepsilon = 0.03$	109.89
CTL	$\varepsilon = 0.02$	56.29
CFD	–	1

4.4.3 Gene Regulatory Network Model

As a final illustration, we consider a multi-scale reaction network (Figure 4.6), which has been used previously to benchmark stochastic methods for biochemical systems modelling [52, 76]. (This model represents a gene regulatory network capable of exhibiting bistability [11], but we do not explore that aspect of the dynamics here.) The reactions, propensities and rate constants are shown in Table 4.5. For this parametrization, four orders of magnitude separate the propensities of the fastest reactions, R_3 and R_4, from those of the slowest reactions, R_1 and R_{12}.

We simulated the model from initial condition $(X_1(0), X_2(0), X_3(0), X_4(0), X_5(0), X_6(0), X_7(0), X_8(0)) = (800, 800, 500, 500, 400, 500, 400, 500)$, on the time interval $[0, 0.1]$, and addressed the sensitivity of the molecular count of species S_2 with respect to parameter

Table 4.5: Gene regulatory network model

R_j	Reaction	Propensity	Nominal rate constant
R_1	$S_3 \xrightarrow{C_1} S_3 + S_1$	$a_1 = C_1 X_3$	$C_1 = 0.16$
R_2	$S_4 \xrightarrow{C_2} S_4 + S_2$	$a_2 = C_2 X_4$	$C_2 = 0.16$
R_3	$S_3 + S_2 \xrightarrow{C_3} S_5$	$a_3 = C_3 X_2 X_3$	$C_3 = 5$
R_4	$S_5 \xrightarrow{C_4} S_3 + S_2$	$a_4 = C_4 X_5$	$C_4 = 3000$
R_5	$S_5 + S_2 \xrightarrow{C_5} S_6$	$a_5 = C_5 X_2 X_5$	$C_5 = 2.5$
R_6	$S_6 \xrightarrow{C_6} S_5 + S_2$	$a_6 = C_6 X_6$	$C_6 = 1600$
R_7	$S_1 \xrightarrow{C_7} \oslash$	$a_7 = C_7 X_1$	$C_7 = 0.1$
R_8	$S_4 + S_1 \xrightarrow{C_8} S_7$	$a_8 = C_8 X_1 X_4$	$C_8 = 2$
R_9	$S_7 \xrightarrow{C_9} S_4 + S_1$	$a_9 = C_9 X_7$	$C_9 = 3000$
R_{10}	$S_7 + S_1 \xrightarrow{C_{10}} S_8$	$a_{10} = C_{10} X_1 X_7$	$C_{10} = 2.5$
R_{11}	$S_8 \xrightarrow{C_{11}} S_7 + S_1$	$a_{11} = C_{11} X_8$	$C_{11} = 1600$
R_{12}	$S_2 \xrightarrow{C_{12}} \oslash$	$a_{12} = C_{12} X_2$	$C_{12} = 0.1$

C_3, with a perturbation size of $h = 0.01$ (i.e. 0.2% of the nominal value).

Figure 4.7 (a-b) shows the mean and standard deviation of the molecular count of species

S_2 as simulated by the next reaction method and adaptive tau-leap technique (over a range of tolerances ε). The results are similar to the previous examples. Figure 4.7 panels (c-d) show the mean and standard deviation of the sensitivity estimators generated by the CFD and CTL. As before, accuracy is dependent on the tolerance threshold ε. The speed-up of the tau-leaping approach (Table 4.6) is reduced compared with the previous examples, but is still considerable.

Table 4.6: Gene regulatory network model: efficiency gain of CTL over CFD, for approximating the sensitivity of the abundance of species S_2 with respect to C_3, for $h = 0.01$. The time interval is $[0, 0.1]$

Method	Tolerance	Efficiency gain
CTL	$\varepsilon = 0.15$	16.43
CTL	$\varepsilon = 0.10$	9.28
CTL	$\varepsilon = 0.05$	2.57
CFD	–	1

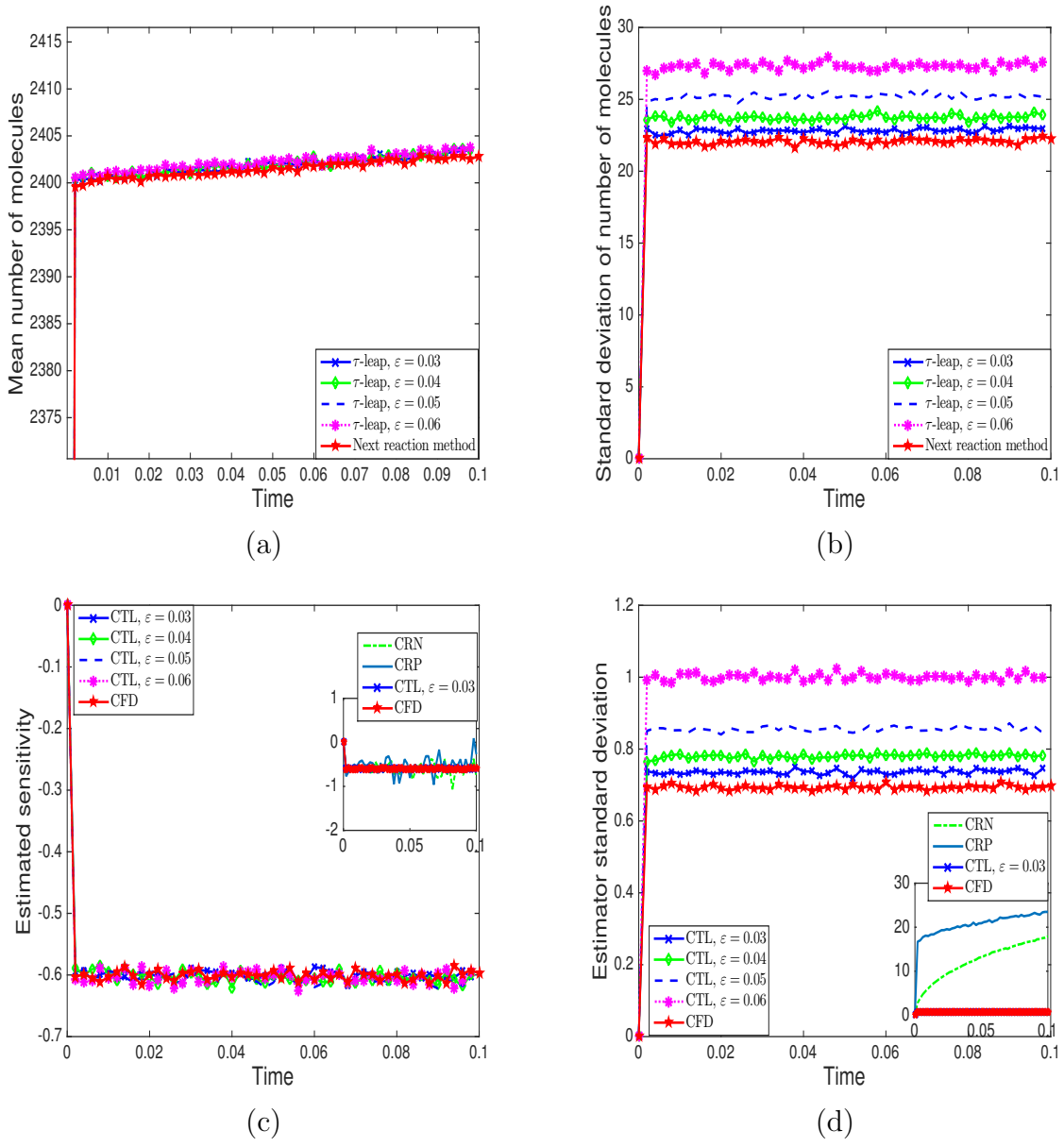


Figure 4.2: Closed reaction chain model. Ensembles of 10000 sample paths were generated on the time-interval $[0, 0.1]$, starting from initial condition $(X_1(0), X_2(0), X_3(0))$=(2000, 1000, 100) with parameters as in Table 4.1. (a-b) Mean and standard deviation of the molecular count for species S_1, determined by the next reaction method and the adaptive tau-leaping algorithm with various tolerances ε. (c-d) Mean and standard deviation of the finite-difference estimators of the sensitivity of the abundance of S_1 to the parameter C_1, calculated by the CRN, CRP, CFD, and CTL methods.

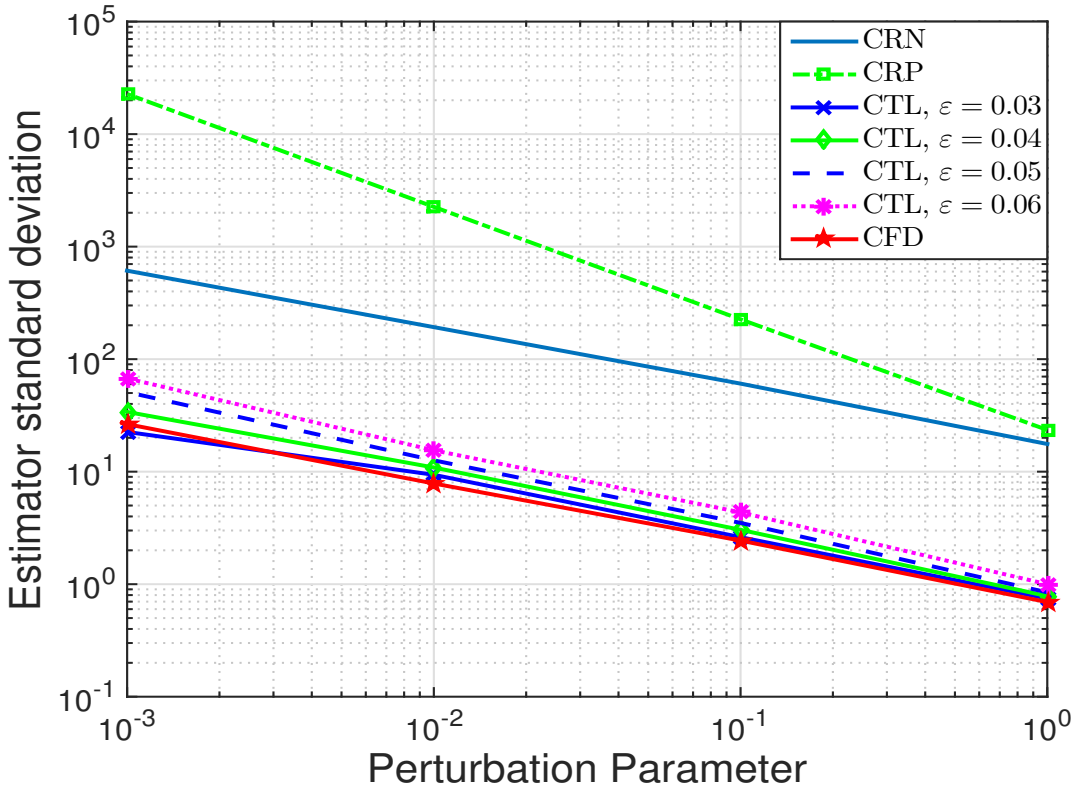


Figure 4.3: Dependence of variability of the sensitivity estimator on the perturbation size h for the two-step reaction chain. The CFD and CTL estimators exhibit comparable variability with an $O(h^{-1})$ dependence on the perturbation size. The estimators generated by the CRN and CRP methods are considerably more variable.

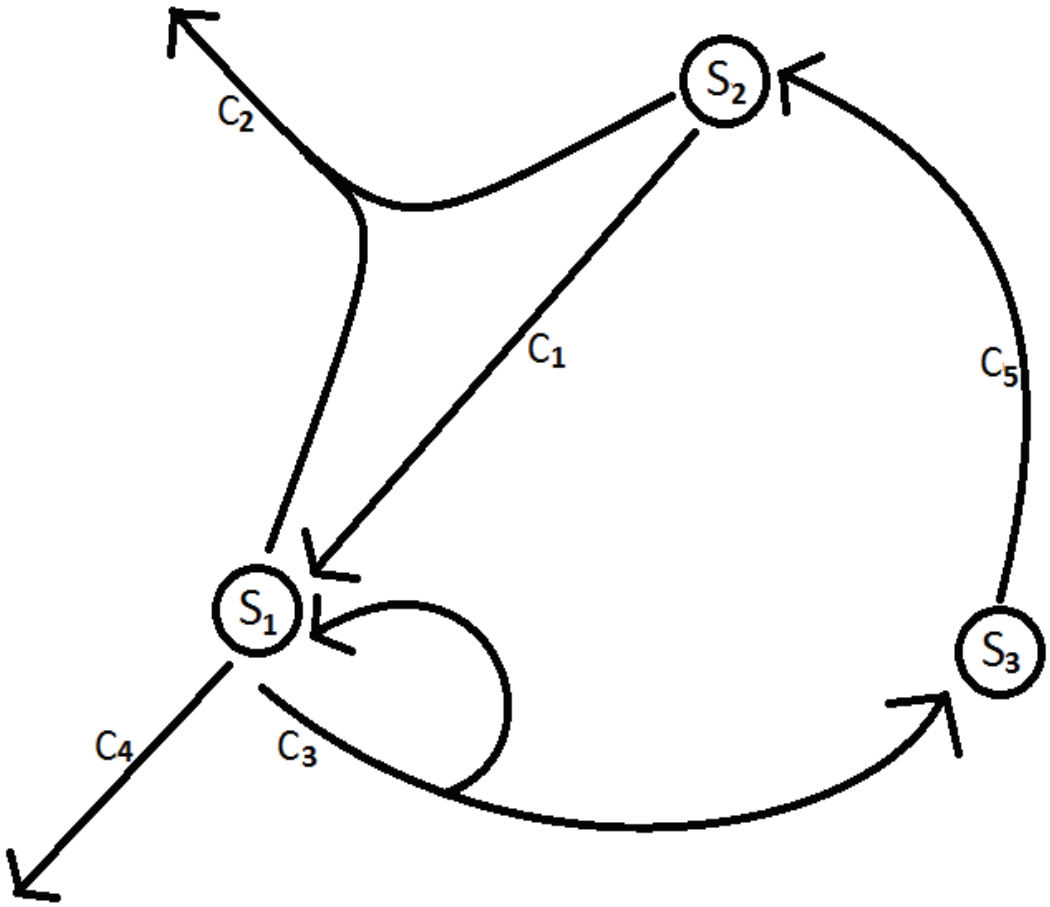


Figure 4.4: Oregonator reaction network.

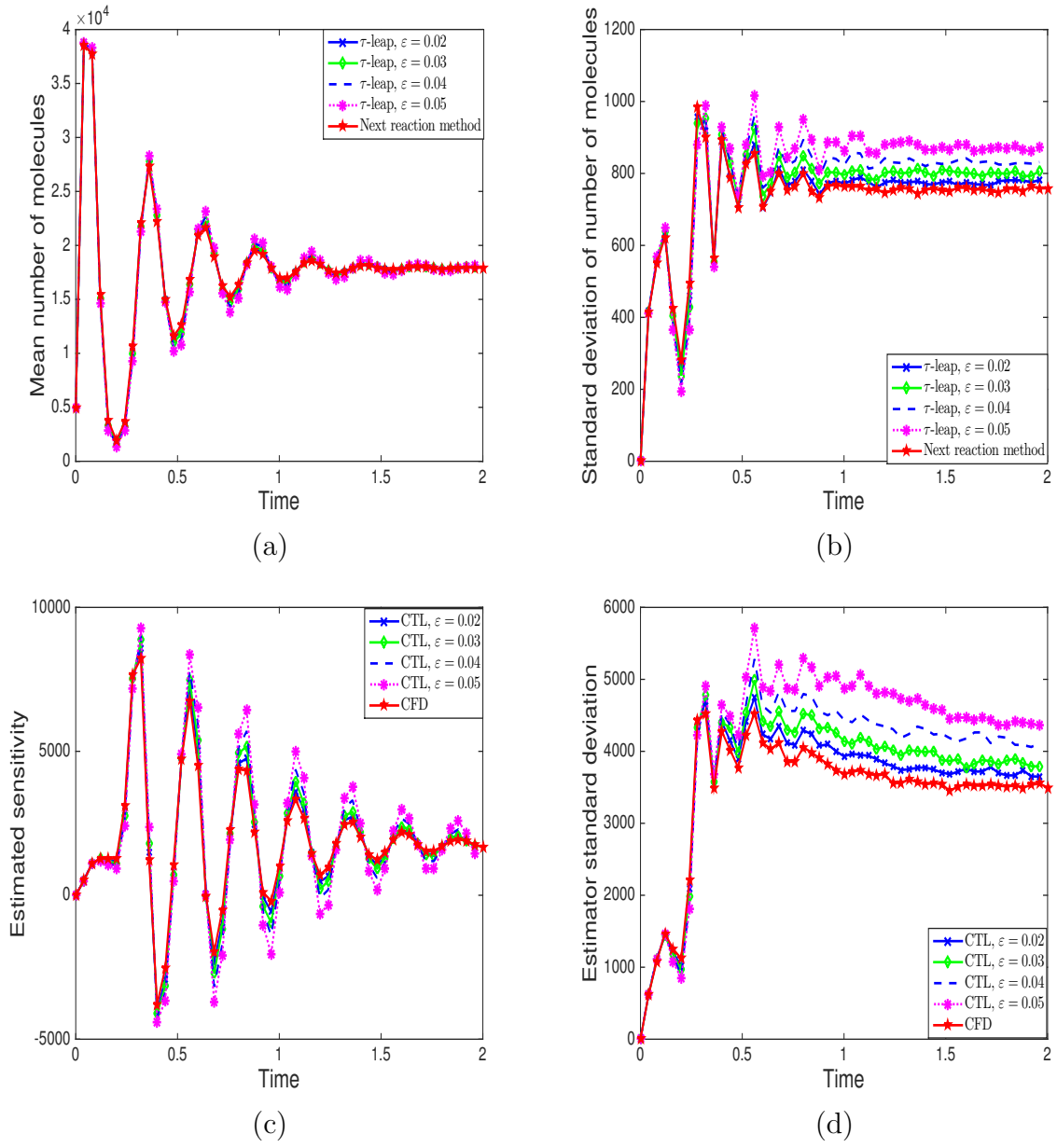


Figure 4.5: Oregonator model. Ensembles of 10000 sample paths were generated on the time-interval $[0, 2]$, starting from initial condition $(X_1(0), X_2(0), X_3(0)) = (5000, 400, 800)$ with parameters as in Table 4.3. (a-b) Mean and standard deviation of the molecular count of S_1, determined by the next reaction method and the adaptive tau-leaping scheme with various tolerances ε. (c-d) Mean and standard deviation of the finite-difference estimators of the sensitivity of the abundance of S_1 with respect to the parameter C_1, determined by the CFD and the CTL methods.

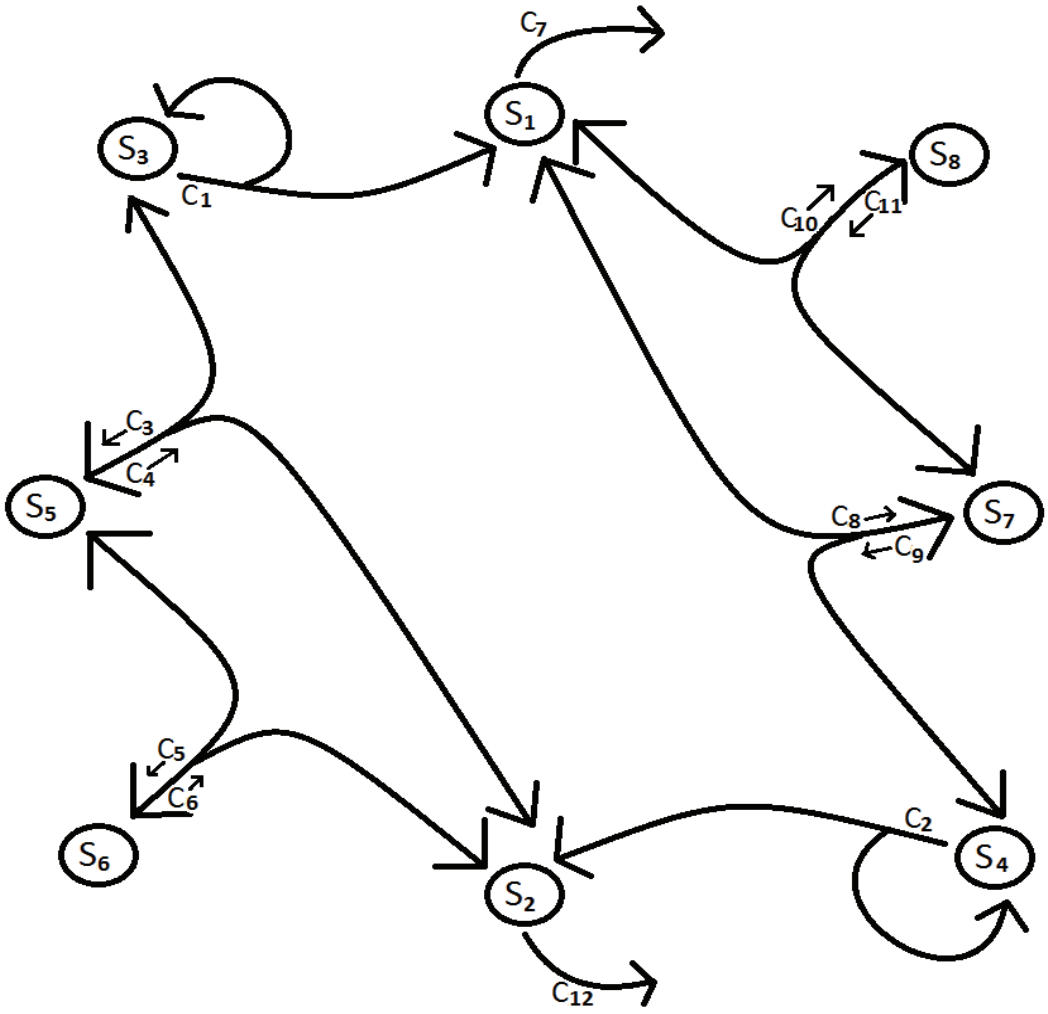


Figure 4.6: Gene regulatory reaction scheme diagram.

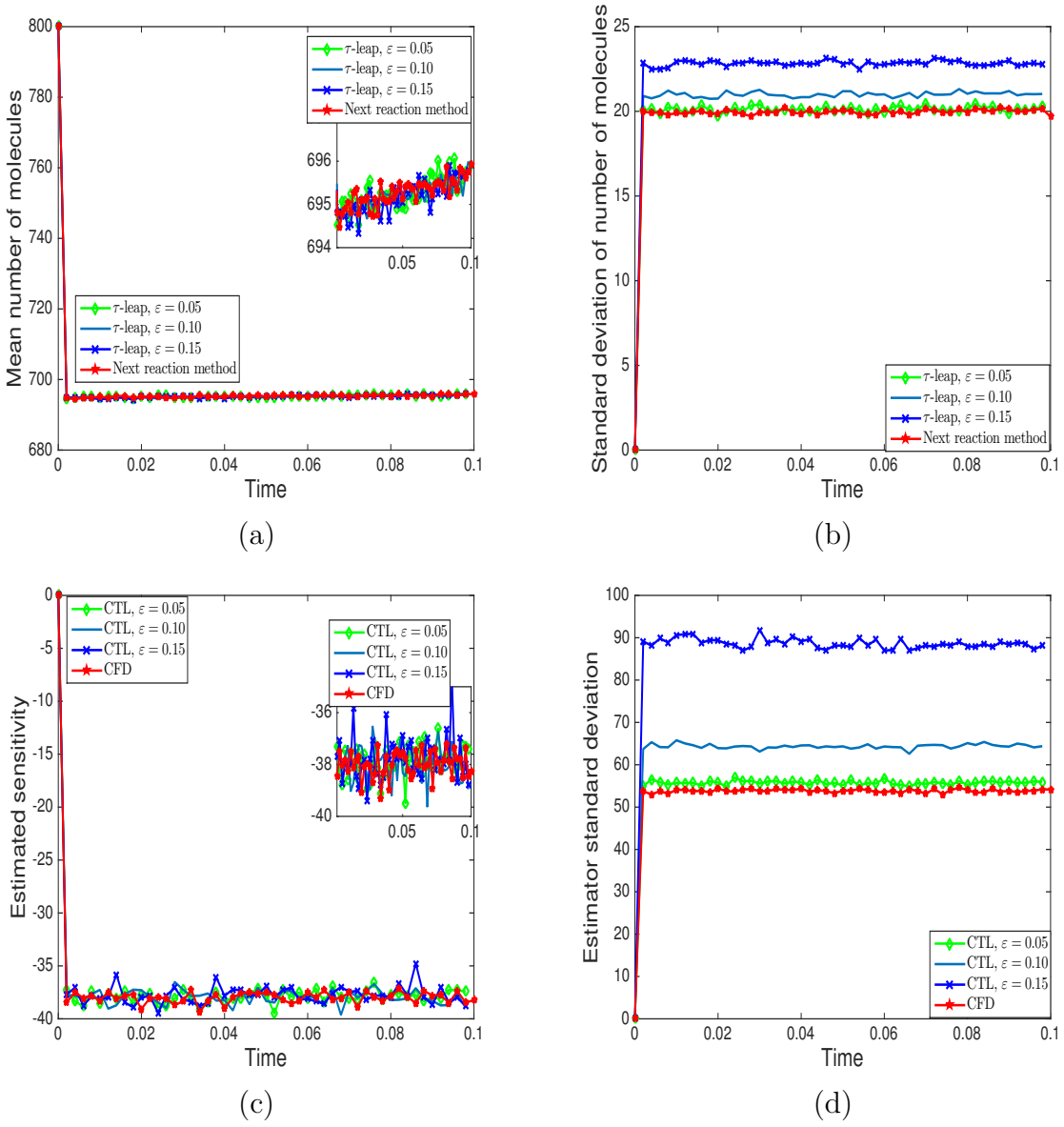


Figure 4.7: Gene regulatory network model. Ensembles of 10000 sample paths were generated on the time interval $[0, 0.1]$, starting from initial condition $(X_1(0), X_2(0), X_3(0), X_4(0), X_5(0), X_6(0), X_7(0), X_8(0))$=(800, 800, 500, 500, 400, 500, 400, 500) with parameters as in Table 4.5. (a-b) Mean and standard deviation of the molecular count of S_2, determined by the next reaction method and the adaptive tau-leaping strategy with various tolerances ε. (c-d) Mean and standard deviation of the finite-difference estimators of the sensitivity of the abundance of S_2 to the parameter C_3, calculated by the CFD, and the CTL.

Chapter 5

Adaptive Coupled Implicit Tau-Leaping Method

5.1 Introduction

The materials in this Chapter are reproduced directly from the jointly-authored publication by Morshed, Ingalls and Ilie [84]. In this chapter we propose a novel strategy, the Coupled Implicit Tau-leaping (CIT) method [84], for estimating local sensitivities. This method is computationally efficient when applied to stiff to very stiff stochastic biochemical systems.

Recall that stiff systems involve quickly changing dynamics, where fast and slow time scales are well separated [15, 93] with the fastest mode being stable as was discussed in Section 2.10. The implicit-tau leaping scheme [93] was designed for such stiff biochemical systems. When used for solving stiff systems, the explicit-tau leaping method may be-

come unstable unless τ is chosen to be small enough such that the accuracy requirements associated with the fast dynamics are satisfied.

Evaluation of the propensity function, a_j at the current known state x causes the tau-leaping method to be an explicit method which was discussed in depth in Section 2.7.3. The time steps for the explicit tau-leaping strategy are limited to the fastest mode, as such it is not suitable for stiff biochemical systems. The tau-leaping strategy for the CME shows a similar instability due to large time steps as the explicit Euler method applied to ordinary differential equations. This is expected, since the tau-leaping strategy is a generalization of the explicit Euler scheme to discrete stochastic well-stirred biochemical systems. To address this issue, Rathinam et al. [93] developed the implicit tau-leaping method which overcomes the poor stability of the explicit strategy and allows larger time steps. The implicit tau-leaping strategy produces an accurate numerical solution for the slow variables in the system, with larger time steps sizes than the explicit tau-leaping scheme for stiff discrete stochastic systems. Furthermore, the mean for the fast variable on the slow manifold is accurate.

5.2 Implicit Tau-Leaping

Recall, in Section 2.7.1, we discussed that exact Monte Carlo simulation algorithms [35, 41, 43] for the Chemical Master Equations (CME) are often computationally expensive on problems of practical interest. An approximate technique which reduces the computational cost of solving the CME is the tau-leaping method, due to Gillespie [42]. This technique was discussed in Section 2.7.3.

Many biochemical systems arising in applications are stiff, displaying both slow and fast dynamics, with the fast modes being stable. However, the explicit tau-leaping strategy is impractical for stiff systems, as its time-step is limited to the fastest mode. To deal with this challenge, Rathinam et al. [93] proposed the *implicit tau-leaping method*. The implicit tau-leaping technique overcomes the stability issue of the explicit strategy, allowing larger steps in time. Consequently, for stiff stochastic biochemical systems, it is more efficient than the explicit method while maintaining a similar accuracy. In fact, the scheme is semi-implicit, being implicit only in the mean part of each term $P_j(a_j, \tau)$, i.e. $a_j\tau$. If $X(t) = \mathbf{x}$, the implicit tau-leaping method updates the system state as

$$\mathbf{X}(t+\tau) = \mathbf{x} + \sum_{j=1}^{M} \nu_j a_j(\mathbf{X}(t+\tau))\tau + \sum_{j=1}^{M} \nu_j \left[P_j(a_j(\mathbf{x}), \tau) - a_j(\mathbf{x})\tau\right]. \tag{5.1}$$

5.3 Stepsize Selection for Implicit Tau-Leaping

Reversible reactions are those that can occur going from reactants to products and vice versa. They can reach an equilibrium between reactants and products. When this occurs for some reversible reactions while the rest of the system is still undergoing significant variation, the system is said to be in *partial equilibrium*. Partial equilibrium occurs when the forward and backward propensities of the reversible reaction are approximately equal: their difference should be much smaller than the propensities themselves. More precisely, if the propensities of the reversible reactions are denoted by $a_+(\mathbf{x})$ and $a_-(\mathbf{x})$, the partial

equilibrium condition (Cao et al. [15]) is

$$|a_+(\mathbf{x}) - a_-(\mathbf{x})| \leq \delta \min\{a_+(\mathbf{x}), a_-(\mathbf{x})\}, \tag{5.2}$$

for some small quantity $\delta > 0$. (In the implementations below we used $\delta = 0.05$.)

We make use of the step-size selection strategy introduced by Cao et al. [15]. For those reactions that are not in partial equilibrium, we demand that the mean and variance of each reactant population X_i should satisfy

$$|X_i(t+\tau) - x_i| \leq \max\{\varepsilon x_i/g_i, 1\} \tag{5.3}$$

where ε is the given tolerance, and the scalar g_i represents the highest order at which species S_i reacts (see Cao et al. [14] for further details).

Following Cao et al. [14], we arrive at an efficient implementation of this leap condition [14] by classifying all reaction that are not in partial equilibrium as critical or non-critical, as follows. We begin by specifying the value of a control parameter, n_c. (Typically $n_c \in [2, 20]$). If a reactant is within n_c firings of producing a zero population, it is called a *critical reaction*. Let us denote by J_{cr}, J_{ncr}, and J_{ne} the set of indices of critical, non-critical reactions and not in partial equilibrium reactions, respectively; denote $J_{necr} = J_{ncr} \bigcap J_{ne}$ the index set of the reaction channels which are non-critical and not in partial equilibrium.

The leap condition (5.3) is implemented for the non-critical reactions, by choosing the time-step τ as:

$$\tau = \min_i \left\{ \frac{\max\{\varepsilon x_i/g_i, 1\}}{|\hat{\mu}_i(\mathbf{x})|}, \frac{\max\{\varepsilon x_i/g_i, 1\}^2}{\hat{\delta}_i^2(\mathbf{x})} \right\},$$

where

$$\hat{\mu}_i(\mathbf{x}) = \sum_{j \in J_{necr}} \nu_{ij} a_j(\mathbf{x}),$$

$$\hat{\delta}_i^2(\mathbf{x}) = \sum_{j \in J_{necr}} \nu_{ij}^2 a_j(\mathbf{x}).$$

5.4 Coupled Implicit Tau-Leaping (CIT)

This section introduces our novel technique (Morshed et al. [84]) for approximating the local sensitivities for stochastic discrete models of biochemical kinetics. This method is effective and accurate for stiff to very stiff models (involving multiple scales in time). Stiff systems are often encountered in applications, as biochemical systems regularly involve both fast and slow reactions. In contrast with the existing finite-difference schemes [2, 96], which utilize exact stochastic simulation algorithms to generate the nominal and perturbed trajectories, our strategy computes coupled paths using the (approximate) implicit tau-leaping strategy. The coupling we employ is related to 3.11, which is used in the CFD method [2]. This coupling shares similarities to the coupling in [69] and is applied in [3] for designing multi-level Monte Carlo methods for well-stirred stochastic biochemical systems. The coupled tau-leaping (CTL) method [83] uses finite-differences to estimate the sensitivities and the (approximate) explicit tau-leaping strategy to generate the coupled trajectories. However, the CTL was designed for biochemical networks that are at most moderately stiff. As opposed to these approaches, the novel CIT technique [84] involves solving implicit equations. For stiff to very stiff models, the new CIT strategy allows much larger time-steps than the previous methods. Consequently, the CIT algorithm is expected

to be significantly more efficient than the existing finite-difference estimators for such systems. Our CIT method is very accurate for approximating the sensitivity of the mean $E\left(X(t)\right)$.

In the CIT algorithm [84], the coupled (i.e. nominal and perturbed) implicit tau-leaping trajectories are generated as follows

$$\begin{aligned} X^c(t+\tau) &= \mathbf{x}^c + \sum_{j=1}^{M} \nu_j [(a_j^c(X^c(t+\tau)) - a_j^c(\mathbf{x}^c))\tau + P_{1,j}(m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h})\tau) \\ &\quad + P_{2,j}((a_j^c(\mathbf{x}^c) - m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h}))\tau)] \end{aligned} \tag{5.4}$$

$$\begin{aligned} X^{c+h}(t+\tau) &= \mathbf{x}^{c+h} + \sum_{j=1}^{M} \nu_j [(a_j^{c+h}(X^{c+h}(t+\tau)) - a_j^{c+h}(\mathbf{x}^{c+h}))\tau + P_{1,j}(m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h})\tau) \\ &\quad + P_{3,j}((a_j^{c+h}(\mathbf{x}^{c+h}) - m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h}))\tau)] \end{aligned} \tag{5.5}$$

with $X^{c+h}(t) = \mathbf{x}^{c+h}$ and $X^c(t) = \mathbf{x}^c$. The Poisson random variables $P_{1,j}$, $P_{2,j}$ and $P_{3,j}$ are independent. We denoted by $m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h}) = \min\left\{a_j^c(\mathbf{x}^c), a_j^{c+h}(\mathbf{x}^{c+h})\right\}$. The contribution of the shared term, $P_{1,j}(m_{j,c,h}(\mathbf{x}^c, \mathbf{x}^{c+h})\tau)$, is expected to be significant, thus leading to a strong coupling. A consequence of this strong coupling is the reduced variance observed for this method (as shown in the next section). Once the Poisson terms are generated, Newton's method is applied to solve numerically each implicit equation, (5.4) for $X^c(t+\tau)$ and (5.5) for $X^{c+h}(t+\tau)$, respectively.

For advancing the numerical solution, the CIT utilizes an extension of the adaptive time-stepping strategy introduced by Cao et al. [15], for the implicit tau-leaping method, as outlined in the previous section. A candidate leap is computed for the critical and non-

critical reactions, independently, on each of the nominal and perturbed trajectories, and then the smallest leap size is chosen as the next step.

CIT Algorithm

1. **Initialize simulation parameters:** assign a value to the tolerance for tau-leaping ε, the tolerance for Newton's method, TOL, the critical threshold n_c, the final time T and the partial equilibrium parameter δ.

2. **Initialize sample paths:** initialize the time $t \leftarrow 0$ and the states $X^{c+h} \leftarrow \mathbf{x}_0$ and $X^c \leftarrow \mathbf{x}_0$.

3. While $t < T$

   (a) **Compute the propensity functions:** $a_j^{c+h}(X^{c+h})$ and $a_j^c(X^c)$ for each $j = 1, \ldots, M$.

   (b) **Partial equilibrium condition:** for each set of reversible reactions in both systems, use the propensities to determine if the pair is in partial equilibrium, given by the condition $|a_+(x) - a_-(x)| \leq \delta \min\{a_+(x), a_-(x)\}$.

   (c) **Find the set of critical reactions for the nominal and perturbed trajectories:** for each non-partial equilibrium reaction R_j in the two systems, with propensity $a_j^c(X^c) > 0$ or $a_j^{c+h}(X^{c+h}) > 0$, determine

   $$L_j = \min_{i \in [1,N]; v_{ij} < 0} \left\lfloor \frac{x_i}{|\nu_{ij}|} \right\rfloor \qquad \lfloor \cdot \rfloor \text{ is the floor function (greatest integer less than).}$$

and set $J_{ncr} = \{j : L_j \geq n_c\}$, the set of non-critical reaction indexes. (L_j is the maximum number of reactions R_j that can occur without exhausting one of its reactants on either the nominal or the perturbed trajectory.)

(d) **Compute candidate stepsizes, ${\tau_1}^c$ and τ_1^{c+h}, for the non-critical and not in partial equilibrium reactions:** If no non-critical reactions occur ($J_{necr} = \emptyset$), set $\tau_1^c = \tau_1^{c+h} = \infty$. Otherwise, determine the set of indices I_{ncr} of species that are reactants of non-critical reactions. For every $i \in I_{ncr}$ and on each of the nominal and perturbed paths:

i. Set ψ_i to be the highest order at which the reactant S_i appears in a non-critical reaction.

ii. Compute g_i as follows:

A. If $\psi_i = 1$, take $g_i = 1$

B. If $\psi_i = 2$, take $g_i = 2$, unless the left hand side of the reaction is $S_i + S_i$, in which case take $g_i = \left(2 + \frac{1}{x_1 - 1}\right)$.

C. If $\psi_i = 3$, take $g_i = 3$, unless the left hand side of the reaction is $S_i + S_i + S_j$, in which case take $g_i = \frac{3}{2}\left(2 + \frac{1}{x_i - 1}\right)$, or the reaction is $S_i + S_i + S_i$, in which case take $g_i = \left(3 + \frac{1}{x_i - 1} + \frac{2}{x_i - 2}\right)$.

iii. If at least one reversible reaction has reached partial equilibrium: evaluate the auxiliary quantities $\hat{\mu}_i(\mathbf{x})$ and $\hat{\delta}_i^2(\mathbf{x})$ according to

$$\hat{\mu}_i(\mathbf{x}) = \sum_{j \in J} \nu_{ij} a_j^c(\mathbf{x}),$$

$$\hat{\delta}_i^2(\mathbf{x}) = \sum_{j \in J} \nu_{ij}^2 a_j^c(\mathbf{x}),$$

for $J = J_{necr}$. Here $J_{necr} = J_{ne} \cap J_{ncr}$, the reactions that are both non-critical and not in partial equilibrium.

Finally, find the first τ candidate ($\tau_1^{(c)}$ and $\tau_1^{(c+h)}$) for each system, using:

$$\tau = \min_i \left\{ \frac{\max\{\varepsilon x_i / g_i, 1\}}{|\hat{\mu}_i(\mathbf{x})|}, \frac{\max\{\varepsilon x_i / g_i, 1\}^2}{\hat{\delta}_i^2(\mathbf{x})} \right\}.$$

(e) **Compute candidate stepsizes, $\tau_2^{(c)}$ and $\tau_2^{(c+h)}$, for the critical reactions:** let $a_0^{cr,(c)}(X^c)$ and $a_0^{cr,(c+h)}(X^{c+h})$ be the sum of the critical reaction propensities for the nominal and perturbed paths, respectively. Sample $\xi_1^{(c)}$ and $\xi_1^{(c+h)}$ from the uniform distribution on $[0, 1]$, and calculate each system's second τ candidate ($\tau_2^{(c)}$ and $\tau_2^{(c+h)}$) with

$$\tau_2^{(c)} = (1/a_0^{cr,(c)}(X^c)) \ln(1/\xi_1^{(c)}) \ ,$$

$$\tau_2^{(c+h)} = (1/a_0^{cr,(c+h)}(X^{c+h})) \ln(1/\xi_1^{(c+h)}) \ .$$

(f) **Determine the next stepsize and the number of critical reactions:** Let $\tau_1 = \min\{\tau_1^{(c)}, \tau_1^{(c+h)}\}$ and $\tau_2 = \min\{\tau_2^{(c)}, \tau_2^{(c+h)}\}$.

i. If $\tau_1^{(c)} < \tau_2^{(c)}$ and $\tau_1^{(c+h)} < \tau_2^{(c+h)}$, no critical reaction occurs. Set $\tau = \tau_1$ and $k_j^c = k_j^{c+h} = 0$ for all critical reactions.

ii. else if $\tau_2^{(c)} < \tau_2^{(c+h)}$, one critical reaction fires on the nominal path. Sample ξ_2 from the uniform distribution on $[0, 1]$. Choose j_{cr} as the smallest integer

satisfying $\sum_{\ell \le j, \ell \in J_{cr}} a_\ell^c(X^c) > \xi_2 a_0^{cr,(c)}$. Take $\tau = \tau_2$, $k_{j_{cr}}^c = 1$, $k_{j_{cr}}^{c+h} = 0$ and $k_j^c = k_j^{c+h} = 0$ for all the other critical reactions.

iii. else if $\tau_2^{(c+h)} < \tau_2^{(c)}$, one critical reaction fires on the perturbed path. Sample ξ_2 from the uniform distribution on $[0, 1]$. Choose j_{cr} as the smallest integer satisfying $\sum_{\ell \le j, \ell \in J_{cr}} a_\ell^{c+h}(X^{(c+h)}) > \xi_2 a_0^{cr,(c+h)}$. Take $\tau = \tau_2$, $k_{j_{cr}}^c = 0$, $k_{j_{cr}}^{c+h} = 1$ and $k_j^c = k_j^{c+h} = 0$ for all the other critical reactions.

iv. else a single critical reaction occurs on each of the coupled paths. Sample ξ_2 from the uniform distribution on $[0, 1]$. Choose j_{cr} as the smallest integer satisfying $\sum_{\ell \le j, \ell \in J_{cr}} a_\ell^c(X^{(c)}) > \xi_2 a_0^{cr,(c)}$. Take $\tau = \tau_2$, $k_{j_{cr}}^c = k_{j_{cr}}^{c+h} = 1$ and $k_j^c = k_j^{c+h} = 0$ for all the other critical reactions.

(g) **Step over the non-critical reactions:** For each $j \in J_{ncr}$, compute $m_j = \min(a_j^c(X^c), a_j^{c+h}(X^{c+h}))$.

i. Generate samples from Poisson distributions

$$\begin{aligned} P_{1,j} &= \text{Poisson}(m_j \tau), \\ P_{2,j} &= \text{Poisson}((a_j^c(X^c) - m_j)\tau), \\ P_{3,j} &= \text{Poisson}((a_j^{c+h}(X^{c+h}) - m_j)\tau). \end{aligned} \tag{5.6}$$

ii. Apply Newton's method, with tolerance TOL, to solve each of the systems

$$U = X^c + \sum_{j \in J_{ncr}} \{[a_j^c(U) - a_j^c(X^c)]\tau + P_{1,j} + P_{2,j}\}\nu_j,$$

$$V = X^{c+h} + \sum_{j \in J_{ncr}} \{[(a_j^{c+h}(V) - a_j^{c+h}(X^{c+h})]\tau + P_{1,j} + P_{3,j}\}\nu_j.$$

where $P_{1,j}$, $P_{2,j}$ and $P_{3,j}$ are given by 5.6.

iii. Update $X^c \leftarrow U$, $X^{c+h} \leftarrow V$.

(h) **Implement the step:** update the time $t \leftarrow t + \tau$ and the system states

$$X^c \leftarrow X^c + \sum_{j \in J_{cr}} k_j^c \nu_j,$$
$$X^{c+h} \leftarrow X^{c+h} + \sum_{j \in J_{cr}} k_j^{c+h} \nu_j.$$

(i) **Approximate sensitivity on the sample path:** $Z = (f(X^{c+h}) - f(X^c))/h$ at current time.

5.5 Numerical Results

This section compares the coupled implicit tau-leaping (CIT) method [84] with the coupled finite-difference (CFD) strategy on some examples of stiff biochemical systems. Recall that, of the published finite-difference techniques for estimating the sensitivities, the CFD technique provides estimates with the lowest variance [2].

In our comparisons, we use ensembles of 10,000 paths of the CFD and of the new CIT methods, respectively. We apply the CIT algorithm as described above with tolerance $\varepsilon = 0.05$, TOL=0.01, and $\delta = 0.05$. We show that the CIT method produces smaller variances than the CFD strategy for the first two models and similar variances for the third model. The CIT estimator is found to be significantly faster than the CFD. The

Table 5.1: Decay-dimerization model

R_j	Reaction	Propensity	Nominal rate constant
R_1	$S_1 \xrightarrow{C_1} \oslash$	$a_1 = C_1 X_1$	$C_1 = 0.05$
R_2	$S_1 + S_1 \xrightarrow{C_2} S_2$	$a_2 = C_2 X_1 (X_1 - 1)/2$	$C_2 = 50$
R_3	$S_2 \xrightarrow{C_3} S_1 + S_1$	$a_3 = C_3 X_2$	$C_3 = 10^6$
R_4	$S_2 \xrightarrow{C_4} S_3$	$a_4 = C_4 X_2$	$C_4 = 0.05$

efficiency is measured by

$$\text{Speed-up over CFD } = \frac{\text{CPU(CFD)}}{\text{CPU (CIT)}}.$$

5.5.1 Decay-dimerization Model

The decay-dimerization model of [93] consists of three molecular species involved in four chemical reactions (Figure 5.1). The reactions and propensities are given in Table 5.1, along with a set of nominal values for the rate constants.

The system was simulated on the time-interval $[0, 1]$, with initial conditions ($X_1(0)$, $X_2(0)$, $X_3(0)$)=(400, 800, 0) and the parameter $n_c = 10$. The mean of the state variable X_2 (i.e. the number of S_2 molecules), for the adaptive implicit tau-leaping algorithm and for the next reaction method, are plotted in Figure 5.2(a); Figure 5.2(b) shows the standard deviation of this state variable. The estimated sensitivity of S_2 with respect to the parameter

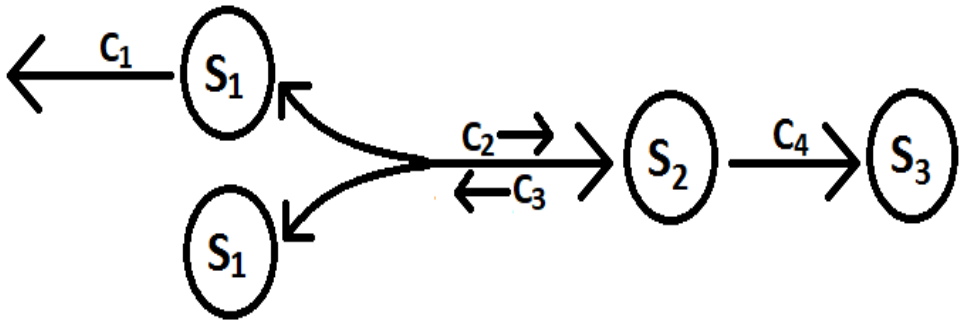


Figure 5.1: Decay-dimerization model reaction chain.

C_2 and that its standard deviation are shown in Figure 5.2(c-d). The perturbation parameter is $h = 0.05$ (i.e. 0.1% of the nominal parameter value). Figure 5.2(d) demonstrates that the variance of the CIT estimator small compared to that of the CFD, demonstrating accuracy. Moreover, the speed-up of CIT scheme over the CFD technique for estimating sensitivities for this particular simulation on the time interval is $[0, 1]$ is

$$\text{Speed-up over CFD} = 9632.70.$$

5.5.2 Genetic Positive Feedback Loop Model

We next consider a simple model of positive feedback in gene expression (Figure 5.3), as presented in [88]. Referring to Table 5.2, x represents a monomeric protein, y the protein dimer, d_0 - the unoccupied regulatory site on the gene coding for x, d_r the dimer-occupied site, and m, the $mRNA$ transcript. The reactions, propensities and a set of nominal parameter values are included in the table.

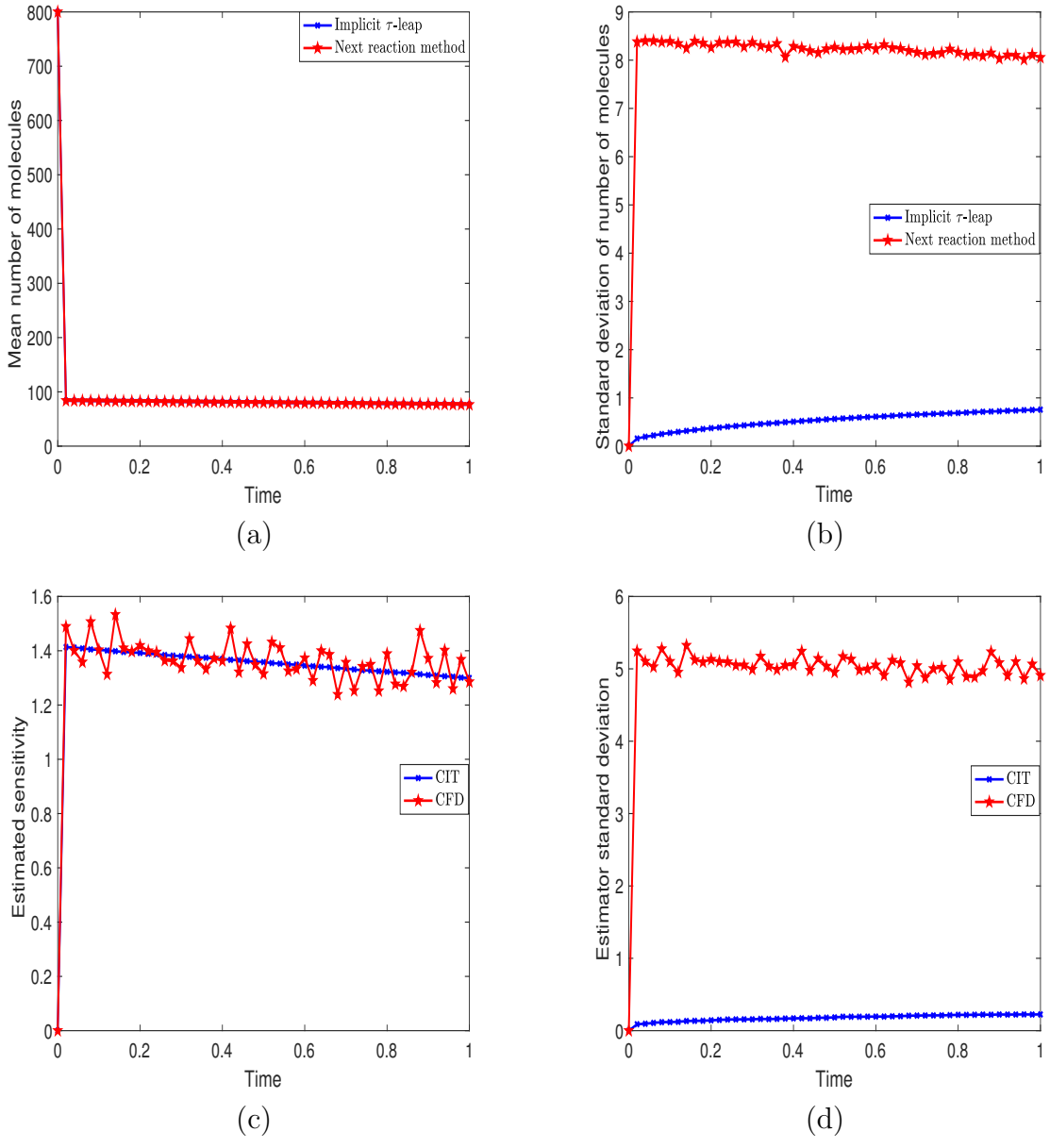


Figure 5.2: Decay-dimerization model: 10,000 trajectories were generated on the time-interval $[0, 1]$, with initial condition $(X_1(0), X_2(0), X_3(0)) = (400, 800, 0)$ and parameters in Table 5.1. (a-b) The mean and standard deviation of the number of molecules for species S_2 were calculated by the next reaction method and the adaptive Implicit tau-leaping algorithm. (c-d) The finite-difference estimates of the sensitivity of the abundance of S_2 with respect to C_2, and the standard deviation of the estimators, for the CFD and CIT.

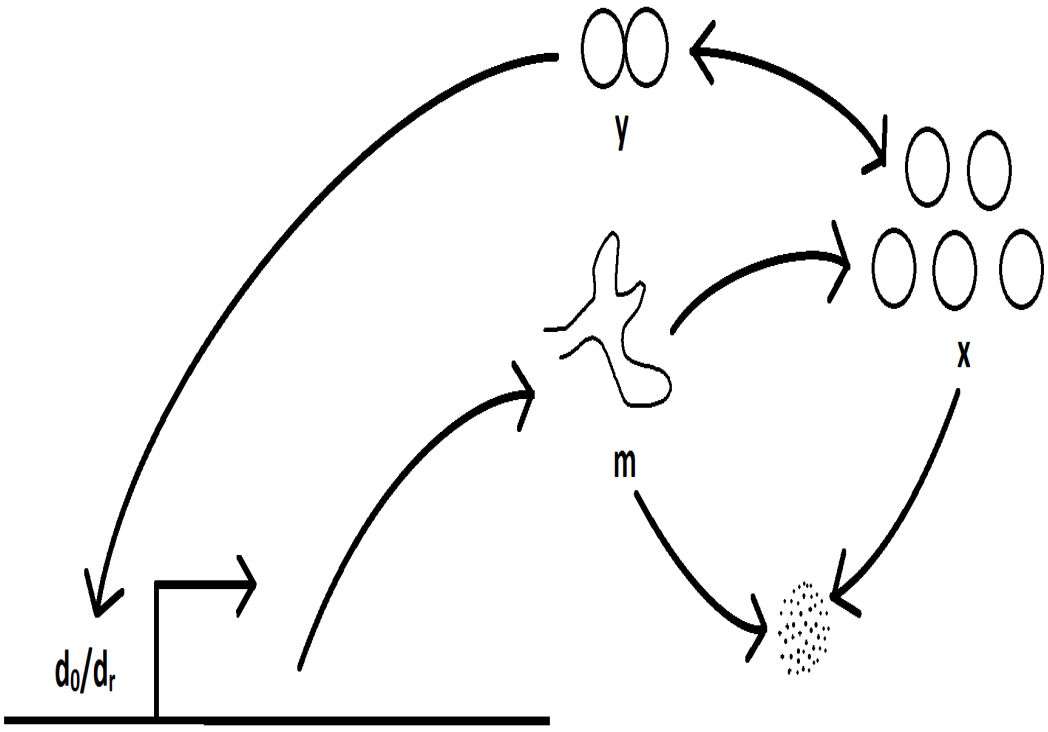


Figure 5.3: Schematic diagram of Genetic positive feedback loop model.

Table 5.2: Genetic positive feedback loop model

R_j	Reaction	Propensity	Nominal rate constant
R_1	$x + x \xrightarrow{C_1} y$	$a_1 = C_1 X(X-1)/2$	$C_1 = 5000$
R_2	$y \xrightarrow{C_2} x + x$	$a_2 = C_2 Y$	$C_2 = 10^6$
R_3	$y + d_0 \xrightarrow{C_3} d_r$	$a_3 = C_3 Y D_0$	$C_3 = 5000$
R_4	$d_r \xrightarrow{C_4} y + d_0$	$a_4 = C_4 D_r$	$C_4 = 10^6$
R_5	$d_0 \xrightarrow{C_5} d_0 + m$	$a_5 = C_5 d_0$	$C_5 = 10$
R_6	$d_r \xrightarrow{C_6} d_r + m$	$a_6 = C_6 D_r$	$C_6 = 20$
R_7	$m \xrightarrow{C_7} m + x$	$a_7 = C_7 M$	$C_7 = 1$
R_8	$x \xrightarrow{C_8} \oslash$	$a_8 = C_8 X$	$C_8 = 0.8$
R_9	$m \xrightarrow{C_9} \oslash$	$a_9 = C_9 M$	$C_9 = 7$

We ran simulations from initial molecular amounts of $(X_1(0), X_2(0), X_3(0), X_4(0), X_5(0)) = (10, 20, 10, 40, 0)$ over the time-interval $[0, 2]$, with $n_c = 10$.

Figure 5.4(a) presents the evolution of the mean amount of the x molecules over $10,000$ paths, generated with the coupled implicit tau-leaping algorithm and the next reaction method, respectively. The standard deviation of the molecular count of x as a function of time, for each of the two algorithms, is shown in Figure 5.4(b). The behaviours of the estimated sensitivity of the x molecular numbers with respect to the parameter C_1, using

the CIT and the CFD methods are presented in Figure 5.4(c), whereas the corresponding standard deviations of the CIT and CFD estimators are given in Figure 5.4(d). The simulations are performed with a perturbation $h = 0.5$ (i.e. 0.01% of the nominal parameter value). From Figure 5.4(d), we observe that the CIT estimator variance is low compared to the variance of the CFD estimator, therefore the sensitivity estimation of the new CIT method is more accurate. This result is confirmed by Figure 5.4(c). In addition, for the set of parameters used, the speed-up, on time interval $[0, 2]$, of the CIT over the CFD is significant .

$$\text{Speed-up over CFD} = 2656.43 \ .$$

5.5.3 Collins Toggle Switch Model

The Collins toggle switch [54] is a gene regulatory network that exhibits bistability: two genes, each encoding a repressor of the other. Referring to figure (Figure 5.5) the species p_1 and p_2 are gene's protein products, while m_1 and m_2 denote the corresponding $mRNA$ transcripts. The parameters α_1 and α_2 denote the maximal transcription rates. Furthermore, β and γ are the degrees of nonlinearity in the repression mechanisms. Gene 1 and Gene 2 repress the expression of each other, thereby leading to a bistable system. The system is perfectly bistable when $\alpha_1 = \alpha_2$ and the maximal expression rates are adequately large. The stiffness parameter of the model is defined by k, where the propensity of mRNA transcription and degradation is proportional to the value of k. For increased values of k, the transcription and degradation rate of mRNA increase thereby the stiffness of the

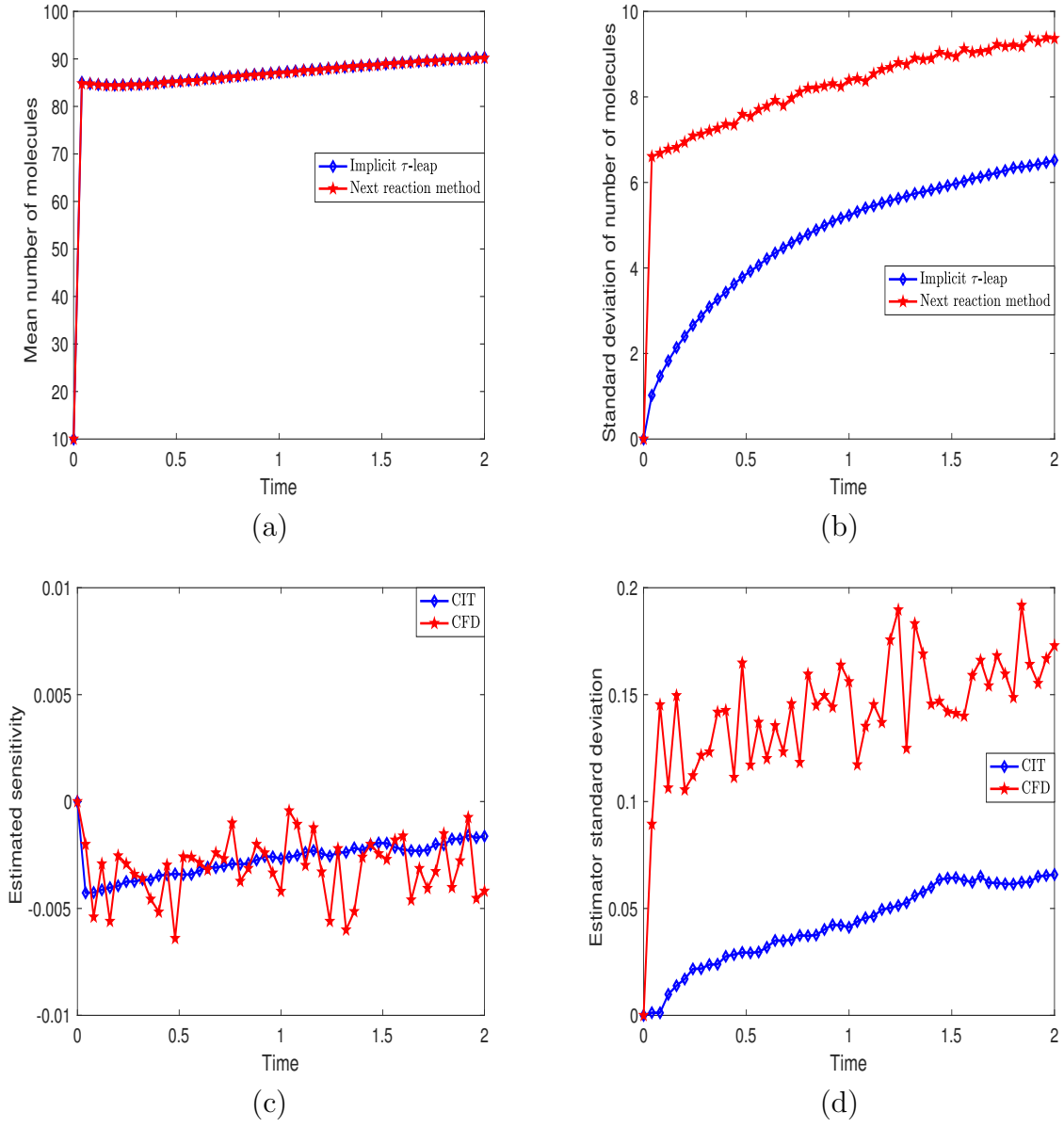


Figure 5.4: Genetic positive feedback loop model. 10000 sample paths with initial condition $(X_1(0), X_2(0), X_3(0), X_4(0), X_5(0)) = (10, 20, 10, 40, 0)$ and parameters as in Table 5.2 were generated on the time-interval $[0, 2]$. (a-b) The mean and standard deviation of the number of molecules for species x were calculated by the next reaction method and the adaptive Implicit tau-leaping algorithm. (c-d) The mean and standard deviation of the finite-difference estimators determined via the CFD and Implicit tau leaping methods, of the sensitivity of the abundance of x to the parameter C_1.

model increases. Table 5.3 lists the reactions, their propensities and rate constants.

Table 5.3: Collin's toggle switch model

R_j	Reaction	Propensity	Nominal rate constant
R_1	$\oslash \xrightarrow{C_1} m_1$	$a_1 = k\frac{\alpha_1}{1+(X_2)^\beta}$	$C_1 = \alpha_1 = 28.98$, $\beta = 4$
R_2	$m_1 \xrightarrow{C_2} \oslash$	$a_2 = kC_2X_3$	$C_2 = 0.23$
R_3	$m_1 \xrightarrow{C_3} p_1 + m_1$	$a_3 = C_3X_3$	$C_3 = 0.23$
R_4	$p_1 \xrightarrow{C_4} \oslash$	$a_4 = C_4X_1$	$C_4 = 0.23$
R_5	$\oslash \xrightarrow{C_5} m_2$	$a_5 = k\frac{\alpha_2}{1+(X_1)^\gamma}$	$C_5 = \alpha_2 = 28.98$, $\gamma = 4$
R_6	$m_2 \xrightarrow{C_6} \oslash$	$a_6 = kC_6X_4$	$C_6 = 0.23$
R_7	$m_2 \xrightarrow{C_7} p_2 + m_2$	$a_7 = C_7X_4$	$C_7 = 0.23$
R_8	$p_2 \xrightarrow{C_8} \oslash$	$a_8 = C_8X_2$	$C_8 = 0.23$

This system was integrated on the time-interval $[0, 2000]$, with initial conditions $X(0) = (76, 75, 60, 60)$ and $n_c = 5$. Sample trajectories for all the species, simulated with the underlying implicit tau-leaping method are shown in Figure 5.6(a-d). The mean and standard deviation number of p_1 molecules for the proposed implicit tau-leaping algorithm and for the the next reaction method are plotted in Figure 5.7(a) and Figure 5.7(b), respectively. Figures 5.7(c-d) present the finite-difference estimation of the sensitivity of the p_1 molec-

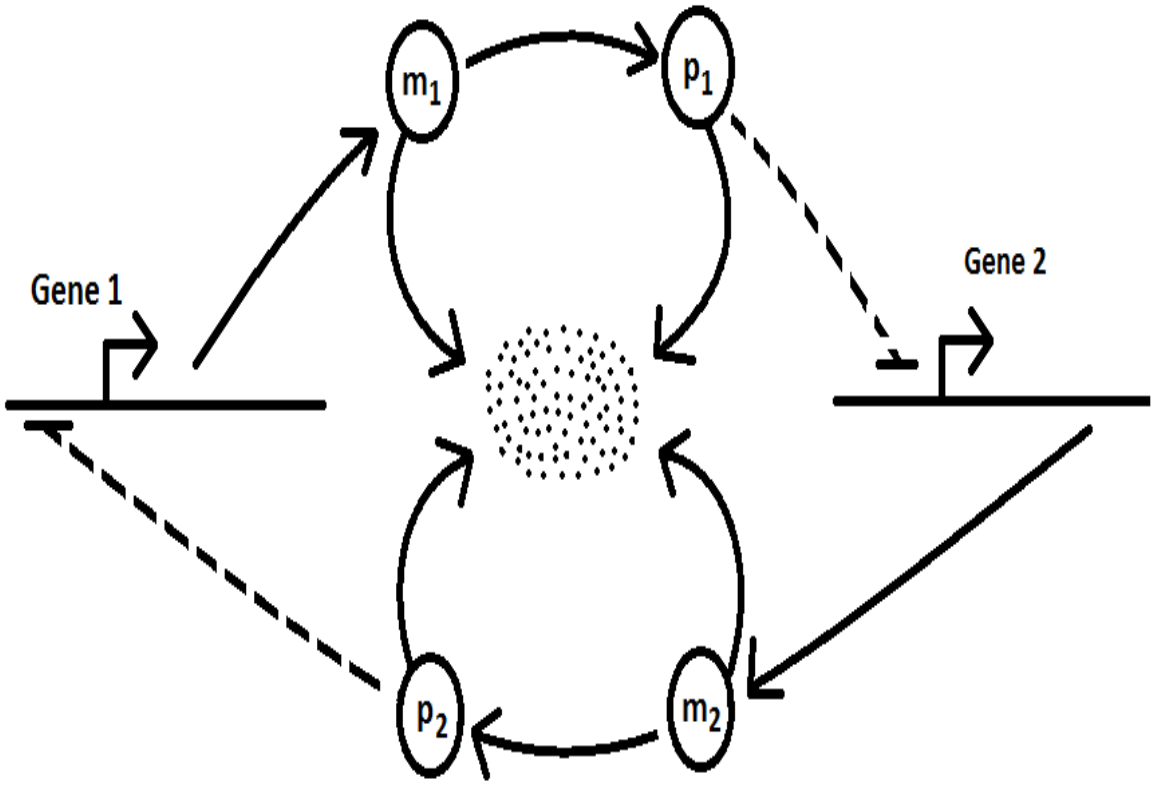


Figure 5.5: Collin's Toggle Switch model reaction scheme diagram.

ular amount with respect to the parameter C_1 and the estimator's standard deviation for each of the CIT and CFD algorithms. In these simulations, the perturbation parameter is $h = 0.05$ (i.e. 0.2% of the nominal parameter value). The estimation of the sensitivity is similar for the CIT and the CFD methods, while the standard deviation of the CIT estimator is slightly larger than that of the CFD estimator. However, for the set of parameters in Table 5.3, the speed-up of the CIT over the CFD is 74-fold. In addition, the performance of the CIT and CFD methods was studied for various degrees of stiffness in the system and the results were reported in Table 5.4. For the parameters tested which lead to a stiff to very stiff biochemical model, we obtained a speed-up of the new CIT strategy compared to the existing CFD method of up to 468 times.

For non-stiff models, the CIT algorithm will perform no better than the CFD method. For this model, a similar computational time for the two algorithms is obtained when the propensities of the fastest and slowest reactions are separated by about two orders of magnitude.

As shown in panel (d) of Figures 5.2, 5.4, and 5.7, the variance of the CIT estimator is not always comparable to that of the CFD estimator (smaller in the first two examples, larger in the third). For first two models, we observe that our CIT method is more accurate and far more efficient than the existing CFD strategy. For the third model, when the value of stiffness parameter k grows, our CIT method becomes increasingly more efficient than the CFD scheme. On the other hand, for the Collins toggle switch model, the variance of the CIT estimator is slightly larger than that of the CFD. The implicit tau-leaping scheme damps the noise for systems reaching a steady state [93]. However, for the toggle switch model, the implicit tau-leaping scheme does not cause noise reduction. Trajectories

frequently switch between two states, the model exhibiting bi-stable behaviour. This behaviour restricts the noise damping property of the implicit tau-leaping scheme and leads to a slightly larger variance of the CIT algorithm than that of the CFD, unlike for the previous two models. According to our numerical experiments, we conclude that our CIT method is expected to be more accurate and significantly more efficient than the CFD technique, when the stiff system reaches a steady-state.

Table 5.4: Collin's toggle switch model: the speed-up of the CIT compared to the CFD for estimating the sensitivity of p_1 with respect to C_1 for $h = 0.05$ on time interval $[0, 2000]$ of CIT over the CFD.

Method	Stiffness parameter k	Speed-up
CIT	300	10.16
CIT	1000	74.71
CIT	3000	468.43
CFD	–	1

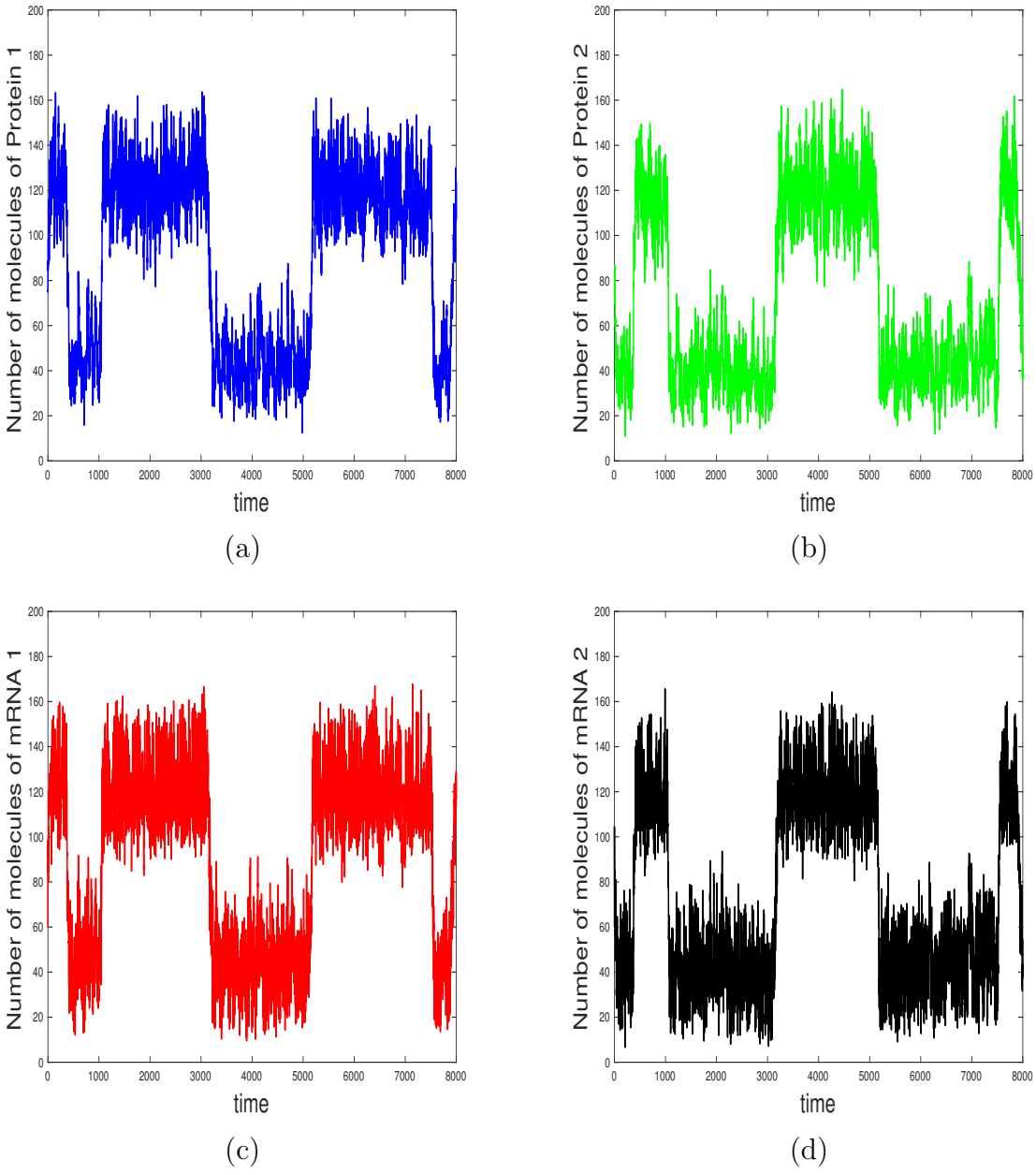


Figure 5.6: Collin's toggle switch model: A sample path of all species with initial condition $(X_1(0), X_2(0), X_3(0), X_4(0)) = (76, 75, 60, 60)$ and the parameters in Table 5.3 generated with the Implicit tau-leaping method on the time-interval $[0, 8000]$.

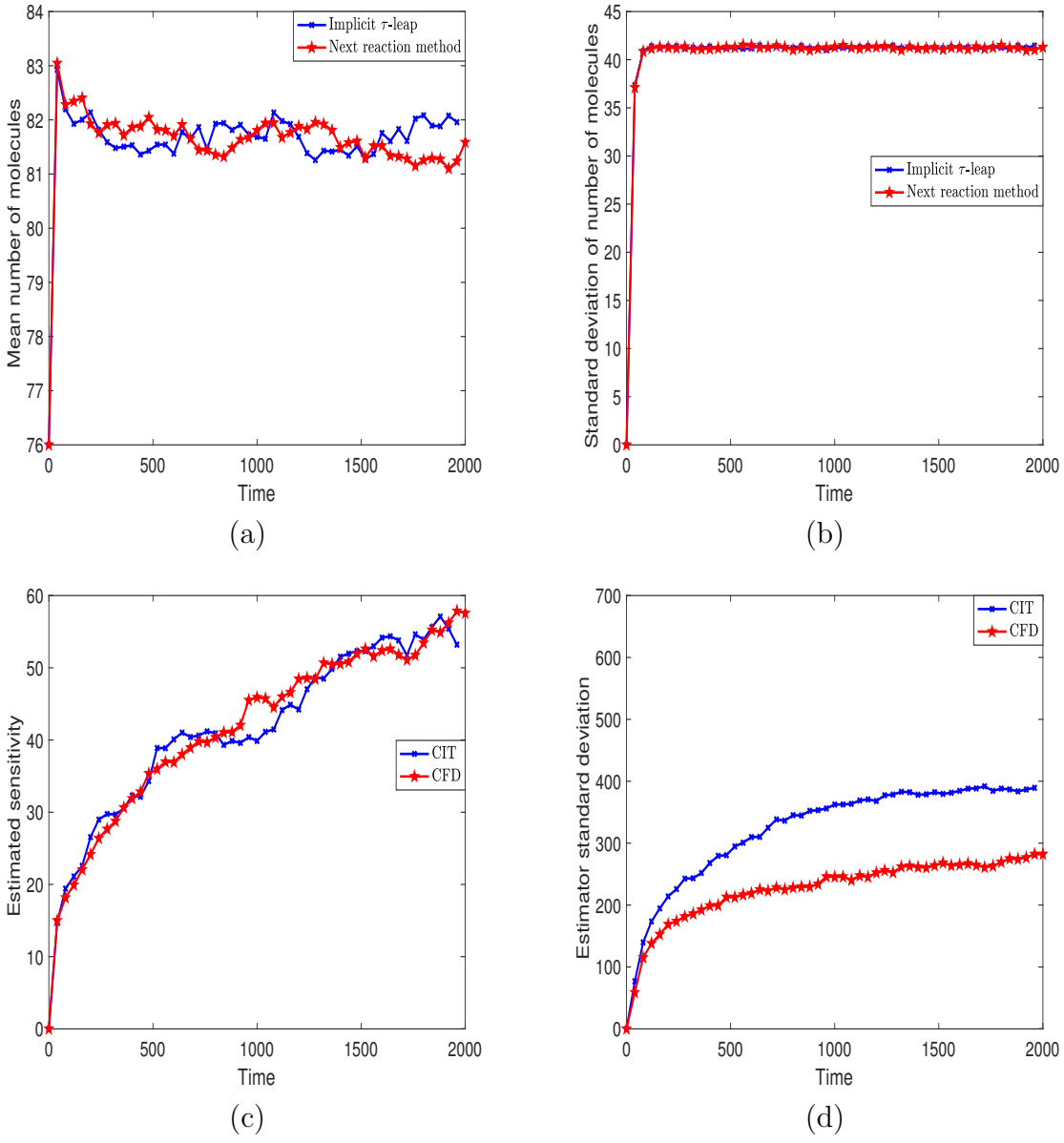


Figure 5.7: Collin's toggle switch model. 10000 sample paths with initial condition $(X_1(0), X_2(0), X_3(0), X_4(0)) = (76, 75, 60, 60)$ and parameters as in Table 5.3 were generated on the time-interval $[0, 2000]$. (a-b) The mean and standard deviation of the number of molecules for species p_1 were calculated by the next reaction method and the adaptive Implicit tau-leaping algorithm. (c-d) The Mean and standard deviation of the finite-difference estimators determined via the CFD and Implicit tau leaping methods, of the sensitivity of the abundance of p_1 to the parameter C_1.

Chapter 6

Identifiability Analysis

6.1 Introduction

One of the many objectives of constructing a model is to predict how a physical system will behave in the future. When designing a mathematical model of a physical system, the key question is that of determining the quality of the estimated parameter values based on available experimental data. Perhaps the biggest question facing the scientist is whether the construction of a unique mathematical model is even possible and whether a unique set of parameters can be found to parameterize the model in a way that is consistent with observable data [4].

If the simulated results are consistent to the actual observations, we can say that the model provides a good representation of reality. However simple consistency with observations may not be enough for the model to be sufficiently useful if the parameters of such model

cannot be uniquely determined [4]. This means that a useful model cannot provide the same results for multiple sets of input parameters. The quality of the estimated data is therefore very important. In the literature this is the problem of identifiability of a model.

To assess the quality of estimated parameter values, identifiability analysis plays an important role [4, 68, 89, 98, 121]. The identifiability analysis can be used to assess the quality of a unique set of model parameters $C = (c_1, c_2, ..., c_M)$ which is closest to the observations. The parameters can be globally or locally identifiable. A globally identifiable parameter requires the ability to uniquely determine model parameters, given an ideal set of observations (data which is free of errors) over the entire parameter space [4]. In contrast, a locally identifiable parameter C requires a unique output for each set of values of the parameter only in the neighborhood of C. The parameters, which result in the output of a model, may not be the only ones in the entire parameter space (there could be countably many parameter values in the entire parameter space which lead to the same model output) [4]. If there are uncountably many values for parameters $C = (c_1, c_2, ..., c_M)$, which give the same output, then the parameter C is not identifiable.

Generally, identifiability consists of two types of analysis [4]. First, structural identifiability analysis, which is also known as a priori identifiability analysis, investigates the theoretical possibility of finding a unique (globally or locally) set of parameter values which are most similar to the observations (where we assume that the observations are free of noise or errors). Second, a posteriori (practical identifiability), investigates the practical possibility of finding a unique (globally or locally) set of parameter values which are most similar to the observations with the available data (but the available data can be noisy or may have errors). The a posteriori analysis can be very important in real life situations because

measurement of observables are always associated with error.

An important approach to determining the identifiability of model parameters relates the sensitivity of the model output to changing model parameters (which was described further in Chapter 3). In the current approach we use the sensitivity analysis approach adopted from [68, 98, 122]. In this approach a high level of identifiability is associated with model parameters c_j if small changes in parameter values lead to a large impact on the overall model output. In addition, confidence intervals for the values of the model parameters are calculated using the methods adopted from [4, 33]. These are discussed in greater detail in Section 6.5. The main goal of this analysis is to generalize the identifiability approach, based on a sensitivity matrix, to stochastic models of biochemical systems.

6.2 Identifiability Approaches for Deterministic Model

There are many approaches in the literature to showing identifiability of a deterministic model. The existence of so many approaches (in the literature) is mostly due to the large differences between all available models. Since the different models differ in terms of structure, complexity and their applications, there is currently no single method or technique that can be used to determine identifiability with every model.

An important practical identifiability approach by Brun et al. [98] uses a sensitivity analysis. This approach attempts to understand how certain properties of the system change when variations are introduced into the model's parameters. A parameter's sensitivity is a measure of how much change in the system output results due to varying the parameter.

Note that parameter's sensitivity was extensively described in Chapter 3. Recall that, a certain parameter of a model is regarded as highly sensitive if a small change in the parameter results in a large change in the system's outcome.

In the work of Brun et al. [98], all identifiability indices are computed by using the sensitivity matrix. In their work, they introduced two types of identifiability measures. The first type of identifiability measures is based on the model output sensitivity to single parameters:

$$S_{ij} = \frac{\partial x_i}{\partial c_j}. \tag{6.1}$$

The first derivative term is known as the first order sensitivity coefficients and the matrix:

$$S = \frac{\partial X}{\partial C} = \begin{pmatrix} \frac{\partial x_1}{\partial c_1} & \frac{\partial x_1}{\partial c_2} & \cdots & \frac{\partial x_1}{\partial c_M} \\ \frac{\partial x_2}{\partial c_1} & \ddots & & \frac{\partial x_2}{\partial c_M} \\ \vdots & & \ddots & \vdots \\ \frac{\partial x_N}{\partial c_1} & \frac{\partial x_N}{\partial c_2} & \cdots & \frac{\partial x_N}{\partial c_M} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1M} \\ s_{21} & \ddots & & s_{2M} \\ \vdots & & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NM} \end{pmatrix}. \tag{6.2}$$

Where:

$$s_{ij} = \frac{\partial x_i}{\partial c_j} = \frac{x_i(t, c_j + \Delta c_j) - x_i(c_j)}{\Delta c_j}. \tag{6.3}$$

Brun et al. [98] uses the sensitivity matrix to develop the identifiability indexes. An alternative approach by Brun et al. [98] considered the influence of the entire parameter set on the output of the model. The objective of this approach was to determine if there

was a linear dependence between the columns of the relative sensitivity matrix $\overline{S}_j$. The relative sensitivity can be defined by

$$\overline{S}_{ij} = \frac{\partial x_i}{\partial c_j}\frac{c_j}{x_i}. \tag{6.4}$$

The jth column of the relative sensitivity matrix is given by

$$\overline{S}_j = \left(\frac{\partial x_1}{\partial c_j}\frac{c_j}{x_1}, \frac{\partial x_2}{\partial c_j}\frac{c_j}{x_2}, ... \frac{\partial x_N}{\partial c_j}\frac{c_j}{x_N}\right)^T \tag{6.5}$$

for all species with respect to the parameter c_j.

In the literature a number of approaches have been developed for the purpose of understanding how identifiability can translate to stochastic models. However these are typically far more complex and require very careful consideration. In fact, a complete approach to defining identifiability in a stochastic case has not yet been developed in the relevant literature. In the subsequent section we provide a brief introduction to treating identifiability in stochastic models. We outline the details of this approach in Section 6.5.

6.3 Fisher Information Matrix (FIM) and Cramer-Rao Bounds

The evolution of the state variable $\mathbf{x} = (x_1, x_2, ..., x_N)$ depends to a large extent on the values of model parameters $C = (c_1, c_2, ..., c_M)$. The Fisher Information Matrix (FIM)

presents a way of determining the amount of information that can be obtained about the unobservable model parameters $C = (c_1, c_2, ..., c_M)$ from the observable state of the system $\mathbf{x} = (x_1, x_2, ..., x_N)$.

By definition, a deterministic model gives the same output for a specific input even if the procedure is repeated many times. This is different than in the stochastic model where the output will be different even with the same input (that is true because the input is a random variable chosen from a probability distribution). In the deterministic case, the parameters and their relationships are provided by a system of differential equations and do not involve any random variables. However in a physical system, the measurement of a state (state of the system) will always include some level of error. The error part will have values within a region in the parameter space (even in the deterministic model). The error part of the measurement is associated with unknown parameters $\hat{C} = (\hat{c}_1, \hat{c}_2, ..., \hat{c}_M)$ (sometimes known as the nuisance parameters [4]). The nuisance parameters represent those unknown model parameters, which may include measurement error. The main objective is to find out how much uncertainty there is in the nuisance parameters or how much useful information is contained in each measurement with respect to each parameter of the state of the system. The FIM provides a way to identify the amount of information of a measurement with respect to a specific parameter. The matrix provides no information about the uncertainty within the measurement. The level of uncertainty within the measurement is estimated by the Cramer-Rao bound which provides the lowest uncertainty level of the parameters. The FIM matrix was defined as a function of the sensitivity matrix by Ashyraliyev et al. [4]. Recall the sensitivity matrix from equation 6.2:

$$S = \frac{\partial X}{\partial C} = \begin{pmatrix} \frac{\partial x_1}{\partial c_1} & \frac{\partial x_1}{\partial c_2} & \cdots & \frac{\partial x_1}{\partial c_M} \\ \frac{\partial x_2}{\partial c_1} & \ddots & & \frac{\partial x_2}{\partial c_M} \\ \vdots & & \ddots & \vdots \\ \frac{\partial x_N}{\partial c_1} & \frac{\partial x_N}{\partial c_2} & \cdots & \frac{\partial x_N}{\partial c_M} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1M} \\ s_{21} & \ddots & & s_{2M} \\ \vdots & & \ddots & \vdots \\ s_{N1} & s_{N2} & \cdots & s_{NM} \end{pmatrix}. \tag{6.6}$$

The FIM [4] can be defined as $F = S^T S$. The eigenvalues of this matrix are related to the identifiability of the model parameters (if the eigenvalues are zero then the model parameters are not identifiable). The inverse of the FIM is related to the uncertainty of the measurement with respect to each parameter. The uncertainty of the measurement with respect to parameter is provided by the standard deviation and can be bound by:

$$error_i = \sqrt{var(c_i)} \geq \sqrt{F_{ii}^{-1}(c_i)} \geq \frac{1}{\sqrt{F_{ii}(c_i)}}. \tag{6.7}$$

The above statement (equation 6.7) represents the Cramer-Rao bound for the lowest uncertainty level of the parameters $C = (c_1, c_2, ..., c_M)$.

6.4 Identifiability Approaches for Stochastic Model

In this section we consider the Chemical Master Equation (CME) model for biochemical reactions network, which was described explicitly in Section 2.6. Recall that the CME consists of a set of ordinary differential equations. The state of the system is given by a state vector $\mathbf{x} = (x_1, x_2, ..., x_N)$ where each component of the vector x_i denotes the number of molecules of type i.

A number of models can be constructed using the CME approach (such as biochemical kinetic models). An important problem here is to establish if such models are identifiable with respect to the set of parameters $C = (c_1, c_2, ..., c_M)$. Recall that in the deterministic case, the identifiability of the model with respect to its parameters could be done using several different approaches based on the computation of the sensitivity matrix which was described further in Section 6.2.

In the stochastic case, modeled by the CME, the deterministic sensitivity matrix is no longer meaningful. This is because in the stochastic case, model variables (such as the output) represent random variables drawn based on probability density function $f(x, t)$ instead of a single value. Determining identifiability will need to use this stochastic information in order to provide useful analysis. The definition of identifiability requires that for each set of input parameter values a unique value of model outputs is given. However in the stochastic case identifiability must be defined with respect to unique distributions instead of a single output value [68]. A simple way to use the stochastic information is to compute the sensitivity of probability density function to changes in the parameter values from the expectation value as:

$$S = E\left[\frac{\partial f(x, C, t)}{\partial c}\right]. \tag{6.8}$$

This is similar to the approach taken by Komorowski et al. [68] where the FIM matrix is constructed as an expectation value of the rate of change (with respect to each parameter) of the distribution. As a result, the FIM can be interpreted as the measure of how the distribution changes in response to the changes in parameter values (this is appropriate in

the stochastic case).

The construction of the FIM for a general CME model is quite complex and can be done using a Monte Carlo approach which was described further in Section 2.7.1 and used in the current analysis. In order to reduce/avoid the complexity of computation, Komorowski et al. [68] used the linear noise approximation method (LNA) [68] to construct the FIM without having to use Monte Carlo methods. Unfortunately, this approach (LNA) can only be used when populations are very large. This is a significant drawback of this approach because it does not provide a useful alternative to the Monte Carlo approach when the populations are small. In practical situations when populations are not large enough, no other methods are available and the Monte Carlo approach must be used to construct the FIM. In the current approach we are interested in constructing an FIM for populations of any size. The use of Monte Carlo is therefore a part of the current approach.

The big question is then whether methods used for identifiability in the deterministic model can be used with the expected value measures such as the FIM (as constructed by Komorowski et al. [68]). Komorowski et al. [68] provided a justification that the FIM can be seen as a sensitivity matrix whose eigenvalues are associated with the identifiability of the model. The number of eigenvalues, which are not zero, provides information about the number of parameters which are identifiable in the model.

The expectation value of the distribution provides the first order information while the higher moments are associated with variance, kurtosis, etc., of the distribution. Komorowski et al. [68] shows that the diagonal elements of the inverse of the FIM can be used to provide the lower bound on the variance of the sensitivity of the distribution to

changes in the parameter values. This is known as the Cramer-Rao inequality which was discussed further in Section 6.3. The main idea behind the inequality (Cramer-Rao) is to provide an additional tool for parameter identifiability. Identifiability estimators (like the eigenvalues of the FIM) with smaller variance have higher identifiability than those with higher variance even if they both have the same eigenvalues.

6.5 Current Approach: Application of Monte Carlo Approaches to Sensitivity Estimation to Identifiability for CME Models

In the current analysis, we attempt to identify and implement a simple approach, based on the sensitivity matrix, using a Monte Carlo algorithm. An approach based on the sensitivity matrix was shown by Yao et al. [122] to provide a useful method to ranking model parameters based on their identifiability. The sensitivity matrix can also be used to obtain a confidence interval within which the true value of the model parameters can be found. In the current analysis the parameter ranking algorithm was adopted from [68, 98, 122]). The method to compute the confidence intervals for all of the model parameter values were adopted from [4, 33].

In the paper by Komorowski et al. [68], where they concluded that a lot of information about the identifiability of a stochastic model can be obtained from the FIM matrix. The FIM is an expectation value of the sensitivity of the distribution (from which all observable values are chosen) to the model parameters. This strongly suggests that a distribution

can be build by running the sensitivity calculation over many trials using a Monte Carlo algorithm which was described further in Section 2.7.1. In the current approach, the distribution is constructed from the sensitivity algorithm for finding the sensitivity matrix using the MC approach. In addition, the variability of the distribution (from the Monte Carlo process) can be used to add new information about the identifiability of the model using Cramer-Rao bounds (which was described further in Section 6.3) on the variability of the distribution. Distributions with high variance are less identifiable.

The goal of the current approach is to approximate the FIM by constructing the sensitivity matrix and using it as an identifiability tool to assess the quality of the estimated parameter values and finding the confidence intervals for true values of the model parameters. In the subsequent section we describe the details of this approach.

6.5.1 Procedure for Determining the Sensitivity Score

In Section 6.2, the construction of the sensitivity matrix was outlined in the context of the deterministic model. In Section 6.4, an approach to constructing a sensitivity matrix for a stochastic model was introduced as an expected value of a distribution. Finally, in this section, we expand the analysis to construct a column-wise time dependant sensitivity matrix (each column of the matrix representing a different parameter). We define, the time dependant column sensitivity matrix:

$$S(t) = \frac{\partial X}{\partial C}(t) = \begin{pmatrix} \frac{\partial X}{\partial C}(t_1) \\ \frac{\partial X}{\partial C}(t_2) \\ \vdots \\ \frac{\partial X}{\partial C}(t_{n_T}) \end{pmatrix} \tag{6.9}$$

where, the sensitivity component for each time point $t_i \in \{t_1, t_2, ..., t_{n_T}\}$ is given by the matrix:

$$\frac{\partial X}{\partial c}(t_k) = \begin{pmatrix} \frac{\partial x_1}{\partial c_1}(t_k) & \frac{\partial x_1}{\partial c_2}(t_k) & \cdots & \frac{\partial x_1}{\partial c_M}(t_k) \\ \frac{\partial x_2}{\partial c_1}(t_k) & \ddots & & \frac{\partial x_2}{\partial c_M}(t_k) \\ \vdots & & \ddots & \vdots \\ \frac{\partial x_N}{\partial c_1}(t_k) & \frac{\partial x_N}{\partial c_2}(t_k) & \cdots & \frac{\partial x_N}{\partial c_M}(t_k) \end{pmatrix}. \tag{6.10}$$

In the above, each column is associated with a specific parameter c_j .In component form we can write the sensitivity matrix as:

$$s_{ij}(t_k) = \frac{\partial x_i}{\partial c_j}(t_k). \tag{6.11}$$

In the above, x_i represents the ith model output at time t_k. We define a relative sensitivity matrix as:

$$\overline{s}_{ij}(t_k) = \frac{\partial x_i}{\partial c_j}(t_k)\frac{c_j}{x_i}. \tag{6.12}$$

We seek a measure of sensitivity score that can be ordered with respect to the degree of influence of each parameter on the model output. For each column of $\overline{s}_{ij}(t_k)$ we define the

norm as:

$$\tilde{s}_j = \sqrt{\sum_{k=1}^{n_T} \sum_{i=1}^{n} (\overline{s}_{ij}(t_k))^2}. \tag{6.13}$$

The norm of each column represents the sensitivity score of the corresponding parameter.

6.5.2 Procedure for Determining the Identifiability Score

In the previous Section 6.5.1, the column relative sensitivity matrix was constructed to estimate the sensitivity score. In this section we introduce an algorithm (as originally proposed by Yao et al. [122]) to rank the model parameters based on their identifiability score. This approach uses the relative sensitivity column matrix.

The procedure to rank each parameter's influence on the model output is done using the orthogonalization procedure by Yao et al. [122]. The key idea of the procedure is to rank the identifiability score of each parameter (column of $\overline{s}_{ij}(t_k)$) based on how it correlates with other parameters. Parameters with a high degree of identifiability will have a low level of correlation with the other parameters.

The vector X_1 represents the column of $\overline{s}_{ij}(t_k)$ with the largest value of the norm $\tilde{s}_j$ and

can be written as:

$$X_1 = Maxcol\{\tilde{s}_{ij}(t_k)\} = Maxcol \begin{pmatrix} \overline{s}_{11}(t_1) & \cdots & \overline{s}_{1m}(t_1) \\ \vdots & \ddots & \vdots \\ \overline{s}_{n1}(t_1) & \cdots & \overline{s}_{nm}(t_1) \\ \vdots & \ddots & \vdots \\ \overline{s}_{11}(t_{n_T}) & \cdots & \overline{s}_{1m}(t_{n_T}) \\ \vdots & \ddots & \vdots \\ \overline{s}_{n1}(t_{n_T}) & \cdots & \overline{s}_{nm}(t_{n_T}) \end{pmatrix} \tag{6.14}$$

where, in the above, Maxcol represents the column of the matrix with the largest norm $\tilde{s}_j$. This also corresponds to the parameter c_j with the largest identifiability. The norm $\tilde{s}_j$ represents the identifiability score of the jth parameter. The projection of each other column j in the sensitivity matrix $\overline{s}_{ij}(t_k)$ onto direction orthogonal to X_1 is given by the residual matrix:

$$R_2 = \overline{s} - X_1 \left(\frac{X_1^T \cdot \overline{s}}{X_1^T \cdot X_1} \right). \tag{6.15}$$

The residual provides the measure of how uncorrelated each of the parameters is to X_1. That is, the column of $\overline{s}_{ij}(t_k)$ which correspond to the largest value of R_2 corresponds to the lowest correlation and therefore the highest identifiability among the remaining parameters. The column of R_2 with the largest norm therefore corresponds to the second highest identifiable parameter. The vector X_1 is then augmented with the column of $\overline{s}_{ij}(t_k)$ which corresponds to that parameter. This augmented matrix gives the matrix X_2. The

residual of the projection of each column of $\overline{s}_{ij}(t_k)$ onto X_2 is then given by the matrix:

$$R_3 = \overline{s} - X_2 \left(X_2^T \cdot X_2\right)^{-1} \cdot X_2^T \overline{s}. \tag{6.16}$$

The next most identifiable parameter corresponds to the column with the highest norm of R_3. The process is repeated until the identifiability score of each parameter is computed.

6.5.3 Estimation of Eigenvalues and Collinearity Index

In the current approach, the distribution is constructed from the algorithm for finding the sensitivity matrix many times using the MC approach. The expectation value of that distribution was used previously to rank model parameters based on their identifiability. In addition, the expectation value of the sensitivity matrix can be further used to provide additional insight into the identifiability of the model parameters by estimating the corresponding eigenvalues Komorowski et al. [68]. The number of the corresponding eigenvalues, which are non-zero (or above some threshold value) can be used to estimate the number of parameters that are identifiable.

In order to ensure that the eigenvalues are relatively meaningful, the sensitivity matrix must be renormalized. Recall that, in Section 6.5.1, the column relative sensitivity matrix $\overline{S}$ was constructed. First, the sensitivity matrix should be computed relatively to the other parameters and then normalized. The relative sensitivity ensures to distinguish large and small sensitivities. Normalization ensures that these are relative sensitivities, so that the choice of units does not play a role in determining their values. The normalization ensures

that the magnitudes of the sensitivities are standardized within a specified range of numbers (typically between 0 and 1).

The normalized sensitivity matrix is given by:

$$\hat{S}_{ij} = \frac{\overline{S}_{ij}}{\|\overline{S}_j\|}. \tag{6.17}$$

Where the vector

$$\|\overline{S}_j\| = \sqrt{(\overline{S}_{1j})^2 + (\overline{S}_{2j})^2 + ... + (\overline{S}_{Nj})^2} \tag{6.18}$$

represents the Euclidean norm of the jth column. Large value of the norm $\|\overline{S}_j\|$ of the jth column of the sensitivity matrix $\overline{S}_{ij}$ shows that, when all other parameters are fixed, a small change in parameter c_j will have a large impact on the overall model output. It suggests that the parameter may be highly identifiable, provided there isn't a problem with correlation among parameters.

In order to compute the eigenvalues, a square sensitivity matrix is needed, which is composed from the normalized sensitivity matrix and its transpose. The eigenvalues of the matrix $\hat{S}^T\hat{S}$ can be determined and provide a measure of the linear dependency between the sensitivity functions. An eigenvalue λ_k, which is close to zero, can be associated with low identifiability of model parameters. Therefore, a useful measure of identifiability is related to the minimum eigenvalue (the model identifiability is restricted by its smallest eigenvalue). The collinearity index is defined by:

$$\gamma_k = \sqrt{\frac{1}{\min \lambda_k}}. \tag{6.19}$$

That is, if the impacts on the model output, due to changing one parameter (or several parameters), results in the same model output impact by changing another set of parameters then the model is not identifiable with respect to those parameters. This is because the change to the output x_i, when varying a parameter c_j, can always give the same output by varying a linear combination of the other parameters. When the lowest eigenvalue is closer to zero, the model is less identifiable. That means, a higher collinearity index indicates a lower model parameter identifiability. Recommended critical threshold values for collinearity index γ_k lie in the range $\gamma_k \in [5, 20]$ [98].

6.5.4 Estimation of the Confidence Intervals

A key aim of this analysis is not just to rank the identifiability of the model parameters but also to estimate their values. Since the actual (true) value of the model parameters can usually not be estimated, we can construct confidence intervals in which the true value of each parameter is expected to occur within a specified confidence level. For instance, when choosing the significance level of $\alpha = 0.05$ we require the true value of the parameter to appear within the boundaries of the confidence interval 95% of the time.

In general, a few approach are available to construct such intervals. A key approach [4] relies on the construction of an absolute sensitivity matrix (which was described further in Sections 3.2.1 and 6.2) $s_{ij}(t_k) = \frac{\partial x_i}{\partial c_j}(t_k)$ as specified before (rather than, the relative

sensitivity matrix $\overline{s}_{ij}(t_k)$). If we assume that the experimental errors in the model are independent and normally distributed then we can define the least square error

$$SSE(c) = \sum_i \sum_k \frac{\left(y^i_{obs}(t_k) - y^i_{sim}(c, t_k)\right)^2}{\left(\sigma^i(t_k)\right)^2} \tag{6.20}$$

where, $y^i_{obs}(t_k)$ and $y^i_{sim}(c, t_k)$ represent the ith mean observed and simulated outputs respectively at time t_k and the standard deviation $\sigma^i(t_k)$ is associated with the observed output values. We can use the least square error to obtain the lower bound on the radius of the 95% confidence interval for the parameter:

$$\Delta_{c_j} = \frac{m}{n-m} SSE(c) \cdot F_{0.05}(m, n-m) \left(\sqrt{\left(s^T(c)s(c)\right)_{ij}}\right)^{-1} \tag{6.21}$$

where m and n correspond to the number of parameters and number of observations respectively. The function $F_{0.05}(m, n-m)$ represents the 95% inverse of cumulative Fisher distribution with m and $n-m$ degrees of freedom. The diagonal element of a matrix is defined as $(*)_{ij}$. Given an estimate of the parameter c, we expect to find the true value of the parameter c_j within the confidence interval $\left[c_j - \Delta_{c_j}, c_j + \Delta_{c_j}\right]$ 95% of the time. It is convenient to report the relative estimates of the parameters as $\left(\frac{\Delta_{c_j}}{\hat{c_j}} \times 100\%\right)$ instead of the absolute confidence intervals.

An additional approach for estimating the lower bound on the radius of the 95 % confidence interval for a parameter can be obtained [33] using the Fisher information matrix as follows:

$$\Delta_{c_j} = 1.96\sqrt{\left(FIM^{-1}\right)_{jj}}. \tag{6.22}$$

In the above, the measurement errors are assumed to be independent and normally distributed. The FIM can be constructed as follows:

$$FIM = s^T W s \tag{6.23}$$

where W represents the inverse of the measurement covariance matrix. The diagonal elements of W (the measurement variances) are calculated for each observation, while the off-diagonal terms (covariance) can be set to zero. As in the case of the previous confidence estimator, the relative values of the parameter estimates may be used.

6.6 Numerical Results

6.6.1 Constitutive Gene Expression Model

Our preliminary analysis is of a single gene expression model which represents gene transcription and gene translation (Figure 6.1).

The single gene expression model [86] consists of two species and four reactions. The reactions along with the propensities and parameter values are given in Table 6.1. Here X_i represents the number of molecular of species S_i. The production rate of mRNA was dependent upon k_r, and the translation rate of mRNA to protein corresponds to k_p. The mRNA and protein degradation rate was given by γ_r and γ_p respectively.

Given the kinetic parameters in Table 6.1 with three different initial conditions ($X_1(0)$, $X_2(0)$)=(5, 5), (40, 500), (100, 1000) the system was simulated on the time-interval $[0, 5]$.

Table 6.1: Constitutive gene expression model

R_j	Reaction	Propensity	Nominal rate constant
R_1	$\varnothing \xrightarrow{k_r} mRNA$	$a_1 = k_r$	$k_r = 20$
R_2	$mRNA \xrightarrow{k_p} Pro + mRNA$	$a_2 = k_p X_1$	$k_p = 10$
R_3	$mRNA \xrightarrow{\gamma_r} \varnothing$	$a_3 = \gamma_r X_2$	$\gamma_r = 1.2$
R_4	$Pro \xrightarrow{\gamma_p} \varnothing$	$a_1 = \gamma_p X_2$	$\gamma_p = 0.7$

In this analysis, the mRNA and protein abundance levels were being observed on the time-interval $[0, 5]$.

Three simulated experiments (since actual experimental measurement were not available) corresponding with the three different initial conditions were completed in triplicate on the time-interval $[0, 5]$. At ten time points in the time interval, the abundance of mRNA and protein in each triplicate experiment were recorded. Then for each experiment, the mean and variance at each time point were calculated.

Since, three experiments corresponding with three different initial conditions were conducted at ten different time points on the same time-interval $[0, 5]$ therefore, 60 observations were collected overall for use in the SSE calculations (which was described further in Section 6.5.4 in equation 6.20) and 60 measurement variances were used to construct the 60×60 diagonal inverse measurement covariance matrix to obtain the FIM.

In the simulation, we also collected 120 observations: one mean and one variance at each

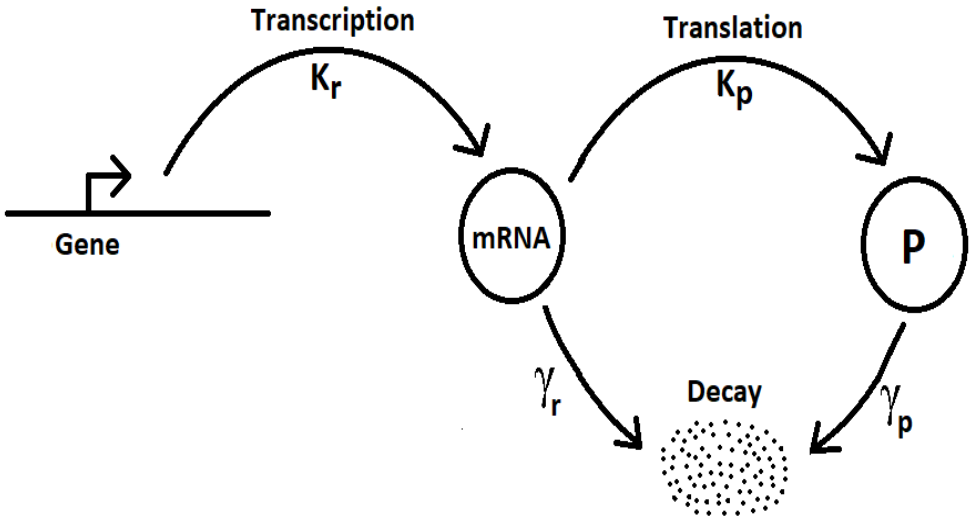
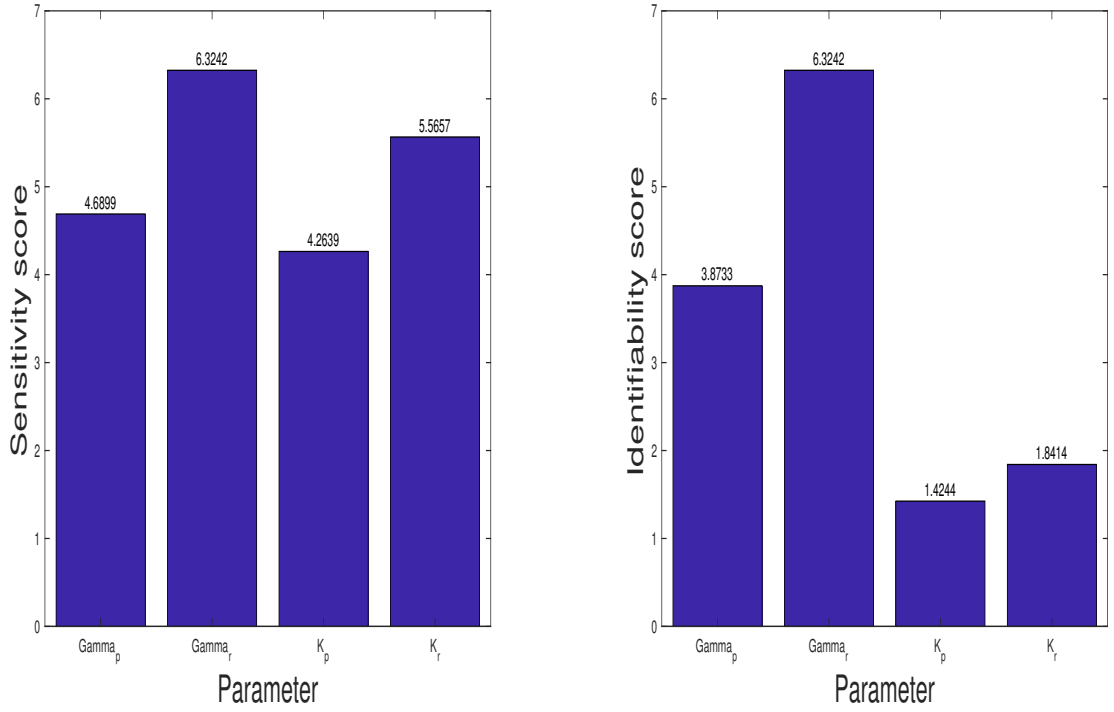


Figure 6.1: Constitutive gene expression model reaction scheme diagram.

observation, calculated over a large ensemble using the same initial conditions as in the three experiments. Data was collected at the same ten time points over the same time-interval, $[0, 5]$. Then the corresponding sensitivity, identifiability score and the associated confidence interval for each parameter are presented in Table 6.2. The corresponding sensitivity and identifiability score for each parameter are presented in Figure 6.2. The corresponding eigenvalues are 0.030, 0.113, 0.820, 3.039 and the collinearity index γ_k is 5.799.

The aim of the analysis is to compute the sensitivity of model output to each parameter, their identifiability score and to estimate the region within which the true values of the model parameters can be found 95% of the time. The results indicated the highest degree of identifiability for the γ_r parameter, meaning that the the model outputs are most sensitive with respect to this parameter. The next highest identifiability score is associated with



(a) (b)

Figure 6.2: Constitutive gene expression model (when considering both species observations). Ensembles of 10000 sample paths with initial condition $(X_1(0), X_2(0))$=(5, 5), (40, 500), (100, 1000) and parameters as in Table 6.1 were generated on the time-interval $[0, 5]$. (a-b) The sensitivity and identifiability score for each parameters.

Table 6.2: Constitutive gene expression model: Uncertainty analysis (when considering 10 different time points for both species' observations)

Parameter	Parameter value	Sensitivity score	Identifiability score	95% CI Δ	95% CI FIM
k_r	20	5.57	1.84	2.46%	13.11%
k_p	10	4.26	1.42	1.49%	10.73%
γ_r	1.2	6.32	6.32	1.70%	11.13%
γ_p	0.7	4.69	3.87	1.13%	10.38%

the γ_p parameter and represents the next highest identifiability score. The second highest sensitivity score is not associated with the γ_p parameter. This is due to the fact that the computation of the identifiability score has taken into account its correlation between the remaining parameters. The identifiability score for the remaining parameters are computed in the same way.

For each parameter, we were interested in computing the confidence interval. Since, for any model, we can never have a complete certainty about the exact values of the model parameters. A confidence interval was obtained around each given parameter value such that the true value of the parameter would be found somewhere within this interval 95% of the time and outside of this interval the remaining 5% of the time. Two methods were used to estimate the confidence interval. Both methods indicated that the true value of the parameters would be found within a small radius Δ of the given parameter values. The small confidence interval indicates a relatively large level of confidence in the given

parameters. That means that the provided parameters were well estimated, given that we do not expect for the true values of these parameters to be outside of their very small confidence intervals more than 5% of the time.

In order to ensure that the algorithm is behaving as expected, tests of the results in special cases were considered. This provides a simple but useful way to gain confidence that the use of this algorithm in more complex scenarios may be reasonable. For these reasons, if we consider only protein observations, the highest identifiability parameter was associated with the protein parameter γ_p, as expected. The analysis had shown a similarly small confidence intervals. Similarly when considering only mRNA observations, the highest identifiability parameter was associated with the mRNA parameter γ_r, again as expected. However, the analysis had now revealed large confidence intervals associated with the protein parameters. This means that the given parameter values are no longer good estimates of their true value. It was observed that the sensitivities of the protein parameters were essentially zero, and associated confidence intervals approaches infinity. This is due to the the fact that mRNA plays a significant role within the model (the mRNA parameter appears in 3 out of the four model equations).

If we consider the protein observations only, then we have in total 30 observations. Then the corresponding sensitivity, identifiability score and the associated confidence interval for each parameter are presented in Table 6.3. The corresponding sensitivity and identifiability score for each parameter are presented in Figure 6.3. The corresponding eigenvalues are 0.023, 0.037, 0.274, 3.666 and the collinearity index γ_k is 6.564.

Again, when considering the mRNA observations only, there are in total 30 observations.

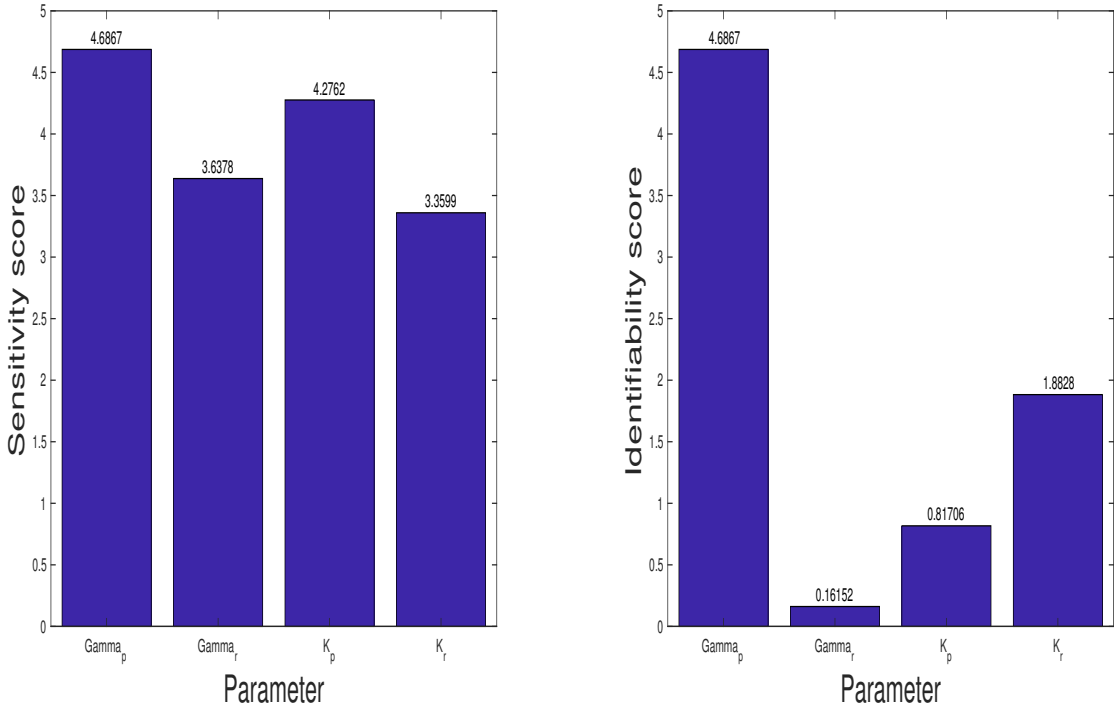


Figure 6.3: Constitutive gene expression model (when considering the protein observations only). Ensembles of 10000 sample paths with initial condition $(X_1(0), X_2(0))$=(5, 5), (40, 500), (100, 1000) and parameters as in Table 6.1 were generated on the time-interval $[0, 5]$. (a-b) The sensitivity and identifiability score for each parameters.

Table 6.3: Constitutive gene expression model: Uncertainty analysis (when considering 10 different time points for the protein observations only)

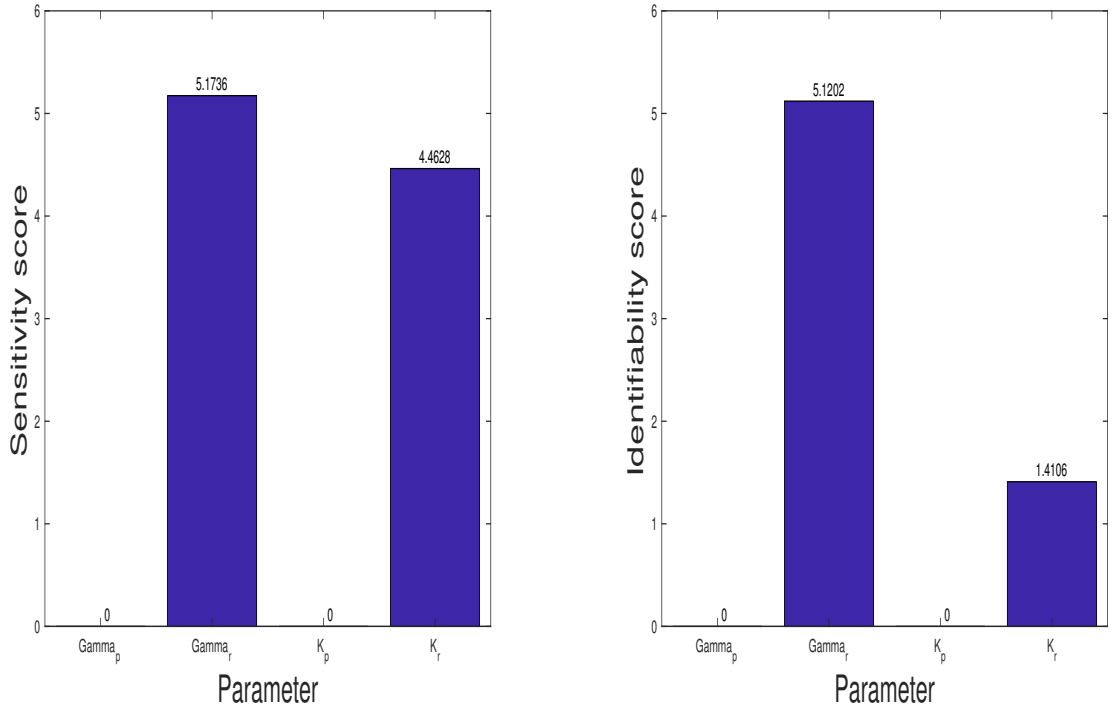
Parameter	Parameter value	Sensitivity score	Identifiability score	95% CI Δ	95% CI FIM
k_r	20	3.36	1.88	2.99%	10.48%
k_p	10	4.28	0.82	1.78%	16.36%
γ_r	1.2	3.64	0.16	2.05%	17.55%
γ_p	0.7	4.69	4.69	1.34%	6.81%

Furthermore, the corresponding sensitivity, identifiability score and associated confidence interval for each parameter are presented in Table 6.4. The corresponding sensitivity and identifiability score for each parameter are presented in Figure 6.4.

Table 6.4: Constitutive gene expression model: Uncertainty analysis (when considering 10 different time points for the mRNA observations only)

Parameter	Parameter value	Sensitivity score	Identifiability score	95% CI Δ	95% CI FIM
k_r	20	4.46	1.41	14.80%	8.36%
k_p	10	0	0	Inf %	Inf %
γ_r	1.2	5.17	5.17	10.52%	17.25%
γ_p	0.7	0	0	Inf %	Inf %

Another three simulated experiments corresponding with three different initial conditions



(a) (b)

Figure 6.4: Constitutive gene expression model (when considering the mRNA observations only). Ensembles of 10000 sample paths with initial condition ($X_1(0)$, $X_2(0)$)=(5, 5), (40, 500), (100, 1000) and parameters as in Table 6.1 were generated on the time-interval $[0, 5]$. (a-b) The sensitivity and identifiability score for each parameters.

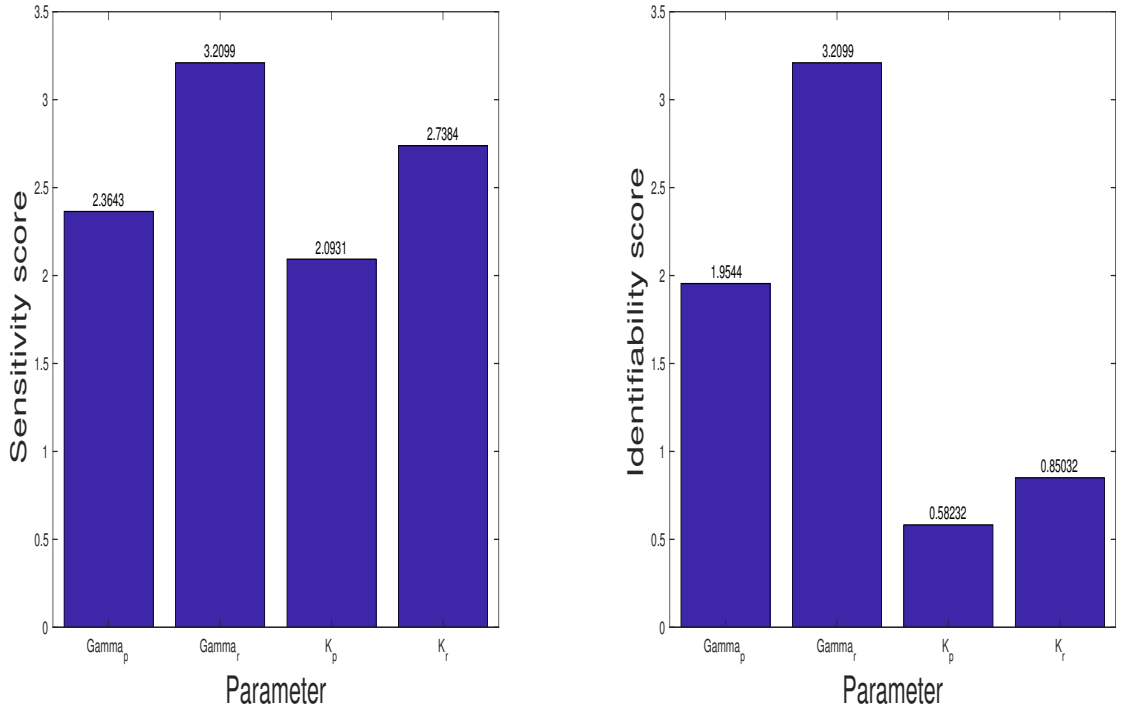
were completed in triplicate on the time-interval $[0, 4]$. This time, however, noise was added to the data in order to approximate real observational data (since actual experimental measurement were not available). The addition of noise to the simulated data resulted in a larger confidence interval than before. This was expected as the noise acted as a source of additional random error. At three time points in the time interval $[0, 4]$, the abundance of mRNA and protein in each triplicate experiment were recorded. Then for each experiment, the mean and variance at each time point were calculated. The corresponding sensitivity, identifiability score and the associated confidence interval for each parameter are presented in Table 6.5. The corresponding sensitivity and identifiability score for each parameter are presented in Figure 6.5. The corresponding eigenvalues are 0.0213, 0.095, 0.846, 3.038 and the collinearity index γ_k is 6.845.

This Constitutive Gene Expression Model was used to test the usefulness of the identifiability approach. This approach used to determine out of four parameters, which parameter plays the most vital role in the model. Simulations revealed that the mRNA degradation rate parameter given by γ_r represents the most identifiable parameter in the model.

The algorithms were simulated numerically using MATLAB (Mathworks) [82]. The MATLAB code will be available upon request after publication.

6.6.2 Lac Induction Model

Using flow cytometry experiments and computational analysis [86], a parameter set was identified to describe single-cell dynamics of green fluorescent protein (GFP) controlled by the lac operon under IPTG induction in vivo (Figure 6.6). The response of the system



(a) (b)

Figure 6.5: Constitutive gene expression model (when noise was added to the data). Ensembles of 10000 sample paths with initial condition $(X_1(0), X_2(0))$=(5, 5), (40, 500), (100, 1000) and parameters as in Table 6.1 were generated on the time-interval $[0, 4]$. (a-b) The sensitivity and identifiability score for each parameters.

Table 6.5: Constitutive gene expression model (when noise was added to the data): Uncertainty analysis (when considering 3 different time points for both species' observations)

Parameter	Parameter value	Sensitivity score	Identifiability score	95% CI Δ	95% CI FIM
k_r	20	2.74	0.85	30.52%	16.71%
k_p	10	2.09	0.58	19.05%	31.26%
γ_r	1.2	3.21	3.21	20.57%	13.81%
γ_p	0.7	2.36	1.95	13.80%	23.22%

at multiple time points and several IPTG levels was then explored. Focus was placed on the simplest consistent model of diffusion of IPTG into the cell and production and degradation of LacI and GFP. Diffusion of IPTG was given by

$$[IPTG]_{in} = [IPTG]_{out}(1 - exp(-rt))$$

where, r and t are the diffusion rate parameter and time respectively. Production and degradation of both LacI and GFP were given by four basic reactions, R_1, R_2, R_3 and R_4.

The Lac Induction model [86] consists of two species and four reactions. The reactions along with the propensities and parameter values are given in Table 6.6. Here X_i represents the number of molecular species S_i.

The production rate of LacI is a constant which corresponds to the constitutive expression C_1 , where $C_1 = K_L$. The degradation rate of LacI is dependant on IPTG concentration.

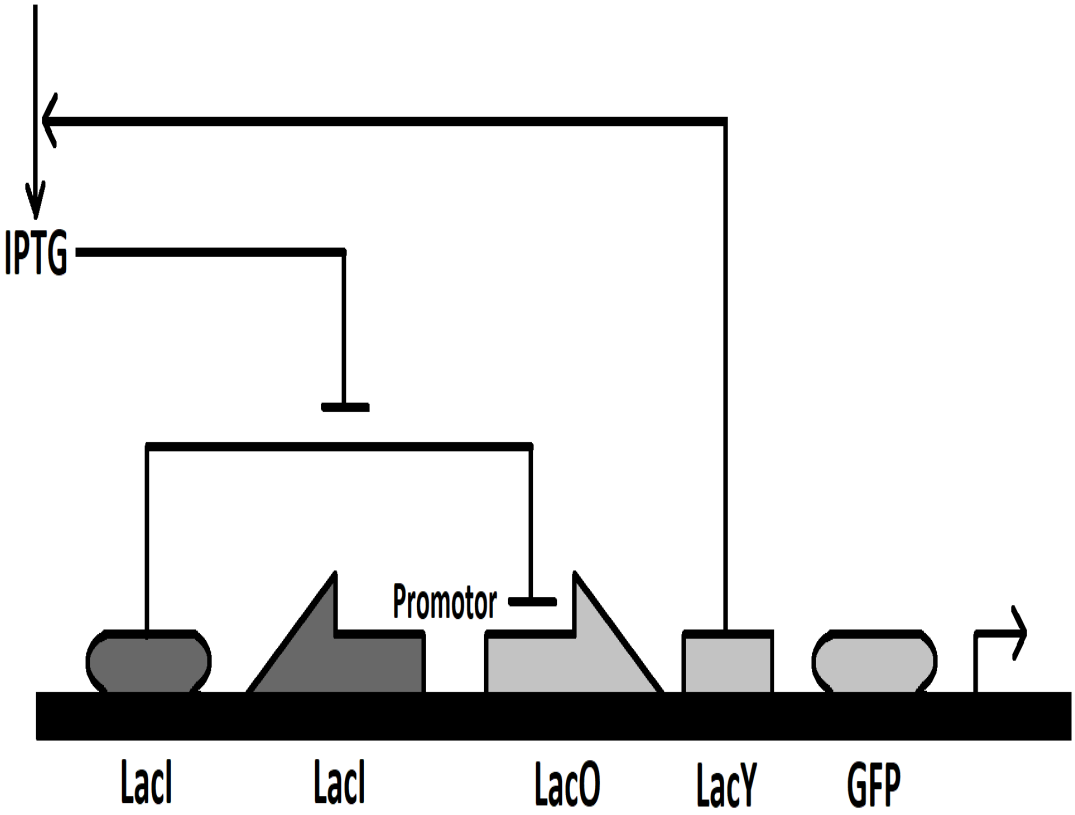


Figure 6.6: Schematic diagram of Lac induction model.

Table 6.6: Lac induction model

R_j	Reaction	Propensity	Nominal rate constant
R_1	$\varnothing \xrightarrow{C_1} LacI$	$a_1 = C_1 = k_L$	$k_L = 1.7 \times 10^{-3}$
R_2	$LacI \xrightarrow{C_2} \varnothing$	$a_2 = C_2 X_1 = \delta_L X_1$	$\delta_L^0 = 3.1 \times 10^{-4}$ $\delta_L^1 = 5.0 \times 10^{-2}$ $r = 2.8 \times 10^{-5}$
R_3	$\varnothing \xrightarrow{C_3} GFP$	$a_3 = C_3 = \frac{K_G}{1+\alpha[X_1]^\eta}$	$K_G = 1.0 \times 10^{-1}$ $\alpha = 1.3 \times 10^4$ $\eta = 2.1$
R_4	$GFP \xrightarrow{C_4} \varnothing$	$a_1 = C_4 X_2 = \delta_G X_2$	$\delta_G = 3.8 \times 10^{-4}$

This rate is assumed to have form

$$C_2 = \delta_L \times [LacI],$$

where δ_L depends upon positive real parameters for the regulatory system, δ_L^0 and δ_L^1 such that,

$$C_2 = [\delta_L^0 + \delta_L^1 [IPTG]_{in}] \times [LacI] = [\delta_L^0 + \delta_L^1 [[IPTG]_{out} \times (1 - exp(-rt))] \times [LacI].$$

The production rate of GFP is a nonlinear function dependent on the concentration of LacI, given by

$$C_3([LacI]) = \frac{K_G}{1 + \alpha[LacI]^\eta}$$

where, K_G, α and η are positive real parameters of the unrepressed GFP production rate, the LacI occupancy strength and the Hill coefficient respectively. The Hill coefficient accounts for the cooperative binding of LacI. The GFP degradation rate is fixed to the concentration of GFP, given by

$$C_4 = \delta_G \times [GFP].$$

Given the kinetic parameters in Table 6.6 with initial conditions $(X_1(0), X_2(0)) = (500, 500)$, the system was simulated on the time-interval $[0, 5]$ hrs. In this analysis, the GFP abundance levels are being observed at several time points (0 hours, 3 hours, 4 and 5 hours). The results are presented for the amount of GFP with varying level of extracellular IPTG induction (5, 10, 20, 40 and 100 μM).

In this Lac induction model, Munsky et al. [86] conducted an experiment of GFP expression on two different days and collected data at different measurement times (t=0,3,4,5 hours) after induction. In this experiment different levels of extracellular IPTG induction (5,10,20, 40 and 100μM) were used. For the purpose of the current analysis, the data was collected from the published histogram figures [86]. Web Plot Digitizer software was used to collect the data from the publication and used directly with the current analysis. The identifiability analysis revealed that out of eight parameters, two parameters given by α and η had shown zero sensitivity and therefore zero identifiability. The analysis also

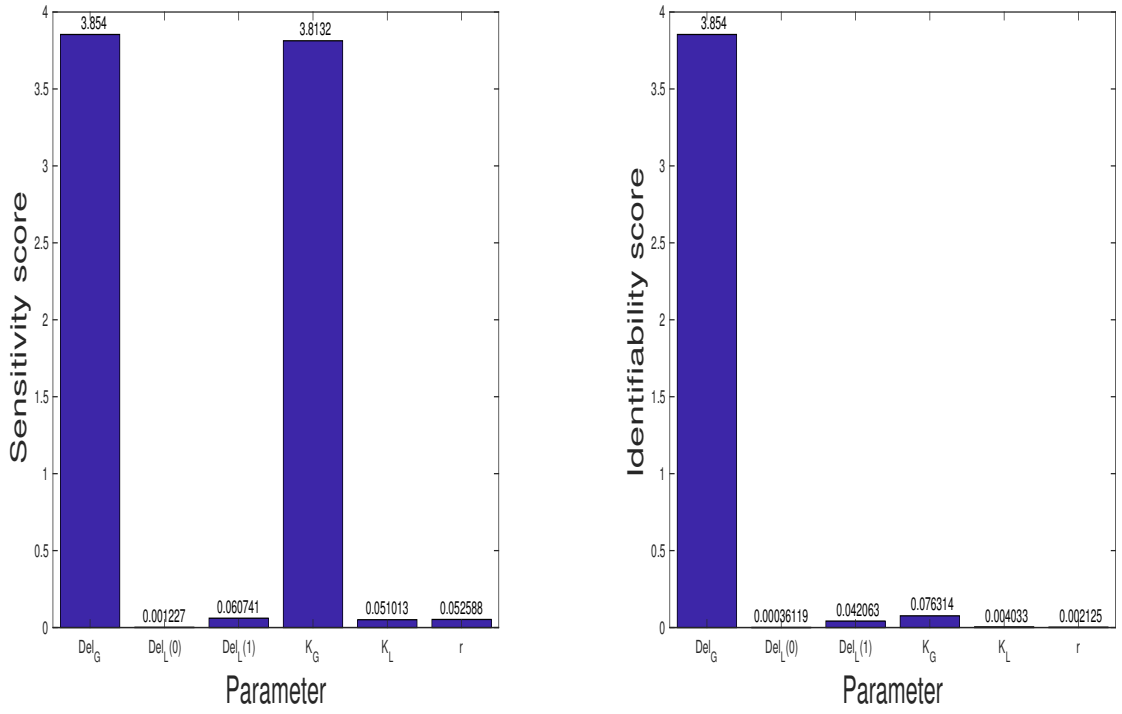


Figure 6.7: Lac induction model. Ensembles of 10000 sample paths with initial condition $(X_1(0), X_2(0)) = (500, 500)$ and parameters as in Table 6.6 were generated on the time-interval $[0, 5]$hrs. (a-b) The sensitivity and identifiability score for each parameters.

Table 6.7: Lac induction model: Uncertainty analysis

Parameter	Parameter value	Sensitivity score	Identifiability score	95% CI Δ	95% CI FIM
k_L	1.7×10^{-3}	0.05	0.004	74708.13%	7080.46%
δ_L^0	3.1×10^{-4}	0.001	0.0004	2767330.27%	120457.92%
δ_L^1	5.0×10^{-2}	0.06	0.04	65161.77%	11047.47%
r	2.8×10^{-5}	0.05	0.002	75195.38%	10894.80%
k_G	1.0×10^{-1}	3.81	0.08	1021.69%	688.66%
δ_G	3.8×10^{-4}	3.85	3.85	1010.35%	677.42%
α	1.3×10^{4}	0	0	Inf %	Inf %
η	2.1	0	0	Inf %	Inf %

revealed that only two parameters δ_G and k_G had shown high levels of identifiability. The remaining parameters had shown low levels of identifiability. The corresponding sensitivity, identifiability score and the associated confidence interval for each parameter are presented in Table 6.7. The corresponding sensitivity and identifiability score for each parameter are presented in Figure 6.7. The corresponding non-zero (6 out of 8 parameters) eigenvalues are 0.00005, 0.0005, 0.005, 0.14, 0.98, 4.87 and the collinearity index γ_k is 141.62.

The results reveal that parameters with high level of identifiability have relatively smaller confidence intervals. On the other hand, the parameters with very low levels of identifiability show confidence intervals which tend to infinity. The direct collecting of data from

the plots in [86] likely resulted in large data errors and inaccuracies. The result of this error could be a key contributor to the large confidence intervals seen in the results of the confidence intervals of the δ_G and k_G parameters.

Chapter 7

Conclusions

There are various mathematical models that can be used to describe the dynamics of a biochemical system. Depending on the essential features present in the system, different models may be more appropriate. Stochastic models must be used to capture the random fluctuations observed in these systems. The presence of noise in a system can be a significant factor in determining its behavior. The Chemical Master Equation is a valuable stochastic model of biochemical kinetics. Solutions to the CME can be probabilistically simulated using the exact stochastic simulation algorithm (SSA) but it is computationally expensive. Tau-leaping methods can be used in order to speed-up the simulation of biochemical systems. More sophisticated techniques are necessary for dealing with systems which manifest stiffness.

As an important mathematical tool, sensitivity analysis can serve as a foundation for the formulation, characterization, and verification of models. Sensitivity analysis is used to

identify important reaction rate parameters that are essential to a system's dynamics. A number of approaches to sensitivity analysis of stochastic discrete models of biochemical kinetics have been developed [2, 46, 96]. We discussed the finite-difference based numerical approaches of sensitivity analysis for the stochastic model of well-stirred biochemical systems and made comparisons among these methods in Chapter 3. We concluded that the CFD algorithm performs better in determining sensitivity for non-stiff biochemical systems.

We also presented an application of adaptive tau-leaping to sensitivity analysis in Chapter 4. Our proposed finite-difference based method for estimating sensitivity for stochastic models of biochemical systems, named Coupled Tau-Leaping (CTL) [83], produces the nominal and perturbed trajectories with strong coupling. Our analysis showed that among finite-difference sensitivity estimators the Coupled Finite Difference (CFD) method proposed by Anderson [2] provides higher accuracy for estimating the sensitivity of a biochemical system. That is why we compared our novel method to the Coupled Finite Difference method. Numerical tests showed that our novel algorithm is significantly more efficient than the Coupled Finite Difference (CFD) algorithm while producing sensitivity estimators that are of similar accuracy. These results showed that our Couple Tau-Leaping (CTL) method outperforms the CFD method when applied to moderately stiff stochastic models of biochemical networks with molecular populations bounded away from zero.

Implicit tau-leaping schemes are preferred over exact Monte Carlo simulations for stochastic models of biochemical systems that are mathematically stiff. The implicit methods are more efficient for accurately determining the slow variables and the mean behaviour of the fast variables of the system. When stiffness is encountered, the Monte Carlo stochastic

simulations have to take very small step-sizes, whereas the implicit tau-leaping method can take large step-sizes and maintains the solution close to the slow manifold.

We also proposed another finite-difference based method for estimating sensitivity for stochastic discrete models of biochemical kinetics named Coupled Implicit Tau-Leaping (CIT) [84] in Chapter 5. This method uses the adaptive implicit τ-leaping strategy to simulate the nominal and perturbed trajectories. Our CIT algorithm produces a strong coupling between the nominal and the perturbed paths to enhance the accuracy of the estimation. Our numerical tests showed that the novel CIT method greatly reduced the computational cost when compared to the CFD while maintaining similar accuracy. Therefore, the CIT method is a better choice than the CFD method for estimating the sensitivity of stochastic models of biochemical reaction networks that are considerably stiff to very stiff.

When designing a model of a physical system, it is very important to identify all of the relevant model parameters and find their mathematical relationship to all of the observations. Identifiability analysis provides a clear way of determining which parameters in the model are well estimated. Identifiability also provides a way to assess the quality of the estimated parameter values. A result with a high level of identifiability provides confidence that the model behaviour is relatively reliable. In any physical models the confidence intervals provide us with a way of estimating the region (in parameter space) where the true values of the model parameters lie. For this reason, identifiability plays an essential role in model parametrization. Lastly, we presented an identifiability approach for discrete stochastic models of biochemical systems by using the sensitivity matrix in Chapter 6. This approach was used as an identifiability tool to assess the quality of estimates.

In the future, there are several directions in which we would like to extend our research. We are interested in finding a single sensitivity technique which can be applied to both stiff and non-stiff biochemical systems. This method will ideally be computationally more accurate and efficient than existing methods.

There are many approaches in the literature for determining identifiability of a model. There is currently no single method or technique that can be used to determine identifiability for every model. We are interesting in extending our techniques to finding identifiability of stochastic models of biochemical kinetics with higher moments.

References

[1] Anderson, D. F., (2007). A modified next reaction method for simulating chemical systems with time dependent propensities and delays, *J. Chem. Phys.* 214107.

[2] Anderson, D.F., (2012). An efficient finite difference method for parameter sensitivities of continuous time Markov chains, *SIAM J. Numer. Anal.* 50, 2237.

[3] Anderson, D.F., Higham, D.J., (2012). Multilevel Monte Carlo for continuous time Markov chains, with applications in biochemical kinetics, *SIAM: Multiscale Modeling and Simulation* 10(1), 146 –179.

[4] Ashyraliyev M., Fomekong-Nanfack Y., Kaandorp J.A., Blom J.G., (2009). Systems biology: parameter estimation for biochemical models. *FEBS J.*, 276(4):886-902.

[5] Arkin, A.P. , Plyasunov, P., (2007). Efficient stochastic sensitivity analysis of discrete event systems. *Journal of Computational Physics*, 221:724–738.

[6] Arkin, A.P., Ross, J., McAdams, H.H., (1998). Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage-infected Escherichia coli Cells , *Genetics*, Vol. 149, pp. 1633–1648.

[7] Ascher, UM., Petzold, LR., (1998). Computer Methods for Ordinary Differential Equations and Differential Algebraic Equations. *Philadelphia: Soc. Ind. Appl. Math (SIAM).*

[8] Burrage, K., Tian, T., Burrage, P., (2004). A multi-scaled approach for simulating chemical reaction systems. *Prog. Biophys. Mol.* Biol. 85, 217-234.

[9] Blake, W.J., Kaern, M., Cantor, C.R., Collins, J.J. (2003). Noise in eukaryotic gene expression, *Nature* 422, 633-637.

[10] Berry, H., (2002). Monte Carlo simulations of enzyme reactions in two dimensions: Fractal kinetics and spatial segregation.*Biophys. J.* 83, 1891-1901.

[11] Burrage, K., Hegland, M., MacNamara, S. and Sidje, R. B., (2006). *Proceedings of the Markov 150th Anniversary Conference*, edited by A. N. Langville and W. J. Stewart (Boson Books, Raleigh, NC), 21, 721–738.

[12] Crampin, E.J., Schnell, S., (2004). New approaches to modelling and analysis of biochemical reactions, pathways and networks, *Progress in Biophysics & Molecular Biology* 86, 1174.

[13] Cao, Y., Gillespie, D.T., Petzold, L.R., (2005). Avoiding negative populations in explicit Poisson tau-leaping. *J. Chem. Phys.*123:054104.

[14] Cao, Y., Gillespie D.T., Petzold L.R., (2006). Efficient stepsize selection for the tau-leaping simulation method. *J. Chem. Phys.*124:044109.

[15] Y. Cao, D.T. Gillespie, L. Petzold, (2007). Adaptive explicit-implicit for the tau-leaping with automatic tau-selection, *J. Chem. Phys.* 126, 224101.

[16] Chis, O., Banga, J., Balsa, E., (2011). Structural Identifiability of Systems Biology Models: A Critical Comparison of Methods. *PLoS ONE*, Volume 6, Issue 11, e27755.

[17] Chatterjee, A.,, Vlachos, D., Katsoulakis, M., (2005). Binomial distribution based τ leap accelerated stochastic simulation. *J. Chem. Phys.* 122:024112.

[18] Cox, B.G., (1994). Modern Liquid Phase Kinetics. Oxford University Press, Oxford.

[19] Clegg, J.S., (1984). Properties and metabolism of the aqueous cytoplasm and its boundaries. *Am. J. Physiol.* 246, R133-R151.

[20] Christoph Z., Sven, S., Jurgen, P: (2014). Exploiting intrinsic fluctuations to identify model parameters, *IET Syst. Biol.*, Vol. 9, Iss. 2, pp. 64-73.

[21] Calef, D.F., Deutch, J.M., (1983). Diffusion-controlled reactions. *Ann. Rev. Phys. Chem.* 34, 493-524.

[22] Cao, Y., Petzold, LR., (2005). Trapezoidal tau-leaping formula for the stochastic simulation of chemically reacting systems. *Proc. Found. Syst. Biol. Eng.* (FOSBE 2005), pp. 149-52.

[23] Cao, Y., Gillespie, D.T., and Petzold, L., (2005). The Slow-scale Stochastic Simulation Algorithm, *J. Chem. Phys.*, Vol. 122, pp. 01411601-01411618.

[24] Degasperi, A. and Gilmore, S. (2008). Sensitivity analysis of stochastic models of bistable biochemical reactions. In *International School on Formal Methods for the*

Design of Computer, Communication and Software Systems (1–20). Springer Berlin Heidelberg.

[25] Ethier, S. N. and Kurtz, T. G., (1986). *Markov Processes: Characterization and Convergence*, Wiley, New York.

[26] Epstein, I.R., Pojman, J.A., (1998). An introduction to nonlinear chemical dynamics: oscillations, waves, patterns, and chaos. Oxford University Press, Oxford.

[27] Espenson, J.H., (1995). Chemical Kinetics and Reaction Mechanisms. McGraw-Hill, Singapore.

[28] Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S., (2002). Stochastic Gene Expression in a Single Cell, *Science*, Vol. 297, pp. 1183–1186.

[29] Elowitz, M.B., Leibler, S., (2000). A Synthetic Oscillatory Network of Transcriptional Regulators, *Nature*, Vol. 403, pp. 335–338.

[30] Federoff, N. and Fontana, W., (2002). Small numbers of big molecules, *Science* **297**, 1129–1131.

[31] Field, R.J. and Noyes, R.M. (1974). Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction, *J. Chem. Phys.* 60: 18771–71884.

[32] Glasserman, P., (2003). Monte Carlo methods in nancial engineering. Springer, USA.

[33] Gadkar, K.G., Gunawan, R., and Doyle III, F.J.: (2005). Iterative approach to model identification of biological networks, *BMC Bioinformatics,* 6: 155.

[34] Gardiner, C. W., (2009). *Stochastic Methods: a Handbook for the Natural and Social Sciences*, Springer, Berlin.

[35] Gibson, M. A. & Bruck, J. (2000). Efficient exact stochastic simulation of chemical systems with many species and many channels. *J.Phys. Chem.* 105, 1876-1889.

[36] Gillespie, D.T., Mangel, M., (1981). Conditioned averages in chemical-kinetics. *J. Chem. Phys.* 75, 704-709.

[37] Gillespie, D.T., (2000). The Chemical Langevin Equation, *J. Chem. Phys.*, Vol. 113, pp. 297–306.

[38] Gillespie, D.T., (1992). A rigorous derivation of the Chemical Master Equation, *Physica A*, Vol. 188, 402–425.

[39] Gillespie, D.T., (1992). *Markov processes, an introduction for physical scientists*, Academic Press, INC. New York.

[40] Gillespie, D.T., (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, *J. Comp. Phys.*, Vol. 22, 403–434.

[41] Gillespie, D.T., (1977). Exact stochastic simulation of coupled chemical reactions, *J. Phys. Chem.*, Vol. 81, 2340–2361.

[42] Gillespie, D.T., (2001). Approximate accelerated stochastic simulation of chemically reacting systems, *J. Chem. Phys.*, Vol. 115, 1716–1733.

[43] Gillespie, D.T., (1976). A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions, *J. Comp. Phys.*, Vol. 22, pp. 403–434.

[44] Gillespie, DT, Petzold LR. (2003). Improved leap-size selection for accelerated stochastic simulation. *J. Chem. Phys* . 119:8229-34.

[45] Gregory J. McRae, James W. Tilden, John H. Seinfeld: (1982). Global sensitivity analysis: a computational implementation of the Fourier Amplitude Sensitivity Test (FAST), *Comp. Chem. zeng.*, 6, 15.

[46] Gunawan, R., Cao, Y., Petzold, L., and Doyle, F. J. (2005). Sensitivity analysis of discrete stochastic systems. *Biophysical Journal*, 88(4), 2530–2540.

[47] Higham, D.J., (2008). Modeling & simulating chemical reactions, *SIAM Rev.* 50(2), pp-347-368.

[48] Hisashi K., Brian L. M., William T., (2011). Probability, Random Processes, and Statistical Analysis: Applications to Communications, Signal Processing, Queueing Theory and Mathematical Finance, Cambridge University Press, Technology & Engineering.

[49] Higham, D. J, (2001). An algorithmic introduction to numerical simulation of stochastic differential equations, *SIAM Rev.*, 43(3), pp.525-546.

[50] Heinrich, R., Schuster, S., (1996). The Regulation of Cellular Systems. Chapman & Hall, New York.

[51] Halling, P.J., (1989). Do the laws of chemistry apply to living cells, *Trends Biochem. Sci.* 14, 317-318.

[52] Hu, Y. and Li, T., (2009). Highly accurate tau-leaping methods with random corrections, *J. Chem. Phys.*, Vol. 130,124109.

[53] Ingalls, B., (2004). Autonomously oscillating biochemical systems: parametric sensitivity of extrema and period. *IET Systems Biology* 1: 62–70.

[54] Ingalls, B. P., (2013). Mathematical Modeling in Systems Biology: an introduction, MIT Press, Cambridge, Massachusetts.

[55] Ilie, S., Morshed, M., (2013). Automatic Simulation of the Chemical Langevin Equation, *Applied Mathematics*, Special Issue on Numerical Analysis, Ed. Chris Cannings, Vol.4 No. 1A, 235-241.

[56] Ilie, S., Morshed, M., (2015). Adaptive time-stepping using control theory for the Chemical Langevin Equation, Journal of Applied Mathematics, Vol.2015, Article ID 567275.

[57] Ilie, S., and Teslya, A., (2012). An Adaptive Stepsize Method for the Chemical Langevin Equation, *J. Chem. Phys.*, Vol. 136, pp. 184101-184115.

[58] Ilie, S., (2012). Variable Time-stepping in the Pathwise Numerical Solution of the Chemical Langevin Equation, *J. Chem. Phys.*, Vol. 137, pp. 234110-234119.

[59] Ilie, S., Enright, W.H., and Jackson, K.R., (2009). Numerical Solution of Stochastic Models of Biochemical Kinetics, *Canadian Applied Mathematics Quarterly*, Vol. 17, No. 3, pp. 523 – 554.

[60] Jahnke, T., (2010). An adaptive wavelet method for the CME, *SIAM J. Scient. Comput.* 31.

[61] Jong, H. de, (2002). Modeling and simulation of genetic regulatory systems, *J. Comput. Biol.* 9(1), 67-103.

[62] Kaern, M., Elston, T.R., Blake, W.J., & Collins, J.J., (2005). Stochasticity in gene expression: from theories to phenotypes,*Nat. Rev. Genet.* 6: 451464.

[63] Kitano, H., (2002). Computational Systems Biology, *Nature*, Vol. 420, pp. 206-210.

[64] Kuthan, H., (2001). Self-organisation and orderly processes by individual protein complexes in the bacterial cell. *Prog. Biophys. Mol. Biol.* 75, 1-17.

[65] Kerker, M., (1974). Brownian movements and molecular reality prior to 1900. *J. Chem. Educ.* 51, 764-768.

[66] Kopelman, R., (1986). Rate-processes on fractals: Theory, simulations, and experiments. *J. Stat. Phys.* 42, 185-200.

[67] Kloeden, P.E., and Platen, E., (1992). Numerical Solution of Stochastic Differential Equations, Springer-Verlag, Berlin.

[68] Komorowski, M., Costa, M.J., Rand, D.A., Stumpf, M.P.H. and Halliday, K., (2011). Sensitivity, robustness, and identifiability in stochastic chemical kinetics models, *Proceedings of the National Academy of Sciences of the United States of America - PNAS*, vol 108, no. 21, pp. 8645-50.

[69] Kurtz, T.G., (1982). *Representation and approximation of counting processes*, in *Advances in Filtering and Optimal Stochastic Control*, Lecture Notes in Control and Inform. Science **42**, Springer, Berlin, 177 - 191.

[70] Kurtz, T. G., (1981). Approximation of Population Processes, CBMS-NSF Regional Conf. Ser. in Appl. Math. 36, SIAM, Philadelphia, 1981.

[71] Lesley T. MacNeil, Albertha J.M. Walhout, (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression, *Genome Res.*, 21(5): 645-657.

[72] Li, Y., H. & Petzold, L., (2004). Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *J. Chem. Phys.* 121, 4059-4067.

[73] Li, T., (2007). Analysis of Explicit Tau-leaping Schemes for Simulating Chemically Reacting Systems, *SIAM Multi. Model. Simul.*, Vol. 6, pp. 417- 436.

[74] Leimkuhler, B. and Reich, S. (2005). Simulating Hamiltonian Dynamics, Cambridge University Press, Cambridge, UK.

[75] Luby-Phelps, K., Castle, P.E., Taylor, D.L., Lanni, F., (1987). Hindered diffusion of inert tracer particles in the cytoplasm of mouse 3T3 cells. *Proc. Natl. Acad. Sci. U.S.A.* 84, 4910-4913.

[76] Lu, Haokai., Li, Peng., (2012). Stochastic projective methods for simulating stiff chemical reacting systems, *Computer Physics Communications* 183 1427–1442.

[77] McAdams, H.H., Arkin, A., (1997). Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 94, 814-819.

[78] Maheshri N, O'Shea E.K, (2007). Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu. Rev. Biophys. Biomol. Struct.* 36: 413434.

[79] Morton-Firth, C.J., (1998). Stochastic simulation of cell signalling pathways. Ph.D. thesis, University of Cambridge.

[80] Morton-Firth, C.J., Bray, D., (1998). Predicting temporal fluctuations in an intracellular signalling pathway. *J. Theor. Biol.* 192, 117-128.

[81] Minton, A.P., (1993). Molecular crowding and molecular recognition. *J. Mol. Recognit.* 6, 211-214.

[82] MATLAB, The Language of Technical Computing, www.mathworks.com.

[83] Morshed, M., Ingalls, B. and Ilie, S., (2017). An efficient finite-difference strategy for sensitivity analysis of stochastic models of biochemical systems, *Biosystems* 151, pg. 43-52.

[84] Morshed, M., Ingalls, B. and Ilie, S., (2017). An effective implicit finite-difference method for sensitivity analysis of stiff stochastic discrete biochemical systems, (in press, *IET Systems Biology*, Dec., 2017).

[85] Meng T. C., Somani, S., and Dhar, P., (2004). Modeling and simulation of biological systems with stochasticity. *In Silico Biology*, 4:293-309.

[86] Munsky, B., Trinth, B., Khammash, M., (2009). Listening to the noise: Random fluctuations reveal gene network parameters, *Mol. Syst. Biol.* 5, Article ID 318.

[87] Mukund, T., Alexander, V. O., (2001). Intrinsic noise in gene regulatory networks, *PNAS*, vol. 98, no. 15, pp. 8614-8619.

[88] M. Bennett, D. Volfson, L. Tsimring, J. Hasty, (2007). Transient dynamics of genetic regulatory networks, *Biophysical Journal*, 92 (10) 3501 – 3512.

[89] Oana-Teodora Chis, Julio R. Banga, and Eva Balsa-Canto, (2011). Structural Identifiability of Systems Biology Models: A Critical Comparison of Methods, *PLoS One.*, 6(11): e27755.

[90] Peter S. Swain, Michael B. Elowitz, and Eric D. Siggia, (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression, *PNAS*, vol. 99, no. 20, pp. 12795-12800.

[91] Press W, Flannery B, Teukolsky S, Vetterling W., (1986). *Numerical Recipes: The Art of Scientific Computing.* Cambridge, UK: Cambridge Univ. Press.

[92] Qian, H., Elson, E.L., (2002). Single-molecule enzymology: stochastic Michaelis-Menten kinetics. *Biophys. Chem.* 101-102, 565-576.

[93] Rathinam, M., Petzold, L.R., Cao, Y., Gillespie, D.T., (2003). Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method, *J Chem. Phys.*, Vol. 119, Nov. 24, 12784-12794.

[94] Rathinam, M., Cao, Y., Petzold, L.R. & Gillespie, D.T., (2005). Consistency and stability of tau-leaping schemes for chemical reaction systems. *SIAM Multiscale Modeling & Simulation.* Vol. 4, No. 3, pp. 867-895.

[95] Rao, C.V., and Arkin, A.P., (2003). Stochastic Chemical Kinetics and the Quasi-steady-state Assumption: Application to the Gillespie Algorithm, *J. Chem. Phys.*, Vol. 118, pp. 4999–5010.

[96] Rathinam, M., Sheppard, P. W., and Khammash, M., (2010). Efficient computation of parameter sensitivities of discrete stochastic chemical reaction networks, *J Chem. Phys.* 132, 034103.

[97] Rosenwasser, E. and Yusupov, R., (2002). *Sensitivity of automatic control systems.* CRC Press, Boca Raton.

[98] Roland Brun, Peter Reichert, Hans R. Knsch., (2001). Practical identifiability analysis of large environmental simulation models, *Water Resources Research,* vol. 37, issue 4, p.1015-1030.

[99] Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmuller, U., Timmer, J., (2009). Stuctural and practical indentifiability analysis of partially observed dynamical models by exploiting the profile likelihood, *Bioinformatics*, Vol. 35, Issue 15, PP-1923-1929.

[100] Ruano, M.V., Ribes, J., De, Pauw, D.J., Sin, G., (2006). Parameter subset selection for the dynamic calibration of activated sludge models (ASMs): experience versus systems analysis, *Water Sci Technol.* 56(8): 107-15.

[101] Scalettar, B.A., Abney, J.R., Hackenbrock, C.R., (1991). Dynamics, structure, and functions are coupled in the mitrocondrial matrix. *Proc. Natl. Acad. Sci. U.S.A.* 88, 8057-8061.

[102] Schnell, S., Turner, T.E., (2004). Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. *Prog. Biophys. Mol. Biol.* 85, 235-260.

[103] Srere, P., Jones, M.E., Mathews, C., (1989). *Structural and Organizational Aspects of Metabolic Regulation.* Alan R. Liss, New York.

[104] Stamatakis, M., and Mantzaris, N.V., (2009). Comparison of Deterministic and Stochastic Models of the lac Operon Genetic Network. *Biophys J.*, Vol. 96(3): pp. 887–906.

[105] Samant, A., and Vlachos, D., (2005). Overcoming Stiffness in Stochastic Simulation Stemming from Partial Equilibrium: a Multiscale Monte-Carlo Algorithm, *J. Chem. Phys.*, Vol. 123, pp. 144114-144122.

[106] Schlick, T., (2002). *Molecular Modeling and Simulation,* Springer-Verlag, Berlin.

[107] Saltelli, A., Ratto., M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S., (2009). *Global Sensitivity Analysis: The Primer*, Wiley, New York.

[108] Sheppard, P. W., Rathinam, M., and Khammash, M., (2012). A pathwise derivative approach to the computation of parameter sensitivities in discrete stochastic chemical systems. *J. Chem. Phys.*, 136:034115.

[109] Srivastava, R., Anderson, D.F., and Rawlings, J.B., (2013). Comparison of finite difference based methods to obtain sensitivities of stochastic chemical kinetic models. *J. Chem. Phys.*, 138:074110.

[110] Sanft, K. R., Gillespie, D. T., and Petzold, L. R., (2011). Legitimacy of the stochastic Michaelis-Menten approximation. *IET Systems Biology*, 5(1), 58–69.

[111] Tönsing, C., Timmer, J. and Kreutz, C., (2014). Cause and cure of sloppiness in ordinary differential equation models. *Physical Review E*, 90(2), 023303.

[112] Tian, T., Burrage, K., (2004). Binomial leap methods for simulating stochastic chemical kinetics. *J. Chem. Phys.* 121:10356-64.

[113] Theodore J. Perkins., (2009). Maximum likelihood trajectories for continuous-time markov chains. Ottawa Hospital Research Institute.

[114] Tian, T., Burrage, K., (2004b). Bistability and switching in the lysis/lysogeny genetic regulatory network of bacteriophage. *J. Theor. Biol.* 227, 229 - 237.

[115] Van Kampen, N. G., (1992). *Stochastic processes in physics and chemistry* (Vol. 1). Elsevier.

[116] Varma, A., Morbidelli, M. and Wu, H., (1999). *Parametric Sensitivity in Chemical Systems*, Cambridge University Press, Cambridge, UK.

[117] Varma, A., Morbidelli, M., (1997). Mathematical Methods in Chemical Engineering, New York: Oxford University Press.

[118] Verkman, A.S., (2002). Solute and macromolecule diffusion in cellular aqueous compartments. *Trends Biochem. Sci.* 27, 27-33.

[119] Wilkinson, D.J., (2006). *Stochastic modelling for systems biology*, Chapman & Hall/CRC.

[120] W.J. Blake, M. Kaern, C.R. Cantor and J.J. Collins, (2003). Noise in Eukaryotic Gene Expression, *Nature*, Vol. 422, pp. 633–637.

[121] Yue H., Brown M., Knowles J., Wang H., Broomhead, D.S., Kell, D.B., (2006). Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: a case study of an NF-kappaB signalling pathway, *Molecular Biosystems*, 2(12): 640-9.

[122] Yao, K. Z., Shaw, B. M., Kou, B., McAuley K. B., and Bacon, D. W., (2003). Modeling Ethylene/Butene Copolymerization with Multi-site Catalysts: Parameter Estimability and Experimental Design, Polymer reaction engineering, Vol. 11, No. 3, pp. 563 - 588.

[123] Zi, Z., (2011). Sensitivity analysis approaches applied to systems biology models. *IET Systems Biology*, 5(6), 336–346.