ACCURATE PARAMETRIC SENSITIVITY OF STOCHASTIC BIOCHEMICAL
SYSTEMS

by

Farid Gassoumov

Bachelor of Science, Ryerson University, 2013

A thesis

presented to Ryerson University

in partial fulfillment of the
requirements for the degree of
Masters of Applied Mathematics
in the Program of
Masters of Applied Mathematics

Toronto, Ontario, Canada, 2015

©Farid Gassoumov 2015

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

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

Abstract

Accurate parametric sensitivity of stochastic biochemical systems

Masters of Applied Mathematics 2015

Farid Gassoumov

Masters of Applied Mathematics

Ryerson University

Computational and Systems Biology are experiencing a rapid development in recent years. Mathematical and computational modelling are critical tools for studying cellular dynamics. Molecular interactions in a cell may display significant random fluctuations when some key species have low amounts (RNA, DNA), making the traditional model of the deterministic reaction rate equations insufficient. Consequently, stochastic models are required to accurately represent the biochemical system behaviour. Nonetheless, stochastic models are more challenging to simulate and analyse than the deterministic ones. Parametric sensitivity is a powerful tool for exploring the system behaviour, such as system robustness with respect to perturbations in its parameters. We present an accurate method for estimating parametric sensitivities for stochastic discrete models of biochemical systems using a high order Coupled Finite Difference scheme and illustrate its advantages compared to the existing techniques.

Acknowledgements

I would like to express my endless gratitude to my supervisor, Dr. Silvana Ilie, for her in depth knowledge, guidance and incredible patience. Furthermore I would like to thank all the wonderful people at the mathematics department at Ryerson University, particularly Dr Katrin Rohlf, Dr. Jean-Paul Pascal and last but not least, Dr. Larry Kolasa. Your knowledge and support has made the years at Ryerson some of the best years of my life.

Dedication

To my loving parents, Jane and Nazim, and to my beloved fiance Ayten.

Contents

1	Introduction	1
2	Background	6
2.1	Chemical Master Equation	7
2.2	Stochastic Simulation Algorithm	9
2.3	Tau-Leaping Method	13
2.4	Normal Tau-Leaping Method	15
2.5	Chemical Langevin Equation	16
2.6	Reaction Rate Equations	17
3	Parametric sensitivity	20
3.1	Motivation	20
3.2	Sensitivity analysis for the CME	21
3.2.1	Finite perturbation approximations of the CME	21
3.2.2	Independent Random Numbers method	22
3.2.3	Random Time Change	22
3.2.4	Common Reaction Path	23
3.2.5	Common Random Number method	24
3.2.6	Coupled Finite Difference	25
3.3	Sensitivity analysis for CLE	28
3.3.1	Euler-Maruyama method for a generic SDE	29
3.3.2	Euler-Maruyama method for CLE	30
3.4	Sensitivity analysis of RRE	30
4	Numerical Results	32
4.1	Simple population model	33
4.2	Stiff Non-linear System 1	37
4.3	Stiff Non-linear System 2	41
5	Conclusion	45

References	52
Glossary	53

List of Figures

4.1	Population model: Mean comparison	34
4.2	Population model: Means magnified	35
4.3	Population model: Standard deviation comparison	35
4.4	Stiff model 1: Means comparison	38
4.5	Stiff model 1: Means magnified	39
4.6	Stiff model 1: Standard deviation comparison	39
4.7	Stiff model 2: Means comparison	42
4.8	Stiff model 2: Means magnified	43
4.9	Stiff model 2: Standard deviation comparison	43

List of Appendices

Appendix A

47

Chapter 1

Introduction

The design and analysis of predictive models of cellular processes are important problems in Computational and Systems Biology. Often in the literature simplifying assumptions are adopted, that the systems are well-stirred and at thermal equilibrium. In this thesis, we will make the same assumptions. A frequently used approach to modelling cellular interactions characterizes the evolution of molecular concentrations by ordinary differential equations (ODE). However, to take into account the randomness of collisions inside the cells by the reacting molecules, and to account for the stochastic behaviour of small populations of important reactant species, a deterministic representation is generally not enough. A better way to model the dynamics of these biochemical systems with accuracy is to take the stochastic approach, utilizing Markov processes [20]. Below we will discuss several applications from Molecular Biology which can be studied using the stochastic and discrete approach.

Many components of living cells are present in low quantities. Such components include DNA and important regulatory molecules [35]. Therefore, stochastic effects, or noise, may be responsible for the large amounts of cell-cell variations exhibited by isogenic populations [40, 14]. Through an experiment designed with constructed strains of *E. coli* bacteria, the detection and discrimination of the stochastic effects was made possible [11]. The types of noise are classified as inherent to biochemical processes of gene expression (the intrinsic noise) and the fluctuations in other cellular components (extrinsic noise). The intrinsic noise for any given gene can be operationally defined as the extent to which two identical copies of the gene differ. Each gene expression is controlled by the concentrations, states and locations of molecules such as enzymes, polymerases and regulatory proteins, while the extrinsic noise arises from the fluctuations in these molecules. Both types of noise greatly influence the overall variation. The amplitude of the variation is controlled by the transcription rate, regulatory dynamics and genetic factors. Due to the inherent stochasticity of the above factors, stochastic modelling is the most accurate approach to understanding and predicting gene expression [44].

One more application of stochastic modelling in gene regulatory networks arises from the time delay in gene transcription [15, 8]. Genetic activity is controlled by molecular signals that determine the time and frequency of the transcription for a given gene. The stochastic factor comes from the environmental influences that produce additional signals, or by the signals from other cells interfering and influencing the

future course of cellular events. In biochemical regulatory networks, the time it takes for these signals and controlling influences is determined by the inevitable delays due to the molecular concentrations changes, either accumulation or a decline. If we define the genetically coupled links as the links where protein product encoded by one gene regulates the expression of the other gene, we can see that the time delay in genetically coupled links depends on the time needed for the protein concentration to achieve the size required for the next level in the cascade [38]. Conversely, the delay after the controlling promoter turns off depends on the time it takes the protein concentration to decay below the effective range. The large differences in time between successful events in regulatory cascades across a cell population and the random pattern of expression of competitive effectors [39], suggest a stochastic approach. The expected time pattern of protein production from the controlled gene was investigated using the stochastic formulation of chemical kinetics [19, 20]. The study of this delay in gene transcription [51] is beyond the scope of this thesis.

Another interesting biological application of stochastic modelling comes from the analysis of circadian rhythms, or internal clocks. Circadian rhythms with varying components exist in all living organisms ranging from complex organisms such as humans to the most basic ones, such as cyanobacteria [34]. Our concern is with the mechanisms that can account for circadian rhythms at the cellular level. Although circadian oscillators exist in complex multicellular organisms as well as in single-cell organisms, it is thought that most occur in single cells [9, 41, 49]. The challenge once again is due to the fact that in cells, a given transcribed gene is either present in singular or at most a small number of copies, and its interaction with a transcriptional regulator is not accurately modelled by deterministic differential equations [10]. Rather, because the number of copies of an expressed gene, and at some times the numbers of transcription factor molecules in a cell is small, such interactions are more accurately described employing a stochastic approach, as has been done for a number of existing models [53, 12, 25, 17], using the Gillespie's algorithm [18, 19].

When dealing with discrete stochastic models, the most popular approach is to characterize the evolution of the probability that the system is in a particular state by a system of ordinary differential equation (ODE) known as the Chemical Master Equation (CME) [20]. This discrete stochastic model of biochemical systems has been employed in several biological settings with remarkable success. Since the Chemical Master Equation is too complex to solve analytically for most applications, we employ Monte Carlo techniques to approximate its solution. The CME can be simulated accurately using one of the two stochastic simulation algorithms proposed by Gillespie [18, 19]. The two algorithms for Monte Carlo simulation of stochastic trajectories of such a chemical system are called the First Reaction and the Direct Method, known also as the Stochastic Simulation Algorithm (SSA) [19]. The First Reaction algorithm proceeds by generating a random event time for each reaction channel from the exponential distributions of each. It's alternative, the Direct Method, generates the time of the next reaction from an exponential distribution with the sum of the channel propensities as the characteristic time as well as the index of that reaction. In both methods, the system is updated according to the stoichiometry of the chosen reaction channel, and the simulation time is updated to the generated event time. Gillespie established that the two algorithms are equivalent and could be used interchangeably. The computational cost of these two exact Monte Carlo methods is prohibitive for systems where reactions run at very high rates. The high computational cost is due to the fact that the SSA must simulate every reaction event. When large populations of some chemical species are present, and/or fast reactions are involved in the system, the number of simulations of reaction events becomes very

large. A common approach to speed up the simulation of CME is the tau-leaping strategy proposed by Gillespie [22], which approximates the solution of CME with very little loss of precision by taking time steps that are much longer than a single reaction. We use a predefined time step τ and in that time interval, we fire simultaneously all the reactions that would have normally taken place. The tau-leaping method serves as both an accelerated numerical solution to CME and also as a link, connecting CME to a less precise, but much more efficient model, called the Chemical Langevin Equation (CLE) [21]. CLE is a system of non-linear Itô stochastic differential equations (SDE) [33] with multiplicative noise, containing considerably fewer components, making its simulation much faster than that of any SSA algorithm. The derivation of the Chemical Langevin Equation comes from the Euler-Maruyama discretization of the continuous time problem involving independent scalar Brownian motions [29].

SDEs are a widely studied class of problems, with numerous published studies in Systems Biology as well as Physics and Chemistry, and there has been a consistent interest in Langevin type equations in the literature. The CLE approach is utilized in modelling enzymatic reactions as well as the gene transcriptions [31, 1]. The CLE contains a deterministic and a stochastic component, and gives rise to the Reaction Rate Equations (RRE) when the stochastic component becomes negligible compared to the deterministic one [52]. Despite the accuracy of RRE being limited to the first order reactions, the deterministic RRE have been employed successfully in predicting the behaviour of chemical reaction systems. However, these equations often give inaccurate predictions when some molecular species have low population numbers. Important examples of practical applications which can not be modelled by the RRE are genetic networks, where RNA and DNA have low amounts of molecules. The use of RRE is justified when the system volume and species populations become very large, while the concentrations remain constant.

A potent strategy for reducing the computational cost of simulating complex biochemical systems is to use hybrid methods. There is a multitude of hybrid methods for solving the Chemical Master Equation, such as the methods of Alfonsi et.al [13], Cao et al. [6], Haseltine & Rawlings [7], Hellander & Lötstedt [27], Kielh et al. [28], MacNamara et al. [16], Mattheyses & Simmons [36], Puchalka & Kierzek [37], Rao & Arkin [42], Salis & Kaznessis [43], Samant & Vlachos [47], Weinan et al. [48]. A hybrid method utilizes a combination of models and/or a combination of exact and approximate algorithms for them [30, 32]. The key is to preserve the probabilistic behaviour while providing greater computational efficiency for those reaction types that occur with high rates. The hybrid approach may merge the Monte Carlo algorithm of discrete simulation with the time-step integration of ordinary differential equations. Thus, we avoid the approximations necessary to add stochastic noise terms to differential equations. This is done by means of dividing the set of reaction channels in a simulation into two regimes: continuous and discrete. The continuous regime reactions are modelled using differential equations. As for the rest of the model, the events of lower frequency, such as transcription, translation and molecular signalling, are retained in the stochastic regime so that important consequences of their stochastic behaviour are preserved. The continuous regime must satisfy the accuracy and stability conditions of the numerical solution technique: the number of instances of each molecular species in a reaction in the continuous regime must be large, and the number of reaction events of each reaction occurring within one time step of the numerical solver must be much larger than one. If either condition is not satisfied for a reaction, then it must be handled in the discrete regime.

Chemical reaction models are most commonly dependent on a set of kinetic parameters, values of which

could be known or unknown, depending on application. In some cases, such as gene regulatory networks, accurate measurements of the kinetic parameter values may not be available. The output of the system is heavily dependent on some of these parameters, necessitating the study of its behaviour with respect to their perturbations. Parametric sensitivity analysis studies the effects of the perturbations in such parameters on the system behaviour as a whole. It allows to identify critical reactions and helps in reducing the model. Sensitivity analysis methods can be categorized in two ways, with each way possessing its own benefits and/or drawbacks. These categories quantify the effects of finite or infinitesimal perturbations [46], each being appropriate for different applications. The finite perturbation approach perturbs the parameter of interest by a small but quantifiable amount. The perturbed and nominal trajectories are simulated using several algorithms. The finite difference approximation methods are used to obtain the sensitivity of the parameter to the said perturbation. The infinitesimal perturbations on the other hand deal with the parameter of interest perturbed by a vanishingly small amount, which is equivalent to taking a partial derivative with respect to the parameter of choice. We will consider the sensitivity analysis for the Chemical Master Equation, the Chemical Langevin Equation and Reaction Rate Equations respectively.

For the sensitivity of the Chemical Master Equation we employ the finite difference approximation approach. The methods for computing the parametric sensitivity for the finite perturbations are based on estimating the derivative of the function of the propensities a_j with respect to the perturbed kinetic parameter c_j [26]. Since the CME essentially consists of a very large ODE system that is impossible to solve analytically for all but the simplest cases, we will employ exact or approximate stochastic methods to simulate the nominal and perturbed trajectories, and calculate the sensitivity by means of finite differences. One method for estimating sensitivities for the CME is based on a representation of the biochemical system dynamics due to Anderson ([3]) and it is called the Coupled Finite Difference (CFD) technique. We propose a new fourth order Coupled Finite Difference technique and show that it has improved accuracy compared to the first order one. We also present the existing Common Reaction Path method (CRP) due to Rathinam et. al. [46] that utilizes the Random Time Change algorithm [4] and the Common Reaction Path method [46]. These methods are applicable to discrete stochastic models, and could produce biased results.

There are two approaches for calculating the infinitesimal sensitivity in the stochastic chemical kinetics. One approach is the likelihood ratio, which will not be covered in this thesis. The other approach is the pathwise differentiation, or infinitesimal perturbation analysis, which is our method of choice when computing the parametric sensitivity of the Chemical Langevin Equation. The infinitesimal sensitivity is defined as the partial derivative of the function of propensities with respect to the kinetic parameter of interest [24]. The pathwise derivative approach produces the sensitivity estimates by differentiating each simulation outcome with respect to the kinetic parameter of choice. This approach is limited to a certain class of problems, in particular to continuous stochastic models of biochemical systems and produces unbiased first order sensitivities.

When very large numbers of each species are present, the CLE model of well-stirred chemical kinetics could be further reduced [30] into a set of Ordinary Differential Equations (ODEs). The reduction is done by neglecting the stochastic component of the CLE, made possible by the earlier assumption. The resulting ODEs, the Reaction Rate Equations (RRE), provide a deterministic and continuous model with dimension equal to the number of reacting species. An important restriction for the RREs is that they are only

accurate for first order reactions. The sensitivity analysis of ODEs and in particular, RREs, has been studies extensively in the past. Methods for parametric sensitivity may be found in [5], while strategies for estimating the sensitivity in the reaction rate equations are referenced in [52].

The outline of this thesis is as follows: Chapter 2 provides detailed background information of CME, SSA, Tau Leaping, CLE and RRE. Chapter 3 describes the sensitivity analysis for both finite and infinitesimal perturbations and introduces the new fourth order CFD strategy. Chapter 4 contains the numerical results showing the advantages of the CFD method analysed. The conclusion and the topics for the future research are presented in Chapter 5.

Chapter 2

Background

One of the most accurate models for describing the stochastic chemical kinetics of well-stirred systems is by the Chemical Master Equation (CME). The Chemical Master Equation is a stochastic discrete model. It has been successfully employed to study critical biological processes, such as gene expression and regulation, as well as circadian rhythms.

Let us consider a system of N chemical species $S_1, \dots, S_N$, interacting with each other via M types of chemical reactions $R_1, \dots, R_M$. A molecular dynamics approach to solving such a system would involve tracking the position and velocity of each molecule in the system - a daunting task. In this thesis, we make the assumptions that the system of interest is well-stirred, of constant volume and in thermal (but not chemical) equilibrium. In a well-stirred system of constant volume, most of the molecular collisions that take place are elastic, resulting in the positions of the molecules becoming uniformly randomized in space, and the velocities thermally randomized in accordance with the Maxwell-Boltzmann distribution [23]. This will allow to forgo tracking the positions and the velocities of the molecules and focus entirely on the changes in the molecular populations, significantly simplifying the mathematical model. This simplified model is more efficient to simulate numerically. If the biochemical system is well-stirred, it is sufficient to represent it at time t , by the state vector:

$$X(t) = [X_1(t), X_2(t), \dots, X_N(t)]^T,$$

where $X_i(t)$ is the number of molecules of the species S_i at time t . If initially, at $t_0 = 0$, the system state is $X(t) = X_0$, we wish to study the evolution in time of the state vector $X(t)$. Note that $X(t)$ is a discrete Markov process.

When a reaction R_j takes place, the system is changed. We keep track of the changes by means of the state-change vector:

$$\nu_j \equiv (\nu_{1j}, \dots, \nu_{Nj})^T,$$

where ν_j is the change in the number of S_i molecules due to one reaction R_j . The matrix having as columns the state-change vectors is called the stoichiometric matrix, and is represented as $\Sigma = \{\nu_{ij}\}_{1 \leq i \leq N, 1 \leq j \leq M}$, where N is the number of reactants and M is the number of reactions involved in the biochemical process. The reaction R_j can also be described by means of its propensity function. Define $a_j(x)dt$ to be the probability

that, given that at time t the system was in state x , one reaction R_j happened in $[t, t + dt]$. Its form is justified from the principles of kinetic theory and it is constructed as follows:

- First order reaction: $S_m \xrightarrow{c_j}$ product, has propensity $a_j(X(t)) = c_j X_m(t)$.
- Second order reaction: $S_m + S_n \xrightarrow{c_j}$ product, where $m \neq n$. The propensity is $a_j(X(t)) = c_j X_m(t) X_n(t)$.
- Dimerization: $S_m + S_m \xrightarrow{c_j}$ product. The propensity is given by $a_j(X(t)) = \frac{1}{2} c_j X_m(t) (X_m(t) - 1)$.

The propensity for the first order reaction implies that the probability of it taking place is directly proportional to the number of molecules of species S_m available in the system. Similarly, the propensity for the second order reaction implies the probability of the reaction taken place is directly proportional to the number of molecules of species S_m and S_n , respectively. The propensity for the dimerization reaction when two identical chemical species interact, can be understood from combinatorics. The factor $\frac{1}{2} X_m(t) (X_m(t) - 1)$ represents the number of ways we can chose an unordered pair of objects from the total of $X_m(t)$. The reactions are assumed to be instantaneous events. We do not consider reactions of third order or above, as they have been shown experimentally to be a chemical combination of the above three reaction types.

2.1 Chemical Master Equation

The Chemical Master Equation model of well-stirred biochemically reacting systems was introduced by Gillespie [19, 20]. In order to derive it, we need the following definition: Let $P(x, t)$ be the probability that at time t the system is in state X , if $X(t_0) = X_0$. We wish to compute this probability for all possible system states. If $X(t + dt) = x$ at $t + dt$, then the following cases are possible:

1. $X(t) = x$ and there was no reaction happening in $[t, t + dt]$
2. $X(t) = x - \vec{\nu}_j$ and one reaction R_j happened in the time interval $[t, t + dt]$
3. At time t the system was several reactions away from the state x .

Assume that dt is small enough such that at most one reaction can happen during the time interval $[t, t + dt]$.

Before proceeding we need to introduce several terms to help compute the probabilities for each of the above scenarios: let H_0 be the probability that no reaction fired in the time interval $[t, t + dt]$, H_j denote the probability that exactly one reaction R_j took place in the time interval $[t, t + dt]$, where $j = 1, 2, \dots, M$ and H_{M+1} be the probability that more than one reaction occurs in the time interval $[t, t + dt]$.

From the law of total probability we know that if:

1. The events (H_j) for $j = 0, 1, 2, \dots, M + 1$ are disjoint, (i.e. no more than one event takes place) and
2. The events (H_j) $j = 0, 1, 2, \dots, M + 1$ are exhaustive, i.e. one of them must occur,

then the probability that an event A takes place is:

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

where $P(A|H_j)$ is the probability that A happens, given that H_j occurs. Take A in this case to be the event that the system is in state x at the time $t + dt$. For $1 \leq j \leq M$, H_j is defined as the event that, at time t , the system was in state $x - \nu_j$. Also, consider H_0 to be the event that the system was in state x at time t , while H_{M+1} may be chosen to be the event that the system was at least two reactions away from x at time t . For any $1 \leq j \leq M$, $P(A|H_j)$ is the probability of the reaction R_j to occur over the time interval $[t, t + dt]$. From the definition of the propensity function, we get:

$$P(A|H_j) = a_j(x - \nu_j)dt$$

where $j = 1, 2, \dots, M$.

Also, $P(A|H_{M+1}) = 0$ because H_{M+1} contains all the states which are more than one reaction away from x and we assumed this is not possible over the step dt .

Similarly, $P(A|H_0)$ is the probability of no reaction firing during the $[t, t + dt]$ interval. The probability of no reaction taking place is the total probability, i.e. 1, from which the probability of any reaction happening is subtracted, namely: $P(A|H_0) = 1 - \text{probability that one reaction happened}$:

$$P(A|H_0) = 1 - \sum_{j=1}^{M+1} a_j(x)dt.$$

Combining the above equations, we derive:

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

Subtracting $P(x, t)$ from both sides leads to:

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

Dividing by dt we obtain:

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

As $dt \rightarrow 0$ we notice that the left side of the equation becomes a time derivative, therefore:

$$\frac{d}{dt}P(x, t) = \sum_{j=1}^M (a_j(x - \nu_j)P(x - \nu_j, t) - a_j(x)P(x, t)). \quad (2.1)$$

The system of ordinary differential equations (ODE) (2.1) is called the *Chemical Master Equation* (CME) (see also Gillespie [20]). Each ODE describes the probability of the system to be at a particular state. The Chemical Master Equation is a refined model of well-stirred biochemical systems. It provides a stochastic and discrete representation of the system dynamics as the time-evolution equation for the probability of the system to be in one state.

2.2 Stochastic Simulation Algorithm

Chemical Master Equation (CME) completely determines $P(x, t | x_0, t_0)$, the probability that at the time t the system is in state x , given that at time t_0 it was in state x_0 . Unfortunately, the Chemical Master Equation has a very high dimension, due to it being a set of coupled ODE's with a separate equation for every possible combination of reacting species. Therefore, it is not feasible to solve the CME directly, either analytically or computationally, unless dealing with the simplest of cases. There is an approach however, to solve the CME indirectly. Rather than integrating the entire set of ODEs, single realizations of the state vector can be computed in such a way that the resulting probability is in exact agreement with that given by the CME. This approach, known as the Stochastic Simulation Algorithm (SSA), was proposed by Gillespie in [19, 18]. Before discussing the algorithm in detail, let's first recall the definition of the exponential distribution:

Definition 1. *An exponential distribution, also known as the memoryless probability distribution, is used to model continuous random variables such as waiting times or lifetimes. It is a process in which events occur continuously and independently at a constant average rate. In general, the exponential probability density function is given by*

$$f(X) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \quad (2.2)$$

where μ is the only parameter and it is the mean of the distribution. The memoryless property arises from the fact that, in exponential distribution,

$$P(X > (s + t) | X > t) = P(X > s)$$

In addition, we shall need the following results (see also [54]).

Lemma 2.2.1. *If $X_k \sim \text{Exp}(\mu_k)$, $k = 1, 2, \dots, n$, are independent exponential random variables then:*

$$X_0 \equiv \min_k(X_k) \sim \text{Exp}(\mu_0) \text{ where } \mu_0 = \sum_{k=1}^n \mu_k.$$

Proof. First note that for $X \sim \text{Exp}(\mu)$ we have $P(X > x) = e^{-\mu x}$. Then:

$$P(X_0 > x) = P(\min_k(X_k) > x)$$

which equals to:

$$P([X_1 > x] \cap [X_2 > x] \cap \dots \cap [X_n > x])$$

$$\begin{aligned}
&= \prod_{k=1}^n P(X_k > x) \\
&= \prod_{k=1}^n e^{-\mu_k x} \\
&= e^{-x \sum_{k=1}^n \mu_k} \\
&= e^{-\mu_0 x}
\end{aligned}$$

Thus $P(X_0 \leq x) = 1 - e^{-\mu_0 x}$, therefore $X_0 \sim \text{Exp}(\mu_0)$. $\square$

Lemma 2.2.2. *Suppose that $X \sim \text{Exp}(\mu)$ and $Y \sim \text{Exp}(\lambda)$ are independent random variables. Then*

$$P(X < Y) = \frac{\mu}{\mu + \lambda}$$

Proof. The following holds

$$\begin{aligned}
P(X < Y) &= \int_0^\infty P(X < Y | Y = y) f(y) dy \\
&= \int_0^\infty P(X < y) f(y) dy \\
&= \int_0^\infty (1 - e^{-\mu y}) \lambda e^{-\lambda y} dy \\
&= \frac{\mu}{\mu + \lambda}
\end{aligned}$$

$\square$

We will also need to establish the likelihood of a particular exponential random quantity of an independent set to be the smallest:

Lemma 2.2.3. *If $X_k \sim \text{Exp}(\mu_k)$, $k = 1, 2, \dots, n$, are independent random variables, let j be the index of the smallest of the X_k . Then j is a discrete random variable with probability mass function PMF*

$$\pi_k = \frac{\mu_k}{\mu_0}, k = 1, 2, \dots, n$$

where $\mu_0 = \sum_{k=1}^n \mu_k$

Proof. We observe that

$$\begin{aligned}
\pi_j &= P(X_j \leq \min_{k \neq j} (X_k)) \\
&= P(X_j < Y)
\end{aligned}$$

where $y = \min_{k \neq j}(X_k)$, so that $Y \sim \text{Exp}(\mu_{-j})$, where $\mu_{-j} = \sum_{k \neq j} \mu_k$

$$\begin{aligned} &= \frac{\mu_j}{\mu_0} \\ &= \frac{\mu_j}{\mu_j + \mu_{-j}} \end{aligned} \tag{2.3}$$

□

Equation (2.3) is the consequence of lemma (2.2.2)

In order to derive Gillespie's algorithm, we shall introduce the probability below:

- Let $P_0(\tau|x, t)$ be the probability that given $X(t) = x$, no reaction takes place in $[t, t + \tau]$.
- Assume that the time interval of interest is $[t, t + \tau + d\tau]$ and that the events over the period $[t, t + \tau]$ are independent of the events over the period $[t + \tau, t + \tau + d\tau]$.
- Then, the probability of no reaction taking place in $[t, t + \tau + d\tau]$ = probability of no reaction in $[t, t + \tau]$ and no reaction in $[t + \tau, t + \tau + d\tau]$ = [Probability of no reaction in $[t, t + \tau]$] x [Probability of no reaction in $[t + \tau, t + \tau + d\tau]$] = [Probability of no reaction in $[t, t + \tau]$] x [1 - sum of probabilities one reaction over $[t + \tau, t + \tau + d\tau]$]
- This in turn = Probability of no reaction in $[t, t + \tau]$ x $[1 - \sum_{j=1}^M a_j(x)d\tau]$.

The independent events assumption allows for the use of multiplication in the derivation above.

Following from the definition of the propensity function, we obtain the following:

$$P_0(\tau + d\tau|x, t) = P_0(\tau|x, t)[1 - \sum_{j=1}^M a_j(x)d\tau].$$

Thus,

$$P_0(\tau + d\tau|x, t) - P_0(\tau|x, t) = -P_0(\tau|x, t)a_0(x)d\tau,$$

where we denoted by $a_0(x) = \sum_{j=1}^M a_j(x)$.

Dividing by $d\tau > 0$ on both sides of the equation we obtain:

$$\frac{P_0(\tau + d\tau) - P_0(\tau|x, t)}{d\tau} = -P_0(\tau|x, t)a_0(x). \tag{2.4}$$

Letting $d\tau \rightarrow 0$ in (2.4) leads to the linear scalar ODE:

$$\frac{d}{d\tau} P_0(\tau|x, t) = -P_0(\tau|x, t)a_0(x). \tag{2.5}$$

The equation (2.5) is subject to the initial condition $P_0(0|x, t) = 1$. Solving the linear scalar ODE (2.5) with this initial condition gives:

$$P_0(\tau|x, t) = e^{-a_0(x)\tau}. \quad (2.6)$$

Next step is to find the joint probability $p(\tau, j|x, t)d\tau$, which is essential for the SSA. It is defined as the probability that the next reaction will be the R_j reaction and that it will take place in the time interval $[t + \tau, t + \tau + d\tau]$.

By definition $p(\tau, j|x, t)d\tau$ is the joint probability that, given $X(t) = x$, the next reaction

(1) happens in $[t + \tau, t + \tau + d\tau]$ and

(2) that reaction is R_j .

Probability of (1) and (2) = (Probability that no reaction happens in $[t, t + \tau]$ x (Probability that reaction R_j happens in $[t + \tau, t + \tau + d\tau]$).

Using the definition of $P_0(\tau|x, t)$ above and that of a propensity function, we derive:

$$p(\tau, j|x, t)d\tau = P_0(\tau|x, t)a_j(x)d\tau, \quad (2.7)$$

which once again follows from the earlier assumption that events are independent during the two time periods $[t, t + \tau]$ and $[t + \tau, t + \tau + d\tau]$. Substituting (2.6) into (2.7) will give:

$$p(\tau, j|x, t) = a_j(x)e^{-a_0(x)\tau}.$$

The above equation can be rewritten as:

$$p(\tau, j|x, t) = \frac{a_j(x)}{\sum_{k=1}^M a_k(x)} \left[\left(\sum_{k=1}^M a_k(x) \right) e^{-\sum_{k=1}^M a_k(x)\tau} \right]. \quad (2.8)$$

Here we denoted by $a_0(x) = \sum_{k=1}^M a_k(x)$ the sum of all propensities corresponding to state x . We note that $\frac{a_j(x)}{a_0(x)}$ is the density of a discrete distribution, which depends on j , the index of the next reaction, while $[a_0(x)e^{-a_0(x)\tau}]$ is the density of an exponential distribution, which depends on τ , the time to the next reaction.

In conclusion, we now have a joint density function for two independent random variables, one of which being the next reaction index and the other being the time until next reaction. From equation (2.8), we could write $p(\tau, j|x, t)$ as the product of those two independent density functions, subject to the earlier assumptions.

From lemma (2.2.3) we know that "j" is the index of the first reaction occurring after time t .

The above derivations lead to the following algorithm due to Gillespie [18, 19], also known as the stochastic simulation algorithm (SSA).

Gillespie Algorithm The steps are as follows:

1. Initialize the system: at $t = 0$, $X(0) = \mathbf{x}_0$.
2. At time t compute all the propensities $a_j(x(t))$, for $1 \leq j \leq M$.
3. Choose two random numbers ξ_1, ξ_2 in $U(0, 1)$ uniformly distributed in $[0, 1]$.
4. Compute j , the index of the next reaction by

$$\min_j \sum_{k=1}^j a_k(x(t)) > \xi_1 a_0(x(t)).$$

5. Compute τ , the time to the next reaction, as

$$\tau = -\frac{\ln \xi_2}{a_0(x)}.$$

6. Update the system at $t + \tau$: $t \leftarrow t + \tau$, $x \leftarrow x + \nu_j$.
7. Go to 2 or stop.

In practice, the termination condition in step 7 is to stop after a certain time elapses. Note that Gillespie's algorithm, also known as the stochastic simulation algorithm, computes at each step the time to the next reaction and the type of that reaction, then updates the system accordingly. Consequently, for generating one trajectory, it simulates all reactions that take place, one at a time. Since many biochemical systems are subject to some reactions that are fast, computing one trajectory for these systems using the SSA is computationally very expensive.

2.3 Tau-Leaping Method

The SSA algorithm is an exact simulation method for reproducing the statistics from the CME. It can be used to solve even the more complex cases, but as explained above, it is often computationally very demanding, since at each iteration we need to obtain a reaction time and a reaction index. Should there be an abundance of molecules in the system and/or presence of fast reactions, so that $a_0(x)$ is very large and the time τ to the next reaction is very small, the simulation requires a large number of random number generations. Before we go into details discussing some techniques to improve the computational cost of stochastic simulation for well-stirred biochemical kinetics, we give the following definition:

Definition 2. A Poisson distribution is a discrete frequency distribution that gives the probability of a number of independent events occurring in a fixed time. Let μ be the average number of times that an event

occurs in a certain period of time or space. The probability of k occurrences of this event is given by:

$$P(X = k) = \frac{\mu^k e^{-\mu}}{k!} \quad (2.9)$$

where $k = 0, 1, 2, \dots$. The mean and standard deviation of the Poisson random variable X are μ and $\sqrt{\mu}$, respectively.

Having to simulate each possible reaction in the system, one at a time, is very inefficient. Thus, it would be helpful, in an effort to speed-up the simulation, to employ an approximate strategy that would sacrifice little precision but greatly reduce the computational cost. One such approximation uses a predefined time step τ and sets $\tau > 0$ in such a way that

$$a_j(X(t + \tau)) \simeq a_j(X(t))$$

for all $j = 1, 2, \dots, M$. In the time interval, $t \leq s \leq t + \tau$ we simultaneously fire all the reactions that would have normally taken place. Starting with a state vector $X(t)$, we approximate the propensity functions $a_j(X(s))$ by $a_j(X(t))$ and use those values to fire an appropriate number of reactions of each type. In order for this approximation to be valid, we need to pick τ to be sufficiently small as to ensure that the propensities do not change significantly during $[t, t + \tau]$. The number of reactions of type R_j in $[t, t + \tau]$ has a Poisson distribution $P_j(a_j(X(t))\tau)$, with mean and variance $a_j(X(t))\tau$. Consequently, the system state may be updated using the following formula:

$$X(t + \tau) = X(t) + \sum_{j=1}^M \nu_j P_j[a_j(X(t))\tau]. \quad (2.10)$$

The equation (2.10) is known as the *tau-leaping method*. This Monte Carlo type scheme was introduced by Gillespie [22]. The validity of the method is subject to the following condition, *the leap condition*:

τ has to be small enough such that for any $1 \leq j \leq M$,

$$|a_j(X(t + \tau)) - a_j(X(t))| < \epsilon |a_j(X(t))|, \quad (2.11)$$

for some small tolerance ϵ . The algorithm to update the system at $t + \tau$ according to the leaping method (2.10) is presented below:

Tau-Leaping Algorithm

1. Initialize the system at $t = t_0, x(t_0) = x_0$
2. For state x at time t , compute all propensities $a_j(x)$ for $j = 1, 2, \dots, M$.
3. Generate $k_j = P_j(a_j(x)\tau)$, the number of reactions R_j , by sampling M independent Poisson distributions $P_j(a_j(x)\tau)$, $1 \leq j \leq M$

4. Update the system state

$$X(t + \tau) = X(t) + \sum_{j=1}^M k_j \nu_j.$$

Set $t \leftarrow t + \tau$. Go to step 2 or stop.

While the tau-leaping method offers the advantage of a reduced simulation time, compared to the Gillespie's algorithm, the strategy needs to be implemented carefully when some population numbers approach zero. Then, the approximate tau-leaping scheme may drive some molecular amounts negative. Maintaining positive molecular numbers will be considered in our future research.

2.4 Normal Tau-Leaping Method

Another definition of interest for efficient stochastic modelling and simulation of biochemical kinetics is presented below.

Definition 3. *A normal distribution, also known as the Gaussian distribution, is a frequency distribution that is approximately mound-shaped. The probability density function of the Gaussian distribution is:*

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{(2\sigma^2)}}, \quad (2.12)$$

where μ and σ are the parameters representing the mean and standard deviation, respectively, and $-\infty < x < \infty$.

In addition to the leap condition, we assume now that τ is large enough such that:

$$a_j(x(t))\tau \gg 1, \text{ for all } 1 \leq j \leq M. \quad (2.13)$$

Given that the condition (2.11) is satisfied, the tau-leaping may be applied. Knowing that propensities are assumed constant over the time period $[t, t + \tau]$, we can find the number of reactions that take place during that time by using a Poisson distribution with mean and variance $a_j(X(t))\tau$. If in addition to (2.11), the condition (2.13) is also valid, then we may use the approximation:

$$P_j(a_j(X(t))\tau) \simeq N_j(a_j(X(t))\tau, a_j(X(t))\tau) = a_j(X(t))\tau + \sqrt{a_j(X(t))\tau} N_j(0, 1), \quad (2.14)$$

where $N_j(0, 1)$ are independent normal distributions with mean 0 and variance 1. That is, the Poisson distribution may be approximated by a normal distribution with the same mean and variance. Thus, the tau-leaping method leads to the following approximate strategy:

$$X(t + \tau) = X(t) + \sum_{j=1}^M \nu_j \left[a_j(x(t))\tau + \sqrt{a_j(x(t))\tau} N_j(0, 1) \right] \quad (2.15)$$

Formula (2.15) is called the *normal tau-leaping method* and it is an approximate strategy for Chemical Master

Equation. Finally, we remark that the two conditions (2.11) and (2.13) are simultaneously satisfied when all molecular populations are sufficiently abundant.

2.5 Chemical Langevin Equation

We begin with the following definition useful for diffusion processes:

Definition 4. *A Wiener process, also known as the scalar standard Brownian motion over a time period $[0, T]$ is a random variable $W(t)$ that depends continuously on $t \in [0, T]$ and has the following properties:*

- 1) $W(0) = 0$ with probability 1.
- 2) The Wiener increment $W(t) - W(s) \sim \sqrt{t-s} \cdot N(0, 1)$ for $0 \leq s < t$ is normally distributed with mean 0 and variance $t - s$.
- 3) For $0 < u < v < s < t$ the Wiener increments $W(v) - W(u)$ and $W(t) - W(s)$ are independent random variables.

The tau-leaping method, aside from being an efficient computational approach, also serves as a point of connection of the CME to a coarser-grained model, i.e. a system with fewer components, known as the Chemical Langevin Equation. To make that connection, let's revisit the Poisson distribution for the tau-leaping method from the previous section. From our background knowledge in probability, we know that the Poisson distribution with a large mean can be well approximated with a Normal distribution with the same mean and variance, if $a_j(x(t))\tau \gg 1$. As observed above, that leads to

$$P_j(a_j(x(t))\tau) \simeq a_j(x(t))\tau + \sqrt{a_j(x(t))\tau} N_j(0, 1), \quad (2.16)$$

where $N_j(0, 1)$ are independent normal distributions with mean 0 and variance 1.

Substituting (2.16) into (2.10) we get:

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

and thus:

$$X(t + \tau) = X(t) + \left[\sum_{j=1}^M \vec{\nu}_j [a_j(X(t))] \right] \tau + \sum_{j=1}^M \vec{\nu}_j \sqrt{a_j(X(t))\tau} N_j(0, 1).$$

Note that we are now using real-valued normal random variables rather than integer-valued Poisson random variables. Take $\tau \rightarrow 0$, the factor $\sqrt{\tau} N_j(0, 1) \rightarrow dW_j$, where W_j are independent Wiener processes (Brownian motion) for $1 \leq j \leq M$. We then get the following equation:

$$dX(t) = \sum_{j=1}^M \vec{\nu}_j a_j(X(t)) dt + \sum_{j=1}^M \sqrt{a_j(X(t))} dW_j. \quad (2.17)$$

The model (2.17) is a system of non-linear Itô stochastic differential equations with multiplicative noise, called the *Chemical Langevin Equation* (CLE) (see Gillespie [21]). Its solution is a stochastic process with N components that describe the time dependence of the amount of each species. The CLE is a stochastic continuous model of well-stirred biochemical systems. In this setting, the state vector $X(t)$ is a continuous Markov process.

Since the CLE is derived from the tau-leaping method, there are two conditions that need to be met for this model to be valid:

1. τ small enough such that

$$a_j(X(t + \tau)) \simeq a_j(X(t))$$

2. τ is large enough such that

$$a_j(X(t))\tau \gg 1$$

These conditions typically occur when all $a_j(X(t))$ are large. In practice, the CLE is applicable when $X_i \geq 100$ for any $i = 1, 2, \dots, N$. Note that the Chemical Langevin Equation is an approximation of the more refined model of the Chemical Master Equation. The CLE has a much lower dimension than the CME, that is N , the number of reacting species.

2.6 Reaction Rate Equations

It is evident from the equation of the CLE that it contains a deterministic and a stochastic part. If we were to ignore the latter (that is the random fluctuations), we would get a set of ODEs that could be used for modelling chemical kinetic systems with very large molecular numbers. The resultant equations are known as the reaction rate equations (RRE), and have been successfully used in modelling the behaviour of chemical reaction systems. Though their accuracy is limited to first order reactions, the deterministic RRE often provides us with satisfactory results. We note however that these equations may give inaccurate predictions when some molecular species have low population numbers. Important examples of practical applications which can not be modelled by the RRE are genetic networks, where RNA and DNA have low amounts of molecules.

In order to obtain the RRE, we will start by computing the expected value of our CLE (2.17):

$$E[dX(t)] = E \left[\sum_{j=1}^M \nu_j a_j[X(t)] dt + \sum_{j=1}^M \nu_j \sqrt{a_j[X(t)]} dW_j \right]$$

We can rewrite the equation above as:

$$E[dX(t)] = d[EX(t)] = \sum_{j=1}^M \nu_j E[a_j(X(t))dt] + \sum_{j=1}^M \nu_j E\left[\sqrt{a_j(X(t))}dW_j(t)\right].$$

But it is well known that, for a Wiener process W_j , $E[\sqrt{a_j(X(t))}dW_j(t)] = 0$, as $W_j(t+dt) - W_j(t) \sim \sqrt{dt}N_j(0,1)$, is normally distributed with mean zero [33]. Thus we can find the rate of change of $E(X(t))$ to be

$$dE(X(t)) = \sum_{j=1}^M \nu_j E[a_j(X(t))dt],$$

or, alternatively,

$$\frac{dE(X(t))}{dt} = \sum_{j=1}^M \nu_j E[a_j(X(t))].$$

Denoting $Y = E(X)$ the expected value of X , we get the following from the equation above:

$$\frac{dY}{dt} = \sum_{j=1}^M \nu_j E[a_j(X(t))].$$

Now, we want the expression to be in the form

$$\frac{dY}{dt} = \sum_{j=1}^M \nu_j a_j(E(X))$$

which, if $E(a_j(X)) = a_j(E(X))$ could be rewritten as

$$\frac{dY}{dt} = \sum_{j=1}^M \nu_j [a_j(Y)].$$

Next we are going to establish when such an assumption could be made. For first order reactions with propensities of the form $a_j(x) = c_j X_k$, it is evident that:

$$E[a_j(X)] = E(c_j X_k) = c_j (E(X_k)) = c_j Y_k = a_j[E(X)],$$

thus

$$E[a_j(x)] = a_j[E(x)].$$

Therefore, we conclude that the average behaviour of the solution of the CLE may be predicted by the RRE when all reactions are of first order, and all molecular population numbers are very large. Thus, we obtain

the reaction rate equation for first order reactions as:

$$\frac{dY(t)}{dt} = \sum_{j=1}^M \nu_j a_j(Y(t)).$$

However, for the second order reactions with propensities represented by $a_j(x) = c_j X_k X_l$, our assumption may not be valid, as

$$E(a_j(X)) \neq a_j(E(X))$$

in general. Consequently, it can not be argued that the RRE solution is the "average" solution from the CLE or CME.

The justification of modelling in terms of concentrations and instantaneous rates of change are predicated on the assumption that a large number of molecules are present. Consider the thermodynamic limit, in which the volume of the system, V , and the species populations, $X_i(t)$, tend to infinity, but the species concentrations $X_i(t)/V$ stay constant. In this scenario, the deterministic coefficients in the Chemical Master Equation grow like the system size but the stochastic coefficients grow like the square root of the system size. Since the stochastic terms become negligible with respect to the deterministic terms, in the thermodynamic limit, the reaction rate equations model may be applied. The RRE is a significant simplification of the CME, being a continuous deterministic model of size equal to the number of biochemical species in the system.

Chapter 3

Parametric sensitivity of biochemical systems

3.1 Motivation

The behaviour of the biochemical reaction network may depend greatly on the values of its kinetic parameters or the initial concentrations. Minuscule perturbations of the said parameters could lead to dramatic changes in the system output. As such, it is of utmost importance to study and characterize the effects of such perturbations. Parametric sensitivity of biochemical systems studies the dependence of the system dynamics on the reaction rate parameters or the initial conditions. It is an essential analysis tool in kinetic modelling and allows us to deduce some important properties of the system. Parametric sensitivity could be employed to quantify the robustness of the model with respect to small perturbations in its parameters, assist in identifying which components of the model are actively contributing to the system dynamics and is critical in model reduction.

In a deterministic system, the (local) first order sensitivity of a quantity $X_i(t, c_j)$ depending on the time t and a parameter c_j is defined as

$$s_{i,j} = \frac{\partial X_i}{\partial c_j}(t, c_j).$$

The larger the sensitivity to the parameter, the greater the change that the system may experience due to the perturbation in the said parameter. Conversely, obtaining small values for the sensitivity of a system is an indication of system robustness with respect to the perturbations in that parameter, allowing to reduce the model or identify crucial reactions. In this thesis, we will compare finite-difference approximation methods for the parametric sensitivity of the Chemical Master Equation with the deterministic sensitivity of the Reaction Rate Equation and a pathwise sensitivity method for the Chemical Langevin Equation. A method for computing sensitivities, utilizing the Random Time Change representation of discrete state Markov processes, introduced by Rathinam et al.[46] will also be discussed. This method, called the Common Reaction Path (CRP), will then be compared to the strategies described above.

3.2 Sensitivity analysis for the CME

Sensitivity analysis for the Chemical Master Equation is a challenging task since the model is discrete and stochastic. As such, the existing sensitivity analysis techniques for ODEs (continuous deterministic models) or for SDEs (continuous stochastic models) no longer apply. To overcome this obstacle, we will be taking the Monte Carlo approach to computing sensitivities. The existing Monte Carlo based methods of estimating sensitivity are the following:

- Finite Perturbation methods:
 - Independent Samples
 - Correlated Samples (with the following methods):
 1. Common Reaction Path
 2. Common Random Numbers
 3. Coupled Finite Differences
- Infinitesimal Perturbation methods:
 1. Likelihood Ratio (Girsanov)
 2. Pathwise Differentiation (if possible)

The methods listed above have their features and drawbacks, which we will analyse in detail and provide numerical results for several of them later in the thesis.

3.2.1 Finite perturbation approximations of the CME

When performing Monte Carlo based sensitivity analysis of stochastic chemical networks, we are typically interested in the sensitivity of the expected value of a function f of the state $X(T)$, at some given final time T . We wish to establish how sensitive the $E[f(X(T, c_j))]$ is to perturbations in the kinetic parameter c_j . In general, this sensitivity may be approximated by the finite-difference scheme

$$\frac{E[f(X(T, c_j + \epsilon, \omega))] - E[f(X(T, c_j, \omega))]}{\epsilon}$$

with ϵ representing a small perturbation in the parameter c_j and ω being a realization of the sample trajectories space Ω . As $\epsilon \rightarrow 0$, the expression above limits to the partial derivative

$$\frac{\partial}{\partial \epsilon} E[f(X(T, c_j + \epsilon, \omega))]$$

evaluated at $\epsilon = 0$. Let us define a random variable S , the expected value of which we set out to compute,

$$S = \frac{E[f(X(T, c_j + \epsilon, \omega))] - E[f(X(T, c_j, \omega))]}{\epsilon}$$

Since we usually can not obtain the distribution for X analytically on account of it being a stochastic process, we can compute a sample of independent realizations of S and then estimate its expected value. The independent sample paths could be generated in a number of ways, as long as the realizations of SSA are exact. In this thesis, we will discuss the Independent Random Number approach (IRN), and the Common Reaction Path (CRP) method using the Random Time Change (RTC) algorithm. Moreover, we shall describe the Coupled Finite Difference approximation method utilizing the tau-leaping algorithm.

3.2.2 Independent Random Numbers method

Should we choose to generate the sample of $f(X(T, c_j + \epsilon))$ independent of the sample of $f(X(T, c_j))$, we would be taking the Independent Random Number approach. A general algorithm for such a method is as follows:

Independent Random Numbers Algorithm

1. For $k = 1$ to N_{max} where N_{max} is the number of trajectories that we are simulating:
2. Initialize a random number generator with a random seed.
3. Let $c = c_j$ and use any SSA algorithm to compute $X(T, c_j)$. All the random numbers we require will be generated by successive calls to the random number generator.
4. Let $c = c_j + \epsilon$ then use the same SSA algorithm as in step (3) to obtain $X(T, c_j + \epsilon)$. The random numbers needed are generated independent of the numbers in step (2) and (3) by successive calls to the random number generator. We can accomplish this by continually drawing from the same stream without resetting the state, or by reinitializing the random number generator with a different random seed.
5. Compute the sensitivity for the k^{th} trajectory by finding $S_k = \frac{f(X(T, c_j + \epsilon)) - f(X(T, c_j))}{\epsilon}$
6. End for loop in k .

This method has a major drawback. We would have to simulate the system N_{max} times to generate accurate estimates when using IRN. Recall from sampling techniques for a population that a sample mean of variable S , labelled $\bar{S}$ has a standard deviation of $\bar{\sigma}_z = \frac{\sigma_z}{\sqrt{N_{max}}}$. Since we wish to obtain results with the lowest possible variance, a high number of simulations is in order, which could get very expensive computationally.

3.2.3 Random Time Change

The Random Time Change representation [4] is applicable to any system described by a continuous Markov process with a probability mass function evolving in time according to the CME (2.1). To do so, we must associate each reaction channel with its own internal clock [46]. The rate of the clock is governed by the propensity function of the reaction. Note that the RTC is another exact stochastic simulation algorithm for the CME. Commonly, two independent streams of random numbers are used for the Monte

Carlo simulations. IRN use may result in a statistical estimator with a large variance - computation time increases as larger samples are required. A strategy for calculating the sensitivities, the Common Reaction Path due to Rathinam et al [46], utilizes the RTC to reduce the computational cost of the simulation. It can be used to compute the nominal and perturbed trajectories, utilizing the same underlying sequence of random numbers.

The Random Time Change Algorithm

1. Initialize the system: $i = 0, t_0 = 0, x(t_0) = x_0, S_j = 0, k_j = 1, I_+^j = E_1^j$ - Initial conditions known.
2. Exit if terminal condition is reached.
3. Calculate $a_j(X(T_i))$ for $j = 1, \dots, M$.
4. Compute $T_{i+1} = T_i + \min \left(\frac{I_+^j - S_j}{a_j(X(T_i))} \right)$.
5. Set $X(T_{i+1}) = X(T_i) + \nu_j$.
6. Set $S_j = a_j X(t_i)(T_{i+1} - T_i)$ and increment k_j .
7. Set $I_+^j = I_+^j + E_{k_j}^j$ and increment i .

It is worth noting that the RTC draws from M independent, parallel streams of random numbers $E_1^j, E_2^j, \dots, (\text{ for } j = 1, \dots, M)$ associated with each reaction channel instead of drawing them from a single stream. This is crucial when performing the parametric sensitivity, by ensuring that keeping the same M parallel sequences of random numbers is equivalent to keeping the same paths for the processes.

3.2.4 Common Reaction Path

The exact simulation algorithm (RTC) is utilized by the Common Reaction Path (CRP), a technique for estimating the sensitivities (see Rathinam et al [46]). If we treat the sequence of samples from the k^{th} stream of random numbers during the simulation as the reaction path of the k^{th} reaction channel, and recognize that collectively the set of reaction paths uniquely determines the evolution of the state, we see that the method was aptly named. Let us assume the existence of M independent streams of unit rate exponential random numbers. Suppose the j^{th} stream is accessed by $E = \text{Rand}(j)$ and N_{max} is the total number of generated trajectories. We will randomly seed each stream before each new trajectory is generated, using the current system clock as the seed. This is done to ensure the independence of the streams. In order to share the reaction paths between the simulations of processes with kinetic parameters c_j and $c_j + \epsilon$, we generate a trajectory $X(c_j)$ and store the random numbers E_k^i in an M dimensional array. It is imperative that the random numbers are ordered exactly as they were generated during the simulation and into the row corresponding to their reaction channel. The trajectory of $X(c_j + \epsilon)$ is then generated using the numbers from the array in the same order. An alternative, more practical approach, is to seed each stream and to store the seed before generating a trajectory for $X(c_j)$, then reseed each stream with identical seeds before generating $X(c_j + \epsilon)$. The algorithm for this approach is outlined below:

Common Reaction Path Algorithm

1. For $i = 1:N_{max}$ do:
2. Seed the streams.
3. Let $c = c_j$ then run the RTC simulation algorithm to generate $X(T, c_j)$
4. Reseed the streams with the same seed as in the previous step.
5. Let $c = c_j + \epsilon$, run the RTC algorithm to compute $X(T, c_j + \epsilon)$.
6. Compute the sensitivity for the i^{th} trajectory by using $S_k = \frac{f(X(T, c_j + \epsilon)) - f(X(T, c_j))}{\epsilon}$.
7. End for loop in i .

3.2.5 Common Random Number method

An alternative to the inaccurate and expensive IRN method, namely the Common Random Number, or CRN, is designed to reduce the variance of the variable S by introducing dependence among the random variables being estimated in the numerical simulations. This is achieved by using the same stream of random numbers to generate the samples of $X(T, c_j)$ and $X(T, c_j + \epsilon)$. The method is applied to the SSA solver due to Gillespie [18, 19]. Estimating the sensitivity by employing the CRN is achieved by using the same stream of uniform random numbers generated at each step. We reseed the random number generator before computing $X(T, c_j + \epsilon)$ with the same seed used to simulate $X(T, c_j)$ at each step. A general algorithm for estimating the sensitivity using CRN is as follows.

Common Random Number Algorithm

1. For $k = 1:N_{max}$ do:
2. Generate a random seed w .
3. Seed the random number generator (Command *rand* was used in Matlab, to generate a uniform random number in $[0,1]$).
4. Let $c = c_j$, run the SSA algorithm to compute $X(T, c_j)$, with all the required random numbers generated by successive calls to the "rand" command.
5. Reseed the random number generator using the same seed as before.
6. Let $c = c_j + \epsilon$, run the same SSA algorithm to compute $X(T, c_j + \epsilon)$, with reusing the random numbers generated at steps (3) and (4).
7. Compute the sensitivity for the i^{th} trajectory by using $S_k = \frac{f(X(T, c_j + \epsilon)) - f(X(T, c_j))}{\epsilon}$
8. End for loop in i .

This method was shown to be less accurate than the CRP method [46] and was not used in our simulations.

3.2.6 Coupled Finite Difference

In the Coupled Finite-Difference (CFD) approach for the Chemical Master Equation, we estimate the sensitivities with respect to the reaction rate parameters, denoted by c_j , for $1 \leq j \leq M$. The key idea for this technique appears in the work of Anderson (see [50]). The initial conditions for the sensitivities are

$$s_{i,j}(t_0) = 0.$$

and the estimation is done using first and fourth order finite difference approximations. The advantages of the fourth order approximation are illustrated in the numerical results section. Among others, it is a more precise method. Before we describe the coupling, let us recall a useful property of independent Poisson random variables:

Proposition 3.2.1. *If P is an independent Poisson random variable with mean and variance α and Q is an independent Poisson random variable with mean and variance β , then $P + Q$ is a Poisson random variable with mean and variance $\alpha + \beta$.*

First order Coupled Finite Difference Let $X(t)$ be the nominal trajectory, subject to the kinetic parameters $c_1, c_2, \dots, c_M$, where $1 \leq j \leq M$. The tau-leaping method is

$$X(t + \tau) = X(t) + \sum_{j=1}^M \nu_j P_j^1(a_j(X(t)), c_1, \dots, c_M).$$

Now let $Y(t)$ be the perturbed trajectory, subject to the kinetic parameters $c_1 + \epsilon, c_2, \dots, c_M$, where $j \in [1, M]$. We perturbed the first kinetic parameter a small quantity ϵ to obtain $c_1 + \epsilon$. Then we can derive the tau-leaping approximation

$$Y(t + \tau) = Y(t) + \sum_{j=1}^M \nu_j P_j^2(a_j(X(t)), c_1 + \epsilon, c_2, \dots, c_M).$$

Let a_j represent the propensities of the nominal trajectory and b_j represent the propensities of the perturbed trajectory, respectively.

Define

$$M_j = P_j^1((\min a_j, b_j)\tau)$$

then

$$K_j = P_j^2((a_j - \min(a_j, b_j))\tau) + M_j$$

and

$$L_j = P_j^3((b_j - \min(a_j, b_j))\tau) + K_j$$

For example, for a trajectory where $a_j \leq b_j$ we have:

$$K_j = P_j^1(a_j(X)\tau).$$

$$L_j = K_j + P_j^2((b_j(Y) - a_j(X))\tau).$$

To verify that L_j is a Poisson random variable with mean and variance $b_j\tau$, we compute:

$$L_j = P_j^1((b_j(Y) - a_j(X))\tau) + P_j^2(a_j(X)\tau).$$

Since the Poisson random variables are independent, we can rewrite the above as:

$$L_j = P_j((b_j(Y) - a_j(X))\tau + a_j(X)\tau) = P_j(b_j(Y)\tau).$$

After every τ interval, we update the values of X and Y :

$$X(t + \tau) = X(t) + \sum_{j=1}^M K_j \nu_j. \quad (3.1)$$

$$Y(t + \tau) = Y(t) + \sum_{j=1}^M L_j \nu_j. \quad (3.2)$$

The sensitivity is then computed using the first order finite difference formula. The sensitivity with respect to the kinetic parameter c_j is thus estimated as:

$$s_{i,l} = \frac{Y_i(t, c_l + \epsilon) - X_i(t, c_l)}{\epsilon}. \quad (3.3)$$

First order Coupled Finite Difference Algorithm:

1. Initialize the system, assigning values to X_0 and t_0 .
2. Assign values to ϵ and the step size τ .
3. For $k = 1:N_{max}$, set $X = X_0$, $Y = X_0$
4. While $t < T$, compute $a_j(x(t, c_l))$ as the nominal propensity and $b_j(y(t, c_l + \epsilon))$, the perturbed propensity with respect to perturbation parameter c_l .
5. Update X using (3.1) and Y using (3.2)

$$X(t + \tau, c_l) = X(t, c_l) + \sum_{j=1}^M K_j \nu_j$$

$$Y(t + \tau, c_l) = Y(t, c_l) + \sum_{j=1}^M L_j \nu_j$$

reusing some Poisson random variables.

6. End while loop.

7. Compute the sensitivity for the k^{th} trajectory by using $S_k = \frac{f(Y(T, c_l + \epsilon)) - f(X(T, c_l))}{\epsilon}$
8. End for loop in k .

Fourth Order Coupled Finite Difference For the fourth order finite-difference approximation, we simulate four coupled trajectories, labelled $X(t)$, $Y(t)$, $Z(t)$ and $U(t)$ respectively. The X trajectory corresponds to the perturbation of $c_l - 2\epsilon$ of the kinetic parameter c_l , the Y trajectory corresponds to the perturbation of $c_l - \epsilon$, the Z trajectory corresponds to the perturbation of $c_l + \epsilon$ and the U trajectory corresponds to the perturbation of $c_l + 2\epsilon$. The respective propensities for $X(t)$, $Y(t)$, $Z(t)$ and $U(t)$ are defined as $a_j(X)$, $b_j(Y)$, $d_j(Z)$ and $e_j(U)$ for $1 \leq j \leq M$. If, for example, $a_j \leq b_j \leq d_j \leq e_j$, then

$$\begin{aligned} K_j &= P_j^1(a_j(X)\tau) \\ L_j &= K_j + P_j^2((b_j(Y) - a_j(X))\tau) \\ M_j &= L_j + P_j^3((d_j(Z) - b_j(Y))\tau) \\ N_j &= M_j + P_j^4((e_j(U) - d_j(Z))\tau) \end{aligned}$$

To verify that N_j has a Poisson distribution with mean $e_j(U)\tau$, we apply proposition (3.2.1):

$$\begin{aligned} N_j &= P_j^4(e_j(U) - d_j(Z))\tau + P_j^3((d_j(Z) - b_j(Y))\tau) + P_j^2((b_j(Y) - a_j(X))\tau) + P_j^1(a_j(X)\tau) \\ &= P_j[e_j(U) - d_j(Z) + d_j(Z) - b_j(Y) + b_j(Y) - a_j(X) + a_j(X)]\tau = P_j(e_j(U)\tau). \end{aligned}$$

After every τ interval, we update the values of X , Y , Z and U as:

$$X(t + \tau) = X(t) + \sum_{j=1}^M K_j \nu_j. \quad (3.4)$$

$$Y(t + \tau) = Y(t) + \sum_{j=1}^M L_j \nu_j. \quad (3.5)$$

$$Z(t + \tau) = Z(t) + \sum_{j=1}^M M_j \nu_j. \quad (3.6)$$

$$U(t + \tau) = U(t) + \sum_{j=1}^M N_j \nu_j. \quad (3.7)$$

The sensitivities are evaluated by the fourth order centered finite difference approximation:

$$s_{i,l}(t) \approx \frac{-X_i(t, c_l + 2\epsilon) + 8X_i(t, c_l + \epsilon) - 8X_i(t, c_l - \epsilon) + X_i(t, c_l - 2\epsilon)}{12\epsilon} \quad (3.8)$$

where $X_i(t, c_l + 2\epsilon)$ is the number of molecules of species X_i at time t and on a realization of the tau-leaping method corresponding to the perturbation $c_l + 2\epsilon$ of the reaction rate constant c_l . Similarly, $X_i(t, c_l + \epsilon)$, $X_i(t, c_l - \epsilon)$ and $X_i(t, c_l - 2\epsilon)$ are the numbers of molecules of species X_i at time t and on a realization of the tau-leaping method corresponding to the perturbation $c_l + \epsilon$, $c_l - \epsilon$ and $c_l - 2\epsilon$ respectively.

Fourth order Coupled Finite Difference Algorithm:

1. Initialize the system, assigning values to X_0 and t_0 .
2. Assign values to ϵ and the step size τ .
3. For $k = 1:N_{max}$, set $X = X_0$, $Y = X_0$, $Z = X_0$, $U = X_0$.
4. While $t < T$, define $a_j(X, c_l - 2\epsilon)$ as the propensity associated with $c = c_l - 2\epsilon$, $b_j(Y, c_l - \epsilon)$ as the propensity associated with $c = c_l - \epsilon$, $d_j(Z, c_l + \epsilon)$ as the propensity associated with $c = c_l + \epsilon$ and $e_j(U, c_l + 2\epsilon)$ as the propensity associated with $c = c_l + 2\epsilon$.
5. Update $X(t + \tau, c_l - 2\epsilon)$, $Y(t + \tau, c_l - \epsilon)$, $Z(t + \tau, c_l + \epsilon)$ and $U(t + \tau, c_l + 2\epsilon)$ according to (3.4), (3.5), (3.6) and (3.7) respectively, reusing some Poisson random numbers.
6. Compute the sensitivity for the k^{th} trajectory using the fourth order finite difference approximation (3.8)
7. End for loop in k .

For both the first and the fourth order CFD methods for estimating the sensitivity for the Chemical Master Equation model, we employ the tau-leaping method (2.10). Note that the trajectories are not independent, they are obtained using common random numbers to advance the numerical solution with the (Poisson) tau-leaping method.

3.3 Pathwise Sensitivity method for CLE

For the Chemical Langevin Equation, we are estimating the local sensitivities by a technique known as the pathwise sensitivity [24]. Take ω a realization of the sample trajectories space Ω and c_j a kinetic parameter. The pathwise sensitivity is interpreted as the pathwise derivative of $X(t, c_j, \omega)$ with respect to c_j , $(\partial X / \partial c_j)(t, c_j, \omega)$, when ω is fixed. Given that the derivative exists, it can be derived as follows:

$$d\left(\frac{\partial X}{\partial c_l}\right) = \sum_{j=1}^M \nu_j \left[\frac{\partial a_j(X)}{\partial X} \frac{\partial X}{\partial c_l} + \frac{\partial a_j(X)}{\partial c_l} \right] (t) dt + \sum_{j=1}^M \nu_j \left[\frac{1}{2\sqrt{a_j(X)}} \left(\frac{\partial a_j(X)}{\partial X} \frac{\partial X}{\partial c_l} + \frac{\partial a_j(X)}{\partial c_l} \right) \right] (t) dW_j \quad (3.9)$$

Since ν_j is independent of the kinetic parameter c_l . This approach is valid when every molecular species is significantly large, a condition required for the Langevin regime. To calculate the local sensitivities, we solve the system of stochastic differential equations (SDE) (3.9) and (2.17) to compute for the state vector

$X(t)$ and the sensitivities $\frac{\partial X(t)}{\partial c_j}$. Since our starting molecular concentrations are independent of the kinetic parameters, we initialize the sensitivities as $\frac{\partial X}{\partial c_j}(0) = 0$.

3.3.1 Euler-Maruyama method for a generic SDE

We present below a numerical method for approximating the solution of stochastic differential equations, the Euler-Maruyama scheme:

Consider a discretization

$$0 = t_0 < t_1 < \dots < t_n < \dots < t_L = T < \infty$$

of the time interval $[0, T]$. We shall approximate a process $X = X(t)$, $t \in [0, T]$ satisfying the SDE:

$$dX(t) = f(t, X(t))dt + \sum_{j=1}^M g_j(t, X(t))dW^j(t) \quad (3.10)$$

on $t \in [0, T]$ with initial value $X(0) = X_0$. The Euler scheme or Euler-Maruyama scheme, is a widely used numerical scheme for solving (3.10). We denote by X_n the approximation of $X(t_n)$. The Euler-Maruyama strategy gives

$$X_{n+1} = X_n + f(t_n, X_n)(t_{n+1} - t_n) + \sum_{j=1}^M g_j(t_n, X_n)(W^j(t_{n+1}) - W^j(t_n))$$

for $n \in \{0, 1, \dots, L\}$ with initial value $X(0) = X_0$. We shall also write

$$\tau_n = t_{n+1} - t_n$$

for the n^{th} increment of the time discretization and call

$$\tau = \max_{0 \leq n \leq L} \tau_n$$

the maximum step size. We shall consider, for simplicity, equidistant time discretizations with

$$t_n = n\tau.$$

The sequence

$$(X_n), 0 \leq n \leq L$$

of values of the Euler approximation (3.3.1) at the instants of the time discretization

$$t_0, t_1, \dots, t_L$$

can be recursively computed. We need to generate the random increments of the Wiener process $W(t)$ as

$$\Delta W^j(n) = W^j(t_{n+1}) - W^j(t_n) = \sqrt{\tau} N_j(0, 1),$$

for $0 \leq n \leq L$. Here $N_j(0, 1)$ are independent Gaussian random variables, with mean 0 and variance 1. Thus

$$E(\Delta W_n) = 0$$

3.3.2 Euler-Maruyama method for CLE

Applied specifically to the CLE, the Euler-Maruyama method is:

$$X_{n+1} = X_n + \sum_{j=1}^M \nu_j a_j(X_n) \tau + \sum_{j=1}^M \sqrt{a_j(X_n) \tau} \Delta W_n^j, \quad (3.11)$$

where

$$\Delta W_n^j = W_{t_{n+1}}^j - W_{t_n}^j \sim \sqrt{\tau} N_j(0, 1).$$

For the sensitivity equation (3.9), we approximate as follows:

$$S_{n+1,l} = S_{n,l} + \sum_{j=1}^M \nu_j \left[\frac{\partial a_j(X_n)}{\partial x} S_{n,l} + \frac{\partial a_j(X_n)}{\partial c_l} \right] \tau + \sum_{j=1}^M \nu_j \left[\frac{1}{2\sqrt{a_j(X_n)}} \left(\frac{\partial a_j(X_n)}{\partial x} S_{n,l} + \frac{\partial a_j(X_n)}{\partial c_l} \right) \right] \Delta W_n^j \quad (3.12)$$

for $1 \leq l \leq M$ and $0 \leq n \leq L$. The above formula was obtained by applying the Euler-Maruyama scheme to the sensitivity equation (3.9).

3.4 Sensitivity of Reaction Rate Equations

Recall from the background section that the RRE are a set of ODEs that could be represented as

$$\frac{dY(t)}{dt} = \sum_{j=1}^M \nu_j a_j(Y(t)), \quad (3.13)$$

which could be rewritten as

$$\frac{dy}{dt} = f(t, y, c_l) \quad (3.14)$$

with $y(0) = y_0$ being the initial conditions given by the system. We take the derivative of (3.14) with respect to the kinetic parameter c_l , as follows:

$$\frac{\partial(dy/dt)}{\partial c_l} = \frac{d}{dt} \frac{\partial y}{\partial c_l} = \frac{\partial f}{\partial y} \frac{\partial y}{\partial c_l} + \frac{\partial f}{\partial c_l}. \quad (3.15)$$

Note that y and f are smooth functions with respect to c_l . Also, the initial sensitivities are:

$$\frac{\partial y}{\partial c_j}(0) = 0.$$

To estimate the solution to y and $\frac{\partial y}{\partial c_j}$, we solve the system of equations (3.14) and (3.15). Utilizing this approach to RRE, we take a partial derivative of (3.13) with respect to c_l and obtain:

$$\frac{d}{dt} \frac{\partial y}{\partial c_l} = \sum_{j=1}^M \nu_j \left(\frac{\partial a_j(y(t))}{\partial y} \frac{\partial y}{\partial c_l} + \frac{\partial a_j(y(t))}{\partial c_l} \right) \quad (3.16)$$

Since ν_j is independent of the kinetic parameter c_j and thus can be taken outside the bracket. This method is applicable to all RRE models and is known as the forward sensitivity analysis.

Combining the systems of ordinary differential equations (3.13) and (3.16), we can solve for y and $\frac{\partial y}{\partial c_j}$. Depending on the type of model we have chosen, we could use a non-stiff or stiff ode solver to obtain the solutions. In non-stiff systems we have used ode45, while in stiff systems, ode15s was employed.

Chapter 4

Numerical Results

We shall demonstrate the advantages of the sensitivity analysis methods developed above on three examples of realistic biochemical systems. The goal of this thesis is to illustrate the accuracy of the sensitivity analysis methods for both non-stiff and stiff systems. We define a stiff system as a system that has well separated time scales, i.e. the propensity functions differ drastically for the reactions involved, with the faster time scale being stable. The presence of the fast stable modes accounts for the initial conditions resulting in trajectories that, after a short rapid transient, lead to a stable manifold. In the stable manifold, the slow modes are responsible for the determination of dynamics of the system, while the fast modes are no longer present [45]. Stiff systems are computationally challenging, requiring the choice of very small time-steps to maintain accuracy.

4.1 Simple population model

Consider the following system of biochemical reactions [50], involving two reactions:

The propensities and state-change vectors corresponding to this biochemical system are given in Table 4.1.

Reaction channel	Reaction propensity	State-change vector
R_1	$a_1(X) = c_1 X_1$	$\nu_1 = (-1, 1, 0)^T$
R_2	$a_2(X) = c_2 X_2$	$\nu_2 = (0, -1, 1)^T$

Table 4.1: The Expected value of a population of species reaction model.

- The parameters are: $C = [2, 1]$.
- The initial conditions are: $X(t_0) = [100, 10, 1]$.
- The vector of propensities is $a(X) = [c_1 X_1, c_2 X_2]^T$.

Figure 4.1: Mean sensitivities for the population model: CFD means vs CLE means vs RRE. ($\tau = 10^{-3}$)

Simulations used for this model: For all the sensitivity methods for stochastic models of this biochemical system, 10000 trajectories were simulated on the time interval $[0, 1]$. The step size $\tau = 10^{-3}$ was used for the first and fourth order CFD method for the CME, and the pathwise sensitivity method for the CLE. The $\epsilon = 0.2$ was used for the CFD and the CRP methods. Also, the sensitivity for the Reaction Rate Equations were computed on the time interval $[0, 1]$.

Method	Mean Sensitivity: X_1	Mean Sensitivity: X_2	Mean Sensitivity: X_3
CRP	-12.3560	3.2045	9.1515
1 st order CFD	-12.2420	3.0670	9.1750
4 th order CFD	-13.6563	3.8490	9.8072
CLE	-13.5382	3.8139	9.7243
RRE	-13.5131	3.7943	9.7188

Table 4.2: Mean sensitivity: $\tau = 1 \times 10^{-3}$ at $t = 1$ second.

Figure 4.2: Magnification of the mean plot: CFD 4th order vs CLE vs RRE. ($\tau = 10^{-3}$)

Figure 4.3: Standard deviation of the sensitivities for the population model: CFD vs CLE std. ($\tau = 10^{-3}$)

Method	Standard deviation: X_1	Standard deviation: X_2	Standard deviation: X_3
CRP	13.5201	22.5232	20.7807
1 st order CFD	4.2656	5.4664	4.9691
4 th order CFD	3.1737	3.9196	3.6228
CLE	1.5226	1.2100	0.6245
RRE	N/A	N/A	N/A

Table 4.3: Standard deviation of sensitivity: $\tau = 1 \times 10^{-3}$ at $t = 1$ second.

Population model system results In Figure 4.1 we plotted the mean sensitivities for each reacting species, with respect to the kinetic parameter c_1 . For the Coupled Finite Difference and the Chemical Langevin Equation sensitivities, identical time step of $\tau = 0.001$ was used. Comparing the sensitivity analysis methods for this model shows an excellent agreement. The fourth order CFD method provides a much more accurate mean sensitivity estimate than the first order CFD, matching the CLE and the RRE methods almost exactly, as seen in Figure 4.2. As evident from the Figure 4.3, the fourth order CFD method also has a significantly lower variance than the first order method, thus requiring fewer simulations to be accurate. Table 4.2 allows us to compare the mean sensitivities for the methods of interest, including the CRP, once again showing an excellent agreement between the methods and see that the CRP is the least accurate one among them. Table 4.3 confirms the advantage of the fourth order CFD method vs the first order CFD method.

It is important to note that while the sensitivity method for the CLE generally applies to systems with large population numbers, the sensitivity strategies for the CME apply to larger classes of problems. In particular, the CFD techniques apply to any system modelled using the CME, for which the tau-leaping method may be employed.

4.2 Stiff Non-linear System 1

We analyse below a stiff, non-linear reaction systems [2], represented via the following reaction network:

with kinetic parameters $c_1 = 10^3$, $c_2 = 10^3$, $c_3 = 10^{-5}$, $c_4 = 10$, $c_5 = 1.0$ and $c_6 = 10^6$. The initial conditions are: $S_1(0) = 10^3$, $S_2(0) = 10^3$ and $S_3(0) = 10$. The propensities and state-change vectors corresponding to this biochemical system are given in Table 4.4.

Reaction channel	Reaction propensity	State-change vector
R_1	$a_1(X) = c_1 X_1 X_2$	$\nu_1 = (-1, -1, 1)^T$
R_2	$a_2(X) = c_2 X_3$	$\nu_2 = (1, 1, -1)^T$
R_3	$a_3(X) = c_3 X_1 X_3$	$\nu_3 = (-1, 1, -1)^T$
R_4	$a_4(X) = c_4 X_2$	$\nu_4 = (1, -1, 1)^T$
R_5	$a_5(X) = c_5 X_2 X_3$	$\nu_5 = (1, -1, -1)^T$
R_6	$a_6(X) = c_6 X_1$	$\nu_6 = (-1, 1, 1)^T$

Table 4.4: The first stiff non-linear biochemical reaction model.

Simulations used for this model: For the first stiff model, the first and fourth order coupled finite difference methods for the CME, the CRP and the pathwise sensitivity technique for the CLE were simulated with 10000 trajectories on the time interval $[0, 0.01]$. The step size $\tau = 10^{-5}$ was used for the first and fourth order CFD, and the CLE sensitivity strategies. In addition, the Reaction Rate Equations method sensitivity method was applied on the time interval $[0, 0.01]$. The $\epsilon = 0.02$ was used for the CFD and the CRP methods.

Figure 4.4: Mean sensitivities for the Stiff model 1: CFD means vs CLE means vs RRE. ($\tau = 10^{-5}$)

Method	Mean Sensitivity: X_1	Mean Sensitivity: X_2	Mean Sensitivity: X_3
CRP	-62.9500	-65.7600	62.8200
1 st order CFD	-65.7000	-66.2000	65.6000
4 th order CFD	-65.2750	-66.5750	65.1250
CLE	-63.8920	-65.5556	63.7684
RRE	-63.9100	-65.5620	63.771

Table 4.5: Mean sensitivity: $\tau = 1 \times 10^{-5}$ at $t = 0.01$ second.

Figure 4.5: Magnification of the mean plot: CFD vs CLE vs RRE. ($\tau = 10^{-5}$) Figure 4.6: Standard deviation of the sensitivities for the stiff model 1: CFD vs CLE std. ($\tau = 10^{-5}$)

Method	Standard deviation: X_1	Standard deviation: X_2	Standard deviation: X_3
CRP	77.6500	85.4238	77.6104
1 st order CFD	33.9804	35.1966	34.0264
4 th order CFD	23.5925	24.4999	23.5300
CLE	1.5314	1.2158	0.6272
RRE	N/A	N/A	N/A

Table 4.6: Standard deviation of sensitivity: $\tau = 1 \times 10^{-5}$ at $t = 0.01$ second.

Stiff system 1 results For the first stiff system, we plotted the mean sensitivities for each reacting species, with respect to kinetic parameter c_1 , as seen in Figure 4.4. For the Coupled Finite Difference and the Chemical Langevin Equation sensitivities, identical time step of $\tau = 10^{-5}$ was used. Comparing the sensitivity analysis methods for this model shows an excellent agreement. For this model, the fourth order CFD method and the first order CFD method both provided a similarly accurate mean sensitivity estimate. The fourth order CFD method matched the CLE and the RRE methods almost exactly, as seen in Figure 4.5. As evident from the Figure 4.6, the fourth order CFD method has a lower variance than the first method, making it the method of choice for future applications. Table 4.5 allows us to compare the mean sensitivities for the methods of interest, including the CRP, once again showing an excellent agreement between the methods. Table 4.6 confirms the advantage of the fourth order CFD method vs the first order CFD method and the CRP.

4.3 Stiff Non-linear System 2

The second stiff, non-linear reaction systems [45] consists of the following biochemical reaction system:

with kinetic parameters $c_1 = 1$, $c_2 = 10$, $c_3 = 10^3$ and $c_4 = 0.1$, with the initial conditions $S_1(0) = 400$, $S_2(0) = 798$, and $S_3(0) = 1$. The propensities and state-change vectors corresponding to this biochemical system are given in Table 4.7.

Reaction channel	Reaction propensity	State-change vector
R_1	$a_1(X) = c_1 X_1$	$\nu_1 = (-1, 0, 0)^T$
R_2	$a_2(X) = c_2 X_1(X_1 - 1)$	$\nu_2 = (-2, 1, 0)^T$
R_3	$a_3(X) = c_3 X_2$	$\nu_3 = (2, -1, 0)^T$
R_4	$a_4(X) = c_4 X_2$	$\nu_4 = (0, -1, 1)^T$

Table 4.7: The stiff non-linear biochemical reaction model.

Simulations used for this model: The sensitivity techniques for the second stiff model are tested the same way. The first and fourth order coupled finite difference method, and the CRP scheme for the CME and the pathwise sensitivity method for the CLE are simulated with 10000 trajectories on the time interval $[0, 0.01]$. The first order and the fourth order CFD strategies and the CLE sensitivity scheme were applied using the same step size $\tau = 10^{-5}$. Moreover, the sensitivity was estimated by employing the Reaction Rate Equations sensitivity scheme for the time interval $[0, 0.01]$. The $\epsilon = 0.1$ was used for the CFD and the CRP methods.

Figure 4.7: Mean sensitivities for the Stiff model 2: CFD means vs CLE means vs RRE. ($\tau = 10^{-5}$)

Method	Mean Sensitivity: X_1	Mean Sensitivity: X_2	Mean Sensitivity: X_3
CRP	-18.6350	9.5600	-0.0410
1 st order CFD	-17.7330	8.9600	0.0090
4 th order CFD	-17.6140	8.8905	0.0075
CLE	-17.7021	8.9297	0.0087
RRE	-17.70181	8.9310	0.0088

Table 4.8: Mean sensitivity: $\tau = 1 \times 10^{-5}$ at $t = 0.01$ second.

Figure 4.8: Magnification of the mean plot: CFD vs CLE vs RRE. ($\tau = 10^{-5}$) Figure 4.9: Standard deviation of the sensitivities for the stiff model 2: CFD vs CLE std. ($\tau = 10^{-5}$)

Method	Standard deviation: X_1	Standard deviation: X_2	Standard deviation: X_3
CRP	198.1781	100.0408	6.8376
1 st order CFD	7.4415	3.7842	0.2999
4 th order CFD	5.1594	2.6263	0.1935
CLE	0.6015	0.3009	0.0050
RRE	N/A	N/A	N/A

Table 4.9: Standard deviation of sensitivity: $\tau = 1 \times 10^{-5}$ at $t = 0.01$ second.

Stiff system 2 results For the second stiff system, we first tested the mean sensitivities for each reacting species with respect to the kinetic parameter c_1 . The mean sensitivities after 10000 runs came out very nearly zero, indicating that c_1 is not one of the primary parameters in this model. Testing the model against c_2 , however, produced non-zero results, as seen in Figure 4.7. For the Coupled Finite Difference and the Chemical Langevin Equation sensitivities, an identical time step of $\tau = 10^{-5}$ was used. The mean sensitivities for this model show an excellent agreement for all methods. Once again, the fourth order CFD method and the first order CFD method both provided a similarly accurate mean sensitivity estimate. Both the first and the fourth order CFD methods matched the CLE and the RRE methods almost exactly, as seen in Figure 4.8. The fourth order CFD once again had a lower standard deviation compared to the first order method, as per Figure 4.9. The CFD methods have significantly lower standard deviation than the CRP, thus leading to much more efficient simulations for the same accuracy of the numerical solution. We once again confirmed that the fourth order CFD has significantly lower variance than the first method, confirming it as the method of choice for future research. Table (4.8) allows us to compare the mean sensitivities for the methods of interest, including CRP, once again showing an excellent agreement between the methods. Table (4.9) further confirms the advantage of the fourth order CFD method vs the first order CFD method.

Chapter 5

Conclusion

Stochastic modelling and simulation have been extensively employed for studying significant problems in Computational and Systems Biology, such as genetic regulatory networks, signalling pathways and cellular dynamics. This thesis considered some widely used stochastic models of well-stirred biochemical systems, namely the Chemical Master Equation and the Chemical Langevin Equation. The Chemical Master Equation is a stochastic discrete model which may be applied to generic biochemical systems, in particular to those with some low molecular amounts of some species. Often for such problems, the deterministic continuous model of the reaction rate equations fails to accurately predict the system behaviour. For the systems with relatively large population numbers for all species, the Chemical Master Equation may be reduced to the continuous stochastic model of the Chemical Langevin Equation.

The focus of this thesis was to design and analyse parametric sensitivity methods for the Chemical Master Equation. Sensitivity analysis is a powerful tool for studying biochemical systems, utilized among others for model reduction and validation. We proposed a new fourth order Coupled Finite Difference method for estimating sensitivities for the CME and compared to the existing sensitivity methods for this model and for the Chemical Langevin Equation or reaction rate equations models, when possible. Our tests in several biochemical systems of interest show that the new method is very accurate. We analysed both stiff and non-stiff problems with excellent results.

Our future work will analyse semi-implicit parametric sensitivity strategies for the Chemical Master Equation and the Chemical Langevin Equation.

Appendix A

Markov Processes

To study the chemical kinetics in general, and the dynamics of genetic and biochemical networks in particular, certain level of familiarity with the theory of stochastic processes is mandatory. A stochastic process is a random variable, i.e. the state of the biochemical network, which evolves through time. The state could be continuous or discrete, and it can may evolve through time in a continuous or discrete way. A Markov process is a stochastic process which possesses the property that the future behaviour depends only on the current state of the system. The dynamic behaviour of the biochemical networks can be effectively modelled by a Markov process, so we will discuss the theory below.

A.1 Discrete State Markov Process

The set $\{\theta^{(t)} | t = 0, 1, 2, \dots\}$ is a *discrete time stochastic process*. The state space S is such that $\theta^{(t)} \in S, \forall t$ and may be discrete or continuous.

Definition 5. A (first order) Markov chain is a stochastic process with the property that the future states are independent of the past states given the present state. Formally, for $A \subseteq S, n = 0, 1, 2, \dots$, we have

$$\begin{aligned} P\left(\theta^{(n+1)} \in A | \theta^{(n)} = x, \theta^{(n-1)} = x_{n-1}, \dots, \theta^{(0)} = x_0\right) \\ = P\left(\theta^{(n+1)} \in A | \theta^{(n)} = x\right), \end{aligned}$$

$$\forall x, x_{n-1}, \dots, x_0 \in S$$

Since the past states provide no information about the future state if the present state is known, the behaviour of the chain is determined by $P(\theta^{(n+1)} \in A | \theta^{(n)} = x)$. This generally depends on A, n and x . However, if there is no n dependence, so that

$$P\left(\theta^{(n+1)} \in A | \theta^{(n)} = x\right) = P(x, A), \forall n,$$

then the Markov chain is said to be (time) *homogeneous*, and the *transition kernel*, $P(x, A)$ determines the behaviour of the chain. Note that $\forall x \in S$, $P(x, \cdot)$ is a probability measure over S .

A.2 Continuous State Markov Process

When dealing with biochemical networks, we will generally assume them to have a discrete state. However, sometimes it is useful to regard the state of certain quantities as continuous. To do that, we need to extend the concept of discrete state Markov chain to the continuous state case. This extension is analogous to the generalization of discrete random quantities to that of continuous random quantities. We will still be working with discrete time, but we are allowing the state space S of the Markov chain to be continuous, i.e. $S \in \mathbb{R}$

For the homogeneous chain, we can define

$$P(x, A) = P\left(\theta^{(n+1)} \in A | \theta^{(n)} = x\right).$$

For continuous state spaces we always have $P(x, \{y\}) = 0$, so in this case we define $P(x, y)$ by

$$\begin{aligned} P(x, y) &= P\left(\theta^{(n+1)} \leq y | \theta^{(n)} = x\right) \\ &= P\left(\theta^{(1)} \leq y | \theta^{(0)} = x\right), \forall x, y \in S, \end{aligned}$$

the conditional cumulative distribution function (CDF). This is the distributional form of the transition kernel for the continuous state space Markov chains. We can also define the corresponding conditional density

$$p(x, y) = \frac{\partial}{\partial y} P(x, y)$$

Where $x, y \in S$. This can be used to define the density form of the transition kernel of the chain. The $p(x, y)$ is the conditional density for the next state (with variable y) given that the current state is x , and so it could be written as $p(y|x)$.

References

- [1] M.A. Altinkaya and E.E. Kuruoglu. Modeling enzymatic reactions via chemical langevin-levy equation. *20th Signal Processing and Communications Applications Conference (SIU)*, 2012.
- [2] D.F. Anderson. Incorporating postleap checks in tau-leaping. *American Institute of Physics*, 128(1):054103, 2007.
- [3] D.F. Anderson. An efficient finite difference method for parameter sensitivities of continuous time markov chains. *SIAM Journal on Numerical Analysis*, 50(5):2237 - 2258, 2012.
- [4] D.F. Anderson, B. Ermentrout, and P.J. Thomas. Stochastic representations of ion channel kinetics and exact stochastic simulation of neuronal dynamics. *Journal of Computational Neuroscience*, 38:67-82, 2015.
- [5] U.M. Ascher and L.R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Cambridge University Press, 1999.
- [6] W.J. Blake, M. Kaern, C.R. Cantor, and J.J. Collins. Noise in eukaryotic gene expression. *Nature*, 422:633-637, 2003.
- [7] Y. Cao, D.T. Gillespie, and L. Petzold. The slow scale stochastic simulation algorithm. *Journal of Chemical Physics*, 122:014116, 2005.
- [8] C. Chapon. Role of the catabolite activator protein in the maltose regulon of escherichia coli. *Journal of Bacteriology*, 150(2):722-729, 1982.
- [9] J.C. Dunlap. Molecular basis for circadian clocks. *Cell*, 96:271-290, 1999.
- [10] R. Edwards, R. Illner, and V. Paetkau. A model for generating circadian rhythm by coupling ultradian oscillators. *Theoretical Biology and Medical Modeling*, 3:12, 2006.
- [11] M.B. Elowitz, A.J. Levine, E.D Siggia, and P.S. Swain. Stochastic gene expression in a single cell. *The Journal of Chemical Physics*, 10, 2002.
- [12] M.B. Elowitz and Leibler S. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335-338, 2000.

- [13] A. Alfonsi et al. Adaptive simulation of hybrid stochastic and deterministic models for biochemical systems. *ESAIM*, 14:1-13, 2005.
- [14] D. Yu et al. An efficient recombination system for chromosome engineering in escherichia coli. *Proceedings of the National Academy of Sciences of U.S.A*, 97(11):5978-5983, 2000.
- [15] O. Yarchuk et al. Stochastic gene expression in a single cell. *Journal of Molecular Biology*, 226:581-596, 1992.
- [16] T.R. Kiehl et al. Hybrid simulation of cellular behavior. *Bioinformatics*, 20:316-322, 2004.
- [17] D.B. Forger and C.S. Peskin. Stochastic simulation of the mammalian circadian clock. *Proc Natl Acad Sci USA*, 102:321-324, 2005.
- [18] D.T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 2:403-434, 1976.
- [19] D.T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340-2361, 1977.
- [20] D.T. Gillespie. A rigorous derivation of the chemical master equation. *Statistical Mechanics and its Applications*, 188:404-425, 1992.
- [21] D.T. Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297, 2000.
- [22] D.T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting species. *Journal of Chemical Physics*, 115:1716-1733, 2001.
- [23] D.T. Gillespie. Stochastic simulation of chemical cinetics. *Annual Review of Physical Chemistry*, 58:35-55, 2007.
- [24] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, 2004.
- [25] D. Gonze, J. Halloy, J.C. Leloup, and A. Goldbeter. Stochastic models for circadian rhythms: Effect of molecular noise on periodic and chapotic behaviour. *C R Biol*, 326:189-203, 2003.
- [26] R. Gunawan, L. Petzold Y. Cao, and F. J. Doyle III. Sensitivity analysis of discrete stochastic systems. *Biophysical Journal*, 88:2530-2540, 2005.
- [27] E.L. Haseltine and J.B. Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *Journal of Chemical Physics*, 117(15):6959-6969, 2002.
- [28] A. Hellander and P. Lötstedt. Hybrid method for the chemical master equation. *Journal of Computational Physics*, 227(1):100-122, 2007.
- [29] D.J. Higham. Modeling and simulating chemical reactions. *SIAM Reviews*, 50:, 347-368, 2008.

- [30] S. Ilie, W.H. Enright, and K.R. Jackson. Numerical solution of stochastic models of biochemical kinetics. *Canadian Applied Mathematics Quarterly*, 17(3):523-554, 2009.
- [31] R. Khanina and D.J. Higham. Chemical master equation and langevin regimes for a gene transcription model. *Theoretical Computer Science*, 408(1):31-40, 2008.
- [32] T.R. Kiehl, R.M. Mattheyses, and M.K. Simmons. Hybrid simulation of cellular behavior. *Bioinformatics*, 20:316-322, 2004.
- [33] P.E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Springer-Verlag Berlin Heidelberg New York, 1992.
- [34] J.C. Leloup and A. Goldbeter. Toward a detailed computational model for the mammalian circadian clock. *Proc Natl Acad Sci USA*, 100:7051-7056, 2003.
- [35] R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in escherichia coli via the lacr/o, the tetr/o and arac/i1-i2 regulatory elements. *The Journal of Chemical Physics*, 25(3):1203-1210, 1997.
- [36] S. MacNamara, A.M. Bersani, K. Burrage, and R.B. Sidje. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the stochastic simulation algorithm and chemical master equation. *Journal of Chemical Physics*, 129:095105, 2008.
- [37] T. Mattheyses and M. Simmons. Hybrid simulation of cellular behavior. *Bioinformatics*, 20:316-322, 2004.
- [38] H.H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Journal of Biochemistry*, 94:814-819, 1997.
- [39] H.H. McAdams, A. Arkin, and J.Ross. Stochastic mechanisms in gene expression. *Genetics*, 149:1633-1648, 1998.
- [40] B.J. Meyer, R. Maurer, and M.Ptashne. Gene regulation at the right operator (or) bacteriophage lambda. i. or3 and autogenous negative control by repressor. *Journal of Molecular Biology*, 139(2):147, 1980.
- [41] I. Mihalcescu, W. Hsing, and S. Leibler. Resilient circadian oscillator revealed in individual cyanobacteria. *Nature*, 430:81-85, 2004.
- [42] J. Puchalka and A.M. Kierzek. Bridging the gap between stochastic and deterministic regimes in the kinetic simulations. *Biophysics*, 86:1357-1372, 2004.
- [43] C.V. Rao and A.P. Arkin. Stochastic chemical kinetics and the quasi-state assumption: Application to the gillespie algorithm. *Journal of Chemical Physics*, 118:4999-5010, 2003.
- [44] J.M. Raser and E.K. O'Shea. Noise in gene expression: Origin, consequences, and control. *Science*, 309:2010-2013, 2005.

-
- [45] M. Rathinam, L.R. Petzold, Y. Cao, and D.T. Gillespie. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *The Journal of Chemical Physics*, 119:12784, 2003.
- [46] M. Rathinam, P.W. Sheppard, and M. Khammash. Efficient computation of parameter sensitivities of discrete stochastic reaction networks. *The Journal of Chemical Physics*, 136:034115, 2012.
- [47] H. Salis and Y. Kaznessis. Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *Journal of Chemical Physics*, 122(5):054103, 2005.
- [48] A. Samant and D. Vlachos. Overcoming stiffness in stochastic simulation stemming from partial equilibrium: a multiscale monte-carlo algorithm. *Journal of Chemical Physics*, 123:144114, 2005.
- [49] U. Schibler and F. Naef. Cellular oscillators: Rhythmic gene expression and metabolism. *Current Opinion Cell Biology*, 53:401-417, 2005.
- [50] R. Srivasta, D.A. Anderson, and J.B. Rawlings. Comparison of finite difference based methods to obtain sensitivities of stochastic chemical kinetics models. *The Journal of Chemical Physics*, 138:074110-1, 2013.
- [51] T. Tian, P.M. Burrage, K. Burrage, and M. Carletti. Stochastic delay differential equations for genetic regulatory networks. *Journal of Computational and Applied Mathematics*, 205(2):696-707, 2007.
- [52] A. Varma, M. Morbidelli, and H. Wu. *Parametric Sensitivity in Chemical Systems*. Cambridge University Press, 1999.
- [53] J.M. Vilar, H.Y. Kueh, N. Barkai, and S. Leibler. Mechanisms of noise-resistance in genetic oscillators. *Proc Natl Acad Sci USA*, 99:5988-5992, 2002.
- [54] D.J Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman and Hall, CRC Press, 2012.