

# NUMERICAL STUDIES OF IMPLICIT TAU LEAPING METHODS FOR STOCHASTIC BIOCHEMICAL SYSTEMS

by

Fauzia Jabeen

Master of Science in Bioinformatics, Muhammad Ali Jinnah, 2013

Master of Science in System Engineering, Quaid-i-Azam, 2000

Master of Science in Physics, Punjab, 1997

Bachelor of Science, Punjab, 1994

A thesis

presented to Ryerson University

in partial fulfillment of the  
requirements for the degree of

Master of Science

in the program of

Applied Mathematics

Toronto, Ontario, Canada, 2018

©Fauzia Jabeen, 2018

## **Author's Declaration**

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.

# **Numerical studies of Implicit Tau Leaping methods for stochastic biochemical systems**

Master of Science, 2018

Fauzia Jabeen

Applied Mathematics

Ryerson University

## **Abstract**

Deterministic models of chemical reactions systems have been used successfully in studying chemical kinetics problems. However, in biochemical systems (e.g. cellular systems in biology), small molecular population sizes of some key reacting species can lead to results that cannot be predicted by the traditional deterministic models. It has been found that such processes involve intrinsic randomness that can be better modeled by stochastic models. Chemical Master Equation (CME) is an accurate stochastic model of well-stirred biochemical systems. We investigate reliable and efficient simulation methods for the CME, namely the implicit tau-leaping method. The tau-leaping algorithms were tested on several models of practical interest such as the Schlögl model and the Goldbeter-Koshland switch and compared to the exact methods. We observed that, for systems not reaching steady state, the implicit tau-leaping strategy is accurate.

## Acknowledgements

I would like to bow my head before Allah Almighty, the Most Gracious, and the Most Merciful, whose benediction bestowed upon me noble parents, talented teachers, provided me sufficient opportunity, and enabled me to undertake and carry out this research work

This dissertation was undertaken at the Faculty of Science, Ryerson University under the supervision of Dr. Silvana Ilie, to whom I am indebted for suggesting the subject and for her encouragement, guidance and support throughout the course of this thesis. Her advice, discussion and effective comments have always been a source of motivation for me.

I would like to thanks to all wonderful people at mathematics department at Ryerson University, particularly Dr. Katrin Rohlf, Dr. Foivos Xanthos and Dr. Pablo Olivers. Your knowledge and support helped me through my time in Ryerson University. Also thanks to Teresa Lee, Luisa Chan, Cadene Henry, Kathy Peter and Steve Kanellis for their administrative support.

A special thanks to my husband Dr. Anwar Majid Mirza and our loving kids Misha and Ibrahim for their love, continuous support, prayers and encouragement. This accomplishment would have not been possible without them.

## Dedication

*Dedicated to my beloved parents, Muhammad Bahadur Khan and Nazir Begum  
(May Allah grant them a high place in Janat-ul-Firdous- Ameen)*

# Table of Contents

<b>Author’s Declaration</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Dedication</b>	<b>v</b>
<b>List of Tables</b>	<b>viii</b>
<b>List of Figures</b>	<b>ix</b>
<b>List of Appendices</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Stochastic vs. Deterministic Modelling . . . . .	2
1.2 Literature Survey . . . . .	3
1.3 Thesis Focus and the Findings . . . . .	6
1.4 Outline of the Thesis . . . . .	7
<b>2 Background</b>	<b>8</b>
2.1 Well-Stirred Biochemical Systems . . . . .	8
2.2 Stochastic and Deterministic Modeling . . . . .	10
2.2.1 Markov Processes . . . . .	13
2.3 Chemical Master Equation (CME) . . . . .	17
2.3.1 Stochastic Simulation Algorithm (SSA) . . . . .	19
2.3.2 Tau-Leaping Method . . . . .	26

2.4	Chemical Langevin Equation . . . . .	27
2.4.1	Stochastic Differential Equations . . . . .	27
2.4.2	Derivation of the Chemical Langevin Equation . . . . .	30
2.4.3	Euler-Maruyama Method . . . . .	32
2.5	Reaction Rate Equation (RRE) . . . . .	34
<b>3</b>	<b>Tau-Leaping Methods</b>	<b>43</b>
3.1	What is Tau-Leaping? . . . . .	43
3.2	Explicit Tau-Leaping . . . . .	45
3.2.1	Algorithm for Explicit tau-Leaping . . . . .	47
3.2.2	How accurate is the explicit tau-leaping strategy? . . . . .	47
3.3	Implicit Tau-Leaping . . . . .	48
3.3.1	Stiffness in Continuous Deterministic ODE Systems . . . . .	49
3.3.2	Stiffness in Discrete-Stochastic Chemical Kinetics: Implicit Tau-Leaping Method . . . . .	50
3.3.3	Implicit Tau-Leaping Algorithm . . . . .	51
3.4	Trapezoidal Tau-Leaping Method . . . . .	53
3.4.1	Trapezoidal Tau-Leaping Algorithm . . . . .	54
<b>4</b>	<b>Numerical Results and Discussion</b>	<b>55</b>
4.1	The Schlögl Model . . . . .	55
4.2	The Goldbeter-Koshland Switch . . . . .	66
4.3	Decay Dimerization . . . . .	74
<b>5</b>	<b>Conclusion and Future Work</b>	<b>77</b>
	<b>Appendix A</b>	<b>79</b>
	<b>References</b>	<b>85</b>

# List of Tables

2.2	Comparison of the CME and the CLE Models . . . . .	34
2.3	RRE and Rate constants for different types of reactions . . . . .	39
4.1	The reactions and propensities in the Schlögl model. . . . .	56
4.2	The Goldbeter-Koshland switch parameters. . . . .	67
4.3	The toggle switch parameters. . . . .	74



# List of Figures

2.1	Time intervals on a sample trajectory path . . . . .	21
3.1	A schematic diagram of the SSA trajectory for $X_k(t)$ . . . . .	44
4.1	The deterministic rate of change of molecules with time using the RRE for the Schlögl model, indicating the roots of the rate function. . . . .	58
4.2	Continuous, deterministic solution of the Schlögl model showing bi-stability of the system. . . . .	59
4.3	The RRE solutions of the Schlögl model using different initial conditions. . .	60
4.4	Several trajectories of the Schlögl model simulated with the SSA. . . . .	61
4.5	Markov jumps in an SSA solution of the Schlögl model. . . . .	61
4.6	SSA solutions of the Schlögl model with different initial conditions. . . . .	63
4.7	The state switching in some SSA trajectories for the Schlögl model. Only 10 trajectories are shown. . . . .	64
4.8	The state switching in explicit tau-leaping trajectories of the Schlögl model. 10 trajectories are shown. . . . .	64
4.9	Schlögl model: Comparison of the probability distributions at $T = 4$ obtained using 10,000 simulations with the SSA and the $\tau$ -leaping methods. . . . .	65
4.10	Schlögl Model: Comparison of the mean values as functions of time, for 10,000 trajectories using the SSA and the $\tau$ -leaping methods. . . . .	65
4.11	Schlögl Model: Comparison of the standard deviation values as functions of time for 10,000 trajectories using the SSA and the $\tau$ -leaping methods. . . . .	66
4.12	Goldbeter Koshland switch - Ten trajectories for the substrate $S$ ( $X_1$ ) as a function of time, using the SSA and the tau-leaping methods. . . . .	68

4.13	Goldbeter Koshland switch - Ten trajectories for the product $P(X_4)$ as a function of time, using the SSA and the tau-leaping methods. . . . .	69
4.14	Goldbeter-Koshland switch - Comparison of probability distributions for all 6 species using $10^4$ trajectories for each of the SSA, explicit, implicit and trapezoidal tau-leaping methods computed at $T = 5$ . . . . .	71
4.15	Goldbeter-Koshland switch - Comparison of the mean values for all 6 species as functions of time, using $10^4$ trajectories for each of the SSA, explicit, implicit and trapezoidal tau-leaping methods. . . . .	72
4.16	Goldbeter-Koshland switch - Comparison of the standard deviation values for all 6 species as functions of time, using $10^4$ trajectories for each of the SSA, explicit, implicit and trapezoidal tau-leaping methods. . . . .	73
4.17	Decay Dimerization - SSA trajectories. . . . .	75
4.18	Decay Dimerization - Comparison of the probability distribution of $S_1$ using 10,000 trajectories of the SSA and the implicit tau-leaping. . . . .	76
4.19	Decay Dimerization - Comparison of the probability distribution of $S_2$ using 10,000 trajectories of the SSA and the implicit tau-leaping. . . . .	76

# List of Appendices

Appendix A - Propositions

79

# Chapter 1

## Introduction

The last few decades have witnessed a tremendous growth in the field of biological sciences, especially at the cellular level. A lot of attention has been paid to biochemical processes taking place at the molecular level. For example, “gene expression” is a process in which the hereditary information stored in sequence of DNA, called genes, is transcribed on and transported through a messenger RNA (mRNA) and eventually translated to produce proteins or RNA [1]. Proteins are macro-molecules performing many different functions in different cells. A single mRNA can be used a number of times to produce different types of proteins. It is interesting to note that the nucleus of every somatic cell (i.e. every cell in a macroorganism except the reproduction cells) contains the entire genome for that organism. Red blood cells (RBCs), muscle cells, neurons are all different types of cells, but the nuclei of all those cells contain the entire genome for that individual. Gene expression processes taking place in the nucleus of each of those different cells produces only those proteins that are required for the structural integrity and functioning of that cells. Hemoglobin (a protein used in RBCs to carry oxygen) is only produced in RBCs and not in the neurons. In this sense, the gene expression processes taking place within the nuclei are highly regulated processes.

In 1976, Spudich and Koshland [22], while working on isogenic (genetically identical) bacteria grown in homogeneous nutrient and environment conditions, found that individual

bacteria retain individual (and yet different) swimming patterns throughout their life cycle. They attributed this phenomenon to the Poissonian variations in the low number of molecules available during the cell division and gene expression processes. They estimated that there were only 6 – 14 mRNA molecules per bacteria cell. This was the first time the stochastic effects were observed experimentally, due to the intrinsic random noise in cellular processes involving very small number of molecules [32].

There has been a growing awareness and evidence that stochasticity plays an essential role in the biochemical processes taking place at the cellular level [29, 31, 23]. In 2002, Elowitz and his coworkers [25], in their landmark experiment were successful in separating and measuring intrinsic and extrinsic (external environmental) noise by studying two fluorescent (colored) proteins in an isogenic population of E-coli bacterial cells.

The stochasticity due to the random intrinsic noise in the low molecular population levels, plays an important role in biochemical reactions taking place at the cellular level [29, 30, 31]. There has been a great demand to develop mathematical models and computation approaches for the quantitative analysis and assessment of the biological (in-vivo) experiments. This is the main topic of this thesis.

## **1.1 Stochastic vs. Deterministic Modelling**

Mathematically models of the biochemical processes taking place at the cellular level, may be deterministic or stochastic. The former is based on the use of ordinary differential equations (ODEs) to represent time variations in the concentrations (moles per unit volume) of the reacting species and it is called the reaction rate equation (RRE). The later approach is more accurate for systems involving low molecular counts in some species. The effects of the intrinsic noise have not been considered significant enough to have any overall effect on the outcomes of chemical experiments [24, 3]. The deterministic approach was found to be sufficient for a variety of biochemical situations. This approach was used, for example,

for describing enzyme kinetics following chemical laws of mass action and rate equations, and is still taught in standard biochemistry textbooks [3, 4]. Such models have been found successful for systems with large number of molecules, involving concentrations in enzymatic reactions that take place in the cytoplasm or reacting solutions in test tubes. At this scale, the statistical noise can be ignored as compared to the average behavior of the molecules. Continuous real valued variables are used in this framework of deterministic modelling.

As the number of molecules becomes small, the intrinsic noise may increase. Due to the random nature of molecular interactions, the number of molecules of different species reacting biochemically changes abruptly. Stochastic and discrete variables need to be used at this scale. Stochastic models are required in situation arising e.g., in gene expression reactions [29]. In this thesis, we are interested in investigating these stochastic models at the scale with small number of molecules in some species.

Between these two extremes of the number of molecules, an intermediate scale exists. At this scale, the intrinsic noise is just becoming significant enough, while the system can still be measured in terms of continuous real valued variables. Continuous stochastic models based on stochastic differential equations are used at this intermediate scale [10].

## 1.2 Literature Survey

The field of biochemical kinetics (involving the study of the rates of biochemical reactions taking place in living organisms under different conditions) has been strongly associated with chemical kinetics in chemistry. Chemical reactions have been studied under laboratory (test tube) conditions in terms of concentrations in moles per volume [3, 4]. Ever since Ludwig Wilhelm (1850) introduced the use of ODEs to model chemical reaction kinetics for the first time, the continuous deterministic approach has been very successfully used to solve a large number of problems in chemistry [34, 13]. It is no surprise that this approach was considered to solve biochemical kinetic problems as well. In fact, probably, the first biochemical kinetic

problem involving enzymes, (known as the Michaelis-Menton model), was studied in 1913 using ODEs. Little attention was paid to the fact that chemical reactions involve discrete molecules that move around quite randomly interacting with each other through elastic collisions and sometimes via inelastic collisions involving chemical reactions.

Some work was initiated during 1950's and 60's to model chemical reactions using stochastic approach based on Markovian jump processes. Initial formulation of the Chemical Master Equation (CME) was carried out by McQuarrie in 1967 [35]. The CME is a set of differential equations giving the time rate of change of probabilities of a chemically reacting system being in each possible state of the system. A more rigorous mathematical treatment of the CME was provided by Gillespie in 1992 [36]. Oppenheim et.al. [37] in 1969 demonstrated and Kurtz [38] proved later in 1972, that the stochastic model based on a Markov chain converges to the deterministic ODE model under the thermodynamic limit of infinite volume  $\Omega$  and infinite number of molecules  $N$  while keeping their ratio  $N/\Omega$  and temperature fixed. The stochastic formulation based on the CME gives full description of the big chemically reacting system. At each point in time, each probabilistic differential equation represents one possible state of the system. The number of possible states can increase exponentially, making the solution of the CME very complex and mathematically intractable [32].

To simulate the CME, Gillespie proposed two Monte Carlo method based algorithms called the first reaction method in 1976 [39] and the direct method in 1977 [40]. The direct method is also called the Stochastic Simulation Algorithm (SSA) or Gillespie's algorithm. In the first reaction method, a random event time is generated for each reaction from the exponential distributions of each reaction channel. In the direct method, the time to next reaction along with the index of that reaction are simulated. Gillespie established that the two algorithms are equivalent. The SSA exactly simulates the CME in the sense that the probability distribution obtained by running very large (infinty) number of trajectories using the SSA will be identical to the one obtain from the CME. The SSA did not get much attention initially until the work of McAdams and Arkin [29] in 1997 established the importance

of stochastic mechanisms in gene expression, some two decades later. Gillespie’s algorithm has been extensively used ever since in biochemical kinetic problems [32]. The computational cost of Gillespie’s algorithm becomes prohibitively high in situations where chemical reactions take place at very fast rates for a large number of chemical species participating in the biochemical process. The algorithm simulates every reaction event, and the time taken between such events becomes very small.

A solution to reduce the high computational cost of the SSA was suggested by Gillespie [11] in 2001 via the use of the tau-leaping approach. The tau-leaping algorithm gives much faster results but at the cost of losing exactness of the SSA. A fixed time step  $\tau$  is chosen so that more reactions can take place during this time interval. The number of reactions of a certain type over this step interval can be estimated using a Poisson distribution with a parameter equal to the product of  $\tau$  and the reaction propensity. The time step  $\tau$  is chosen in such a way that the leap condition of very small variations in the propensity functions during this interval is ensured. Gillespie’s tau-leaping method is also called the explicit tau-leaping. This approximate approach is inefficient on stiff systems. Biochemical systems are often stiff, having both fast and slowly varying components. Usually, the fast reactions are transient and they settle down quite rapidly as compared to the slow reactions. To model them properly, a small tau-leap is required which slows down the explicit tau-leap simulation and increases the simulation time dramatically.

Rathinam et.al [15] proposed the implicit tau-leaping method in 2003 to deal with stiff systems, having both slow and fast time scales. They modified the explicit tau-leaping method by using the Poissonian stochastic term at the beginning of the time step and the deterministic term at the end of the step. A converged state of the system is evaluated iteratively using, for example, Newton’s method before moving on to the new time step. For the situations when the number of the molecules becomes very large and the stochastic component becomes less significant, the explicit tau-leaping method reduces to Euler’s method and the implicit tau leaping method reduces to the implicit Euler’s method for solving ODEs



[8].

Cao and Petzold (2005) found that the implicit tau-leaping damps the stochastic noise and reduces the variances for large time steps as compared to the SSA simulations [21]. They proposed a trapezoidal tau-leaping method based on the trapezoidal rule for solving ODEs.

Tau-leaping methods also provides the first step towards the link between discrete stochastic CME/SSA model and the continuous-deterministic RRE/ODE model. Under the assumptions that the time step  $\tau$  satisfies the leap condition (propensity functions variations are very small) and that the product of  $\tau$  with the propensity functions is large, the Poissonian stochastic term can be replaced by a term with a normal distribution. This leads to a system of stochastic differential equations (SDE) discretized using the Euler-Maruyama method [14]. This set of SDEs is called the Chemical Langevin Equation (CLE) and this model is applicable for systems with large molecular counts [13, 41].

If we consider a system where the stochastic intrinsic noise could be ignored due to the large number of molecules such that the ratio of the number of molecules with volume is fixed under constant thermodynamic conditions, then the CLE is reduced to a set of ODEs representing the reaction rate equations [13, 14]. This is inline with the original findings of Oppenheim et.al [37] and Kurtz [38].

### 1.3 Thesis Focus and the Findings

The focus of this thesis is to implement and study the stochastic simulation methods for biochemically reacting systems. Different tau-leaping approaches are compared with the SSA simulations.

Previously, it has been stated that the SSA/tau-leaping interlacing is important to maintain correct statistics of implicit tau-leaping method for all well-stirred biochemical systems, as the implicit tau-leaping method alone damps the noise in the fast system variables [15]. We made the following important observation.

*For models not reaching a steady state (e.g. bistable systems, systems with large noise) interlacing is not needed, as the Implicit tau-leaping strategy is accurate by itself.*

This observation has been tested by conducting simulations for several models including the Shlögl model, and the Goldbeter-Koshland switch. For models reaching a steady-state, such as the decay-dimerization system, the noise in the fast dynamics is reduced.

## 1.4 Outline of the Thesis

The outline of the thesis is as follows: Chapter 2 gives the detailed background for the CME, SSA, tau-leaping methods, CLE and RRE. Tau leaping methods are described in detail in Chapter 3. Numerical results are given for several models in Chapter 4. Chapter 5 discusses the results of the thesis and recommendations for future work.

# Chapter 2

## Background

In this chapter, we introduce the modeling of the chemical reactions taking place in biochemical systems under the well-stirred assumption. We discuss the widely used model of the Chemical Master Equation (CME) in detail, followed by its solution with exact Monte-Carlo Method known as the Stochastic Simulation Algorithm (SSA). The approximatic stochastic simulation technique for solving the CME, the tau-leaping method, is introduced. Different tau-leaping methods are presented in the next chapter. We also derive the Chemical Langevin Equation (CLE) and a numerical scheme for solving it, the Euler Maryama method. The reaction rate equation (RRE) model is derived for a system in thermodynamic limit.

### 2.1 Well-Stirred Biochemical Systems

We consider a biochemical system consisting of  $N$  chemical species  $\{S_1, S_2, \dots, S_N\}$ . These molecules interact with each other through  $M$  chemical reactions  $\{R_1, R_2, \dots, R_M\}$ . We assume that the molecules are confined to a constant volume  $\Omega$ , in thermal equilibrium, at fixed temperature  $T$ . The system is assumed well-stirred. The overall state of the biochemical

system can be represented by the *state vector*,

$$\mathbf{X}(t) = \begin{bmatrix} X_1(t) \\ X_2(t) \\ \vdots \\ X_N(t) \end{bmatrix} \quad (2.1)$$

where  $X_i(t)$  is a nonnegative integer giving the number of molecules of species of type  $S_i$  present at time  $t$ .

When a chemical reaction takes place, a molecule of species A may, for example, combine with a molecule of species B to produce a new molecule of species C. Depending on the nature of the chemical reaction, molecules of the same species can combine together or decompose into other molecules. Also the chemical reactions could be reversible. The various types of chemical reactions are discussed in Section 2.2. We assume that the number of molecules of each species present in the system at the initial time  $t = 0$  are given. In other words, the initial state of the system  $\mathbf{X}(t)$  at  $t=0$  is completely known. As the molecules move around the system, interacting with each other, their number changes, resulting into a change of the state of the system. Our aim is to describe how the number of molecules (and hence the state  $\mathbf{X}(t)$ ) evolve as time progresses.

To model this problem, in principle, both the position and velocity of each molecule changing with time should be considered under the laws of physics. In this regard, we need to keep track of every molecular collision and the outcome of every interaction. However, this approach becomes computationally too expensive, when the overall number of molecules becomes too large and/or when the molecular dynamics over a long period of time is of interest. To simplify, the biochemical system is considered to be well stirred. This means that we are ignoring the spatial positions and velocities of individual molecules in the system and are only interested in their counts at a particular instant  $t$  in time. The well-stirred system approximation leads us to consider that the majority of the molecular collisions are

non-reactive (elastic) in nature, resulting:

- in random but uniform distribution of molecules within the volume  $\Omega$  after the majority of collisions are elastic in nature, and
- in random velocities of the molecules following a Maxwell-Boltzmann distribution of thermal equilibrium.

This approximation greatly simplifies the problem. We, therefore, are able to ignore most of non-reactive molecular collisions and concern ourselves only with the events involving changes due to chemical interactions between the molecules.

## 2.2 Stochastic and Deterministic Modeling

Traditionally, deterministic formulations have been used to model the behaviour of biochemical systems with fair amount of success. This usually involves setting up of a set of  $N$  ordinary differential equations (ODEs) for the  $N$  chemical species participating in the biochemical process. Each ODE expresses the time rate change of the molecular concentrations (i.e. number of molecules per unit volume,  $\frac{X_i(t)}{\Omega}$ ) as a function of the concentration of all the species along with the chemical rate constants:

$$\frac{dX_i}{dt} = f_i(X_1, X_2, \dots, X_N) \quad (2.2)$$

These equations are called the "Reaction Rate Equations" (RREs). The analytic solution of these equations may be obtained only for a few modelled situations. For realistic problems involving large number of reacting species, numerical methods are employed to solve them computationally. The numerical solutions obtained for the deterministic model of the RRE are generally,

1. Single valued (same number of molecules per unit volume is obtained on repeated runs of the numerical solution).

2. Real valued (the molecular concentration and hence the number of molecules always turns out to be real valued rather than integer).
3. The number of molecules changes continuously as a function of time (this infers directly from the original ODE's that is a continuous deterministic model in time).

Both experimental and theoretical studies conducted during the late nineteen nineties and over the turn of century give conclusive evidence that statistical random noise must be taken into account for biochemical systems involving small number of big molecules at the cellular level [23, 28, 29, 30, 31].

**Remark.** *Before going into details of stochastic modeling of well-stirred biochemical reactions, some of the relevant terms are defined and explained here.*

The number of molecules corresponding to a particular species change over times, as different (or sometimes same) molecules react chemically. Each chemical reaction  $R_j$  can be characterized in terms of the state-change vector  $\nu_j$  and the propensity function  $a_j$  of the reaction.

**State Change Vector** In the *state change vector*

$$\nu_j = \begin{bmatrix} \nu_{1j}(t) \\ \nu_{2j}(t) \\ \vdots \\ \nu_{ij}(t) \\ \vdots \\ \nu_{Nj}(t) \end{bmatrix} \quad (2.3)$$

$\nu_{ij}$  represents the change in the  $S_i$ ,  $i$ th species number of molecules, as the chemical reaction  $R_j$  takes place. Therefore, if the system is in state  $\mathbf{x}$  and one  $R_j$  chemical reaction takes place, the state of the system changes from  $\mathbf{x}$  to  $\mathbf{x} + \nu_j$ .

**Propensity Function** The *propensity function* is defined as:

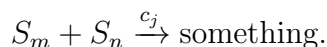
$a_j(\mathbf{x})dt$  = the probability, given that  $\mathbf{X}(t) = \mathbf{x}$ , such that one  $R_j$  will occur somewhere inside the volume  $\Omega$  in the next infinitesimal time interval from time  $t$  to time  $t + dt$ . (2.3)

**Reaction Types** The chemical reactions are considered to be distinct and instantaneous in nature. They are generally of two types: *unimolecular* and *bimolecular*.

Unimolecular reactions involves a single molecule as



while bimolecular reactions involve two molecules



Trimolecular and other higher order reactions are considered to be made up sequences of two or more reactions of unimolecular or bimolecular type. The propensity function for both unimolecular and bimolecular reactions may be derived from the laws of chemical kinetics.

- For a unimolecular reaction, if there are  $x_m$  molecules of a species  $S_m$  in the system, then the probability that some one of them will undergo the unimolecular reaction in the next infinitesimal time step  $dt$  is  $x_m c_j dt$ , where  $c_j$  is the reaction rate constant for this chemical reaction. Therefore, the propensity function for the unimolecular reaction is  $a_j(\mathbf{x}) = c_j x_m$ .
- For a bimolecular reaction in which molecules from two distinct species  $S_m$  and  $S_n$  react to produce something, the probability that one of the  $(x_m, x_n)$  pairs inside the volume  $\Omega$  will react in the next  $dt$  is  $x_m x_n c_j dt$ , where  $c_j$  is the reaction rate constant for this type of chemical reaction. Therefore, the propensity function in this case is

$a_j(\mathbf{x}) = c_j x_m x_n$ . In the particular case, when the two molecular species are the same, the number of distinct molecular pairs of  $S_m$  molecules is  $\frac{1}{2}x_m(x_m - 1)$  and hence the propensity function becomes  $a_j(\mathbf{x}) = c_j \frac{1}{2}x_m(x_m - 1)$ .

Now let us look into stochastic modeling of biochemical reactions on well-stirred systems.

### 2.2.1 Markov Processes

A stochastic process possessing the property that given the current state of the system, its future behaviour does not depend on the past states, is called a *Markov Process* named after the Russian mathematician Adre Markov (1856-1922). In other words, if the time evolution of a Markov process needs to be predicted, then only the information about the current state of the system is needed (and no information about the past states is required). When the system transits from one state to another in a Markov process, a "Markov Chain" of events is created. It must be pointed out that the dynamic behavior of many biochemical systems can be accurately modelled with the help of Markov processes. In particular, we are interested in the processes involving discrete state changes in continuous time (i.e., a continuous-time stochastic process involving random variables indexed by a continuous-time variable).

To formalize, let us consider a Markov process that can be in one of the states from the state space  $S = \{1, 2, \dots, r\}$ . Let the state of the system at any time  $t$  be  $X(t) = x \in S$ . At a later time,  $t + t'$ , the system instantaneously jumps to state  $X(t + t') = y \in S$ . The future time evolution of the system is characterized by the transition kernel

$$K(x, t, y, t') = \mathcal{P}(x, y)$$



where the *transition probability matrix*  $\mathcal{P}(.,.)$  is defined as

$$\mathcal{P} = \begin{pmatrix} \mathcal{P}(x_1, x_1) & \mathcal{P}(x_1, x_2) & \dots & \mathcal{P}(x_1, x_r) \\ \mathcal{P}(x_2, x_1) & \mathcal{P}(x_2, x_2) & \dots & \mathcal{P}(x_2, x_r) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{P}(x_r, x_1) & \mathcal{P}(x_r, x_2) & \dots & \mathcal{P}(x_r, x_r) \end{pmatrix} \quad (2.4)$$

such that  $\mathcal{P}(x, y)$  is the conditional probability

$$\mathcal{P}(x, y) = \mathcal{P}_r(X(t + t') = y | X(t) = x), \quad x, y \in S, \quad t \in [0, \infty)$$

Here  $\mathcal{P}$  is an  $r \times r$  matrix (a real valued function on  $S \times S$ ) satisfying

$$\mathcal{P}(x, y) \geq 0 \quad x, y \in S \quad (2.5)$$

and

$$\sum_{y \in S} \mathcal{P}(x, y) = 1 \quad x, y \in S \quad (2.6)$$

This means that the elements of this matrix are always nonnegative real numbers by equation (2.5) and that the sum of the elements in any row of this matrix is 1 by equation (2.6). The matrix  $\mathcal{P}(.,.)$  is called a *Markov* or *Stochastic* matrix.

If the transition kernel  $K$  does not explicitly depend on  $t$ , the Markov process is called *homogeneous* and the transition kernel can be expressed as

$$\mathcal{K}(x, y, t') = \mathcal{P}(t'). \quad (2.7)$$

For  $t' = 0$ , the transition probability matrix is an identity matrix,

$$\mathcal{P}(0) = I, \quad (2.8)$$

as there will be no transition from one state to another (different) state, with zero time interval.

### Chapman Kolmogorov Equation

According to the Chapman Kolmogorov Theorem [2], if  $X_n$ , where  $n = 0, 1, 2, \dots$  is a homogeneous Markov Chain, then

$$\mathcal{P}_r(X_{m+n} = j | X_0 = i) = \sum_{k \in S} \mathcal{P}_r(X_{m+n} = j | X_m = k) \mathcal{P}_r(X_m = k | X_0 = i) \quad (2.9)$$

In other words, the conditional probability that the Markov process transits from state  $i$  to state  $j$  in  $m + n$  steps is equal to the sum of the product of the conditional probabilities of reaching an intermediate state  $k$  from  $i$  in  $m$  steps and from state  $k$  to  $j$  in  $n$  steps. Equation (A.1) is also called the Chapman Kolmogorov equation. For completeness, the proof is given in Appendix A. Therefore, the transition kernel in equation (2.7) can be written as

$$K(i, j, t + t') = \sum_{k=1}^r K(i, k, t) K(k, j, t') \quad (2.10)$$

or, in terms of transition probability (stochastic) matrix, as

$$\mathcal{P}(t + t') = \mathcal{P}(t) \mathcal{P}(t') = \mathcal{P}(t') \mathcal{P}(t). \quad (2.11)$$

### Transition Rate Matrix (Propensity Matrix)

The *transition rate matrix*  $Q$  is defined as the derivative of  $\mathcal{P}(t')$  with respect to time at  $t' = 0$ , i.e.,

$$Q = \left. \frac{d\mathcal{P}(t')}{dt'} \right|_{t'=0} \quad (2.12)$$

$$\begin{aligned}
 &= \lim_{\delta t \rightarrow 0} \frac{\mathcal{P}(\delta t) - \mathcal{P}(0)}{\delta t} \\
 &= \lim_{\delta t \rightarrow 0} \frac{\mathcal{P}(\delta t) - I}{\delta t}
 \end{aligned} \tag{2.13}$$

Rearranging above, we obtain the infinitesimal transition matrix

$$\begin{aligned}
 Qdt &= \mathcal{P}(dt) - I \\
 \mathcal{P}(dt) &= Qdt + I
 \end{aligned} \tag{2.14}$$

As  $\mathcal{P}(dt)$  is a stochastic matrix, the rate matrix  $Q$  must satisfy:

1. the off-diagonal elements of  $\mathcal{P}(dt)$  and  $Qdt$  must be the same,
2. the rows of  $Q$  must sum to zero.

**Forward Kolmogorov Equation** The rate of change of the stochastic matrix with time can be written as (see also [2] for details)

$$\begin{aligned}
 \frac{d\mathcal{P}(t)}{dt} &= \lim_{\delta t \rightarrow 0} \frac{\mathcal{P}(t + \delta t) - \mathcal{P}(t)}{\delta t} \\
 &= \lim_{\delta t \rightarrow 0} \frac{\mathcal{P}(t)\mathcal{P}(\delta t) - \mathcal{P}(t)}{\delta t} \quad \text{using equation (2.11)} \\
 &= \mathcal{P}(t) \lim_{\delta t \rightarrow 0} \frac{\mathcal{P}(\delta t) - I}{\delta t} \\
 &= \mathcal{P}(t)Q \quad \text{using (2.13).}
 \end{aligned}$$

Thus,

$$\frac{d\mathcal{P}(t)}{dt} = \mathcal{P}(t)Q. \tag{2.15}$$

Or, in terms of the transition kernel,

$$\frac{dK(i, j, t)}{dt} = \sum_{k=1}^r K(i, k, t)q_{kj} \tag{2.16}$$

This is called *Kolmogorov's forward equation*.

**Backward Kolmogorov Equation** [2] We begin again with the rate of change of the stochastic matrix with respect to time as

$$\frac{d\mathcal{P}(t)}{dt} = \lim_{\delta t \rightarrow 0} \frac{\mathcal{P}(t + \delta t) - \mathcal{P}(t)}{\delta t}$$

Using equations (2.11) and (2.13), we derive

$$\begin{aligned} \mathcal{P}(t + \delta t) &= \mathcal{P}(\delta t)\mathcal{P}(t) \\ \frac{d\mathcal{P}(t)}{dt} &= \lim_{\delta t \rightarrow 0} \frac{\mathcal{P}(\delta t)\mathcal{P}(t) - \mathcal{P}(t)}{\delta t} \\ &= \lim_{\delta t \rightarrow 0} \frac{\mathcal{P}(\delta t) - I}{\delta t} \mathcal{P}(t) \\ &= Q\mathcal{P}(t) \quad \text{By using equation(2.13)} \end{aligned} \tag{2.17}$$

In terms of the transition kernel, we can write

$$\frac{dK(i, j, t)}{dt} = \sum_{k=1}^r q_{i,k} K(k, j, t) \tag{2.18}$$

This is called *Kolmogorov's backward equation*.

## 2.3 Chemical Master Equation (CME)

We now derive the Chemical Master Equation [36, 2] (CME). We assume that, initially, the system was in state  $\mathbf{X}(t_0) = \mathbf{x}_0$ . We define

$$P(\mathbf{x}, t | \mathbf{x}_0, t_0) = \text{the probability that the system state is } \mathbf{X}(t) = \mathbf{x}, \text{ given that } \mathbf{X}(t_0) = \mathbf{x}_0 \tag{2.19}$$

We are interested in finding out the time-evolution of this probability function, which is governed by the Chemical Master Equations (CME). To derive this equation, we begin with

the Kolmogorov's forward equation (see also Wilkinson, [2]),

$$\frac{d}{dt}P(\mathbf{x}, t|\mathbf{x}_0, t_0) = \sum_{\{\mathbf{x}' \in \mathcal{M}\}} q_{\mathbf{x}', \mathbf{x}} P(\mathbf{x}', t|\mathbf{x}_0, t_0) \quad (2.20)$$

where  $\mathcal{M}$  is a countable state space and  $q_{\mathbf{x}', \mathbf{x}}$  is an element of the state transition rate matrix  $\mathbf{Q}$ , and is defined by

$$q_{\mathbf{x}', \mathbf{x}} dt = \text{the probability that the transition from state } \mathbf{x}' \text{ to state } \mathbf{x} \text{ will take place} \\ \text{within the infinitesimal time interval } [t, t + dt) \quad (2.21)$$

The Kolmogorov's forward equation (2.20) can be written as

$$\begin{aligned} \frac{d}{dt}P(\mathbf{x}, t|\mathbf{x}_0, t_0) &= \sum_{\{\mathbf{x}' \in \mathcal{M}\}} q_{\mathbf{x}', \mathbf{x}} P(\mathbf{x}', t|\mathbf{x}_0, t_0) \\ &= \sum_{\{\mathbf{x}' \in \mathcal{M} | \mathbf{x}' \neq \mathbf{x}\}} q_{\mathbf{x}', \mathbf{x}} P(\mathbf{x}', t|\mathbf{x}_0, t_0) + q_{\mathbf{x}, \mathbf{x}} P(\mathbf{x}, t|\mathbf{x}_0, t_0) \end{aligned} \quad (2.22)$$

Using the property that the rows of  $\mathbf{Q}$  must each sum to zero [2], we have

$$\begin{aligned} \sum_{\{\mathbf{x}' \in \mathcal{M}\}} q_{\mathbf{x}, \mathbf{x}'} &= 0 \\ \implies \sum_{\{\mathbf{x}' \in \mathcal{M} | \mathbf{x}' \neq \mathbf{x}\}} q_{\mathbf{x}, \mathbf{x}'} + q_{\mathbf{x}, \mathbf{x}} &= 0 \\ \implies q_{\mathbf{x}, \mathbf{x}} &= - \sum_{\{\mathbf{x}' \in \mathcal{M} | \mathbf{x}' \neq \mathbf{x}\}} q_{\mathbf{x}, \mathbf{x}'} \end{aligned} \quad (2.23)$$

Substituting equation (2.23) into (2.22), we get

$$\begin{aligned} \frac{d}{dt}P(\mathbf{x}, t|\mathbf{x}_0, t_0) &= \sum_{\{\mathbf{x}' \in \mathcal{M}|\mathbf{x}' \neq \mathbf{x}\}} q_{\mathbf{x}', \mathbf{x}} P(\mathbf{x}', t|\mathbf{x}_0, t_0) - \sum_{\{\mathbf{x}' \in \mathcal{M}|\mathbf{x}' \neq \mathbf{x}\}} q_{\mathbf{x}, \mathbf{x}'} P(\mathbf{x}, t|\mathbf{x}_0, t_0) \\ &= \sum_{\{\mathbf{x}' \in \mathcal{M}|\mathbf{x}' \neq \mathbf{x}\}} [q_{\mathbf{x}', \mathbf{x}} P(\mathbf{x}', t|\mathbf{x}_0, t_0) - q_{\mathbf{x}, \mathbf{x}'} P(\mathbf{x}, t|\mathbf{x}_0, t_0)] \end{aligned} \quad (2.24)$$

It could be noted that the summation is over all transitions (reactions) from  $j = 1, \dots, M$ , and  $\mathbf{x}' = \mathbf{x} - \nu_j$ . Also that the transition rate matrix elements are propensities such that,

$$\begin{aligned} q_{\mathbf{x}, \mathbf{x}'} &= a_j(\mathbf{x}) \\ q_{\mathbf{x}', \mathbf{x}} &= a_j(\mathbf{x} - \nu_j) \end{aligned} \quad (2.25)$$

Therefore, equation (2.24) can be written as

$$\frac{d}{dt}P(\mathbf{x}, t|\mathbf{x}_0, t_0) = \sum_{\{\mathbf{x}' \in \mathcal{M}|\mathbf{x}' \neq \mathbf{x}\}} [a_j(\mathbf{x} - \nu_j) P(\mathbf{x} - \nu_j, t|\mathbf{x}_0, t_0) - a_j(\mathbf{x}) P(\mathbf{x}, t|\mathbf{x}_0, t_0)] \quad (2.26)$$

This is called the *Chemical Master Equation* (CME). It is a set of coupled ordinary differential equations (ODE) giving the time evolution of the system state, starting from some initial state at time  $t_0$ . Time is a continuous variable, while the state of the system changes discretely, in a stochastic manner. It must be noted here that the state vector is  $N$  dimensional and that this number could be large. Also, solving the CME (2.26) means solving a large set of ODEs, with one ODE for each of the possible states. Arkin et. al. [28] have reported that, in their model of  $\lambda$ -phage, a realistic population size lead to  $10^{70}$  possible states while solving the problem using the Chemical Master Equation.

### 2.3.1 Stochastic Simulation Algorithm (SSA)

As mentioned earlier, the Chemical Master Equation (CME) is a system of coupled ordinary differential equations. The size of this system is equal to the number of all possible states

of the system involving  $M$  chemical reactions. Due to this extremely high dimension, the CME cannot be handled via analytic or computational techniques for most realistic applications. Alternate approaches have been developed including Gillespie's algorithm [39, 40], tau-leaping methods [11] etc. to deal with this challenge.

In this section, the exact solution approach of the stochastic simulation algorithm (SSA) (also known as Gillespie's algorithm) is described. The theoretical justifications of the algorithm is presented. This is followed by the description of the algorithm itself along with the difficulties encountered when using this algorithm.

**Derivation of the Gillespie's Algorithm** The main idea behind this exact method is to simulate one solution trajectory at a time correctly rather than solving the CME for the probability distribution of all possible states simultaneously. Such trajectories can be simulated by following the "exact probability distribution", reflecting the corresponding probability distribution given by the CME.

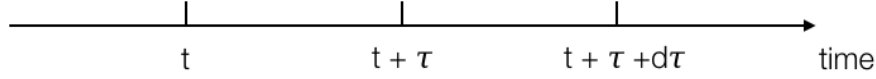
To simulate a trajectory of the system, the entire time span of the trajectory can be considered to be composed of two types of time intervals:

1. time intervals during which no biochemical reaction takes place,
2. time intervals during which one chemical reaction takes place.

We therefore consider two time intervals as depicted in Figure 2.1 . The first time interval starts from time  $t$  and ends at time  $[t + \tau)$ . The second interval starts from time  $[t + \tau)$  and ends at time  $[t + \tau + d\tau)$  where  $d\tau$  is an infinitesimal time step. Corresponding to these two intervals, the following two probabilities are defined:

$$\begin{aligned} \mathcal{P}_0(\tau|\mathbf{x}, t) = & \text{the probability that no reaction takes place during the time} \\ & \text{interval } [t, t + \tau), \text{ provided that } X(t) = \mathbf{x}, \end{aligned} \tag{2.27}$$

and

**Figure 2.1:** Time intervals on a sample trajectory path

$$p(\tau, j | \mathbf{x}, t) = \text{the probability that the next reaction takes place during the time interval } [t + \tau, t + \tau + d\tau) \text{ and that it is the } j^{\text{th}} \text{ reaction, provided that } X(t) = \mathbf{x}. \quad (2.28)$$

The above two events takes place independently, without affecting the occurrence of one another. This makes possible to replace an “and” with a “product” while working with the probabilities. Using the definition of the propensity function  $a_j(X(t))$ , we know that the probability of the  $j^{\text{th}}$  reaction taking place in a time step of size  $d\tau$  is  $a_j(X(t))d\tau$ . The construction of the propensities for various types of chemical reactions have already been discussed in Section 2.2. We can now write:

$$\begin{aligned} \mathcal{P}_0(\tau + d\tau | \mathbf{x}, t) &= \text{Probability that no reaction takes place over } [t, t + \tau + d\tau) \\ &= \text{Probability that no reaction occurs over } [t, t + \tau) \\ &\quad \text{and no reaction occurs over } [t + \tau, t + \tau + d\tau) \\ &= \text{Probability that no reaction occurs over } [t, t + \tau) \\ &\quad \times \text{Probability that no reaction occurs over } [t + \tau, t + \tau + d\tau) \\ &= \text{Probability that no reaction occurs over } [t, t + \tau) \\ &\quad \times (1 - \text{sum of probabilities that each reaction} \\ &\quad \text{takes place over } [t + \tau, t + \tau + d\tau)) \end{aligned}$$

$$\mathcal{P}_0(\tau + d\tau | \mathbf{x}, t) = \mathcal{P}_0(\tau | \mathbf{x}, t) \left( 1 - \sum_{k=1}^M a_k(\mathbf{x}) d\tau \right) \quad (2.29)$$

From equation (2.29),

$$\begin{aligned} a_{\text{sum}}(\mathbf{x}) &= \sum_{k=1}^M a_k(\mathbf{x}) \quad \text{then,} \\ \mathcal{P}_0(\tau + d\tau | \mathbf{x}, t) &= \mathcal{P}_0(\tau | \mathbf{x}, t) (1 - a_{\text{sum}}(\mathbf{x}) d\tau) \end{aligned} \quad (2.30)$$



Re-arranging, we have

$$\frac{\mathcal{P}_0(\tau + d\tau|\mathbf{x}, t) - \mathcal{P}_0(\tau|\mathbf{x}, t)}{d\tau} = -a_{sum}(\mathbf{x})\mathcal{P}_0(\tau|\mathbf{x}, t)$$

In the limiting case when  $d\tau \rightarrow 0$ , we get

$$\frac{d\mathcal{P}_0(\tau|\mathbf{x}, t)}{d\tau} = -a_{sum}(\mathbf{x})\mathcal{P}_0(\tau|\mathbf{x}, t)$$

This is an ODE with the initial condition  $\mathcal{P}_0(0|\mathbf{x}, t) = 1$ . Solving

$$\frac{d\mathcal{P}_0(\tau|\mathbf{x}, t)}{\mathcal{P}_0(\tau|\mathbf{x}, t)} = -a_{sum}(\mathbf{x})d\tau$$

$$\ln \mathcal{P}_0(\tau|\mathbf{x}, t) = -a_{sum}(\mathbf{x})\tau + C$$

$$\mathcal{P}_0(\tau|\mathbf{x}, t) = e^{-a_{sum}(\mathbf{x})\tau + C} = Ae^{-a_{sum}(\mathbf{x})\tau}$$

Using the initial condition  $\mathcal{P}_0(0|\mathbf{x}, t) = 1$  gives  $A = 1$ . Therefore

$$\mathcal{P}_0(\tau|\mathbf{x}, t) = e^{-a_{sum}(\mathbf{x})\tau} \tag{2.31}$$

Now from equation (2.28)

$$\begin{aligned} p(\tau, j|\mathbf{x}, t)d\tau &= \text{the probability that no reaction takes place over } [t, t + \tau) \\ &\quad \text{and that the } j^{th} \text{ reaction takes place over } [t + \tau, t + \tau + d\tau) \\ &= \text{the probability that no reaction takes place over } [t, t + \tau) \\ &\quad \times \text{the probability that the } j^{th} \text{ reaction takes place over } [t + \tau, t + \tau + d\tau) \\ &= \mathcal{P}_0(\tau|\mathbf{x}, t) \times a_j(\mathbf{x})d\tau. \end{aligned} \tag{2.32}$$

From equation (2.31)

$$\begin{aligned} p(\tau, j | \mathbf{x}, t) &= e^{-a_{sum}(\mathbf{x})\tau} \times a_j(\mathbf{x}) \\ \implies p(\tau, j | \mathbf{x}, t) &= a_j(\mathbf{x}) e^{-a_{sum}(\mathbf{x})\tau} \end{aligned}$$

We can write this as

$$p(\tau, j | \mathbf{x}, t) = (a_{sum}(\mathbf{x}) e^{-a_{sum}(\mathbf{x})\tau}) \times \left( \frac{a_j(\mathbf{x})}{a_{sum}(\mathbf{x})} \right) \quad (2.33)$$

From equation (2.33), it could be inferred that the joint probability density function  $p(\tau, j | \mathbf{x}, t)$  is constituted of two density functions. The first density function gives the time  $\tau$  to the next reaction with probability density  $a_{sum}(\mathbf{x}) e^{-a_{sum}(\mathbf{x})\tau}$ . The second gives the index  $j$  of the next reaction having probability density  $\frac{a_j(\mathbf{x})}{a_{sum}(\mathbf{x})}$ .

The three propositions (given in Appendix A) results are used in obtaining equations (2.32) and (2.33) to further derive the Gillespie's algorithm. Note that (2.33) is the main equation for deriving the stochastic simulation algorithm (SSA). The equation (A.7) shows that  $\frac{a_j}{a_{sum}}$  corresponds to a discrete random variable, that means picking one reaction with index  $j$  corresponding to the reaction which happens first (see also Appendix A, Proposition 2) to  $a_j(X)$ . The term  $a_{sum}(X) e^{(-a_{sum}(X)\tau)}$  is the density function for a continuous random variable with an exponential distribution. In general, an exponentially distributed random variable with parameters  $a > 0$  is characterized by the density function

$$\begin{cases} ae^{-ax} & \text{for } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Hence if  $X$  is exponentially distributed with parameter  $a$ , then for any  $0 < c < d$

$$\begin{aligned} P(c < X < d) &= \int_c^d ae^{-ax} dx \\ &= e^{-ac} - e^{-ad} \end{aligned}$$

Now

$$\begin{aligned}
 P(c < X < d) &= P(-d < -X < -c) \\
 &= P(e^{-ad} < e^{-aX} < e^{-ac}) \\
 &= e^{-ac} - e^{-ad}
 \end{aligned} \tag{2.34}$$

So the random variable

$$Z := e^{-aX}$$

has a uniform  $(0, 1)$  distribution. This means that the probability of  $Z$  lying in any subinterval of  $(0, 1)$  is given by the length of the subinterval. Hence  $X$  may be written as

$$Z = e^{-aX}$$

So

$$X = \frac{1}{a} \ln\left(\frac{1}{Z}\right) \tag{2.35}$$

where  $Z$  is uniform  $(0, 1)$  with  $a = a_{sum}(X(t))$ .

The above results and the propositions in Appendix A lead us to the following Monte Carlo method called Gillespie's Algorithm [39, 40].

**Gillespie's Algorithm (SSA)** We can summarize the stochastic simulation algorithm in the following steps.

1. Initialize the state vector  $X = x_0$  at initial time  $t_0$ .
2. Calculate the propensity functions and their sum,  $a_k(X)$  for  $1 \leq k \leq M$  and  $a_0(X) = \sum_{k=1}^M a_k(X)$
3. Draw two independent samples  $r_1$  and  $r_2$  from  $U(0, 1)$ .

4. Then calculate the time to the next reaction with the following formula

$$\tau = \frac{1}{a_{sum}(X(t))} \ln \frac{1}{r_1}$$

5. Calculate the index of the next reaction  $\sum_{k=1}^{j-1} a_k(X) < r_2 a_0(X) \leq \sum_{k=1}^j a_k(X)$
6. Update system state  $X \leftarrow X + \nu_j$  and current time from  $t \leftarrow t + \tau$  and go to step 2 or STOP according to the termination criteria.

### Issues Encountered while Executing the SSA

Gillespie's algorithm is an exact method to solve the CME. However, the main issue while executing the SSA is the execution time taken to run a sufficiently large number of trajectories to produce a reasonable statistics. Typically, tens of thousands of trajectories need to be simulated to produce accurate estimates of the probability distributions.

At each iteration of the algorithm, we need to calculate the time taken for the next reaction and the reaction index. This involves updating of the propensity functions and the state vector. The time  $\tau$  taken until the next reaction depends inversely on the sum of propensity functions i.e.  $a_{sum}(\mathbf{x})$ . If this sum is large, the time  $\tau$  becomes small. For a system with a large number of molecules and having some reactions taking place at large rate (fast reactions), the propensity function sum ( $a_{sum}(\mathbf{x})$ ) becomes very large resulting in a very small time step  $\tau$ . The simulation execution time becomes large due to very small time steps  $\tau$ . Also a large number of random numbers needs to be generated. For systems of practical interest, due to the enormous number of reaction events encountered when very fast reactions are present, the detailed construction of every reaction event becomes a very time consuming task.

### 2.3.2 Tau-Leaping Method

In this section, a brief introduction to the tau-leaping method [11] is given. A detailed description of the leaping methods is presented in Chapter 3. One important topic for tau-leaping approximate methods is the selection of the time step  $\tau$ . The time step  $\tau$  is chosen such that the leap condition is satisfied. The leap condition ensures that each of the propensity functions changes only by a small amount over the time step  $\tau$ .

If  $\tau$  satisfies the leap condition, then the number of reactions  $R_j$  taking place between  $[t, t + \tau)$  may be approximated by a Poisson random variable  $\mathcal{P}_j(a_j(X(t))\tau)$  with parameter  $a_j(X(t))\tau$ . Then the system state may be updated as

$$X(t + \tau) = X(t) + \sum_{j=1}^M \nu_j \mathcal{P}_j(a_j(X(t))\tau) \quad (2.36)$$

This represents the tau-leaping method.

#### Problems Encountered While Using Tau-Leaping Simulations

A number of tau-leaping algorithms have been proposed and used to simulate biochemical systems approximately [11, 15, 21, 63]. However, there are several problems encountered in such simulations. Two major issues are:

1. The choice of the time step  $\tau$  — Determination of the time interval  $[t, t + \tau)$  in advance such that the leap condition is satisfied.
2. The tau-leaping methods should be applied if  $\tau$  can be chosen large enough such that a large number of reactions happen between  $[t, t + \tau)$ , so that the leaping method is much more efficient than the SSA.

## 2.4 Chemical Langevin Equation

Paul Langevin (1872-1946), a French physicist, developed the Langevin equation to describe the Brownian motion of a particle in a fluid. The form of Langevin equation used in modeling of chemically reacting systems showing stochastic behavior is called the Chemical Langevin equation (CLE)[41]. This section begins with an introduction to stochastic differential equations. This is followed by the derivation of the Chemical Langevin equations for a well stirred biochemically reacting system. A comparison of the CME and the CLE models concludes this section.

### 2.4.1 Stochastic Differential Equations

**Brownian Motion / Wiener Process** The Brownian motion also called the Wiener process is a continuous-time stochastic process  $W(t)$  ( $0 \leq t \leq T$ ) satisfying:

1. At time  $t = 0$ ,  $W(0) = 0$  with probability 1.
2. For any  $0 \leq s < t < u < v \leq T$  the increments  $W(t) - W(s)$  and  $W(v) - W(u)$  are independent random variables.
3. For any  $0 \leq s \leq t \leq T$  the increment  $W(t) - W(s)$  has a normal distribution with mean zero and variance  $(t - s)$ . In other words,

$$W(t) - W(s) \sim \sqrt{t - s}N(0, 1)$$

where  $N(0, 1)$  represents a normally distributed random variable with zero mean and unit variance.

For numerical simulation purposes, the continuous time variable is discretized over the whole time interval  $[0, T]$  of the simulation such that  $t_j = j\delta t$  where  $j = 1, 2, \dots, N$  and  $\delta t = T/L$

for some positive interger  $L$ . Corresponding to each  $t_j$ , the Wiener process is generated such that  $W_j = W(t_j)$  with  $W_0 = W(t_0) = 0$ . Also from conditions (2) and (3) above

$$W_j = W_{j-1} + \sqrt{\delta t} N(0, 1) \quad j = 1, 2, \dots, N$$

where  $N(0, 1)$  is drawn from a normal distribution with zero mean and unit variance.

**Stochastic Calculus and Stochastic Integrals** For dynamical systems without noise, the time evolution of the system is often modeled using ordinary differential equations. In case of stochastic systems, the problem often involves stochastic differential equations. The Brownian motion (or the Wiener process) is not differentiable in the traditional sense. Therefore, stochastic calculus is used for this purpose. There are two main approaches used in this calculus called Itô stochastic calculus and Stratonovich stochastic calculus. The approaches are equivalent and can be converted into one another [10].

**Itô's Approach** Named after the Japanese mathematician Kiyosi Itô (1915-2006), in this approach, Itô integrals and the change of variable formula (chain rule for stochastic variables) called Itô lemma are used. An Itô process is defined as an adapted stochastic process that can be expressed as

$$X(t) = X(0) + \int_0^t f(X(t))dt + \int_0^t g(X(s))dW(s) \quad (2.37)$$

where  $f$  and  $g$  are scalar functions,  $X(t)$  is a random variable with the initial condition  $X(0)$  and  $W(s)$  is a Wiener process. According to Itô lemma (stochastic chain rule), if  $f$  is twice differentiable and  $X$  is an Itô process, then

$$df(X(t)) = f'(X(t))dX(t) + \frac{1}{2}f''(X(t))g^2(t)dt \quad (2.38)$$

In ordinary calculus, the integral  $\int_0^T h(t)dt$  for a suitable scalar function  $h(t)$  is approximated

by the Reimann sum

$$\sum_{j=0}^{N-1} h(t_j)(t_{j+1} - t_j)$$

where  $t_j = j\delta t$  represents the discretized grid for the time interval  $[0, T]$  for  $j = 0, 1, 2, \dots, N$ .

We can write

$$\int_0^T h(t)dt = \lim_{\delta t \rightarrow 0} \sum_{j=0}^{N-1} h(t_j)(t_{j+1} - t_j)$$

For a Weiner process  $W(t)$ , the stochastic integral  $\int_0^T h(t)dW(t)$  can be approximated as the sum

$$\int_0^T h(t)dW(t) \sim \sum_{j=0}^{N-1} h(t_j)(W(t_{j+1}) - W(t_j)) \quad (2.39)$$

This is called the Itô integral formula.

**Stratonovich Approach** This approach is named after the Russian mathematician Stratonovich (1930-1997). This approach is usually used in Physics, where physical laws are being considered. Unlike the Itô approach, in this approach Stratonovich integrals are defined such that the "chain rule" (change of variables in differentiation) of the ordinary calculus holds. In the Stratonovich approach, the Reimann sum approximation

$$\int_0^T h(t)dt = \lim_{\delta t \rightarrow 0} \sum_{j=0}^{N-1} h\left(\frac{t_j + t_{j+1}}{2}\right)(t_{j+1} - t_j)$$

for calculating integrals is extended to determine the stochastic integral

$$\int_0^T h(t)dW(t) \sim \sum_{j=0}^{N-1} h\left(\frac{t_j + t_{j+1}}{2}\right)(W(t_{j+1}) - W(t_j)) \quad (2.40)$$

**Stochastic Differential Equations** An ordinary differential equation (ODE) for a continuous function  $X(t)$  is usually written as

$$\frac{dX(t)}{dt} = f(x(t))$$



with the initial condition  $x(t_0) = x_0$ , for a well defined function  $f(x)$ . It has the solution

$$X(t) = x_0 + \int_0^t f(X(s))ds$$

A stochastic differential equation (SDE) for a random variable  $X(t)$  can be written as

$$dX(t) = f(X)dt + g(X)dW(t) \quad (2.41)$$

with the initial condition  $X(t_0) = X_0$ . Here  $f$  and  $g$  are scalar functions while  $X(t)$  and  $W(t)$  are stochastic processes. It can be solved using the integral equation

$$X(t) = X_0 + \int_0^t f(X(s))ds + \int_0^t g(X(s))dW(s) \quad (2.42)$$

### 2.4.2 Derivation of the Chemical Langevin Equation

With the background presented in Section 2.4.1, we (can now) derive the Chemical Langevin Equation for biochemical kinetics problems. In this thesis, Gillespie's approach [41] has been followed. An approach based on finding first and second moments of the CME using parametric equation has also been adopted [16] to derive the CLE. The details of this approach are given in the articles [10, 6].

**Gillespie's Approach** The tau-leaping methods introduced in Section 2.3.2 plays a very important role of connecting the discrete, stochastic CME/SSA formulation with the continuous deterministic RRE/ODE (reaction rate equations) formulation via continuous stochastic formulation of the Chemical Langevin equation. In the tau-leaping method, a time step of size  $\tau > 0$  is chosen small enough such that the leap condition is satisfied, that is the propensity functions  $a_j(X(t))$  remaining almost constant during the interval  $[t, t + \tau)$ . The probability of occurrence of the reaction  $R_j$  during this interval is given by a Poisson distribution with parameter  $a_j(X(t))\tau$ . Then, the state of the system may be approximated

by

$$X(t + \tau) = X(t) + \sum_{j=1}^M \nu_j \mathcal{P}_j(a_j(X)\tau)$$

This is the explicit form of the tau-leaping method. Here  $X(t)$  is a discrete stochastic variable. If the time step  $\tau$  may be chosen large enough, such that  $a_j(X)\tau \gg 1$ , then the Poisson distribution may be approximated by a normal distribution with the mean and variance both equal to  $a_j(X)\tau$ . i.e.,

$$\mathcal{P}_j(a_j(X)\tau) \longrightarrow N(a_j(X)\tau, a_j(X)\tau),$$

or

$$\mathcal{P}_j(a_j(X)\tau) \approx a_j(X)\tau + \sqrt{a_j(X)\tau} Z_j \quad (2.43)$$

where  $Z_j$  represents independent normal distributions with mean zero and variance 1. It should be noted that the Poisson distribution is discrete random variable. The Gaussian distribution is continuous and with real values that could be both positive and negative. Under the assumption  $a_j(X)\tau \gg 1$  for all  $1 \leq j \leq M$ , and denoting the normal distribution with mean  $\mu$  and variance  $\sigma^2$  by  $N(\mu, \sigma^2)$ , we derive

$$\mathcal{P}_j(a_j(X)\tau) \approx N_j(a_j(X)\tau, a_j(X)\tau)$$

The discrete stochastic variable  $X(t)$  representing the number of molecules becomes a continuous stochastic variable. It follows from the explicit tau-leaping formula

$$\begin{aligned}
X(t + \tau) &= X(t) + \sum_{j=1}^M N_j(a_j(X)\tau, a_j(X)\tau)\nu_j \\
&= X(t) + \sum_{j=1}^M \left[ a_j(X)\tau + \sqrt{a_j(X)\tau} N_j(0, 1) \right] \nu_j \\
&= X(t) + \sum_{j=1}^M \nu_j a_j(X)\tau + \sum_{j=1}^M \nu_j \sqrt{a_j(X)} N_j(0, 1) \sqrt{\tau}
\end{aligned} \tag{2.44}$$

Taking  $\tau \rightarrow dt$  in equation (2.44), we obtain

$$dX(t) = \sum_{j=1}^M \nu_j a_j(X) dt + \sum_{j=1}^M \nu_j \sqrt{a_j(X)} dW_j(t) \tag{2.45}$$

where  $W_j$  are independent Weiner processes. This is a stochastic differential equation called the Chemical Langevin Equation (CLE). This model is valid when the molecular counts of each species is sufficiently large. Note that the CLE is an  $N$ -dimensional stochastic differential equation which is a reduction of the CME in the regime of large molecular numbers.

### 2.4.3 Euler-Maruyama Method

Given a stochastic differential equation (SDE) in the form (2.41)

$$dX(t) = f(X)dt + g(X)dW$$

with  $X(0) = X_0$  and  $0 \leq t < T$ . Its solution can be written in the integral form (2.42)

$$X(t) = X_0 + \int_0^t f(X(s))ds + \int_0^t g(X(s))dW(s)$$

To solve this equation numerically, we discretize the time interval  $[0, T]$  using  $\tau = T/L$  for some positive integer  $L$ . Let us denote  $X(j\tau) = X_j$  for  $j = 0, 1, 2, \dots, L-1$ . Then we can

approximate

$$X_{j+1} = X_j + f(X_j)\delta t + g(X_j)(W((j+1)\tau) - W(j\tau)) \quad j = 1, 2, \dots, L \quad (2.46)$$

This is the *Euler-Maruyama* (EM) method for numerically solving this SDE. It could be noted that for  $g \equiv 0$ , the above method becomes Euler's method for solving ODEs. In case of the CLE (2.45), the Euler-Maruyama method gives

$$X(t + \tau) = X(t) + \tau \sum_{j=1}^M \nu_j a_j(X(t)) + \sqrt{\tau} \sum_{j=1}^M \nu_j \sqrt{a_j(X(t))} N_j(0, 1) \quad (2.47)$$

The continuous stochastic variable  $X(t)$  is estimated at the discrete instants of time  $\{0, \tau, 2\tau, \dots, (L-1)\tau\}$  giving the sequence of random numbers  $\{X_0, X_1, \dots, X_{L-1}\}$ . The computational algorithm for numerically solving the CLE using the Euler-Maruyama method follows these steps.

Step 0 Initialize the number of species  $M$ , number of molecules for each species  $X_j(0)$ , stoichiometric coefficients  $\nu_j$  and propensities  $a_j(X(0))$  for  $j = 1, 2, \dots, M$

Step 1 Draw  $M$  independent samples from the zero mean, unit variance normal distribution  $Z_j \sim N_j(0, 1)$ , for  $j = 1, 2, \dots, M$

Step 2 Calculate

$$X(t + \tau) = X(t) + \tau \sum_{j=1}^M \nu_j a_j(X(t)) + \sqrt{\tau} \sum_{j=1}^M \nu_j \sqrt{a_j(X(t))} Z_j$$

Step 3 Update time  $t \leftarrow t + \tau$

Step 4 Check for the stopping criteria  $t \geq T$ . If false return to Step 1, otherwise stop.

A comparison of the models based on the Chemical Master Equation (CME) and the reduced Chemical Langevin Equation (CLE) model is given in Table 2.2.

CME	CLE
A microscopic, discrete, stochastic model	A macroscopic, continuous, stochastic model
Model is represented as a set of linear ODEs	Model is represented as a set of nonlinear SDEs.
One ODE for each state of the system	One SDE for each chemical species of the system
State space and hence the number of ODEs is very large	Reduced dimensions of the system (equal to the number of chemical species)
Number of molecules is strictly represented as nonnegative integer	Real valued random variables represent the number of molecules
Probability distributions over a large number of discrete states are used	Continuous probability distribution for each chemical species is used
Completely stochastic in nature, the system jumps from one state to another randomly	Contains the stochastic as well as deterministic components as parts of the SDEs

**Table 2.2:** Comparison of the CME and the CLE Models

There is an intermediate regime between the discrete stochastic (CME) and continuous deterministic (RRE) regimes where the intrinsic noise is still important but the number of molecules is large enough to describe the biochemical reaction kinetics by continuous models. This regime is treated with the CLE. Since the number of molecules is large (few hundreds or more), the size of the state space becomes too large to be handled by most computational algorithms based on CME. Methods based on using CLE are employed in this regime [16, 41].

## 2.5 Reaction Rate Equation (RRE)

Chemical reactions performed as experiments at the laboratory scale (macroscopic scale) are typically modelled using a set of coupled ordinary differential equations called the reaction rate equations (RRE). The amount of a chemical species is measured in terms of the concentration (moles per litre and written as  $M$ ). The number of molecules in one mole of a

chemical species is given by the Avogadro's number  $\eta_A = 6.023 \times 10^{23}$  molecules per mole. If  $\Omega$  is the volume, then the concentration  $y_i(t)$  of a chemical species in moles per litre can be converted into  $X_i(t)$  molecules as

$$X_i(t) = y_i(t) \times \eta_A \times \Omega \quad (2.48)$$

To determine the RREs, the law of mass action and rate law are used. The reaction rate is given in terms of the change in the concentration of a reactant divided by the stoichiometric coefficient. For example, in a chemical reaction,



The reaction rate is given by

$$r = -\frac{1}{n} \frac{\Delta A}{\Delta t} = -\frac{1}{m} \frac{\Delta B}{\Delta t} \quad (2.50)$$

The instantaneous reaction rate is given by

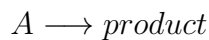
$$r = -\frac{1}{n} \frac{dA}{dt} = -\frac{1}{m} \frac{dB}{dt} \quad (\text{moles per litre per second}) \quad (2.51)$$

The negative sign is used as in the reaction, the reactants (A or B above) are being consumed (i.e.  $\frac{dA}{dt}$  or  $\frac{dB}{dt}$  are negative), making the overall reaction rate positive. According to the rate law, for simple reactions, the reaction rate is proportional to the concentrations of the reactants raised to some power. For example, in case of reaction (2.49), the rate law gives

$$r = kA^n B^m \quad (2.52)$$

It must be emphasized here that the above form of the rate law holds for simple reactions. For complex (multi-step) reactions the overall form of the rate law is different from (2.52),

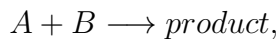
and the powers could be different from stoichiometric coefficients  $n$  and  $m$ . In practical situations, a complex reaction is decomposed into simple reactions, or the rate law is determined experimentally. In case of equation (2.52), the order of the reaction is  $n + m$  (sum of the stoichiometric coefficients) and  $k$  is the constant of proportionality called the rate constant. The units of the rate constants  $k$  depend upon the order of the reaction. For a first order reaction



the units of the rate constant can be determined from the rate law

$$\begin{aligned} -\frac{dA}{dt} &= kA \\ \implies \frac{M}{\text{sec}} &= [k] \times M \implies [k] = \text{sec}^{-1} \end{aligned} \quad (2.53)$$

For a second order reaction,



the rate law gives

$$\begin{aligned} -\frac{dA}{dt} &= kAB \\ \implies \frac{M}{\text{sec}} &= [k] \times M \times M \implies [k] = M^{-1}\text{sec}^{-1} \end{aligned} \quad (2.54)$$

i.e., the units of  $k$  are in moles per second.

**CLE and the thermodynamic limit** Assume now that the system volume  $\Omega$  and the species population size  $X_i(t)$  increase to infinity (keeping the temperature constant) such that  $\frac{X_i(t)}{\Omega}$  remains constant. This is called *thermodynamic limit*. The two terms on the right hand side of the CLE (2.45) behave differently in the thermodynamic limit. The first term (deterministic) grows with the size of the system, while the second term (stochastic) increases as the square root of the system size. Overall, the deterministic term dominates

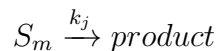
and the stochastic term can be neglected, to give

$$\frac{dy(t)}{dt} = \sum_{j=1}^M \nu_j a_j(y(t)) \quad (2.55)$$

This is the reaction rate equation (RRE), where  $y(t)$  represents the continuous, real valued, deterministic variable. It must be noted that in equation (2.55), the propensity functions  $a_j(y(t))$  uses rate constants  $c_j$ 's, while the reaction rate equation written at macroscopic scale, the rate constants  $k_j$ 's are used.

**Relationship between the Macroscopic and Microscopic Rate Constants** To determine the relation between the macroscopic rate constants  $k_j$ 's and the microscopic rate constants  $c_j$ 's, we use the study of the following reaction types.

### First Order Reactions



According to the rate law, the deterministic rate is

$$-\frac{dy_m}{dt} = k_j y_m \quad (2.56)$$

In terms of the number of molecules using equation (2.48)

$$\begin{aligned} -\frac{1}{\eta_A \Omega} \frac{dX_m}{dt} &= k_j \frac{X_m}{\eta_A \Omega} \\ \implies \frac{dX_m}{dt} &= -k_j X_m \end{aligned} \quad (2.57)$$

Using the RRE (2.55)

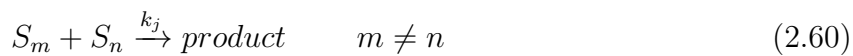
$$\begin{aligned} \frac{dX_m}{dt} &= \nu a_j(X_m) \\ &= (-1)c_j X_m \end{aligned} \quad (2.58)$$



Equating the right hand side of equations (2.57) and (2.58), we get

$$c_j = k_j \quad (2.59)$$

## Second Order Reactions



In this case, the rate law gives

$$-\frac{dy_m}{dt} = k_j y_m y_n \quad (2.61)$$

Converting the concentrations into the number of molecules

$$\begin{aligned} -\frac{1}{\eta_A \Omega} \frac{dX_m}{dt} &= k_j \frac{X_m}{\eta_A \Omega} \frac{X_n}{\eta_A \Omega} \\ \implies \frac{dX_m}{dt} &= -\frac{k_j}{\eta_A \Omega} X_m X_n \end{aligned} \quad (2.62)$$

Using the RRE (2.55), we get

$$\begin{aligned} \frac{dX_m}{dt} &= \nu a_j X_m \\ &= (-1) c_j X_m X_n \end{aligned} \quad (2.63)$$

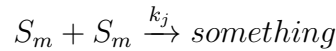
and thus from (2.62) and (2.63), we obtain

$$c_j = \frac{k_j}{\eta_A \Omega} \quad (2.64)$$

**Table 2.3:** RRE and Rate constants for different types of reactions

Reaction Order	Scale	Reaction	Propensity	$c_j$ vs $k_j$
First Order	Microscopic /molecular	$S_m \xrightarrow{c_j} product$	$k_j y_m$	$c_j = k_j$
	Macroscopic /molar	$S_m \xrightarrow{k_j} product$	$c_j X_m$	
Second Order	Microscopic /molecular	$S_m + S_n \xrightarrow{c_j} product$	$k_j y_m y_n$	$c_j = \frac{k_j}{\eta_A \Omega}$
	Macroscopic /molar	$S_m + S_n \xrightarrow{k_j} product$	$c_j X_m X_n$	
Dimerization	Microscopic /molecular	$S_m + S_m \xrightarrow{c_j} product$	$2k_j y_m^2$	$c_j = \frac{2k_j}{\eta_A \Omega}$
	Macroscopic /molar	$S_m + S_m \xrightarrow{k_j} product$	$2c_j X_m (X_m - 1)$	

### Dimerization Reactions



The deterministic rate law gives

$$-\frac{1}{2} \frac{dy_m}{dt} = k_j y_m^2$$

In terms of the number of molecules, we can write

$$-\frac{1}{2} \frac{1}{\eta_A \Omega} \frac{dX_m}{dt} = k_j \frac{X_m^2}{\eta_A^2 \Omega^2}$$

$$\implies \frac{dX_m}{dt} = -\frac{2k_j}{\eta_A\Omega} X_m^2 \quad (2.65)$$

From the RRE (2.55)

$$\frac{dX_m}{dt} = \nu c_j \frac{X_m(X_m - 1)}{2}$$

$$\frac{dX_m}{dt} = (-2)c_j \frac{X_m(X_m - 1)}{2}$$

$$\frac{dX_m}{dt} = c_j X_m(X_m - 1) \sim c_j X_m^2 \quad (2.66)$$

and therefore from equations (2.65) and (2.67), we have

$$c_j \approx \frac{2k_j}{\eta_A\Omega} \quad (2.67)$$

and for zeroth order function the propensity function is given as

$$a_j(X, c_j) = c_j$$

Table 2.3 summarizes the types of reactions along with the propensities in concentrations and molecular forms and the relationship between the rate constants used. So the RRE, under simplified assumptions, is a reduction of the CME model, when there are very large numbers of molecules of each species present in the system. However, in general, the solution of the RRE may not be the average solution of the CLE or the CME. Indeed, for non-independent random variables  $X, Y$ , in general,

$$E[XY] \neq E[X]E[Y]$$

Let us now consider the time derivative of the expected value  $E(X(t))$ . From taking the

expectation in the Chemical Master equation 2.26, we derive

$$\begin{aligned}
\frac{\partial}{\partial t} E(X(t)) &= \frac{\partial}{\partial t} \sum_{X \in \mathcal{M}} X p(X, t) \\
&= \sum_{X \in \mathcal{M}} X \frac{\partial}{\partial t} p(X, t) \\
&= \sum_{X \in \mathcal{M}} X \sum_{j=1}^M [a_j(X - \nu_j, c_j) p(X - \nu_j, t) - a_j(X, c_j) p(X, t)] \\
&= \sum_{j=1}^M \left[ \sum_{X \in \mathcal{M}} X a_j(X - \nu_j, c_j) p(X - \nu_j, t) - \sum_{X \in \mathcal{M}} X a_j(X, c_j) p(X, t) \right] \\
&= \sum_{j=1}^M \left[ \sum_{X \in \mathcal{M}} (X + \nu_j) a_j(X, c_j) p(X, t) - \sum_{X \in \mathcal{M}} X a_j(X, c_j) p(X, t) \right] \tag{2.68} \\
&= \sum_{j=1}^M [E((X(t) + \nu_j) a_j(X(t), c_j)) - E(X(t) a_j(X(t), c_j))] \\
&= \sum_{j=1}^M E(\nu_j a_j(X(t), c_j)) \\
&= \sum_{j=1}^M \nu_j E(a_j(X(t), c_j))
\end{aligned}$$

If all reactions are of zero or first order, then we can use the linearity of expectation to get

$$E(a_j(X(t), c_j)) = a_j(E(X(t), c_j))$$

By substituting it into (2.68), we obtain

$$\frac{\partial}{\partial t} E(X(t)) = \sum_{i=1}^M \nu_i a_i(E(X(t), c_i)) \tag{2.69}$$

Now denoting the average state vector by  $y(t) = E(X(t))$  we get,

$$\frac{d}{dt}y(t) = \sum_{i=1}^M \nu_i a_i(y(t), c_i)$$

or

$$\frac{d}{dt}y(t) = \nu a(y(t), c) \tag{2.70}$$

which is the reaction rate equation. We remind the reader that in the above equation,  $\nu$  is the stoichiometric matrix,  $a$  is the vector of propensity functions and  $c$  is the stochastic rate constant vector.

So when all reactions are zero and first order, the deterministic solution will correctly describe the expected value of the solution of CME. However, it will not give any insight into variability. Moreover, the average solution of the CME may be different from the RRE solution if reactions of order 2 or higher are present in the system.

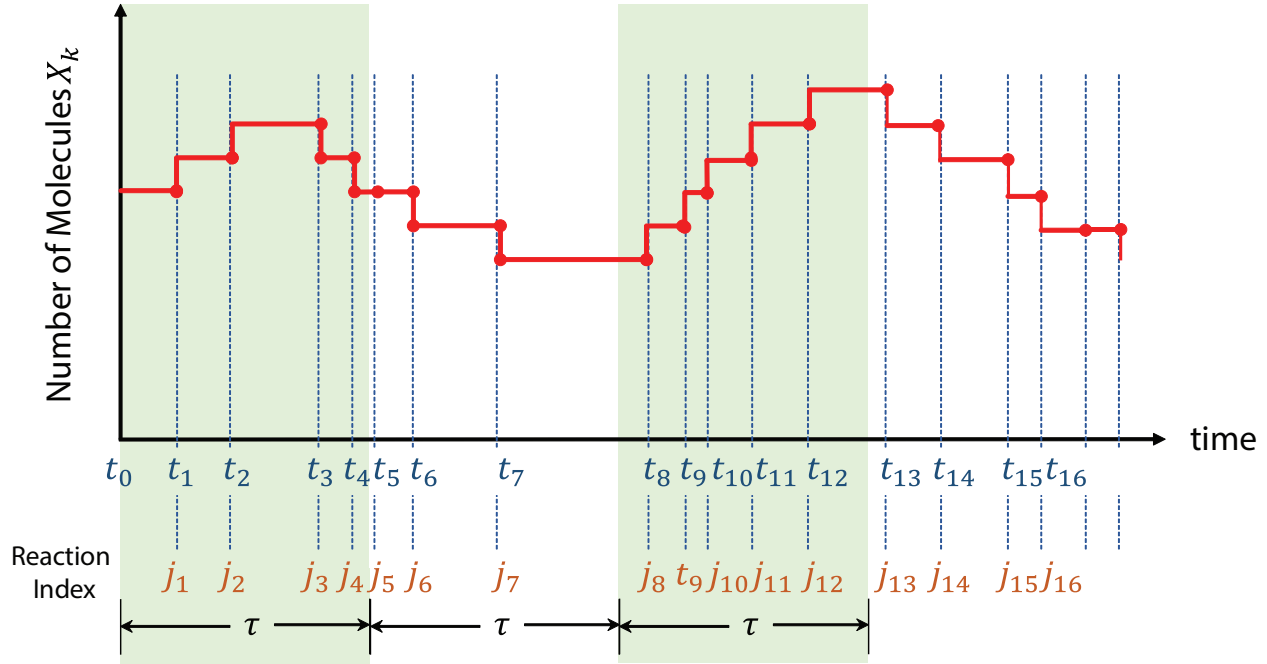
# Chapter 3

## Tau-Leaping Methods

The exact Monte Carlo methods for solving the Chemical Master Equation, like the SSA, simulate every reaction in the system to form a solution trajectory. Often, these exact methods are computationally very expensive for realistic biochemical processes involving fast reactions, as pointed out in Section 2.3.1. To overcome these high computational costs, approximate solution methods were proposed such as the tau-leaping methods [11, 15, 21]. Tau-leaping methods with their various aspects are described in this Chapter. The details of the three important tau-leaping approaches, namely the explicit [11], implicit [15] and trapezoidal [21] tau-leaping strategies, are given. The algorithms of these methods are presented here while the numerical results are given in chapter 4.

### 3.1 What is Tau-Leaping?

Consider a well-stirred biochemically reacting system as described in Section 2.1. A schematic diagram of the time evolution of the  $S_k$  molecular population,  $X_k(t)$ , is illustrated in Figure 3.1. Starting with the initial population  $X_k(t_0)$ , the number of molecules changes at successive instants  $t_1, t_2, t_3, \dots$ . At each of these time instants, a biochemical reaction involving species  $k$  takes place that changes the number of molecules,  $X_k(t)$ . Let us label these chem-



**Figure 3.1:** A schematic diagram of the SSA trajectory for  $X_k(t)$

ical reactions  $j_1, j_2, j_3, \dots$  respectively. Each of these labels can take up a value from 1 to  $M$ , corresponding to the chemical reactions  $R_1, R_2, R_3, \dots, R_M$ . A "history axis" consisting of the pairs  $(t, j)$  completely describe a realization of time evolution of  $X(t)$ . If we only monitor the history axis, the complete trajectory of  $X_k(t)$  can be constructed starting from the initial value  $X_k(t_0)$ .

In the (exact) SSA procedure, the stepping in time occurs randomly, in accordance with the accuracy of each reaction. Refer to Figure 3.1 and consider subdividing the time-axis in fixed time intervals of length  $\tau$ . It can be noticed that a number of reactions of different types affecting  $X_k(t)$  take place during an interval of length  $\tau$ . If the reactions are fast, the number of reaction events per interval of step  $\tau$  will be large. For slow reactions, the number of reactions per  $\tau$  interval will be smaller. Knowing the exact SSA trajectory and the history axis, we can exactly determine the population size at the end of each  $\tau$  interval. In simulations, we are not interested in the entire history axis. Rather, the information on

the number of molecules  $X_k(t)$  at the end of each interval of length  $\tau$  is sufficient. The question arises: can we determine  $X_k(t)$  at the end of  $\tau$  interval? If so, and if we can leap over many reaction over the time step  $\tau$ , then significant speed up gains of the simulation are obtained.

Let us assume that  $K$  reactions took place during  $[t, t + \tau)$ . Then

$$X(t + \tau) = X(t) + \nu_{j_1} + \nu_{j_2} + \dots \nu_{j_k} \quad (3.1)$$

where each of  $j_1, j_2, \dots, j_k$  can take up values from 1 to  $M$ . If we collect the reactions of same type and  $R_1$  occurred  $L_1$  times,  $R_2$ ,  $L_2$  times, and so on, then equation (3.1) can be written as

$$X(t + \tau) = X(t) + \nu_1 L_1 + \nu_2 L_2 + \dots \nu_M L_M,$$

thus,

$$X(t + \tau) = X(t) + \sum_{j=1} \nu_j L_j, \quad (3.2)$$

where  $L_j$  is the number of times reaction  $R_j$  took place during the interval  $[t, t + \tau)$ , to change  $X_k(t)$  by the state-change vector  $\nu_j$ . If we could estimate all  $L_j$ , over  $[t, t + \tau)$  then, we can estimate  $X_k$  at  $t + \tau$  knowing  $X_k(t)$ . The method to estimate  $L_j$  was proposed by Gillespie in the form of the explicit tau-leaping algorithm [11].

## 3.2 Explicit Tau-Leaping

Let us assume that  $\tau > 0$  is some time step and

$$L_j(\tau|\mathbf{x}, t) = \text{the number of reactions of type } R_j \text{ that took place during the interval } [t, t + \tau) \text{ provided that } \mathbf{X}(t) = \mathbf{x}. \quad (3.3)$$



Then, the number of molecules at the end of the leap are given by

$$\mathbf{X}(t + \tau) = \mathbf{x} + \sum_{j=1}^M \nu_j L_j(\tau | \mathbf{x}, t). \quad (3.4)$$

We are interested in finding a good approximation for  $L_j(\tau | \mathbf{x}, t)$ . Recall that the product  $a_j(\mathbf{X}(t))dt$  gives the probability that the reaction  $R_j$  will fire during  $[t, t + dt)$  for an infinitesimal step  $dt$ . If  $a_j(\mathbf{X}(t))$  does not change significantly during  $[t, t + \tau)$ , then  $a_j(\mathbf{X}(t))\tau$  is a good average estimate of the number of  $R_j$  reactions fired during this interval. In fact, drawing from a Poisson distribution with mean (and variance)  $a_j(\mathbf{X}(t))\tau$  is a good approximation for  $L_j(\tau | \mathbf{x}, t)$ . According to Gillespie [11], the following condition is sufficient.

**Leap Condition:** Choose  $\tau > 0$  small enough such that during the interval  $[t, t + \tau)$ , no individual propensity function  $a_j(\mathbf{X}(t))$  changes significantly.

Under the leap condition,  $L_j(\tau | \mathbf{x}, t)$  can be approximated as

$$L_j(\tau | \mathbf{x}, t) \sim \mathcal{P}_j(a_j(\mathbf{x})\tau) \quad (3.5)$$

if  $\mathbf{X}(t) = \mathbf{x}$ . Here  $\mathcal{P}_j(a_j(\mathbf{x})\tau)$  are independent Poisson random variables with means (and variances)  $a_j(\mathbf{x})\tau$ . Then equation (3.4) can be written as

$$\mathbf{X}(t + \tau) = \mathbf{x} + \sum_{j=1}^M \nu_j \mathcal{P}_j(a_j(\mathbf{x})\tau). \quad (3.6)$$

This is called the explicit tau-leaping method [11]. As discussed in Section 2.1, this strategy has similarities with the Euler's method applied to the RRE, which is an ODE. Euler's method has been developed to solve non-stiff system of ODEs. By contrast, the implicit Euler method was developed to efficiently approximate the solution of stiff systems. As expected, the tau-leaping formula given in equation (3.6) is not efficient for stiff systems (as discussed in Section 3.3). Therefore, it is called the 'explicit' tau-leaping method.

### 3.2.1 Algorithm for Explicit tau-Leaping

The algorithm for the explicit tau-leaping scheme can be summarised as:

Step 1 Initialize the state vector  $\mathbf{x} = x_0$  for time  $t = t_0$  and other system parameters.

Step 2 Choose an approximate value for the leap size  $\tau$ .

Step 3 Calculate the propensities  $a_j(\mathbf{x})$  for  $j = 1, 2, \dots, M$

Step 4 Draw  $M$ , Poisson random numbers  $l_j$  with mean (and variance)  $a_j(\mathbf{x})\tau$  each i.e.,

$$l_j = \mathcal{P}(a_j(\mathbf{x})\tau) \text{ for } j = 1, 2, \dots, M.$$

Step 5 Update the system state using the explicit tau-leaping formula:

$$\mathbf{x} \leftarrow \mathbf{x}(t) + \sum_{j=1}^M \nu_j L_j$$

Step 6 Update the current time from  $t \leftarrow t + \tau$ . If the updated time is less than the simulation end time, go to Step 3, otherwise STOP.

### 3.2.2 How accurate is the explicit tau-leaping strategy?

When the exact approach of the SSA is replaced by the (faster) explicit tau-leaping method, accuracy is lost. This loss of accuracy is due to the following (see also [32]):

- (i) We do not know the exact number of reactions that take place during the tau-leap interval, we approximate these numbers.
- (ii) We do not calculate, when a reaction took place during the tau-leap interval. Therefore, the variations in the state vector are not known during this time interval.

- (iii) The occurrence of a reaction changes the state vector. The increase or decrease in the number of molecules of one species influences the probability of occurrence of other reactions.

Another way to analyze the loss of accuracy and the speed-up, is through comparing the size of  $\tau$  with the that of the SSA:

- (a) None of the propensity functions  $a_j$ , for  $j = 1, 2, \dots, M$  depend upon  $\mathbf{X}(t)$  (trivial case). In this situation, the leap condition is always satisfied for any  $\tau$  and the  $\tau$ -leaping method will be exact.
- (b) The propensity functions are linearly or quadratically dependent on the molecular populations. If there are large number of molecules, reactions changing these numbers by one or two molecules will not significantly change the propensities. Therefore, the tau-leap method can run faster than the exact SSA.
- (c) For a leap size  $\tau$  comparable or less than the inverse of the sum of propensities (i.e.  $\frac{1}{\sum_{j=1}^M a_j}$ ), we approach the Gillespie algorithm simulation and there is no gain in computational time.

### 3.3 Implicit Tau-Leaping

Stiffness is an attribute of many biochemical systems in which reactions operate over very different time scales, for instance “fast” and “slow”, where the “fast” time scale is stable. Initially, a system follows a transient mode, which is short and rapid. The system then quickly moves into the slow reaction mode, which determines the overall long time dynamics of the system. Stiffness of a system requires both deterministic and stochastic models be solved using numerical methods designed specifically for such systems. Both the SSA and the explicit tau-leaping algorithms are inefficient for stiff systems [7, 15]. We first discuss

how stiffness is treated numerically in systems modelled using ODEs. The approach is then extended to the discrete-stochastic systems.

### 3.3.1 Stiffness in Continuous Deterministic ODE Systems

A stiff system is considered to be in a sort of equilibrium (but may not be in static) state. When any of the system variables are slightly perturbed, the system responds quite rapidly to get back to the equilibrium state. The stiff systems have a long (slow) quasi equilibrium phase and a transient (fast) short time phase after a perturbation or at the start up. The stiffness of the system depends upon how well distinct these two phases are. Consider an ODE model representing a stiff system:

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}, t), \quad \mathbf{x}(0) = \mathbf{x}_0 \quad (3.7)$$

The numerical solution for this system can be obtained using the "explicit Euler method" using

$$\mathbf{x}(t + \tau) = \mathbf{x}(t) + \tau \mathbf{f}(\mathbf{x}(t), t) \quad (3.8)$$

The accuracy of the solution  $\mathbf{x}$  at time  $t + \tau$  depends on the value of the time step  $\tau$ . The solution will approach the exact solution (more accurate) as  $\tau$  gets smaller.

For large  $\tau$ , the solution deviates away from the exact solution, resulting in large errors. In case of stiff systems, where the fast time scale is dominant at the beginning, we need to use small  $\tau$  for maintaining the statisticity of the simulation [7]. The stepsize taken by the explicit numerical method is restricted by the fast mode of the solution.

This restriction on the explicit Euler method makes it very slow as time steps greater than the time scale of the fast mode lead to unstable results. Can we use large time steps without causing numerical instability? The answer to this question lies in the use of "the implicit Euler method" [70]. This method approximates the derivative employing a backward difference formula instead of the forward difference formula used by the explicit Euler

method. Therefore, instead of equation (3.8), we consider

$$\mathbf{x}(t + \tau) = \mathbf{x}(t) + \tau \mathbf{f}(\mathbf{x}(t + \tau), t + \tau). \quad (3.9)$$

This strategy requires to find  $\mathbf{x}(t + \tau)$  iteratively using, for example, Newton's method for solving the implicit equation (3.9). This iterative process is repeated at every time step. This technique allows us to obtain stable numerical solutions using large time steps  $\tau$ , making the implicit Euler's method more efficient than the explicit Euler scheme for stiff problems. Remark that, the implicit Euler's method is unconditionally stable.

### 3.3.2 Stiffness in Discrete-Stochastic Chemical Kinetics: Implicit Tau-Leaping Method

The tau-leap size limitation of the explicit tau-leaping scheme for stiff problems was overcome by the implicit tau-leaping method proposed by Rathinam (2003)[15]. In the explicit tau-leaping strategy (3.6), the Poisson random variable  $\mathcal{P}(a_j, \tau)$  are evaluated at the start of the time step  $X(t)$ . In the implicit tau-leaping scheme, this Poisson random variable is considered to be made up of two parts:

1. The first part being the mean value  $a_j \tau$  of  $\mathcal{P}_j$ .
2. The second part being the zero mean random variable  $\mathcal{P}_j - a_j \tau$ .

The first part is evaluated at the unknown state  $\mathbf{X}(t + \tau)$  and the second part at the known state  $\mathbf{X}(t)$ , to give the implicit tau-leaping method:

$$\mathbf{X}(t + \tau) = \mathbf{x} + \sum_{j=1}^M (\tau a_j(\mathbf{X}(t + \tau)) + \mathcal{P}_j(a_j(\mathbf{x})\tau) - \tau a_j(\mathbf{x})) \nu_j \quad (3.10)$$

given that  $\mathbf{X}(t) = \mathbf{x}$ . The random variables  $\mathcal{P}_j$  are statistically independent. As in the case of the implicit Euler method for the deterministic case, equation (3.10) is required to be

solved iteratively using a variant of Newton's method for the solution of nonlinear systems of equations.

In case of large molecular number, the Poisson distribution is approximated by the Normal distribution to give

$$\mathbf{X}(t + \tau) = \mathbf{x} + \tau \sum_{j=1}^M \nu_j a_j(\mathbf{X}(t + \tau)) + \sqrt{\tau} \sum_{j=1}^M \nu_j \sqrt{a_j(\mathbf{x}(t))} N_j(0, 1) \quad (3.11)$$

In the thermodynamic limit, the stochastic term can be ignored to give

$$\mathbf{X}(t + \tau) = \mathbf{X}(t) + \tau \sum_{j=1}^M \nu_j a_j(\mathbf{X}(t + \tau)), \quad (3.12)$$

which is the implicit Euler method for solving the deterministic RRE. Rathinam et.al. [15] and Gillespie et.al. [7] noted that the natural fluctuations of the fast variables are *damped out by the implicit tau-leaping method* due to the use of large time step. According to [15], this observation applies to all well-stirred biochemical systems.

### 3.3.3 Implicit Tau-Leaping Algorithm

Starting with the implicit tau-leaping formula (3.10), we introduce

$$\mathbf{Z}(t) = \mathbf{x} + \sum_{j=1}^M [\mathcal{P}_j(a_j(\mathbf{x})\tau) - \tau a_j(\mathbf{x})] \nu_j \quad (3.13)$$

such that  $\partial \mathbf{Z} / \partial \mathbf{Y} = 0$ , where  $\mathbf{Y} = \mathbf{X}(t + \tau)$ . Then equation (3.10) can be written as

$$\mathbf{G}(\mathbf{Y}) = -\mathbf{Y} + \mathbf{Z} + \sum_{j=1}^M [\nu_j a_j(\mathbf{Y})\tau] = 0 \quad (3.14)$$

According to the Newton's method,

$$\mathbf{Y}^{(n+1)} = \mathbf{Y}^n - [\mathbf{J}(\mathbf{Y}^{(n)})]^{-1} \mathbf{G}(\mathbf{Y}^{(n)}), \quad (3.15)$$

where the Jacobian matrix  $\mathbf{J}$  is given by

$$\mathbf{J}(\mathbf{Y}) = \begin{pmatrix} \frac{\partial G_1}{\partial Y_1} & \frac{\partial G_1}{\partial Y_2} & \cdots & \frac{\partial G_1}{\partial Y_N} \\ \frac{\partial G_2}{\partial Y_1} & \frac{\partial G_2}{\partial Y_2} & \cdots & \frac{\partial G_2}{\partial Y_N} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial G_N}{\partial Y_1} & \frac{\partial G_N}{\partial Y_2} & \cdots & \frac{\partial G_N}{\partial Y_N} \end{pmatrix} \quad (3.16)$$

In the implicit tau-leaping, equation (3.15) is solved iteratively with some stopping criteria (such as the desired tolerance level is achieved) to obtain the state vector at the end of tau-leap interval. The steps involved for implementing the implicit tau-leaping scheme are summarized in the following.

1. Initialize the state vector  $\mathbf{x} = \mathbf{x}_0$  at time  $t = t_0$ . Other system parameters including the rate parameters, the steps 5 and 6, the tolerance  $\epsilon$  are specified.
2. Calculate  $\mathbf{Z}(t)$  using equation (3.13) by generating  $M$  Poisson random numbers with means  $a_j(\mathbf{x})\tau$  for  $j = 1, 2, \dots, M$ .
3. Determine the vector  $\mathbf{G}(\mathbf{Y})$  with the initial guess of  $\mathbf{Y}$  as  $\mathbf{x}$  using (3.14)
4. Calculate the Jacobian matrix using (3.16).
5. Update  $\mathbf{Y}$  using (3.15)
6. Check if error is below the tolerance  $\epsilon$ . If the error is larger, go to step (2) with  $\mathbf{x} = \mathbf{Y}$ . Otherwise store the state vector  $\mathbf{Y}$  and CONTINUE.
7. Update current time from  $t$  to  $t + \tau$ . If the updated time is less than the simulation end time, goto step (2), otherwise STOP.

### 3.4 Trapezoidal Tau-Leaping Method

To overcome the damping effect (the variance generated is much smaller than the exact value for large step sizes) of the implicit tau-leaping scheme, Cao and Petzold [21] proposed the trapezoidal tau-leaping in 2005. They extended the idea of the trapezoidal rule for solving numerically ODEs to modify the implicit tau-leaping. The trapezoidal rule was selected for the following reasons:

1. The explicit and implicit Euler formulae are both first order accurate, while the trapezoidal rule is second order accurate, when applied to ODEs.
2. The trapezoidal rule is A-stable [70, 71] while the implicit Euler method is L(linearly) stable.
3. The trapezoidal rule does not suffers from the damping effect.

To obtain the trapezoidal tau-leaping method, consider the implicit tau-leaping scheme (3.10) and replace  $\tau$  by  $\frac{\tau}{2}$  in the two summation terms appearing on the right hand side:

$$\mathbf{X}(t + \tau) = \mathbf{x} + \sum_{j=1}^M \left[ \frac{\tau}{2} a_j(\mathbf{X}(t + \tau)) + \mathcal{P}_j(a_j(\mathbf{x})\tau) - \frac{\tau}{2} a_j(\mathbf{x}) \right] \nu_j, \quad (3.17)$$

given  $\mathbf{X}(t) = \mathbf{x}$ . This is the trapezoidal tau-leaping method. The only difference between this method and the implicit tau-leaping is in terms of the coefficients of  $a_j(\mathbf{x})$  and  $a_j(\mathbf{X}(t + \tau))$ ,  $M$  Poisson random numbers are generated and Newton's iterative method is used to obtain the update state vector at the end of each leap. The summation on the right hand side of equation (3.17) gives a real number which needs to be rounded to the nearest integer to give the state vector  $\mathbf{X}(t + \tau)$ .



### 3.4.1 Trapezoidal Tau-Leaping Algorithm

Following the steps laid down in [21], the trapezoidal tau-leaping algorithm can be summarized as:

1. Initialization: Set the initial number of molecules  $\mathbf{x} = \mathbf{x}_0$ , at time  $t = 0$ ; set the leap size  $\tau$  and other system parameters.
2. Calculate the propensity functions  $a_j(\mathbf{x})$  for  $j = 1, 2, \dots, M$ .
3. Generate  $M$  independent Poisson random numbers  $\mathcal{P}_j(a_j(\mathbf{x})\tau)$  with mean  $a_j(\mathbf{x})\tau$ .
4. Solve equation (3.17) using Newton's method.
5. Calculate

$$L_j = \text{ROUND} \left( \left[ \mathcal{P}(a_j(\mathbf{x})\tau) - \frac{\tau}{2}a_j(\mathbf{x}) + \frac{\tau}{2}a_j(\mathbf{X}(t + \tau)) \right] \right) \quad (3.18)$$

where  $\text{ROUND}(c)$  function rounds each element of  $c$  to the nearest integer. In case of tie, it will round the element to the largest integer.

6. Update the state of the system using

$$\mathbf{X} \leftarrow \mathbf{x} + \sum_{j=1}^M L_j \nu_j \quad (3.19)$$

and update  $t \leftarrow t + \tau$ .

7. Check if  $t$  has reached the end of the simulation. STOP if yes, otherwise go to step (2).

# Chapter 4

## Numerical Results and Discussion

In this Chapter, we study the accuracy of the tau-leaping methods and, in particular, of the implicit tau-leaping method. In [15], it was stated that the implicit tau-leaping strategy is damping the noise in the fast variables for *all* models of well-stirred biochemical systems. However, we observed that this is not the case for biochemically reacting systems which do not reach a steady state, but have a different qualitative behavior. Examples include models with bi-stability where, due to the noise some paths switch between deterministic steady-states. Our claim is, that for these systems, the implicit tau-leaping methods are accurate in both fast and slow variables.

### 4.1 The Schlögl Model

The Schlögl model [72, 73, 74] serves as an excellent example of a chemically reacting system showing bi-stable behavior. The Schlögl reacting system is composed of three chemical species (labeled  $A$ ,  $B$  and  $S$ ) reacting through four chemical reactions. The chemical reactions along with their respective propensities are given in Table 4.1. The number of molecules of the species  $A$  and  $B$  are kept constant during the entire time span of the experiment. If  $X_A$ ,  $X_B$  and  $X$  represent the number of molecules of the three species  $A$ ,  $B$  and  $S$ , then the

state vector  $\mathbf{X}$  can be written as

$$\mathbf{X} = \begin{pmatrix} X \\ X_A \\ X_B \end{pmatrix}$$

The stoichiometric matrix  $\boldsymbol{\nu}$ , for the reacting system will be:

$$\boldsymbol{\nu} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix},$$

where the four columns of the stoichiometric matrix  $\boldsymbol{\nu}$  represent the four state change vectors  $\boldsymbol{\nu}_1$ ,  $\boldsymbol{\nu}_2$ ,  $\boldsymbol{\nu}_3$  and  $\boldsymbol{\nu}_4$  corresponding to the four reactions  $R_1$ ,  $R_2$ ,  $R_3$  and  $R_4$ , respectively.

Case	Reaction	Propensity
1	$R_1 : A + 2S \xrightarrow{c_1} 3S$	$a_1(X) = c_1 X_A X(X-1)/2$
2	$R_2 : 3S \xrightarrow{c_2} A + 2S$	$a_2(X) = c_2 X(X-1)(X-2)/6$
3	$R_3 : B \xrightarrow{c_3} S$	$a_3(X) = c_3 X_B$
4	$R_4 : S \xrightarrow{c_4} B$	$a_4(X) = c_4 X$

**Table 4.1:** The reactions and propensities in the Schlögl model.

For our simulations, the number of molecules of species  $A$  and  $B$  have been taken to be  $X_A = 10^5$  and  $X_B = 2 \times 10^5$ , respectively. The initial number of molecules of species  $S$  is  $X(0) = 250$ . The stochastic reaction rate parameters are  $c_1 = 3 \times 10^{-7}$ ,  $c_2 = 10^{-4}$ ,  $c_3 = 10^{-3}$  and  $c_4 = 3.5$ . The interval of integration is  $[0, 4]$ .

We begin with the continuous deterministic description of this model. According to the RRE (2.55)

$$\frac{dy(t)}{dt} = \sum_{j=1}^M \nu_j a_j(y(t))$$

where the number of molecules are represented by a continuous real-valued variable  $y(t)$ .

For the Schlögl model, there is only one reacting species which varies in time, namely  $S$ . Therefore, the RRE can be written as

$$\frac{dy(t)}{dt} = \nu_{11}a_1(y(t)) + \nu_{12}a_2(y(t)) + \nu_{13}a_3(y(t)) + \nu_{14}a_4(y(t)) \quad (4.1)$$

Substituting the stoichiometric coefficients, we derive

$$\frac{dy(t)}{dt} = a_1(y(t)) - a_2(y(t)) + a_3(y(t)) - a_4(y(t))$$

After substituting the propensity functions according to Table 4.1 and simplifying, we get

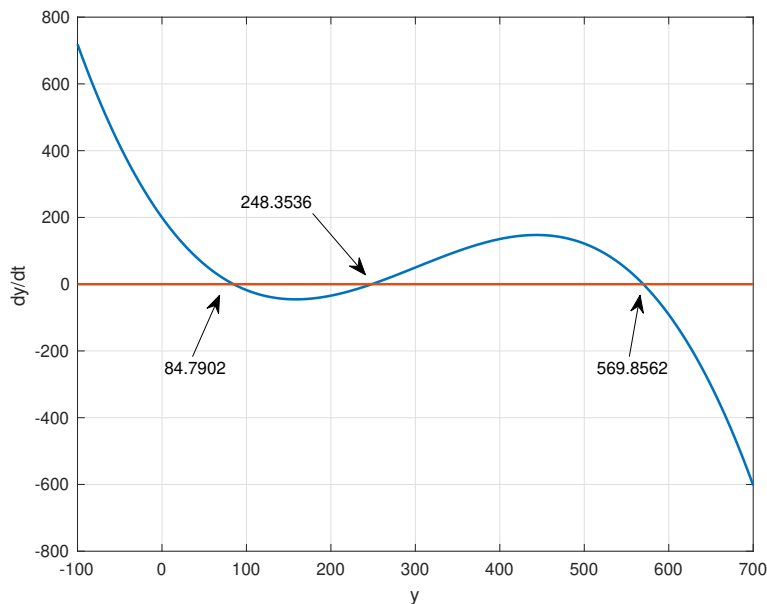
$$\frac{dy(t)}{dt} = -\left(\frac{c_2}{6}\right)y^3 + \left(\frac{c_1X_A}{2} + \frac{c_2}{2}\right)y^2 - \left(\frac{c_1X_A}{2} + \frac{c_2}{3} + c_4\right)y + c_3X_B \quad (4.2)$$

The following ODE is obtained after substituting the constant parameter values,

$$\frac{dy(t)}{dt} = -1.666667 \times 10^{-5}y^3 + 1.505 \times 10^{-2}y^2 - 3.515033y + 200 \quad (4.3)$$

The right hand side of this ODE is a cubic polynomial with three roots:  $r_1 = 84.7902$ ,  $r_2 = 248.3536$  and  $r_3 = 569.8562$  as shown in Figure 4.1. The middle root  $r_2$  corresponds to the unstable equilibrium, while the other two ( $r_1$  and  $r_3$ ) represent the two stable equilibrium values. Therefore, the Schlögl model serves as an excellent example of a *bistable system*. This bistable behavior is shown in Figure 4.2, where the solution of equation (4.3) was plotted with slightly different values of the initial number of molecules ( $X_0 = 248$  and  $X_0 = 249$ ). It can be noticed that the solutions converge to the two stable equilibrium roots  $r_1$  and  $r_3$ .

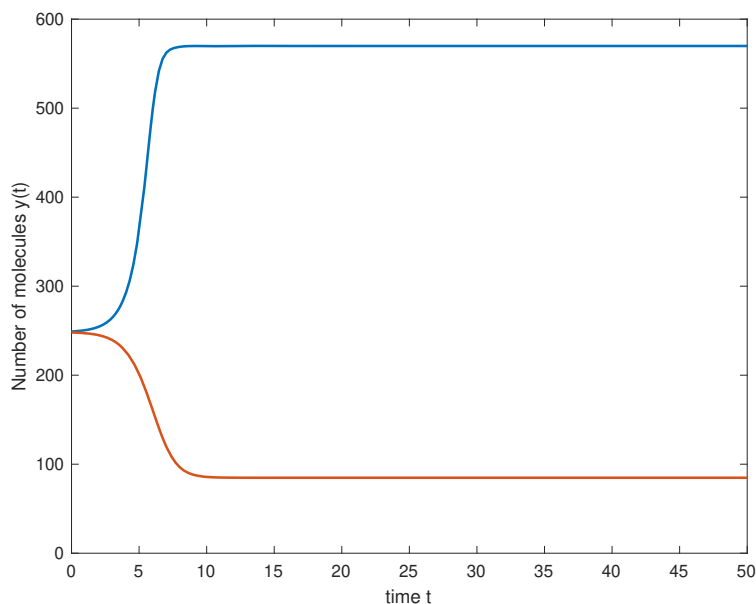
An interesting question can be asked at this stage: At some randomly chosen instant of time, near which of the two stable states, will the system be? To answer this question, we solve the RRE (4.3), repeatedly with different initial conditions. The results are shown in the form of a family of curves shown in Figure 4.3, all being converging to either of two stable



**Figure 4.1:** The deterministic rate of change of molecules with time using the RRE for the Schlögl model, indicating the roots of the rate function.

states. It can be noted that, if the system starts with an initial value  $X_0 < r_2$  (the unstable equilibrium root value), then it always converges to the stable state corresponding to the root  $r_1$ . Similarly, if  $X_0 > r_2$ , the system always converges to the stable state corresponding to the root  $r_3$ . Therefore, deterministically, it is the initial condition that dictates the future behavior of the system. Once the system is in any of the two stable states, it stays in that state. There is no question of jumping of the system from one state to the other.

We now turn to the discrete, stochastic model of the CME for this system. We first run the SSA simulation for the Schlögl model from time  $t = 0$  to  $t = 10$ . Figure 4.4 shows ten trajectories with  $X_0 = 250$  as the initial condition. Each trajectory represents the state of the system as it evolves in time. A closeup of one of the trajectories is shown in Figure 4.5. As expected, the state of the system does not follow a smooth continuous curve (as was found in the deterministic case). The number of molecules changes abruptly at random moments in time, resulting in jumps in the state of the system. Figure 4.4 shows repeated runs of the



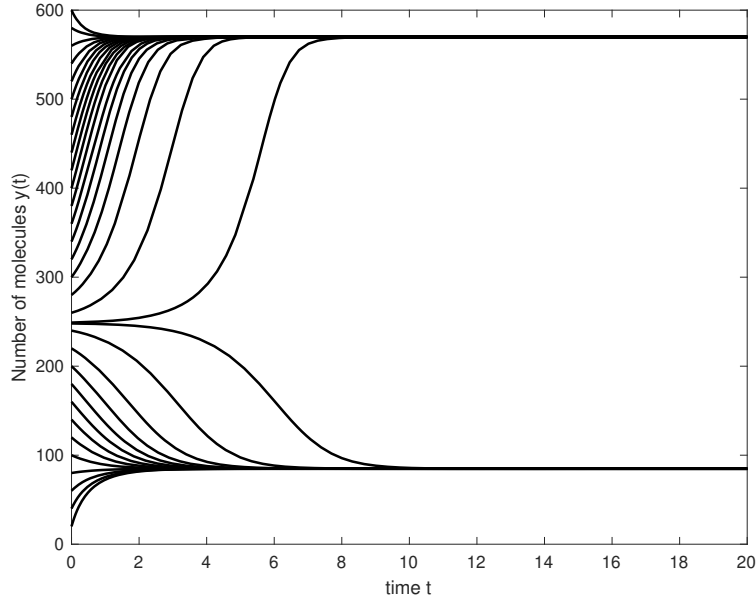
**Figure 4.2:** Continuous, deterministic solution of the Schlögl model showing bi-stability of the system.

SSA and each time a new trajectory is formed. This is unlike the deterministic case, where only one trajectory is formed on repeated runs of the RRE simulation with the same initial condition. Starting with the same initial condition, a trajectory for the stochastic model can evolve into any of the two stable states.

The SSA was run with different initial number of molecules (taken from the set):

$$\{50, 100, 150, 230, 248, 270, 400, 450, 500\}$$

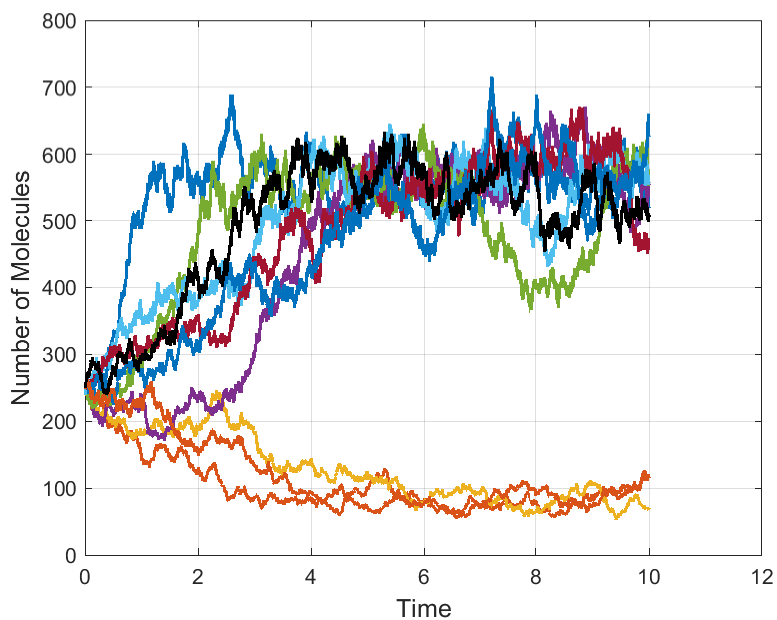
The simulation results are shown in Figure 4.6. For initial values far away from the unstable equilibrium point ( $r_2$ ), the SSA trajectories converge toward either of the two stable states ( $r_1$  or  $r_3$ , but not both) that is nearer to the initial value. However, as the initial number of molecules approach the unstable equilibrium value ( $r_2$ ), the chance that the system can converge to any of the two stable states increases. We can determine the probability that the system will converge to a stable state.



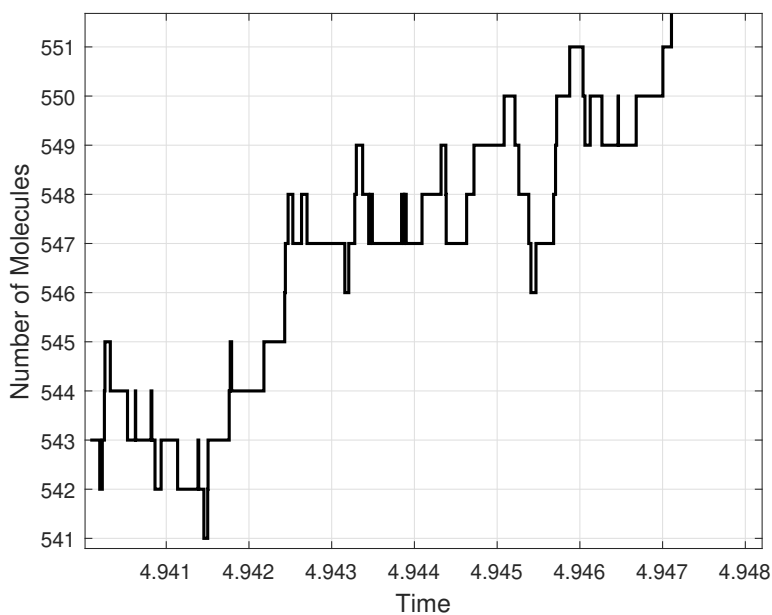
**Figure 4.3:** The RRE solutions of the Schlögl model using different initial conditions.

Another interesting behavior found in the stochastic modeling is that of *State Switching*, whereby a system being in one of the stable states for some time, spontaneously changes to the other state. This is unlike the behaviour for the deterministic model, where once a system converges to a stable state, it does not switch to the other state at a later time. We ran the SSA simulations for a longer period of time (from  $t = 0$  to  $t = 40$ ) and have observed the switching of state (as shown in Figure 4.7). The simulations show a trajectory where the system (after attaining the upper stable state) suddenly switches to the lower stable state. This behavior is only possible due to the intrinsic noise of the system. This is in total contrast with the deterministic picture (see Figure 4.3).

We also simulated the Schlögl model with the explicit tau-leaping method. The state switching was also observed in the simulation results as shown in Figure 4.8. In this case, for one of the trajectories, after attaining a stable state, the system moves spontaneously from the lower stable state to the upper stable state. This shows that even with approximate stochastic simulations, the intrinsic noise of the system plays its role and the system exhibits



**Figure 4.4:** Several trajectories of the Schlögl model simulated with the SSA.



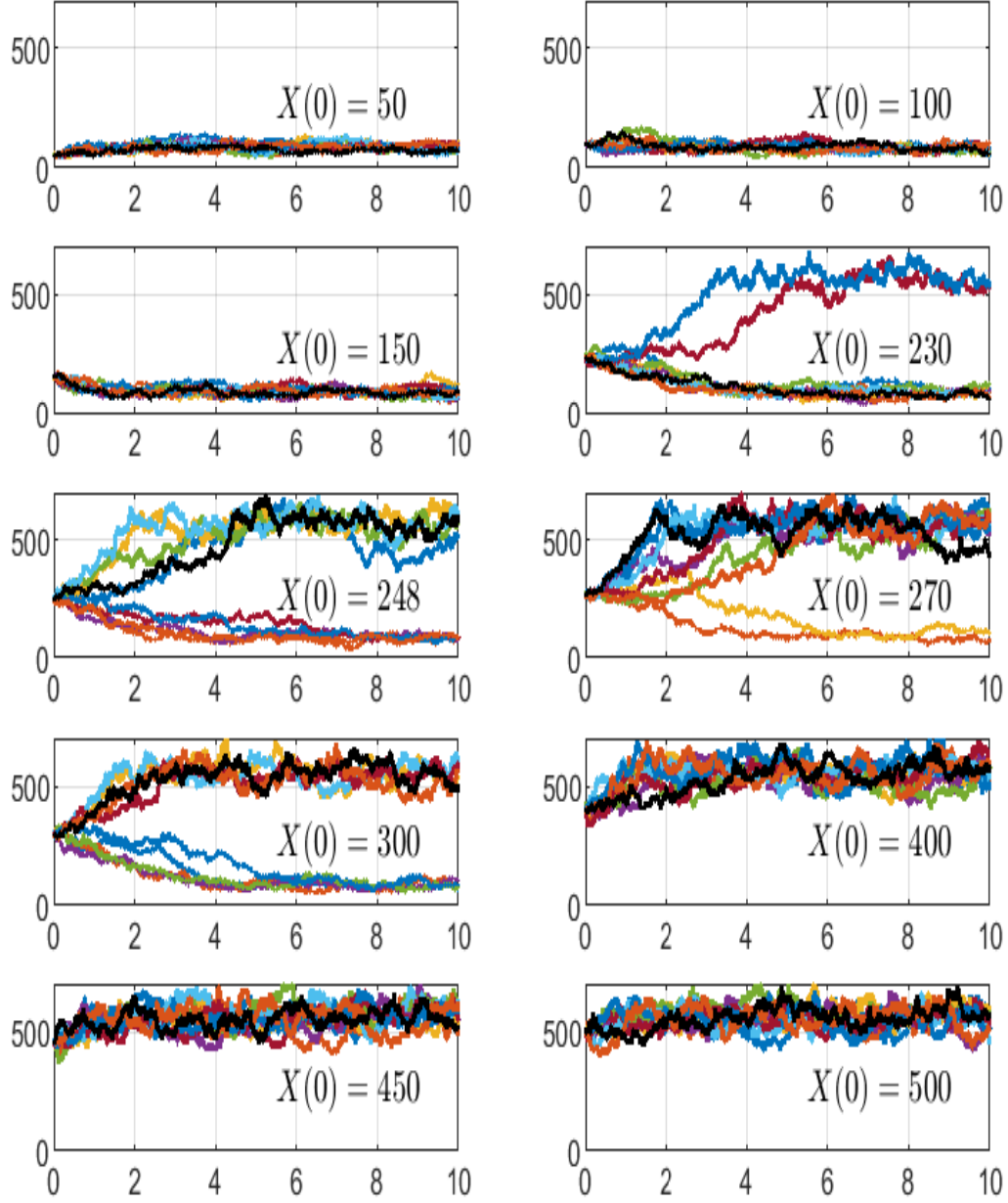
**Figure 4.5:** Markov jumps in an SSA solution of the Schlögl model.



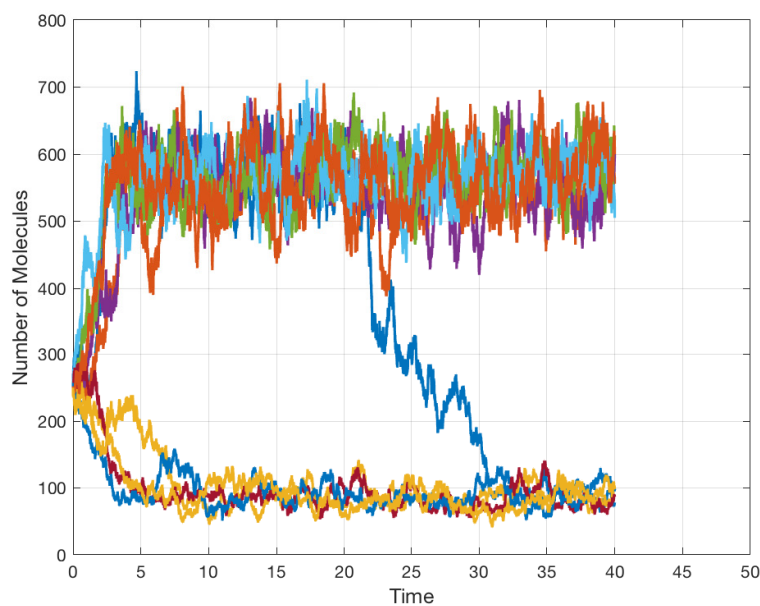
the state switching behavior.

We have implemented the three tau-leaping methods (i.e., explicit, implicit and trapezoidal tau-leaping) along with the SSA as discussed in Chapter 3 in MATLAB®. All simulations were carried out for  $10^4$  trajectories for each method. Each trajectory was obtained for the time interval from  $t = 0$  to  $t = 4$ . A leap of size  $\tau = 0.0125$  was used in all the simulations using tau-leaping methods. The probability distributions generated using the SSA, the explicit, implicit and trapezoidal tau-leaping strategies with fixed stepsize of  $\tau = 0.0125$  are shown in Figure 4.9. The approximate tau-leaping methods closely follows the SSA results. Compared to the other two tau-leaping methods, the trapezoidal tau-leaping method overshoots at some values of  $X$  near the two peaks. Moreover, the mean and standard deviations of the number of  $S$  molecules are calculated for 10,000 trajectories obtained using the three tau-leaping methods. The mean values as functions of time obtained from the three tau-leaping methods along with the corresponding values obtained from the SSA are plotted in Figure 4.10. The approximate mean values are very close to those from the SSA runs. The standard deviations obtained for all four methods as functions of time are shown in Figure 4.11.

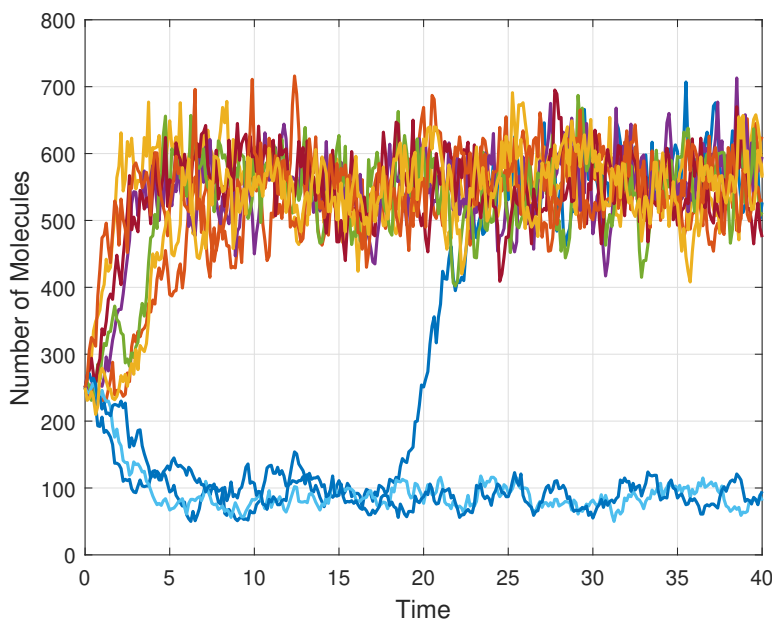
The interesting thing to note is that the standard deviations obtained with the tau-leaping methods match very well with those obtained with the SSA. Therefore, it is inferred that the implicit tau-leaping method by itself gives results comparable to those of the trapezoidal tau-leaping method which is considered to be a higher order method (see Chapter 3). The mean values obtained using the implicit tau-leap strategy are also found to be very close to those obtained with the trapezoidal tau-leap method as shown in Figure 4.10. Thus, for this bi-stable model, the noise is not damped by the implicit tau-leaping strategy and no additional technique is required to maintain accuracy.



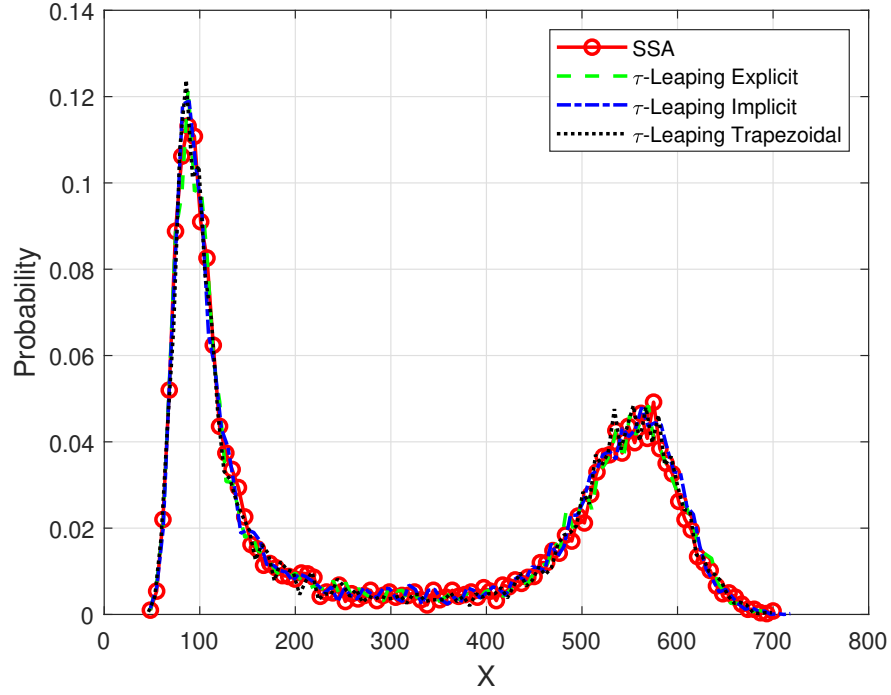
**Figure 4.6:** SSA solutions of the Schlögl model with different initial conditions.



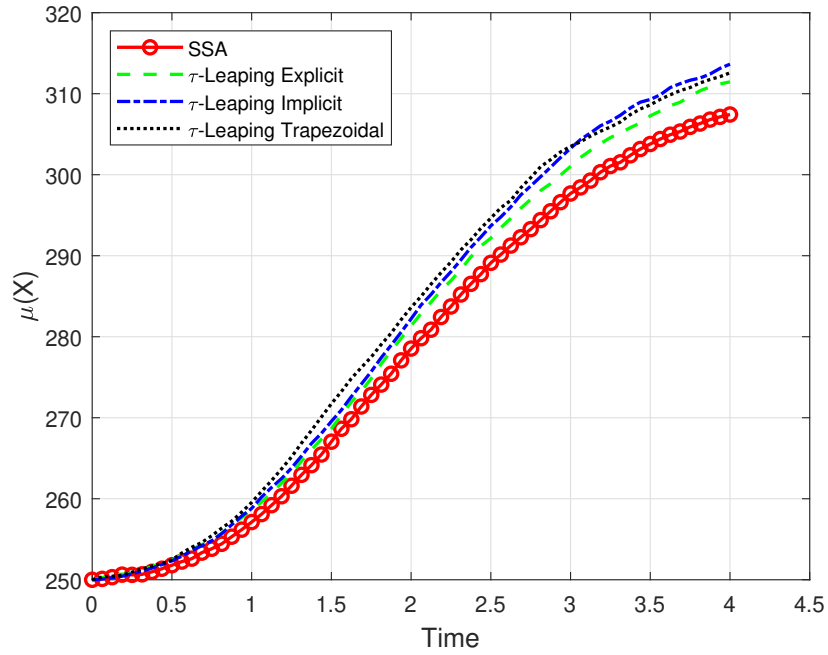
**Figure 4.7:** The state switching in some SSA trajectories for the Schlögl model. Only 10 trajectories are shown.



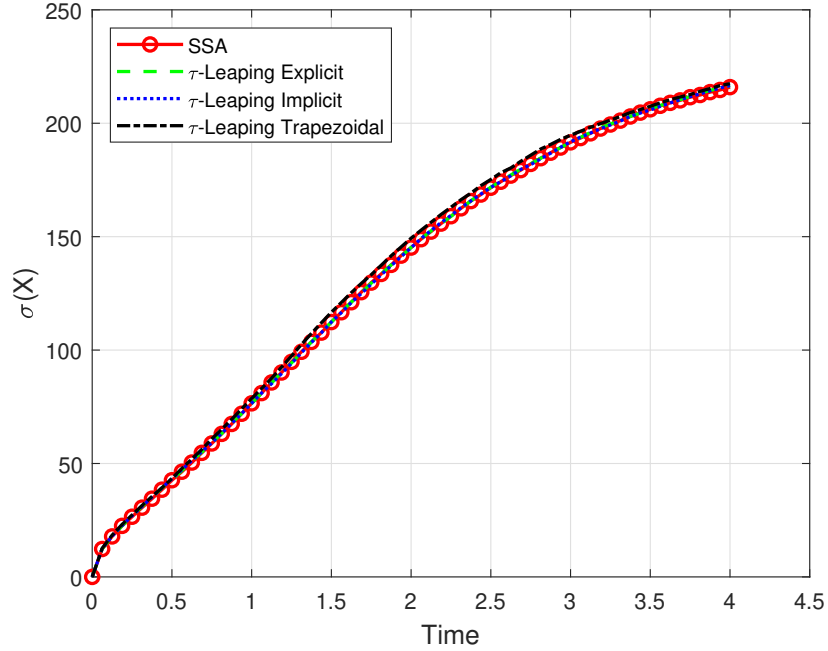
**Figure 4.8:** The state switching in explicit tau-leaping trajectories of the Schlögl model. 10 trajectories are shown.



**Figure 4.9:** Schlögl model: Comparison of the probability distributions at  $T = 4$  obtained using 10,000 simulations with the SSA and the  $\tau$ -leaping methods.



**Figure 4.10:** Schlögl Model: Comparison of the mean values as functions of time, for 10,000 trajectories using the SSA and the  $\tau$ -leaping methods.



**Figure 4.11:** Schlögl Model: Comparison of the standard deviation values as functions of time for 10,000 trajectories using the SSA and the  $\tau$ -leaping methods.

## 4.2 The Goldbeter-Koshland Switch

Goldbeter and Koshland (1981) studied this biochemically reacting system involving protein modification by effector and modifying enzymes [75]. This system, also called the Goldbeter-Koshland switch (GK switch), consists of a pair of substrate ( $S$ )- product ( $P$ ) proteins that are interconverted by two enzymes ( $E_1$  and  $E_2$ ), making two intermediate compound molecules ( $K_1$  and  $K_2$ ). The GK switch was studied by Melykuti *et al* [16] using stochastic modeling. Ahn *et. al.* [76] also implimented the GK switch using the SSA, the explicit and the implicit tau leaping methods with a different set of initial conditions. The GK switch consists of  $N = 6$  chemically reacting species subjected to  $M = 6$  chemical reactions. The chemical reactions along with the propensity functions are given in Table 4.2. The state of the system is represented by the vector  $(X_1, X_2, X_3, X_4, X_5, X_6)^T$ , where  $X_j$  for  $j = 1, 2, \dots, 6$  are the number of molecules of the reacting species corresponding to  $(S, E_1, K_1, P, E_2, K_2)$ .

Label	Reaction	Propensity
$R_1$	$S_1 + S_2 \xrightarrow{c_1} S_3$	$a_1(X) = c_1 X_1 X_2$
$R_2$	$S_3 \xrightarrow{c_2} S_1 + S_2$	$a_2(X) = c_2 X_3$
$R_3$	$S_3 \xrightarrow{c_3} S_4 + S_2$	$a_3(X) = c_3 X_3$
$R_4$	$S_4 + S_5 \xrightarrow{c_4} S_6$	$a_4(X) = c_4 X_4 X_5$
$R_5$	$S_6 \xrightarrow{c_5} S_4 + S_5$	$a_5(X) = c_5 X_6$
$R_6$	$S_6 \xrightarrow{c_6} S_1 + S_5$	$a_6(X) = c_6 X_6$

**Table 4.2:** The Goldbeter-Koshland switch parameters.

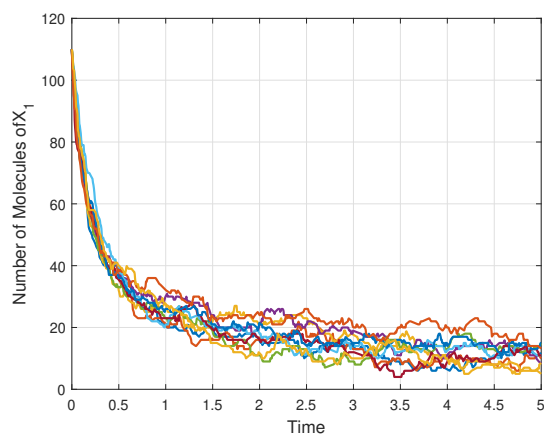
The stoichiometric matrix is given by:

$$\boldsymbol{\nu} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & 1 & -1 & -1 \end{pmatrix}$$

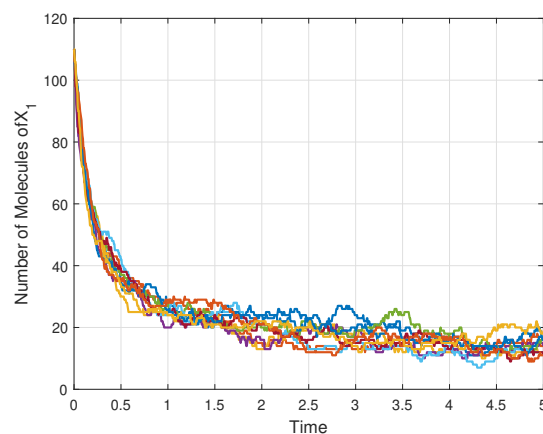
where the six columns of the matrix  $\boldsymbol{\nu}$  represent the six state change vectors  $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3, \boldsymbol{\nu}_4, \boldsymbol{\nu}_5$  and  $\boldsymbol{\nu}_6$  for the six reactions  $R_1, R_2, R_3, R_4, R_5$  and  $R_6$ , respectively.

The reaction rate constant values used to simulate the GK switch are:  $c_1 = 0.05, c_2 = 0.1, c_3 = 0.1, c_4 = 0.01, c_5 = 0.1$  and  $c_6 = 0.1$ . The initial state of the system is considered to be  $(110, 100, 30, 30, 100, 30)^T$ . Ten trajectories using the SSA, the explicit, the implicit and the trapezoidal tau-leaping schemes are plotted in Figure 4.12 for the substrate  $S$  ( $X_1$ ) and in Figure 4.13 for the product  $P$  ( $X_4$ ). The trajectories obtained using the tau-leaping methods are similar to those using the SSA.

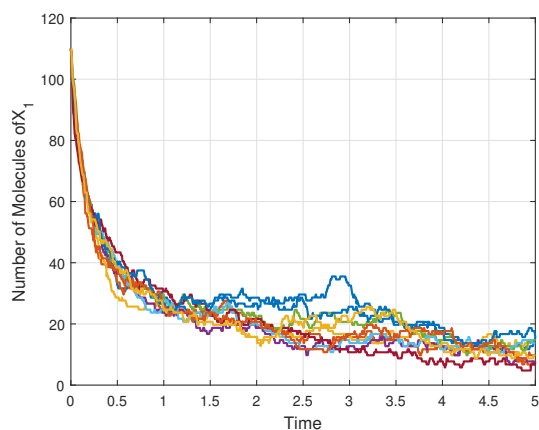
Figure 4.14 shows the probability distributions obtained from simulating the three tau-leaping methods and the SSA for all the chemical species. Each simulation was carried out



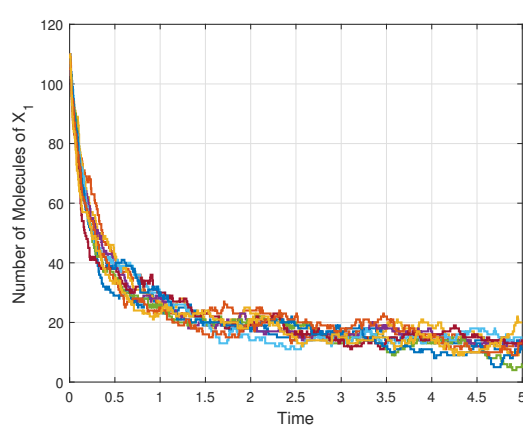
(a) SSA trajectories.



(b) Explicit tau-leaping

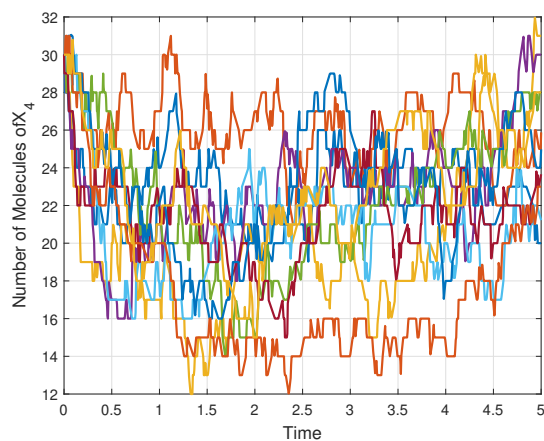


(c) Implicit tau-leaping

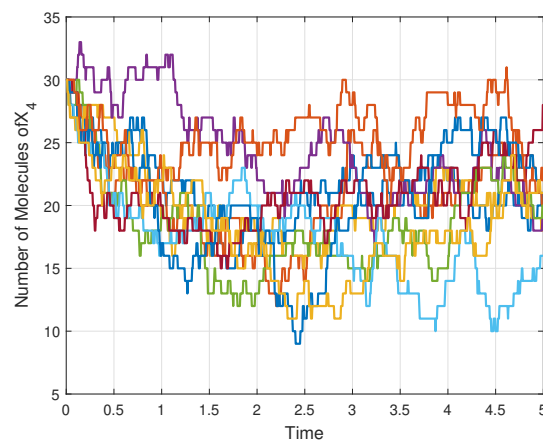


(d) Trapezoidal tau-leaping

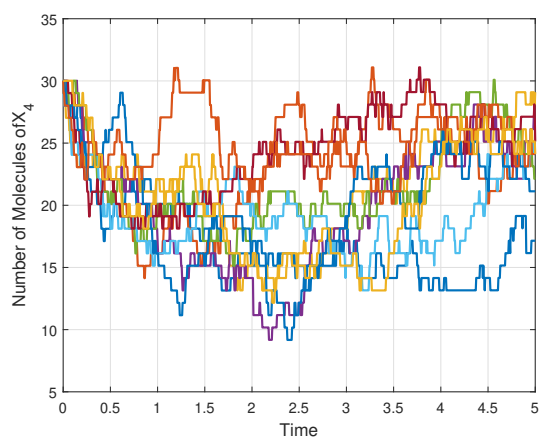
**Figure 4.12:** Goldbeter Koshland switch - Ten trajectories for the substrate  $S$  ( $X_1$ ) as a function of time, using the SSA and the tau-leaping methods.



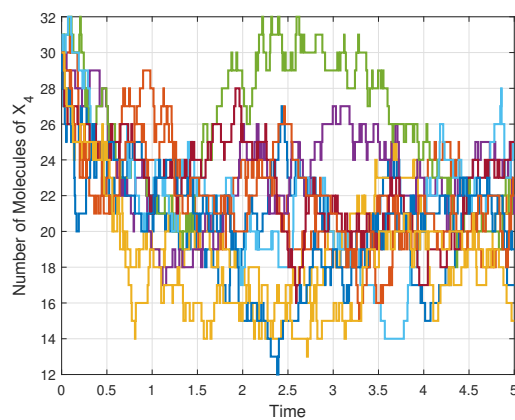
(a) SSA trajectories.



(b) Explicit tau-leaping



(c) Implicit tau-leaping



(d) Trapezoidal tau-leaping

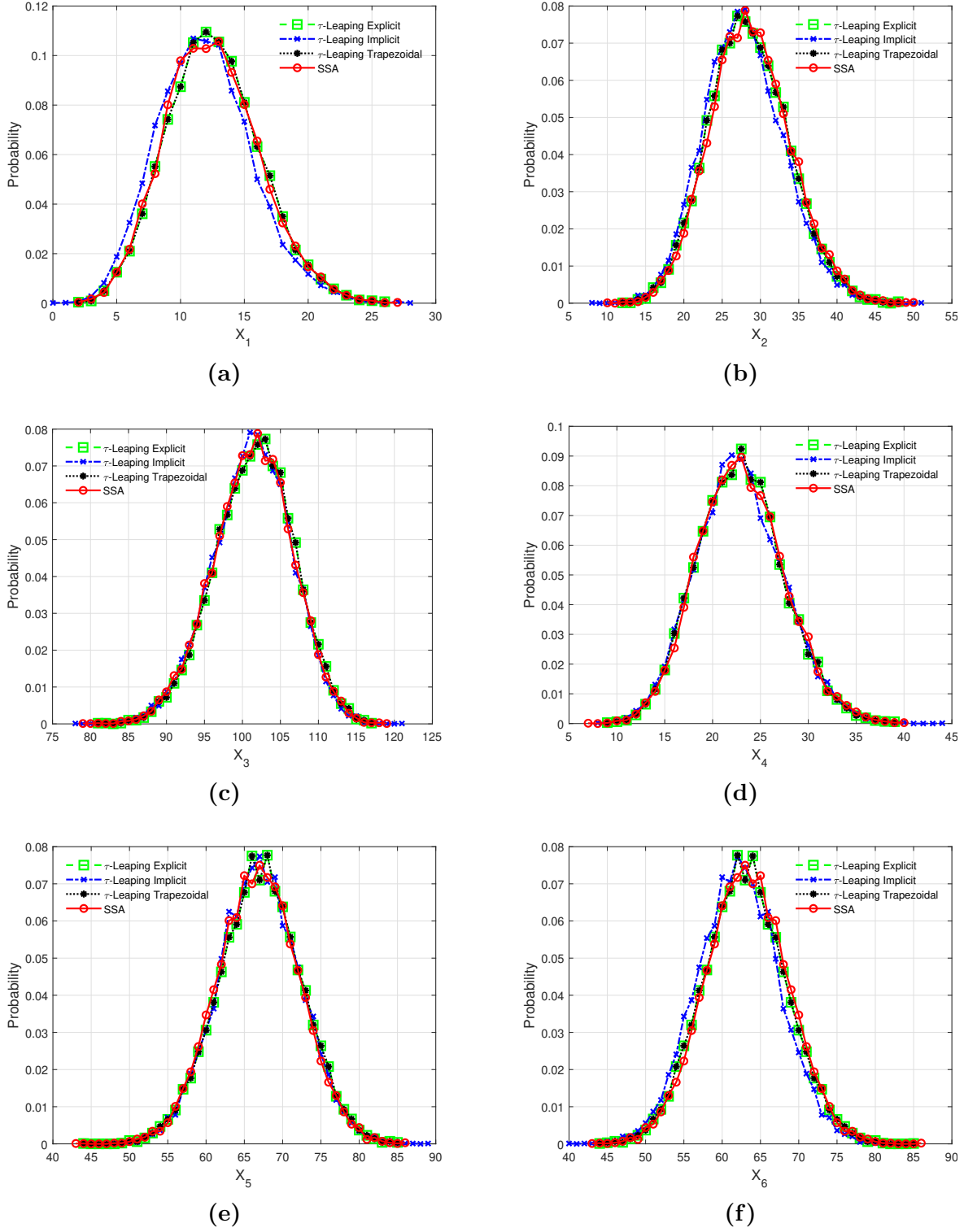
**Figure 4.13:** Goldbeter Koshland switch - Ten trajectories for the product  $P(X_4)$  as a function of time, using the SSA and the tau-leaping methods.



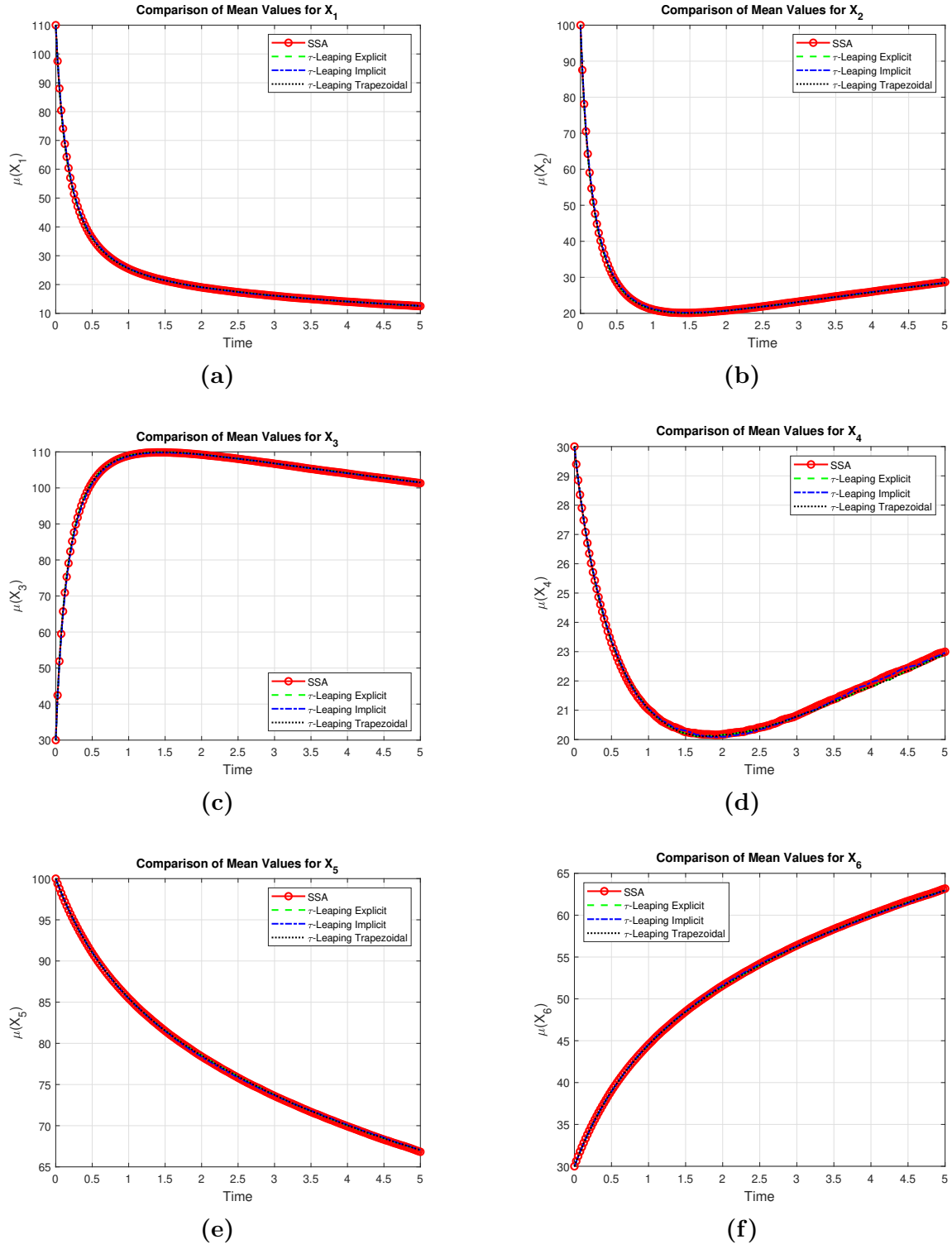
for 10,000 trajectories. Each trajectory was obtained from  $t = 0$  to  $t = 5$ . A step size  $\tau = 0.005$  was used for the three tau-leaping methods. More oscillations in the explicit and the trapezoidal tau-leap are found near the peak values of the probability distributions. The implicit tau-leaping method gives results closer to those obtained with the exact SSA near the peak values.

The mean and the standard deviation of the number of molecules as functions of time are estimated for 10,000 trajectories obtained with the SSA and the tau-leaping methods. These means and standard deviations are plotted for each reacting species in Figures 4.15 and 4.16. The tau-leaping estimations of the mean and standard deviation  $X_t$ -values match closely those computed using the exact SSA. It can be inferred that the implicit tau-leaping gives very accurate results compared to the exact strategy.

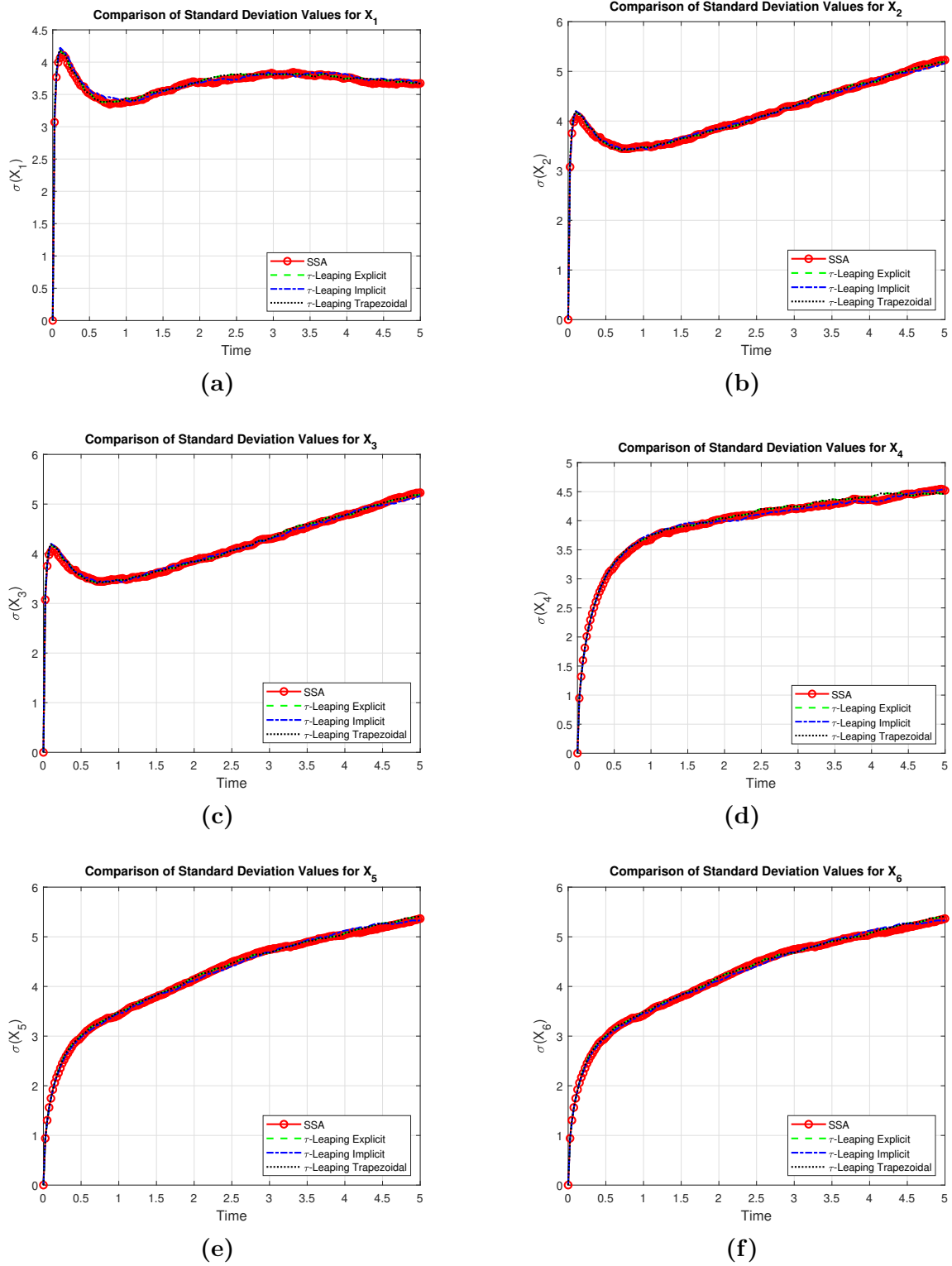
In conclusion, for this model, the implicit tau-leaping scheme does not reduce the noise in the system, as predicted in [15].



**Figure 4.14:** Goldbeter-Koshland switch - Comparison of probability distributions for all 6 species using  $10^4$  trajectories for each of the SSA, explicit, implicit and trapezoidal tau-leaping methods computed at  $T = 5$ .



**Figure 4.15:** Goldbeter-Koshland switch - Comparison of the mean values for all 6 species as functions of time, using  $10^4$  trajectories for each of the SSA, explicit, implicit and trapezoidal tau-leaping methods.



**Figure 4.16:** Goldbeter-Koshland switch - Comparison of the standard deviation values for all 6 species as functions of time, using  $10^4$  trajectories for each of the SSA, explicit, implicit and trapezoidal tau-leaping methods.

### 4.3 Decay Dimerization

This model was originally proposed by Gillespie [11] in 2001 and used by Rathinam [15] in 2003. There are  $N = 3$  chemical species interacting through  $M = 4$  chemical reactions. The reactions along with the propensities functions are given in Table (4.3). The reaction  $R_1$  is an irreversible isomerization. The species  $S_1$  decays with a rate constant  $c_1$ . Reactions  $R_2$  and  $R_3$  form a reversible dimerization of the monomer  $S_1$  into an unstable dimer  $S_2$ . The unstable dimer can also convert into a stable species  $S_3$ . The rate constants  $c_2$  and  $c_3$  are large compared to the other two rate constants. Therefore the reactions  $R_2$  and  $R_3$  take place at a faster rate. The presence of both fast and slow scales makes the problem of decay dimerization a stiff problem.

Label	Reaction	Propensity
$R_1$	$S_1 \xrightarrow{c_1} 0$	$a_1 = c_1 X_1$
$R_2$	$S_1 + S_1 \xrightarrow{c_2} S_2$	$a_2 = c_2 X_2 (X_2 - 1)/2$
$R_3$	$S_2 \xrightarrow{c_3} S_1 + S_1$	$a_3 = c_3 X_2$
$R_4$	$S_2 \xrightarrow{c_4} S_3$	$a_4 = c_4 X_2$

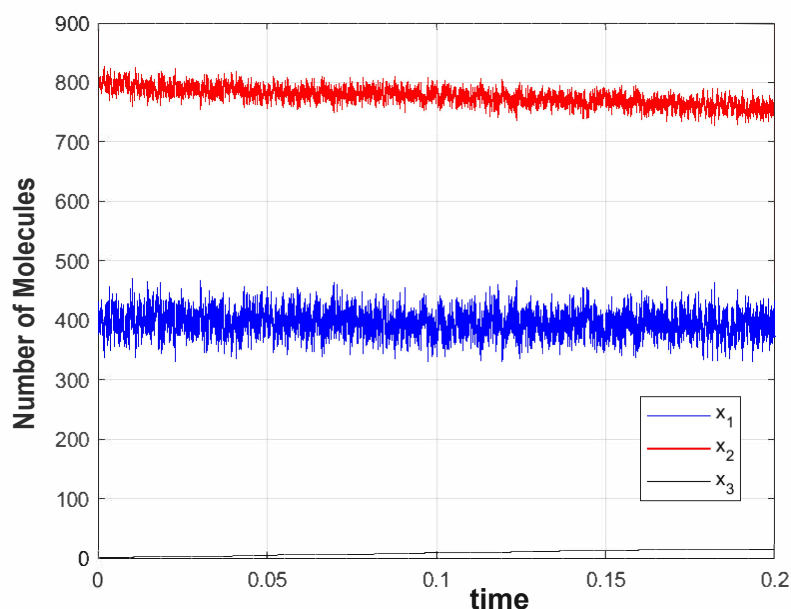
**Table 4.3:** The toggle switch parameters.

Figure 4.17 shows three representative trajectories when the SSA is run for this problem from  $t = 0$  to  $t = 0.2$  with rate constants  $c_1 = 1.0$ ,  $c_2 = 10.0$ ,  $c_3 = 1000.0$  and  $c_4 = 0.1$ . The initial number of molecules of the three species are  $X_1(0) = 400$ ,  $X_2(0) = 798$  and  $X_3(0) = 0$ . The stoichiometric matrix is given by:

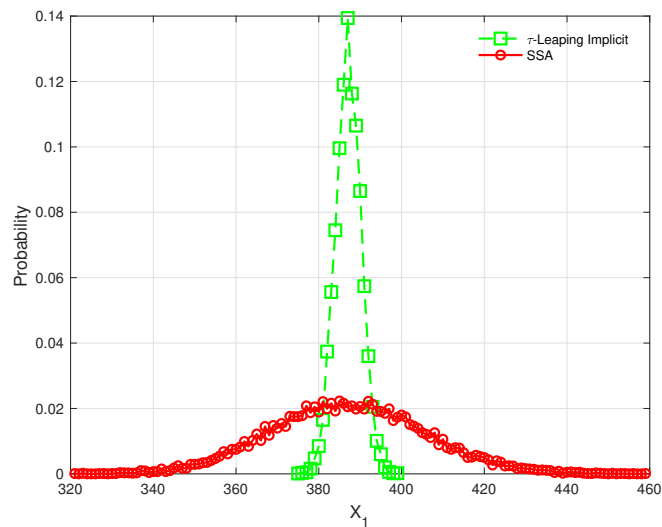
$$\nu = \begin{pmatrix} -1 & -2 & 2 & 0 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The figure indicates a fast decrease in the number of molecules of species  $S_1$  and  $S_2$  and a slow

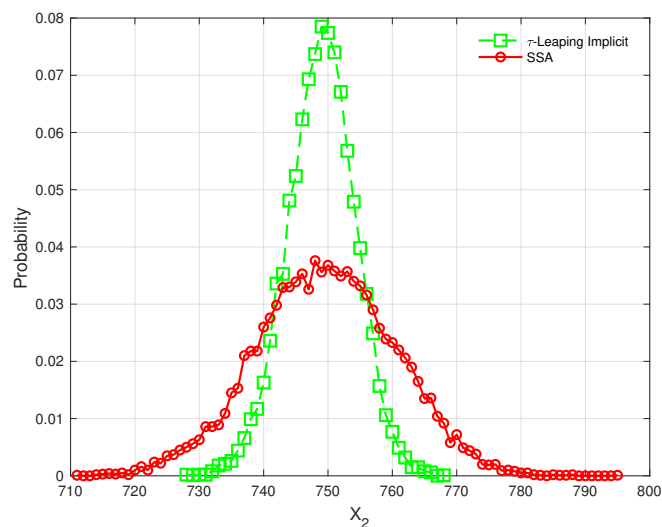
growth in the number of molecules of the stable species  $S_3$ . Figure 4.18 and Figure 4.19 give a comparison of the probability distributions for  $S_1$  and  $S_2$  species using 10,000 trajectories of the SSA and implicit tau-leaping method. The results indicate that the implicit tau-leaping method accurately models the non-steady state behavior of the system. Therefore, it could be inferred that for the non-steady state part of stiff problems, the implicit tau-leaping method is itself accurate and it does not require an interlacing with the exact SSA. The implicit tau-leaping gives a speed up of two orders of magnitude compared to the SSA simulation.



**Figure 4.17:** Decay Dimerization - SSA trajectories.



**Figure 4.18:** Decay Dimerization - Comparison of the probability distribution of  $S_1$  using 10,000 trajectories of the SSA and the implicit tau-leaping.



**Figure 4.19:** Decay Dimerization - Comparison of the probability distribution of  $S_2$  using 10,000 trajectories of the SSA and the implicit tau-leaping.

# Chapter 5

## Conclusion and Future Work

Stochastic modeling of biochemical processes at the level of a single cell is an important area of research. In particular, cellular processes involving some molecular species in low numbers (e.g. DNA or RNA) necessitate stochastic models for an accurate description of the behaviour. However, stochastic models are expensive to simulate, thus effective numerical methods to approximate their solution are desirable. Among the stochastic models of well-stirred biochemical system, among the most accurate and widely used, is the Chemical Master Equation. For the Chemical Master Equation, Gillespie proposed an exact Monte Carlo method, namely the stochastic simulation algorithm. This method is expensive on practical application, when fast reactions are present. Approximate tau-leaping strategies for the CME were developed.

In this thesis, three tau-leaping methods to find approximate solutions for CME have been investigated. The explicit tau-leaping method has improved the computational efficiency compared to the exact SSA for nonstiff models. The implicit and trapezoidal tau-leaping methods are stable for stiff systems having both slow and fast scales and more efficient than the SSA and explicit tau-leap strategies.

We studied in detail the behavior of the implicit tau-leaping method. Previous work stated that the implit tau-leaping scheme damped the noise in the fast variables. However, we



observed that, for systems not reaching steady-state, the implicit tau-leaping scheme is accurate for all variables.

We tested the behavior of the implicit tau-leaping method on two models not reaching a steady state: the Schlögl model (a bi-stable system) and the Goldbeter-Koshland switch (a stochastic switch system), both with large levels of noise and not reaching a steady-state. We compared the implicit tau-leaping simulation results with those of the SSA and other tau-leaping strategies and found that all variables (slow and fast) were accurately estimated.

By contrast, the implicit tau-leaping method reduces the noise of the fast variables for systems reaching a steady-state, as previously observed in the literature. We illustrated this by testing on a decay-dimerization model.

Stochastic simulation of biochemical reaction systems is a broad area. The development of efficient and accurate simulation algorithms to solve problems of practical interest modelled using the CME remains an active area of research. In the tau-leaping methods, the choice of the stepsize plays an important role in improving the efficiency of the simulations. This is one area where in which the current work can be extended in the future.

# Appendix A

## Propositions

*In this appendix, we have defined the terms used in this thesis.*

### Chapman Kolmogorov Equation / Theorem

According to Chapman Kolmogorov theorem, if  $X_n$ , where  $n = 0, 1, 2, \dots$  is a homogeneous Markov Chain, then

$$\mathcal{P}_r(X_{m+n} = j | X_0 = i) = \sum_{k \in S} \mathcal{P}_r(X_{m+n} = j | X_m = k) \mathcal{P}_r(X_m = k | X_0 = i) \quad (\text{A.1})$$

In other words, the conditional probability that the Markov process transits from state  $i$  to state  $j$  in  $m + n$  steps is equal to the sum of the product of the conditional probabilities of reaching an intermediate state  $k$  from  $i$  in  $m$  steps and from state  $k$  to  $j$  in  $n$  steps. Equation (A.1) is also called Chapman Kolmogorov Equation. For the complete sake, the proof is given in appendix A

**Proof** The left hand side can be written (introducing that the Markov process passes through an intermediate state  $K$ ) as:

$$\begin{aligned}
\mathcal{P}_r(X_{m+n} = j | X_0 = i) &= \mathcal{P}_r(U_{k \in S}(X_{m+n} = j, X_m = k) | X_0 = i) \\
&= \mathcal{P}_r(U_{k \in S}(X_{m+n} = j, X_m = k | X_0 = i)) \\
&\text{(conditional probability distribution over addition)} \\
&= \sum_{k \in S} \mathcal{P}_r(X_{m+n} = j, X_m = k | X_0 = i) \\
&\text{(countable additivity)} \\
&= \frac{\sum_{k \in S} \mathcal{P}_r(X_{m+n} = j, X_m = k, X_0 = i)}{\mathcal{P}_r(X_0 = i)} \\
&\text{(conditional probability definition: Wilkinson Def 3.4)} \\
&= \frac{\sum_{k \in S} \mathcal{P}_r(X_{m+n} = j, X_m = k, X_0 = i)}{\mathcal{P}_r(X_0 = i)} \times \frac{\mathcal{P}(X_m = k, X_0 = i)}{\mathcal{P}(X_m = k, X_0 = i)} \\
&= \frac{\sum_{k \in S} \mathcal{P}_r(X_{m+n} = j, X_m = k, X_0 = i)}{\mathcal{P}_r(X_0 = i)} \times \frac{\mathcal{P}(X_m = k, X_0 = i)}{\mathcal{P}(X_m = k, X_0 = i)} \\
&= \frac{\sum_{k \in S} \mathcal{P}_r(X_{m+n} = j, X_m = k, X_0 = i)}{\mathcal{P}_r(X_m = k, X_0 = i)} \times \frac{\mathcal{P}(X_m = k, X_0 = i)}{\mathcal{P}(X_0 = i)} \\
&\text{(using conditional probability definition: Wilkinson Def 3.4)} \\
&= \sum_{k \in S} \mathcal{P}_r(X_{m+n} = j | X_m = k, X_0 = i) \times \mathcal{P}_r(X_m = k | X_0 = i) \\
&\text{(using the definition of Markov processes)} \\
&= \sum_{k \in S} \mathcal{P}_r(X_{m+n} = j | X_m = k) \mathcal{P}_r(X_m = k | X_0 = i) \\
&= \text{R.H.S}
\end{aligned}$$

### Propositions used in SSA

So far we know that the number of reactions occuring in a short time interval is approximately Poisson (large number of reactions (i.e.  $n$  is large) and probability of each reaction is very small (i.e.  $p$  is very small)). The number of reactions in different intervals are inde-

pendent of one another as well.

**Proposition** Consider a Poisson process with rate  $\lambda$ . Let  $T$  be the time to the first event (after zero). Then

$$T \sim \text{Exp}(\lambda)$$

**Proof** Let  $N_t$  be the number of events in the interval  $(0, t]$  for given fixed  $t > 0$ . By definition of Poisson

$$N_t \sim P_o(\lambda t)$$

Consider the cumulative distribution function (CDF) of  $T$ ,

$$\begin{aligned} F_T(t) &= P(T \leq t) \\ &= 1 - P(T > t) \\ &= 1 - P(N_t = 0) \\ &= 1 - \frac{(\lambda t)^0 e^{-\lambda t}}{0!} \\ &= 1 - e^{-\lambda t} \end{aligned}$$

This is the distribution function of an  $\text{Exp}(\lambda)$  random quantity.

$$T \sim \text{Exp}(\lambda) \tag{A.2}$$

So the time to first event of a Poisson process is an exponential random variable. But then using the independence properties of the Poisson process, it should be reasonably clear that the time between any two such events has the same exponential distribution. Thus the times between events of the Poisson process are exponential.

Using equation (A.2) we have

$$X_i \sim \text{Exp}(a_i)$$

where  $X_i$  for  $i = 1, 2, 3, \dots, n$  are independent variables.

Here we use another proposition which states that

**Proposition** If  $X_i \sim \text{Exp}(a_i)$ ,  $i = 1, 2, \dots, n$  are independent random variables, then

$$X_0 \equiv \min_i \{X_i\} \sim \text{Exp}(a_{\text{sum}}(X)) \quad (\text{A.3})$$

where

$$a_{\text{sum}}(X) = \sum_{i=1}^n a_i(X)$$

**Proof :**

We know that  $X \sim \text{Exp}(a)$  so we have  $P(X > x) = e^{-\lambda x}$ . Then

$$\begin{aligned} P(X_0 > x) &= P(\min_i X_i > x) \\ &= P([X_1 > x] \cap [X_2 > x] \cap \dots \cap [X_n > x]) \\ &= \prod_{i=1}^n P(X_i > x) \\ &= \prod_{i=1}^n e^{-a_i x} \\ &= e^{-x \sum_{i=1}^n a_i} \\ &= e^{-a_{\text{sum}} x} \end{aligned}$$

So

$$P(X_0 \leq x) = 1 - e^{-a_{\text{sum}} x} \quad (\text{A.4})$$

and hence

$$X_0 \sim \text{Exp}(a_{\text{sum}}(X)) \quad (\text{A.5})$$

**Lemma** Suppose that  $X_1 \sim \text{Exp}(a_i)$  and  $X_2 \sim \text{Exp}(a_j)$  are independent random vari-

ables. so

$$\begin{aligned}
 P(X_1 < X_2) &= \int_0^\infty P(X_1 < X_2 | X_2 = y) f(y) dy \\
 &= \int_0^\infty P(X_1 < y) f(y) dy \\
 &= \int_0^\infty (1 - e^{-a_i y}) a_j e^{-a_j y} dy \\
 &= \frac{a_i}{a_i + a_j} \\
 &= \frac{a_i}{a_{sum}}
 \end{aligned} \tag{A.6}$$

so we can say

$$P(X_1 < X_2) = \frac{a_i(X)}{a_{sum}(X)} \tag{A.7}$$

Using Lemma 1, we can make the following proposition

**Proposition** If  $X_i \sim \text{Exp}(a_i)$ , where  $i = 1, 2, 3, \dots, n$  are independent random variables, let  $j$  be the index of the smallest of the  $X_i$ . then  $j$  is a discrete random variable with PMF

$$\pi_i = \frac{a_i}{a_{sum}}, \quad i = 1, 2, \dots, n \tag{A.8}$$

where

$$a_{sum} = \sum_{i=1}^n a_i$$

**Proof :**

$$\pi_j = P(X_j < \min_{i \neq j} \{X_i\}) \tag{A.9}$$

where

$$Y = \min_{i \neq j} \{X_i\}$$

so that

$$Y \sim \text{Exp}(a_{-j})$$

where

$$a_{-j} = \sum_{i \neq j} a_i$$

Using equation (A.7), we can write

$$\begin{aligned} &= \frac{a_j}{a_j + a_{-j}} \\ &= \frac{a_j}{a_{sum}} \\ \pi_j &= \frac{a_j}{a_{sum}} \end{aligned} \tag{A.10}$$

This gives the likelihood of a particular exponentail random quantity of an independent collection being the smallest. By using (A.5)(A.7) and (A.10), (??) can be written in the standard form as:

$$p(\tau, j | \mathbf{x}, t) = \left( \frac{a_j(\mathbf{x})}{a_{sum}(\mathbf{x})} \right) a_{sum}(\mathbf{x}) e^{-a_{sum}(\mathbf{x})\tau} \tag{A.11}$$

# References

- [1] O'Connor, C.M and Adams, J.U., *Essentials of Cell Biology*, 2010: Cambridge, MA, NPG Education.
- [2] Wilkinson, D.J, *Stochastic Modeling for Systems Biology*, 2006: Chapman & Hall/CRC.
- [3] Berg J., Tymoczko J., and Stryer L., *Biochemistry*, 2006: W.H. Freeman New York.
- [4] Nelson, D., and Cox, M., *Lehninger Principles of Biochemistry*, 2004: W.H. Freeman New York.
- [5] Gillespie, D.T., *Stochastic Simulation of Chemical Kinetics*, Annu. Rev. Phys. Chem. 2007;58:35-55.
- [6] Wilkinson, D.J., *Stochastic Modelling for Quantitative Description of Heterogeneous Biological Systems*, Nature Review Genetics, 2009, 10: p 122-133.
- [7] Gillespie, D.T., Petzold, L.R., *Numerical Simulation for Biochemical Kinetics*, in System Modelling in Cell Biology, From Concepts to Nuts and Bolts, MIT Press, 2006.
- [8] Ilie, S., Enright W.H., Kenneth R.J., *Numerical Solution of Stochastic Models of Biochemical Kinetics*, Canadian Applied Mathematics Quarterly, 2009: 17(3):523-554.
- [9] Cao, Y., Petzold, L.R., *Slow Scale Tau-Leaping Method*, Computer Methods in Applied Mechanics and Engineering, 2008, 197:3472-3479.



- 
- [10] Higham, D.J., *An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations*, Society for Industrial and Applied Mathematics, 2001: 43:525-546.
- [11] Gillespie, D.T., *Approximate Accelerated Stochastic Simulation of Chemically Reacting Systems*, J. Chem Phys, 2001: 115:1716-1733.
- [12] Yang, Y., Rathinam, M., *Tau-leaping of Stiff Stochastic Chemical Systems via Local Central Limit Approximation*, J. Computational Physics, 2013: 242:p581-606.
- [13] Gillespie, D.T., Hellander, A., Petzold, L.R., *Perspective: Stochastic Algorithms for Chemical Kinetics*, J. Chem Phys, 2013: 138:170901 doi: 10.1063/1.4801941.
- [14] Higham, D.J., *Modeling and Simulating Chemical Reactions*, Society for Industrial and Applied Mathematics Rev, 2008: 50(2):347-368.
- [15] Rathinam, M., Petzold, L.R., Cao, Y., Gillespie, D.T., *Stiffness in Stochastic Chemically Reacting Systems: The implicit Tau-leaping Method*, J. Chem Phys, 2003: 119:12784-12794.
- [16] Melykuti, B., Burrage, K., Zygalakis, K.C., *Fast Stochastic Simulation of Biochemical Reaction Systems by Alternative Formulations of the Chemical Langevin Equation*, J. Chem Phys, 2010: 132:164109.
- [17] Gunawan, R., Cao, Y., Petzold, L.R., Doyle, F.J., *Sensitivity Analysis of Discrete Stochastic Systems*, [?] Biophysical, 2005: 88:2530-2540.
- [18] Cao, Y., Gillespie, D.T., Petzold, L.R., *Efficient Stepsize Selection for the Tau-leaping Simulation Method*, J. Chem Phys, 2006: 124:044109.
- [19] Cao, Y., Gillespie, D.T., Petzold, L.R., *Adaptive Explicit-implicit Tau-leaping Method with Automatic Tau Selection*, J. Chem Phys, 2007: 126:224101.

- 
- [20] Gillespie, D.T., Petzold, L.R., *Improved Leap-size Selection for Accelerated Stochastic Simulation*, J. Chem Phys, 2003, 119, 8229.
- [21] Cao, Y., Petzold, L.R., *Trapezoidal Tau-Leaping Formula for the Stochastic Simulation of Biochemical Systems*, Proc. of Foundations of Systems Biology in Engineering, 2005: p 149-152.
- [22] Spudich, J.L., & Koshland Jr, D.E., *Non-genetic Individuality: Chance in the Single Cell*, Nature, 1976: 262:467-471.
- [23] Elowitz, M.B., Leibler, S., *A Synthetic Oscillatory Network of Transcriptional Regulators*, Nature, 2000: 403(6767):335-338.
- [24] Chen, W.w., Niepel, M., and Sorger, P.K., *Classic and Contemporary Approaches to Modelling Biochemical Reactions*, Genes Dev, 2010: 24:1861-1875.
- [25] Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S., *Stochastic Gene Expression in a Single Cell*, Science, 2002, 297:1183-1186.
- [26] Swain, P.S., Elowitz, M.B., & Siggia, E.D., *Intrinsic and Extrinsic Contributions to Stochasticity in Gene Expression*, Proceedings of the National Academy of Sciences, 2002, 99, p:2795-12800.
- [27] Turner, T.E., Schnell, S., Burrage, K., *Stochastic Approaches for Modelling in Vivo Reactions*, Computational Biology and Chemistry, 2004, 28, p:165-178.
- [28] Arkin, A.P., Ross, J., McAdams, H.H., *Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage  $\lambda$ -Infected Escherichia Coli Cells*, Genetics, 1998, 149, p:1633-48.
- [29] McAdams, H.H., Arkin, A.P., *Stochastic Mechanisms in Gene Expression*, Proceedings of the National Academy of Sciences, 1997, 94:814-819.

- [30] Fedoroff, N., Fontana, W., *Small Numbers of Big Molecules*, Science, 2002, p: 1129-1131.
- [31] McAdams, H.H., Arkin, A., *It's a Noisy Business! Genetic Regulation at the Nanomolar Scale*, Trends Genet., 1999, 15(2):65-69.
- [32] Szkely, Jr. T., Burrage, K., *Stochastic Simulation in Systems Biology*, Computational and Structural Biotechnology, 2014, 12:14-25.
- [33] Voit, E., Martens, H., Omholt, S.W., *150 Years of the Mass Action Law*, PLOS Comput. Biol., 2015, 11(1)e1004012.
- [34] Zambelli, S., *Chemical Kinetics, an Introduction* in Chemical Kinetics Vivek Patel (Ed), 2012, Intech Open. doi: 10.5772/37081.
- [35] McQuarrie, *Stochastic Approach to Chemical Kinetics* J. Appl. Prob., 1967, 4(3):413-478.
- [36] Gillespie, D.T., *A Rigorous Derivation of the Chemical Master Equation*, Physics A, 1992, 188:404-425.
- [37] Oppenheim, I., Shuler, K.E., Weiss, G.H., *Stochastic and Deterministic Formulation of Chemical Rate Equation*, J. Chem Phys, 1969, 50(1):460-466.
- [38] Kurtz, T.G., *The Relationship Between Stochastic and Deterministic Models for Chemical Reactions*, J. Chem Phys, 1972, 57(7):2976-2978.
- [39] Gillespie, D.T., *A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions*, Comput. Phys., 1976, 22:403-434.
- [40] Gillespie, D.T., *Exact Stochastic Simulation of Coupled Chemical Reactions*, J. Chem Phys, 1977, 81:2340-2361.
- [41] Gillespie, D.T., *The Chemical Langevin Equation*, J. Chem Phys, 2000, 113:297-306

- 
- [42] Kitano, H. *Computational Systems Biology*, Nature, 2002, 420:206-210.
- [43] Raser, J.M., & O'Shea, E.K., *Noise in Gene Expression: Origins, Consequences, and Control*, Science, 2005, 309:2010-2013.
- [44] Gillespie, D.T., *Markov Processes: an Introduction for Physical Scientists*, Academia, New York, 1992.
- [45] Li, H., Cao, Y., Petzold, L.R. and Gillespie, D.T., *Algorithms and Software for Stochastic Simulation of Biochemical Reacting Systems*, Biotechnol. Prog., 2007, 24:56-61.
- [46] Paulsson, J., Berg, O., & Ehrenberg, M., *Stochastic Focusing: Fluctuation-enhanced Sensitivity of Intracellular Regulation*, Proc. Nati. Acad. Sci. USA, 2000, 97:7148-7153.
- [47] Gibson, M.A. & Bruck, J., *Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels*, J. Chem Phys A., 2000: 104:1876-1889.
- [48] Kiehl, T.R., Mattheyses, R.M., & Simmons, M.K., *Hybrid Simulation of Cellular Behavior*, Bioinformatics, 2004: 20:316-322.
- [49] Alfonsi, A., Cances, E., Turinici, G., dl Ventura, B. & Huisinga, W., *Adaptive Simulation of Hybrid Stochastic and Deterministic Models for Biochemical Systems*, ESAIM: Proc., 2005: 14:1-13.
- [50] Puchalka, J. & Kierzek, A.M., *Bridging the Gap Between Stochastic and Deterministic Regimes in the Kinetic Simulations of the Biochemical Reaction Networks*, J. Biophys., 2004: 86:1357-1372.
- [51] Rao, C.V., & Arkin, A., *Stochastic Chemical Kinetics and the Quasi-steady-state Assumption: Application to the Gillespie Algorithm*, J. Chem Phys, 2003, 118:4999-5010.

- [52] Haseltine, E.L. & Rawlings, J.B., *Approximate Simulation of a System of Coupled Fast and Slow Reactions for Stochastic Chemical Kinetics*, J. Chem Phys, 2002: 117:6959-6969.
- [53] Salis, H. & Kaznessis, Y., *Accurate Hybrid stochastic Simulation of a System of Coupled Chemical or Biochemical Reactions*, J. Chem Phys, 2005: 122:054103.
- [54] Cao, Y., Gillespie, D.T. & Petzold, L.R., *Multiscale Stochastic Simulation Algorithm with Stochastic Partial Equilibrium Assumption for Chemically Reacting Systems*, J. Comp. Phys., 2005, 206:305-411.
- [55] Samant, A. & Vlachos, D.G., *Overcoming Stiffness in Stochastic Simulation Stemming from Partial Equilibrium: a Multiscale Monte Carlo Algorithm*, J. Chem Phys, 2005: 123:144114.
- [56] Weinan, E., Liu, D., & Vanden-Eijnden, E., *Nested Stochastic Simulation Algorithm for Chemical Kinetic Systems with Disparate Rates*, J. Chem Phys, 2005: 123:194107.
- [57] Weinan, E., Liu, D., & Vanden-Eijnden, E., *Nested Stochastic Simulation Algorithms for Chemical Kinetic Systems with Multiple Time Scales* J. Chem Phys, 2007: 221:158-180.
- [58] Hastings, W.K., *Monte Carlo Sampling Methods Using Markov Chains and their Applications* Biometrika, 1970, 57:97-109.
- [59] Haseltine, E.L. & Rawlings, J.B., *On the Origins of Approximations for Stochastic Chemical Kinetics*, J. Chem Phys, 2005: 123:144917.
- [60] Salis, H. & Kaznessis, Y., *Equation-free Probabilistic Steady-state Approximation: Dynamic Application to the Stochastic Simulation of Biochemical Reaction Networks*, J. Chem Phys, 2005: 123:214106.

- [61] Cao Y., Gillespie, D.T., & Petzold L.R., *The Slow-scale Stochastic Simulation Algorithms*, J. Chem Phys, 2005: 122:014116.
- [62] Cao Y., Gillespie, D.T., & Petzold L.R., *Avoiding Negative Populations in Explicit Poisson Tau-leaping*, J. Chem Phys, 2005: 123:054104.
- [63] Tian, T., & Burrage, K., *Binomial Leap Methods for Simulating Stochastic Chemical Kinetics*, J. Chem Phys, 2004, 121:10356-10364.
- [64] Gillespie, D.T., *Stochastic Chemical Kinetics*, Dordrecht: Springer , 2005, In Handbook of Materials Modeling, ed. S Yip, pp. 1735-1752.
- [65] Gillespie, D.T., *The Multivariate Langevin and Fokker-Planck Equations*, Am. J. Phys., 1996, 64:1246-1257.
- [66] Cao Y., Li, H., & Petzold L.R., *Efficient Formulation of the Stochastic Simulation Algorithm for Chemically Reacting Systems*, J. Chem Phys, 2004: 121:4059-4067.
- [67] Samoilov, M., Plyasunov, S. & Arkin, A.P., *Stochastic Amplification and Signaling in Enzymatic Futile Cycles through Noise Induced Bistability with Oscillations*, Proc. Natl. Acad. Sci USA, 2005, 102:2310-2315.
- [68] Raser, J.M. & O'Shea, E.K., *Control of Stochasticity in Eukaryotic Gene Expression*, Science, 2004: 304:1811-1814.
- [69] Blake, W.J., Karen, M., Cantor, C.R., & Collins, J.J., *Noise in Eukaryotic Gene Expression*, Nature, 2003, 422:633-637.
- [70] Bradie, B., *A Friendly Introduction to Numerical Analysis*, Pearson Prentice Hall, NJ, 2006.
- [71] Ascher, U., Petzold, L.R., *Computer Methods for Ordinary Differential Equations and Differential-algebraic Equations*, SIAM, 1998.

- [72] Schlögl, F., *Chemical Reaction Models for Non-equilibrium Phase Transitions*, Z. Physik. 1972, 253:147-161.
- [73] Matheson, I., Walls, D.F., Gardiner, W., *Stochastic Models of First-order Nonequilibrium Phase Transitions in Chemical Reaction*, J. Statistical Physics, 1975, 12(1):21-34.
- [74] Rathinam, M., Petzold, L., Cao, Y., Gillespie, D.T., *Consistency and Stability of Tau-leaping Schemes for Chemical Reaction Systems*, SIAM Multi. Model. Simul., 2005, 4(3):867-895.
- [75] Goldbeter, A., Koshland, D.E., *An Amplified Sensitivity Arising from Covalent Modification in Biological Systems*, Proc. Natl. Acad. Sci. USA, 1981, 87(11): 6840-6844.
- [76] Ahn, T., Cao, Y., Watson, L.T., *Stochastic Simulation Algorithms for Chemical Reactions*, Int. Conf. Bioinfo. Comput. Biol. BIOCOMP, 2008, Las Vegas, USA.
- [77] Gardner, T.S., Cantor, C.R., Collins, J.J., *Construction of a Genetic Toggle Switch in Escherichia Coli*, Nature, 2000, 403(6767):339-342.
- [78] Yang, Y., Rathinam, M., *Tau-leaping of Stiff Stochastic Chemical Systems via Local Central Limit Approximation*, J. Comput. Phys., 2013, 242:581-606.