Computational Methods for Parametric Sensitivities of Stochastic Chemical Reaction Networks

By

Elizabeth Skubak Wolf

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Mathematics)

at the

UNIVERSITY OF WISCONSIN – MADISON

2014

Date of final oral examination: May 13, 2014

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

David F. Anderson, Assistant Professor, Mathematics

James B. Rawlings, Professor, Chemical and Biological Engineering

Benedek Valko, Associate Professor, Mathematics

Bret Hanlon, Assistant Professor, Statistics

Saverio Spagnolie, Assistant Professor, Mathematics

Abstract

Stochastic dynamical system models are increasingly being used to help understand the behavior of biochemical networks at the cellular level. In this thesis, we study the most common such models: continuous time Markov chains (CTMCs). In particular, we focus on computational methods for estimating parametric sensitivities of biochemical CTMC models. Sensitivities, which are derivatives of model output quantities with respect to model parameters, are a useful analytical tool. One important application is in the setting of parameter estimation, in which sensitivities can be utilized to significantly improve the efficiency of parameter-space optimization algorithms.

The main contribution of this dissertation is the introduction of a hybrid pathwise method for the estimation of parametric sensitivities. The new hybrid method combines elements from the three main types of procedures for sensitivity estimation, and has a number of desirable qualities not simultaneously possessed by any other such method. First, it is unbiased. Second, it is applicable to nearly any physically relevant biochemical CTMC model. Third, and as we demonstrate on several numerical examples, it is often the most efficient method for the problem at hand, particularly if one wishes to estimate the full gradient of parametric sensitivities.

While the main contribution of this thesis is the introduction of the new hybrid method, advances are also provided for the efficient computation of second order sensitivities. In particular, we present an efficient second order finite difference method, which has already appeared in the literature in [42].

Acknowledgements

I would foremost like to thank my advisor, David F. Anderson, for being so helpful and encouraging. I would also like to thank Thomas G. Kurtz for always being willing to share his expertise; Benedek Valko and Timo Seppalainen for teaching their probability classes so well; and Bret Hanlon and James Rawlings for many helpful discussions. Finally, I would like to thank Chris Wolf, for being so understanding and supportive.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Deterministic ordinary differential equation (ODE) models have historically been the primary choice for modeling biochemical systems. These models, however, often provide a poor representation of system dynamics, particularly when molecular abundances are small. Over the past two decades, the availability of biological data has grown dramatically, and increasingly this data has shown disagreement from the predictions of ODE models. Most notably, in the mid-1990s, molecular biologists began developing fluorescent protein reporters, making it possible to quantitatively measure molecular abundances, even in systems occurring within a single cell. For example, using these techniques, Raj et al. show in [30] that "gene expression in mammalian cells is subject to large, intrinsically random fluctuations." Such variations, which have also been found in yeast and bacteria, are often further amplified by feedback mechanisms, and can be crucial to the overall fate of the cell [9]. Furthermore, and in stark contrast to ODE models, stochastic models can admit the possibility of extinction, for example, or allow for switching in a bistable genetic system [41]. Therefore, while stochastic models of dynamical systems are not a new addition to the field of biology, they are becoming an increasingly popular tool for understanding biochemical systems.

The most common choice of stochastic model in the biochemical setting is a continuous-time Markov chain (CTMC). In the biology literature, a CTMC model is usually described by its Kolmogorov forward equation (1.4), commonly called the chemical master equation, whereas in this work we primarily focus on the stochastic equation (1.2). It is often difficult or impossible to solve explicitly for the distribution of a CTMC, particularly when the state space is infinite. On the other hand, CTMCs are straightforward to simulate, using either the stochastic simulation algorithm (SSA) [14] or the next reaction method (NRM) [1, 12]. Therefore, Monte Carlo methods are widely used to study these CTMCs, and underlie the computational methods presented in this thesis.

The models under consideration typically include a number of parameters, and it is often important to understand the role of these parameters in model dynamics. Given some CTMC $X_t(\theta)$ depending on a vector of parameters θ, suppose that f is some output of the system, which we also allow to depend on θ. For example, one can take $f(\theta, X_{t\in[0,T](\theta)}) = X_{i,T}(\theta)$, which gives the abundance of the i^{th} system species at time T. More generally, one may be interested in some path functional $f(\theta, X_{t\in[0,T](\theta)}) = \int_a^b h(\theta, X_s(\theta))ds$. In this thesis, we study computational methods for estimating the parametric sensitivities

$$\frac{\partial}{\partial\theta_i}\mathbb{E}[f(\theta, X_{t\in[0,T]}(\theta))] \quad \text{and/or} \quad \frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathbb{E}[f(\theta, X_{t\in[0,T]}(\theta))],$$

which quantitatively describe how perturbations in the parameter θ affect the system response function of interest. See Section 1.1 for more details pertaining to the models considered in this thesis.

Sensitivity analysis has a long history in the fields of discrete event systems and

stochastic control, where sensitivities can be useful in choosing parameters for optimal system performance [6, 15, 40]. Sensitivities are also important in financial mathematics, where they are known as "the Greeks," and are useful, for instance, for evaluating hedging strategies [13]. For surveys on sensitivities in these fields, see for example [7, 11].

Sensitivity analysis is a valuable tool in the biochemical setting, and there has therefore been an increasing focus on developing methods tailored to the models that arise from biochemical systems. See, for example, [2, 19, 26, 32, 37, 39]. A biologist may be interested in the first order sensitivity describing how the expected abundance of a certain protein will be changed if, say, the rate of the transcription of mRNA decreases. Second order sensitivities provide concavity information, which is useful for finding roots or extrema of means of a given system output. Moreover, in many biological models, the only means of determining appropriate rate parameters is through experimentation. If the model provides a reasonable approximation of the system, sensitivities can be used to help in this estimation. First, they can be used to determine whether a parameter is identifiable from available data [22]. For example, if only protein abundances can be measured, but the sensitivity of protein abundance to a parameter of interest is zero or near-zero, the data will be unable to isolate a single appropriate parameter. Furthermore, to estimate identifiable parameters from data, first and second order sensitivities can be combined to significantly increase the efficiency of a parameter space optimization algorithm [33].

In the remainder of this chapter, we review some necessary background material. In Chapter 2, we describe the three main types of existing computational methods for approximating parameter sensitivities: finite difference methods, the likelihood ratio

method, and pathwise methods. In Chapter 3 we introduce a hybrid method for first order sensitivity estimation, the main work of this thesis. This new method is a hybrid in the sense that it incorporates components of the three different sensitivity methods detailed in Chapter 2. The hybrid method is efficient and unbiased, and unlike many existing methods involving pathwise differentiation, it is widely applicable to CTMCs arising from biochemical systems.

In Chapter 4, we turn our attention to the estimation of second order sensitivities. In particular, we provide an extension of the coupling method found in [2] that is applicable to second order differences, and describe and analyze the resulting method for computing second order sensitivities. Much of the work in Chapter 4 was published in [42]. Finally, we give some conclusions and avenues for future work in Chapter 5.

1.1 Mathematical Model

We say that a process R is a counting process if $R(0) = 0$ and R is constant except for jumps of $+1$. The CTMC models we study are $\mathbb{Z}^d$-valued processes X satisfying the representation

$$X_t = X_0 + \sum_{k=1}^{K} R_k(t)\zeta_k \,, \tag{1.1}$$

where X_0 is the initial state, the R_k are counting processes, $K < \infty$, and $\zeta_k \in \mathbb{Z}^d$. Note that $R_k(t)$ gives the number of jumps of X in the direction ζ_k by time t. Using the random time change representation of Kurtz [25], the counting processes $\{R_k\}$ can be represented via

$$R_k(t) = Y_k\left(\int_0^t \lambda_k(X_s)ds\right),$$

where the Y_k are independent unit-rate Poisson processes and where the $\lambda_k : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ are intensity functions. Thus one representation for X is the solution to the stochastic equation

$$X_t = X_0 + \sum_{k=1}^{K} Y_k \left(\int_0^t \lambda_k(X_s) ds \right) \zeta_k. \tag{1.2}$$

Throughout this thesis, we assume that X is non-explosive, i.e. that with probability one there are only finitely many transitions within a given finite time. Specifically, letting

$$\tau_n = \inf_{t \geq 0} \left\{ \sum_{k=1}^{K} R_k(t) = n \right\},$$

we assume that $\mathbb{P}\left(\lim_{n\to\infty} \tau_n > t\right) = 1$ for each $t > 0$. For more information on this representation, see [3, 4].

Most lattice-valued systems can be represented via such counting processes, each giving the number of jumps in one of finitely many specified directions. In particular, this representation is also useful in the study and control of systems in fields such as queueing theory and the study of population processes. As biochemical reaction networks are the main motivation for this thesis, we use biochemical terminology and examples throughout.

In the biochemical setting, d is the number of chemical species in a system of interest, and the i^{th} coordinate $X_{i,t}$ of $X_t \in \mathbb{Z}^d_{\geq 0}$ denotes the abundance of the i^{th} species at time t. The system changes due to K chemical reactions, each occurring with a rate λ_k, known as an intensity function, or in the biology literature as a propensity function. The reaction vector ζ_k represents the total change to system molecular abundances due to the occurrence of the k^{th} reaction. We may decompose these reaction vectors as $\zeta_k = \nu'_k - \nu_k$, where the product vector ν'_k gives the abundances of molecules produced in

the k^{th} reaction, and the source vector ν_k gives the abundances of molecules consumed. The number of molecules of the i^{th} species required for reaction k is therefore ν_{ki}, so, for the most part, we only consider intensities λ_k for which $\lambda_k(x) = 0$ if $x_i < \nu_{ki}$ for some i. Thus if $X_0 \in \mathbb{Z}^d_{\geq 0}$, such intensities then imply that the state space $\mathcal{S}$ of X is some subset of $\mathbb{Z}^d_{\geq 0}$. That is, the quantities of the constituent species are always nonnegative, and are given as integer abundances rather than concentrations. We will later relax these conditions on the intensities for a process Z introduced in Chapter 3, and allow the states of Z to have negative coordinates.

We give two example biochemical models, though we postpone the discussion of the intensities until Section 1.1.1.

Example 1.1. *Consider a model of gene transcription and translation, in which mRNA (M) is being transcribed and then translated into protein (P), and with both the mRNA and the protein undergoing degradation. Given in a common reaction network representation,*

$$\emptyset \rightarrow M, \quad M \rightarrow M + P, \quad M \rightarrow \emptyset, \quad P \rightarrow \emptyset. \tag{1.3}$$

The first reaction represents the transcription of a strand of mRNA, and the second represents the translation of an mRNA strand into a protein. The third and fourth reactions represent the degradation of mRNA and protein respectively. The vector-valued process

$$X_t = \begin{bmatrix} X_{M,t} \\ X_{P,t} \end{bmatrix} \in \mathbb{Z}^2_{\geq 0}$$

records the quantities of mRNA and protein respectively at time t. Ordering the four

reactions of (1.3) *from left to right, we have the source vectors*

$$\nu_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \nu_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \nu_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \nu_4 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and the product vectors

$$\nu_1' = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \nu_2' = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \nu_3' = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \nu_4' = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Therefore

$$\zeta_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \zeta_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \zeta_3 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \zeta_4 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

◇

Example 1.2. *Consider some molecule A that binds with itself to form a molecule B, and suppose that A can also bind with B to form some molecule C:*

$$A + A \rightarrow B, \quad A + B \rightarrow C.$$

Note that $A + A$ is often written as $2A$. Then the state space lies within $\mathbb{Z}^3_{\geq 0}$ and

$$\nu_1 = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \nu_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \nu_1' = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \nu_2' = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \zeta_1 = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}, \zeta_2 = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}.$$

◇

The CTMC satisfying the stochastic equation (1.2) is often described by its Kolmogorov forward equation (or chemical master equation),

$$\frac{d}{dt}P_{X_0}(x,t) = \sum_{k=1}^{K} P_{X_0}(x-\zeta_k,t)\lambda_k(x-\zeta_k)1_{\{x-\zeta_k\in\mathbb{Z}^d_{\geq 0}\}} - P_{X_0}(x,t)\sum_{k=1}^{K}\lambda_k(x), \tag{1.4}$$

where $P_{x_0}(x,t) = P(X_t = x|X_0 = x_0)$ is the probability of being in state $x \in \mathbb{Z}^d_{\geq 0}$ at time $t \geq 0$ given an initial condition of x_0.

Additionally, the CTMC satisfying the stochastic equation (1.2) can be described by its generator, which is the operator $\mathcal{A}$ defined via

$$(\mathcal{A}f)(x) = \sum_k \lambda_k(x)(f(x+\zeta_k) - f(x)) \tag{1.5}$$

for bounded $f : \mathbb{R}^d \to \mathbb{R}$. We remind the reader of Dynkin's formula, which holds for more general functions f:

$$\begin{aligned}\mathbb{E}[f(X_t)] &= \mathbb{E}[f(X_0)] + \mathbb{E}\left[\int_0^t (\mathcal{A}f)(X_s)ds\right] \\ &= \mathbb{E}[f(X_0)] + \mathbb{E}\left[\int_0^t \sum_{k=1}^{K}\lambda_k(X_s)[f(X_s+\zeta_k) - f(X_s)]ds\right].\end{aligned} \tag{1.6}$$

For example, for processes X that satisfy the growth condition we give in Section 3.4, which is nearly all biologically relevant processes, the equation (1.6) holds for functions f that grow at most polynomially. See [3, 10, 19].

1.1.1 Parameters and Sensitivities

We now suppose that the intensities are dependent on some vector of parameters $\theta \in \mathbb{R}^R$. The number of parameters R is often equal to the number of reactions K, though since this need not be the case we will distinguish between R and K throughout. We consider the parametrized family of models $X_t(\theta)$ satisfying

$$X_t(\theta) = X_0(\theta) + \sum_{k=1}^{K} Y_k \left(\int_0^t \lambda_k(\theta, X_s(\theta)) ds \right) \zeta_k. \tag{1.7}$$

Throughout this thesis, we will assume that $X_0(\theta) \equiv X_0$ is fixed, though several of the included methods can be extended to allow the initial condition to be random and/or θ-dependent. We also use $\mathcal{A}^\theta$ to represent the generator of $X(\theta)$ as in (1.5).

In the biochemical setting, θ commonly represents the mass-action constants of the K reactions:

Definition 1.3 (Mass-action kinetics)**.** *We say that the intensities are of mass-action form if*

$$\lambda_k(\theta, x) = \theta_k \prod_{i=1}^{d} \nu_{ki}! \binom{x_i}{\nu_{ki}} = \theta_k g_k(x) \tag{1.8}$$

where $\theta_k > 0$ and where the g_k, which notably are functions of x only, are defined by the above equation. Recall that ν_{ki} gives the number of molecules of the i^{th} species used as the source of the k^{th} reaction.

Note that these intensities are therefore proportional to the number of distinct subsets of the molecules in the system that can form the source for the reaction, which implicitly assumes that the system is well-mixed. For example, if the k^{th} reaction requires no source molecules, the propensity is simply given by θ_k; that is, $g_k \equiv 1$. If the reaction

is unary, i.e. requires a single source molecule, the propensity is $\theta_k x_i$, with $g_k(x) = x_i$ for some $i \in \{1, \ldots, d\}$. If it is binary, i.e. requires two source molecules, the propensity takes the form of either $\theta_k x_i x_j$ or $\theta_k x_i(x_i - 1)$ for some $i, j \in \{1, \ldots, d\}$, in which case $g_k(x)$ is either $x_i x_j$ or $x_i(x_i - 1)$ respectively.

We return to our examples from pages 6 and 7, now giving each reaction a mass-action rate parameter. By convention, these parameters are written above the reaction arrows.

Example 1.4. *The mRNA transcription and translation model*

$$\emptyset \xrightarrow{\theta_1} M, \quad M \xrightarrow{\theta_2} M + P, \quad M \xrightarrow{\theta_3} \emptyset, \quad P \xrightarrow{\theta_4} \emptyset$$

under mass-action kinetics has intensities given by

$$\lambda_1(\theta, X) = \theta_1, \quad \lambda_2(\theta, X) = \theta_2 X_M, \quad \lambda_3(\theta, X) = \theta_3 X_M, \quad \textit{and} \quad \lambda_4(\theta, X) = \theta_4 X_P.$$

With the ζ_k as in Example 1.1, the process X is represented as in (1.7) *by*

$$X_t(\theta) = X_0(\theta) + Y_1\big(\theta_1 t\big)\zeta_1 + Y_2\left(\int_0^t \theta_2 X_{M,s}\, ds\right)\zeta_2 + Y_3\left(\int_0^t \theta_3 X_{M,s}\, ds\right)\zeta_3 + Y_4\left(\int_0^t \theta_4 X_{P,s}\, ds\right)\zeta_4.$$

◇

Example 1.5. *The model*

$$A + A \xrightarrow{\theta_1} B, \quad A + B \xrightarrow{\theta_2} C$$

has intensities

$$\lambda_1(\theta, X) = \theta_1 X_A(X_A - 1), \quad \textit{and} \quad \lambda_2(\theta, X) = \theta_2 X_A X_B.$$

Then X is given by

$$X_t(\theta) = X_0(\theta) + Y_1\left(\int_0^t \theta_1 X_{A,s}(X_{A,s} - 1)ds\right)\zeta_1 + Y_2\left(\int_0^t \theta_2 X_{A,s} X_{B,s}\, ds\right)\zeta_2.$$

$\diamond$

There are realistic models in which a given parameter affects the rates of more than one reaction. There are also choices of kinetics other than mass-action kinetics. Therefore, we do not assume mass-action kinetics in this thesis unless explicitly stated.

Let f be some measurable function of $X_{t\in[0,T]}$, the path of our process X through some fixed terminal time $T < \infty$. We will also allow f to depend explicitly on the parameter θ. For the results in the thesis, we will require mild regularity conditions on f (see Condition 3.26). For example, we may be interested in

1. $f(\theta, X_{t\in[0,T]}(\theta)) = X_{i,T}(\theta)$, the abundance of the i^{th} species at time T,
2. $f(\theta, X_{t\in[0,T]}(\theta)) = \frac{1}{T}\int_0^T X_{i,s}(\theta)\, ds$, a time-average of the i^{th} abundance,
3. $f(\theta, X_{t\in[0,T]}(\theta)) = \int_a^b h(X_s(\theta))ds$ for some $h : \mathbb{R}^d \to \mathbb{R}$ and some $0 \le a \le b \le T$, or
4. $f(\theta, X_{t\in[0,T]}(\theta)) = \mathbf{1}(X_T(\theta) \in A)$ for some set A.

Define

$$J(\theta) := \mathbb{E}[f(\theta, X_{t\in[0,T]}(\theta))]. \tag{1.9}$$

We are interested in estimating derivatives of J with respect to the components of θ. That is, for $i, j \in \{1, \ldots, R\}$ we wish to estimate

$$\frac{\partial}{\partial \theta_i} J(\theta) \quad \text{and/or} \quad \frac{\partial^2}{\partial \theta_i \partial \theta_j} J(\theta)$$

at some fixed value θ_0 of interest (we will generally write θ rather than θ_0 if the context is clear).

Example 1.6. *Consider the simple decay model $S \xrightarrow{\theta} \emptyset$ with X_t denoting the abundance of S at time t and with a fixed initial abundance X_0. Then the single reaction vector is $\zeta = -1$. We take a mass-action intensity of $\lambda(\theta, x) = \theta x$. Using Dynkin's formula* (1.6) *with $f(x) = x$, we have that*

$$\mathbb{E}[X_t(\theta)] = \mathbb{E}[X_0] + \mathbb{E}\left[\int_0^t \theta X_s[(X_s - 1) - X_s]ds\right] = X_0 - \int_0^t \mathbb{E}[\theta X_s]ds.$$

Solving, we see that

$$\mathbb{E}[X_t(\theta)] = X_0 e^{-\theta t} \quad \textit{and so} \quad \frac{d}{d\theta}\mathbb{E}[X_t(\theta)] = -X_0 t e^{-\theta t}.$$

$\diamond$

For models with nonlinear intensities, on the other hand, parametric sensitivities generally must be estimated. This is often done by utilizing simulation and Monte Carlo estimation, which we review in Section 1.2.

1.1.2 The Standard Coupling

Coupling is a proof technique that allows two random variables to be compared by constructing them on the same probability space. This main idea can also be used in a computational Monte Carlo framework as a way to reduce variance. In nearly every method presented in this thesis, we will couple two CTMCs of the form (1.2). As the specific CTMCs to be coupled will change among our different applications, we present this standard coupling using two general processes X and Z in the random time change representation (1.2). We assume both are d-dimensional processes with the same reaction vectors ζ_k; it is the intensities of the two processes that may differ. Our goal is to reduce the variance of the difference $X - Z$.

The main idea of the coupling is illustrated in the following toy example. Suppose that X and Z are homogeneous Poisson processes with rates 13.1 and 13, respectively. That is, both processes have a single reaction with $\zeta = 1$, but X increases with constant rate 13.1 while Z increases with constant rate 13. If we wish to study the difference between X and Z, we could use independent, unit-rate Poisson processes Y_1 and Y_2 and define

$$X_t = Y_1(13.1t) \quad \text{and} \quad Z_t = Y_2(13t).$$

Then $\mathbb{E}(X_t - Z_t) = 0.1t$ and

$$\text{Var}(X_t - Z_t) = \text{Var}(Y_1(13.1t)) + \text{Var}(Y_2(13t)) = 26.1t.$$

We would like to lower this variance, so instead, we define

$$X_t = Y_1(13t) + Y_2(0.1t) \quad \text{and} \quad Z_t = Y_1(13t).$$

Then we still have $\mathbb{E}(X_t - Z_t) = \mathbb{E}Y_2(0.1t) = 0.1t$ as needed, but now

$$\text{Var}(X_t - Z_t) = \text{Var}(Y_2(0.1t)) = 0.1t \ll 26.1t. \tag{1.10}$$

Note that we took the intensity of Y_1, the shared counting process, to be the minimum of the two original intensities.

In general, we want to consider two processes X and Z given as in (1.1) and (1.2) by

$$X_t = X_0 + \sum_{k=1}^{K} R_k^X(t)\zeta_k \quad \text{and} \quad Z_t = Z_0 + \sum_{k=1}^{K} R_k^Z(t)\zeta_k \,, \tag{1.11}$$

where the counting processes R_k^X have intensities λ_k^X and where the R_k^Z have intensities λ_k^Z. We split the counting processes R_k^X and R_k^Z into sub-processes to be shared among the two CTMCs X and Z. Specifically, for $b_\ell \in \{0, 1\}$ we define the new counting processes $R_{k,[b_1,b_2]} = Y_{k,[b_1,b_2]}\left(\int_0^t \Lambda_{k,[b_1,b_2]}(X_s, Z_s)ds\right)$, where the $Y_{k,[b_1,b_2]}$ are independent unit-rate Poisson processes and where

$$\begin{aligned}
\Lambda_{k,[1,1]}(x, z) &= \lambda_k^X(x) \wedge \lambda_k^Z(z), \\
\Lambda_{k,[1,0]}(x, z) &= \lambda_k^X(x) - \lambda_k^X(x) \wedge \lambda_k^Z(z), \\
\Lambda_{k,[0,1]}(x, z) &= \lambda_k^Z(x) - \lambda_k^X(x) \wedge \lambda_k^Z(z),
\end{aligned} \tag{1.12}$$

where we have used the notation $a \wedge b := \min\{a, b\}$. Note that for each $k = 1, \ldots, K$ we

have

$$R_k^X \stackrel{d}{=} R_{k,[1,1]} + R_{k,[1,0]} \quad \text{and} \quad R_k^Z \stackrel{d}{=} R_{k,[1,1]} + R_{k,[0,1]}$$

so that the processes given by

$$\begin{aligned} X_t &= X_0 + \sum_{k=1}^{K} \left(R_{k,[1,1]} + R_{k,[1,0]}\right) \zeta_k \quad \text{and} \\ Z_t &= Z_0 + \sum_{k=1}^{K} \left(R_{k,[1,1]} + R_{k,[0,1]}\right) \zeta_k, \end{aligned} \tag{1.13}$$

have the same distribution as the processes in (1.11) [3]. The binary numbers simply give a convenient indexing scheme: the process X is constructed using all the counting processes $R_{k,[b_1,b_2]}$ in which $b_1 = 1$. Similarly, Z uses those in which $b_2 = 1$.

Because X and Z share the counting processes $R_{k,[1,1]}$, many of the jumps in the paths of X and Z will occur simultaneously. These shared jumps significantly lower the variance of the difference $X - Z$, in a manner similar to (1.10) above. Indeed, in the difference $X - Z$, the $R_{k,[1,1]}$ are completely canceled. Moreover, because the rate of each $R_{k,[1,1]}$ is the minimum of the two relevant intensities, the processes X and Z share jumps as often as possible: we cannot allow the rates of the $R_{k,[1,1]}$ to be any larger, because otherwise one of the other counting processes $R_{k,[0,1]}$ and $R_{k,[1,0]}$ would have a negative intensity (see (1.12)). These "auxiliary" counting processes $R_{k,[0,1]}$ and $R_{k,[1,0]}$ make up for any remaining intensity. Note that if $\lambda_k^X(x)$ and $\lambda_k^Z(z)$ are close in value, then these remainders are small. In fact, for any pair of states x, z at least one of the two remainder processes for the k^{th} reaction is zero, depending on the value of the minimum.

Finally, we note that this standard coupling constructs a new $2d$-dimensional CTMC

W given by

$$W_t = \begin{bmatrix} X_t \\ Z_t \end{bmatrix},$$

where d is the dimension of the original processes. For each reaction $k = 1, \ldots, K$ of the original process X (or Z), the CTMC W has the three reaction vectors in $\mathbb{Z}^{2d}$

$$\begin{bmatrix} \zeta_k \\ \zeta_k \end{bmatrix}, \begin{bmatrix} \zeta_k \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \zeta_k \end{bmatrix},$$

where each 0 is interpreted as $\vec{0} \in \mathbb{Z}^d$. The intensities at a state $w = (x, z)$ of these three reactions of W are given as in (1.12) by $\Lambda_{k,[1,1]}(x, z)$, $\Lambda_{k,[1,0]}(x, z)$, and $\Lambda_{k,[0,1]}(x, z)$ respectively. That is, W can also be written in the form (1.2), albeit with a different dimension.

1.2 Simulation

For most nonlinear models, one cannot solve for parametric sensitivities as we did in Example 1.6. Instead, sensitivities must be approximated by simulating sample paths and employing Monte Carlo estimation.

1.2.1 Simulation Methods

Exact sample paths can be constructed in several ways. The two most common are the next reaction method (NRM) [1, 12], and the stochastic simulation algorithm (SSA), also known as Gillespie's direct algorithm [14]. We do not explicitly write the θ-dependence in the explanation of these methods, but the parametrized model is simulated in the same

way by simply using the parametrized intensities. Both simulation methods provide exact paths from the correct distribution.

Next Reaction Method

The random time change representation (1.2) very naturally suggests the NRM simulation method. At a given time t, the algorithm uses the integrated intensities $\int_0^t \lambda_k(X_s)ds$, called internal times, to determine what portions of each Y_k have been observed through time t. From this information, the method determines which counting process jumps next, and when that jump occurs. Recall that if $u \sim uniform(0,1)$, then $\log\left(\frac{1}{u}\right) \sim exp(1)$. Also recall that if $r \sim exp(1)$, then $\frac{r}{c} \sim exp(c)$.

ALGORITHM: Next Reaction Method

1. Initialize: set $x = X_0$, set $t = 0$, and set end time T. Also set $S_k = 0$ for each $k = 1, \ldots, K$.

2. Generate K *iid* $u_k \sim uniform(0,1)$. Set $I_k = \log\left(\frac{1}{u_k}\right)$ for each k.

3. Calculate each $\lambda_k(x)$. Set $\Delta = \min_k \frac{1}{\lambda_k(x)}(I_k - S_k)$ and $i = \text{argmin}_k \frac{1}{\lambda_k(x)}(I_k - S_k)$.

4. Stop if $t + \Delta > T$. Otherwise continue.

5. Set $t = t + \Delta$, set $S_k = S_k + \Delta\lambda_k(x)$ for each k, and set $x = x + \zeta_i$.

6. Generate $u \sim uniform(0,1)$ (independently of all other random variables) and set $I_i = I_i + \ln\left(\frac{1}{u}\right)$

7. Return to step 3.

Stochastic Simulation Algorithm

The SSA simulates the embedded discrete time Markov chain in the standard way and then samples the needed exponential holding times [14]. Formally, let $\lambda_0(x) = \sum_{k=1}^{K} \lambda_k(x)$ and for $j = 1, \ldots, K$ let $q_j(x) = \sum_{k=1}^{j} \frac{\lambda_k(x)}{\lambda_0(x)}$, with $q_0(x) := 0$. Note that $q_K(x) = 1$. Define $E_1(x) = [0, q_1(x)]$ and $E_k(x) = (q_{k-1}(x), q_k(x)]$ for $k = 2, \ldots, K$, so that for a given state x the intervals E_k partition the unit interval.

Let Y be a unit-rate Poisson process, and let $\{u_i\}$ be an i.i.d. sequence of uniform$(0, 1)$ random variables that are also independent of Y. Define the process X as

$$\begin{aligned} X_t &= X_0 + \sum_{k=1}^{K} \zeta_k \int_0^t \mathbf{1}\left\{u_{R_0(s-)} \in E_k(X_{s-})\right\} dR_0(s). \\ R_0(t) &= Y\left(\int_0^t \lambda_0(X_s) ds\right). \end{aligned} \tag{1.14}$$

Then X has the same distribution as (1.2) [4]. Note that the indicators, with the intervals and uniform random numbers, are simply encoding the usual "binning" or "dartboard" method of constructing a discrete random variable. The counting process R_0 determines the jump times T_i as exponential random variables with parameter $\lambda_0(X_{T_i-})$. This leads to the following algorithm:

ALGORITHM: Stochastic Simulation Algorithm

1. Initialize: set $x = X_0$, set $t = 0$, and set end time T.
2. Calculate each $\lambda_k(x)$ and set $\lambda_0(x) = \sum_{k=1}^{K} \lambda_k(x)$
3. Generate two iid $u_1, u_2 \sim uniform(0, 1)$.
4. Set $\Delta = \frac{1}{\lambda_0(x)} \log\left(\frac{1}{u_1}\right)$.

5. Stop if $t + \Delta > T$. Otherwise continue.

6. Find the minimal i such that $u_2 < q_i(x) = \sum_{k=1}^{i} \frac{\lambda_k(x)}{\lambda_0(x)}$.

7. Set $t = t + \Delta$ and set $x = x + \zeta_i$.

8. Return to step 2.

1.2.2 Monte Carlo Estimation

Suppose that one wishes to estimate some unknown quantity μ. To use the Monte Carlo estimation strategy, one first finds or constructs some random variable M such that either $\mathbb{E}[M] = \mu$ or $\mathbb{E}[M] \approx \mu$. Then one obtains some number n of independent, identically distributed samples $M_1, \ldots, M_n$ of M, generally via some simulation method, and computes the sample-mean estimator

$$\hat{M}_n = \frac{1}{n} \sum_{m=1}^{n} M_m,$$

called a Monte Carlo estimator. If $\mathbb{E}[M] = \mu$, then $\hat{M}_n$ is an unbiased estimator for μ; otherwise $\hat{M}_n$ has some nonzero bias $\mathbb{E}[M] - \mu$.

Such estimations are often done with some target variance in mind, or equivalently, some target half-width for a confidence interval. If we denote the target sample variance by V^*, we see that obtaining this variance requires a certain number of samples:

$$\mathrm{Var}(\hat{M}_n) = \mathrm{Var}\left(\frac{1}{n} \sum_{m=1}^{n} M_m\right) = \frac{1}{n}\mathrm{Var}(M) = V^* \quad \Longrightarrow \quad n = \frac{1}{V^*}\mathrm{Var}(M).$$

Thus a smaller target variance requires more samples, as expected. More importantly,

note that $\mathrm{Var}(M)$ plays a key role. A smart choice of M is one for which $\mathrm{Var}(M)$ is small: a lower variance estimator leads directly to a more efficient method.

There is an enormous amount of literature available on Monte Carlo methods, as they are used broadly across nearly all of the sciences. For more information about Monte Carlo methods in this general setting, we recommend [7].

Each of the sensitivity methods discussed or introduced in this thesis uses some Monte Carlo estimator, and so requires path simulation and averaging to produce sensitivity estimates. The main differences between these methods are generally due to the choice of estimator M, and often to the choice of the underlying probability space as well.

Chapter 2

Existing First Order Sensitivity Methods

Recall that we are interested in the output quantity of our CTMC model given as in (1.9) by

$$J(\theta) := \mathbb{E}[f(\theta, X_{t\in[0,T]}(\theta))] \ ,$$

where f is some function of θ and the path through a terminal time T and where $\theta \in \mathbb{R}^R$ is some vector of parameters. For the results in the thesis, we will require mild regularity conditions on f (see Condition 3.26). We wish to estimate

$$\nabla_\theta J(\theta) = \left[\frac{\partial}{\partial\theta_i}\mathbb{E}[f(\theta, X_{t\in[0,T]}(\theta))]\right]_{i=1,\ldots,R}. \tag{2.1}$$

Three main classes of computational methods have emerged to solve this problem for stochastic systems:

1. finite difference methods
2. likelihood ratio methods, and
3. pathwise differentiation methods.

In order to motivate these three methods, we consider a simple example. Let $\mu(\theta) \sim exp(\theta)$ be a 1-dimensional exponential random variable with parameter θ. Since $\mathbb{E}[\mu(\theta)] = \frac{1}{\theta}$, we may solve for $\frac{d}{d\theta}\mathbb{E}[\mu(\theta)] = \frac{d}{d\theta}\frac{1}{\theta} = -\frac{1}{\theta^2}$. We now describe how to estimate $\frac{d}{d\theta}\mathbb{E}[\mu(\theta)]$ using each of the three methods.

1. Finite Difference Methods. Finite difference methods take a very straightforward approach, approximating the sensitivity by estimating a finite difference as the name suggests. To approximate $\frac{d}{d\theta}\mathbb{E}[\mu(\theta)]$ via this method, we would choose some small $h > 0$ and construct a Monte Carlo estimator by sampling from

$$h^{-1}[\mu(\theta + h) - \mu(\theta)], \tag{2.2}$$

where $\mu(\theta + h)$ and $\mu(\theta)$ are exponential random variables constructed on the same probability space. If we use a space on which the two exponentials are independent, the random variable (2.2) has variance

$$\begin{aligned} \mathrm{Var}(h^{-1}[\mu(\theta + h) - \mu(\theta)]) &= h^{-2}\left[\mathrm{Var}(\mu(\theta + h)) + \mathrm{Var}(\mu(\theta))\right] \\ &= h^{-2}\left[\frac{\theta^2 + (\theta + h)^2}{\theta^2(\theta + h)^2}\right] = O(h^{-2}) \ , \end{aligned}$$

which is large for small h. If instead we use a space on which $\mu(1)$ is a unit exponential, and recalling that $\mu(\theta) \stackrel{d}{=} \frac{\mu(1)}{\theta}$ use $\mu(1)$ to construct both required random variables, we

have that

$$\begin{aligned}\mathrm{Var}(h^{-1}[\mu(\theta+h)-\mu(\theta)]) &= \mathrm{Var}\left(h^{-1}\left[\frac{\mu(1)}{\theta+h}-\frac{\mu(1)}{\theta}\right]\right)\\ &= \mathrm{Var}\left(h^{-1}\mu(1)\left[\frac{\theta-(\theta+h)}{\theta(\theta+h)}\right]\right)\\ &= \frac{1}{\theta^2(\theta+h)^2} = O(1).\end{aligned}$$

We discuss finite difference methods in the setting of stochastic processes in Section 2.1.

2. Likelihood Ratio Method. The likelihood ratio method, discussed in detail in Section 2.2, has also been successfully applied to CTMC models. In this method, one constructs paths of the process according to a θ-dependent probability measure. Again considering our example, we take the probability space $(\Omega, \mathcal{F}, P^\theta)$, where $\Omega = \mathbb{R}_{\geq 0}$, $\mathcal{F}$ is the corresponding σ-algebra of Borel sets, and P^θ is the measure defined by $P^\theta(A) = \int_A \theta e^{-\theta x} dx$ for $A \in \mathcal{F}$. On this space the random variable $X(\omega) = \omega$ has an exponential distribution with parameter θ. Then

$$\mathbb{E}[\mu(\theta)] = \mathbb{E}^\theta[X] = \int_\Omega X(\omega) dP^\theta(\omega) = \int_0^\infty x\theta e^{-\theta x} dx,$$

and, differentiating, we obtain

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}[\mu(\theta)] &= \int_0^\infty x(-x\theta e^{-\theta x} + e^{-\theta x}) dx\\ &= \int_0^\infty x(-x+\theta^{-1})\theta e^{-\theta x} dx\\ &= \mathbb{E}^\theta[X(-X+\theta^{-1})].\end{aligned}$$

We could then compute the sensitivity through Monte Carlo estimation by sampling the random variable $X(-X+\theta^{-1})$ under the measure P^θ. Note that the random variable $X(-X+\theta^{-1})$ does have the correct mean:

$$\mathbb{E}^\theta[X(-X+\theta^{-1})] = -\mathbb{E}^\theta[X^2] + \theta^{-1}\mathbb{E}^\theta[X] = -2\theta^{-2} + \theta^{-2} = -\theta^{-2}.$$

We describe this method in detail in Section 2.2 below.

3. Pathwise Differentiation Methods. In contrast to the likelihood ratio method, in a pathwise method one proceeds by using a probability space that is independent of θ, and then uses θ in the construction of the random variable or stochastic process. We again explain the idea on our simple example of an exponential random variable with parameter θ.

As defined on the previous page, we use the space $(\Omega, \mathcal{F}, P^1)$. On this space, the random variable $X(\omega) = \frac{\omega}{\theta}$ has an exponential distribution with parameter θ. Thus we have that

$$\frac{d}{d\theta}\mathbb{E}[\mu(\theta)] = \frac{d}{d\theta}\mathbb{E}^1[X] = \mathbb{E}^1\left[\frac{d}{d\theta}\frac{\omega}{\theta}\right] = \mathbb{E}^1\left[-\frac{\omega}{\theta^2}\right] = \mathbb{E}^1\left[-\frac{X}{\theta}\right].$$

That is, to estimate the desired sensitivity we can sample the random variable $-\frac{X}{\theta}$ under $\mathbb{P}^1$, which has mean $\mathbb{E}\left[-\frac{X}{\theta}\right] = -\frac{1}{\theta}\mathbb{E}[X] = -\theta^{-2}$ as expected.

Unfortunately, most existing pathwise methods can only be applied to a small fraction of the CTMCs that arise as models for biochemical systems. In Section 2.3, we introduce pathwise methods in more detail, and discuss the difficulties of pathwise methods in the CTMC setting.

Many of the methods described below can be extended to higher order sensitivities. In particular, second order sensitivities are addressed in more detail in Chapter 4.

2.1 Finite Difference Methods

Finite difference methods for approximating parameter sensitivities start with the simple observation that for smooth functions J we may approximate a derivative by perturbing the parameter vector in the relevant direction by some small $\epsilon > 0$, so that

$$\frac{\partial}{\partial \theta_i} J(\theta) \approx \frac{J(\theta + \epsilon e_i) - J(\theta)}{\epsilon}. \tag{2.3}$$

where e_i is the vector with a 1 in the i^{th} position and 0 elsewhere. To estimate the full gradient one simply estimates the entries $\frac{\partial}{\partial \theta_i} J(\theta)$ for $i = 1, \ldots, R$ one by one; thus for simplicity for the remainder of this section we assume that $R = 1$. The Monte Carlo estimator for the right-hand side of (2.3) is then

$$\hat{M}_n(\epsilon) = \frac{1}{n} \sum_{m=1}^{n} M_m(\epsilon), \tag{2.4}$$

where

$$M(\epsilon) = \frac{f(\theta + \epsilon, X_{t \in [0,T]}(\theta + \epsilon)) - f(\theta, X_{t \in [0,T]}(\theta))}{\epsilon}, \tag{2.5}$$

and where $M_m(\epsilon)$ denotes the m^{th} sample of $M(\epsilon)$. Note that $\hat{M}_n(\epsilon)$ is a biased estimator, since the approximation (2.3) has an error of $O(\epsilon)$. For ease of exposition and notation, we have used the forward difference (2.3). In practice, however, one may choose to use the central difference instead, which has a bias of only $O(\epsilon^2)$.

Note that one realization of $M(\epsilon)$ requires that we have two simulated paths, one of $X_{t\in[0,T]}(\theta)$ and one of $X_{t\in[0,T]}(\theta+\epsilon)$. Thus

$$\begin{aligned}
\mathrm{Var}(\hat{M}_n(\epsilon)) &= n^{-2}\sum_{m=1}^{n}\mathrm{Var}(M(\epsilon)) \\
&= n^{-1}\epsilon^{-2}\mathrm{Var}\big(f(\theta+\epsilon, X_{t\in[0,T]}(\theta+\epsilon)) - f(\theta, X_{t\in[0,T]}(\theta))\big) \\
&= n^{-1}\epsilon^{-2}\Big[\mathrm{Var}\big(f(\theta+\epsilon, X_{t\in[0,T]}(\theta+\epsilon))\big) \\
&\qquad + \mathrm{Var}\big(f(\theta, X_{t\in[0,T]}(\theta))\big) \\
&\qquad - 2\mathrm{Cov}\big(f(\theta, X_{t\in[0,T]}(\theta+\epsilon)), f(\theta, X_{t\in[0,T]}(\theta))\big)\Big].
\end{aligned} \tag{2.6}$$

We could simply compute the two paths independently. The associated method is then known as the Independent Random Numbers (IRN) method, or the crude Monte Carlo method. Because the covariance term in (2.6) is then zero, the variance of $\hat{M}_n(\epsilon)$ is $O(n^{-1}\epsilon^{-2})$. The goal of any coupling method in this context, therefore, is to lower the variance of $\hat{M}_n(\epsilon)$ by coupling the two processes $X_t(\theta)$ and $X_t(\theta+\epsilon)$. In particular, if $f(\theta, X_{t\in[0,T]}(\theta))$ and $f(\theta+\epsilon, X_{t\in[0,T]}(\theta+\epsilon))$ are positively correlated, the covariance term in (2.6) decreases the variance of the estimator $\hat{M}_n(\epsilon)$. The three finite difference methods below achieve this goal using different couplings and with varying amounts of effectiveness. In our setting, any of these coupling methods should always be used instead of the IRN method, since the extra effort of implementing some kind of coupling is usually more than made up for by the resulting gain in efficiency.

2.1.1 Common Random Numbers Method

As described in Section 1.2.1, the stochastic simulation algorithm uses a Poisson process Y (equivalently, an independent sequence of exponential random variables) to determine the holding times, and a sequence $\{u_i\}$ of independent uniform$(0,1)$ random variables to determine the path of the embedded chain. The common random numbers (CRN) method couples the two processes $X_t(\theta)$ and $X_t(\theta+\epsilon)$ in the finite difference $M(\epsilon)$ in (2.5) by using the same instance of Y and the same sequence $\{u_i\}$ in the construction of both paths in (1.14). However, a change in θ may cause some u_i to fall into different intervals for the two paths, an occurrence that becomes more likely with larger times t and that causes the two paths to "de-couple." Therefore, for moderate or large times, the finite difference Monte Carlo estimator constructed with CRN paths may not be substantially more efficient than computing the paths independently.

2.1.2 Common Reaction Path Method

While the model (1.2) can be simulated using the stochastic simulation algorithm, as noted in Section 1.2.1 the random time change representation is more naturally suited to the next reaction method. The Common Reaction Path (CRP) method [32] couples the two paths $X_t(\theta)$ and $X_t(\theta+\epsilon)$ by using the same unit-rate Poisson processes in the random time change representation. That is,

$$\begin{aligned} X_t(\theta) &= X_0(\theta) + \sum_{k=1}^{K} Y_k\left(\int_0^t \lambda_k(\theta, X_s(\theta))ds\right)\zeta_k, \\ X_t(\theta+\epsilon) &= X_0(\theta+\epsilon) + \sum_{k=1}^{K} Y_k\left(\int_0^t \lambda_k(\theta+\epsilon, X_s(\theta+\epsilon))ds\right)\zeta_k, \end{aligned} \tag{2.7}$$

where the same instances of the K independent Poisson processes Y_k are used in both processes.

Pairs of paths constructed in this way should be similar. Indeed, this method does not use the "binning" method that was troublesome for the CRN method, so even if the embedded chain becomes different in the two CRP paths, this does not always mean the paths will significantly diverge. For moderate times t, this implies that using the CRP method can significantly reduce the variance of the finite difference estimates $M(\epsilon)$, leading to a more efficient method than the CRN or IRN methods.

However, as time progresses, the two paths of (2.7) do begin to diverge, and so similarly to the CRN method, the CRP method generally becomes less effective at reducing variance for larger t. See [2, 32] for a more detailed discussion.

2.1.3 Coupled Finite Difference Method

The Coupled Finite Difference (CFD) method [2] utilizes the standard coupling of Section 1.1.2 for the two processes $X_t(\theta)$ and $X_t(\theta+\epsilon)$ in the finite difference $M(\epsilon)$. That is, using $3K$ independent Poisson processes $Y_{k,(1,1)}$, $Y_{k,(1,0)}$ and $Y_{k,(0,1)}$, the method constructs the pair of processes as:

$$\begin{aligned} X_t(\theta) &= X_0(\theta) + \sum_k \left(R_{k,[1,1]} + R_{k,[1,0]}\right)\zeta_k \\ X_t(\theta+\epsilon) &= X_0(\theta+\epsilon) + \sum_k \left(R_{k,[1,1]} + R_{k,[0,1]}\right)\zeta_k, \end{aligned} \tag{2.8}$$

where for $b_\ell \in \{0,1\}$ the $R_{k,[b_1,b_2]} = Y_{k,[b_1,b_2]}\left(\int_0^t \Lambda_{k,[b_1,b_2]}(X_s(\theta), X_s(\theta+\epsilon))ds\right)$ are counting processes with intensities

$$\Lambda_{k,[1,1]}(x,y) = \lambda_k(\theta,x) \wedge \lambda_k(\theta+\epsilon,y),$$
$$\Lambda_{k,[1,0]}(x,y) = \lambda_k(\theta,x) - \lambda_k(\theta,x) \wedge \lambda_k(\theta+\epsilon,y),$$
$$\Lambda_{k,[0,1]}(x,y) = \lambda_k(\theta+\epsilon,y) - \lambda_k(\theta,x) \wedge \lambda_k(\theta+\epsilon,y).$$

This coupling often leads to significant variance reduction of the finite difference estimates, even for large times. In particular, suppose that f is a function of the process solely at the terminal time T, so that $J(\theta) = \mathbb{E}[f(X_T(\theta))]$. Then under mild conditions, as proved in [2], the CFD method lowers the variance of the numerator of $M(\epsilon)$ to $O(\epsilon)$. This lowers $\mathrm{Var}(M(\epsilon))$ to $O(\epsilon^{-1})$ and yields $\mathrm{Var}(\hat{M}_n(\epsilon)) = O(n^{-1}\epsilon^{-1})$, a full order (in ϵ) lower than if the paths had been simulated independently, as computed using (2.6). The variance of a CFD estimate is often a full order smaller than the variances of CRN and CRP estimates as well, since the variance of the CRN and CRP estimates will often approach a value of order $O(n^{-1}\epsilon^{-2})$. An example of this behavior is shown in Figure 2.1. Thus, in most cases the CFD method is currently the most efficient finite difference method for computing first derivative sensitivities. See [2] for details.

2.2 The Likelihood Ratio Method

The likelihood ratio (LR) method constructs paths of the process according to a θ-dependent probability measure. That is, rather than constructing the path with the parameter θ, we let the probability measure on the path space be θ-dependent. In

particular, we will express the expectation in (2.1) with a θ-dependent path density, and then exchange the order of the derivative and the expectation. This density arises through a Girsanov change of measure transformation. For more details on this subject, see [23, 24, 29].

First consider a filtered probability space $(\Omega, \{\mathcal{F}_t\}_{t\geq 0}, Q)$ under which the counting processes $\{N_k, k = 1, \ldots, K\}$ are independent unit-rate Poisson processes. Recall that K is the number of reactions in our system; we think of $N_k(t)$ as the number of occurrences of reaction k. Then $X_t = X_0 + \sum_{k=1}^{K} N_k(t)\zeta_k$ is determined. While paths of X_t may leave the desired state space $\mathcal{S}$ (e.g. paths may leave the positive orthant), we construct a new measure under which X has the correct distribution and in particular under which such paths have probability zero.

Let $Z(t) = \{Z_k(t) : k = 1, \ldots, K\}$ be a nonnegative, cadlag function adapted to the filtration $\{\mathcal{F}_t\}$. Then we may define the Girsanov exponential

$$G(t) = \prod_{k=1}^{K} \exp\left(\int_0^t \log(Z_k(s-))dN_k(s) - \int_0^t (Z_k(s) - 1)ds\right). \tag{2.9}$$

In particular, we will take the $Z_k(t)$ to be our intensities $\lambda_k(X_t)$. Indeed, these rate functions are cadlag, and are nonanticipating because X_t is completely determined by the processes $\{N_k\}$ through time t. We define the rates to be zero if the process is outside $\mathcal{S}$, so that the rates are also nonnegative. This also implies that $G(t) = 0$ on any path that leaves $\mathcal{S}$ at some point through time t (under Q). With this choice for the Z_k, we can simplify to see that

$$G(t) = \exp\left(\sum_{\ell=0}^{N(t)-1} \log \lambda_{k_\ell}(\hat{X}_\ell) - \sum_{k=1}^{K} \int_0^t (\lambda_k(X_s) - 1)ds\right) \tag{2.10}$$

where $N(t) := \sum_k N_k(t)$ is the total number of jumps by time t, k_ℓ indicates that the ℓ^{th} jump was due to N_ℓ, and where $\hat{X}_\ell$ is the ℓ^{th} state of embedded chain of the process X with initial state $\hat{X}_0$.

Recall that we have assumed that our process X is non-explosive; that is, given

$$\tau_n = \inf_{t \geq 0} \left\{ \sum_{k=1}^{K} Y_k \left(\int_0^t \lambda_k(X_s) ds \right) = n \right\}$$

we assume that $\mathbb{P}\left(\lim_{n\to\infty} \tau_n > t\right) = 1$ for each $t > 0$. For a fixed time T it can be shown that $\mathbb{E}^Q[G(T)] = 1$, so that we may define the probability measure P via $dP = G(T)dQ$. This new measure P is absolutely continuous with respect to Q (written $P \ll Q$). See [24]. Since $G = \frac{dP}{dQ}$ is a Radon-Nikodym derivative, or density, we have for a function h of the path of X through some time T that

$$\mathbb{E}^P[h(X_{t\in[0,T]})] = \mathbb{E}^Q\left[h(X_{t\in[0,T]})G(T)\right].$$

Furthermore, under P, the processes $\{N_k\}$ have the same distribution as the solutions to the system

$$\left\{ R_k(t) = Y_k \left(\int_0^t \lambda_k(X_s) ds \right) \right\},$$

i.e., under P the process X has the same distribution as (1.2). See [24]. In particular, since $G(t) = 0$ on any path that leaves $\mathcal{S}$ at some point through time t, the set of such paths is a P-null set.

Considering the parameter θ, we can similarly obtain a family of probability measures $P^\theta \ll Q$ and a family of associated densities $G(\theta) = \frac{dP^\theta}{dQ}$. That is, under P^θ the process

X has the same distribution as the process $X(\theta)$ as given in (1.7), and similarly we have

$$\mathbb{E}^{P^\theta}[h(X_{t\in[0,T]})] = \mathbb{E}^{Q}\left[h(X_{t\in[0,T]})G(\theta,T)\right].$$

Fix some $i \in \{1,\dots,R\}$. As long as the exchange of the derivative and expectation is valid,

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}\mathbb{E}^{P^\theta}[h(X_{t\in[0,T]})] &= \frac{\partial}{\partial\theta_i}\mathbb{E}^{Q}\left[h(X_{t\in[0,T]})G(\theta,T)\right] \\
&= \mathbb{E}^{Q}\left[h(X_{t\in[0,T]})\frac{\partial}{\partial\theta_i}G(\theta,T)\right] \\
&= \mathbb{E}^{Q}\left[h(X_{t\in[0,T]})\frac{\frac{\partial}{\partial\theta_i}G(\theta,T)}{G(\theta,T)}G(\theta,T)\right] \\
&= \mathbb{E}^{Q}\left[h(X_{t\in[0,T]})\left(\frac{\partial}{\partial\theta_i}\log G(\theta,T)\right)G(\theta,T)\right] \\
&= \mathbb{E}^{P^\theta}\left[h(X_{t\in[0,T]})\frac{\partial}{\partial\theta_i}\log G(\theta,T)\right].
\end{aligned} \tag{2.11}$$

We give sufficient conditions for the validity of this exchange below. First, recall that we wish to estimate sensitivities of functions $f(\theta, X_{t\in[0,T]})$ which may depend explicitly on the parameter. This can be done similarly using the product rule, again assuming that the exchange of the derivative and expectation is valid:

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}\mathbb{E}^{P^\theta}[f(\theta,X_{t\in[0,T]})] &= \frac{\partial}{\partial\theta_i}\mathbb{E}^{Q}\left[f(\theta,X_{t\in[0,T]})G(\theta,T)\right] \\
&= \mathbb{E}^{Q}\left[G(\theta,T)\frac{\partial}{\partial\theta_i}f(\theta,X_{t\in[0,T]}) + f(\theta,X_{t\in[0,T]})\frac{\partial}{\partial\theta_i}G(\theta,T)\right] \\
&= \mathbb{E}^{Q}\left[G(\theta,T)\left(\frac{\partial}{\partial\theta_i}f(\theta,X_{t\in[0,T]}) + f(\theta,X_{t\in[0,T]})\frac{\partial}{\partial\theta_i}\log G(\theta,T)\right)\right] \\
&= \mathbb{E}^{P^\theta}\left[\frac{\partial}{\partial\theta_i}f(\theta,X_{t\in[0,T]}) + f(\theta,X_{t\in[0,T]})\frac{\partial}{\partial\theta_i}\log G(\theta,T)\right].
\end{aligned} \tag{2.12}$$

If there exists some neighborhood $\Theta \subset \mathbb{R}^R$ of θ where the following sufficient conditions are true, the exchange of the derivative and expectation is valid:

(1) $\mathbb{E}^Q\left[f(\theta, X_{t\in[0,T]})G(\theta,T)\right] = \mathbb{E}^{P^\theta}[f(\theta, X_{t\in[0,T]})]$ exists for each $\theta \in \Theta$ (i.e., the quantity whose sensitivity we would like to estimate exists),

(2) Q-almost surely, $\frac{\partial}{\partial\theta_i}\left[f(\theta, X_{t\in[0,T]})G(\theta,T)\right]$ exists for each $\theta \in \Theta$, and

(3) $\left|\frac{\partial}{\partial\theta_i}\left[f(\theta, X_{t\in[0,T]})G(\theta,T)\right]\right|$ can be bounded uniformly on Θ by some random variable in $L^1(Q)$.

This is a generalization of the Leibniz integral rule. The differentiability of G follows as long as the intensities λ are differentiable, and we will assume that f is differentiable in θ. The first and third conditions are more difficult, though are believed to be true for most reasonable choices of (non-explosive) process and functions f. For nearly all biologically relevant processes and functions f, we prove that the exchange is valid in Section 3.4.3.

Consequently, assuming that we may exchange the derivative and expectation, we may use the Monte Carlo estimator

$$\hat{M}_n = \frac{1}{n}\sum_{m=1}^{n} M_m \quad \text{with} \quad M = \frac{\partial}{\partial\theta_i} f(\theta, X_{t\in[0,T]}) + f(\theta, X_{t\in[0,T]})H_i(\theta,T),$$

where M_m is the m^{th} sample of M and where $H_i(\theta,T) := \frac{\partial}{\partial\theta_i}\log G(\theta,T)$. Note that the path samples are taken with respect to our the measure P^θ.

The random variable $H_i(\theta,T)$ is commonly known as a weighting function, or simply weight. This weight is simple to compute during path simulation, since, by parametrizing

and differentiating (2.10) we have that

$$H_i(\theta, T) = \sum_{\ell=0}^{N(T)-1} \frac{\frac{\partial}{\partial \theta_i}\lambda_{k_\ell}(\theta, \hat{X}_\ell)}{\lambda_{k_\ell}(\theta, \hat{X}_\ell)} - \sum_{k=1}^{K} \int_0^T \frac{\partial}{\partial \theta_i}\lambda_k(\theta, X_s)ds. \tag{2.13}$$

Under the mass-action assumption, this simplifies to

$$H_i(\theta, T) = \frac{1}{\theta_i}\left(N_i(T) - S_i(T)\right)$$

where $S_i(T) = \int_0^T \lambda_i(\theta, X(s))ds$ is the internal time at (real) time T of reaction i. Recall that $N_i(T)$ gives the number of times the i^{th} reaction has occurred by time T.

The likelihood ratio method is consequently quite simple to implement. As one simulates a path of the process $X_{t\in[0,T]}$ at the original, or nominal, parameter value of θ, one should also compute $H_i(\theta, t)$ and the other quantities required for the estimate M. Either simulation method of Section 1.2.1 may be used. For an explicit algorithm, see, for example [28] or [37]. Note that unlike finite difference methods, the likelihood ratio method produces an unbiased estimator.

While the LR method is easy to implement and unbiased, it is generally quite inefficient due to a high variance. A helpful observation comes from the fact that the weight function is a mean zero random variable, and can therefore be used as a control variate (see [7]). However, even with this addition to the method, the variance can still be prohibitively large, particularly for large simulation times, and so for many models this method is not practical.

Finally, we note that in our CTMC setting the likelihood ratio method is also known as the Girsanov method, after the change of measure derivation. A slightly different

derivation explains the name likelihood ratio method. Recall that for a given θ, we constructed the measure $P^\theta \ll Q$ with density $G(\theta) = \frac{dP^\theta}{dQ}$. Then, if $P^\theta \ll P^{\theta_0}$, the Radon-Nikodym derivative $\frac{dP^\theta}{dP^{\theta_0}}$ is given by $\frac{G(\theta,t)}{G(\theta_0,t)}$. That is, the measure P^{θ_0} at the fixed, nominal value of θ_0 can be used as the reference measure. Then noting that

$$\frac{\partial}{\partial\theta_i}\frac{G(\theta,t)}{G(\theta_0,t)}\bigg|_{\theta_0} = \frac{\partial}{\partial\theta_i}\log G(\theta,t)\bigg|_{\theta_0}$$

we see that the same weight function is derived. This ratio of densities $\frac{G(\theta,t)}{G(\theta_0,t)}$ is known as a likelihood ratio; its derivative is often called a score function. This latter change of measure is the one described in [7] or [28]. Alternatively, one can explicitly construct a reference measure similar to P^{θ_0} without appealing to the Girsanov exponential. See [16] for further details.

This likelihood ratio method, in the context of finding sensitivities in stochastic processes, arose in the discrete event literature in the late 1980s. It is also known as the score function method, the importance sampling method, and the "what if" method. In addition to the references above, see, for example, [5, 17, 34, 35] for the likelihood ratio method in this and other settings.

2.3 Pathwise Methods

Pathwise (PW) methods have a much more limited scope, and are invalid in many settings. However, a rule of thumb in the literature suggests that pathwise methods, when they apply, are more efficient than the likelihood ratio method.

When using a pathwise method, one proceeds differently than when using the likelihood ratio method, by viewing the construction of the path as depending on θ while using a probability measure that does not depend on θ. For example, one can think of the space as consisting of K independent streams of independent unit exponential random variables, each stream determining one of the K Poisson processes Y_k in the random time change representation (1.7), and constructing the path through the next reaction method. To obtain a pathwise estimate, one then proceeds by switching the order of the derivative and the expectation in the sensitivity.

However, we must take care. Since the process X is a CTMC, the paths are piecewise constant, with discrete jumps. Suppose that f is a function solely of the process at the terminal time T, so that $J(\theta) = \mathbb{E}[f(X_T(\theta))]$. Then it is most always the case that

$$\frac{\partial}{\partial \theta_i}\mathbb{E}[f(X_T(\theta))] \neq \mathbb{E}\left[\frac{\partial}{\partial \theta_i} f(X_T(\theta))\right],$$

since the derivative of X is zero wherever it exists. Thus we cannot build a Monte Carlo estimator by sampling $\frac{\partial}{\partial \theta_i} f(X_T(\theta))$. Instead, we can first replace $f(X_T(\theta))$ with some function with the same mean but that is smoother in θ. In this case, we briefly describe two options for this replacement, but then describe a major problem that makes both choices invalid for many of our biochemical models.

The Regularized Pathwise Derivative (RPD) method presented in [37] estimates $\frac{\partial}{\partial \theta_i} J(\theta)$ using derivatives of the time averaged random variable

$$L(\theta) = \frac{1}{2w}\int_{T-w}^{T+w} f(X_s(\theta))ds \approx f(X_T(\theta)), \tag{2.14}$$

where w is some fixed window size. That is, the RPD method uses the Monte Carlo estimator

$$\hat{M}_n = \frac{1}{n}\sum_{m=1}^{n} M_m \quad \text{where} \quad M = \frac{\partial}{\partial \theta_i} L(\theta) \tag{2.15}$$

and where M_m is the m^{th} sample of M. When the method applies, it is generally very efficient, though it gives a biased estimate, with the size of the bias a function of the size of w. Specifically, a smaller w typically leads to a smaller bias but a larger variance.

Alternatively, though we are unaware of this specific method in the literature, one can use Dynkin's formula (1.6) to derive an unbiased estimator. For this method, we let $L(\theta) = f(X_0) + \int_0^T (\mathcal{A}^\theta f)(X_s(\theta,\omega))ds$, so that $\mathbb{E}[L(\theta)] = \mathbb{E}[f(X_T(\theta)]$, and use a Monte Carlo estimator constructed as in (2.15). While unbiased when it applies, this estimator tends to have higher variance than the RPD estimator. We shall refer to this method as the Dynkin pathwise method.

In both cases, we introduced an integral in the hope that the resulting random variable L would become smoother in θ than $f(X_T(\theta))$. Assuming that the paths are constructed using the next reaction method, we can argue that these $L(\theta)$ (both special cases of a more general functional) are almost surely differentiable. However, even after our attempts at smoothing, neither is a valid estimator in general: in many cases we will still have that $\frac{\partial}{\partial \theta_i}\mathbb{E}[L(\theta)] \neq \mathbb{E}\left[\frac{\partial}{\partial \theta_i} L(\theta)\right]$. In the next chapter, we explore why this is so, and provide a method that avoids the problem.

We also note that Gupta and Khammash, in [19], analytically derive a different pathwise estimator using a limiting argument. The method provides an unbiased estimate, though the algorithm is complicated to implement, as the method requires a significant

amount of auxiliary simulation, which may also significantly reduce the method's efficiency. Pathwise methods have also been explored in the discrete event system literature, where such methods are known as (infinitesimal) perturbation analysis, or IPA or PA for short. See for example [20], or [15], in which Glasserman presents a pathwise method for sensitivities of CTMCs that applies to a strictly larger subset of CTMC models than the two methods discussed above. However, there does not seem to be a way to extend this method to all biochemical models. Similarly, there is a class of sensitivity methods for discrete event systems referred to as Smoothed Perturbation Analysis (SPA). See for example [18]. These methods condition on some information $\mathcal{F}$ to construct the estimator $\mathbb{E}[f(X_t(\theta))|\mathcal{F}]$ of the quantity $\mathbb{E}[f(X_t(\theta))]$. The choice of $\mathcal{F}$, however, is key, and there does not appear to be a choice of $\mathcal{F}$ so that the pathwise method is both applicable and tractable for our biochemical CTMC models (1.7).

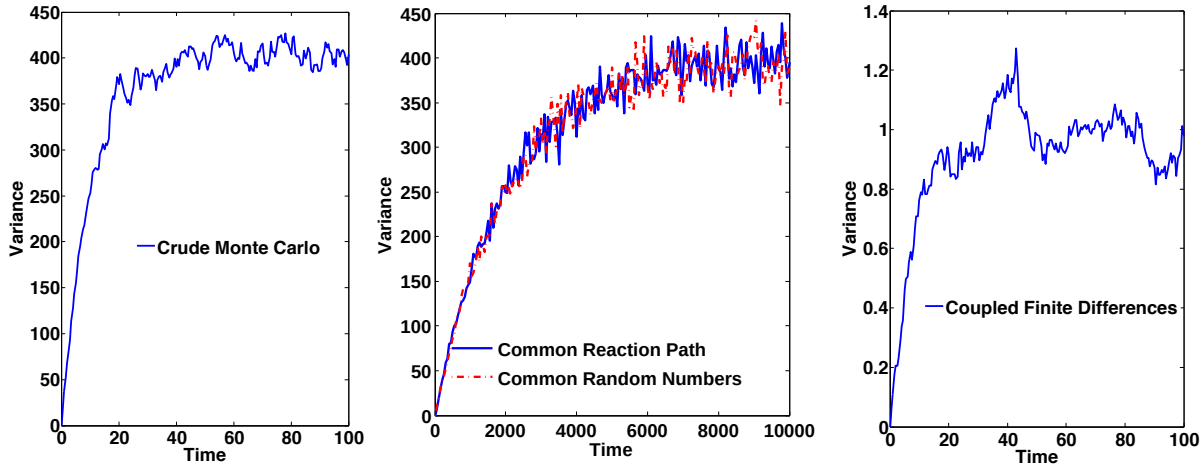

Figure 2.1: Convergence in time of the variance of the different first order finite difference sensitivity estimators $\hat{M}_n(\epsilon)$ of $\frac{\partial}{\partial\theta_1}\mathbb{E}[X_t(\theta)]$ on the birth–death model $\emptyset \xrightarrow{\theta_1} \mathcal{S}$, $\mathcal{S} \xrightarrow{\theta_2} \emptyset$ at $(\theta_1, \theta_2) = (2, 0.1)$ and $X_0 = 0$, where X_t gives the abundance of $\mathcal{S}$ at time t. The paper [2] compares these methods on this and other examples. A perturbation of $\epsilon = .01$, the centered difference, and 10^3 sample paths were used. The variance axes have two drastically different scales, and we have also used two different time scales to better demonstrate these convergences. Note that the limiting value of the CFD method is much smaller than the limiting value of the three other finite difference methods. For this example, CRP and CRN are approximately equivalent; this is not always the case [32].

Chapter 3

A Hybrid Pathwise Method for First Order Sensitivities

In Section 2.3, we introduced two main pathwise methods for sensitivity estimation, the Dynkin pathwise method, and the Regularized Pathwise Derivative (RPD) method (as in [37]). Both are special cases of a more general functional L, so we now study pathwise methods in this more general framework.

For non-explosive CTMCs X given by the random time change representation (1.7), and for $a < b$, we are interested in computing the θ-sensitivities of expectations of functionals of paths of the form

$$L_X(\theta) := \int_a^b F(\theta, X_s(\theta))ds, \tag{3.1}$$

where F is a function of the state of the process that may also depend explicitly on θ. For the results in this thesis, we will require F to satisfy certain regularity conditions; see Condition 3.11.

Note that if we find a method for computing sensitivities of the more general functionals (3.1), by taking $a = 0$, $b = T$ some fixed, finite time, and $F = \mathcal{A}^\theta f$, we can compute sensitivities of the form $\frac{\partial}{\partial \theta_i}\mathbb{E}[f(X_T(\theta))]$ by the Dynkin pathwise method of the

previous section. We could also use the RPD estimator (2.14) by taking $a = T - w$, $b = T + w$, and $F = \frac{1}{2w} f$. This will be revisited in more generality in Section 3.3.

Pathwise methods can be applicable to stochastic processes built on a probability space that is θ-independent. We use a filtered probability space $(\Omega, \{\mathcal{F}_t\}_{t\geq 0}, Q)$ under which $\{Y_k, k = 1, \ldots, K\}$ are independent unit-rate Poisson processes. The path of X is then a function of both ω and of θ as in the random time change representation (1.7). This is in contrast to the likelihood ratio method, in which paths are determined by the probability space alone.

The goal of our pathwise method is to construct a Monte Carlo estimator for the sensitivity $\frac{\partial}{\partial \theta_i}\mathbb{E}[L_X(\theta)]$ by obtaining n samples $M_1, \ldots, M_n$ of the random variable $M = \frac{\partial}{\partial \theta_i} L_X(\theta)$ and computing the sample average $\hat{M}_n = \frac{1}{n}\sum_{m=1}^{n} M_m$. Of course, we only expect $\hat{M}_n$ to approximate $\frac{\partial}{\partial \theta_i}\mathbb{E}[L_X(\theta)]$ if

$$\frac{\partial}{\partial \theta_i}\mathbb{E}[L_X(\theta)] = \mathbb{E}\left[\frac{\partial}{\partial \theta_i} L_X(\theta)\right].$$

If this equation holds, a single path of X can be used in the pathwise estimators of each entry of the gradient by computing $\frac{\partial}{\partial \theta_i} L_X(\theta)$ for each $i = 1, \ldots, R$.

We will show in Lemma 3.14 that at a given value of θ the function L_X is a.s. differentiable. That is, defining $D(\theta, h) := h^{-1}(L_X(\theta + h) - L_X(\theta))$, as $h \to 0$ we have

$$D(\theta, h) \xrightarrow{a.e.} \frac{\partial}{\partial \theta_i} L_X(\theta). \tag{3.2}$$

Since $\lim_{h\to 0}\mathbb{E}[D(\theta, h)] = \frac{d}{d\theta}\mathbb{E}[L_X(\theta)]$ by definition, if the convergence (3.2) is also in

the mean, then as needed,

$$\frac{\partial}{\partial\theta_i}\mathbb{E}[L_X(\theta)] = \mathbb{E}\left[\frac{\partial}{\partial\theta_i}L_X(\theta)\right]. \tag{3.3}$$

However, as discussed in Section 2.3 in the context of the Dynkin pathwise and RPD methods, even though the function L_X is an integral, L_X is *not* always smooth enough for this L^1 convergence to hold. Thus (3.3) is false for many of our CTMC models (1.7), for which pathwise methods are therefore not applicable. In Section 3.1, we explore how pathwise methods can fail on our CTMC models, and give the condition under which pathwise methods are valid. We prove this validity in Section 3.2.

In Section 3.3 we introduce the new hybrid method, the main work of this thesis, which allows us to incorporate a pathwise method in the estimation of sensitivities of the CTMC models (1.7) in general. The hybrid method works by first constructing a CTMC Z that approximates X, but also such that pathwise methods are valid for Z. Then, coupling X and Z in the standard manner of Section 1.1.2, we note that

$$E[L_X(\theta)] = \mathbb{E}[L_X(\theta) - L_Z(\theta)] + \mathbb{E}[L_Z(\theta)].$$

To estimate the sensitivity $\frac{\partial}{\partial\theta_i}E[L_X(\theta)]$, we compute separately the two sensitivities $\frac{\partial}{\partial\theta_i}\mathbb{E}[L_X(\theta) - L_Z(\theta)]$ and $\frac{\partial}{\partial\theta_i}\mathbb{E}[L_Z(\theta)]$, the first using the likelihood ratio method, and the second using a pathwise method. The benefit of this strategy is that both estimators will tend to have a low variance. Indeed, the variance of the first estimator will be small because X and Z are coupled, and the variance of the second estimator will be small because it utilizes a pathwise method. Overall, the low variance of these estimators leads

to a more efficient sensitivity method.

More details are provided in Section 3.3. In Section 3.4, we prove some moment results for our CTMC models, which we then use to prove the validity of the likelihood ratio piece our hybrid estimates. We also use these moment bounds in Section 3.5 to prove a limiting result regarding the hybrid method. Finally, in Section 3.6 we give numerical examples of the hybrid method and compare its performance with existing methods.

3.1 The Non-Interruptive Condition

Recall that $D(\theta, h) = h^{-1}(L_X(\theta + h) - L_X(\theta))$. As discussed above, we would like $D(\theta, h) \xrightarrow{L_1} \frac{\partial}{\partial \theta_i} L_X(\theta)$ to hold, so we wish to find some L^1 function dominating $D(\theta, h)$. Because $D(\theta, h)$ includes a division by h, we will certainly run into difficulty if L_X has a jump discontinuity at θ. Unfortunately, L_X can indeed have a jump discontinuity. To help illustrate how, as well as to suggest the condition necessary to ensure that L_X is continuous, we consider a simple example.

Example 3.7. *Consider the model*

$$A \xrightarrow{\theta} \emptyset, \quad A \xrightarrow{1} B$$

where we suppose that $X_t(\theta) \in \mathbb{Z}^2_{\geq 0}$ gives the abundances of A and B respectively, with an initial condition of $X_0(\theta) = (1, 0)$. Define $L_X(\theta) = \int_0^T X_{B,s}(\theta) ds$.

Labeling $A \to \emptyset$ and $A \to B$ as reactions 1 and 2 respectively, consider the system at time 0. By the random time change representation (1.7)*, the time until the first reaction*

is $\min\{\frac{e_1}{\theta}, e_2\}$*, where* e_1 *and* e_2 *give the times of the first jumps of* Y_1 *and* Y_2 *respectively. Both* e_1 *and* e_2 *have a unit exponential distribution; what is important here is that they do not depend on* θ*.*

Then

$$L_X(\theta) = \int_0^T X_{B,s}(\theta)ds = (T - e_2)^+ \mathbf{1}\left(\frac{e_1}{\theta} > e_2\right)$$

has a jump discontinuity at $\theta = \frac{e_1}{e_2}$ *if* $e_2 < T$*.*

We can also see that $\frac{d}{d\theta}L_X(\theta)$ *is zero where it exists. However,* $\frac{d}{d\theta}\mathbb{E}[L_X(\theta)]$ *is not zero, as will be shown in Example 3.6.2. Thus* (3.3) *does not hold, and pathwise methods are not valid on* L_X*.*

◇

The question arises: why did an attempt at a pathwise method fail in the previous example, and how can they fail in general? We look closely at this problem, since it will provide a good intuition for later proofs.

A change in θ changes the intensities, which in turn causes two different types of change in $L_X(\theta)$. First, changes in θ will change the jump times of X. These changes alone are not problematic, as the jump times, for the most part, simply scale with the intensities, which we will assume to be θ-differentiable. Second, however, a change in θ can change the embedded chain of the process. This is a more drastic change in the process, and causes the more serious problem of possible discontinuities in L.

Assuming we construct our paths using the next reaction method, however, changes in the embedded chain do not necessarily cause a discontinuity. Indeed, suppose that we have in hand realizations of the K Poisson processes Y_k. Consider two consecutive jumps of X, say reactions 1 and 2, occurring in that order. As θ increases (or similarly for decreases), the time between these jumps may decrease to the point where the process

sees reaction 2 occur immediately after reaction 1. A further increase in θ may switch the order of the reactions, since, considering the internal times of the two relevant reactions, reaction 2 now happens first; reaction 1 is "just about" to fire as well, which should result in reaction 1 occurring immediately after reaction 2. If the two reaction occurrences simply switch order, the change in the path is in a sense still minimal, and in fact L_X remains continuous in θ throughout. This is often referred to as a *crossover* in the discrete event literature, and we will use the same terminology here. The argument above is formalized in Lemma 3.16.

A crossover is not what occurred in our Example 3.7: we cannot have the two possible jumps $A \to B$ and $A \to \emptyset$ simply cross over each other, because if the single A molecule decays first, the intensity of reaction $A \to B$ becomes zero (and vice versa). Returning to our discussion from the previous paragraph, suppose that an increase in θ has now caused reaction 2 to occur before reaction 1. Then reaction 1 is just about to occur because of the internal time structure *unless the occurrence of reaction 2 prevents the occurrence of reaction 1* by setting the intensity λ_1 to zero. If such an *interruption* occurs, L_X is discontinuous and pathwise methods are not applicable.

Interestingly, if we had used an SSA instead of an NRM construction this behavior regarding the order changes of reactions due to changes in θ is very different. In the SSA case, problems other than interruptions can arise because of the "dartboard" or "binning" method of choosing a state: there is no similar sense of competing jumps. That is, the random time change representation with NRM behaves differently than an SSA construction when we consider small perturbations of the parameter, even though the two methods produces paths with the same distribution.

From our argument above, in order for a pathwise method to be valid, we will require the following condition prohibiting such interruptions. It formalizes the notion that the occurrence of a reaction cannot set the rate of another to zero. Recall that $\mathcal{S}$ is the state space of X.

Condition 3.8 (Non-Interruptive)**.** *The intensities $\lambda_k, k = 1, \ldots, K$ of X satisfy this condition on a set Θ if at every state x in the state space $\mathcal{S}$ and every $\theta \in \Theta$, if $\lambda_k(\theta, x) > 0$ and $\lambda_\ell(\theta, x) > 0$ for $k \neq \ell$, then $\lambda_k(\theta, x + \zeta_\ell) > 0$ and $\lambda_\ell(\theta, x + \zeta_k) > 0$.*

3.2 A Pathwise Method for Non-Interruptive CTMCs

Let $\hat{X}_i(\theta)$ denote the i^{th} state in the embedded discrete time chain of the process $X_t(\theta)$, and let T_i^θ be the i^{th} jump time, with $T_0^\theta = 0$. We let $N(\theta) := N(\theta, b)$ be the number of jumps of X by time b. Assume that F is differentiable in θ with some mild regularity conditions (see Condition 3.11). We are interested in computing the θ-sensitivities of expectations of functionals of paths of the form

$$L_X(\theta) := \int_a^b F(\theta, X_s(\theta))ds = \sum_{i=0}^{N(\theta)} F(\theta, \hat{X}_i(\theta))[T_{i+1}^\theta \wedge b - T_i^\theta \vee a]^+, \tag{3.4}$$

where $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. The summation is simply the integral of the piece-wise constant integrand, restricted to $[a, b]$. Since X is non-explosive, N^θ is almost surely finite.

In Section 3.2.1, we calculate explicit forms of the derivatives of L_X with respect to θ_i. In Section 3.2.2, we give additional conditions we will require, including that

the intensities are bounded. We show, in Sections 3.2.3 and 3.2.4, that with the Non-Interruption Condition 3.8 and the additional conditions of Section 3.2.2, the pathwise method is valid; i.e., that as in (3.3) we have

$$\frac{d}{d\theta_i}\mathbb{E}[L_X(\theta)] = \mathbb{E}\left[\frac{d}{d\theta_i}L_X(\theta)\right].$$

In the subsequent sections of this chapter, we show how to incorporate the pathwise method into a general sensitivity method, one that *does not* require the process to have bounded intensities or the non-interruptive condition.

For notational convenience, in this section we take θ to be 1-dimensional.

3.2.1 Path Derivatives

We compute here explicit forms of the derivative of $L_X(\theta)$ for use in the method and in its proof. The states in the embedded chain are discrete-valued, so for any i, the derivative $\frac{d}{d\theta}\hat{X}_i(\theta)$ is zero wherever it exists. Therefore, from (3.4) we have that

$$\begin{aligned}\frac{d}{d\theta}L_X(\theta) = \sum_{i=0}^{N(\theta)}\Bigg[&[T_{i+1}^\theta \wedge b - T_i^\theta \vee a]^+\left(\frac{\partial}{\partial\theta}F(\theta,\hat{X}_i(\theta))\right)\\ &+ F(\theta,\hat{X}_i(\theta))\frac{\partial}{\partial\theta}[T_{i+1}^\theta \wedge b - T_i^\theta \vee a]^+\Bigg],\end{aligned} \tag{3.5}$$

where the partials of the function F are with respect to the first variable.

The terms in the top line of (3.5) are straightforward. The terms in the second line require the derivatives of the jump times T_i^θ, so we now focus on their derivation. Define $\Delta_i^\theta = T_{i+1}^\theta - T_i^\theta$ to be the holding time of the process in the i^{th} state (so that the indexing begins at 0). We first derive $\frac{\partial}{\partial\theta}\Delta_i^\theta$ as in [37]. Below, we show how to then compute the

derivatives of the $[T^\theta_{i+1} \wedge b - T^\theta_i \vee a]^+$ as in (3.5).

Let $S^\theta_k(t) = \int_0^t \lambda_k(\theta, X_s(\theta))ds$, and let $I^k_+(t) = \inf\left\{r \geq S^\theta_k(t) : Y_k(r) > Y(S^\theta_k(t))\right\}$. Then

$$\Delta^\theta_i = \min_k \left\{ \frac{I^k_+(T^\theta_i) - S^\theta_k(T^\theta_i)}{\lambda_k(\theta, \hat{X}_i(\theta))} \right\}.$$

Let k_i be the argmin in the above expression, so that k_i is is the index of the reaction that accounted for the i^{th} reaction of this realization of the process. We can then derive using the product rule that

$$\begin{aligned} \frac{\partial}{\partial\theta}\Delta^\theta_i &= -\frac{I^{k_i}_+ - S^\theta_{k_i}(T^\theta_i)}{\lambda_{k_i}(\theta, \hat{X}_i(\theta))^2}\frac{\partial}{\partial\theta}\lambda_{k_i}(\theta, \hat{X}_i(\theta)) - \lambda_{k_i}(\theta, \hat{X}_i(\theta))^{-1}\frac{\partial}{\partial\theta}S^\theta_{k_i}(T^\theta_i) \\ &= -\frac{\Delta^\theta_i}{\lambda_{k_i}(\theta, \hat{X}_i(\theta))}\frac{\partial}{\partial\theta}\lambda_{k_i}(\theta, \hat{X}_i(\theta)) - \lambda_{k_i}(\theta, \hat{X}_i(\theta))^{-1}\frac{\partial}{\partial\theta}S^\theta_{k_i}(T^\theta_i). \end{aligned} \tag{3.6}$$

Note that $\frac{\partial}{\partial\theta}T^\theta_0 = 0$. For $i > 0$, the definition of Δ^θ_i implies that

$$\frac{\partial}{\partial\theta}T^\theta_i = \sum_{j=0}^{i-1}\frac{\partial}{\partial\theta}\Delta^\theta_j.$$

Now, for $t \in \left[T^\theta_i, T^\theta_{i+1}\right]$ we have for any k that $S^\theta_k(t) = S^\theta_k(T^\theta_i) + \lambda_k(\theta, \hat{X}_i(\theta))(t - T^\theta_i)$. We can use this to find that

$$\frac{\partial}{\partial\theta}S^\theta_k(T^\theta_i) = \frac{\partial}{\partial\theta}S^\theta_k(T^\theta_{i-1}) + \Delta^\theta_{i-1}\frac{\partial}{\partial\theta}\lambda_k(\theta, \hat{X}_{i-1}(\theta)) + \lambda_k(\theta, \hat{X}_{i-1}(\theta))\frac{\partial}{\partial\theta}\Delta^\theta_{i-1}. \tag{3.7}$$

The $\{\Delta^\theta_i\}$ and $\{S^\theta_i\}$ values can be solved for recursively given that the initial $S^\theta_k(T^\theta_0)$ values are 0 for each k. Let $i_a \in \mathbb{N}$ be maximal such that $T^\theta_{i_a} < a$; that is, the i^{th}_a jump

is the last jump to occur before time a. Then

$$\frac{\partial}{\partial\theta}[T^\theta_{i+1}\wedge b - T^\theta_i \vee a]^+ = \begin{cases} 0 & i < i_a \quad \text{or} \quad i > N(\theta) \\ \sum_{j=0}^{i_a} \frac{\partial}{\partial\theta}\Delta^\theta_j = \frac{\partial}{\partial\theta}T^\theta_{i_a+1} & i = i_a \\ \frac{\partial}{\partial\theta}\Delta^\theta_i & i_a \le i < N(\theta) \\ -\sum_{j=0}^{N(\theta)-1} \frac{\partial}{\partial\theta}\Delta^\theta_j = -\frac{\partial}{\partial\theta}T^\theta_N & i = N(\theta) \end{cases} \tag{3.8}$$

can easily be computed as needed in (3.5).

This leads to the following pathwise algorithm for simulating a path of $X_t(\theta)$ and computing $\frac{d}{d\theta}L_X(\theta)$. This algorithm is a more general version of the algorithm in [37].

ALGORITHM: Pathwise Derivative Algorithm

1. Initialize: set $x = X_0$, set $t = 0$, and set end time T. Set $S_k = 0$ and $dS_k = 0$ for each $k = 1, \ldots, K$, and set $dL = 0$. Set $flag = 0$ and set $dT = 0$.

2. Generate K *iid* $u_k \sim uniform(0,1)$. Set $I_k = \log\left(\frac{1}{u_k}\right)$ for each k.

3. Calculate each $\lambda_k(x)$. Set $\Delta = \min_k \frac{1}{\lambda_k(x)}(I_k - S_k)$ and $i = \text{argmin}_k \frac{1}{\lambda_k(x)}(I_k - S_k)$.

4. If $t + \Delta > T$, go to Step 11. Otherwise continue to Step 5.

5. Set $d\Delta = -\frac{\Delta}{\lambda_i(\theta,x)}\frac{\partial}{\partial\theta}\lambda_i(\theta,x) - \frac{dS_i}{\lambda_i(\theta,x)}$. Then set $dT = dT + d\Delta$.

6. Set $dL = dL + \Delta\frac{\partial}{\partial\theta}F(\theta,x) + F(\theta,x)A$, where $A = \begin{cases} 0 & t < a \\ dT & t > a \quad \text{and} \quad flag = 0 \\ d\Delta & \text{otherwise} \end{cases}$

   If $t > a$ and $flag = 0$, set $flag = 1$.

7. Set $t = t + \Delta$. Also, for each k set $S_k = S_k + \Delta\lambda_k(x)$ and set $dS_k = dS_k + \Delta\frac{\partial}{\partial\theta}\lambda_k(\theta, x) + \lambda_k(\theta, x)d\Delta$.

8. Set $x = x + \zeta_i$.

9. Generate $u \sim uniform(0, 1)$ and set $I_i = I_i + \ln\left(\frac{1}{u}\right)$

10. Return to step 3.

11. Set $dL = dL + (T - t)\frac{\partial}{\partial\theta}F(\theta, x) - F(\theta, x)dT$.

The quantity dL gives $\frac{d}{d\theta}L_X(\theta)$ for the simulated instance of the path of X. Note that this algorithm is only valid using the next reaction method as given here. See the discussion at the end of Section 3.1.

3.2.2 Other Conditions

We now formalize the other conditions we will require for our results. Recall that $\mathcal{S}$ is the state space of X.

Condition 3.9 (Bounded Intensities)**.** *The intensities $\lambda_k, k = 1, \ldots, K$ of X satisfy this condition on some set Θ if there exists some constant Γ_M such that for all k and all $x \in \mathcal{S}$ we have*

$$\sup_{\theta\in\Theta}\sup_{x\in\mathcal{S}}\lambda_k(\theta, x) \leq \Gamma_M.$$

Condition 3.10. *The intensities $\lambda_k, k = 1, \ldots, K$ of X satisfy this condition on some set Θ if they are differentiable in θ on Θ and:*

i. There exists some constant Γ_m not depending on x such that for all $x \in \mathcal{S}$ and all

k we have

$$\sup_{\theta\in\Theta}\lambda_k(\theta,x)\neq 0 \Rightarrow \sup_{\theta\in\Theta}\frac{1}{\lambda_k(\theta,x)}\leq \Gamma_m.$$

That is, for a given $x\in\mathcal{S}$, either the rates $\lambda_k(\theta,x)$ are identically zero on Θ, or they must be bounded away from zero uniformly in θ and x.

ii. There exists some constant Γ' such that for all $x\in\mathcal{S}$, all k, and all $i=1,\ldots,R$ we have

$$\sup_{\theta\in\Theta}\sup_{x\in\mathcal{S}}\left|\frac{\partial}{\partial\theta_i}\lambda_k(\theta,x)\right|\leq\Gamma'.$$

The constants in Conditions 3.9 and 3.10 do not depend on k, but this is not a restriction as K and R are finite and so one can simply take maxima. We will assume for convenience that Γ_M,Γ_m, and Γ' are at least 1.

Note that if the intensities follow mass-action kinetics (Definition 1.3), part i of Condition 3.10 holds automatically: for any fixed k, if $\lambda_k(\theta,x)=\theta_k g_k(x)$ is nonzero, it is bounded from below by θ_k, since $x\in\mathcal{S}\subset\mathbb{Z}^d_{\geq 0}$. Furthermore, again assuming the intensities follow mass-action kinetics, part ii of Condition 3.10 is implied by the Bounded Intensity Condition 3.9 as long as each θ_k is strictly positive. Indeed, if Condition 3.9 holds with mass-action intensities $\lambda_k(\theta,x)=\theta_k g(x)$, the state space $\mathcal{S}$ must be compact, and the $g_k(x)$ are then bounded for $x\in\mathcal{S}$; then note that

$$\frac{\partial}{\partial\theta_i}\lambda_k(\theta,x)=\begin{cases} g_k(x) & i=k \\ 0 & \text{otherwise.}\end{cases}$$

As mentioned, we also need a few mild regularity conditions on F. Note that we do not require F to be bounded. Here and throughout we use the notation $\|x\|$, for $x\in\mathbb{R}^d$,

to denote the 1-norm, $\|x\| = \sum_{i=1}^{d} x_i$.

Condition 3.11. *The function $F : \Theta \times \mathbb{R}^d \to \mathbb{R}$ satisfies this condition if it is differentiable in θ on Θ so that:*

A. There exist constants $C_1 > 1$ and $c_1 > 1$ such that $\sup_{\theta\in\Theta} |F(\theta, x)| \leq C_1(1 + \|x\|^{c_1})$.

B. There exist constants $C_2 > 1$ and $c_2 > 1$ such that $\sup_{\theta\in\Theta} \left|\frac{\partial}{\partial\theta} F(\theta, x)\right| \leq C_2(1+\|x\|^{c_2})$.

3.2.3 Preliminary Results

Recall that $N(\theta, b) = N(\theta)$ is the number of jumps of the process through time b. Because the total rate is bounded uniformly in θ, we have the following result. It is an immediate corollary of a result we prove in Section 3.4.

Proposition 3.12. *Given the process $X_t(\theta)$ satisfying* (1.7) *and with intensities satisfying the Bounded Intensity condition 3.9 on some set Θ containing θ, for any finite t and $q \in [1, \infty)$ we have*

$$\mathbb{E}\left[\sup_{\theta\in\Theta} N(\theta)^q\right] < \infty \quad \text{and} \quad \mathbb{E}\left[\sup_{\theta\in\Theta}\sup_{s\in[0,t]} \|X_s(\theta)\|^q\right] < \infty.$$

Both results in the proposition follow because N is bounded by a Poisson random variable. See Section 3.4 for details.

The bound on the intensity also implies the existence of exponential moments of N:

Lemma 3.13. *Given the process $X_t(\theta)$ satisfying* (1.7) *and with intensities satisfying the Bounded Intensity condition 3.9 on some set Θ containing θ, for any fixed $c \in [1, \infty)$, we have the exponential moments $\mathbb{E}\left[\sup_{\theta\in\Theta} c^{N(\theta)}\right] < \infty$.*

Proof. By assumption, the total number of jumps of the process through time b is stochastically bounded uniformly in θ by a Poisson random variable $\hat{N}$ with parameter $\tilde{\Gamma} = bK\Gamma_M$, so that $\mathbb{E}\left[\sup_{\theta\in\Theta} c^{N(\theta)}\right] \leq \mathbb{E}[c^{\hat{N}}]$. Now note that for any fixed, finite c we have

$$\mathbb{E}[c^{\hat{N}}] = \sum_{m=0}^{\infty} c^m \mathbb{P}(\hat{N} = m) = \sum_{m=0}^{\infty} c^m \frac{\tilde{\Gamma}^m}{m!} e^{-\tilde{\Gamma}} = e^{-\tilde{\Gamma}} \sum_{m=0}^{\infty} \frac{(c\tilde{\Gamma})^m}{m!} = e^{-\tilde{\Gamma}} e^{c\tilde{\Gamma}} < \infty.$$

□

We will use these moment bounds in the proof of Theorem 3.19. We now turn to preliminary results concerning the differentiability of L_X.

Lemma 3.14. *Suppose we are given the process $X_t(\theta)$ satisfying* (1.7)*. Further suppose there exists a neighborhood Θ of θ such that X satisfies the Non-Interruptive condition 3.8 on Θ, and such that its intensities satisfy Conditions 3.9 and 3.10 on Θ. Let $F : \Theta \times \mathbb{R}^d \to \mathbb{R}$ be some measurable function which is differentiable in the first variable, and let $L_X(\theta) = \int_a^b F(\theta, X_s(\theta))ds$ be defined as in* (3.4)*. Then at any $\theta \in \Theta$, $L_X(\theta)$ is a.s. differentiable.*

Proof. Fix some $\theta \in \Theta$. Note that $N(\theta) < \infty$ almost surely. Furthermore, for a.e. ω there is some neighborhood of θ on which $N(\theta)$ is constant. Therefore, considering (3.5), we only require that F is differentiable in θ, which we assumed, and that the holding times $[T_{i+1}^{\theta} \wedge b - T_i^{\theta} \vee a]^+$ are also a.s. differentiable in θ. The latter will follow from a.s. differentiability of the jump times T_i^{θ}, which we now show, since the probability that a jump occurs at exactly a or b is zero.

The probability that a path has two or more jumps occurring at the same time is also

zero. Therefore, there is almost surely some $h_0(\theta,\omega) > 0$ such that for each $h \in (-h_0, h_0)$ the paths $X_t(\theta + h, \omega)$, constructed using the next reaction method, have the same embedded discrete chain through time b. Explicitly, letting $I = \{0, 1, \ldots, N(\theta)\}$, one may take

$$h_0(\theta,\omega) < \min\left(1, \frac{\min_{i\in I}[T_{i+1}(\theta) \wedge b - T_i(\theta)]}{\max_{k,i\in I} \sup_{0<\delta<1} \lambda_k(\theta+\delta, \hat{X}_i(\theta))}, \frac{[T_{N(\theta)+1}(\theta) - b]}{\max_k \sup_{0<\delta<1} \lambda_k(\theta+\delta, \hat{X}_{N(\theta)}(\theta))}\right).$$

Indeed, for some embedded state to change, some holding time must first decrease to zero; the rate at which this occurs cannot exceed the quantity given in the relevant denominator above. The quantity on the right is positive almost surely (unless all the rates begin at zero, in which case the process is unchanged for all time regardless of h and the result is trivial).

Now for $h \in (-h_0, h_0)$ the embedded path is fixed, and so the jump times T_i^θ change solely by scaling due to changes in the rates (as in the computations in Section 3.2.1). Thus the jump times T_i^θ are differentiable since the rates are differentiable.

□

We have not yet used the Non-Interruptive Condition 3.8. This additional condition will allow us to show that L_X remains continuous through a crossover of jumps. We will require this continuity in the proof of Theorem 3.19 to use the mean value theorem. We first turn to proving that the jump times of a non-interruptive processes are continuous in Lemma 3.15 before proving the continuity result regarding L_X in Lemma 3.16.

Lemma 3.15. *Suppose we are given the process $X_t(\theta)$ satisfying (1.7). Further suppose there exists a neighborhood Θ of θ such that X satisfies the Non-Interruptive condition 3.8 on Θ. Then the jump times T_i^θ are continuous in θ on Θ.*

Proof. This is clear by the derivations in Section 3.2.1, except possibly at values of θ at which the embedded chain of X changes. Because X is non-interruptive, this can only occur at a crossover of two jumps. So consider perturbing θ in a small interval around the occurrence of a crossover of two jumps i and $i+1$. As θ changes, the distance between T_i^θ and T_{i+1}^θ decreases until the two times are equal. As θ changes further, the order of the jumps switches, but even though the jump times may now denote the time of a different reaction, T_i^θ and T_{i+1}^θ still change continuously in θ. □

The Non-Interruptive condition is used crucially here, since we only need to consider what occurs at a crossover. At an interruption, the two reactions are not able to occur in either order, since the first reaction sets the rate of the second reaction to zero. This causes a jump discontinuity in one of the jump times, and in L as well.

Lemma 3.16. *Suppose we are given the process $X_t(\theta)$ satisfying* (1.7)*. Further suppose there exists a neighborhood Θ of θ such that X satisfies the Non-Interruptive condition 3.8 on Θ, and such that its intensities satisfy Conditions 3.9 and 3.10 on Θ. Let $F : \Theta \times \mathbb{R}^d \to \mathbb{R}$ be some measurable function which is differentiable in the first variable, and let $L_X(\theta) = \int_a^b F(\theta, X_s(\theta))ds$ be defined as in* (3.4)*. Then for any $\theta \in \Theta$ and for $h > 0$ such that $(\theta - h, \theta + h) \subset \Theta$, with probability $1 - O(h^2)$ we have that $L_X(\theta)$ is continuous and piecewise differentiable on $(\theta - h, \theta + h)$.*

Proof. There are two parts to the proof. First, we show that if on the interval $(\theta-h, \theta+h)$ no more than one change occurs to the embedded chain of the path of X on the interval $[a, b]$, then $L_X(\theta)$ is continuous on that interval. Second, we require that the probability of two or more such changes is $O(h^2)$.

So suppose that there is at most one change to the embedded chain on the interval $[a, b]$. Such a change can occur in one of two ways:

(i) two (or more) jump times T_i^θ cross over on the path of X through time b, which causes a chain in the path of X, or

(ii) some jump time T_i^θ enters or exits the interval $[a, b]$, which changes the states of the path that appear in L_X.

What we must show is that L_X is continuous even at these occurrences. Recall from (3.4) that

$$L_X(\theta) = \sum_{i=0}^{N} F(\theta, \hat{X}_i(\theta))[T_{i+1}^\theta \wedge b - T_i^\theta \vee a]^+ . \tag{3.9}$$

Also recall that F is continuous in θ by assumption, and that the T_i^θ are continuous by Lemma 3.15.

Suppose that (i) occurs at θ^*. Then two reactions k and ℓ occur at the same time. (The case when three or more jumps occur together is essentially the same.) Further suppose these reactions occur as the i^{th} and $(i+1)^{\text{st}}$ jumps. Then at θ^*, there is a discontinuity in $\hat{X}_i(\theta)$: from one side it is $\hat{X}_{i-1}(\theta)+\zeta_k$ and from the other it is $\hat{X}_{i-1}(\theta)+\zeta_\ell$. However, by the Non-Interruptive condition, the two reactions can occur in either order, and the net result of the two reactions is the same regardless: $\zeta_k + \zeta_\ell$ is added to the system. That is, $\hat{X}_{i+1}(\theta) \equiv X_{i-1}(\theta) + \zeta_k + \zeta_\ell$ on the whole interval, and furthermore, this crossover of jumps affects no other states of the embedded chain.

Then in the summation (3.9), any given term changes continuously except possibly the i^{th} term,

$$F(\theta, \hat{X}_i(\theta))[T_{i+1}^\theta \wedge b - T_i^\theta \vee a]^+. \tag{3.10}$$

But at θ^*, we have $T_{i+1}^{\theta^*} = T_i^{\theta^*}$, so that the term is actually zero at the point of discontinuity of $\hat{X}_i(\theta)$, and $L_X(\theta)$ remains continuous at θ^* as needed.

Now suppose that at θ^*, case (ii) occurs. Since an additional jump time appears in the interval $[a, b]$ at θ^*, the possible issue is that an extra term may show up in the summation. However, this additional jump time is either equal to a or b, so that the holding time $[T_{i+1}^{\theta} \wedge b - T_i^{\theta} \vee a]^+$ is zero. Then L_X is again continuous at θ^* as needed.

Under both cases, L_X is piecewise differentiable, because as can be seen by the derivations in Section 3.2.1, L_X is differentiable except possibly at values of θ at which the embedded chain changes, which by assumption occurs at most once.

Finally, we require that the probability of two or more changes occurring to the embedded chain of the path of X on the interval $[a, b]$ is $O(h^2)$. For a proof, we refer the reader to the second part of the proof of Appendix 5.B in [15], page 120.

□

3.2.4 Statements and Proofs of the Main Results

We now prove two results bounding the derivatives of the internal times of the reactions and the holding times of the paths, before proving the main result in Theorem 3.19.

Lemma 3.17. *Suppose we are given the process $X_t(\theta)$ satisfying* (1.7)*. Suppose there exists a neighborhood Θ of θ such that X satisfies the Non-Interruptive condition 3.8 on Θ, and such that its intensities satisfy Conditions 3.9 and 3.10 on Θ. Then for each i from 0 to $N_b(\theta)$ we have*

$$M_i := \max_k \max_{j \le i} \left| \frac{\partial}{\partial \theta} S_k^{\theta}(T_j^{\theta}) \right| \le \Gamma' b (2\Gamma_M \Gamma_m)^i,$$

where $\frac{\partial}{\partial\theta}S_k^\theta(t)$ *is as in* (3.7).

Proof. Consider (3.6) and (3.7) and recall that for each k we have $\frac{\partial}{\partial\theta}S_k^\theta(T_0^\theta)=0$. Then

$$\left|\frac{\partial}{\partial\theta}\Delta_0^\theta\right| = \left|\frac{\Delta_0^\theta}{\lambda_{k_i}(\theta,\hat{X}_0(\theta))}\frac{\partial}{\partial\theta}\lambda_{k^0}(\theta,\hat{X}_0(\theta))\right| \le \Delta_0^\theta\Gamma'\Gamma_m.$$

Then for any k,

$$\frac{\partial}{\partial\theta}S_k^\theta(T_1^\theta) = \Delta_0^\theta\frac{\partial}{\partial\theta}\lambda_k(\theta,\hat{X}_0(\theta)) + \lambda_k(\theta,\hat{X}_0(\theta))\frac{\partial}{\partial\theta}\Delta_0^\theta,$$

so that

$$M_1 = \max_k\left|\frac{\partial}{\partial\theta}S_k^\theta(T_1^\theta)\right| \le \Delta_0^\theta\Gamma' + \Gamma_M\Delta_0^\theta\Gamma'\Gamma_m \le 2\Gamma'\Gamma_m\Gamma_M\Delta_0^\theta.$$

Similarly, for a given i we have

$$\begin{aligned}\left|\frac{\partial}{\partial\theta}\Delta_i^\theta\right| &\le \left|\frac{\Delta_i^\theta}{\lambda_{k_i}(\theta,\hat{X}_i(\theta))}\frac{\partial}{\partial\theta}\lambda_{k_i}(\theta,\hat{X}_i(\theta))\right| + \left|\lambda_{k_i}(\theta,\hat{X}_i(\theta))^{-1}\frac{\partial}{\partial\theta}S_{k_i}^\theta(T_i^\theta)\right| \\ &\le \Delta_i^\theta\Gamma'\Gamma_m + \Gamma_m M_{i-1}.\end{aligned}$$

Therefore, using that

$$\frac{\partial}{\partial\theta}S_k^\theta(T_i^\theta) = \frac{\partial}{\partial\theta}S_k^\theta(T_{i-1}^\theta) + \Delta_{i-1}^\theta\frac{\partial}{\partial\theta}\lambda_k(\theta,\hat{X}_{i-1}(\theta)) + \lambda_k(\theta,\hat{X}_{i-1}(\theta))\frac{\partial}{\partial\theta}\Delta_{i-1}^\theta$$

and noticing that the M_i are nondecreasing, we see that

$$\begin{aligned} M_i &\leq M_{i-1} + \Gamma' \Delta_{i-1}^{\theta} + \Gamma_M \left| \frac{\partial}{\partial \theta} \Delta_{i-1}^{\theta} \right| \\ &\leq M_{i-1} + \Gamma' \Delta_{i-1}^{\theta} + \Gamma_M (\Delta_{i-1}^{\theta} \Gamma' \Gamma_m + \Gamma_m M_{i-2}) \\ &\leq M_{i-1} + \Gamma' \Delta_{i-1}^{\theta} + \Gamma_M (\Delta_{i-1}^{\theta} \Gamma' \Gamma_m + \Gamma_m M_{i-1}) \\ &\leq 2\Gamma_M \Gamma_m M_{i-1} + 2\Gamma' \Gamma_M \Gamma_m \Delta_{i-1}^{\theta}. \end{aligned}$$

Iterating this inequality, we see that

$$M_i \leq (2\Gamma_M \Gamma_m)^{i-1} 2\Gamma' \Gamma_M \Gamma_m \sum_{j=0}^{i-1} \Delta_j^{\theta} \leq \Gamma' b (2\Gamma_M \Gamma_m)^i.$$

□

Corollary 3.18. *Suppose we are given the process $X_t(\theta)$ satisfying* (1.7)*. Suppose there exists a set Θ containing θ such that X satisfies the Non-Interruptive condition 3.8 on Θ, and such that its intensities satisfy Conditions 3.9 and 3.10 on Θ. Then for each i from 0 to $N_T(\theta)$ we have*

$$\left| \frac{\partial}{\partial \theta} \Delta_i^{\theta} \right| \leq 2\Gamma' b \Gamma_m (2\Gamma_M \Gamma_m)^i,$$

where $\frac{\partial}{\partial \theta} \Delta_i^{\theta}$ is as in (3.6)*.*

Proof. By (3.6), Condition 3.10, and Lemma 3.17, we have that

$$\begin{aligned}\left|\frac{\partial}{\partial\theta}\Delta_i^\theta\right| &\le \left|\frac{\Delta_i^\theta}{\lambda_{k_i}(\theta,\hat{X}_i(\theta))}\frac{\partial}{\partial\theta}\lambda_{k_i}(\theta,\hat{X}_i(\theta))\right| + \left|\lambda_{k_i}(\theta,\hat{X}_i(\theta))^{-1}\frac{\partial}{\partial\theta}S_{k_i}^\theta(T_i^\theta)\right| \\ &\le b\Gamma_m\Gamma' + \Gamma_m\left|\frac{\partial}{\partial\theta}S_{k_i}^\theta(T_i^\theta)\right| \\ &\le b\Gamma_m\Gamma' + \Gamma_m\Gamma' b(2\Gamma_M\Gamma_m)^i \\ &\le 2\Gamma' b\Gamma_m(2\Gamma_M\Gamma_m)^i.\end{aligned}$$

□

Theorem 3.19. *Suppose we are given the process $X_t(\theta)$ satisfying* (1.7)*. Further suppose there exists a neighborhood Θ of θ such that X satisfies the Non-Interruptive condition 3.8 on Θ, and such that its intensities satisfy Conditions 3.9 and 3.10 on Θ. Given a measurable function $F : \Theta \times \mathbb{R}^d \to \mathbb{R}$ that satisfies Condition 3.11, we have for L_X as in* (3.4) *that* $\dfrac{d}{d\theta}\mathbb{E}[L_X(\theta)] = \mathbb{E}\left[\dfrac{d}{d\theta}L_X(\theta)\right]$.

Proof. Let $\tilde{h}$ be the infimum over h for which two or more changes occur to the embedded chain of the path of X on the interval $[a, b]$. Note that $\tilde{h} > 0$ is positive with probability 1, as in the proof of Lemma 3.14. Without loss of generality, we choose $\tilde{h}$ so that the interval $(\theta - \tilde{h}, \theta + \tilde{h})$ is contained in Θ. We must prove the middle equality in

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}[L_X(\theta)] &= \lim_{h\to 0}\mathbb{E}[h^{-1}[L_X(\theta+h) - L_X(\theta)]] \\ &= \mathbb{E}\left[\lim_{h\to 0}h^{-1}[L_X(\theta+h) - L_X(\theta)]\right] \\ &= \mathbb{E}\left[\frac{d}{d\theta}L_X(\theta)\right].\end{aligned}$$

We write

$$\begin{aligned}\mathbb{E}[h^{-1}[L_X(\theta+h)-L_X(\theta)]] =& \mathbb{E}[h^{-1}[L_X(\theta+h)-L_X(\theta)]\mathbf{1}(h<\tilde{h})] \\ &+ \mathbb{E}[h^{-1}[L_X(\theta+h)-L_X(\theta)]\mathbf{1}(h\geq\tilde{h})].\end{aligned} \tag{3.11}$$

Consider the first term. By the proof of Lemma 3.16, since at most one change occurs to the embedded chain we know that L_X is continuous and piecewise differentiable on $(\theta-\tilde{h},\theta+\tilde{h})$. By a generalized mean value theorem (e.g. [8]) we have that

$$\left|h^{-1}[L_X(\theta+h)-L_X(\theta)]\mathbf{1}(h<\tilde{h})\right| \leq \sup_{\theta\in\Theta}\left|\frac{d}{d\theta}L_X(\theta)\right|,$$

where the supremum is over those points at which the derivative exists. Since we know that $h^{-1}[L_X(\theta+h)-L_X(\theta)] \overset{a.s.}{\to} \frac{d}{d\theta}L_X(\theta)$ as $h\to 0$ by Lemma 3.14, if we can show that the supremum on the right has finite expectation, by the dominated convergence theorem we would have that $\mathbb{E}[h^{-1}[L_X(\theta+h)-L_X(\theta)]] \to \mathbb{E}[\frac{d}{d\theta}L_X(\theta)]$. We will also show that the second term in (3.11) goes to zero as $h\to 0$, which then proves the result.

Recall that

$$\begin{aligned}\left|\frac{d}{d\theta}L_X(\theta)\right| &= \left|\sum_{i=0}^{N}[T_{i+1}(\theta)\wedge b - T_i(\theta)\vee a]^+\left(\frac{\partial}{\partial\theta}F(\theta,\hat{X}_i(\theta))\right)\right. \\ &\qquad \left. + F(\theta,\hat{X}_i(\theta))\frac{\partial}{\partial\theta}[T_{i+1}(\theta)\wedge b - T_i(\theta)\vee a]^+\right| \\ &\leq \left|\sum_{i=0}^{N}[T_{i+1}(\theta)\wedge b - T_i(\theta)\vee a]^+\left(\frac{\partial}{\partial\theta}F(\theta,\hat{X}_i(\theta))\right)\right| \\ &\qquad + \left|\sum_{i=0}^{N}F(\theta,\hat{X}_i(\theta))\frac{\partial}{\partial\theta}[T_{i+1}(\theta)\wedge b - T_i(\theta)\vee a]^+\right|\end{aligned}$$

We now consider these two terms separately. By Condition 3.11 and Lemma 3.12, the

term on the left is bounded uniformly on Θ by a quantity of finite expectation:

$$\begin{aligned}
&\left| \sum_{i=0}^{N} [T_{i+1}(\theta) \wedge b - T_i(\theta) \vee a]^+ \left(\frac{\partial}{\partial \theta} F(\theta, \hat{X}_i(\theta)) \right) \right| \\
&\qquad \leq \sum_{i=0}^{N} T_{i+1}(\theta) \wedge b - T_i(\theta) \vee a]^+ \left| \frac{\partial}{\partial \theta} F(\theta, \hat{X}_i(\theta)) \right| \\
&\qquad \leq C_2 \sum_{i=0}^{N} [T_{i+1}(\theta) \wedge b - T_i(\theta) \vee a]^+ (1 + \|\hat{X}_i^\theta\|^{c_2}) \\
&\qquad \leq C_2 (b-a)(1 + \max_{i \leq N} \|\hat{X}_i^\theta\|^{c_2}) \\
&\qquad \leq C_2 (b-a)(1 + \sup_{\theta \in \Theta} \sup_{s \in [0,b]} \|X_s(\theta)\|^{c_2}).
\end{aligned}$$

For the second term, from (3.8) and our work in Lemma 3.17 we have for any i the rather crude bound

$$\left| \frac{\partial}{\partial \theta} [T_{i+1}(\theta) \wedge b - T_i(\theta) \vee a]^+ \right| \leq \sum_{j=0}^{N} \left| \frac{\partial}{\partial \theta} \Delta_j \right|.$$

Therefore,

$$\begin{aligned}
&\left| \sum_{i=0}^{N} F(\theta, \hat{X}_i(\theta)) \frac{\partial}{\partial \theta} [T_{i+1}(\theta) \wedge b - T_i(\theta) \vee a]^+ \right| \\
&\qquad \leq C_1 \sum_{i=0}^{N} (1 + \|\hat{X}_i^\theta\|^{c_1}) \left| \frac{\partial}{\partial \theta} [T_{i+1}(\theta) \wedge b - T_i(\theta) \vee a]^+ \right| \\
&\qquad \leq C_1 (1 + \max_{i \leq N} \|\hat{X}_i^\theta\|^{c_1}) \sum_{i=0}^{N} \sum_{j=0}^{N} \left| \frac{\partial}{\partial \theta} \Delta_j \right| \\
&\qquad \leq C_1 (1 + \max_{i \leq N} \|\hat{X}_i^\theta\|^{c_1}) \sum_{i=0}^{N} \sum_{j=0}^{N} 2\Gamma' T \Gamma_m^2 (2\Gamma_M \Gamma_m)^j \\
&\qquad \leq C_1 (1 + \sup_{\theta \in \Theta} \sup_{s \in [0,b]} \|X_s(\theta)\|^{c_1}) N^2 2\Gamma' T \Gamma_m (2\Gamma_M \Gamma_m)^N.
\end{aligned}$$

By Lemmas 3.12 and 3.13, and repeated applications of the Cauchy-Schwarz inequality, we see that this quantity is bounded uniformly in θ on Θ by a quantity of finite expectation as needed.

Finally, we must show that $\mathbb{E}[h^{-1}[L_X(\theta+h)-L_X(\theta)]\mathbf{1}(h \geq \tilde{h})]$ goes to zero as $h \to 0$. By using the Cauchy-Schwarz inequality, we see that

$$\mathbb{E}\left[h^{-1}[L_X(\theta+h)-L_X(\theta)]\mathbf{1}(h \geq \tilde{h})\right]^2 \leq h^{-2}\mathbb{E}\big[[L_X(\theta+h)-L_X(\theta)]^2\big]P(h \geq \tilde{h}).$$

Since by Lemma 3.16 we have $P(h \geq \tilde{h}) = O(h^2)$, and since $[L_X(\theta+h)-L_X(\theta)] \xrightarrow{a.s} 0$, we are done by the dominated convergence theorem if we can show that $[L_X(\theta+h)-L_X(\theta)]^2$ is bounded by an integrable function.

By Condition 3.11 on F, for any $\theta \in \Theta$ we have that

$$\begin{aligned}
[L_X(\theta)]^2 &= \left(\int_a^b F(\theta, X_s^M(\theta))ds\right)^2 \\
&\leq (b-a)\int_a^b \big(F(\theta, X_s^M(\theta))\big)^2 ds \\
&\leq (b-a)\int_a^b C_1^2(1+\|X_s^M(\theta)\|^{c_1})^2 ds \\
&\leq C_1^2(b-a)^2(2+2\sup_{\theta\in\Theta}\sup_{s\in[0,b]}\|X_s(\theta)\|^{2c_1}),
\end{aligned} \tag{3.12}$$

where the final line follows because $(a+b)^2 \leq 2a^2+2b^2$. This final line also has finite expectation by Lemma 3.12. Also note that the bound (3.12) is uniform, so that it holds for $|L_X(\theta+h)^2|$ as well. Then as needed,

$$|L_X(\theta+h)-L_X(\theta)|^2 \leq 2[L_X(\theta+h)]^2+2[L_X(\theta)]^2 \leq 4\sup_{\theta\in\Theta}[L_X(\theta)]^2,$$

which has finite expectation by the comment above. □

3.3 The Hybrid Pathwise Method

We now propose a new method for computing sensitivities of our CTMC models (1.7) in the form of a hybrid pathwise and likelihood ratio method. In particular, this hybrid method will be valid even for models with interruptions.

Since pathwise-only methods cannot be applied to any process with interruptions, we will first construct a process Z that approximates X. The process Z will also be a d-dimensional CTMC, and will have the same reactions as X, but Z will be non-interruptive. We will also require that the rates of Z are bounded. (Note that we are *not* requiring the original process X to be non-interruptive nor have bounded rates.) Since Z is non-interruptive with bounded intensities, by Theorem 3.19 we may use the pathwise method of Section 3.2 for computing sensitivities of

$$L_Z(\theta) := \int_a^b F(\theta, Z_s(\theta))ds. \tag{3.13}$$

Of course, estimates of sensitivities of L_Z will not in general be the same as those of L_X, as we will not have that $\mathbb{E}[L_Z(\theta)] = \mathbb{E}[L_X(\theta)]$. Therefore, we build X and Z on the same probability space by using the standard coupling, and then note, using linearity of the expectation and the derivative, that for $i = 1, \ldots, R$ we have

$$\frac{\partial}{\partial\theta_i}\mathbb{E}[L_X(\theta)] = \frac{\partial}{\partial\theta_i}\mathbb{E}[L_X(\theta) - L_Z(\theta)] + \frac{\partial}{\partial\theta_i}\mathbb{E}[L_Z(\theta)].$$

To estimate the first term on the right, we will use the likelihood ratio method. Since Z

approximates X, and since the two processes are coupled, the variance of this difference will be small. We use a pathwise method to estimate the second term.

In Section 3.3.1, we discuss the use of the likelihood ratio method on the difference of the coupled processes X and Z. Because the coupled process is a different CTMC with new intensities involving minima, we must take some care. In Section 3.3.2, we give the construction of the approximate process. In Section 3.3.3, we give the hybrid estimators.

3.3.1 The Likelihood Ratio Method for a Coupled Process

We wish to use the likelihood ratio method to estimate the sensitivity $\frac{\partial}{\partial\theta_i}\mathbb{E}[L_X(\theta) - L_Z(\theta)]$, where the processes X and Z are coupled as in Section 1.1.2. We have not yet described how to construct Z, which will have the same reactions as X but with possibly different intensities. First, we need to know what is required of the rates of Z for the likelihood ratio method to be valid on the coupled process. In particular, recalling our discussion in Section 2.2, we require that the rates of the coupled process be differentiable in θ.

Recall the standard coupling from Section 1.1.2 and let $W = (X, Z)$ be the coupled process. In particular, for each original reaction $k = 1, \ldots, K$ with reaction vector $\zeta_k \in \mathbb{Z}^d$, the coupled process has three relevant new reactions. The rates of these new reactions at a state $w = (x, z)$ are as follows, where we use λ^X and λ^Z to denote the intensities of X and Z respectively, and where we have parametrized the Λ in the obvious

way:

$$\begin{aligned}
\Lambda_{k,[1,1]}(\theta, x, z) &= \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z), \\
\Lambda_{k,[1,0]}(\theta, x, z) &= \lambda_k^X(\theta, x) - \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z), \\
\Lambda_{k,[0,1]}(\theta, x, z) &= \lambda_k^Z(\theta, x) - \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z).
\end{aligned} \tag{3.14}$$

Assuming that the rates of both X and Z are differentiable, it is easy to see that the new rates for the coupled process are piecewise differentiable. However, because these intensities involve minima of the original rates, there may be values of θ and $w = (x, z)$ where the derivative does not exist. In particular, this can occur if for some k the two rates in the minimum $\lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z)$ are equal, since at such points the left- and right-hand derivatives may be different. Unfortunately, even for some simple models this can occur on a set of paths with nonzero probability.

For example, suppose that our model follows mass-action kinetics as in Section 1.1.1. At first glance it seems to be a reasonable idea to set $\lambda_k^Z(\theta, z)$ to be some fixed, small number δ if $\lambda_k^X(\theta, z) = 0$. This choice would ensure that Z cannot have interruptions, since no reaction would ever have a rate of zero. However, this choice will be problematic when we wish to use the likelihood ratio method on (X, Z). Indeed, supposing that at some state z_0 the rate $\lambda_k^X(\theta, z)$ is zero, we have that

$$\Lambda_{k,[1,1]}(\theta, x, z_0) = \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z_0) = \theta_k g_k(x) \wedge \delta.$$

Then for any x at which $\lambda_k^X(\theta, x) = \delta$, the derivative of $\Lambda_{k,[1,1]}(\theta, x, z_0)$ from one side is $g_k(x)$ but from the other side is zero.

One way to avoid this problem is to set $\lambda_k^Z(\theta, z) = \theta_k \delta$ whenever $\lambda_k^X(\theta, z) = 0$.

Then the θ_k factors out of the minimum, or put another way, the left- and right-hand derivatives are equal. This suggests the condition we need to require in general:

Condition 3.20. *Suppose for some $k \in \{1, \ldots, K\}$ and some $x, z \in \mathbb{Z}^d$, we have that $\lambda_k^X(\theta, x) = \lambda_k^Z(\theta, z)$. Then we require for each $i = 1, \ldots, R$ that $\frac{\partial}{\partial \theta_i}\lambda_k^X(\theta, x) = \frac{\partial}{\partial \theta_i}\lambda_k^Z(\theta, z)$.*

This ensures the differentiability of the minima involved in the intensities Λ, and so ensures the differentiability of each Λ. That is, Condition 3.20 implies the second requirement in the list of Leibniz conditions given on page 33. As before, the first and third of the Leibniz conditions are more difficult to prove. As the coupled process (X, Z) is a new CTMC, the process of proving the first and third conditions is no different than before; in particular, the proof of these conditions for a large class of models, as we will give in Section 3.4.3, also applies to the coupled process.

Let the *type* $j \in \{1, 2, 3\}$ represent $[1, 1], [1, 0], [0, 1]$ respectively. Also, let $\hat{X}_i$ and $\hat{Z}_i$ be the states of X and Z respectively at the time of the i^{th} jump of the coupled process $W = (X, Z)$, taking care to note that the i^{th} jump of W is not necessarily the i^{th} jump of either X or Z. As in Section 2.2, the weighting function for the coupled process W can be found via a straightforward calculation:

$$H_i(\theta, T) = \sum_{j=1}^{3} \left[\sum_{i=0}^{N-1} \frac{\frac{\partial}{\partial \theta_i}\Lambda_{k_i,j}(\theta, \hat{X}_i, \hat{Z}_i)}{\Lambda_{k_i,j}(\theta, \hat{X}_i, \hat{Z}_i)} - \sum_{k=1}^{K} \int_0^t \frac{\partial}{\partial \theta_i}\Lambda_{k,j}(\theta, W_s) ds \right],$$

where $N = N(\theta, T)$ is the total number of jumps of W through time T. Under the

mass-action assumption, $H_i(\theta, t)$ is given by

$$H_i(\theta, T) = \sum_{j=1}^{3} \left[\frac{1}{\theta_i} \left(N_{i,j} - S_{i,j}(T) \right) \right],$$

where $N_{i,j}$ is the number of jumps of reaction i of type j by time T, and where $S_{i,j}(T) = \int_0^T \Lambda_{i,j}(\theta, X_s, Z_s) ds$ is the internal time at (real) time T of the reaction of W of reaction i and type j.

We note that Condition 3.20 also ensures absolute continuity as needed for the work of Section 2.2. That is, in the notation of that section, $P^\theta \ll P^{\theta_0}$. This follows because Condition 3.20 ensures that on some neighborhood of θ_0, for any fixed states x and z and any k, a change in θ cannot change the argmin of $\Lambda_{k,[1,1]}(\theta, x, z) = \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z)$. This fact implies that on this neighborhood a change in θ cannot change which of the rates $\Lambda_{k,[1,0]}(\theta, x, z)$ and/or $\Lambda_{k,[0,1]}(\theta, x, z)$ is zero (see (3.14)). Therefore, any path of $W = (X, Z)$ that is possible under P^θ is also possible under $\mathbb{P}^{\theta_0}$.

3.3.2 Construction of an Approximate Process

We wish to define a CTMC Z that approximates X, but such that Z satisfies the Non-Interruptive condition 3.8 and has bounded rates so that we may use a pathwise method to estimate $\frac{\partial}{\partial \theta_i} \mathbb{E}[L_Z(\theta)]$. Furthermore, we wish to define the rates of Z so that we may use the likelihood ratio method to estimate $\frac{\partial}{\partial \theta_i} \mathbb{E}[L_X(\theta) - L_Z(\theta)]$, as discussed in the previous section.

We construct such a Z explicitly in the case of mass-action kinetics. Recall that under mass-action kinetics, θ is a vector of positive rate parameters, one for each reaction. Choose some very large M. As suggested in the previous section, we may take for a

given k the intensity

$$\lambda_k^Z(\theta, z) = \begin{cases} \theta_k \delta_k & \text{if } z_i < \nu_{ki} \text{ for any } i \text{ such that } \nu_{ki} > 0 \\ \theta_k M & \text{if } \lambda_k^X(\theta, z) \geq \theta_k M \\ \lambda_k^X(\theta, z) & \text{otherwise} \end{cases} \tag{3.15}$$

for some choices of $\delta_k > 0$. Then for the most part the rates of Z are defined as they are for X, so that Z is a reasonable approximation to X. Moreover, recall that we assumed that $\lambda_k^X(\theta, x) = 0$ if $x_i < \nu_{ki}$ for some i. The first case in 3.15, therefore, prevents the process Z from having an interruption. This intensity is also bounded as required, and one can easily check that Condition 3.10 is also satisfied. Thus a pathwise method is valid for $\frac{\partial}{\partial \theta_i}\mathbb{E}[L_Z(\theta)]$ as desired. Finally, Condition 3.20 holds as is required to use the likelihood ratio method to estimate $\frac{\partial}{\partial \theta_i}\mathbb{E}[L_X(\theta) - L_Z(\theta)]$.

It will now be possible that some reaction occurs in Z even if, for some ℓ, the ℓ^{th} coordinate of Z is less than $\nu_{k\ell}$. That is, a reaction can occur in Z even if the number of molecules usually required is not present. Consequently, Z may have physically meaningless values, such as states with negative coordinates. Even though biologically speaking this is nonsensical, mathematically it is perfectly valid that the CTMC Z leaves the positive orthant. In fact, allowing Z to reach these states is crucial to the hybrid method, as they are what allow the method to be non-interruptive. In implementation, we simply must take care to reasonably define the intensities of Z on these physically meaningless states of $\mathbb{Z}^d$.

Concerning the choice of the δ_k: one should take δ_k to be small, but not too small. While the best choice for the δ_k may be model-dependent, numerical experiments show

that $\delta_k \equiv 1$ is a reasonable choice. If δ_k is too large, the process Z may cease to be a good approximation of X, which will cause the variance of the likelihood ratio estimate of $\frac{\partial}{\partial \theta_i}\mathbb{E}[L_X(\theta) - L_Z(\theta)]$ to be large. On the other hand, making δ_k is too small makes it very rare that the process Z makes a jump that the process X cannot make. As $L_Z(\theta)$ may be significantly different on such paths, this may increase the variance of the pathwise estimate of $\frac{\partial}{\partial \theta_i}\mathbb{E}[L_Z(\theta)]$.

Concerning the choice of M: one should take M very large. Indeed, as will be discussed at the beginning of Section 3.5, we do not believe that the bounded rates are necessary for validity of the pathwise method of Section 3.2, though thus far we have been unable to prove this. Note, however, that this restriction does not change the processes X for which the hybrid method is valid; it only changes the way in which we must construct the approximate process Z.

In the case of more general intensities, as discussed at the beginning of this section, we must require that Z satisfy all of the following on some neighborhood Θ of θ:

1. Z satisfies the Non-interruptive condition 3.8,
2. the intensities of Z satisfy the bounded rate condition 3.9,
3. the intensities of Z satisfy the additional conditions 3.10, and
4. the intensities of Z are defined such that the differentiability condition 3.20 is satisfied.

As already discussed, 1–4 are satisfied for (3.15).

Remark 3.1. *For some processes in which the network structure is apparent, one may be able to improve the performance of the method by taking advantage of this structure. For*

example, if a certain reaction can never be interrupted, there is no problem in allowing that rate to be zero for Z. For example, see the model in Section 3.6.2.

3.3.3 The Hybrid Pathwise Estimator

Recall that we have the equation

$$\frac{\partial}{\partial\theta_i}\mathbb{E}[L_X(\theta)] = \frac{\partial}{\partial\theta_i}\mathbb{E}[L_X(\theta) - L_Z(\theta)] + \frac{\partial}{\partial\theta_i}\mathbb{E}[L_Z(\theta)].$$

We have now constructed Z so that a pathwise method is valid for computing the second term on the right, as proved in Section 3.2.4. Furthermore, we have coupled the processes X and Z so that the likelihood ratio method is valid on the first term on the right, as will be proved for our biochemical models of interest in Section 3.4.3.

We can therefore estimate the sensitivity $\frac{\partial}{\partial\theta_i}\mathbb{E}[L_X(\theta)]$ using a sum of two independent Monte Carlo estimators

$$\hat{M}^{LR}_{n_1} + \hat{M}^{PW}_{n_2},$$

where

$$\hat{M}^{LR}_{n_1} = \frac{1}{n_1}\sum_{m=1}^{n_1} M^{LR}_m \quad \text{and} \quad \hat{M}^{PW}_{n_2} = \frac{1}{n_2}\sum_{m=1}^{n_2} M^{PW}_m$$

with

$$M^{LR} = \int_a^b \left(\frac{\partial}{\partial\theta_i}F(\theta, X_s) - \frac{\partial}{\partial\theta_i}F(\theta, Z_s)\right) ds + (L_X(\theta) - L_Z(\theta))H_i(\theta, b)$$

and

$$M^{PW} = \frac{\partial}{\partial \theta_i} L_Z(\theta),$$

where $H_i(\theta, b)$ is the likelihood ratio weight of the coupled process $W = (X, Z)$ through time b. Thus $\hat{M}_{n_1}^{LR}$ is computed as discussed in Section 3.3.1, and $\hat{M}_{n_2}^{PW}$ is computed by the algorithm given in Section 3.2.1, though of course using the process Z rather than X.

It is important that the paths for the two estimators be computed independently. This is because the probability measures underlying the two methods are not the same: the measure used in the likelihood ratio estimate is θ-dependent, whereas the measure in the pathwise estimate is not.

A Note on Two Related Estimators

In many cases, the sensitivity we wish to estimate is of the form $\frac{\partial}{\partial \theta_i}\mathbb{E}[f(X_T(\theta))]$. THat is, f is a function of the path at some fixed time T rather than the entire path through T. In this case, to use the hybrid method we begin from the equation

$$\frac{\partial}{\partial \theta_i}\mathbb{E}[f(X_T(\theta))] = \frac{\partial}{\partial \theta_i}\mathbb{E}[f(X_T) - f(Z_T)] + \frac{\partial}{\partial \theta_i}\mathbb{E}[f(Z_T(\theta))], \tag{3.16}$$

where the non-interruptive process Z is constructed as in Section 3.3.2. As in the general case, we use the likelihood ratio method on the coupled difference, though note that the estimator is then simpler, since the function f is no longer a path functional and does not depend explicitly on θ. We cannot immediately use a pathwise estimator on the

second term, since $\frac{\partial}{\partial\theta_i} f(Z_T(\theta))$ is zero wherever it exists. Instead, we use one of the two method which we introduced in Section 2.3.

1. **The Dynkin Hybrid Method**. This method, in addition to using the likelihood ratio method on the first term in 3.16, uses Dynkin's formula 1.6 to estimate the second term. That is, since

$$\mathbb{E}[f(Z_T(\theta))] = \mathbb{E}[f(Z_0(\theta))] + \mathbb{E}\left[\int_0^T (\mathcal{A}_Z^\theta f)(Z_s(\theta))ds\right],$$

   we use the function $L_Z(\theta)$ as in (3.4) by taking $a = 0$, $b = T$, and $F = \mathcal{A}_Z^\theta f$, where $\mathcal{A}_Z^\theta$ is the generator of the approximate process Z. Then we can compute $\frac{\partial}{\partial\theta_i}\mathbb{E}[f(Z_T(\theta))]$ by using Section 3.2. The expression of the derivatives is as in (3.5) through (3.8), though of course with the process Z in place of X.

2. **The Regularized Pathwise Derivative (RPD) Hybrid Method**. In addition to using the likelihood ratio method on the first term in 3.16, we can use the RPD method [37] on the second term. Note that

$$f(Z_T(\theta)) \approx \frac{1}{2w}\int_{T-w}^{T+w} f(Z_s(\theta))ds,$$

   so that we may use the function $L_Z(\theta)$ as given in (3.4) by taking $a = T - w$, $b = T + w$, and $F = \frac{1}{2w}f$ to obtain the random variable on the right-hand side. Then because Z is non-interruptive, we can use the method of Section 3.2 with the time-averaged random variable to estimate the desired sensitivity

$$\frac{\partial}{\partial\theta_i}\mathbb{E}[f(Z_T(\theta))] \approx \frac{1}{2w}\mathbb{E}\left[\frac{\partial}{\partial\theta_i}\int_{T-w}^{T+w} f(Z_s(\theta))ds\right].$$

The expression of the derivatives is then also as in (3.5) through (3.8), except using the process Z.

While the RPD hybrid method will produce a biased estimate of the sensitivity $\frac{\partial}{\partial\theta_i}f(Z_T(\theta))$, and so also of $\frac{\partial}{\partial\theta_i}f(X_T(\theta))$, the bias is controllable in the "window" size w. For many models we expect these RPD hybrid estimates to have a lower variance than the estimates obtained by using the Dynkin hybrid method, so that if a small bias is acceptable, the RPD hybrid method may be preferred for reasons of efficiency. In particular, a larger w increases the bias but decreases the variance of the estimates.

Take care to note that unless the process X satisfies the Non-Interruptive condition 3.8, the RPD method alone is *not* a valid way to estimate $\frac{\partial}{\partial\theta_i}\mathbb{E}[f(X_T(\theta))]$. That is, if interruptions may occur in the paths of X then

$$\lim_{w\to 0}\frac{1}{2w}\mathbb{E}\left[\frac{\partial}{\partial\theta_i}\int_{T-w}^{T+w}f(X_s(\theta))ds\right] \neq \frac{\partial}{\partial\theta_i}\mathbb{E}[f(X_T(\theta))].$$

See Section 3.6 for examples demonstrating this fact. As discussed in Example 3.7, the Dynkin pathwise method alone is also not valid for estimating sensitivities of models with interruptions. Instead, we must in general use a hybrid method with the approximate process Z as presented in this section.

3.4 Processes With at Most Linear Growth

In this section, we give results showing that the likelihood ratio portion of the hybrid method is valid for a large class of CTMC models, including our biochemical CTMC models of interest. Recall that the validity of the pathwise portion of the hybrid method

was proved in Section 3.2. To prove the results of this section, we will first establish moment bounds on the CTMCs X and Z. These bounds will also allow us to prove a related limiting argument in Section 3.5.

The recent papers [19] and [31] present two different proofs that, under certain conditions on overall growth, the CTMC X has finite moments at finite times. Though the proofs are different, both use Dynkin's formula. We present here a different proof of this moment result, utilizing more basic principles. The main idea is to stochastically bound X by a simple process with linear growth, which is easily seen to have finite moments.

Recall that our CTMC models, as in (1.7), satisfy

$$X_t(\theta) = X_0(\theta) + \sum_{k=1}^{K} Y_k \left(\int_0^t \lambda_k(\theta, X_s(\theta)) ds \right) \zeta_k,$$

and that we take $X_0(\theta) \equiv X_0$ fixed.

We use $\|x\|$ to denote the 1-norm on $\mathbb{Z}^d$; that is, $\|x\| = \sum_{i=1}^{d} |x_i|$. We also define $\mathcal{R}_1$ for those indices in $\{1, \ldots, K\}$ such that $\mathbf{1} \cdot \zeta_k > 0$, where $\mathbf{1}$ is the vector of all ones, and $\mathcal{R}_2$ for $\{1, \ldots, K\} \setminus \mathcal{R}_1$. That is, $\mathcal{R}_1$ contains the indices of those reactions that increase the total population $\|X\|$, while reactions in $\mathcal{R}_2$ either decrease the total population or leave it unchanged. For convenience we will assume that $|\mathcal{R}_1| \geq 1$, as otherwise the process is bounded.

We require the following condition on the intensities of the model X.

Condition 3.21 (Linear Growth)**.** *We say the intensities of the CTMC $X_t(\theta)$ as in* (1.7) *satisfy the linear growth condition on the set $\Theta \subset \mathbb{R}^R$ if there exists some constant*

K_1 such that for all $x \in \mathbb{Z}^d_{\geq 0}$

$$\sup_{\theta \in \Theta} \sum_{k \in \mathcal{R}_1} \lambda_k(\theta, x) \leq K_1(1 + \|x\|),$$

and there also exist constants $\ell \geq 1$ and K_2 such that for all $x \in \mathbb{Z}^d_{\geq 0}$

$$\sup_{\theta \in \Theta} \sum_{k \in \mathcal{R}_2} \lambda_k(\theta, x) \leq K_2(1 + \|x\|^\ell).$$

The second statement requires all of the intensities to be at most polynomial in $\|x\|$. The first statement requires that any reaction that adds to the total population is actually at most linear. This is the crucial assumption, but it is reasonable, as otherwise there may be the possibility of explosion in finite time. It is automatically satisfied if the intensities or the process are bounded, such as when some conservation relation holds. It is also automatically satisfied for any mass-action network of chemical reactions that are each at most binary, which is the vast majority of physically relevant models. See the discussion following Definition 1.3.

The inclusion of the supremum over θ is also not an overly strict requirement. Indeed, the supremum is implied if Θ is compact and the intensities are continuous on Θ.

3.4.1 A Bounding Linear Growth Process

We first prove Lemma 3.22, which shows the existence of moments of a linear growth process and a related process. We will then use these processes to bound certain quantities from our original process X in Lemma 3.23.

Lemma 3.22. *Let $\tilde{K}_1$, $\tilde{K}_2$, and ℓ be constants and let Y_1, Y_2 be independent unit-rate*

Poisson processes. Consider the one dimensional birth processes satisfying

$$B_t = 1 + Y_1\left(\int_0^t \tilde{K}_1 B_s ds\right) \quad \textit{and} \quad C_t = Y_2\left(\int_0^t \tilde{K}_2 B_s^\ell ds\right),$$

noting that the process B is in the definition of the process C. Then for fixed $t \geq 0$, $m \geq 1$, and $q \in [1, \infty)$ we have that $\mathbb{P}(B_t \geq m) = (1 - e^{-t\tilde{K}_1})^{m-1} \to 0$ as $m \to \infty$, and moreover the moments $\mathbb{E}[B_t^q]$ and $\mathbb{E}[C_t^q]$ are finite.

Proof. A linear birth process has a negative binomial distribution; see for example [21]. The distribution of the process B is negative binomial with parameters 1 and $e^{-\tilde{K}_1 t}$, which is equivalent to the geometric distribution with parameter $e^{-\tilde{K}_1 t}$. In particular, for $k \geq 0$, we know $\mathbb{P}(B_t = k) = e^{-t\tilde{K}_1}(1 - e^{-t\tilde{K}_1})^{k-1}$. A simple geometric series gives that $\mathbb{P}(B_t \geq m) = (1 - e^{-t\tilde{K}_1})^{m-1} \to 0$ as $m \to \infty$ as needed. We also have the existence of moments of B:

$$\mathbb{E}[B_t^q] = \sum_{k=1}^{\infty} k^q \mathbb{P}(B_t = k) = \frac{e^{-t\tilde{K}_1}}{1 - e^{-t\tilde{K}_1}} \sum_{k=1}^{\infty} k^q (1 - e^{-t\tilde{K}_1})^k$$

which is finite by another geometric series argument. To prove that C also has finite

moments, note that B is increasing, so that

$$\begin{aligned}
\mathbb{E}[C_t^p] &= \mathbb{E}\left[\left(Y_2\left(\int_0^t \tilde{K}_2 B_s^\ell ds\right)\right)^q\right] \\
&\leq \mathbb{E}\left[\left(Y_2\left(\int_0^t \tilde{K}_2 B_t^\ell ds\right)\right)^q\right] \\
&= \mathbb{E}\left[\left(Y_2(t\tilde{K}_2 B_t^\ell)\right)^q\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(Y_2(t\tilde{K}_2 B_t^\ell)\right)^q \middle| B_t\right]\right] \\
&\leq \mathbb{E}\left[C_q \sum_{i=1}^{q} (t\tilde{K}_2 B_t^\ell)^i\right]
\end{aligned}$$

where C_q is constant. This final line follows since, with B_t fixed, $Y_2(t\tilde{K}_2 B_t^\ell)$ is Poisson distributed with rate $t\tilde{K}_2 B_t^\ell$, and the qth moment of a $Poisson(\lambda)$ random variable is bounded by $C_q \sum_{i=1}^q \lambda^i$, where C_q is a constant. Finiteness now follows from the existence of moments for B. □

We return to the process X satisfying the linear growth Condition 3.21. Recall that the set $\mathcal{R}_1$ contains the indices of those reactions that increase the total population $\|X\|$, while reactions in $\mathcal{R}_2$ either decrease the total population or leave it unchanged. We now construct the process X and the processes B and C on the same probability space as follows.

To construct X for some given θ, we will use one counting process for the timing of the reactions in $\mathcal{R}_1$, and another for the timing of the $\mathcal{R}_2$ reactions. In particular, using

independent, unit-rate Poisson processes Y_1 and Y_2, we define

$$\begin{aligned} X_t(\theta) &= X_0 + \sum_{k\in\mathcal{R}_1} \zeta_k \int_0^t \mathbf{1}\left\{u_{R_1^\theta(s-)} \in E_k^1(\theta, X_{s-}(\theta))\right\} dR_1^\theta(s) \\ &\quad + \sum_{k\in\mathcal{R}_2} \zeta_k \int_0^t \mathbf{1}\left\{u_{R_2^\theta(s-)} \in E_k^2(\theta, X_{s-}(\theta)\right\} dR_2^\theta(s), \text{ where} \\ R_1^\theta(t) &= Y_1\left(\int_0^t \sum_{k\in\mathcal{R}^1} \lambda_k(\theta, X_s(\theta))ds\right), \text{ and} \\ R_2^\theta(t) &= Y_2\left(\int_0^t \sum_{k\in\mathcal{R}^2} \lambda_k(\theta, X_s(\theta))ds\right). \end{aligned} \tag{3.17}$$

The partitioning intervals $E_k^1(\theta, x)$ and $E_k^2(\theta, x)$ are defined analogously to (1.14). The counting processes R_1^θ and R_2^θ determine the jump times of X. If R_1^θ jumps at time s, the process X will jump by some reaction in $\mathcal{R}_1$; to decide which one, the "binning" method is used (see Section 1.2.1). Similarly, R_2^θ corresponds to jumps of reactions in $\mathcal{R}_2$. This construction therefore combines elements of both the NRM and SSA methods (see Section 1.2.1); all three produce a processes with distribution equal to that of the solution to the random time change equation of our CTMC model (1.2).

Let $c = \max_{k\in\mathcal{R}_1} \mathbf{1}\cdot\zeta_k$ and note that $\|x+\zeta_k\| \le \|x\| + c$ for all $k \in \mathcal{R}_1$ and $x \in \mathbb{Z}^d_{\ge 0}$ (recall that we are using the 1-norm). Note that as we assumed that $\mathcal{R}_1$ is not empty, we have $c \ge 1$. Also recall that for $k \in \mathcal{R}_2$ we have $\|x+\zeta_k\| \le \|x\|$. Then for any $t \ge 0$,

$$\|X_t(\theta)\| \le \|X_0\| + cR_1^\theta(t). \tag{3.18}$$

Lemma 3.23. *Suppose for θ in some set Θ we are given the processes $\{X_t(\theta)\}$ satisfying* (3.17)*, and with intensities satisfying the linear growth Condition 3.21 on Θ with*

constants K_1, K_2, *and* ℓ. *Define*

$$B_t = 1 + Y_1\left(\int_0^t \tilde{K}_1 B_s ds\right) \quad \textit{and} \quad C_t = Y_2\left(\int_0^t \tilde{K}_2 B_s^\ell ds\right),$$

where the Poisson processes Y_1 *and* Y_2 *are the same processes used in the construction of* X, *and where*

$$\tilde{K}_1 = cK_1(2 + \|X_0\|) \quad \textit{and} \quad \tilde{K}_2 = c^\ell K_2(2 + \|X_0\|)^\ell.$$

Then for any $t \in [0, \infty)$ *and any* $\theta \in \Theta$ *we have*

$$R_1^\theta(t) \le B_t \quad \textit{and} \quad R_2^\theta(t) \le C_t.$$

Furthermore, for any $t \ge 0$, $m \ge 1$, *and* $q \in [1, \infty)$ *we have that* $\mathbb{E}\left[\sup_{\theta\in\Theta}[R_1^\theta(t)]^q\right] < \infty$ *and that* $\mathbb{P}(R_1^\theta(t) \ge m) \to 0$ *uniformly on* Θ.

Proof. Fix some $\theta \in \Theta$. By the linear growth Condition 3.21, the intensity of R_1^θ at time zero is no more than $K_1(1 + \|X_0\|)$, which is bounded by the intensity of B at time zero, $\tilde{K}_1 = cK_1(2 + \|X_0\|)$. The processes B and R_1^θ are both increasing, and the intensity of B is also increasing. Also, since R_1^θ and B are both constructed from Y_1, the first jump times of both are constructed with the same unit exponential random variable. Thus the first jump time of T_1^θ of the process R_1^θ can be no earlier than the first jump time of B. Therefore, we have that $R_1^\theta(t) \le B_t$ for $t \le T_1^\theta$.

So suppose that $R_1^\theta(t) \le B_t$ for $t \le T_n^\theta$, where T_n^θ is the n^{th} jump time of R_1^θ. For any t in this range, we then have by (3.18) and the linear growth Condition 3.21 that

the rate of R_1^θ is bounded by

$$\begin{aligned} K_1(1 + \|X_t(\theta)\|) &\leq K_1(1 + \|X_0\| + cR_1^\theta(t)) \\ &\leq K_1(1 + \|X_0\| + cB_t) \\ &\leq cK_1(2 + \|X_0\|)B_t = \tilde{K}_1 B_t. \end{aligned}$$

The final expression is the rate of B at time t. Since this holds at $t = T_n^\theta$, using the same reasoning as for the first jump, we have that $R_1^\theta(t) \leq B_t$ for $t \leq T_{n+1}^\theta$ as well. Then by induction the process B stochastically dominates the process R_1^θ at all times $t \geq 0$.

To see that $R_2^\theta(t) \leq C_t$, note that the rate of R_2^θ at time $t \geq 0$ is similarly bounded by

$$\begin{aligned} K_2(1 + \|X_t(\theta)\|^\ell) &\leq K_2(1 + \|X_t(\theta)\|)^\ell \\ &\leq K_2(1 + \|X_0\| + cB_t)^\ell \\ &\leq c^\ell K_2(2 + \|X_0\|)^\ell B_t^\ell = \tilde{K}_2 B_t^\ell, \end{aligned}$$

which is the rate of C at time t. The last two statements in the lemma now follow by Lemma 3.22, since $R_1^\theta(t) \leq B_t$. □

3.4.2 Proof of Finite Moments

Corollary 3.24. *Given the process $X_t(\theta)$ satisfying (1.7) and with intensities satisfying the linear growth Condition 3.21 on some set Θ containing θ, define $N_t(\theta)$ to be the total number of jumps of $X(\theta)$ by time t. Then for any $q \in [1, \infty)$ and any $t \in [0, \infty)$ we have the moments*

1. $\mathbb{E}\left[\sup_{\theta\in\Theta}[N_t(\theta)]^q\right] < \infty$,

2. $\mathbb{E}\left[\sup_{\theta\in\Theta}\sup_{s\in[0,t]}\|X_s(\theta)\|^q\right] < \infty$,

3. $\mathbb{E}\left[\sup_{\theta\in\Theta}\sup_{s\in[0,t]}[\lambda_k(\theta, X_s(\theta))]^q\right] < \infty$ *for each* $k = 1, \ldots, K$, *and we have*

4. $\mathbb{P}\left(\sup_{\theta\in\Theta}\sup_{s\in[0,t]}\|X_s(\theta)\| > m\right) \to 0$ *as* $m \to \infty$.

5. $\mathbb{P}\left(\sup_{\theta\in\Theta}\sup_{s\in[0,t]}\lambda_k(\theta, X_s(\theta)) > m\right) \to 0$ *as* $m \to \infty$ *for each* $k = 1, \ldots, K$.

Proof. Define X, B, and C as in Lemma 3.23. Then by Lemma 3.22, we have that $\mathbb{E}[B_t^q]$ and $\mathbb{E}[C_t^q]$ are finite. By Lemma 3.23, $\mathbb{E}[R_1^\theta(t)]^q \leq \mathbb{E}[B_t^q] < \infty$ and similarly for R_2^θ. Since $N_t(\theta) = R_1^\theta(t) + R_2^\theta(t)$, we have that $\sup_{\theta\in\Theta} N_t(\theta) \leq B_t + C_t$ also has finite moments, as as the mixed terms in the expansion of the right hand side may be bounded using the Cauchy-Schwarz inequality.

For the second claim, using Lemma 3.23, the fact that since R_1^θ is increasing, and the bound (3.18), we have that

$$\sup_{s\in[0,t]}\|X_s(\theta)\| \leq \|X_0\| + cR_1^\theta(t) \leq \|X_0\| + cB_t. \tag{3.19}$$

This bound, which has finite moments, holds uniformly on Θ. This also implies that

$$\begin{aligned}\mathbb{P}(\sup_{\theta\in\Theta}\sup_{s\in[0,t]}\|X_s(\theta)\| > m) &\leq \mathbb{P}(\|X_0\| + cB_t > m) \\ &= \mathbb{P}\left(B_t > \frac{m - \|X_0\|}{c}\right)\end{aligned}$$

which goes to 0 as $m \to \infty$ by Lemma 3.22.

Recall that $a \vee b = \max(a, b)$. The third and fifth claims now follow since for all $k = 1, \ldots, K$ and any $\theta \in \Theta$ we have

$$\begin{aligned}\sup_{s\in[0,t]} \lambda_k(\theta, X_s(\theta)) &\leq \sup_{s\in[0,t]} (K_1 \vee K_2)(1 + \|X_s(\theta)\|^\ell) \\ &\leq (K_1 \vee K_2)(1 + \sup_{s\in[0,t]} \|X_s(\theta)\|^\ell).\end{aligned}$$

□

Remark 3.2. *We wish to point out that the results of this section also hold for the process Z approximating X as constructed in Section 3.3.2. Even though the process Z may end up outside of the positive orthant $\mathbb{Z}^d_{\geq 0}$, the intensities of those reactions that push Z further from the positive orthant are bounded by construction. We do not give the details here, but the process Z can be similarly bounded in terms of the linear growth process B and the auxiliary process C. Thus, we will say that the process Z satisfies the linear growth Condition 3.21 even though its state space lies outside of $\mathbb{Z}^d_{\geq 0}$.*

3.4.3 Proof of the Likelihood Ratio Method

In Section 2.2, we showed that the likelihood ratio sensitivity method is valid as long as the derivative and expectation can be exchanged. As discussed on page 33, however, in the setting of CTMCs the validity of this exchange is nontrivial to prove, and many treatments of the method in the literature either assume the exchange is valid or assume that the intensities are bounded.

Theorem 3.27 below shows explicitly that this exchange of the derivative and expectation is valid for nearly any CTMC that would arise from a biochemical system. In

addition to the mild regularity conditions on the intensities described below, we will require that the intensities satisfy the linear growth condition 3.21. As we discussed when we introduced this condition, any system following mass-action kinetics with at most binary reactions (which is most any physically relevant system) satisfies this condition. Furthermore, Theorem 3.27 proves that for these systems we may also use the likelihood ratio method on the coupled process $W = (X, Z)$ in the hybrid method of Section 3.3.3 to estimate $\frac{\partial}{\partial \theta_i}\mathbb{E}[L_X(\theta) - L_Z(\theta)]$. This is discussed further below.

Recall that $\mathcal{S}$ is the state space of X. We will require the following condition on the intensities of X:

Condition 3.25. *The intensities $\lambda_k, k = 1, \ldots, K$ of X satisfy this condition on $\Theta \subset \mathbb{R}^R$ if each λ_k is differentiable in θ_i for all $i = 1, \ldots, R$, and:*

i. There exist constants p_1 and C_1 such that for all k and all $x \in \mathcal{S}$ we have

$$\sup_{\theta\in\Theta} \lambda_k(\theta, x) \neq 0 \Rightarrow \sup_{\theta\in\Theta} \frac{1}{\lambda_k(\theta, x)} \leq C_1(1 + \|x\|^{p_1}).$$

That is, for a fixed x, if the rates $\lambda_k(\theta, x)$ are not identically zero on Θ, they must be bounded away from zero.

ii. There exist constants p_2 and C_2 such that for all k, all $x \in \mathcal{S}$, and all $i = 1, \ldots, R$,

$$\left| \sup_{\theta\in\Theta} \frac{\partial}{\partial \theta_i} \lambda_k(\theta, x) \right| \leq C_2(1 + \|x\|^{p_2}).$$

Suppose X satisfies mass-action kinetics as in (1.8) and the linear growth condition 3.21. Then as discussed in relation to Condition 3.10, the intensities of X satisfy part

i of this condition. Part ii is satisfied since $\frac{\partial}{\partial\theta_i}\lambda_k(\theta,x) = g_k(x)$, which is polynomial in the coordinates of x. Also note that this condition is a weaker version of Condition 3.10. Then the approximate process Z from the hybrid method also satisfies Condition 3.25.

We also require the following regularity condition on the functional f:

Condition 3.26. *The $\mathbb{R}$-valued function f of θ and of paths $X_{t\in[0,T]}$ through some time T satisfies this condition on Θ if $\frac{\partial}{\partial\theta_i}f(\theta,\cdot)$ exist on Θ and*

A. there exist constants $p_A > 1$ and $C_A > 1$ such that $|f(\theta, X_{t\in[0,T]})| \leq C_A(1 + \sup_{t\in[0,T]}\|X_t\|^{p_A})$ for all $\theta\in\Theta$, and

B. there exist constants $p_B > 1$ and $C_B > 1$ such that for all i we have $|\frac{\partial}{\partial\theta_i}f(\theta, X_{t\in[0,T]})| \leq C_B(1 + \sup_{t\in[0,T]}\|X_t\|^{p_B})$ for all $\theta\in\Theta$.

Theorem 3.27. *Suppose we are given the process $X(\theta)$ satisfying (1.7) and with intensities satisfying the linear growth Condition 3.21 and Condition 3.25 on some neighborhood $\Theta\subset\mathbb{R}^R$ of θ. Also suppose f satisfies Condition 3.26 on Θ. Then*

$$\frac{\partial}{\partial\theta_i}\mathbb{E}^{P^\theta}[f(\theta, X_{t\in[0,T]})] = \mathbb{E}^{P^\theta}\left[\frac{\partial}{\partial\theta_i}f(\theta, X_{t\in[0,T]}) + f(\theta, X_{t\in[0,T]})H_i(\theta,T)\right],$$

where $H_i(\theta,T)$ is as in (2.13).

Before proving this theorem, we show that it implies the validity of the likelihood ratio estimate used in the hybrid method on the coupled process $W = (X,Z)$ given that the original process X satisfies mass-action kinetics and the linear growth Condition 3.21. Under these conditions on X, as already discussed both X and Z satisfy both Conditions 3.21 and 3.25. Consider the coupled process $W = (X,Z)$ of Section 3.3.1.

The intensities of W are given by

$$\Lambda_{k,[1,1]}(\theta, x, z) = \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z),$$
$$\Lambda_{k,[1,0]}(\theta, x, z) = \lambda_k^X(\theta, x) - \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z),$$
$$\Lambda_{k,[0,1]}(\theta, x, z) = \lambda_k^Z(\theta, x) - \lambda_k^X(\theta, x) \wedge \lambda_k^Z(\theta, z).$$

Since these intensities are bounded by the maximum of the original intensities $\lambda_k^X(\theta, x)$ and $\lambda_k^Z(\theta, z)$, the rates of W satisfy the linear growth Condition 3.21 as well. To see that they also satisfy Condition 3.25, note that the derivatives of the rates of W are similarly bounded by the derivatives of $\lambda_k^X(\theta, x)$ and $\lambda_k^Z(\theta, z)$, so that part ii holds. Part i holds as well. To see this, recall that under mass-action kinetics, by our construction of Z the θ_k factors out of each rate of W. The remaining factor does not depend on θ and is discrete-valued. Thus the relevant hypotheses of Theorem 3.27 hold for W.

Finally, since we wish to estimate $\frac{\partial}{\partial \theta_i}\mathbb{E}[L_X(\theta) - L_Z(\theta)]$ using the likelihood ratio method, we define $f(\theta, W_{t\in[0,b]}) = \int_a^b F(\theta, X_s)ds - \int_a^b F(\theta, Z_s)ds$. We must show that $f(\theta, W_{t\in[0,b]})$ satisfies Condition 3.26. This will follow because we chose F to satisfy condition 3.11. First note that $\|W\| = \|X\| + \|Z\|$, and that for any $c \geq 1$ we therefore

have that $\|W\|^c \geq \|X\|^c + \|Z\|^c$. Now, for each $\theta \in \Theta$,

$$\begin{aligned}
|f(\theta, W_{t\in[0,b]})| &\leq \int_a^b |F(\theta, X_s)|ds + \int_a^b |F(\theta, Z_s)|ds \\
&\leq \int_a^b C_1(1 + \|X_s\|^{c_1})ds + \int_a^b C_1(1 + \|Z_s\|^{c_1})ds \\
&\leq \int_a^b C_1(2 + \|X_s\|^{c_1} + \|Z_s\|^{c_1})ds \qquad (3.20) \\
&\leq (b-a)C_1(2 + \sup_{s\in[0,b]} \|W_s\|^{c_1}) \\
&\leq 2(b-a)C_1(1 + \sup_{s\in[0,b]} \|W_s\|^{c_1}).
\end{aligned}$$

Using the same argument as above with $\frac{\partial}{\partial\theta_i} f(\theta, W_{t\in[0,b]})$, one shows that f satisfies part *ii* of Condition 3.26 as well.

Proof of Theorem 3.27. One way to prove this result would be to show that the three Leibniz conditions from page 33 hold. However, it is quite difficult in practice to find a Q-integrable function that bounds $\frac{\partial}{\partial\theta} G(\theta, T)$ uniformly on Θ as needed for the third condition, where $G(\theta)$ was defined as the density $\frac{dP^\theta}{dQ}$ in Section 2.2. Therefore, we prove the result in a different manner.

Note that

$$G(\theta, T) = \exp\left(\sum_{i=0}^{N-1} \log \lambda_{k_i}(\hat{X}_i) - \sum_{k=1}^{K} \int_0^T (\lambda_k(X_s) - 1)ds\right)$$

and

$$H_i(\theta, T) = \frac{\partial}{\partial\theta_i} \log G(\theta, T) = \sum_{\ell=0}^{N-1} \frac{\frac{\partial}{\partial\theta_i}\lambda_{k_\ell}(\theta, \hat{X}_\ell)}{\lambda_{k_\ell}(\theta, \hat{X}_\ell)} - \sum_{k=1}^{K} \int_0^T \lambda_k(\theta, X_s)ds$$

so that unless some $\lambda_{k_\ell}(\theta, \hat{X}_\ell) = 0$, which has probability zero under P^θ, we have by

Condition 3.25 that

$$\begin{aligned}\sup_{\theta\in\Theta}|H_i(\theta,T)| \leq NC_1(1+\sup_{t\in[0,T]}\|X_s\|^{p_1})C_2(1+\sup_{t\in[0,T]}\|X_s\|^{p_2}) \\ + KT\max_k \sup_{\theta\in\Theta}\sup_{s\in[0,t]}\lambda_k(\theta,X_s),\end{aligned} \tag{3.21}$$

which has finite moments by Corollary 3.24 (use the Cauchy-Schwarz inequality on the first term).

Define $N_k(t)$ to be the number of times the k^{th} reaction has occurred in the path X_T by time T. For brevity in notation, we will let $\vec{N} := \vec{N}(T) \in \mathbb{Z}^K_{\geq 0}$ such that the k^{th} coordinate of $\vec{N}(T)$ is given by $N_k(t)$. Then

$$\begin{aligned}\mathbb{E}^{P^\theta}[f(\theta,X_{t\in[0,T]})] &= \mathbb{E}^Q\left[f(\theta,X_{t\in[0,T]})G(\theta,T)\right] \\ &= \sum_{\vec{n}\in\mathbb{Z}^K_{\geq 0}}\mathbb{E}^Q\left[\mathbf{1}(\vec{N}=\vec{n})f(\theta,X_{t\in[0,T]})G(\theta,T)\right].\end{aligned} \tag{3.22}$$

Then the exchange of the expectation and derivative is valid for a single term:

$$\begin{aligned}&\frac{\partial}{\partial\theta_i}\mathbb{E}^Q\left[\mathbf{1}(\vec{N}=\vec{n})f(\theta,X_{t\in[0,T]})G(\theta,T)\right] \\ &\quad= \mathbb{E}^Q\left[\mathbf{1}(\vec{N}=\vec{n})G(\theta,T)\left(\frac{\partial}{\partial\theta_i}f(\theta,X_{t\in[0,T]})+f(\theta,X_{t\in[0,T]})H_i(\theta,T)\right)\right].\end{aligned} \tag{3.23}$$

To see this, note that since the total number of jumps is fixed (and each $\|\zeta_k\| < \infty$), $\|X_s\|$ is bounded for all $s \in [0,T]$. Thus the integrand on the left side of (3.23) is bounded by Condition 3.26 on f and by the definition of G. The integrand on the right is similarly bounded, since $H(\theta,T)$ is infinite only if some rate $\lambda_{k_\ell}(\theta,\hat{X}_\ell) = 0$, in which case $G(\theta,T) = 0$. Thus on this single term the Lipschitz conditions are satisfied, as both

f and the λ_k are assumed to be differentiable in θ, so that (3.23) holds as claimed.

Recall from Section 3.4 that some reaction index $k \in \mathcal{R}^1$ if and only if an occurrence of the k^{th} reaction increases the total population $\|X\|$. We define the random variable $R_1 := R_1(t) = \sum_{k\in\mathcal{R}_1} N_k(t)$, and for any fixed $\vec{n} \in \mathbb{Z}^K_{\geq 0}$ we define $r_1(\vec{n}) = \sum_{k\in\mathcal{R}_1} n_k$. Now let

$$s_r(\theta, T) := \sum_{r_1(\vec{n})=r} \mathbb{E}^Q \left[\mathbf{1}(\vec{N} = \vec{n}) f(\theta, X_{t\in[0,T]}) G(\theta, T)\right] \quad \text{and} \quad S_j(\theta, T) := \frac{\partial}{\partial\theta_i} \sum_{r=1}^{j} s_r(\theta, T).$$

That is, the summation on the left is over all vectors $\vec{n} \in \mathbb{Z}^K_{\geq 0}$ such that the sum of the coordinates corresponding to $\mathcal{R}_1$ reactions is equal to r.

First, we claim that

$$\frac{\partial}{\partial\theta_i} s_r(\theta, T) = \sum_{r_1(\vec{n})=r} \frac{\partial}{\partial\theta_i} \mathbb{E}^Q \left[\mathbf{1}(\vec{N} = \vec{n}) f(\theta, X_{t\in[0,T]}) G(\theta, T)\right]. \tag{3.24}$$

Second, we also claim that the partial sums S_j converge uniformly on Θ. Rewriting (3.22), we see that

$$\begin{aligned}
\mathbb{E}^{P^\theta}[f(\theta, X_{t\in[0,T]})] &= \sum_{\vec{n}\in\mathbb{Z}^K_{\geq 0}} \mathbb{E}^Q \left[\mathbf{1}(\vec{N} = \vec{n}) f(\theta, X_{t\in[0,T]}) G(\theta, T)\right] \\
&= \sum_{r=1}^{\infty} \sum_{r_1(\vec{n})=r} \mathbb{E}^Q \left[\mathbf{1}(\vec{N} = \vec{n}) f(\theta, X_{t\in[0,T]}) G(\theta, T)\right] \\
&= \sum_{r=1}^{\infty} s_r(\theta, T).
\end{aligned}$$

Therefore, our two claims imply that

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\mathbb{E}^{P^\theta}[f(\theta, X_{t\in[0,T]})] &= \frac{\partial}{\partial\theta_i}\sum_{r=1}^{\infty} s_r(\theta, T) \\ &= \sum_{r=1}^{\infty}\frac{\partial}{\partial\theta_i}s_r(\theta, T) \\ &= \sum_{r=1}^{\infty}\frac{\partial}{\partial\theta_i}\sum_{r_1(\vec{n})=r}\mathbb{E}^Q\left[\mathbf{1}(\vec{N}=\vec{n})f(\theta, X_{t\in[0,T]})G(\theta, T)\right] \\ &= \sum_{r=1}^{\infty}\sum_{r_1(\vec{n})=r}\frac{\partial}{\partial\theta_i}\mathbb{E}^Q\left[\mathbf{1}(\vec{N}=\vec{n})f(\theta, X_{t\in[0,T]})G(\theta, T)\right]\end{aligned}$$

where the second line follows from standard results (e.g., [36]) and our claim that the partial sums converge uniformly. Then using (3.23) we obtain

$$\begin{aligned}&\frac{\partial}{\partial\theta_i}\mathbb{E}^{P^\theta}[f(\theta, X_{t\in[0,T]})] \\ &\quad= \sum_{r=1}^{\infty}\sum_{r_1(\vec{n})=r}\mathbb{E}^Q\left[\mathbf{1}(\vec{N}=\vec{n})G(\theta, T)\left(\frac{\partial}{\partial\theta_i}f(\theta, X_{t\in[0,T]}) + f(\theta, X_{t\in[0,T]})H_i(\theta, T)\right)\right] \\ &\quad= \mathbb{E}^Q\left[G(\theta, T)\left(\frac{\partial}{\partial\theta_i}f(\theta, X_{t\in[0,T]}) + f(\theta, X_{t\in[0,T]})H_i(\theta, T)\right)\right] \\ &\quad= \mathbb{E}^{P^\theta}\left[\frac{\partial}{\partial\theta_i}f(\theta, X_{t\in[0,T]}) + f(\theta, X_{t\in[0,T]})H_i(\theta, T)\right]\end{aligned}$$

as needed.

First, to show (3.24), we again consider the Lipschiz conditions, this time to justify interchanging the derivative and the summation in $s_r(\theta, T)$. Now

$$s_r(\theta, T) = \sum_{r_1(\vec{n})=r}\mathbb{E}^Q\left[\mathbf{1}(\vec{N}=\vec{n})f(\theta, X_{t\in[0,T]})G(\theta, T)\right] = \mathbb{E}^{P^\theta}\left[\mathbf{1}(R_1=r)f(\theta, X_{t\in[0,T]})\right].$$

This quantity is finite for all $\theta\in\Theta$ by Condition 3.26 and since under P^θ if $R_1 = r$ then

for all $s \leq T$ we know $\|X_s\| \leq \|X_0\| + cr$, where $c = \max_{k \in \mathcal{R}_1} \mathbf{1} \cdot \zeta_k$. Using (3.23) and the fact that $G(\theta, T)$ is nonnegative, we obtain the following bound:

$$\begin{aligned}
&\left| \frac{\partial}{\partial \theta_i} \mathbb{E}^Q \left[\mathbf{1}(\vec{N} = \vec{n}) f(\theta, X_{t\in[0,T]}) G(\theta, T) \right] \right| \\
&\quad = \left| \mathbb{E}^Q \left[G(\theta, T) \mathbf{1}(\vec{N} = \vec{n}) \left(\frac{d}{d\theta_i} f(\theta, X_{t\in[0,T]}) + f(\theta, X_{t\in[0,T]}) H_i(\theta, T) \right) \right] \right| \\
&\quad \leq \mathbb{E}^Q \left[G(\theta, T) \mathbf{1}(\vec{N} = \vec{n}) \left| \frac{d}{d\theta_i} f(\theta, X_{t\in[0,T]}) + f(\theta, X_{t\in[0,T]}) H_i(\theta, T) \right| \right] \\
&\quad = \mathbb{E}^{P^\theta} \left[\mathbf{1}(\vec{N} = \vec{n}) \left| \frac{d}{d\theta_i} f(\theta, X_{t\in[0,T]}) + f(\theta, X_{t\in[0,T]}) H_i(\theta, T) \right| \right] \\
&\quad \leq \mathbb{E}^{P^\theta} \left[\mathbf{1}(\vec{N} = \vec{n}) \left(C_B(1 + \sup_{t\in[0,T]} \|X_t\|^{p_B}) + C_A(1 + \sup_{t\in[0,T]} \|X_t\|^{p_A}) \sup_{\theta\in\Theta} \left| H_i(\theta, T) \right| \right) \right],
\end{aligned}$$

where the last line follows from Condition 3.26. Note that this bound is uniform in θ. We now show that this bound is summable:

$$\begin{aligned}
&\sum_{r_1(\vec{n})=r} \left| \frac{\partial}{\partial \theta_i} \mathbb{E}^Q \left[\mathbf{1}(\vec{N} = \vec{n}) f(\theta, X_{t\in[0,T]}) G(\theta, T) \right] \right| \qquad (3.25) \\
&\leq \sum_{r_1(\vec{n})=r} \mathbb{E}^{P^\theta} \left[\mathbf{1}(\vec{N} = \vec{n}) \left(C_B(1 + \sup_{t\in[0,T]} \|X_t\|^{p_B}) + C_A(1 + \sup_{t\in[0,T]} \|X_t\|^{p_A}) \sup_{\theta\in\Theta} \left| H_i(\theta, T) \right| \right) \right] \\
&\leq \mathbb{E}^{P^\theta} \left[\mathbf{1}(R_1 = r) \left(C_B(1 + \sup_{t\in[0,T]} \|X_t\|^{p_B}) + C_A(1 + \sup_{t\in[0,T]} \|X_t\|^{p_A}) \sup_{\theta\in\Theta} \left| H_i(\theta, T) \right| \right) \right] \\
&\leq \mathbb{E}^{P^\theta} \left[C_B(1 + \sup_{t\in[0,T]} \|X_t\|^{p_B}) \right] + \mathbb{E}^{P^\theta} \left[C_A(1 + \sup_{t\in[0,T]} \|X_t\|^{p_A}) \sup_{\theta\in\Theta} \left| H_i(\theta, T) \right| \right].
\end{aligned}$$

The first expectation is finite by Corollary 3.24. The second expectation is also finite, using the Cauchy-Schwarz inequality, Corollary 3.24 and the bound on H given by (3.21). Thus the derivative and sum in (3.24) commute as needed.

Then what remains to show is our second claim, that the partial sums S_j converge

uniformly on Θ. For $\ell < m$ we have

$$\begin{aligned} S_m(\theta,T) - S_\ell(\theta,T) &= \sum_{r=\ell+1}^{m} \frac{\partial}{\partial\theta_i} s_r(\theta,T) \\ &= \sum_{r=\ell+1}^{m} \frac{\partial}{\partial\theta_i} \sum_{r_1(\vec{n})=r} \mathbb{E}^Q \left[\mathbf{1}(\vec{N}=\vec{n}) f(\theta, X_{t\in[0,T]}) G(\theta,T) \right] \\ &= \sum_{r=\ell+1}^{m} \sum_{r_1(\vec{n})=r} \frac{\partial}{\partial\theta_i} \mathbb{E}^Q \left[\mathbf{1}(\vec{N}=\vec{n}) f(\theta, X_{t\in[0,T]}) G(\theta,T) \right]. \end{aligned}$$

Using a bound from (3.25), we have

$$\begin{aligned} &\left| S_m(\theta,T) - S_\ell(\theta,T) \right| \le \sum_{r=\ell+1}^{m} \sum_{r_1(\vec{n})=r} \left| \frac{\partial}{\partial\theta_i} \mathbb{E}^Q \left[\mathbf{1}(\vec{N}=\vec{n}) f(\theta, X_{t\in[0,T]}) G(\theta,T) \right] \right| \\ &\le \sum_{r=\ell+1}^{m} \mathbb{E}^{P^\theta} \left[\mathbf{1}(R_1 = r) \left(C_B(1 + \sup_{t\in[0,T]} \|X_t\|^{p_B}) + C_A(1 + \sup_{t\in[0,T]} \|X_t\|^{p_A}) \sup_{\theta\in\Theta} \left| H_i(\theta,T) \right| \right) \right] \\ &= \mathbb{E}^{P^\theta} \left[\mathbf{1}(\ell < R_1 \le m) \left(C_B(1 + \sup_{t\in[0,T]} \|X_t\|^{p_B}) + C_A(1 + \sup_{t\in[0,T]} \|X_t\|^{p_A}) \sup_{\theta\in\Theta} \left| H_i(\theta,T) \right| \right) \right] \\ &\le \sqrt{P^\theta(\ell < R_1 \le m)} \\ &\qquad \times \sqrt{\mathbb{E}^{P^\theta} \left[\left(C_B(1 + \sup_{t\in[0,T]} \|X_t\|^{p_B}) + C_A(1 + \sup_{t\in[0,T]} \|X_t\|^{p_A}) \sup_{\theta\in\Theta} \left| H_i(\theta,T) \right| \right)^2 \right]} \end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality. Now, similarly to (3.25) above, the second square root has finite expectation uniformly in θ by using the bound $(a+b)^2 \le 2a^2 + 2b^2$ and another application of the Cauchy-Schwarz inequality. Furthermore, the first square root goes to zero uniformly in θ as $\ell \to \infty$, by Lemmas 3.22 and 3.23. That is, by choosing ℓ large enough $|S_m(\theta,T) - S_\ell(\theta,T)|$ can be made arbitrarily small uniformly on Θ, so the partial sums converge uniformly as needed by the Cauchy criterion. $\square$

3.5 A Limiting Argument for the Hybrid Pathwise Method

By the proof in Section 3.2.4, we know that we may use the pathwise method for non-interruptive processes with bounded intensities to get estimates of model sensitivities. While Section 3.1 shows that the non-interruptive condition 3.8 is necessary, we do not believe that bounded intensities are necessary; however, we have been unable to prove this fact analytically.

We are able to show the following weaker limit. Given some process X, we define the related process X^M by "capping" the intensities at some large value M. We then show that, under mild conditions, a given sensitivity of X^M converges to the sensitivity of X as $M \to \infty$. By Remark 3.2, this result holds for the process Z as constructed for the hybrid method as well. Since the result only concerns the values of the sensitivities, not the validity of any given sensitivity method, this theorem says nothing about which sensitivity methods are valid on the limiting model.

Theorem 3.28. *Suppose we are given the process $X_t(\theta)$ satisfying* (1.7)

$$X_t(\theta) = X_0(\theta) + \sum_{k=1}^{K} Y_k \left(\int_0^t \lambda_k(\theta, X_s(\theta)) ds \right),$$

where the intensities λ_k satisfy the linear growth Condition 3.21 and Condition 3.25 on some neighborhood $\Theta \subset \mathbb{R}^R$ of θ. Also suppose F satisfies Condition 3.11 on Θ. For each $M \in \mathbb{N}$, define the CTMC $X^M(\theta)$ satisfying

$$X_t^M(\theta) = X_0^M(\theta) + \sum_{k=1}^{K} Y_k \left(\int_0^t \lambda_k^M(\theta, X_s^M(\theta)) ds \right),$$

where $\lambda_k^M(\theta, x) = \lambda_k(\theta, x) \wedge M$ *and where* $X_0(\theta) = X_0^M(\theta) = X_0$ *is fixed. For each* M*, also define*

$$L(\theta, M) := \int_a^b F(\theta, X_s^M(\theta))ds$$

similarly to (3.4)*. Then for each* $i = 1, \ldots, R$*,*

$$\lim_{M\to\infty} \frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta, M)] = \frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta)],$$

where

$$L(\theta) := \int_a^b F(\theta, X_s(\theta))ds.$$

Proof. For each θ, construct the processes $X^M(\theta)$ with the same realizations of the independent unit-rate Poisson processes Y_k as used for X. Then $X^M(\theta) \to X(\theta)$ a.s. as $M \to \infty$. Indeed, for a.e. path of X, there is some M_0 such that the intensity is bounded by M_0 through time b. Then for $M > M_0$ the two paths X and X^M are then the same through time b, since they are the same at time zero and since for any state x along the path of X through b, we have by assumption that

$$\lambda_k^M(\theta, x) = \lambda_k(\theta, x) \wedge M = \lambda_k(\theta, x).$$

This implies that $L(\theta, M) \to L(\theta)$ a.s. as well. Then we must show that as $M \to \infty$ we have (1) that $\mathbb{E}[L(\theta, M)] \to \mathbb{E}[L(\theta)]$ and (2) that $\frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta, M)]$ converges uniformly on Θ. Then by standard results (for example, see [36]), $\lim_{M\to\infty} \frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta, M)] = \frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta)]$ as needed.

Throughout the reminder of this proof, we will use several of the bounds from the

construction in Section 3.4.1, in which we showed that the process X and several related quantities have finite moments by stochastically bounding the process X using certain processes B and C. In particular, note that if B and C are constructed in that manner, then they can also be used to stochastically bound X^M uniformly for any M.

By Condition 3.11 on F we have

$$\begin{aligned} |L(\theta, M)| &= \left| \int_a^b F(\theta, X_s^M(\theta)) ds \right| \\ &\leq \int_a^b \left| F(\theta, X_s^M(\theta)) \right| ds \\ &\leq \int_a^b C_1(1 + \|X_s^M(\theta)\|^{c_1}) ds \\ &\leq C_1(b-a)(1 + \sup_{s \in [0,b]} \|X_s^M(\theta)\|^{c_1}). \end{aligned} \tag{3.26}$$

Thus by (3.19), we have

$$\sup_M |L(\theta, M)| \leq C_1(b-a)(1 + (\|X_0\| + cB_b)^{c_1}), \tag{3.27}$$

and so by Lemmas 3.22 and 3.23, the quantity on the right has finite moments and is independent of M. By dominated convergence, we have the statement (1), that $\mathbb{E}[L(\theta, M)] \to \mathbb{E}[L(\theta)]$ on Θ.

For a stochastic process Z, let $f(\theta, Z_{s \in [0,b]}) = \int_a^b F(\theta, Z_s) ds$. We will be considering $f(\theta, X_{s \in [0,b]}(\theta))$ and $f(\theta, X^M_{s \in [0,b]}(\theta))$. The fact that f satisfies Condition 3.26 for the processes X and X^M follows from an argument very similar to that in (3.20). Also,

since the λ_k satisfy Conditions 3.21 and 3.25, so do the λ_k^M. Then by Theorem 3.27,

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta,M)] &= \frac{\partial}{\partial\theta_i}\mathbb{E}^{P^\theta}[f(\theta,X^M_{s\in[0,b]})] \\ &= \mathbb{E}^{P^\theta}\left[\frac{\partial}{\partial\theta_i}f(\theta,X^M_{s\in[0,b]}) + f(\theta,X^M_{s\in[0,b]})H_i(\theta,b,X^M)\right],\end{aligned} \tag{3.28}$$

where we have written $H_i(\theta,b,X^M)$ to denote the weighting function (2.13) for the process X^M. Then

$$H_i(\theta,b,X^M) = \sum_{\ell=0}^{N-1}\frac{\frac{\partial}{\partial\theta_i}\lambda^M_{k_\ell}(\theta,\hat{X}^M_\ell)}{\lambda^M_{k_\ell}(\theta,\hat{X}^M_\ell)} - \sum_{k=1}^{K}\int_0^b \lambda^M_k(\theta,X^M_s)ds,$$

and as in (3.21), we have the bound

$$\begin{aligned}\sup_{\theta\in\Theta}|H_i(\theta,b,X^M)| \le NC_1(1+\sup_{t\in[0,b]}\|X^M_s\|^{p_1})C_2(1+\sup_{t\in[0,b]}\|X^M_s\|^{p_2}) \\ + Kb\max_k\sup_{\theta\in\Theta}\sup_{s\in[0,b]}\lambda^M_k(\theta,X^M_s).\end{aligned} \tag{3.29}$$

Therefore,

$$\begin{aligned}\sup_M\sup_{\theta\in\Theta}|H_i(\theta,b,X^M)| \le NC_1(1+\|X_0\|+cB_b^{p_1})C_2(1+\|X_0\|+cB_b^{p_2}) \\ + Kb\tilde{K}_1B_b,\end{aligned} \tag{3.30}$$

where recall that $\tilde{K}_1B_s$ is the intensity at time s of the bounding linear growth process B as in Section 3.4.1.

Since $X^M\to X$ a.s. as $M\to\infty$, we also have, as $M\to\infty$, the a.s. convergences

$$f(\theta,X^M_{s\in[0,t]})\to f(\theta,X_{s\in[0,t]}),$$

$\frac{\partial}{\partial\theta_i} f(\theta, X^M_{s\in[0,t]}) \to \frac{\partial}{\partial\theta_i} f(\theta, X_{s\in[0,t]})$, and

$H_i(\theta, T, X^M) \to H_i(\theta, T, X)$.

What is left to show is statement (2), that $\frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta, M)]$ converges uniformly on Θ. So consider for $M_1 < M_2$ the quantity

$$\sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta, M_1)] - \frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta, M_2)]\right|.$$

We will show that this quantity tends to zero as $M_1 \to \infty$, which gives the desired uniform convergence through the Cauchy criterion. By (3.28) we have that

$$\begin{aligned}
&\left|\frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta, M_1)] - \frac{\partial}{\partial\theta_i}\mathbb{E}[L(\theta, M_2)]\right| \\
&\qquad = \left|\mathbb{E}^{P^\theta}\left[\frac{\partial}{\partial\theta_i} f(\theta, X^{M_1}_{s\in[0,b]}) + f(\theta, X^{M_1}_{s\in[0,b]}) H_i(\theta, b, X^{M_1})\right]\right. \\
&\qquad\qquad \left. - \mathbb{E}^{P^\theta}\left[\frac{\partial}{\partial\theta_i} f(\theta, X^{M_2}_{s\in[0,b]}) + f(\theta, X^{M_2}_{s\in[0,b]}) H_i(\theta, b, X^{M_2})\right]\right| \\
&\qquad \le \left|\mathbb{E}^{P^\theta}\left[\frac{\partial}{\partial\theta_i} f(\theta, X^{M_1}_{s\in[0,b]}) - \frac{\partial}{\partial\theta_i} f(\theta, X^{M_2}_{s\in[0,b]})\right]\right| \qquad (3.31) \\
&\qquad\qquad + \left|\mathbb{E}^{P^\theta}\left[f(\theta, X^{M_1}_{s\in[0,b]}) H_i(\theta, b, X^{M_1}) - f(\theta, X^{M_2}_{s\in[0,b]}) H_i(\theta, b, X^{M_2})\right]\right|.
\end{aligned}$$

We consider the two terms separately. Note that both terms are zero if X^{M_1} and X^{M_2} are equal through time b, an event that is guaranteed by our construction of these processes if $\sup_{s\in[0,b]} \lambda_k^{M_1}(\theta, X_s) \le M_1$ for each k.

We can bound the first term in (3.31) by

$$\left|\mathbb{E}^{P^\theta}\left[\frac{\partial}{\partial\theta_i}f(\theta, X^{M_1}_{s\in[0,b]}) - \frac{\partial}{\partial\theta_i}f(\theta, X^{M_2}_{s\in[0,b]})\right]\right| \\ \leq \mathbb{E}^{P^\theta}\left[\left|\frac{\partial}{\partial\theta_i}f(\theta, X^{M_1}_{s\in[0,b]}) - \frac{\partial}{\partial\theta_i}f(\theta, X^{M_2}_{s\in[0,b]})\right|\mathbf{1}\left(\sup_{s\in[0,b]}\lambda_k^{M_1}(\theta, X_s) \geq M_1 \text{ for some } k\right)\right].$$

To show that this tends to zero as $M \to \infty$, first use the Cauchy-Schwarz inequality. The first factor is uniformly bounded in θ and in M by Condition 3.26 on f and, similarly to (3.27), by the fact that the linear growth process B, which has finite moments, stochastically bounds each X^M. Then note that

$$\begin{aligned}\mathbb{P}&\left(\sup_{s\in[0,b]}\lambda_k^{M_1}(\theta, X_s) \geq M_1 \text{ for some } k\right) \\ &\leq \mathbb{P}\left(\sup_{\theta\in\Theta}\sup_{s\in[0,b]}\lambda_k(\theta, X_s) \geq M_1 \text{ for some } k\right) \\ &\leq \mathbb{P}\left(K_1 B_b \geq M_1\right),\end{aligned} \tag{3.32}$$

which goes to zero by Lemma 3.22. Similarly, consider the second term in (3.31),

$$\left|\mathbb{E}^{P^\theta}\left[f(\theta, X^{M_1}_{s\in[0,b]})H_i(\theta, b, X^{M_1}) - f(\theta, X^{M_2}_{s\in[0,b]})H_i(\theta, b, X^{M_2})\right]\right| \\ \leq \mathbb{E}^{P^\theta}\left[\left|f(\theta, X^{M_1}_{s\in[0,b]})H_i(\theta, b, X^{M_1}) - f(\theta, X^{M_2}_{s\in[0,b]})H_i(\theta, b, X^{M_2})\right| \right. \\ \left. \times\, \mathbf{1}\left(\sup_{s\in[0,b]}\lambda_k^{M_1}(\theta, X_s) \geq M_1 \text{ for some } k\right)\right].$$

Again, use the Cauchy-Schwarz inequality and the fact that the probability of the event in the indicator goes to zero uniformly in θ and M as in (3.32). To see that $\left|f(\theta, X^M_{s\in[0,b]})H_i(\theta, b, X^M)\right|$ is bounded uniformly in M and θ, use the Cauchy-Schwarz

inequality once more, and the fact that (3.26) and (3.30) also provide finite second moment bounds by Lemma 3.22.

□

3.6 Numerical Examples

With the examples in this section, we wish to demonstrate two points: the validity of the pathwise or hybrid methods, and the efficiency of pathwise and hybrid methods. First, we numerically show that pathwise-only methods can fail if we have a process in which interruptions can occur. Second, we compare the efficiency of the methods that are valid for specific biochemical models, including the hybrid method introduced in this thesis.

In the literature for parameter sensitivity estimation for random systems in general, there is a rule of thumb that pathwise methods should be used when they are valid. For many of the models discussed below, pathwise methods are significantly more efficient than other methods, having variances possibly several orders of magnitude smaller than that of other methods. For some models, however, pathwise methods are less efficient than other methods. Future work will involve a wider numerical study to help determine a better framework for choosing the most efficient method for a given model.

3.6.1 The Pathwise Method on Non-Interruptive Processes

If the process X of interest satisfies the Non-Interruptive condition 3.8, the pathwise method of Section 3.2 may be used to estimate sensitivities. That is, there is no need to construct an approximate process Z to use the full hybrid method of Section 3.3.3 since

X is already non-interruptive. We give two such examples here.

Birth–Death

We consider the birth-death model from Figure 2.1 in Chapter 2, given by

$$\emptyset \underset{\theta_2}{\overset{\theta_1}{\rightleftarrows}} \mathcal{S}$$

with mass-action kinetics, and where X_t gives the abundance of $\mathcal{S}$ at time t with $X_0 = 0$. For this process, we can solve to find that

$$\mathbb{E}X_t(\theta) = \frac{\theta_1}{\theta_2}(1 - e^{-\theta_2 t}),$$

$$\frac{\partial}{\partial\theta_1}\mathbb{E}X_t(\theta) = \frac{1}{\theta_2}(1 - e^{-\theta_2 t}) \quad \text{and} \quad \frac{\partial}{\partial\theta_2}\mathbb{E}X_t(\theta) = \frac{\theta_1}{\theta_2}(te^{-\theta_2 t}) - \frac{\theta_1}{\theta_2^2}(1 - e^{-\theta_2 t}).$$

Since X is clearly non-interruptive, we use the Dynkin Pathwise and RPD methods from Section 2.3 to estimate the sensitivity gradient of the quantity $\mathbb{E}X_t(\theta)$ at $\theta_0 = (\theta_1, \theta_2) = (2, 0.1)$. Though the intensity of the model is technically unbounded, the intensities are "bounded in practice:" throughout these simulations no intensity was ever greater than $M = 10^3$, showing that, if we had used the full hybrid method, the approximate process Z would be the same as X with very high probability. Thus we may confidently use both pathwise-only methods, and as can be seen in the table below, both give a good estimate of the true sensitivity.

Figure 3.1 compares estimator variance of the different methods through time 200 for both entries of the gradient. Note that the behavior of the pathwise methods are rather different for each entry. Concerning estimates of $\frac{\partial}{\partial\theta_1}\mathbb{E}X_t(\theta)$, the variance of the

pathwise methods is lower than the likelihood ratio method for these times. However, the variance of the pathwise methods grow at a faster rate, so that for large enough times each pathwise method will become less efficient than the likelihood ratio method. On the other hand, for $\frac{\partial}{\partial\theta_2}\mathbb{E}X_t(\theta)$, the variance of each pathwise estimator is orders of magnitude smaller than the variance of the likelihood ratio estimator, even when a control variate is used. Furthermore, for this sensitivity the variance of the Dynkin pathwise method seems to converge.

Table 3.1 gives more details of the efficiency of the Dynkin Pathwise and RPD methods, compared with the likelihood ratio and CFD methods at $t = 50$, for which from the work above we can calculate the true gradient:

$$\nabla_\theta \mathbb{E}X_{50}(\theta)\big|_{\theta_0=(2,0.1)} = (9.93, -191.91).$$

Comparing the two unbiased methods, we see that while the Dynkin method is slower per path, it is more efficient: at 10^3 paths it is much faster but about as precise as the likelihood ratio method with 10^4 paths. As discussed with Figure 3.1, this analysis will change depending on t. It is more difficult to directly compare the biased methods because of their dependence on w or ϵ. Here and throughout, the half-widths given for estimates are computed for a 95% confidence interval.

Linear Growth

We first consider the linear growth model

$$\mathcal{S} \xrightarrow{\theta} 2\mathcal{S},$$

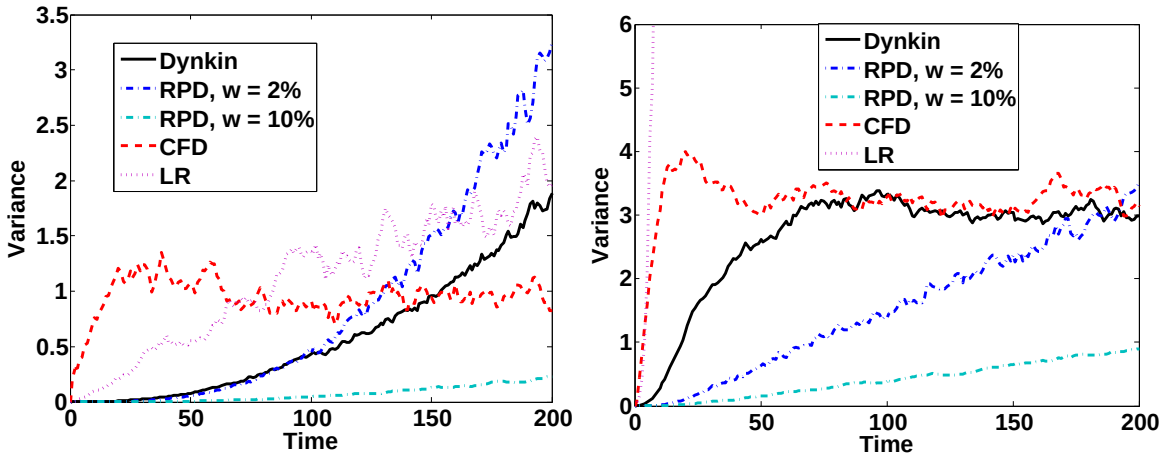

Figure 3.1: Variance of the different sensitivity estimators of (left) $\frac{\partial}{\partial\theta_1}\mathbb{E}X_t(\theta)$ and (right) $\frac{\partial}{\partial\theta_2}\mathbb{E}X_t(\theta)$ on the birth–death model, with 1000 sample paths. A perturbation of $\epsilon = .01$ and the centered difference were used for the CFD method, and the weights were used as control variates in the likelihood ratio method. Note that the pathwise and likelihood ratio method behave rather differently on the two estimates. Also note that the choice of w significantly affects the variance of the RPD estimators: $w = 10\%$ means that w was chosen to be 10% of the time given on the x-axis, so that the paths for the RPD method must actually be simulated until time 220 (or 204 when $w = 2\%$). The likelihood ratio estimator for $\frac{\partial}{\partial\theta_2}\mathbb{E}X_t(\theta)$ has a variance of about 850 at time 200.

Unbiased				
Method	n	CPU	Gradient Estimate	
Dynkin PW	10^3	1.6	(10.3 ± 0.6,	-192.5 ± 3.4)
LR	10^3	0.8	(10.2 ± 14.4,	-204.8 ± 27.1)
Dynkin PW	10^4	14.8	(10.0 ± 0.2,	-191.5 ± 1.0)
LR	10^4	7.3	(10.2 ± 0.5,	-193.3 ± 8.8)

Biased					
Method	$w\epsilon$	n	CPU	Gradient Estimate	
RPD	5	10^3	1.2	(9.6 ± 0.7,	-194.0 ± 3.2)
RPD	1	10^3	1.2	(10.8 ± 1.8,	-193.0 ± 6.3)
CFD	.01	10^3	1.0	(10.4 ± 2.8,	-192.6 ± 5.2)
CFD	.002	10^3	1.0	(9.0 ± 5.8,	-193.0 ± 21.4)
RPD	5	10^4	11.6	(9.9 ± 0.2,	-191.5 ± 1.0)
RPD	1	10^4	11.0	(10.3 ± 0.6,	-191.5 ± 2.0)
CFD	.01	10^4	9.4	(10.2 ± 0.9,	-192.9 ±1.6)
CFD	.002	10^4	9.3	(9.4 ± 1.9,	-193.0 ± 6.7)

Table 3.1: A comparison of the sensitivity methods on the birth–death model. CPU gives computation time in seconds; n is the number of paths simulated. Actual gradient: (9.93,-191.91). The LR method included the weight as a control variate to reduce variance. The CFD method used the centered difference, and used $n/2$ paths per entry for a more fair comparison as it is the only method that cannot reuse paths. For the RPD method, $w = 5$ is 10% and $w = 1$ is 2% of $t = 50$, which matches Figure 3.1.

and let $X_t(\theta)$ denote the number of $\mathcal{S}$ molecules at time t with $X_0(\theta) \equiv 1$. Note that the intensity of $X_t(\theta)$ at any time $t > 0$ is not bounded, so the proof given in Section 3.2.4 does not hold. However, since the model only has one reaction, the derivatives $\frac{\partial}{\partial\theta} S^\theta(T_i^\theta)$ are zero for all i. Though we do not include the proof here, the fact that these derivatives are zero leads to a similar proof that the pathwise method holds, even though the intensity is not bounded.

Note that for this simple model, we have

$$\mathbb{E}X_t = e^{\theta t} \quad \text{and} \quad \frac{\partial}{\partial\theta}\mathbb{E}X_t = te^{\theta t},$$

so that at $t = 5$ we have $e^5 \approx 148.41$ and $5e^5 \approx 742.07$. Table 3.2 gives numerical results comparing our various sensitivity methods. The two pathwise methods have approximately the same efficiency, so that one would likely choose to use the unbiased Dynkin pathwise method.

3.6.2 A Simple Switch

In contrast to the previous models, the following simple switch is one in which the two pathwise-only methods do *not* work:

$$\emptyset \overset{\theta_1}{\leftarrow} A \overset{\theta_2}{\rightarrow} B \overset{\theta_3}{\rightarrow} C,$$

We will suppose that $X_0(\theta) = (a, 0, 0)$ gives the initial abundances of A, B, and C respectively. We estimate the derivative with respect to θ_1 of the mean number of C molecules, $\frac{\partial}{\partial\theta_1}\mathbb{E}X_{C,t}(\theta)$, at $\theta = (\frac{1}{4}, 1, 1)$ and at various times t.

Unbiased			
Method	n	CPU	Estimate
Dynkin PW	10^4	11.7	739.8 ± 14.4
LR	10^4	7.0	719.9 ± 50.4
Dynkin PW	10^5	115.7	741.8 ± 4.6
LR	10^5	71.1	743.6 ± 17.8

Biased							
1%				0.1%			
Method	n	CPU	Estimate	Method	n	CPU	Estimate
RPD PW	10^4	9.5	750.7 ± 15.1	RPD PW	10^4	9.2	734.9 ± 14.8
CFD	10^4	9.4	744.6 ± 23.4	CFD	10^4	8.6	738.2 ± 34.9
RPD PW	10^5	96.2	749.4 ± 4.7	RPD PW	10^5	90.7	738.4 ± 4.7
CFD	10^5	96.2	746.3 ± 7.4	CFD	10^5	100.6	735.1 ± 10.9

Table 3.2: A comparison of the sensitivity methods on the linear growth model. Actual sensitivity: 742.07. CPU gives computation time in seconds; n is the number of paths simulated. The LR method included the weight as a control variate to reduce variance, and for the CFD method the centered difference was used. Note that the methods in the top table are unbiased; for the method in the bottom table, w and h respectively were chosen such that the bias was either 1% or 0.1% of the actual sensitivity value, which for this simple model is a tractable computation: $w = .21, h = .049$ and $w = .065, h = .0155$ respectively. Both methods overestimate the sensitivity, as expected. The variance of the RPD method for this model does not seem to be significantly different for the two choices of w; on the other hand, the variance of the CFD method is substantially different for the two choices of h.

Pathwise-only Methods are Invalid

Since this switch model is linear, we can solve for the sensitivity exactly at $\theta = (\theta_1, 1, 1)$:

$$\frac{\partial}{\partial \theta_1}\mathbb{E}X_{C,t}(\theta) = \frac{a}{\theta_1^2(1+\theta_1)^2}\left((1+\theta_1)^2 e^{-t} - \theta_1^2 - (t\theta_1^2 + (t+2)\theta_1 + 1)e^{-t(1+\theta_1)}\right).$$

We now consider the error of the Dynkin pathwise and RPD methods in computing the sensitivity $\frac{\partial}{\partial \theta_1}\mathbb{E}X_{C,t}(\theta)$ of this switch model using the estimator

$$\frac{\partial}{\partial \theta_1}\frac{1}{2w}\int_{T-w}^{T+w} X_{C,s}(\theta)ds.$$

Note that by Dynkin's formula (1.6), $\mathbb{E}X_{C,t}(\theta) = \mathbb{E}\int_0^t X_{B,s}ds$. Thus we have already showed analytically, in Example 3.7, that for $a = 1$ the Dynkin pathwise method will always return estimates of zero. Considering the RPD method, suppose the time T is taken to be a larger number such as $T = 10$ and w is taken much smaller than T. Then we also expect the RPD method to return sensitivity estimates of zero: the estimate only takes into account jumps that occur in a time window of size $2w$ centered at T, and with high probability all jumps will have been completed by time $T - w$. These errors occur as expected in simulation, as shown in the first plot in Figure 3.2. We also show results when using a time of $t = 2$, with two different choices of window size w for the RPD method; the Dynkin pathwise and RPD methods still do not provide correct estimates. At the very small time of $t = 0.5$, the error of the methods appears to be very small, though it is still noticeable for small initial abundances of A.

In these plots, the same value of w was used for both the RPD and RPD Hybrid methods. Therefore, at small times we can see that the bias of the window size w is

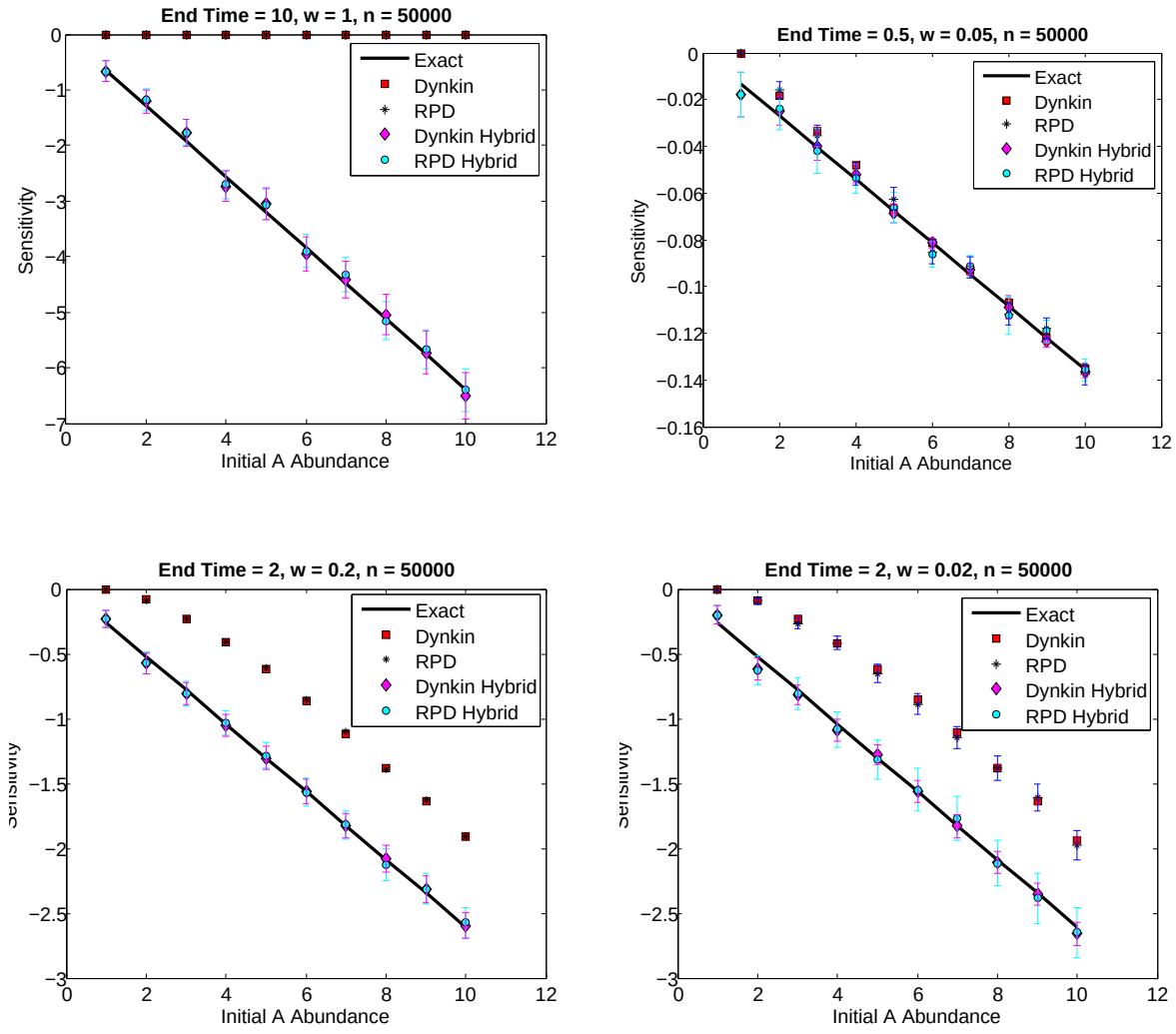

Figure 3.2: A comparison of error and bias of the estimates for the sensitivity of $\mathbb{E}X_{C,t}(\theta) = \mathbb{E}\int_0^t X_{B,s}ds$ with respect to θ_1, the rate parameter controlling the decay of A in the switch model of Section 3.6.2. Except at very small times, the Dynkin and RPD methods have significant error. In the case of the RPD method, we can see that this error is not due to the bias from the window size w by comparing to the RPD hybrid method. n gives the total number of paths simulated.

the major contributing factor to the error of the RPD method, since the RPD hybrid and RPD-only methods show a similar bias. At larger times this is not the case: the RPD method's error is very different from the bias of the RPD hybrid method. Thus the error (rather than bias) of the RPD method varies. In particular, the error increases as interruptions become more likely. For this switch model, while we may not see an interruption if the end time t is small, at larger times the path sees an interruption with a probability approaching 1. Crucially, however, if interruptions are possible, one cannot *a priori* know the size of the error of the RPD method.

The results in Figure 3.2 show numerically that the Regularized Pathwise Differentiation (RPD) method is not valid for models with interruptions, as discussed in Appendix B of [37]. The results also show that the Dynkin pathwise method is not valid on this model. Instead, for such models the hybrid method, or a finite difference or likelihood ratio method, should be used.

Comparison of Valid Methods

For the hybrid methods, for $k = 1, 2$ we take $\delta_k \equiv 1$ and construct Z as in Section 3.3.2 with rates

$$\lambda_1^Z(\theta, z) = \begin{cases} \frac{1}{4} & z_A < 1 \\ \frac{1}{4} z_A & \text{otherwise} \end{cases}, \qquad \lambda_2^Z(\theta, z) = \begin{cases} 1 & z_A < 1 \\ z_A & \text{otherwise} \end{cases}.$$

The statement $z_A < 1$ for the state z is equivalent to $\lambda_1^X(\theta, z) = 0$, since the rates for the original process X are zero if and only if $z_A < \nu_{1A} = 1$, where ν_{1A} is the coordinate representing the number of A molecules in the source vector for the the first reaction.

Note that the process Z may now reach states where the first coordinate is negative. For $k = 3$ we simply set $\lambda_3^Z(\theta, x) = \lambda_k^X(\theta, x) = z_B$. We allow this rate $\lambda_3^Z(\theta, x)$ be zero because the reaction $B \to C$ is never interrupted by some other reaction, so that the Z constructed with these rates is still non-interruptive. This illustrates Remark 3.1. Note that, unlike the abundance of A, the abundances of B and C in Z will never be negative as long as $Z_{B,0}$ and $Z_{C,0}$ are nonnegative. Also note that these rates are uniformly bounded by a.

In Table 3.3, we include the details of the pathwise method simulations from Figure 3.2 with $a = 10$, for $t = 2$ and $t = 10$, and compare to CFD and LR estimators. Here and throughout, the half-width given for the confidence interval the hybrid method estimates is the sum of the half-widths from the two different estimates $\hat{M}^{LR}$ and $\hat{M}^{PW}$ from Section 3.3.3.

3.6.3 Gene Transcription and Translation

We now consider a model of mRNA transcription and its translation into protein. It is a more complex version of Example 1.1 from the introduction, since we will also include the dimerization of the protein. See also [3]. Table 3.4 gives the reactions of the model. Since the model does not satisfy the Non-interruptive Condition 3.8, the table also shows the rates that were used for the approximate process Z in the hybrid methods.

Dimer Abundance Sensitivity

We calculate the sensitivity $\frac{\partial}{\partial \theta_4} \mathbb{E} X_{D,t}(\theta)$, the derivative of the number of dimers at time t with respect to the parameter controlling the rate of mRNA degradation.

We compare the performance of the Dynkin Hybrid method and the RPD Hybrid

$t = 2$			
Method	w or ϵ	CPU	Estimate
Dynkin Hybrid	—	9.3	-2.64 ± .09
LR	—	9.0	-2.65 ± .08
RPD Hybrid	.02	9.2	-2.62 ± .19
RPD Hybrid	.2	9.4	-2.54 ± .10
RPD Hybrid	1	10	-2.50 ± .08
CFD	.1	36	-2.62 ± .04
CFD	.01	36	-2.54 ± .14

$t = 10$			
Method	w or ϵ	CPU	Estimate
Dynkin Hybrid	—	17	-6.57 ± .42
LR	—	10	-6.36 ± .08
RPD Hybrid	.1	18	-6.53 ± .59
RPD Hybrid	1	17	-6.41 ± .38
RPD Hybrid	5	19	-6.45 ± .35
CFD	.1	40	-6.37 ± .06
CFD	.01	39	-6.53 ± .22

Table 3.3: A comparison of the sensitivity methods on the switch model with $a = 10$. CPU gives computation time in seconds. $n = 5 \times 10^4$ paths were simulated; the hybrid methods used $n_1 = 5 \times 10^3, n_2 = 4.5 \times 10^4$ at $t = 2$ and $n_1 = 4 \times 10^4, n_2 = 1 \times 10^4$ at $t = 10$ (see Section 3.3.3). Actual sensitivities $\frac{\partial}{\partial \theta_1} X_{C,t}(\theta)$ at $\theta = (\frac{1}{4}, 1, 1)$ are -2.61 at $t = 2$ and -6.39 at $t = 10$. The horizontal rule separates the unbiased estimates (above) from the biased estimates. The LR method included the weight as a control variate to reduce variance. The CFD method used the centered difference. While the hybrid methods are comparable with the LR method at $t = 2$, at $t = 10$ the LR method is significantly more efficient than either hybrid method. For comparison, at $t = 2$, $\epsilon = .1$, and with 1.4×10^4 paths, the CFD method took 10 seconds to provide an estimate of $-2.56 \pm .08$. At $t = 10$, $\epsilon = .1$, and with 1.4×10^4 paths, the CFD method took 20 seconds to provide an estimate of $-6.48 \pm .09$.

Reaction		λ_k^X	λ_k^Z
transcription	$\emptyset \to M$	θ_1	θ_1
translation	$M \to M + P$	$\theta_2 X_M$	$\begin{cases} \theta_2 & Z_M < 1 \\ \theta_2 \tilde{M} & \theta_2 Z_M \geq \theta_2 \tilde{M} \\ \theta_2 Z_M & \text{otherwise} \end{cases}$
dimerization	$P + P \to D$	$\theta_3 X_P(X_P - 1)$	$\begin{cases} \theta_3 & Z_P < 2 \\ \theta_3 \tilde{M} & Z_P \geq 2 \quad \text{and} \\ & \theta_3 Z_P(Z_P - 1) \geq \theta_3 \tilde{M} \\ \theta_3 Z_P(Z_P - 1) & \text{otherwise} \end{cases}$
degradation	$M \to \emptyset$	$\theta_4 X_M$	$\begin{cases} \theta_4 \tilde{M} & \theta_4 Z_M \geq \theta_4 \tilde{M} \\ \theta_4 Z_M & \text{otherwise} \end{cases}$
degradation	$P \to \emptyset$	$\theta_5 X_P$	$\begin{cases} \theta_5 & Z_P < 1 \\ \theta_5 \tilde{M} & \theta_5 Z_P \geq \theta_5 \tilde{M} \\ \theta_5 Z_P & \text{otherwise} \end{cases}$
degradation	$D \to \emptyset$	$\theta_6 X_P$	$\begin{cases} \theta_6 \tilde{M} & \theta_6 Z_D \geq \theta_6 \tilde{M} \\ \theta_6 Z_D & \text{otherwise} \end{cases}$

Table 3.4: Reactions and rates for the hybrid methods for the dimerization model. We take all initial quantities equal to zero and $\tilde{M} = 10^6$ (we have added a tilde to the constant from Section 3.3.2 for notational reasons). For the process Z to be non-interruptive, only three of the intensities cannot be zero; see Remark 3.1. Note that in Z with $Z_0 = (0, 0, 0)$, the abundance of P may become negative, but abundances of M and D will not.

method to the likelihood ratio method and the CFD method. We also estimated the sensitivity using the RPD method alone, though recall that it is not expected to be a valid method on this model. We choose two different sets of parameter values, as the behavior of the various sensitivity methods is rather different on the two sets. See Table 3.5. In particular, note that the RPD method seems to give the correct answer for the first set of parameters, but not the second. This again shows that, *a priori*, one cannot know whether the RPD method will be valid if the model allows interruptions to occur. Also, note that for the first set of parameters, the hybrid methods are more efficient than the likelihood ratio method, but on the second set, they are comparable.

Integrated Dimerization Rate Sensitivity

We now estimate sensitivities of

$$\int_0^t \lambda_5(\theta, X_s)ds = \int_0^t \theta_5 X_{P,s}(X_{P,s} - 1)ds,$$

the integral of the rate of the dimerization reaction. This quantity is a functional of the path. Therefore, we use the hybrid method of Section 3.3 on this quantity directly. That is, we do not need to use Dynkin's formula or a time average, as we have in previous examples, to turn the quantity into the correct form. Also note that, unlike in previous examples, the functional depends explicitly on θ, which requires the methods to take into account the partial derivative of the functional in both the pathwise and likelihood ratio estimators. Again we compute the desired sensitivity at two different sets of parameter values. As Tables 3.6 and 3.7 show, the hybrid method is by far the most efficient.

$\theta = (200, 10, 0.01, 25, 1, 1)$			
Method	w or ϵ	CPU	Estimate
Dynkin Hybrid	—	89	-1.04 ± .04
LR	—	67	-0.99 ± .16
RPD Hybrid	.05	91	-1.08 ± .13
RPD Hybrid	.5	96	-1.00 ± .03
CFD	.25	112	-1.03 ± .06
CFD	.125	112	-1.01 ± .09
RPD	.05	83	-1.03 ± .11
RPD	.5	91	-1.00 ± .03

$\theta = (2, 10, 0.1, 1, 1, 0.1)$			
Method	w or ϵ	CPU	Estimate
Dynkin Hybrid	—	43	-69.7 ± 5.0
LR	—	30	-68.7 ± 4.2
RPD Hybrid	.3	44	-74.0 ± 6.1
RPD Hybrid	3	47	-70.3 ± 2.6
CFD	.01	50	-68.7 ± 4.9
CFD	.005	51	-78.2 ± 7.1
RPD	.3	37	-57.3 ± 4.1
RPD	3	41	-58.4 ± 1.7

Table 3.5: A comparison of the sensitivity methods on the dimerization model. CPU gives computation time in seconds; $n = 5\times10^3$ paths were simulated. The first sensitivity estimated was $\frac{\partial}{\partial\theta_4}\mathbb{E}X_{D,t}(\theta)$ at $t = 5$ and at the given value of θ. The second sensitivity estimated was $\frac{\partial}{\partial\theta_4}\mathbb{E}X_{D,t}(\theta)$ at $t = 30$ and the given value of θ. The horizontal rule separates the unbiased estimates (above) from the biased estimates. The RPD method is not valid for the model since interruptions can occur: it seems to give a correct estimate for the first set of parameters, but not the second. The LR method included the weight as a control variate. The CFD method used the centered difference. Both hybrid methods used $n_1 = 1000, n_2 = 4000$ (see Section 3.3.3). While the hybrid methods are for the most part more efficient than the LR method for first set of parameters, for the second set of parameters, the hybrid and LR methods are comparable.

$\theta = (200, 10, 0.01, 25, 1, 1)$				
	Hybrid	LR	CFD	CFD
∇_θ	0.57 ± 0.004	0.58 ± 0.04	0.57 ± 0.01	0.55 ± 0.04
	-4.5 ± 0.03	-4.6 ± 0.4	-4.5 ± 0.1	-4.5 ± 0.3
	11.4 ± 0.08	11.5 ± 0.6	11.4 ± 0.3	10.9 ± 0.6
	-55.6 ± 0.3	-55.5 ± 3.9	-55.5 ± 1.7	-56.7 ± 4.8
	3385 ± 19	3394 ± 256	3376 ± 143	3521 ± 434
	0.0 ± 0.0	-0.2 ± 2.1	0 ± 0	0 ± 0
CPU	261	85	138	138
$h\%$	–	–	.05	.01

$\theta = (2, 10, 0.1, 1, 1, 0.1)$				
	Hybrid	LR	CFD	CFD
∇_θ	94.1 ± 2.3	103.2 ± 4.9	108.1 ± 6.3	99.6 ± 12.3
	-184.3 ± 3.9	-209.1 ± 9.9	-210.8 ± 11.2	-204.3 ± 27.3
	23.7 ± 0.5	23.2 ± 2.7	22.1 ± 0.6	22.1 ± 1.5
	-87.5 ± 3.4	-71.5 ± 16.2	-79.6 ± 3.9	-80.0 ± 9.5
	449.1 ± 17.3	549.9 ± 145.4	417.8 ± 38.4	434.6 ± 91.3
	-0.5 ± 1.0	15.8 ± 117.2	0 ± 0	0 ± 0
CPU	125	36	61	63
$h\%$	–	–	.05	.01

Table 3.6: A comparison of sensitivity methods on the dimerization model. Estimates given with 95% confidence. CPU gives computation time in seconds. The first sensitivity gradient estimated was $\nabla_\theta \mathbb{E} \int_0^t \lambda_5(\theta, X_s) ds$ at $t = 5$ and at the first value of θ. The second sensitivity gradient estimated was of the same function, but at $t = 30$ and the second value of θ. The hybrid and LR methods are unbiased; CFD is not. The LR method included the weight as a control variate. The CFD method used the centered difference; the value in the $h\%$ row denotes that a perturbation of $(.05)\theta_i$, for example, was used for the i^{th} gradient entry. All methods used $n = 6000$ total paths; the hybrid method used $n_1 = 5500, n_2 = 500$ (see Section 3.3.3); the CFD methods used 1000 paths per gradient entry. The hybrid method, though it takes longer per path, is the most efficient. The hybrid method is still most efficient even if fewer paths are used, so that the method uses approximately the same computation time as the LR method; see Table 3.7.

$\theta = (200, 10, 0.01, 25, 1, 1)$		
	Hybrid	LR
∇_θ	0.57 ± 0.007 -4.5 ± 0.05 11.3 ± 0.1 -55.6 ± 0.6 3380 ± 34 0.0 ± 0.0	0.58 ± 0.04 -4.6 ± 0.4 11.5 ± 0.6 -55.5 ± 3.9 3394 ± 256 -0.2 ± 2.1
CPU	87	85
paths	$n_1 = 200$ $n_2 = 1800$	$n = 6000$

$\theta = (2, 10, 0.1, 1, 1, 0.1)$		
	Hybrid	LR
∇_θ	98.2 ± 5.0 -187.0 ± 7.5 23.9 ± 1.0 -85.7 ± 5.5 438.8 ± 33.1 5.0 ± 10.3	103.2 ± 4.9 -209.1 ± 9.9 23.2 ± 2.7 -71.5 ± 16.2 549.9 ± 145.4 15.8 ± 117.2
CPU	32	36
paths	$n_1 = 125$ $n_2 = 1375$	$n = 6000$

Table 3.7: A comparison of sensitivity methods on the dimerization model. See also Table 3.6. Estimates given with 95% confidence. CPU gives computation time in seconds. The first sensitivity gradient estimated was $\nabla_\theta \mathbb{E} \int_0^t \lambda_5(\theta, X_s) ds$ at $t = 5$ and at the first value of θ. The second sensitivity gradient estimated was of the same function, but at $t = 30$ and the second value of θ. The LR method included the weight as a control variate. To provide a more fair comparison between these two unbiased methods, we simulate the hybrid method so that approximately the same computation time is used as for the LR method. For the definition of n_1 and n_2, see see Section 3.3.3.

Chapter 4

A Coupling Method for Second Order Sensitivities

With the exception of Section 4.3, the contents of this chapter appeared in [42].

While first derivative sensitivities of CTMCs have been much studied, less focus has been given to finding reasonable algorithms for the computation of sensitivities of higher order. Second derivative sensitivities (the Hessian), however, are also particularly useful. For example, they provide concavity information which is necessary for finding roots or extrema of an expectation. In a more general optimization setting, the Hessian can be used to improve upon a simple steepest-descent method. Variants of Newton and quasi-Newton methods, for instance, include an approximate Hessian to use curvature to better choose the step direction in for the next iterate of the optimization. When the Hessian is positive semi-definite, these methods provably achieve a fast rate of local convergence. Additionally, trust-region based optimization methods can also be markedly improved by including a Hessian estimate [7, 27, 33, 38]. Developing algorithms that successfully integrate these optimization methods in the chemical reaction network and CTMC setting, for example in the context of parameter estimation, is a topic of current research which depends critically on having an efficient method for approximating the Hessian [39].

In Section 4.1, we discuss existing second order methods. In Section 4.2, we present a method for the computation of second order sensitivities of the CTMC (1.7). This method is an extension of the Coupled Finite Difference (CFD) method for first derivative sensitivities (see Section 2.1.3 or [2]) to the calculation of second derivative sensitivities; consequently the new method is termed the CFD2 method. CFD2 estimates have a significantly lower variance than the relevant extensions of any of the other existing first derivative methods of Chapter 2. That is, the CFD2 method requires much less CPU time to produce an approximation within a desired tolerance level. We give analytic results in Section 4.3, and in Section 4.4 we include numerical results from the CFD2 method and other methods for comparison.

4.1 Existing Second Order Sensitivity Methods

Recall that $\theta \in \mathbb{R}^R$ is some vector of parameters, that f is some function of interest, e.g. the abundance of some molecule, and that

$$J(\theta) := \mathbb{E} f(\theta, X_{t\in[0,T]}(\theta)).$$

We wish to estimate second derivatives of J.

One method of estimation is the likelihood ratio (LR) method of Section 2.2. Recall from that section the measures P^θ and Q, with the associated densities $G(\theta, t)$. Assuming that the relevant Leibnitz condition holds regarding integrability and differentiability, and that G is twice-differentiable in θ, we may simply take another derivative of the

equation (2.11) to obtain for some function h of the path that

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\mathbb{E}^{P_\theta}h(X_{t\in[0,T]}) &= \frac{d}{d\theta}\mathbb{E}^{Q}\left[h(X_{t\in[0,T]})\left(\frac{d}{d\theta}\log G(\theta,T)\right)G(\theta,T)\right] \\ &= \mathbb{E}^{Q}\left[h(X_{t\in[0,T]})G(\theta,T)\left(\frac{\partial^2}{\partial\theta^2}\log G(\theta,T)+\left(\frac{d}{d\theta}\log G(\theta,T)\right)^2\right)\right] \\ &= \mathbb{E}^{P_\theta}\left[h(X_{t\in[0,T]})H_2(\theta,T)\right]\end{aligned}$$

where the second order weight $H_2(\theta,t) := \frac{\partial^2}{\partial\theta^2}\log G(\theta,T) + \left(\frac{d}{d\theta}\log G(\theta,T)\right)^2$ is used analogously to the first order weight $H(\theta,T)$. Note that one can similarly write weights for mixed partial derivatives. Furthermore, the functions h can be replaced with functions $f(\theta, X_{t\in[0,T]})$, as in $J(\theta)$ above, that explicitly depend upon the parameters. As in (2.12), one can then use the product rule as long as all θ-derivatives of f are reasonably well-behaved.

The first order likelihood ratio method often produces estimates with prohibitively high variance, and this second order likelihood ratio method, unfortunately, produces estimates with even larger variance. Of course, regardless of the sensitivity method used, the variance of a second order sensitivity estimator will generally be larger than the variance of a first order estimator. Therefore, when computing second order sensitivities, it is more likely that one would be willing to accept a bias in return for a more efficient estimate. We thus turn to finite difference methods, in which the choice of perturbation h allows for some control over the variance. We do note that the hybrid pathwise method from Chapter 3 should also extend to second order sensitivities, though this possibility has not yet been explored.

The remainder of this chapter is thus concerned with finding an efficient computational method for the approximation of a second partial derivative of J, $\frac{\partial^2}{\partial\theta_j\partial\theta_i}J(\theta)$ for some $i, j \in \{1, \ldots, R\}$. Let e_i be the vector with a 1 in the i^{th} position and 0 elsewhere. We may approximate this sensitivity by perturbing the parameter vector in both relevant directions, so that

$$\frac{\partial^2}{\partial\theta_j\partial\theta_i}J(\theta) = \frac{J(\theta + (e_i + e_j)\epsilon) - J(\theta + e_i\epsilon) - J(\theta + e_j\epsilon) + J(\theta)}{\epsilon^2} + O(\epsilon).$$

In our setting, this suggests we approximate the second derivative sensitivity by

$$\begin{aligned}\epsilon^{-2}\mathbb{E}\big[f(\theta + (e_i + e_j)\epsilon, X_{t\in[0,T]}(\theta + (e_i + e_j)\epsilon)) - f(\theta + e_i\epsilon, X_{t\in[0,T]}(\theta + e_i\epsilon))\\ - f(\theta + e_j\epsilon, X_{t\in[0,T]}(\theta + e_j\epsilon)) + f(\theta, X_{t\in[0,T]}(\theta))\big],\end{aligned} \tag{4.1}$$

where $X(\cdot) := X_{t\in[0,T]}(\cdot)$ is the path through the terminal time T. The Monte Carlo estimator for (4.1) with n estimates is then

$$\hat{M}_n(\epsilon) = \frac{1}{n}\sum_{\ell=1}^{n} M_\ell(\epsilon), \tag{4.2}$$

where

$$\begin{aligned}M_\ell(\epsilon) := \epsilon^{-2}\big[f(\theta, X_{t\in[0,T],\ell}(\theta + (e_i + e_j)\epsilon)) - f(\theta, X_{t\in[0,T],\ell}(\theta + e_i\epsilon))\\ - f(\theta, X_{t\in[0,T],\ell}(\theta + e_j\epsilon)) + f(\theta, X_{t\in[0,T],\ell}(\theta))\big],\end{aligned}$$

where, for example, $X_{t\in[0,T],\ell}(\theta)$ is the ℓ^{th} path simulated with parameter choice θ. Thus, for second derivatives, finite difference methods require up to four simulated paths to

produce one estimate, as opposed to the likelihood ratio method, which requires only one path per estimate. When coupling methods are used with the finite difference, however, the variance of the estimates produced are usually significantly lower than the variance of likelihood ratio estimates, as demonstrated in Section 4.4, so that finite different methods often provide much more effective estimators.

Note that if the four relevant processes are computed independently, which is the extension to second order sensitivities of the IRN method, the variance of the estimator $\hat{M}_n(\epsilon)$ is $n^{-1}\mathrm{Var}(M(\epsilon)) = O(n^{-1}\epsilon^{-4})$. This is computed analogously to the work in Section 2.1. As usual the goal of the coupling is to lower the variance of $M(\epsilon)$ by correlating the relevant processes.

The CRN method for second order sensitivities is nearly the same as the first order method discussed in Section 2.1.1: one reuses the same stream of random numbers to produce each of the four paths in (4.1) using the SSA method. Similarly, the CRP method of Section 2.1.2 is extended to second order sensitivities by using same instances of the K independent Poisson processes Y_k in the four paths in (4.1). However, as the numerical examples in Section 4.4 demonstrate, the variance of the CRN and CRP estimators grow significantly in time; in many cases the asymptotic variance is of the same order as the IRN method, because paths constructed using the CRN or CRP coupling techniques tend to de-couple over time. Therefore, we find an extension of the CFD method to second order sensitivities.

For ease of exposition and notation, we have described finite differences using the forward difference (4.1), which as noted has an $O(\epsilon)$ bias. Our formal construction will also use the forward difference. In practice, however, it is no more difficult to use the

central second difference,

$$\begin{aligned}\epsilon^{-2}\big[&f(\theta+(e_i+e_j)\epsilon/2, X_{t\in[0,T]}(\theta+(e_i+e_j)\epsilon/2))\\&-f(\theta+(e_i-e_j)\epsilon/2, X_{t\in[0,T]}(\theta+(e_i-e_j)\epsilon/2))\\&-f(\theta+(e_j-e_i)\epsilon/2, X_{t\in[0,T]}(\theta+(e_j-e_i)\epsilon/2))\\&+f(\theta-(e_i+e_j)\epsilon/2, X_{t\in[0,T]}(\theta-(e_i+e_j)\epsilon/2))\big],\end{aligned} \tag{4.3}$$

which has a bias of only $O(\epsilon^2)$; this is what we have implemented in our numerical examples.

4.2 The 2nd Order Coupled Finite Difference Method

The integral part of this method, called the CFD2 method, is the coupling presented in detail below. We will show that on a large class of functions f, this CFD2 method lowers the variance of the numerator of $M(\epsilon)$ to $O(\epsilon)$, thereby lowering the variance of $M(\epsilon)$ to $O(\epsilon^{-3})$ and yielding $\mathrm{Var}(\hat{M}_n(\epsilon)) = O(n^{-1}\epsilon^{-3})$. For several non-trivial examples, however, the method gives even better performance, lowering the variance of $\hat{M}(\epsilon)$ another order of magnitude to $O(\epsilon^{-2})$, by lowering the variance of the numerator to $O(\epsilon^2)$. We will prove that the CFD2 method achieves a variance of the lower order $O(\epsilon^{-2})$ when the intensities are linear. We do not believe linearity to be necessary; numerical examples such as that in Section 4.4.4 suggest that the method achieves the faster rate on other models as well.

In contrast, every other coupling method we attempted[1] yielded an asymptotic variance for $\hat{M}_n(\epsilon)$ of $O(\epsilon^{-3})$ at best, and in general these couplings were much less efficient

[1]We do not provide a full list of other, less efficient couplings, of which there are many.

than the coupling being proposed here. While the $O(\epsilon^{-3})$ variance of the CFD2 method is sharp, numerical examples such as that in Section 4.4.3 show that even for models for which the CFD2 method is $O(\epsilon^{-3})$, the variance of CFD2 estimates is much lower than estimates computed via other methods.

Furthermore, the paths generated when using the CFD2 method can also be re-used to compute first derivatives of the system, for use in an optimization algorithm, for example. Additionally, the CFD2 algorithm is reasonably straightforward to implement, and, unlike the other finite difference methods discussed above, the CFD2 method produces a coupled process that is still a CTMC, which allows for an analysis using martingale methods.

The crucial piece of the second order Coupled Finite Difference (CFD2) method is the coupling of the four processes in the difference (4.1). We now construct this coupling, which is an extension of the standard coupling of Section 1.1.2. As with the usual standard coupling, note that the new coupling given here may be applicable in other contexts, as it is in no way specific to the four processes of (4.1).

4.2.1 Construction of the Coupling

For the computation of second derivatives, rather than correlated pairs of runs, correlated quartets of runs are used. We suppose we have the four CTMCs of (4.1), with $i, j \in \{1, \ldots, R\}$ fixed, which for convenience of exposition we order as

$$X(\theta + (e_i + e_j)\epsilon),\ X(\theta + e_i\epsilon),\ X(\theta + e_j\epsilon),\ X(\theta). \tag{4.4}$$

We also assume that their initial conditions are equal (i.e., they are equal at $t = 0$), to some fixed value X_0. For each of the four processes above, there is an associated propensity for each of the K reaction channels. For example, the propensity of the kth reaction channel of the first process (the one with parameter choice $\theta + (e_i + e_j)\epsilon$) is

$$\lambda_{k,1} := \lambda_k(\theta + (e_i + e_j)\epsilon, X_t(\theta + (e_i + e_j)\epsilon)).$$

Similarly rename the propensity of the kth reaction channel of the second process (parameter choice $\theta + e_i\epsilon$) by $\lambda_{k,2}$, the third process as $\lambda_{k,3}$, and the fourth process as $\lambda_{k,4}$, as per our ordering (4.4). Note that these propensities are dependent on θ and $X_t(\theta)$, but for notational convenience we will drop either or both of these dependencies in our notation when they are not necessary for the current discussion.

Next, we introduce a coupling of these four processes that will produce an estimator (4.2) with low variance. The main idea is similar to the standard coupling from Section 1.1.2 in that it rests on splitting a counting process into sub-processes, to be shared among the four CTMCs (4.4). We create a sub-process to allow the 1st and 2nd processes of (4.4) to jump simultaneously, one to allow the 1st and 3rd to jump simultaneously, one for the 2nd and 4th, and one for the 3rd and 4th. Additionally, we create a sub-process that allows all four to jump simultaneously. As in the first derivative setting, the rates of these sub-processes will involve minima of the original CTMCs. Finally, we also require four additional sub-processes to make up any "leftover" propensity of the original CTMCs.

Formally, define $R_{k,[b_1,b_2,b_3,b_4]}$ as a counting process, where $b_\ell \in \{0, 1\}$. A jump of $R_{k,[b_1,b_2,b_3,b_4]}$ indicates that the ℓth process in the ordering (4.4) jumps by reaction k if

and only if $b_\ell = 1$, for $\ell \in \{1,2,3,4\}$. For example, $R_{k,[1,1,0,0]}(t)$ counts the number of times the kth reaction has fired simultaneously for the first and second processes of (4.4) (but the third and fourth did not fire), whereas $R_{k,[1,0,1,0]}(t)$ counts the number of times the kth reaction has fired simultaneously for the first and third processes of (4.4) (but the second and fourth did not fire). Define the propensity of $R_{k,[b_1,b_2,b_3,b_4]}$ by $\Lambda_{k,[b_1,b_2,b_3,b_4]}$, so that in the random time change representation,

$$R_{k,[b_1,b_2,b_3,b_4]}(t) = Y_{k,[b_1,b_2,b_3,b_4]}\left(\int_0^t \Lambda_{k,[b_1,b_2,b_3,b_4]}(s)ds\right) \tag{4.5}$$

where the Y's are independent unit-rate Poisson processes and where the propensities are

$$\begin{aligned}
\Lambda_{k,[1,1,1,1]} &= \lambda_{k,1} \wedge \lambda_{k,2} \wedge \lambda_{k,3} \wedge \lambda_{k,4} \\
\Lambda_{k,[1,1,0,0]} &= \lambda_{k,1} \wedge \lambda_{k,2} - \Lambda_{k,[1,1,1,1]} \\
\Lambda_{k,[0,0,1,1]} &= \lambda_{k,3} \wedge \lambda_{k,4} - \Lambda_{k,[1,1,1,1]} \\
\Lambda_{k,[1,0,1,0]} &= (\lambda_{k,1} - \lambda_{k,1} \wedge \lambda_{k,2}) \wedge (\lambda_{k,3} - \lambda_{k,3} \wedge \lambda_{k,4}) \\
\Lambda_{k,[0,1,0,1]} &= (\lambda_{k,2} - \lambda_{k,1} \wedge \lambda_{k,2}) \wedge (\lambda_{k,4} - \lambda_{k,3} \wedge \lambda_{k,4}) \\
\Lambda_{k,[1,0,0,0]} &= (\lambda_{k,1} - \lambda_{k,1} \wedge \lambda_{k,2}) - \Lambda_{k,[1,0,1,0]} \\
\Lambda_{k,[0,1,0,0]} &= (\lambda_{k,2} - \lambda_{k,1} \wedge \lambda_{k,2}) - \Lambda_{k,[0,1,0,1]} \\
\Lambda_{k,[0,0,1,0]} &= (\lambda_{k,3} - \lambda_{k,3} \wedge \lambda_{k,4}) - \Lambda_{k,[1,0,1,0]} \\
\Lambda_{k,[0,0,0,1]} &= (\lambda_{k,4} - \lambda_{k,3} \wedge \lambda_{k,4}) - \Lambda_{k,[0,1,0,1]},
\end{aligned} \tag{4.6}$$

where we recall that $a \wedge b := \min\{a, b\}$.

The proposed coupling is then given by the following:

$$\begin{aligned}
X_t(\theta + (e_i + e_j)\epsilon) &= X_0(\theta) + \sum_k \zeta_k(R_{k,[1,1,1,1]}(t) + R_{k,[1,1,0,0]}(t) + R_{k,[1,0,1,0]}(t) + R_{k,[1,0,0,0]}(t)) \\
X_t(\theta + e_i\epsilon) &= X_0(\theta) + \sum_k \zeta_k(R_{k,[1,1,1,1]}(t) + R_{k,[1,1,0,0]}(t) + R_{k,[0,1,0,1]}(t) + R_{k,[0,1,0,0]}(t)) \\
X_t(\theta + e_j\epsilon) &= X_0(\theta) + \sum_k \zeta_k(R_{k,[1,1,1,1]}(t) + R_{k,[0,0,1,1]}(t) + R_{k,[1,0,1,0]}(t) + R_{k,[0,0,1,0]}(t)) \\
X_t(\theta) &= X_0(\theta) + \sum_k \zeta_k(R_{k,[1,1,1,1]}(t) + R_{k,[0,0,1,1]}(t) + R_{k,[0,1,0,1]}(t) + R_{k,[0,0,0,1]}(t)).
\end{aligned} \tag{4.7}$$

A few comments are in order. First, note that, for example, the marginal process $X_t(\theta+(e_i+e_j)\epsilon)$ above involves all the counting processes in which $b_1 = 1$. Second, each of these marginal processes $X_t(\cdot)$ have the same distribution as the original, uncoupled, processes since the transition rates of the marginal processes have remained unchanged. This can be checked by simply summing the rates of the relevant counting processes, which are all those $\Lambda_{k,[b_1,b_2,b_3,b_4]}$ in which a given $b_\ell = 1$. Third, if f is linear, for example if we are estimating the abundance of a particular molecule, many of the $R_{k,[b_1,b_2,b_3,b_4]}$ are completely cancelled if we now construct the difference (4.1). An example of this will be shown in Section 4.4.1. Fourth, even if $i = j$, the coupling requires two different copies of the process $X_t(\theta + e_i\epsilon)$, one taking the role of $X_t(\theta + e_i\epsilon, t)$ and the other $X_t(\theta + e_j\epsilon)$.

4.2.2 An Alternative Construction of the Coupling

The coupling described in the previous section can be derived in an alternate way, which explains why the method is termed "double coupled." We could first couple the first and second processes of (4.4) using the standard coupling, and then couple the third

and fourth in the same manner. For example, using the $\lambda_{k,\ell}$ as defined in the previous section, the first two processes in (4.4) are constructed as:

$$\begin{aligned} X(\theta + (e_i + e_j)\epsilon, t) &= X_0(\theta) + \sum_k \left(R_{k,[1,1]} + R_{k,[1,0]}\right)\zeta_k \\ X(\theta + e_i\epsilon, t) &= X_0(\theta) + \sum_k \left(R_{k,[1,1]} + R_{k,[0,1]}\right)\zeta_k, \end{aligned} \tag{4.8}$$

where $R_{k,[b_1,b_2]} = Y_{k,[b_1,b_2]}\left(\int_0^t \Lambda_{k,[b_1,b_2]}(s)ds\right)$ are defined analogously to (4.5) and where

$$\begin{aligned} \Lambda_{k,[1,1]}(s) &= \lambda_{k,1}(s) \wedge \lambda_{k,2}(s), \\ \Lambda_{k,[1,0]}(s) &= \lambda_{k,1}(s) - \lambda_{k,1}(s) \wedge \lambda_{k,2}(s), \\ \Lambda_{k,[0,1]}(s) &= \lambda_{k,2}(s) - \lambda_{k,1}(s) \wedge \lambda_{k,2}(s). \end{aligned}$$

The processes defined in (4.8) jump together as often as possible: they share the subprocesses $R_{k,[1,1]}$, each of which runs at a propensity equal to the minimum of the respective propensities of the two original processes. We then expect the variance of the first finite difference $[f(\theta, X_t(\theta + (e_i + e_j)\epsilon)) - f(\theta, X_t(\theta + e_i\epsilon))]\epsilon^{-1}$ to be small since the two processes of (4.8) will remain approximately the same whenever they jump simultaneously via $R_{k,[1,1]}$.

Now note that, together, the two processes (4.8) can be viewed as a new CTMC with dimension $2d$, twice that of that of the original process. The third and fourth processes in (4.4) can be similarly coupled, giving us *two* $2d$-dimensional CTMCs. Finally, we couple these new processes into a single CTMC of dimension $4d$, using the standard coupling a third time. This construction leads to the same process as given in the previous section. The details are left to the interested reader.

4.2.3 The CFD2 Algorithm

We present two algorithms for the simulation of paths coupled as in Section 4.2. The first utilizes NRM simulation, whereas the second utilizes SSA (see Section 1.2.1). As usual, it will be problem specific as to which algorithm is most efficient.

Below, rand(0,1) indicates a uniform[0,1] random variable, independent from all previous random variables. Recall that if $U \sim$ rand(0,1), then $\ln(1/U)/\lambda$ is exponentially distributed with parameter $\lambda > 0$. Also recall that even if i and j are equal, the processes $X(\theta + e_i\epsilon)$ and $X(\theta + e_j\epsilon)$ are still constructed separately. Define the set

$$\begin{aligned} B :=\{&[1,1,1,1],[1,1,0,0],[0,0,1,1],[1,0,1,0],\\ &[0,1,0,1],[1,0,0,0],[0,1,0,0],[0,0,1,0],[0,0,0,1]\} \end{aligned}$$

and note that it will often be convenient to use a *for* loop, from 1 to 9, to enumerate over the vectors in B.

Algorithm-modified next reaction method applied to (4.7).

Initialization: Set $X(\theta+(e_i+e_j)\epsilon) = X(\theta+e_i\epsilon) = X(\theta+e_j\epsilon) = X(\theta) = X_0$ and $t = 0$; for each $k \in \{1,\ldots,M\}$ and each $b \in B$, set $T_{k,b} = 0$ and $P_{k,b} = \ln(1/u_{k,b})$ for $u_{k,b} \sim$ rand(0,1).

Repeat the following steps:

(i) For each k, set

$$\lambda_{k,1} = \lambda_k(\theta + (e_i + e_j)\epsilon, X(\theta + (e_i + e_j)\epsilon))$$
$$\lambda_{k,2} = \lambda_k(\theta + e_i\epsilon, X(\theta + e_i\epsilon))$$
$$\lambda_{k,3} = \lambda_k(\theta + e_j\epsilon, X(\theta + e_j\epsilon))$$
$$\lambda_{k,4} = \lambda_k(\theta, X(\theta))$$

and use to set each of the nine variables $\Lambda_{k,b}$ as above in (4.6).

(ii) For each k and $b \in B$, set

$$\Delta t_{k,b} = \begin{cases} (P_{k,b} - T_{k,b})/\Lambda_{k,b} & , \quad \text{if } \Lambda_{k,b} > 0 \\ \infty & , \quad \text{else} \end{cases}.$$

(iii) Set $\Delta = \min_{k,b}\{\Delta t_{k,b}\}$ and let $\mu := k$ and $\nu := b = [b_1, b_2, b_3, b_4]$ be the indices where the minimum is achieved.

(iv) Set $t = t + \Delta$.

(v) Update state vector variables $X(\theta + (e_i + e_j)\epsilon), X(\theta + e_i\epsilon), X(\theta + e_j\epsilon), X(\theta)$ by adding ζ_μ to the ℓth process if and only if $b_\ell = 1$ in ν.

(vi) For each k and $b \in B$, set $T_{k,b} = T_{k,b} + \Delta \cdot \Lambda_{k,b}$.

(vii) Set $P_{\mu,\nu} = P_{\mu,\nu} + \ln(1/u)$ where $u \sim \text{rand}(0,1)$.

(viii) Return to (i) or quit.

ALGORITHM-GILLESPIE'S DIRECT METHOD APPLIED TO (4.7).

Initialization: Set $X(\theta+(e_i+e_j)\epsilon) = X(\theta+e_i\epsilon) = X(\theta+e_j\epsilon) = X(\theta) = X_0$ and $t = 0$.

Repeat the following steps:

(i) For each k, set

$$\lambda_{k,1} = \lambda_k(\theta + (e_i + e_j)\epsilon, X(\theta + (e_i + e_j)\epsilon))$$
$$\lambda_{k,2} = \lambda_k(\theta + e_i\epsilon, X(\theta + e_i\epsilon))$$
$$\lambda_{k,3} = \lambda_k(\theta + e_j\epsilon, X(\theta + e_j\epsilon))$$
$$\lambda_{k,4} = \lambda_k(\theta, X(\theta))$$

and use to set each of the nine variables $\Lambda_{k,b}$ as above in (4.6).

(ii) Let $\Lambda_0 = \sum_k \sum_b \Lambda_{k,b}$ and $u \sim \text{rand}(0,1)$, and set

$$\Delta = \ln(1/u)/\Lambda_0.$$

(iii) Set $t = t + \Delta$.

(iv) Let $u \sim \text{rand}(0,1)$ and use to select $(\mu, \nu) \in \{(k,b) : k \in \{1,\ldots,M\}, b \in B\}$ where each pair (k,b) is selected with probability $\lambda_{k,b}/\Lambda_0$.[2]

(v) Update state vector variables $X(\theta + (e_i + e_j)\epsilon), X(\theta + e_i\epsilon), X(\theta + e_j\epsilon), X(\theta)$ by adding ζ_μ to the ℓth process if and only if $b_\ell = 1$ in ν.

[2]This is the usual "binning" step of SSA.

(vi) Return to (i) or quit.

4.3 Analytical Results

In this section, we will show that the CFD2 method produces Monte Carlo estimates μ_n such that $\mathrm{Var}(\mu_n(\epsilon)) = O(n^{-1}\epsilon^{-3})$, as discussed in Section 4.1. This result follows immediately from results from the CFD method which show that the difference of the coupled paths in the numerator of the estimates is $O(\epsilon)$.

We will also show that the CFD2 method achieves a variance of the lower order $O(\epsilon^{-2})$ when the intensities are linear. This follows if we can show that the difference of the coupled paths in the numerator of the second difference (4.1) is $O(\epsilon^2)$. We do not believe that linear intensities are a necessary condition on this lower order of variance. However, the faster rate is not always achieved, showing that the $O(n^{-1}\epsilon^{-3})$ variance bound is sharp. Numerical examples in Section 4.4 give examples of both situations.

We prove these results for functions f of a particular form, though again numerical examples suggest that the results hold for other f as well.

4.3.1 Assumptions

Define $F^\theta(x) := \sum_k \lambda_k(\theta, x)\zeta_k$. We assume the following conditions on the intensities, which are essentially as given in [2]. Recall that K is the number of reactions in the model, d is the dimension of the process X, and R is the dimension of the parameter θ.

Condition 4.29. *There exists some $K_1 > 0$ so that for all $k = 1, \ldots, K$, all $x, y \in \mathcal{S}$,*

and all relevant θ we have

$$|\lambda_k(\theta, x) - \lambda_k(\theta, y)| + |F^\theta(x) - F^\theta(y)| \leq K_1|x - y|.$$

Condition 4.30. *There exists some $K_2 > 0$ so that for all $k = 1, \ldots, K$, all $\epsilon \in (0, 1)$, all $i = 1, \ldots, R$, and all relevant θ we have*

$$\sup_{x \in \mathcal{S}} \left[|\lambda_k(\theta + e_i\epsilon, x) - \lambda_k(\theta, x)| + |F^{\theta + e_i\epsilon}(x) - F^\theta(x)| \right] \leq K_2\epsilon.$$

Furthermore, we assume that f is of the following form.

Condition 4.31. *The function $f : \mathbb{R}^d \to \mathbb{R}$ is a C^1 function with bounded first derivative.*

That is, the output f of the system is a function of the state at some fixed terminal time T, rather than the entire path through that time. Then we are estimating second derivative sensitivities of $J(\theta) = f(X_T(\theta))$. Note that very often f is taken to be simply a coordinate map giving the abundance of a given species at a given time; that f certainly satisfies this condition.

Most, if not all, of these conditions can likely be relaxed to local rather than global conditions.

4.3.2 Proofs on the Order of the Variance

We first prove in Theorem 4.33 that the difference of paths in (4.1) is $O(\epsilon)$. This will require only one short lemma. The bulk of the work lies in the proof of Theorem 4.34, which shows that in some cases this variance is only $O(\epsilon^2)$.

Lemma 4.32. *Given the four processes coupled as in (4.7), and satisfying Conditions 4.29 and 4.30, and given some function f satisfying Condition 4.31, for a fixed $T > 0$, there exist some $C_{T,K,f} > 0$ and $D_{T,K,f} > 0$ such that*

$$\mathbb{E} \sup_{t \leq T} |f(X_t(\theta + (e_i + e_j)\epsilon)) - f(X_t(\theta + e_i\epsilon))| \leq C_{T,K,f}\epsilon,$$

$$\mathbb{E} \sup_{t \leq T} |f(X_t(\theta + e_j\epsilon)) - f(X_t(\theta))| \leq C_{T,K,f}\epsilon,$$

$$\mathbb{E} \sup_{t \leq T} |f(X_t(\theta + (e_i + e_j)\epsilon)) - f(X_t(\theta + e_i\epsilon))|^2 \leq D_{T,K,f}\epsilon, \text{ and}$$

$$\mathbb{E} \sup_{t \leq T} |f(X_t(\theta + e_j\epsilon)) - f(X_t(\theta))|^2 \leq D_{T,K,f}\epsilon.$$

Proof. This is proved in the context of the CFD method in previous work of Anderson in [2] on the standard coupling for the first difference; the results apply here since the CFD2 method begins by coupling these pairs of processes as in Section 4.2.2. □

Theorem 4.33. *Given the four processes coupled as in (4.7), and satisfying Conditions 4.29 and 4.30, and given some function f satisfying Condition 4.31, for a fixed $T > 0$, there exists some $C_{T,M,f} > 0$ such that*

$$\mathbb{E} \sup_{t \leq T} |f(X_t(\theta + (e_i + e_j)\epsilon)) - f(X_t(\theta + e_i\epsilon)) - f(X_t(\theta + e_j\epsilon)) + f(X_t(\theta))|^2 \leq C_{T,M,f}\epsilon.$$

Proof. Use that

$$\begin{aligned} &|f(X_t(\theta+(e_i + e_j)\epsilon)) - f(X_t(\theta + e_i\epsilon)) - f(X_t(\theta + e_j\epsilon)) + f(X_t(\theta))|^2 \\ &\qquad \leq 2|f(X_t(\theta + (e_i + e_j)\epsilon)) - f(X_t(\theta + e_i\epsilon))|^2 + 2|f(X_t(\theta + e_j\epsilon)) - f(X_t(\theta))|^2. \end{aligned}$$

The proof follows by properties of the supremeum, and Lemma 4.32. □

As the second moment provides an upper bound on the variance, a direct consequence of this theorem is that the CFD2 method in general provides an estimator with a variance of $O(\epsilon^{-3})$. This bound is sharp.

In many models, however, the variance of the difference of paths may be reduced a full order of magnitude further to scale instead with ϵ^2, so that the difference estimator has instead a variance of $O(\epsilon^{-2})$. We prove this for models whose intensities are linear, though we do not believe this to be a necessary condition. For example, models including unary reactions with mass action kinetics have linear intensities. We will additionally require that the function f is also linear. Note that linear intensities satisfy Conditions 4.29 and 4.30, and linear functions f satisfy Condition 4.31.

Theorem 4.34. *Given the four processes coupled as in (4.7) with intensities linear in both the parameter and state variables, and given some linear function $f : \mathbb{R}^d \to \mathbb{R}$, given any $T > 0$, there is a positive constant $C_{T,M,f}$ such that*

$$\mathbb{E} \sup_{t \leq T} |f(X_t(\theta + (e_i + e_j)\epsilon)) - f(X_t(\theta + e_i\epsilon)) - f(X_t(\theta + e_j\epsilon)) + f(X_t(\theta))|^2 \leq C_{T,M,f}\epsilon^2.$$

The proof proceeds as follows. We first prove a lemma very similar to Lemma 4.32 above, giving bounds on other pairs of the four processes. We note that these preliminary bounds do not require the intensities to be linear. We next prove a messy but useful result in Lemma 4.36, and then Lemmas 4.37 and 4.38 to give L^1 and L^2 bounds on the difference in (4.1). Finally we will be able to prove the theorem itself.

We also note here that the following proofs consider $i \neq j$ so that the result holds for mixed partial derivatives. If, however, one assumes i and j are equal, the proof still

holds (and much of the work simplifies).

Lemma 4.35. *Given the four processes coupled as in (4.7), and satisfying Conditions 4.29 and 4.30, and given some function f satisfying Condition 4.31, for a fixed $T > 0$, there exist some $C_{T,K,f} > 0$ and $D_{T,K,f} > 0$ such that*

$$\mathbb{E}\sup_{t\leq T}|f(X_t(\theta+(e_i+e_j)\epsilon))-f(X_t(\theta))|\leq C_{T,M,f}\epsilon \text{ and}$$

$$\mathbb{E}\sup_{t\leq T}|f(X_t(\theta+e_i\epsilon))-f(X_t(\theta+e_j\epsilon))|\leq C_{T,M,f}\epsilon$$

$$\mathbb{E}\sup_{t\leq T}|f(X_t(\theta+(e_i+e_j)\epsilon))-f(X_t(\theta))|^2\leq D_{T,M,f}\epsilon \text{ and}$$

$$\mathbb{E}\sup_{t\leq T}|f(X_t(\theta+e_i\epsilon))-f(X_t(\theta+e_j\epsilon))|^2\leq D_{T,M,f}\epsilon.$$

Proof. For the first bound, add and subtract $f(X_t(\theta + e_i\epsilon))$ inside the absolute value and use the triangle inequality along with Lemma 4.33, noting that i, j are arbitrary. The other bounds follow similarly, with the last two additionally using that $(x + y)^2 \leq 2x^2 + 2y^2$. □

Lemma 4.36. *Given the four processes coupled as in (4.7) with intensities linear in both the parameter and state variables, given $s > 0$, there exists some non-negative constants C_1 through C_3 such that for any k we have*

$$\begin{aligned}
&|\lambda_k(\theta+(e_i+e_j)\epsilon, X_s(\theta+(e_i+e_j)\epsilon))-\lambda_k(\theta+e_i\epsilon, X_s(\theta+e_i\epsilon))\\
&\quad-\lambda_k(\theta+e_j\epsilon, X_s(\theta+e_j\epsilon))+\lambda_k(\theta, X_s(\theta))|\\
&\qquad\leq C_1|X_s(\theta+(e_i+e_j)\epsilon)-X_s(\theta+e_i\epsilon))-X_s(\theta+e_j\epsilon)+X_s(\theta)|\\
&\qquad\quad+C_2\epsilon|X_s(\theta+(e_i+e_j)\epsilon)-X_s(\theta)|+C_3\epsilon|X_s(\theta+e_i\epsilon)-X_s(\theta+e_j\epsilon)|.
\end{aligned}$$

Proof. Fix k. First we note that the quantity whose mean is to be bounded is clearly no more than

$$|2\lambda_k(\theta + (e_i + e_j)\epsilon, X_s(\theta + (e_i + e_j)\epsilon)) - 2\lambda_k(\theta + e_i\epsilon, X_s(\theta + e_i\epsilon)) \\ - 2\lambda_k(\theta + e_j\epsilon, X_s(\theta + e_j\epsilon)) + 2\lambda_k(\theta, X_s(\theta))|$$

This allows us to now add and subtract several new terms and to group the terms into fours that behave well. Indeed, one can check that without absolute values the following bound is an equality:

$$\begin{aligned}
&|2\lambda_k(\theta + (e_i + e_j)\epsilon, X_s(\theta + (e_i + e_j)\epsilon)) - 2\lambda_k(\theta + e_i\epsilon, X_s(\theta + e_i\epsilon)) \\
&\qquad - 2\lambda_k(\theta + e_j\epsilon, X_s(\theta + e_j\epsilon)) + 2\lambda_k(\theta, X_s(\theta))| \\
&\leq |\lambda_k(\theta + (e_i + e_j)\epsilon, X_s(\theta + (e_i + e_j)\epsilon)) - \lambda_k(\theta + e_i\epsilon, X_s(\theta + (e_i + e_j)\epsilon)) \\
&\qquad - \lambda_k(\theta + e_j\epsilon, X_s(\theta + (e_i + e_j)\epsilon)) + \lambda_k(\theta, X_s(\theta + (e_i + e_j)\epsilon))| \\
&+ |\lambda_k(\theta + (e_i + e_j)\epsilon, X_s(\theta)) - \lambda_k(\theta + e_i\epsilon, X_s(\theta)) - \lambda_k(\theta + e_j\epsilon, X_s(\theta)) + \lambda_k(\theta, X_s(\theta))| \\
&+ |\lambda_k(\theta + (e_i + e_j)\epsilon, X_s(\theta + (e_i + e_j)\epsilon)) - \lambda_k(\theta, X_s(\theta + (e_i + e_j)\epsilon)) \\
&\qquad - \lambda_k(\theta + (e_i + e_j)\epsilon, X_s(\theta)) + \lambda_k(\theta, X_s(\theta))| \\
&+ |\lambda_k(\theta + e_i\epsilon, X_s(\theta + e_j\epsilon)) - \lambda_k(\theta + e_j\epsilon, X_s(\theta + e_j\epsilon)) \\
&\qquad - \lambda_k(\theta + e_i\epsilon, X_s(\theta + e_i\epsilon)) + \lambda_k(\theta + e_j\epsilon, X_s(\theta + e_i\epsilon))| \\
&+ |\lambda_k(\theta + e_i\epsilon, X_s(\theta + (e_i + e_j)\epsilon)) - \lambda_k(\theta + e_i\epsilon, X_s(\theta + e_i\epsilon)) \\
&\qquad - \lambda_k(\theta + e_i\epsilon, X_s(\theta + e_j\epsilon)) + \lambda_k(\theta + e_i\epsilon, X_s(\theta))| \\
&+ |\lambda_k(\theta + e_j\epsilon, X_s(\theta + (e_i + e_j)\epsilon)) - \lambda_k(\theta + e_j\epsilon, X_s(\theta + e_i\epsilon)) \\
&\qquad - \lambda_k(\theta + e_j\epsilon, X_s(\theta + e_j\epsilon)) + \lambda_k(\theta + e_j\epsilon, X_s(\theta))|
\end{aligned}$$

Note that intensities in first term differ only in the first variable of the intensities, and so since λ is linear this term is in fact zero. The same is true of the second term. Also by linearity, the third term can be rewritten and then bound as follows:

$$|\lambda_k((e_i+e_j)\epsilon, X_s(\theta+(e_i+e_j)\epsilon)-X_s(\theta))| \leq C\epsilon|X_s(\theta+(e_i+e_j)\epsilon)-X_s(\theta)|$$

where C is some constant. The fourth term can bounded similarly but with the quantity $|X_s(\theta+e_i\epsilon)-X_s(\theta+e_j\epsilon)|$.

For the fifth term, linearity we have for some C that

$$\begin{aligned}|\lambda_k(\theta+e_i\epsilon, X_s(\theta+(e_i+e_j)\epsilon))-\lambda_k(\theta+e_i\epsilon, X_s(\theta+e_i\epsilon))&\\ -\lambda_k(\theta+e_i\epsilon, X_s(\theta+e_j\epsilon))+\lambda_k(\theta+e_i\epsilon, X_s(\theta))|&\\ = C|X_s(\theta+(e_i+e_j)\epsilon)-X_s(\theta+e_i\epsilon))-X_s(\theta+e_j\epsilon)+X_s(\theta)|.&\end{aligned}$$

The sixth and last term can be bounded similarly to the fifth.

Combining these bounds into gives the result for one k. As there are a finite number of reactions, take the maximum constants to apply for all k. □

Lemma 4.37. *Given the four processes coupled as in (4.7) with intensities linear in both the parameter and state variables, for a fixed $T>0$, there exists some $C_{T,M}>0$ such that*

$$\mathbb{E}\sup_{t\leq T}|X_t(\theta+(e_i+e_j)\epsilon)-X_t(\theta+e_i\epsilon))-X_t(\theta+e_j\epsilon)+X_t(\theta)| \leq C_{T,M}\epsilon^2.$$

Proof. First, from the coupling (4.7), we have

$$\begin{aligned} &X_t(\theta + (e_i + e_j)\epsilon) - X_t(\theta + e_i\epsilon)) - X_t(\theta + e_j\epsilon) + X_t(\theta) \\ &\qquad = \sum_k \zeta_k \left(R_{k,[1,0,0,0]}(t) - R_{k,[0,1,0,0]}(t) - R_{k,[0,0,1,0]}(t) + R_{k,[0,0,0,1]}(t) \right). \end{aligned}$$

Thus, by the triangle inequality and the non-negativity of the processes $R_{k,[b_1,b_2,b_3,b_4]}(t)$,

$$\begin{aligned} &|X_t(\theta + (e_i + e_j)\epsilon) - X_t(\theta + e_i\epsilon)) - X_t(\theta + e_j\epsilon) + X_t(\theta)| \\ &\qquad \leq \sum_k |\zeta_k| \left(R_{k,[1,0,0,0]}(t) - R_{k,[0,1,0,0]}(t) - R_{k,[0,0,1,0]}(t) + R_{k,[0,0,0,1]}(t) \right). \end{aligned}$$

Since the R are counting processes, we have that $R_{k,[b_1,b_2,b_3,b_4]}(t) \leq R_{k,[b_1,b_2,b_3,b_4]}(T)$ for all $t \leq T$. Thus,

$$\begin{aligned} &\sup_{t \leq T} |X_t(\theta + (e_i + e_j)\epsilon) - X_t(\theta + e_i\epsilon)) - X_t(\theta + e_j\epsilon) + X_t(\theta)| \\ &\qquad \leq \sum_k |\zeta_k| \left(R_{k,[1,0,0,0]}(T) - R_{k,[0,1,0,0]}(T) - R_{k,[0,0,1,0]}(T) + R_{k,[0,0,0,1]}(T) \right). \end{aligned}$$

Taking expectations and using the definitions (4.5) and (4.6) we obtain

$$
\begin{aligned}
&\mathbb{E}\sup_{t\le T}|X_t(\theta+(e_i+e_j)\epsilon)-X_t(\theta+e_i\epsilon))-X_t(\theta+e_j\epsilon)+X_t(\theta)| \\
&\quad\le \sum_k |\zeta_k|\mathbb{E}\left(R_{k,[1,0,0,0]}(T)+R_{k,[0,1,0,0]}(T)+R_{k,[0,0,1,0]}(T)+R_{k,[0,0,0,1]}(T)\right) \\
&\quad\le c\sum_k \mathbb{E}\int_0^T \Lambda_{k,[1,0,0,0]}(s)+\Lambda_{k,[0,1,0,0]}(s)+\Lambda_{k,[0,0,1,0]}(s)+\Lambda_{k,[0,0,0,1]}(s)\,ds \\
&\quad= c\sum_k \mathbb{E}\int_0^T (\lambda_{k,1}-\lambda_{k,1}\wedge\lambda_{k,2})+(\lambda_{k,3}-\lambda_{k,3}\wedge\lambda_{k,4}) \\
&\qquad\qquad -2\left((\lambda_{k,1}-\lambda_{k,1}\wedge\lambda_{k,2})\wedge(\lambda_{k,3}-\lambda_{k,3}\wedge\lambda_{k,4})\right) \\
&\qquad\qquad +(\lambda_{k,2}-\lambda_{k,1}\wedge\lambda_{k,2})+(\lambda_{k,4}-\lambda_{k,3}\wedge\lambda_{k,4}) \\
&\qquad\qquad -2\left((\lambda_{k,2}-\lambda_{k,1}\wedge\lambda_{k,2})\wedge(\lambda_{k,4}-\lambda_{k,3}\wedge\lambda_{k,4})\right)\,ds \\
&\quad= c\sum_k \mathbb{E}\int_0^T |(\lambda_{k,1}-\lambda_{k,1}\wedge\lambda_{k,2})-(\lambda_{k,3}-\lambda_{k,3}\wedge\lambda_{k,4})| \\
&\qquad\qquad +|(\lambda_{k,2}-\lambda_{k,1}\wedge\lambda_{k,2})-(\lambda_{k,4}-\lambda_{k,3}\wedge\lambda_{k,4})|\,ds
\end{aligned}
\tag{4.9}
$$

where the last equality follows via the identity $x+y-2(x\wedge y)=|x-y|$.

Fix k and consider one term of this integrand. Relabel the terms within the integrand as $|(a-a\wedge b)-(c-c\wedge d)|+|(b-a\wedge b)-(d-c\wedge d)|$. Then there are four cases dependent on the values of the two minima:

1. $a\le b, c\le d$: Then $|a-a\wedge b-c+c\wedge d|+|b-a\wedge b-d+c\wedge d|=0+|-a+b+c-d|$.

2. $a\le b, d\le c$: Then we obtain $|-c+d|+|-a+b|=c-d+b-a=|a-b-c+d|$.

3. $b\le a, c\le d$: Then we obtain $|a-b|+|-d+c|=a-b+d-c=|a-b-c+d|$.

4. $b\le a, d\le c$: Then we obtain $|a-b-c+d|+0$.

That is, $|a - a\wedge b - c + c\wedge d| + |b - a\wedge b - d + c\wedge d| = |a - b - c + d|$. Therefore, we may rewrite the integrand of (4.9) as $|\lambda_{k,1} - \lambda_{k,2} - \lambda_{k,3} + \lambda_{k,4}|$. Returning to our calculation in (4.9), we see that

$$\begin{aligned}
&\mathbb{E}\sup_{t\le T}|X_t(\theta+(e_i+e_j)\epsilon) - X_t(\theta+e_i\epsilon)) - X_t(\theta+e_j\epsilon) + X_t(\theta)| \\
&\quad\le c\sum_k \int_0^T \mathbb{E}|\lambda_{k,1} - \lambda_{k,2} - \lambda_{k,3} + \lambda_{k,4}|ds \\
&\quad= c\sum_k \int_0^T \mathbb{E}|\lambda_k(\theta+(e_i+e_j)\epsilon, X_s(\theta+(e_i+e_j)\epsilon)) \\
&\qquad\qquad - \lambda_k(\theta+e_i\epsilon, X_s(\theta+e_i\epsilon)) - \lambda_k(\theta+e_j\epsilon, X_s(\theta+e_j\epsilon)) + \lambda_k(\theta, X_s(\theta))|ds \\
&\quad\le c\sum_k \int_0^T C_1\mathbb{E}|X_s(\theta+(e_i+e_j)\epsilon) - X_s(\theta+e_i\epsilon)) - X_s(\theta+e_j\epsilon) + X_s(\theta)| \\
&\qquad\qquad + C_2\epsilon\mathbb{E}|X_s(\theta+(e_i+e_j)\epsilon) - X_s(\theta)| + C_3\epsilon\mathbb{E}|X_s(\theta+e_i\epsilon) - X_s(\theta+e_j\epsilon)|ds \\
&\quad\le c\sum_k \int_0^T C_4\epsilon^2 + C_1\mathbb{E}|X_s(\theta+(e_i+e_j)\epsilon) - X_s(\theta+e_i\epsilon) - X_s(\theta+e_j\epsilon) + X_s(\theta)|ds,
\end{aligned}$$

where the last two lines follow from Lemmas 4.36 and 4.35. Finally, we see that

$$\begin{aligned}
&\mathbb{E}\sup_{t\le T}|X_t(\theta+(e_i+e_j)\epsilon) - X_t(\theta+e_i\epsilon)) - X_t(\theta+e_j\epsilon) + X_t(\theta)| \\
&\quad\le cC_1K\epsilon^2T + cC_2K\int_0^T \mathbb{E}|X_s(\theta+(e_i+e_j)\epsilon) - X_s(\theta+e_i\epsilon) - X_s(\theta+e_j\epsilon) + X_s(\theta)|ds \\
&\quad\le cC_1K\epsilon^2T \\
&\qquad\qquad + cC_2K\int_0^T \mathbb{E}\sup_{r\le s}|X_r(\theta+(e_i+e_j)\epsilon) - X_r(\theta+e_i\epsilon) - X_r(\theta+e_j\epsilon) + X_r(\theta)|ds.
\end{aligned}$$

The proof is finished by applying the Gronwall inequality. □

Lemma 4.38. *Given the four processes coupled as in (4.7) with intensities linear in*

both the parameter and state variables, for a fixed $T > 0$, there exists some $C_{T,M} > 0$ such that

$$\mathbb{E} \sup_{t \leq T} |X_t(\theta + (e_i + e_j)\epsilon) - X_t(\theta + e_i\epsilon)) - X_t(\theta + e_j\epsilon) + X_t(\theta)|^2 \leq C_{T,M}\epsilon^2.$$

Proof. We first consider the same quantity as in the previous lemma, but here we rewrite it in terms of a martingale:

$$\begin{aligned}
&X_t(\theta + (e_i + e_j)\epsilon) - X_t(\theta + e_i\epsilon)) - X_t(\theta + e_j\epsilon) + X_t(\theta) \\
&\quad = \sum_k \zeta_k \left(R_{k,[1,0,0,0]}(t) - R_{k,[0,1,0,0]}(t) - R_{k,[0,0,1,0]}(t) + R_{k,[0,0,0,1]}(t) \right) \\
&\quad = M^{\theta,\epsilon}(t) + \sum_k \zeta_k \int_0^t \Lambda_{k,[1,0,0,0]}(s) - \Lambda_{k,[0,1,0,0]}(s) - \Lambda_{k,[0,0,1,0]}(s) + \Lambda_{k,[0,0,0,1]}(s) ds \\
&\quad = M^{\theta,\epsilon}(t) + \sum_k \zeta_k \int_0^t \left(\lambda_k(\theta + (e_i + e_j)\epsilon, X_s(\theta + (e_i + e_j)\epsilon)) \right. \qquad (4.10) \\
&\qquad \left. -\lambda_k(\theta + e_i\epsilon, X_s(\theta + e_i\epsilon)) - \lambda_k(\theta + e_j\epsilon,_s X(\theta + e_j\epsilon)) + \lambda_k(\theta, X_s(\theta)) \right) ds
\end{aligned}$$

where $M^{\theta,\epsilon}(t)$ is the martingale created by centering this process,

$$\begin{aligned}
M^{\theta,\epsilon}(t) := \sum_k \zeta_k \left(R_{k,[1,0,0,0]}(t) - R_{k,[0,1,0,0]}(t) - R_{k,[0,0,1,0]}(t) + R_{k,[0,0,0,1]}(t) \right. \\
\left. + \Lambda_{k,[1,0,0,0]}(t) - \Lambda_{k,[0,1,0,0]}(t) - \Lambda_{k,[0,0,1,0]}(t) + \Lambda_{k,[0,0,0,1]}(t) \right).
\end{aligned}$$

Note that $M^{\theta,\epsilon}$ has (conditional) quadratic covariation matrix

$$[M^{\theta,\epsilon}]_t = \sum_k \zeta_k (\zeta_k)^T \left(R_{k,[1,0,0,0]}(t) + R_{k,[0,1,0,0]}(t) + R_{k,[0,0,1,0]}(t) + R_{k,[0,0,0,1]}(t) \right)$$

so that, as in Lemma 4.37 with the work of (4.9) and following, we have

$$\begin{aligned}\mathbb{E}[M^{\theta,\epsilon}]_t &= \sum_k \zeta_k(\zeta_k)^T \mathbb{E}\left(R_{k,[1,0,0,0]}(t) + R_{k,[0,1,0,0]}(t) + R_{k,[0,0,1,0]}(t) + R_{k,[0,0,0,1]}(t)\right) \\ &\leq \sum_k \zeta_k(\zeta_k)^T \mathbb{E}\int_0^t C_1\epsilon^2 + C_2\mathbb{E}|X_s(\theta + (e_i+e_j)\epsilon) - X_s(\theta + e_i\epsilon) \\ &\qquad\qquad - X_s(\theta + e_j\epsilon) + X_s(\theta)|ds.\end{aligned}$$

Now, by the Burkholder-Davis-Gundy inequality, we have that for some constant $C > 0$

$$\begin{aligned}\mathbb{E}\sup_{t\leq T}|M^{\theta,\epsilon}(t)|^2 &\leq C\mathbb{E}\sum_i [M_i^{\theta,\epsilon}]_T \\ &\leq C\sum_i\sum_k (\zeta_i)^2 \int_0^T C_4\epsilon^2 + C_1\mathbb{E}|X_s(\theta + (e_i+e_j)\epsilon) - X_s(\theta + e_i\epsilon)) \\ &\qquad\qquad - X_s(\theta + e_j\epsilon) + X_s(\theta)|ds \\ &\leq C_5T\epsilon^2 + C_6\int_0^T \mathbb{E}|X_s(\theta + (e_i+e_j)\epsilon) - X_s(\theta + e_i\epsilon)) - X_s(\theta + e_j\epsilon) + X_s(\theta)|ds \\ &\leq C_7T\epsilon^2\end{aligned} \tag{4.11}$$

where the last line follows from Lemma 4.37.

Returning to (4.10), take absolute values, use the triangle inequality, square, and

several times use the bound $(x+y)^2 \leq 2x^2 + 2y^2$. Using Lemma 4.36 as well, we obtain

$$
\begin{aligned}
&|X_t(\theta + (e_i+e_j)\epsilon) - X_t(\theta + e_i\epsilon)) - X_t(\theta + e_j\epsilon) + X_t(\theta)|^2 \\
&\quad \leq 2|M^{\theta,\epsilon}(t)|^2 + 2\Big| \sum_k \zeta_k \int_0^t \big(\lambda_k(\theta + (e_i+e_j)\epsilon, X_s(\theta + (e_i+e_j)\epsilon)) \\
&\qquad\qquad - \lambda_k(\theta + e_i\epsilon, X_s(\theta + e_i\epsilon)) - \lambda_k(\theta + e_j\epsilon,_s X(\theta + e_j\epsilon)) + \lambda_k(\theta, X_s(\theta))\big) ds\Big|^2 \\
&\quad \leq 2|M^{\theta,\epsilon}(t)|^2 + 2tc \sum_k \int_0^t \big|\lambda_k(\theta + (e_i+e_j)\epsilon, X(\theta + (e_i+e_j)\epsilon, s)) \\
&\qquad\qquad - \lambda_k(\theta + e_i\epsilon, X(\theta + e_i\epsilon, s)) - \lambda_k(\theta + e_j\epsilon, X(\theta + e_j\epsilon, s)) + \lambda_k(\theta, X(\theta, s))\big|^2 ds \\
&\quad \leq 2|M^{\theta,\epsilon}(t)|^2 + 2tcK \int_0^t \Big(C_1 |X_s(\theta + (e_i+e_j)\epsilon) - X_s(\theta + e_i\epsilon)) - X_s(\theta + e_j\epsilon) + X_s(\theta)| \\
&\qquad\qquad + C_2\epsilon |X_s(\theta + (e_i+e_j)\epsilon) - X_s(\theta)| + C_3\epsilon |X_s(\theta + e_i\epsilon) - X_s(\theta + e_j\epsilon)| \Big)^2 ds \\
&\quad \leq 2|M^{\theta,\epsilon}(t)|^2 + 8tcK \int_0^t C_1^2 |X_s(\theta + (e_i+e_j)\epsilon) - X_s(\theta + e_i\epsilon)) - X_s(\theta + e_j\epsilon) + X_s(\theta)|^2 \\
&\qquad\qquad + C_2^2\epsilon^2 |X_s(\theta + (e_i+e_j)\epsilon) - X_s(\theta)|^2 + C_3^2\epsilon^2 |X_s(\theta + e_i\epsilon) - X_s(\theta + e_j\epsilon)|^2 ds \\
&\quad \leq 2 \sup_{t \leq T} |M^{\theta,\epsilon}(t)|^2 \\
&\qquad + 8TcK \int_0^T \sup_{r \leq s} \big(C_1^2 |X_r(\theta + (e_i+e_j)\epsilon) - X_r(\theta + e_i\epsilon)) - X_r(\theta + e_j\epsilon) + X_r(\theta)|^2 \\
&\qquad\qquad + C_2^2\epsilon^2 |X_r(\theta + (e_i+e_j)\epsilon) - X_r(\theta)|^2 + C_3^2\epsilon^2 |X_r(\theta + e_i\epsilon) - X_r(\theta + e_j\epsilon)|^2 \big) ds
\end{aligned}
$$

Since this last line now holds for all $t \leq T$, we may also take a supremum on the left hand side. Taking expectations and using (4.11) and Lemma 4.35 with $f(x) = x$, we

now obtain

$$\begin{aligned}
&\mathbb{E}\sup_{t\le T}|X_t(\theta+(e_i+e_j)\epsilon)-X_t(\theta+e_i\epsilon))-X_t(\theta+e_j\epsilon)+X_t(\theta)|^2\\
&\quad\le 2\mathbb{E}\sup_{t\le T}|M^{\theta,\epsilon}(t)|^2\\
&\qquad+8TcK\int_0^T\big(\mathbb{E}\sup_{r\le s}C_1^2|X_r(\theta+(e_i+e_j)\epsilon)-X_r(\theta+e_i\epsilon))-X_r(\theta+e_j\epsilon)+X_r(\theta)|^2\\
&\qquad\qquad+\mathbb{E}\sup_{r\le s}C_2^2\epsilon^2|X_r(\theta+(e_i+e_j)\epsilon)-X_r(\theta)|^2\\
&\qquad\qquad+\mathbb{E}\sup_{r\le s}C_3^2\epsilon^2|X_r(\theta+e_i\epsilon)-X_r(\theta+e_j\epsilon)|^2\big)ds\\
&\quad\le C_7T\epsilon^2+C_8T\epsilon^3\\
&\quad+8TcK\int_0^T\big(\mathbb{E}\sup_{r\le s}C_1^2|X_r(\theta+(e_i+e_j)\epsilon)-X_r(\theta+e_i\epsilon))-X_r(\theta+e_j\epsilon)+X_r(\theta)|^2\big)ds
\end{aligned}\tag{4.12}$$

and so, by Gronwall's inequality, we have

$$\mathbb{E}\sup_{t\le T}|X_t(\theta+(e_i+e_j)\epsilon)-X_t(\theta+e_i\epsilon)-X_t(\theta+e_j\epsilon)+X_t(\theta)|^2\le C_{T,K}\epsilon^2.\tag{4.13}$$

□

Proof of Theorem 4.34. Use the linearity of f to write

$$\begin{aligned}
&|f(X_t(\theta+(e_i+e_j)\epsilon))-f(X_t(\theta+e_i\epsilon))-f(X_t(\theta+e_j\epsilon))+f(X_t(\theta))|^2\\
&\qquad\le C^2\big|X_t(\theta+(e_i+e_j)\epsilon)-X_t(\theta+e_i\epsilon)-X_t(\theta+e_j\epsilon)+X_t(\theta)\big|^2
\end{aligned}$$

and use the previous result. □

4.4 Numerical Examples

In this section, we compare the double coupled method with the following existing second order methods discussed in Section (4.1).

(a) the usual Independent Random Numbers (IRN) estimator in which the processes of (4.1) are simulated independently, also referred to as the crude Monte Carlo method,

(b) the common random numbers approach (CRN) in which the processes of (4.1) are simulated given the same stream of random numbers using SSA,[3]

(c) the Common Reaction Path (CRP) method in which the processes of (4.1) are coupled by reusing each Y_k in the representation (1.7) of the four processes,

(d) the double coupled (CFD2) method proposed here, which implements the coupling (4.7),

(e) the likelihood ratio method (LR) in which the computed weight function is used as a control variate.

All methods except (b) were simulated using NRM, modified as necessary. We also note that the first four methods use the second finite difference, which has some bias; recall that to reduce this bias we actually simulate the centered difference (4.3), which is accomplished in the same way as the forward difference but with the parameters shifted. The likelihood ratio method is the only one of the four methods we use here that is unbiased; its high variance, however, typically makes the method unusable. Finally, when discussing performance, we will refer to n of (4.2) as the number of estimates.

[3]At each step, the first random number determines the time of the next reaction, and the second determines which occurs; the reactions were listed in a fixed order as given in this paper.

4.4.1 A Simple Birth Process

Consider a pure birth process $A \to 2A$. Here, $\zeta = 1$, and denoting by X_t the number of A molecules at time t, we assume a propensity function $\lambda(\theta, X_t(\theta))$, so that in the random time change representation,

$$X_t(\theta) = X_0 + Y\left(\int_0^t \lambda(\theta, X_s(\theta))ds\right),$$

where, as usual, Y is a unit-rate Poisson process.

Suppose we are interested in the second derivative of $\mathbb{E}X_t$ with respect to θ, so that $f(\theta, x) = x$. We double couple the processes as in (4.7), noting that we are in the special case when $i = j$. This does not change the main idea of the double coupling, but it requires us to distinguish the two original processes with the same parameter value $\theta+\epsilon$; we label them as $X_t^1(\theta + \epsilon)$ and $X_t^2(\theta + \epsilon)$. Ordering as in (4.4), and noting that since there is only one reaction we may drop the subscript k, we find that

$$\begin{aligned}
\lambda_1 &= \lambda(\theta + 2\epsilon, X_t(\theta + 2\epsilon)) \\
\lambda_2 &= \lambda(\theta + \epsilon, X_t^1(\theta + \epsilon)) \\
\lambda_3 &= \lambda(\theta + \epsilon, X_t^2(\theta + \epsilon)) \\
\lambda_4 &= \lambda(\theta, X_t(\theta)),
\end{aligned}$$

and use these to define the Λ's as given in (4.6). The double coupled processes are then

given as

$$\begin{aligned}
X_t(\theta+2\epsilon) &= X_0(\theta) + R_{[1,1,1,1]}(t) + R_{[1,1,0,0]}(t) + R_{[1,0,1,0]}(t) + R_{[1,0,0,0]}(t) \\
X_t^1(\theta+\epsilon) &= X_0(\theta) + R_{[1,1,1,1]}(t) + R_{[1,1,0,0]}(t) + R_{[0,1,0,1]}(t) + R_{[0,1,0,0]}(t) \\
X_t^2(\theta+\epsilon) &= X_0(\theta) + R_{[1,1,1,1]}(t) + R_{[0,0,1,1]}(t) + R_{[1,0,1,0]}(t) + R_{[0,0,1,0]}(t) \\
X_t(\theta) &= X_0(\theta) + R_{[1,1,1,1]}(t) + R_{[0,0,1,1]}(t) + R_{[0,1,0,1]}(t) + R_{[0,0,0,1]}(t).
\end{aligned}$$

Now that we have coupled the processes, note that when we consider the second difference (4.1) for the given f, which is linear, most of the sub-processes cancel. For example, since $R_{[1,1,0,0]}$ is present in both $X_t(\theta+2\epsilon)$, which is positive in the difference, and in $X_t^1(\theta+\epsilon)$, which is negative, $R_{[1,1,0,0]}$ is not present in the second difference. One can easily check that the numerator of the difference (4.1) simplifies in this case to

$$\begin{aligned}
&X_t(\theta+2\epsilon) - X_t^1(\theta+\epsilon) - X_t^2(\theta+\epsilon) + X_t(\theta) \\
&= R_{[1,0,0,0]}(t) - R_{[0,1,0,0]}(t) - R_{[0,0,1,0]}(t) + R_{[0,0,0,1]}(t).
\end{aligned} \tag{4.14}$$

Note that the rates of the four remaining counting processes of (4.14) are usually relatively small; in fact, at any given time at least two of the four must have zero propensity, as can be seen by considering the possible values of the minima involved.

Suppose that $\lambda(\theta, X_t(\theta)) = \theta X_t(\theta)$ is simply a constant times the population at time t. We choose to estimate $\frac{\partial^2 \mathbb{E} X_t}{\partial \theta^2}$ at $t = 5$ and $\theta = 1/2$, with $X_0(\theta) = 1$. We use $\epsilon = 1/50$ for the finite difference methods. For simple examples such as this, one can solve for the derivative explicitly; in this case the actual value is 304.6.

As can be seen in the data in Table 4.1, the variance of the double coupled estimator,

Method	Estimates	Approximation	# updates	CPU time (s)
IRN	100,000	307 ± 447	$\approx 3.7 \times 10^6$	38
CRP	100,000	315 ± 24	$\approx 3.7 \times 10^6$	49
CRN	100,000	282 ± 24	$\approx 3.3 \times 10^6$	32
LR	100,000	311 ± 20	$\approx 1.1 \times 10^6$	37
CFD2	100,000	296 ± 12	$\approx 1.2 \times 10^6$	22

Table 4.1: 95% confidence intervals and computation time for each of the five methods (a) through (e), after 100,000 estimates, on the simple birth model of 4.4.1 (with linear propensity). An ϵ of 1/50 was used for the three finite difference methods. Actual value: 304.6.

manifested in the width of the confidence interval, is smaller than that of the estimators given by the other methods. For instance, for the same number of estimates it gives a confidence interval of half the width of the CRP and CRN methods, which for this single-reaction model, though implemented differently, give equivalent estimators. Here and throughout, confidence intervals are constructed as $\pm 1.96\sqrt{v}$ where v is the variance of the estimator (4.2).

For each method, we also include the CPU time that was required for the simulation, as well as the number of updates made to the system state (the number of times a reaction vector is added to the state vector). The latter is a useful comparison tool, as it provides a measure of the amount of work the method requires, but is not influenced by differences in implementation (such as use of Gillespie vs next reaction algorithms). These differences, on the other hand, often affect CPU time. We do not also provide a random number count for each method, but note here that except for CRN this number is equal to the number of system updates. For CRN, which uses Gillespie's algorithm, two random numbers are used per system update. Finally, the CPU time will certainly

vary by machine; all tests described in this section were run in MATLAB on a Windows machine with a 1.6GHz processor.

4.4.2 mRNA Transcription and Translation

We now examine the performance of the proposed method on a more realistic model. In the following model of gene transcription and translation, first seen in the Introduction of this thesis, mRNA is being transcribed and then translated into protein, while both the mRNA and the protein may undergo degradation. The given constants are in the sense of mass action kinetics, so for example protein is being created at a rate of γ times the number of mRNA molecules:

$$\emptyset \underset{\theta}{\overset{2}{\rightleftarrows}} M \xrightarrow{\gamma} M + P, \qquad P \xrightarrow{1} \emptyset.$$

We assume initial concentrations of zero mRNA and protein molecules. The stochastic equation for this model is

$$\begin{aligned} X(\theta,t) = {} & Y_1(2t)\left(\begin{smallmatrix}1\\0\end{smallmatrix}\right) + Y_2\left(\int_0^t \theta X_M(\theta,s)ds\right)\left(\begin{smallmatrix}-1\\0\end{smallmatrix}\right) \\ & + Y_3\left(\int_0^t \gamma X_M(\theta,s)ds\right)\left(\begin{smallmatrix}0\\1\end{smallmatrix}\right) + Y_4\left(\int_0^t X_P(\theta,s)ds\right)\left(\begin{smallmatrix}0\\-1\end{smallmatrix}\right) \end{aligned}$$

where $X = \left(\begin{smallmatrix}X_M\\X_P\end{smallmatrix}\right)$ gives the numbers of the mRNA and protein molecules respectively. Note that we have moved the parameter t from the subscript for notational convenience.

In subsection 4.4.2, we compute the second derivative of the expected number of protein molecules with respect to θ, while in subsection 4.4.2, we compute the mixed partial of this same quantity with respect to both θ and γ. In subsection 4.4.2, we

compute the second derivative of the *square* of the expected number of protein molecules with respect to θ.

2^{nd} derivative of protein abundance with respect to θ

Suppose we would like to estimate the second derivative of the expected number of protein molecules with respect to θ at a time of $t = 30$ and $\theta = \frac{1}{4}$. Additionally, we fix $\gamma = 10$ and $X_0 = 0$. One can analytically find that $\frac{\partial^2}{\partial\theta^2}\mathbb{E}X_P(30) = 2496$.

First, Table 4.2 gives simulation data as in the previous examples, with two different perturbations, ϵ, of θ used. Note the trade-off between bias and precision: a larger epsilon implies the second finite difference has a larger bias, but, since there is an ϵ^2 in the denominator of the estimator, the variance of the estimator is smaller; for small epsilon it is vice-versa. Table 4.3 shows the relevant data for the likelihood ratio method.

Perhaps more illustrative is Table 4.4, which compares the numbers of estimates and system updates as well as the time required to achieve a 95% confidence interval of a set width. These data give a good idea of the efficiency of the methods, as often one desires the estimate within a given tolerance. We can see that the double coupled method is approximately 25 times faster than CRP, 73 times faster than the often used CRN method, over 100 times faster than the LR method, and over 125 times faster than IRN. Note also that the double coupled method requires drastically fewer estimates to achieve the same confidence, so that, even though the computation of one double coupled estimate requires more time than most of the other methods, as can be seen in Table 4.2, the lower variance leads to very large time savings.

Finally, in Figure 4.1 we include a plot of the variance of the different estimators versus time. Note that the scales on the plots are very different. The plots corresponding

Method	Estimates	$\epsilon = 1/20$	$\epsilon = 1/100$	# updates	CPU time (s)
CRN	1,000	2682 ± 1192	5950 ± 19123	$\approx 1.26 \times 10^7$	46
CRP	1,000	2758 ± 569	-2630 ± 9268	$\approx 1.27 \times 10^7$	70
CFD2	1,000	2655 ± 129	2640± 1001	$\approx 4.68 \times 10^6$	48
CRN	10,000	2453 ± 369	1505 ± 6120	$\approx 1.27 \times 10^8$	457
CRP	10,000	2783 ± 179	2627 ± 2937	$\approx 1.27 \times 10^8$	672
CFD2	10,000	2601 ± 40	2352 ± 282	$\approx 4.68 \times 10^7$	483
CRN	40,000	2386 ± 188	1069 ± 2984	$\approx 5.07 \times 10^8$	1829
CRP	40,000	2745 ± 89	3593 ± 1468	$\approx 5.07 \times 10^8$	2739
CFD2	40,000	2582 ± 20	2512 ± 147	$\approx 1.87 \times 10^8$	1931

Table 4.2: 95% confidence intervals for each of the finite difference methods (b), (c), and (d) for the computation in the mRNA and protein model of subsection 4.4.2. Note that the bias of the second finite difference can be seen when $\epsilon = 1/20$ (the actual value is 2496). Also note that, though for a fixed number of estimates the CFD2 method is not the fastest method, it achieves a much smaller confidence interval. The number of updates and computational time for a fixed number of estimates are essentially independent of ϵ and so the reported values, here and throughout, are the average of the values for the two choices of ϵ.

Estimates	Approximation	# updates	CPU time (s)
1,000	2150 ± 2258	$\approx 4.20 \times 10^6$	14
10,000	2429 ± 729	$\approx 4.19 \times 10^7$	135
40,000	2176 ± 404	$\approx 1.68 \times 10^8$	540

Table 4.3: 95% confidence intervals for the LR method (d) for the computation in the mRNA transcription model computation of subsection 4.4.2. Even though this method is fastest per estimate, note that the variance (and so the width of the confidence interval) is large.

Method	Estimates	Approximation	# updates	CPU time (s)
LR	495,000	2506 ± 120	$\approx 2.1 \times 10^9$	6619
IRN	190,000	2617 ± 120	$\approx 2.4 \times 10^9$	7657
CRN	98,100	2572 ± 120	$\approx 2.6 \times 10^8$	4489
CRP	22,200	2532 ± 120	$\approx 2.8 \times 10^8$	1533
CFD2	1150	2565 ± 120	$\approx 5.8 \times 10^6$	61

Table 4.4: Required estimates, updates, and computational time needed for 95% confidence intervals of $\pm$ 120 for all five methods on the computation of the mRNA transcription model computation of subsection 4.4.2. An ϵ of 1/20 was used for the finite difference methods.

to finite difference methods all appear to converge; the limiting value for the double coupled method, however, is over 20 times smaller than the CRP method, and over 170 times smaller than the CRN and IRN methods. Note also that, as time increases, the CRN variance tends to the same value as the IRN method; this is expected, since we expect the processes to decouple. The variance for CRP behaves similarly, converging to a number of approximately the same order of magnitude as the IRN method, though the value itself is significantly lower in this four-reaction model. The plot for the LR method scales quadratically, as is expected by the form of the estimator (see Chapter VII.3 of [7]). This shows that, for moderate and large times, the double coupled method quickly becomes much more efficient then the other estimators.

Mixed partial of protein abundance

We compare the five methods in the estimation of $\frac{\partial^2}{\partial\gamma\partial\theta}\mathbb{E}X_P(30)$ at $\theta = 1/4$ and $\gamma = 10$, which can be calculated exactly to be -31.8. Table 4.5 shows the approximations and computational complexity of these methods using $5,000$ estimates.

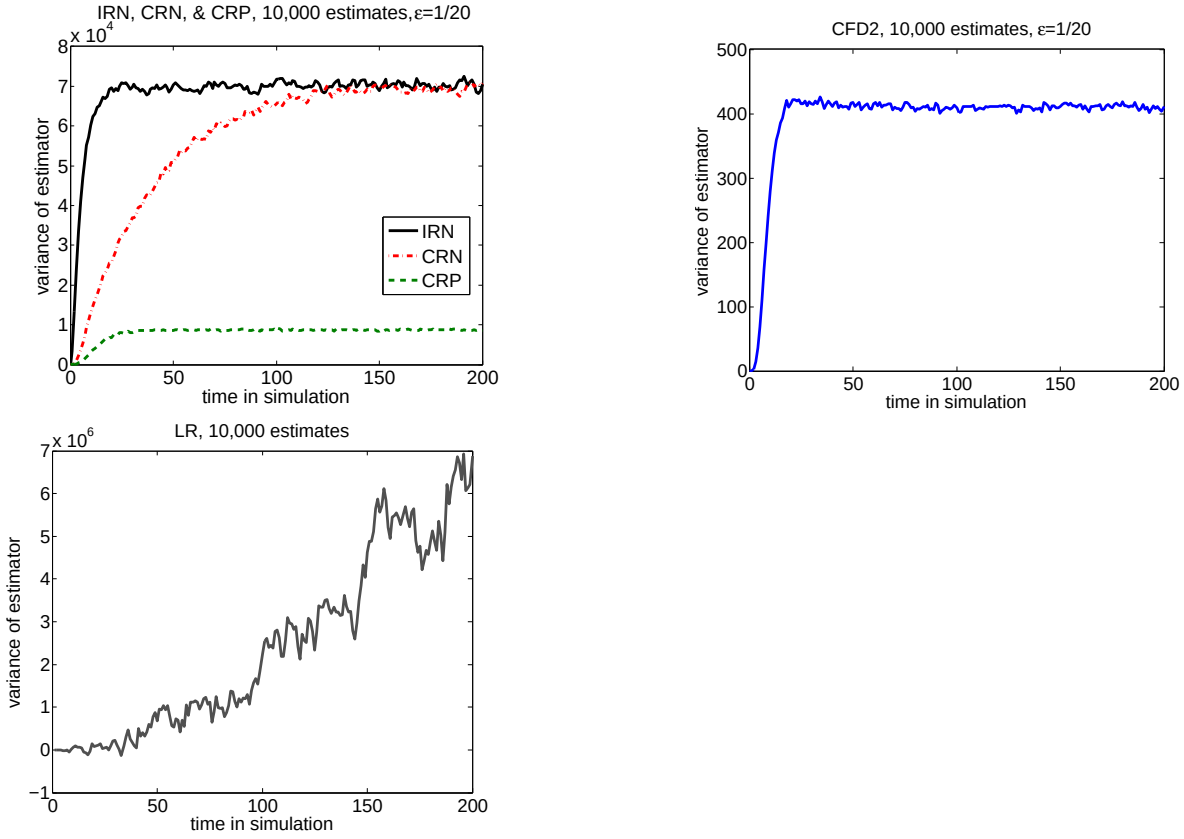

Figure 4.1: Variance versus time of the estimators of the five different methods, applied to the calculation of $\frac{\partial^2}{\partial\theta^2}\mathbb{E}X_P(\theta,t)$ in the mRNA transcription model of subsection 4.4.2. Note that the scales are vastly different.

Method	Estimates	Approximation	# updates	CPU time (s)
IRN	5,000	607 ± 923	$\approx 8.43 \times 10^7$	264
CRN	5,000	191.5 ± 330	$\approx 8.42 \times 10^7$	273
CRP	5,000	21.0 ± 96	$\approx 8.41 \times 10^7$	365
CFD2	5,000	-33.8 ± 4	$\approx 2.25 \times 10^7$	238
LR	5,000	-15.4 ± 113	$\approx 2.10 \times 10^7$	73
LR	17,000	-61.8 ± 68	$\approx 6.72 \times 10^7$	234

Table 4.5: 95% confidence intervals and computational complexity for all five methods, after $5,000$ estimates, for the computation of the mixed partial derivative in the mRNA transcription model as in subsection 4.4.2. An ϵ of $1/25$ was used for the finite difference methods. Additionally, results from the LR method with CPU time approximately that of CFD2 are included for comparison. Actual value: -31.8.

Note that in this example, the LR method outperforms all methods, except CFD2, with respect to computation time. Thus, for comparison, we have also included the results of a test using the LR method in which the CPU time is approximately the same as CFD2; note that the confidence interval for the CFD2 method is much smaller. Figure 4.2 shows variance plots of the CRN and CRP, and CFD2 methods over time in simulation.

2^{nd} derivative of the square of protein abundance with respect to θ

We also calculate, from the mRNA transcription model of Example 4.1, $\frac{\partial^2}{\partial\theta^2}\mathbb{E}(X_P(t)^2)$ at $t = 5$ and $\theta = \frac{1}{4}$, with $\gamma = 10$ and $X_0 = 0$. Note here we are considering a function f of the state space which is non-linear.

In Figure 4.3, we plot the log of the variance of the numerator of the estimator (4.3) versus the log of epsilon. Since we expect, for the double coupled CFD2 method, that

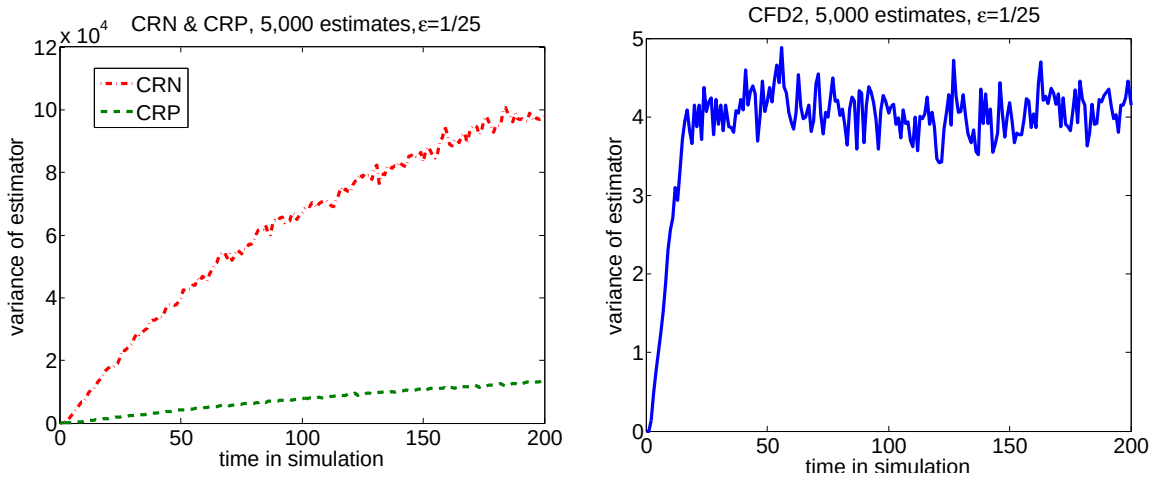

Figure 4.2: Plots of variance over time for the CRN and CRP methods, and the CFD2 method, in computing the mixed partial derivative of the mRNA transcription model of subsection 4.4.2. Note the very different scales. For comparison, the IRN method plateaus at a variance of approximately 5×10^6.

this variance $V(\epsilon)$ should scale like $C\epsilon^p$ for some constants C and p, we see that the slope of $\log(V(\epsilon)) = \log(C) + p\log(\epsilon)$ from our simulations will suggest the value of p. This plot suggests that $p = 2$; since the numerator of the estimator is then divided by ϵ^2 in $\hat{M}(\epsilon)$, this suggests a final variance of $O(n^{-1}\epsilon^{-2})$ for the estimator (4.2) as discussed in Section 4.1.

For comparison, the slope of this log-log plot for the IRN method is zero, as the variance of the numerator does not depend on epsilon, giving a final variance of $O(n^{-1}\epsilon^{-4})$. The slopes for the associated log-log plots for the CRN and CRP estimators will vary with time (discussed further in Section 4.4.4).

The general behavior of the variances over time for the IRN, CRN, and CRP methods can be seen in Figure 4.4.

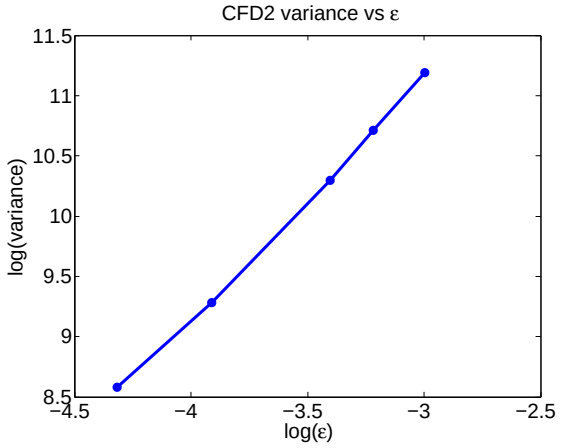

Figure 4.3: This log-log plot of variance versus epsilon (100,000 estimates) for the mRNA transcription model computation of subsection 4.4.2 suggests that the CFD2 method gives an estimator of $O(\epsilon^{-2})$ even though the function f of the system state is non-linear: the slope of the best fit line is 1.98.

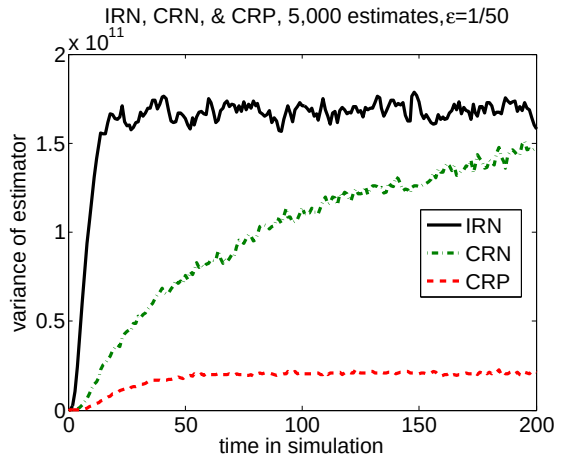

Figure 4.4: Plot of variance over time for 5,000 estimates of the IRN, CRN, and CRP methods for the mRNA transcription model computation of subsection 4.4.2. The variance of the CFD2 method is too small to be seen; at time 200 it is $\approx 3.5 \times 10^8$.

4.4.3 Quadratic Decay

In order to demonstrate that the $O(\epsilon^2)$ convergence rate seen in the previous examples does not universally hold, we consider a pure decay process of a population X_t, so that the sole reaction has $\zeta = -1$ and quadratic propensity $\lambda(\theta, X_t(\theta)) = \theta X_t(\theta)(X_t(\theta) - 1)$, and calculate $\frac{\partial^2 \mathbb{E}X_t(\theta)}{\partial \theta^2}$ with $\theta = 1$ and with initial population $X_0(\theta) = 2000$. Figure 4.5 gives a log-log plot of variance versus epsilon at time 0.001. Since it suggests $p = 1$, this demonstrates that, in this case, the double coupled method provides only $O(n^{-1}\epsilon^{-3})$ convergence as discussed in Section 4.1, showing that rate to be sharp.

As demonstrated in Table 4.6 and in Figure 4.6, however, the double coupled method is still significantly more efficient than existing methods on this model.

4.4.4 Genetic Toggle Switch

Finally, we consider a model of a genetic toggle switch that also appeared in [2] and [32]:

$$\emptyset \underset{1}{\overset{\lambda_1}{\rightleftarrows}} A \ , \quad \emptyset \underset{1}{\overset{\lambda_2}{\rightleftarrows}} B$$

where

$$\lambda_1(t) = \frac{b}{1 + X_B(t)^\beta} \quad \text{and} \quad \lambda_2(t) = \frac{a}{1 + X_A(t)^\alpha},$$

and where $X_A(t)$ and $X_B(t)$ denote the number of gene products from two interacting genes. Note that each gene product inhibits the growth of the other.

We take parameter values of $b = 50, \beta = 2.5, a = 16$ and will differentiate with respect to α. Note that this model does *not* follow mass action kinetics, or have linear propensities. In subsection 4.4.4, we consider a second derivative of $\mathbb{E}X_B$ at a fixed time,

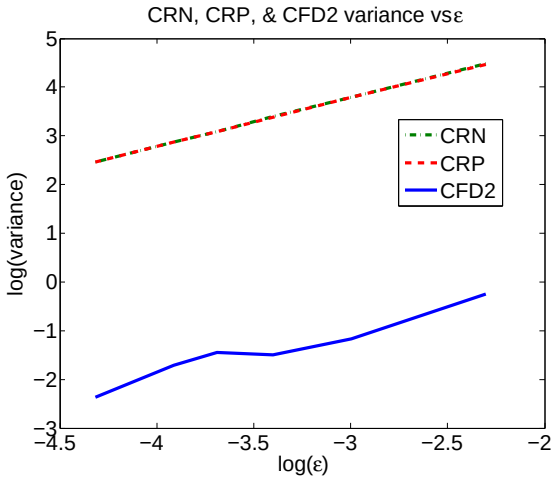

Figure 4.5: This log plot of variance versus epsilon (each point computed to the first of 300,000 estimates or a confidence of ± 10) for the decay model of subsection 4.4.3 suggests that the CFD2 method gives an estimator of only $O(\epsilon^{-3})$: the slope of the best fit line is approximately 0.97. While the CRN and CRP methods also give an estimator of this same rate (the slope of the best fit lines are both $\approx .99$, and in fact the lines are on top of each other), the variance of the estimates from CRN and CRP are significantly higher than those from the CFD2 method, as can be seen by the wide gap between the above curves.

Method	ϵ	Estimates	Approximation	# updates	CPU time (s)
LR	n/a	10,000	1240 ± 1070	$\approx 1.3 \times 10^7$	9
IRN	1/20	10,000	555 ± 218	$\approx 4.8 \times 10^7$	35
CRN	1/20	10,000	585 ± 52	$\approx 4.0 \times 10^7$	30
CRP	1/20	10,000	584 ± 52	$\approx 4.0 \times 10^7$	30
CFD2	1/20	10,000	592 ± 5	$\approx 1.4 \times 10^7$	90
CRN	1/20	272,000	588 ± 10	$\approx 1.1 \times 10^9$	813
CRP	1/20	271,000	589 ± 10	$\approx 1.1 \times 10^9$	862
CFD2	1/20	1,950	592 ± 10	$\approx 2.7 \times 10^6$	17
CRN	1/50	169,500	543 ± 50	$\approx 6.8 \times 10^8$	511
CRP	1/50	169,000	515 ± 50	$\approx 6.8 \times 10^8$	510
CFD2	1/50	1,800	605 ± 50	$\approx 2.5 \times 10^6$	16

Table 4.6: Estimates, ϵ used, and updates and computational time needed for the given 95% confidence intervals for all methods for $\frac{\partial^2 \mathbb{E}X_t(\theta)}{\partial \theta^2}$ at $t = 0.001$ for the quadratic decay model of subsection 4.4.3. The upper half of the table shows the relevant results after the simulation of 10,000 estimates. The lower half of the table shows the results of simulations run until the estimate had a confidence interval of a desired width. The IRN and LR methods were unable to achieve these precisions due to memory constraints. Note again the equivalence of the CRN and CRP methods on a single reaction model.

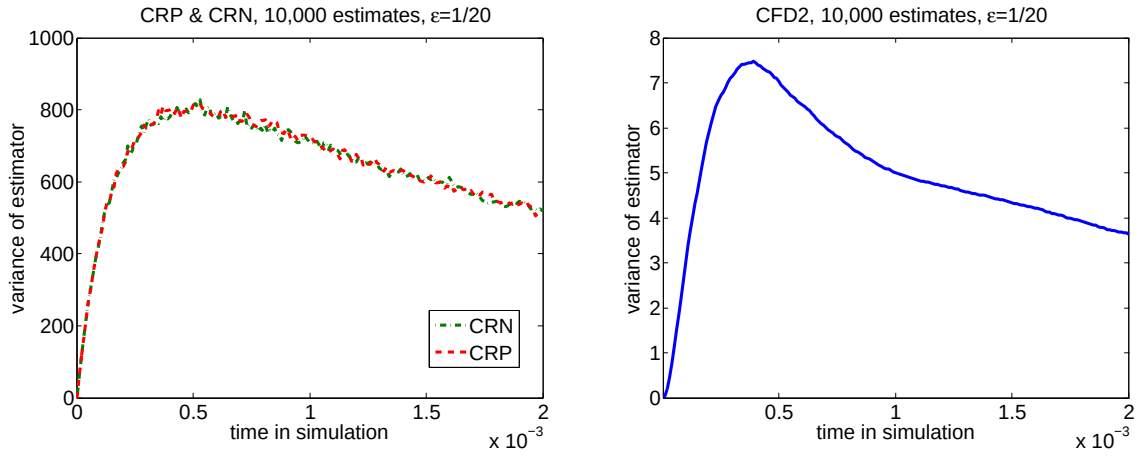

Figure 4.6: The behavior over time of the variance of the estimates of the CRN, CRP, and CFD2 methods on the quadratic decay model of subsection 4.4.3. Note that the variance for the CFD2 method is 100 times smaller than the other two methods, which, as expected, act the same on this model. An ϵ of 1/20 was used and 10,000 estimates were run. The plot of the IRN variance is similar in shape but with a peak variance of 3.1×10^4.

while in subsection 4.4.4, we consider a second derivative of the expected *time average* of X_A up to a given time, which is a functional of the path of X_A rather than simply X_A at some terminal time.

2^{nd} derivative of abundance of B with respect to α

We estimate $\frac{\partial^2 \mathbb{E}X_B(\alpha,t)}{\partial \alpha^2}$ at $\alpha = 1$ and at two times, 5 and 400. In Figure 4.7, we plot the log of the variance of the numerator of the estimator (4.3), using CFD2, versus the log of the perturbation epsilon. As in subsection 4.4.2, the plot clearly suggests that $p = 2$. We also plot the same quantity using CRP and CRN. These slopes, on the other hand, vary with time. For small times both slopes are close to one, but as time increases the slopes decrease, until, for very large times, they are close to zero. This corresponds with the fact that for large times the variances of the CRP and CRN estimates converge to values on the order of the IRN estimate variance, which, as previously noted, is independent of the value of epsilon. The general behavior of the variances over time can be seen in Figure 4.8, where it is seen that CFD2 has a variance that is 16 times lower than CRN and 36 times lower than CRP. Further, we note that for this model the CRP method outperforms the CRN method for small times, while for larger times CRN outperforms CRP.

2^{nd} derivative of time average of abundance of A with respect to α

Finally, we include an example computing a sensitivity of a path functional. That is, the quantity we wish to study is a function of the path of the process $X(s)$ for $s \leq t$, rather than just the terminal value $X(t)$. The only difference in implementation is the need to compute this quantity during the simulation of the path (or to store the path for the

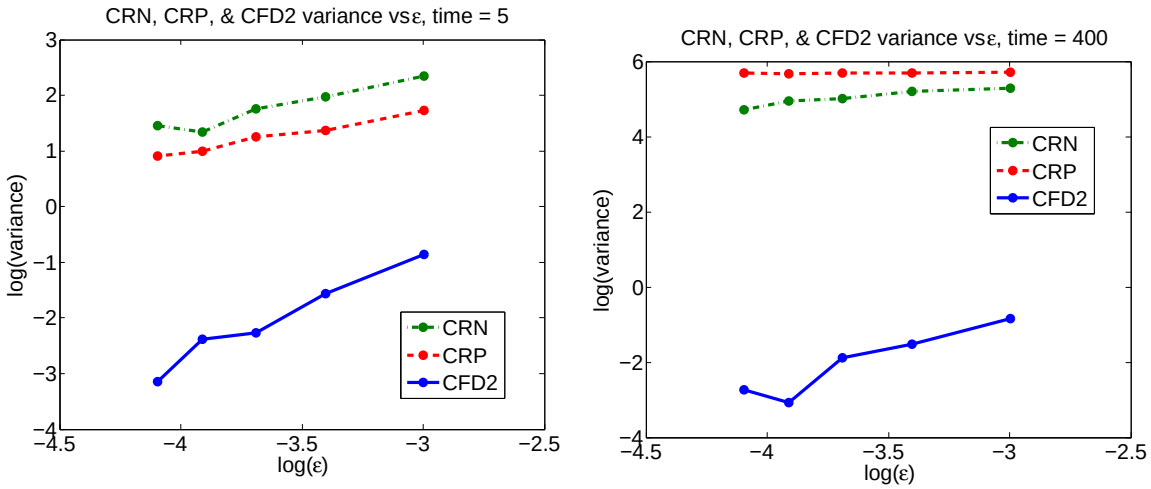

Figure 4.7: These log-log plots (25,000 estimates) of variance versus epsilon for computation on the gene toggle model as in subsection 4.4.4, at two different times, suggest that the double coupled method gives an estimator of $O(\epsilon^{-2})$ even though two of the intensities are nonlinear: the slope of the best fit line for the CFD2 method is approximately 2 (=1.97) at both times. The slope for the CRP and CRN methods, on the other hand, are approximately .74 and .90 respectively at time 5, but are only around .03 and .49 at time 400.

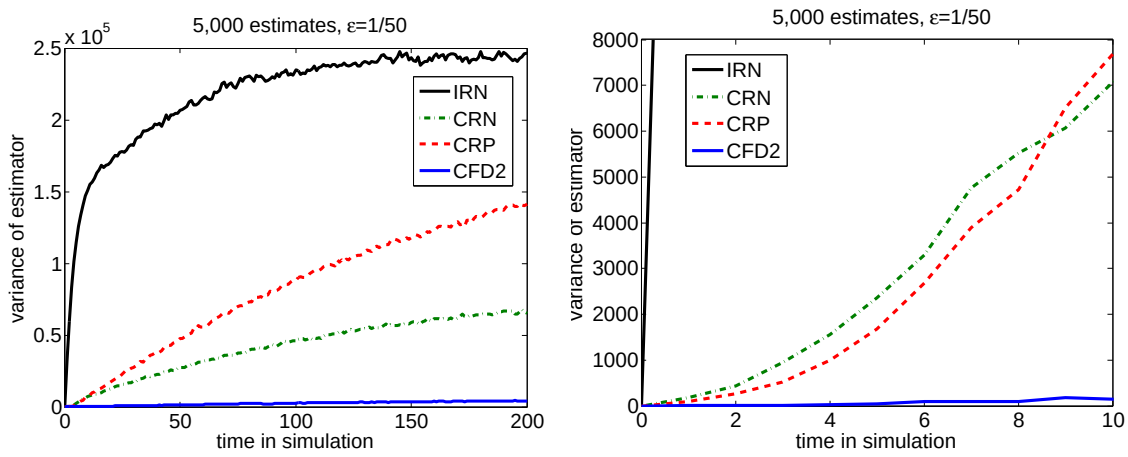

Figure 4.8: Plots of variance over time for 10,000 estimates of the finite difference methods for the gene toggle model as in subsection 4.4.4; the top includes time up to 200, while the bottom provides a close-up view of the plot for times less than 10. The value of the CFD2 variance at time 200 is approximately 4,000, while the CRN variance is approximately 65,000.

computation after its simulation). Table 4.7 shows the estimates of $\frac{\partial^2}{\partial \alpha^2}\mathbb{E}\, t^{-1}\int_0^t X_A(s)ds$ at $t = 30$ using the various finite difference methods, demonstrating the advantage of the double coupled method for these path functional quantities as well. Additionally, Figure 4.9 shows that the overall behavior of the variances of the three finite difference methods remains the same as in the previous examples.

Method	Estimates	Approximation	CPU time (s)
IRN	100,000	-13.8 ± 621	2240
CRN	100,000	-274 ± 146	1441
CRP	100,000	-215 ± 107	3035
CFD2	100,000	-222 ± 26	2722

Table 4.7: 95% confidence intervals and computational complexity for each of the methods (a) through (d), after 100,000 estimates, for the time average computation at $t = 30$ on the gene toggle model of subsection 4.4.4. An ϵ of 1/50 was used.

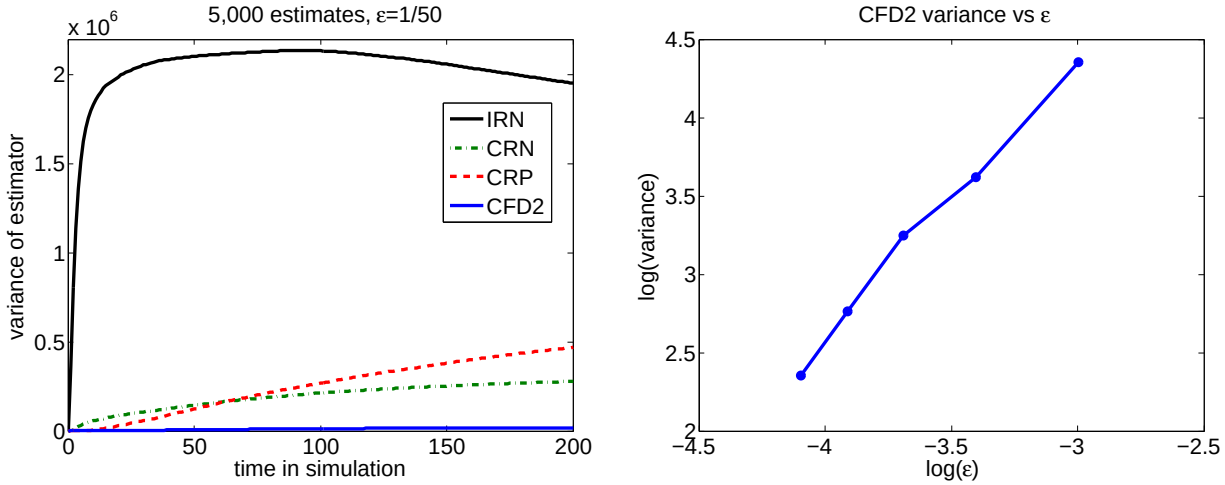

Figure 4.9: At top, a plot of variance over time for 5,000 estimates for the finite difference methods for the path functional computation on the gene toggle model in subsection 4.4.4. The value of the CFD2 variance at time 200 is approximately 19,000, while the CRN variance is approximately 282,000. At bottom, a log-log plot (2,000 estimates) of variance versus epsilon for this computation suggests that the double coupled method gives an estimator converging faster than $O(\epsilon^{-3})$ in this computation as well: the slope of the best fit line for the CFD2 method is approximately 1.78.

Chapter 5

Conclusions and Future Work

This thesis presented a new hybrid method for first order sensitivity estimation for discrete biochemical reaction networks. The hybrid method uses the standard coupling to combine the likelihood ratio and pathwise methods, and provides efficient, unbiased estimation of sensitivities of path functionals. For sensitivities of output quantities at some fixed time, one can use one of two variants of the hybrid method, the Dynkin hybrid method or the RPD hybrid method. As demonstrated on several numerical results, the hybrid method, or one of these variants, is often the most efficient method available. Future work will involve additional numerical experiments in the hope that we can isolate certain characteristics of CTMC models and their intensities that may suggest which sensitivity method is likely to be most efficient for that particular model.

Chapter 4, as in [42], gives the CFD2 method for the computation of second order sensitivities. Through several numerical examples we have demonstrated its advantage over existing methods. Future work will involve studying other methods for the computation of the full sensitivity Hessian. Indeed, in models with a large number R of parameters, the CFD2 method is implausible as it must estimate the Hessian entry by entry, resulting in $O(R^2)$ separate Monte Carlo computations. Another related avenue of future work will explore the possibility of extending the hybrid method of Chapter 3 to the setting of second order sensitivities.

Bibliography

[1] David F. Anderson, *A modified next reaction method for simulating chemical systems with time dependent propensities and delays*, J. Chem. Phys. **127** (2007), no. 21, 214107.

[2] ———, *An efficient finite difference method for parameter sensitivities of continuous time Markov chains*, SIAM: Journal on Numerical Analysis **50** (2012), 2237–2258.

[3] David F. Anderson and Thomas G. Kurtz, *Stochastic analysis of biochemical systems*, to appear.

[4] ———, *Continuous time Markov chain models for chemical reaction networks*, Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology (H. Koeppl et al., ed.), Springer, 2011, pp. 3–42.

[5] H Arsham, A Feuerverger, DL McLeish, J Kreimer, and RY Rubinstein, *Sensitivity analysis and the "what if" problem in simulation analysis*, Mathematical and Computer Modelling **12** (1989), no. 2, 193–219.

[6] Hossein Arsham, *Algorithms for sensitivity information in discrete-event systems simulation*, Simulation Practice and Theory **6** (1998), no. 1, 1–22.

[7] Soren Asmussen and Peter W. Glynn, *Stochastic simulation: Algorithms and analysis*, Springer, 2007.

[8] Jean Alexandre Dieudonné, Jean Dieudonné, France Mathematician, and Jean Dieudonné, *Foundations of modern analysis*, vol. 286, Academic press New York, 1960.

[9] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain, *Stochastic gene expression in a single cell*, Science **297** (2002), no. 5584, 1183–1186.

[10] Stewart N. Ethier and Thomas G. Kurtz, *Markov processes: Characterization and convergence*, 2 ed., John Wiley & Sons, New York, 2005.

[11] Michael C Fu, *What you should know about simulation and derivatives*, Naval Research Logistics (NRL) **55** (2008), no. 8, 723–736.

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

[13] Mike Giles and Paul Glasserman, *Smoking adjoints: Fast Monte Carlo greeks*, Risk **19** (2006), no. 1, 88–92.

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

[15] Paul Glasserman, *Gradient estimation via perturbation analysis*, Kluwer Academic Publishers, 1991.

[16] Peter W Glynn, *Likelihood ratio gradient estimation for stochastic systems*, Communications of the ACM **33** (1990), no. 10, 75–84.

[17] Peter W Glynn and Donald L Iglehart, *Importance sampling for stochastic simulations*, Management Science **35** (1989), no. 11, 1367–1392.

[18] Wei-Bo Gong and Yu-Chi Ho, *Smoothed (conditional) perturbation analysis of discrete event dynamical systems*, Automatic Control, IEEE Transactions on **32** (1987), no. 10, 858–866.

[19] Ankit Gupta and Mustafa Khammash, *Unbiased estimation of parameter sensitivities for stochastic chemical reaction networks*, SIAM Journal on Scientific Computing **35** (2013), no. 6, A2598–A2620.

[20] Yu-Chi Ho, *Performance evaluation and perturbation analysis of discrete event dynamic systems*, Automatic Control, IEEE Transactions on **32** (1987), no. 7, 563–572.

[21] David G Kendall, *On the generalized" birth-and-death" process*, The annals of mathematical statistics (1948), 1–15.

[22] Michal Komorowski, Maria J. Costa, David A. Rand, and Michael P. H. Stumpf, *Sensitivity, robustness, and identifiability in stochastic chemical kinetics models*, PNAS **108** (2011), no. 21, 8645–8650.

[23] Masanori Koyama, *Analysis of stochastically modeled biochemical processes with applications to numerical methods*, Ph.D. thesis, University of Wisconsin – Madison, 2013.

[24] Thomas G. Kurtz, *Lectures on stochastic analysis*, `http://www.math.wisc.edu/~kurtz/735/main735.pdf`.

[25] ———, *Representations of Markov processes as multiparameter time changes*, Ann. Prob. **8** (1980), no. 4, 682–715.

[26] Brian Munsky and Mustafa Khammash, *The finite state projection algorithm for the solution of the chemical master equation*, The Journal of chemical physics **124** (2006), no. 4, 044104.

[27] J. Nocedal and S. J. Wright, *Numerical optimization*, second ed., Springer, New York, 2006.

[28] Sergey Plyasunov and Adam P. Arkin, *Efficient stochastic sensitivity analysis of discrete event systems*, J. Comp. Phys. **221** (2007), 724 – 738.

[29] Philip Protter, *Stochastic integration and differential equations*, Springer, 1990.

[30] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi, *Stochastic mRNA synthesis in mammalian cells*, PLoS biology **4** (2006), no. 10, e309.

[31] Muruhan Rathinam, *Moment growth bounds on continuous time Markov processes on non-negative integer lattices*, arXiv:1304.5169.

[32] Muruhan Rathinam, Patrick W. Sheppard, and Mustafa Khammash, *Efficient computation of parameter sensitivities of discrete stochastic chemical reaction networks*, Journal of Chemical Physics **132** (2010), 034103.

[33] J.B. Rawlings and J.G. Ekerdt, *Chemical reactor analysis and design fundamentals*, Nob Hill Pub., 2002.

[34] Martin I Reiman and Alan Weiss, *Sensitivity analysis for simulations via likelihood ratios*, Operations Research **37** (1989), no. 5, 830–844.

[35] Reuven Y Rubinstein and Alexander Shapiro, *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method*, vol. 346, Wiley New York, 1993.

[36] Walter Rudin, *Principles of mathematical analysis*, McGraw-Hill New York, 1976.

[37] PW Sheppard, M. Rathinam, and M. Khammash, *A pathwise derivative approach to the computation of parameter sensitivities in discrete stochastic chemical systems.*, The Journal of chemical physics **136** (2012), no. 3, 034115.

[38] J.C. Spall, *Feedback and weighting mechanisms for improving Jacobian estimates in the adaptive simultaneous perturbation algorithm*, Automatic Control, IEEE Transactions on **54** (2009), no. 6, 1216–1229.

[39] Rishi Srivastava, David F Anderson, and James B Rawlings, *Comparison of finite difference based methods to obtain sensitivities of stochastic chemical kinetic models*, The Journal of chemical physics **138** (2013), no. 7, 074110.

[40] Stephen G Strickland, *Gradient/sensitivity estimation in discrete-event simulation*, Proceedings of the 25th conference on Winter simulation, ACM, 1993, pp. 97–105.

[41] Tianhai Tian and Kevin Burrage, *Stochastic models for regulatory networks of the genetic toggle switch*, Proceedings of the National Academy of Sciences **103** (2006), no. 22, 8372–8377.

[42] Elizabeth Skubak Wolf and David F. Anderson, *A finite difference method for estimating second order parameter sensitivities of discrete stochastic chemical reaction networks*, J. Chem. Phys. **137** (2012), no. 22, 224112.