EFFICIENCY OF THE GIRSANOV TRANSFORMATION APPROACH FOR PARAMETRIC SENSITIVITY ANALYSIS OF STOCHASTIC CHEMICAL KINETICS

TING WANG* AND MURUHAN RATHINAM†

Abstract. For stochastic reaction networks consisting of several species, Monte Carlo methods are the most suitable for parametric sensitivity analysis. Most of the Monte Carlo methods for sensitivity analysis can be classified into three categories, the pathwise derivative (PD) method, the finite difference (FD) method and the Girsanov transformation (GT) method. It has been numerically observed in the literature that when applicable, the PD method and FD method tend to be more efficient than the GT method. In this work, we provide a theoretical justification for this observation in terms of system size asymptotic analysis. We illustrate our theory via some numerical examples.

Key words. stochastic chemical kinetics, Girsanov change of measure, finite difference, parametric sensitivity, variance analysis, density dependent processes.

AMS subject classifications. Primary: 60H35, 65C99; Secondary: 92C42, 92C45

1. Introduction. Dynamical models of real world systems often contain parameters that need to be determined from observations or measurements. Estimation of parametric sensitivities of dynamical systems is an essential part of the modeling and parameter estimation process. For instance the problem of finding the set of parameters that best fit some observed data can be formulated as an optimization problem over the parameter space where the partial derivatives of the objective function depend on the parametric sensitivities defined as partial derivatives of some system output with respect to the parameters.

In deterministic dynamical systems governed by ordinary differential equation (ODEs) the sensitivities defined by partial derivatives of some function of the state with respect to parameters \( \frac{\partial f(X(t))}{\partial c_k} \) are essentially computed by numerical integration of an auxiliary system of evolution equations obtained by linearization of the original ODEs.

In contrast for stochastic dynamical systems several different approaches exist. Stochastic dynamical systems in continuous time that are Markovian may be described by Kolmogorov’s forward equations which are either partial differential equations (PDEs) or systems of coupled ODEs which frequently involve very large or infinite number of variables. The former arise for instance when the Markov process is governed by an Ito stochastic differential equation (SDE) driven by Brownian motion and the latter arise for instance when the Markov process evolves on the discrete but infinite state space \( \mathbb{Z}^n \) or \( \mathbb{Z}_+^n \). If the dimensionality of the state space of the Markov process is large as in several applications, then direct solution of the forward equation is computationally prohibitive. In these instances Monte Carlo simulation of exact or approximate sample paths of the process is the only feasible approach. In this context several different Monte Carlo approaches exist for the numerical computation of the parametric sensitivities as well. Parametric sensitivities of a stochastic process \( X(t) \)
are typically defined by the partial derivatives

$$\frac{\partial}{\partial c_k} \mathbb{E}(f(X(t))),$$

where $c_k$ are scalar parameters, $f$ is some suitable scalar function of the state space, $\mathbb{E}$ is the expectation and $t > 0$ is some fixed final time. For simplicity we shall focus on one scalar parameter $c$. We shall also focus on Markov processes in continuous time with discrete state space $\mathbb{Z}_n^+$, with certain form of parametric dependence for the intensity functions. Specifically, we shall focus on stochastic chemical kinetic models as the motivating examples. Our analysis will be applicable to models arising in stochastic epidemiology, predator-prey dynamics and other forms of population processes.

All the Monte Carlo methods for computing the sensitivity involve the estimation of the expected value $\mathbb{E}(S(t))$ of some process $S$ at time $t > 0$ via iid sample estimation, where $S(t)$ can be computed easily from the knowledge of system parameters and the sample path of $X$ on the time interval $[0, t]$. As a general reference we suggest [3, 9].

The Monte Carlo methods can broadly be categorized into finite difference (FD) methods [2, 3, 16], pathwise derivative (PD) methods [3, 20] and the likelihood ratio or the Girsanov transformation (GT) methods [3, 14]. In the first two approaches (FD and PD) the set up is such that the stochastic process $X$ depends on the parameter $c$ and the underlying random elements are independent of $c$ which is the same as fixing a probability space $(\Omega, \mathcal{F}, P)$ and having a family of processes $X(t, c, \omega)$ where $\omega \in \Omega$. In the GT approach the process $X$ is regarded as independent of $c$ and one considers a family of probability measures $P(c)$ that depend on $c$, thus the quantity $\mathbb{E}(f(X(t)))$ still has the same dependence on $c$.

The FD methods involve approximation of the partial derivative by the finite difference $\mathbb{E}[f(X(t,c + h)) - f(X(t,c - h))]/(2h)$. We note that the efficiency of a Monte Carlo method for sensitivity depends critically on the variance of $S(t)$. In the case of FD,

$$S(t) = [f(X(t,c + h)) - f(X(t,c - h))]/(2h),$$

and when $f(X(t,c + h))$ and $f(X(t,c - h))$ are strongly positively correlated, one may expect the variance to be small. In the PD method

$$S(t) = \frac{\partial}{\partial c} f(X(t,c)),$$

and the method is applicable provided the derivative exists with probability one, analytical computation of the derivative is possible and the commutation

$$\mathbb{E} \left( \frac{\partial}{\partial c} f(X(t,c)) \right) = \frac{\partial}{\partial c} \mathbb{E}(f(X(t,c))) \quad \text{(1.1)}$$

holds.

In the GT method, supposing the sensitivity at $c = c_0$ is required, one denotes by $P(c)$ the law on the path space, of the process $X$ and under certain conditions the law $P(c)$ is absolutely continuous with respect to $P(c_0)$ for $c$ in a neighborhood of $c_0$. If

$$L(c, t) = \frac{dP(c)}{dP(c_0)},$$
the Radon-Nikodym derivative on the \( \sigma \)-algebra \( \mathcal{F}_t \) generated by the process up to time \( t > 0 \) then one may write

\[
\frac{\partial}{\partial c} \mathbb{E}(f(X(t))) = \frac{\partial}{\partial c} \int_{\Omega} f(X(t)) L(c, t) dP(c_0) = \int_{\Omega} f(X(t)) \frac{\partial}{\partial c} L(c, t) dP(c_0)
\]

(1.2)

provided the differentiation inside the integral is valid. If

\[
Z(t) = \left. \frac{\partial}{\partial c} \right|_{c=c_0} L(c, t),
\]

(1.3)

is analytically tractable then the required sensitivity is given by

\[
\frac{\partial}{\partial c} \mathbb{E}(f(X(t))) = \mathbb{E}_{c_0}[f(X(t))Z(t)],
\]

thus the sensitivity estimator \( S(t) = f(X(t))Z(t) \), where the subscript \( c_0 \) means that the expectation is taken with respect to \( P(c_0) \).

In many applications the GT method is applicable in that the “weight” \( Z(t) \) is easy to compute along with the simulation of \( X(t) \) and the differentiation inside the integral in (1.2) is valid. The PD method is not always valid as the commutation in (1.1) is frequently invalid. In this paper we shall be concerned with processes \( X(t) \) with state space \( \mathbb{Z}_n \) or \( \mathbb{Z}_n^+ \) and in that context \( X(t) \) has sample paths that are piecewise constant and \( \partial f(X(t, c))/\partial c \) is zero with probability one, while the sensitivity \( \partial \mathbb{E}(f(X(t, c)))/\partial c \) is in general non-zero, showing that the commutation in (1.1) is not valid. It is possible to regularize the problem by replacing \( \partial f(X(t, c))/\partial c \) with

\[
\frac{\partial}{\partial c} \frac{1}{2w} \int_{t-w}^{t+w} f(X(s, c)) ds,
\]

to obtain an estimator for which the commutation of derivative with expectation holds for a restricted class of examples [20]. Also see [9] in the context of computing the sensitivity of path integrals.

It has been observed that the PD method when applicable yields an estimator with lower variance than the GT estimator which is applicable in most situations [3, 20]. In the context of processes on integer lattices, the regularized PD (RPD) method is only applicable to a limited class of examples and results in a biased estimator [20]. The FD methods also result in biased estimators. Both the FD and RPD methods also involve the use of method parameters, \( h \) or \( w \), and the smaller these are the less the bias of these methods. However decreasing \( h \) or \( w \) results in an increase in the variance of the FD or RPD estimators respectively. The GT estimator on the other hand is unbiased and does not involve method parameters to be determined. However, it has been observed that in many situations the GT estimator has much larger variance compared to the FD and RPD estimators [3, 14, 16, 20]. To our knowledge, no theoretical explanation has been presented for the large variance of the GT method observed in many applications. In this paper, we provide a theoretical explanation for the large variance. In stochastic chemical kinetics as well as other population models, there is a “system size parameter” \( N \) and in the \( N \to \infty \) these
systems behave deterministically. Our analysis shows that the variance of the GT method grows much faster in \( N \) than the variances of RPD and FD methods.

Recently introduced methods auxiliary path algorithm (APA) \cite{10} and Poisson path algorithm (PPA) \cite{11} do not strictly belong to the three categories mentioned above. While they are closely related to the FD and the PD methods, they provide unbiased estimators similar to the GT. We do not investigate these methods in this paper.

The rest of the paper is organized as follows. In Section 2, we briefly review the stochastic chemical kinetic model as well as the random time change representation. We also present a numerical example to motivate the system size analysis and we present an alternative to the GT method, the centered GT (CGT) method. In Section 3, we describe the system size scaling and state some running assumptions for the rest of the analysis. Section 4 proves some asymptotic results and in Section 5 we present the asymptotic behavior of the sensitivity as well as the variances of the GT, CGT and FD estimators. Section 6 presents numerical results that support the analysis and in Section 7 we present some concluding remarks.

2. Stochastic chemical kinetics and a motivating example. While we focus on the description of stochastic chemical kinetics, we note that analogous models appear in other fields such as epidemiology and predator-prey models.

Let us consider a chemical reaction system consisting of \( m \) reaction channels and \( n \) chemical species \( \{S_1, \ldots, S_n\} \). The \( n \)-dimensional state vector \( X(t) \) characterizes the state of the system where each entry \( X_i(t) \) represents the number of molecules of the species \( S_i \) at time \( t \). The process \( X \) is assumed to be Markov with infinitesimal generator \( A \) given by

\[
(\mathcal{A} f)(x) = \sum_{j=1}^{m} a_j(x)(f(x + \nu_j) - f(x))
\]

for any bounded \( f : \mathbb{R}^n \to \mathbb{R} \).

The firing of a reaction channel \( j \in \{1, \ldots, m\} \) at time \( t \) causes the state to be incremented by the stoichiometric metric vector \( \nu_j \); \( X(t) = X(t-) + \nu_j \) thus we assume that \( X \) is càdlàg, i.e. paths of \( X \) are right continuous with left-hand limits.

For \( j = 1, \ldots, m \) we denote by \( R_j(t) \) the number of firings of the \( j \)-th reaction channel during \((0, t]\). Thus \( X(t) = X(0) + \nu R(t) \) for \( t \geq 0 \), where \( \nu \) is the stoichiometric matrix whose \( j \) column is \( \nu_j \) and \( R(t) = (R_1(t), \ldots, R_m(t))^T \). We note that \( R(0) = 0 \) and \( R_j(t) - R_j(t-) \) is either 0 or 1.

Associated with each reaction channel is an intensity function also known as propensity function in the chemical kinetics literature \( a_j(x), j = 1, \ldots, m \), which is such that given \( X(t) = x \) the probability of one or more firing of reaction channel \( j \) during \((t, t + h]\) is \( a_j(x)h + o(h) \) as \( h \to 0+ \). Following the terminology of \cite{9} we note that \( R_j \) are counting processes which admit the \( \mathcal{F}_t \)-predictable intensity process \( a_j(X(t-)) \) where \( \mathcal{F}_t \) is the filtration generated by \( X \) and \( R \).

One can use the random time change representation \cite{7} to express \( X \) via the stochastic equation

\[
X(t) = X(0) + \sum_{j=1}^{m} Y_j \left( \int_{0}^{t} a_j(X(s)) ds \right) \nu_j \tag{2.1}
\]
where $Y_j$ are independent unit rate Poisson processes. It follows that

$$R_j(t) = Y_j \left( \int_0^t a_j(X(s)) \, ds \right), \quad j = 1, \ldots, m.$$ 

In the description of these processes $X(t)$ and $R(t)$, there are parameters $c = (c_1, \ldots, c_p)$ involved. We assume that these parameters enter only via the intensity functions and not via the stoichiometric vectors. Thus $a_j = a_j(x, c)$ for $j = 1, \ldots, m$ while the $\nu_j$ for $j = 1, \ldots, m$ do not depend on $c$. We are concerned with the determination of parametric sensitivities of the form

$$\frac{\partial}{\partial c_k} \mathbb{E}(f(X(t))),$$

where $t > 0$ and $f : \mathbb{R}^n \to \mathbb{R}$.

In the context of stochastic chemical kinetics, the weight process $Z$ defined by (1.3) is given by [14, 20]

$$Z(t) = \sum_{j=1}^m \int_0^t \frac{\partial a_j(X(s), c_0)}{a_j(X(s), c_0)} \, dR_j(s) - \sum_{j=1}^m \int_0^t \frac{\partial a_j(X(s), c_0)}{c} \, ds.$$ 

(2.2)

It has been observed that the GT method is not very efficient in the sense that the estimator has large variance compared to other methods such as finite difference or pathwise derivatives [5, 20]. To our knowledge a theoretical explanation for this has not been provided in the literature. When the coefficient $c_1$ is really close to 0 it is possible to explain the large variance of the GT estimator [10]. This explanation does not apply when $c_1$ is not close to 0. The GT method appears to have larger variance than other methods even when $c$ is not close to 0.

We also investigate a modified GT method, which we call the centered Girsanov transformation (CGT) method in which we replace the estimator $f(X(t))Z(t)$ with $(f(X(t)) - \mathbb{E}(f(X(t))))Z(t)$. Since $Z(t)$ has zero mean this new estimator has the same mean as the original one and hence is also unbiased. In practice $\mathbb{E}(f(X(t)))$ is not known and needs to be estimated as well. We note that a similar approach was used in [21]. The following simple example is instructive.

**Example.** Consider the pure death model

$$S \xrightarrow{c} \emptyset$$

with intensity function $a(x) = cx$. We want to understand the sensitivity of the expected number of molecule $S$ with respect to the parameter $c$. The population process $X(t)$ and the weight process $Z(t)$ in this example are

$$X(t) = x_0 - \int_{(0,t]} dR(s),$$

$$Z(t) = \int_{(0,t]} \frac{1}{c} dR(s) - \int_0^t X(s) \, ds.$$ 

(2.3)

One can use the Ito formula for processes driven by finite variation processes (see [19]) to write down the equation for $X(t)Z(t)$,

$$X(t)Z(t) = \int_{(0,t]} \left( \frac{X(s-)}{c} - Z(s-) - \frac{1}{c} \right) \, dR(s) - \int_0^t X(s)^2 \, ds.$$

$$\frac{\partial}{\partial c_k} \mathbb{E}(f(X(t))),$$
Taking expected value of both sides leads to (after simplification)

\[ 
E(X(t)Z(t)) = -c \int_0^t E(X(s)Z(s)) \, ds - \int_0^t E(X(s)) \, ds. \tag{2.4} 
\]

Also note that

\[ 
E X(t) = x_0 - \int_0^t E X(s) \, ds. \tag{2.5} 
\]

The last equation and (2.4) form a system of linear ODEs with initial condition \(E X(0) = x_0\) and \(E X(0)Z(0) = 0\). One can easily solve this ODE system to obtain

\[ 
E(X(t)Z(t)) = \frac{x_0}{1-c}(e^{-ct} - e^{-ct}). 
\]

Similarly, to find an analytical formula for \(E(X(t)^2Z(t)^2)\), one can derive a linear system of ODEs involving variables of the form \(E(X(t)^\alpha Z(t)^\beta)\), where \(\alpha\) and \(\beta\) are integers satisfying \(\alpha \leq 2, \beta \leq 2\). The variances of GT and CGT estimators can be shown to be

\[ 
\text{Var}(X(t)Z(t)) = \frac{1}{c^2}(e^{-2ct}x_0^3 - 4e^{-2ct}x_0^2 + 3e^{-2ct}x_0 + 3e^{-2ct}x_0^2c^2 \\
- 2e^{-3ct}x_0 + 3e^{-3ct}x_0^2 + e^{-ct}x_0^2 - e^{-ct}x_0 \\
+ e^{-ct}x_0^2c^2 - 4e^{-2ct}x_0^2c^2x_0 - e^{-3ct}x_0^3), \tag{2.6} 
\]

and

\[ 
\text{Var}((X(t) - E[X(t)])Z(t)) = \frac{1}{c^2}(-2e^{-2ct}x_0^3 + 3e^{-2ct}x_0^2 + e^{-2ct}x_0^2c^2 \\
- 2e^{-3ct}x_0 + e^{-3ct}x_0^2 + e^{-ct}x_0^2 \\
- e^{-ct}x_0 + e^{-ct}x_0^2c^2 - 4e^{-2ct}x_0^2c^2). \tag{2.7} 
\]

Depending on the values of \(c, t\) and \(x_0\) the variance of the GT method may be more or less than that of CGT. However it is instructive to focus on the dependence on the initial state \(x_0\). The variance of GT estimator contains \(x_0^3\) while that of CGT estimator only has \(x_0^2\) involved in the formula. If \(x_0\) is modestly large (say 100), a significant amount of variance reduction can be expected using CGT. On the other hand, if we consider the variance of the FD estimator, we can bound the variance as

\[ 
\frac{1}{h^2} \text{Var}(X(t, c + h) - X(t, c)) \leq \frac{2}{h^2} (\text{Var}(X(t, c + h)) + \text{Var}(X(t, c))). 
\]

Since \(\text{Var}(X(t, c))\) depends on \(x_0\) linearly (see for instance [15]), we expect the variance of the FD estimator to be even smaller than that of the CGT.

These observations motivate the analysis in the rest of this paper. We note that \(x_0\) can be thought of as a measure of system size. We will show that when the system size grows according to the classical scaling the variance of the FD methods grow the smallest in terms of system size, while the variance of the CGT method grows modestly and the variance of the GT method grows the fastest for more general chemical reaction systems.
3. Scaling with system size. As mentioned in the previous section, the system size shall be the key to our analytical explanation for the larger variance of the GT estimator. In this section we set the stage for the system size analysis and state some assumptions that shall be carried throughout the rest of the paper.

We first describe the stochastic mass action form of intensities that commonly arise in stochastic chemical kinetics [8]. If we divide the stoichiometric vector $\nu_j$ into two parts, such that $\nu_j = \nu_j' - \nu_j''$, where $\nu_j'$: the vector number of molecules of each species that are created in the $j$th reaction, $\nu_j''$: the vector number of molecules of each species that are consumed in the $j$th reaction, then the intensity of the $j$th reaction is

$$a_j^N(x, c) = \frac{c_j}{N^{\vert \nu_j''\vert - 1}} \prod_{i} \left( \frac{x_i}{\nu_{ij}''} \right)$$

(3.1)

where $\vert \nu_j''\vert = \sum_{i=1}^{n} \nu_{ij}''$ and $N$ is the volume of the system times Avogadro’s number, $c_j$ is a constant specifying the rate of the reaction. We note that the term $\left( \frac{x_i}{\nu_{ij}''} \right)$ represents the number of ways to choose $\nu_{ij}''$ molecules from $x_i$ molecules of the $i$th species. Thus the intensity functions $a_j^N$ depend on $N$ and $x$ in a specific manner referred to as density dependence [7]. This density dependence leads to a deterministic limiting behavior in the large system size ($N \to \infty$). As a result it is instructive to study the family of processes $X_j^N$ indexed by $N \geq 1$ corresponding to the family of intensity functions $a_j^N$. Following [7] one can represent all these processes on the same sample space via the stochastic equation

$$X_j^N(t) = X_j^N(0) + \sum_{j=1}^{m} Y_j \left( \int_{0}^{t} a_j^N(X_j^N(s)) \, ds \right) \nu_j, \quad N \geq 1,$$

(3.2)

where $Y_j$ are independent unit rate Poisson processes. We also define the corresponding family of vector reaction count processes $R_j^N(t)$ whose $j$th component $R_j^N(t)$ counts the number of reaction events of type $j$ that occurred during $(0, t]$. Thus

$$R_j^N(t) = Y_j \left( \int_{0}^{t} a_j^N(X_j^N(s)) \, ds \right), \quad N \geq 1, \quad j = 1, \ldots, m.$$

We also define the centered processes $M_j^N(t) = (M_j^N(t), \ldots, M_m^N(t))$ by

$$M_j^N(t) = R_j^N(t) - \int_{0}^{t} a_j^N(X_j^N(s)) \, ds, \quad N \geq 1, \quad j = 1, \ldots, m.$$

Now we shall make some explicit assumptions which shall hold throughout the rest of this paper.

**Assumption 1** We assume the following form of parameter dependence on the intensity function. For each $j = 1, \ldots, m$ and $N$,

$$a_j^N(x, c) = c_j b_j^N(x),$$

(3.3)

where $b_j^N : \mathbb{R}^n \to \mathbb{R}$ are such that $b_j^N$ restricted to $\mathbb{Z}_+^n$ are nonnegative. This also implies that there are precisely $m$ parameters, one for each reaction $j$. 

For the analysis in this paper we need not assume the stochastic mass action form, but merely the density dependence which is stated by our Assumption 2.

**Assumption 2** We suppose that for each \( j = 1, \ldots, m \), there exists a function \( a_j(x) \) such that for each compact \( K \subset \mathbb{R}^n_+ \), the collection of functions \( a_j^N(Nx) - Na_j(x) \) is uniformly bounded for \( x \in K \) and \( N \geq 1 \). We note that this implies that for each compact set \( K \subset \mathbb{R}^n_+ \) there exists \( B_K > 0 \) such that

\[
\left| \frac{a_j^N(Nx)}{N} - a_j(x) \right| \leq \frac{B_K}{N}, \quad N \geq 1, \quad x \in K, \quad j = 1, \ldots, m. \tag{3.4}
\]

Defining \( X_N(t) = N^{-1}X^N(t) \), we note that \( X_N \) can be interpreted as the concentration of molecules at time \( t \) for system size \( N \). We note that \( X_N \) are coupled via the following stochastic equations.

\[
X_N(t) = X_N(0) + \sum_{j=1}^{m} N^{-1}Y_j \left( \int_0^t a_j^N(NX_N(s)) \, ds \right) \nu_j. \tag{3.5}
\]

We state the following theorem regarding the limiting behavior of \( X_N \) (see [7] for details).

**Theorem 3.1.** (Law of large number [7]) Suppose for each compact \( K \subset \mathbb{R}^n_+ \),

\[
\sum_{j=1}^{m} |\nu_j| \sup_{x \in K} a_j(x) < \infty,
\]

and \( F(x) = \sum_{j=1}^{m} \nu_j a_j(x) \) is Lipschitz on \( K \), that is, for each \( x, y \in K \),

\[
|F(x) - F(y)| \leq M_K |x - y|.
\]

Suppose \( X_N(0) = x_0 \in \mathbb{R}^n_+ \) for all \( N \). Then for every \( t \geq 0 \),

\[
\lim_{N \to \infty} \sup_{s \leq t} |X_N(s) - X(s)| = 0 \quad \text{a.s.,}
\]

where the deterministic limit \( X \) satisfies \( X(t) = x_0 + \int_0^t F(X(s)) \, ds \).

**Remark:** We note that with fixed initial condition \( X_N(0) = x_0 \) we want \( X_N(0) = Nx_0 \) to belong to \( \mathbb{Z}^n_+ \), which may not hold for all \( N \geq 1 \) but we assume that it holds for a sequence of \( N \) values tending to \( \infty \). For instance if \( x_0 \) is rational this is true. This is adequate for our purposes.

In order to satisfy the conditions stated in Theorem 3.1 we shall assume the following.

**Assumption 3** For each \( j = 1, \ldots, m \), the functions \( a_j : \mathbb{R}^n \to \mathbb{R} \) are continuously differentiable. This automatically implies the Lipschitz condition in Theorem 3.1.

The following assumption is used to facilitate the analysis in this paper. Several, but not all examples in applications satisfy this assumption.

**Assumption 4** We take \( X_N(0) = x_0 \) (deterministic) for all \( N \). We assume that the sequence of concentration processes \( X_N \) is uniformly bounded, that is, there exists a constant \( \Gamma \) such that for all \( t \geq 0 \),

\[
|X_N(t)| \leq \Gamma \quad \text{a.s.} \tag{3.6}
\]

for all \( N \geq 1 \).
We note that if there exists a strictly positive vector $\alpha \in \mathbb{R}_+^m$ so that $\alpha^T \nu_j \leq 0$ for each $j$ then this assumption is satisfied. We note that a form of converse of this statement is also true [17].

Now we turn our attention to the sensitivity. Given $f : \mathbb{R}^n \to \mathbb{R}$, we are interested in computing the sensitivity

$$\frac{\partial}{\partial c} \mathbb{E}(f(X^N(t))),$$

where $c \in (0, \infty)$ is a parameter. In view of Assumption 1, without loss of generality, we shall take $c = c_1$. Then we note that the GT sensitivity estimator is $f(X^N(t))Z^N(t)$ and the CGT estimator is $[f(X^N(t)) - \mathbb{E}(f(X^N(t)))]Z^N(t)$, where we note that $Z^N(t) = M_1^N(t)/c_1$ in this case.

As we are concerned with families of processes indexed by $N$, it makes sense to consider a corresponding family of functions $f^N : \mathbb{R}^n \to \mathbb{R}$ instead of one function $f$ and make reasonable assumptions on $f^N$ and $f$.

To motivate the assumption we make on $f^N$ and $f$ we note that we shall be concerned with $f^N(X_N(t)) = f^N(NX_N(t))$ which we wish to compare with $f(X(t))$. When $f^N(x) = x_1$, one of the components of $x$, we have

$$f^N(NX_N(t))/N = X_{N_1}(t) \to X_1(t) = f(X(t)),$$

with $f(x) = x_1$. Alternatively, if $f^N(x) = x_1^{\alpha}$ for some $\alpha > 0$ we have

$$f^N(NX_N(t))/N^\alpha = (X_{N_1}(t))^{\alpha} \to (X_1(t))^{\alpha} = f(X(t)),$$

with $f(x) = x_1^{\alpha}$. If however $f^N(x) = x_1^2 + x_1$ then we have

$$f^N(NX_N(t))/N^2 = (X_{N_1}(t))^2 + X_{N_1}(t)/N \to (X_1(t))^2 = f(X(t)),$$

where $f(x) = x_1^2$. In this case we note that $f^N(Nx)/N^2 - f(x) = x_1/N$ which tends to 0 as $1/N$, uniformly for $x$ in a compact set. Motivated by this we impose the following assumption.

**Assumption 5** We assume that there exist a function $f$ and a constant $\alpha > 0$ such that for each compact set $K \subset \mathbb{R}_+^m$,

$$|f^N(Nx)/N^\alpha - f(x)| \leq \frac{L_K}{\sqrt{N}}, \quad x \in K, \quad N \geq 1 \quad (3.7)$$

for some constant $L_K > 0$. We remark that the $O(1/\sqrt{N})$ behavior is adequate for our proofs.

4. Large $N$ behavior. In this section we derive results concerning the $N \to \infty$ limit for the various relevant processes.

Throughout the rest of the paper $X(t)$ will denote the solution to the equation

$$X(t) = x_0 + \sum_{j=1}^m \nu_j \int_0^t a_j(X(s))ds, \quad (4.1)$$

where $x_0 \in \mathbb{R}_+^m$ is fixed.

**Lemma 4.1.** We have for each $j = 1, \cdots, m$, there exists $A_j > 0$ such that for all $t > 0$

$$\frac{a_j^N(NX_N(t))}{N} \leq A_j \quad a.s.$$
Proof. By (3.6) in Assumption 4, the processes $X_N$ are contained in a compact set of $\mathbb{R}^n$, say $K$, therefore for each $j$ we have the estimation

$$\sup_{t \geq 0} \frac{a_j^N(NX_N(t))}{N} \leq \sup_{x \in K} \frac{a_j^N(Nx)}{N}$$

Since $N^{-1}a_j^N(Nx)$ converges uniformly to $a_j(x)$ for $x$ in $K$ by (3.3), it is apparent that $\sup_{x \in K} N^{-1}a_j^N(Nx)$ is bounded by continuity of $a_j$. Hence $\sup_{t \geq 0} N^{-1}a_j^N(NX_N(t))$ is bounded by a constant $A_j$. $\Box$

**Lemma 4.2.** For $j = 1, \cdots, m$, and $t > 0$, we have

$$\sup_{s \leq t} \left| \frac{a_j^N(NX_N(s))}{N} - a_j(X(s)) \right| \to 0, \ a.s. $$

Proof. Write

$$\left| \frac{a_j^N(NX_N(s))}{N} - a_j(X(s)) \right| \leq \left| \frac{a_j^N(NX_N(s))}{N} - a_j(X_N(s)) \right| + \left| a_j(X_N(s)) - a_j(X(s)) \right|.$$ 

The first part on the right hand side converges to zero uniformly for $s$ in $[0, t]$ because of Assumptions 2 and 4. To see the second part on the right hand side converges uniformly to 0 on $[0, t]$, note that by Assumption 3 and Assumption 4, $a_j$ is Lipschitz continuous on the compact set $K$ (which contains $X_N$ and $X$), hence the result follows by Theorem 3.1 $\Box$

We define the sequence of scaled processes $R_N(t)$ by $R_N(t) = R^N(t)/N$.

**Lemma 4.3.** For each $j = 1, 2, \cdots, m$ and $t > 0$

$$\sup_{s \leq t} \left| R_N(j) - \int_0^s a_j(X(u))du \right| \to 0 \ a.s. $$

Proof. Recall that $R_j^N(t) = Y_j \left( \int_0^t a_j^N(NX_N(s))ds \right)$. For each $j = 1, \cdots, m$,

$$\sup_{s \leq t} \left| \frac{1}{N} Y_j \left( \int_0^s a_j^N(NX_N(u))du \right) - \int_0^s a_j(X(u))du \right|$$

$$\leq \sup_{s \leq t} \left| \frac{1}{N} Y_j \left( \int_0^s a_j^N(NX_N(u))du \right) - \frac{1}{N} \int_0^s a_j^N(NX_N(u))du \right|$$

$$+ \int_0^t \left| \frac{1}{N} a_j^N(NX_N(u)) - a_j(X(u)) \right| du.$$ 

The second term on the right hand side converges to zero by Lemma 4.2. Setting $\hat{Y}(t) = Y(t) - t$, the first term on the right can be written and then bounded as

$$\sup_{s \leq t} \left| \frac{1}{N} \hat{Y}_j \left( \int_0^s a_j^N(NX_N(u))du \right) \right| \leq \sup_{s \leq t} \left| \frac{1}{N} \hat{Y}_j (NA_j s) \right| \ a.s.$$
where the last term converges to zero by the law of large numbers for Poisson process.

**Lemma 4.4.** For a given \( t > 0 \), suppose that \( f \) is continuous at \( X(t) \). Then

\[
\lim_{N \to \infty} |f^N(NX_N(t))/N^\alpha - f(X(t))| = 0, \quad \text{a.s.} \tag{4.2}
\]

**Proof.** Write

\[
|f^N(NX_N(t))/N^\alpha - f(X(t))| \leq |f^N(NX_N(t))/N^\alpha - f(X_N(t))| + |f(X_N(t)) - f(X(t))|.
\]

The first term converges to zero almost surely by Assumption 4 and (3.7) in Assumption 5. The second term converges to zero by the continuity assumption on \( f \) since \( X_N(t) \) converges to \( X(t) \) almost surely.

Recall the definition of \( M^N \),

\[
M^N(t) = R^N(t) - \int_0^t a^N(NX_N(s))ds.
\]

Note that in general \( M^N(t) \) is an \( m \)-dimensional local martingale for each \( N \), but by Lemma 4.1 it follows that \( \mathbb{E}[R_j^N(t)] \leq NA_j t \) for all \( t > 0 \) which makes \( M^N(t) \) a martingale. We define the scaled processes \( M^N = N^{-1}M^N \) and \( Z^N = N^{-1}Z^N \). We note that \( Z^N(t) = M^N_1(t)/c_1 \) and \( Z_N(t) = M_{N1}(t)/c_1 \).

Let us denote by \( D[0, \infty) \) the space of càdlàg functions mapping from \([0, \infty)\) to \( \mathbb{R}^m \) endowed with the Skorohod topology. We provide a lemma on the weak convergence of \( M^N \).

**Theorem 4.5.** Let \( C(t) = (c_{ij}(t)) \) be a \( m \times m \) matrix of functions, where

\[
c_{ij}(t) = \left\{ \begin{array}{cl}
\int_0^t a_j(X(s))ds & i = j \\
0 & i \neq j
\end{array} \right. \tag{4.3}
\]

Then \( \sqrt{N}M_N \Rightarrow \bar{M} \) on \( D^m[0, \infty) \), where \( \bar{M}(t) \) is an \( m \)-dimensional Gaussian process with independent increments having mean vector and covariance matrix

\[
\mathbb{E}[\bar{M}(t)] = (0, \cdots, 0), \quad \mathbb{E}[\bar{M}(t)\bar{M}(t)^T] = C(t). \tag{4.4}
\]

Furthermore, the scaled Girsanov sensitivity (or weight) process \( \sqrt{N}Z_N \Rightarrow U \) on \( D[0, \infty) \), where

\[
U(t) = \frac{1}{c_1} \bar{M}_1(t). \tag{4.5}
\]

Also since \( U \) has continuous sample paths, for each \( t > 0 \) we have

\[
\sqrt{N}Z_N(t) \Rightarrow U(t).
\]

**Proof.** The proof relies on Theorem B.1 the martingale FCLT in the appendix. Note that each jump of \( \sqrt{N}M_N \) has size \( 1/\sqrt{N} \), therefore,

\[
\lim_{N \to \infty} \mathbb{E} \left[ \sup_{s \leq t} \left| \sqrt{N}M_N(s) - \sqrt{N}M_N(s-) \right| \right] = 0.
\]
Also, for each pair \((i, j)\) with \(i, j = 1, \ldots, m\), and each \(t\), since the jump size for \(M_{N,j}\) is always \(N^{-1}\) and there are no simultaneous jumps, we have the following quadratic covariation

\[
\left[ \sqrt{NM_{N,i}}, \sqrt{NM_{N,j}} \right] (t) = \begin{cases} 
R_{N,j} & i = j \\
0 & i \neq j 
\end{cases} \tag{4.6}
\]

By Lemma 4.3, \(R_{N,j}(t)\) converges almost surely to \(c_{jj}(t) = \int_0^t a_j(X(s))ds\). Then, for each pair \((i, j)\),

\[
\left[ \sqrt{NM_{N,i}}, \sqrt{NM_{N,j}} \right] (t) \to c_{ij}(t)
\]

almost surely and hence in probability. Thus, the weak convergence of \(M_N\) follows from the martingale FCLT.

**Lemma 4.6.** For each \(p \geq 1\), there exists a constant \(\beta(p)\) such that for all \(t > 0\)

\[
\limsup_N \mathbb{E} \left( \sup_{s \leq t} \left| \sqrt{NM_N(s)} \right|^p \right) \leq \beta(p)t^{p/2}. \tag{4.7}
\]

**Proof.** Observe that the quadratic variation of \(\sqrt{NM_N}\) is

\[
\left[ \sqrt{NM_N}, \sqrt{NM_N} \right] (t) = N^{-1} \sum_{j=1}^m Y_j \left( \int_0^t a_j^N(NX_N(s))ds \right).
\]

By the Burkholder-Davis-Gundy inequality (see [15]), there exists a constant \(C(p)\) such that

\[
\mathbb{E} \left( \sup_{s \leq t} \left| \sqrt{NM_N(s)} \right|^p \right) \leq C(p) \mathbb{E} \left( \frac{1}{N} \sum_{j=1}^m Y_j \left( \int_0^t a_j^N(NX_N(s))ds \right) \right)^{p/2} \leq C(p) \mathbb{E} \left( \frac{1}{N} \sum_{j=1}^m Y_j (NA_j t) \right)^{p/2} \leq C(p) N^{-p/2} \left( \mathbb{E} \left( \sum_{j=1}^m Y_j (NA_j t) \right)^p \right)^{1/2},
\]

where we have used Lemma 4.4.

Hence,

\[
\limsup_N \mathbb{E} \left( \sup_{s \leq t} \left| \sqrt{NM_N(s)} \right|^p \right) \leq \limsup_N C(p) N^{-p/2} \left( \mathbb{E} \left( \sum_{j=1}^m Y_j (NA_j t) \right)^p \right)^{1/2}.
\]

First we observe that for \(j = 1, \ldots, m\), the \(p\)th moment of the Poisson random variable \(Y_j(NA_j t)\) is a polynomial of degree \(p\) in \(NA_j t\). Also, noting that \(Y_j\) are independent, we obtain that the right hand side is bounded by a term \(\beta(p)t^{p/2}\), where \(\beta(p)\) is a constant. \(\blacksquare\)
Since $Z^N(t) = c_1^{-1}M^N(t)$, we immediately have the following property regarding process $Z_N$.

**Lemma 4.7.** For each $p \geq 1$, there exists a constant $\gamma(p)$ such that for all $t > 0$,

$$
\limsup_{N} \mathbb{E} \left( \sup_{s \leq t} \sqrt{N} |Z_N(s)| \right)^p \leq \gamma(p)^{p/2}.
$$

Define the process $V_N(t) = \sqrt{N}(X_N(t) - X(t))$. Let us consider the moment of this process on a compact time interval.

**Lemma 4.8.** For each $p \geq 1$, there exist constants $\bar{\beta}(p), K(p)$ such that for all $t > 0$,

$$
\limsup_{N} \mathbb{E} (|V_N(s)|^p) \leq \bar{\beta}(p)^{p/2}t^{p/2}e^{K(p)t}.
$$

**Proof.** Recall that

$$X_N(s) = x_0 + \nu R_N(s)$$

and

$$X(s) = x_0 + \int_0^s \nu a(X(u))du,$$

where $\nu$ is the $n$ by $m$ dimensional stoichiometric matrix. One can write $V_N$ as follows,

$$V_N(s) = \sqrt{N}\nu R_N(s) - \sqrt{N} \int_0^s \nu a(X(u))du$$

$$= \sqrt{N}\nu \left( R_N(s) - \int_0^s \frac{a^N(NX_N(u))}{N} du \right)$$

$$+ \sqrt{N}\nu \left( \int_0^s \frac{a^N(NX_N(u))}{N} - a(X(u))du \right).$$

Note that we denote $M_N(s) = R_N(s) - \int_0^s N^{-1} a^N(NX_N(u)) du$, hence

$$|V_N(s)| \leq \|\nu\| \left| \sqrt{N}M_N(s) \right| + \|\nu\| \int_0^s \sqrt{N} \left| \frac{a^N(NX_N(u))}{N} - a(X(u)) \right| du.$$

We note that

$$\sqrt{N} \left| \frac{a^N(NX_N(u))}{N} - a(X(u)) \right| \leq \sqrt{N} \left| \frac{a^N(NX_N(u))}{N} - a(X_N(u)) \right|$$

$$+ \sqrt{N} |a(X_N(u)) - a(X(u))|.$$

Since $X_N$ lies in a compact set $K$ according to Assumption 4, we have for all $u > 0$,

$$\left| \frac{a^N(NX_N(u))}{N} - a(X_N(u)) \right| \leq \frac{B_K}{N}$$

where we have used Assumption 2 and abuse of notation using $B_K$ from (3.3).
On the other hand, for each $j = 1, \cdots, m$, by Assumption 3, $a_j$ is continuously differentiable and hence it is Lipschitz continuous on the compact set $K$. Hence, there exists a Lipschitz constant $C_j$ such that for all $u > 0$,

$$|a_j(X_N(u)) - a_j(X(u))| \leq C_j |X_N(u) - X(u)|.$$  

It follows that there exists a constant $C$ such that

$$|a(X_N(u)) - a(X(u))| \leq C |X_N(u) - X(u)|.$$  

Therefore,

$$|V_N(s)| \leq \|\nu\| \left( \sqrt{N}M_N(s) + N^{-1/2}B_K s + C \int_0^s \sqrt{N} |X_N(u) - X(u)| \, du \right)$$

$$= \|\nu\| \left( \sqrt{N}M_N(s) + N^{-1/2}B_K s + C \int_0^s |V_N(u)| \, du \right),$$

Applying the inequality $(a + b + c)^p \leq 3^p(a^p + b^p + c^p)$ and the Holder’s inequality, we obtain

$$|V_N(s)|^p \leq (3\|\nu\|)^p \left( \sqrt{N}M_N(s)^p + N^{-p/2}(B_K s)^p + C^p s^{p-1} \int_0^s |V_N(u)|^p \, du \right).$$

Taking expected value of both sides, for $s \in [0, t]$

$$\mathbb{E}|V_N(s)|^p \leq (3\|\nu\|)^p \left( \mathbb{E} \sqrt{N}M_N(s)^p + N^{-p/2}(B_K s)^p \right) + (3\|\nu\|)^p C^p s^{p-1} \left( \int_0^s \mathbb{E}|V_N(u)|^p \, du \right).$$

To estimate the first term of the right hand side, recall that in the proof of Lemma 4.6,

$$\mathbb{E} \left( \sup_{s \leq t} \sqrt{N}M_N(s) \right)^p \leq C(p)N^{-p/2} \left( \mathbb{E} \left( \sum_{j=1}^m Y_j(NA_j t) \right)^p \right)^{1/2}.$$  

For convenience, let us denote

$$\Phi_N(t) = C(p)N^{-p/2} \left( \mathbb{E} \left( \sum_{j=1}^m Y_j(NA_j t) \right)^p \right)^{1/2}.$$  

Therefore,

$$\mathbb{E}|V_N(s)|^p \leq (3\|\nu\|)^p \left( \Phi_N(t) + N^{-p/2}(B_K s)^p + C^p s^{p-1} \left( \int_0^s \mathbb{E}|V_N(u)|^p \, du \right) \right).$$

We note that $\mathbb{E}|V_N(s)|^p$ is continuous in $s$ and applying the Gronwall inequality, we obtain for $s \leq t$,

$$\mathbb{E}|V_N(s)|^p \leq (3\|\nu\|)^p \left( \Phi_N(t) + N^{-p/2}(B_K t)^p \right) e^{(3\|\nu\|)^p C^p s^p}.$$  

Taking supremum over $s \in [0, t]$ and then taking $\limsup_N$, the result follows from same considerations as in the proof of Lemma 4.6. QED.
5. **Scaling of sensitivity and estimator variance.** A measure of accuracy of a Monte Carlo estimator $S$ is the ratio of its standard deviation to the absolute value of its expected value, i.e.,

$$\frac{\sqrt{\text{Var}(S)}}{|\mathbb{E}(S)|}.$$ 

We refer to this quantity as the relative standard error (RSE).

In this section, we study the system size dependence of the sensitivity

$$\frac{\partial}{\partial c} \mathbb{E}(f^N(X^N(t)))$$

and the variances and RSEs of the GT, CGT and FD estimators. In the context of stochastic mass action form of intensities given by (3.1), we note that $c = c_j$ is the deterministic parameter while $c'_N = c_j/N^{\nu''_j} - 1$ is the stochastic parameter. In practice, one would compute sensitivity with respect to the stochastic parameter $c'_N$. The difference between the sensitivity with respect to the stochastic parameter and with respect to the deterministic parameter is merely a scaling factor $N^{\nu''_j} - 1$. Therefore, the RSE is unchanged regardless of whether one considers the sensitivity with respect to the stochastic parameter or the deterministic parameter. From an analytical point of view, it is convenient to study the sensitivity with respect to the deterministic parameter.

Recall that the sensitivity estimator of Girsanov transformation method is

$$f^N(X^N(t))Z^N(t,c)$$

where $f^N : \mathbb{R}^n \rightarrow \mathbb{R}$.

**Theorem 5.1.** In addition to our running assumptions, we assume that $f$ in (3.7) is continuously differentiable. Then for each $t \geq 0$

$$\sup_{s \leq t} \mathbb{E}(f^N(X^N(s))Z^N(s)) = O(N^\alpha).$$

That is, the sensitivity is asymptotically $O(N^\alpha)$ uniformly on $[0,t]$.

**Proof.** It is sufficient to show that $\sup_{s \leq t} \mathbb{E}(f^N(X^N(s))Z^N(s))/N^\alpha$ is bounded in $N$. Instead of working with $\mathbb{E}(f^N(X^N(s))Z^N(s))/N^\alpha$, we use

$$\mathbb{E}\left(\frac{f^N(X^N(s))}{N^\alpha}Z^N(s) - f(X(s))Z^N(s)\right)$$

because they are equal but the latter is easier to work with.

Note that $f$ is continuously differentiable hence Lipschitz on the compact set $K$ corresponding to Assumption 4. Denote by $C_K$ the Lipschitz constant for $f$. Using the assumptions on $f^N$ and $f$ and writing $X^N$ in terms of $V_N$ as before as

$$X^N(s) = NX(s) + \sqrt{NV_N(s)},$$
leads to

$$
\frac{f^N(NX(s) + \sqrt{NV_N(s)})}{N^\alpha} - f(X(s)) \to Z^N(s)\\
\leq \frac{f^N(NX(s) + \sqrt{NV_N(s)})}{N^\alpha} - f(X(s)) \frac{V_N(s)}{\sqrt{N}} \to Z^N(s)\\
+ f \left( X(s) + \frac{V_N(s)}{\sqrt{N}} \right) - f(X(s)) \to Z^N(s)\\
\leq \frac{L_N}{\sqrt{N}} |Z^N(s)| + C_K \frac{|V_N(s)|}{\sqrt{N}} Z^N(s)\\
\leq L_K \sqrt{N} Z^N(s) + \frac{1}{2} C_K \left( |V_N(s)|^2 + N |Z^N(s)|^2 \right).
$$

The result follows from Lemmas 4.7 and 4.8.

**Remark:** We believe that under the $N^\alpha$ scaling, the sensitivity of the stochastic process should limit to the sensitivity of the fluid limit ODE as $N \to \infty$. The above result is weaker than that.

**Theorem 5.2.** In addition to our running assumptions, we assume that $f$ in (3.7) is bounded on every compact set and for a given $t > 0$, $f$ is continuous at $X(t)$. Then we have as $N \to \infty$,

$$
N^{-1-2\alpha} E \left\{ \left( f^N(X^N(t))^2 \right) (Z^N(t))^2 \right\} \to (f(X(t)))^2 \frac{1}{c_1} \int_0^t a_1(X(s)) ds. \quad (S.1)
$$

Furthermore, for each $t > 0$,

$$
\sup_{s \leq t} E \left( (f^N(X^N(s))) Z^N(s) \right)^2 = O(N^{2\alpha+1}).
$$

**Proof.** Lemma 4.7 implies the uniformly integrability of $N^{-1}(Z^N(t))^2$. By Assumption 4 and (3.7) we have that $(f^N(X^N(t))^2)$ is a uniformly bounded sequence. Thus $N^{-1-2\alpha} f^N(X^N(t))^2 \to (f(X(t))^2)$ is uniformly integrable.

By Lemma 4.4 we have that $N^{-2\alpha} f^N(X^N(t))^2$ converges to $(f(X(t))^2)$ almost surely. We also have that $N^{-1} Z^N(t)$ converge weakly to $U(t)$. Thus by Theorem A.2 in appendix and the continuous mapping theorem we have that

$$
N^{-1-2\alpha} f^N(X^N(t))^2 \to (f(X(t))^2 U^2(t).
$$

By Theorem 3.5 from [5], we note that if a uniformly integrable sequence converges weakly then it converges in the mean, hence the result (S.1) follows.

Also, recall that $(f^N(X^N(t))^2) / N^{2\alpha}$ is uniformly bounded, hence

$$
N^{-2\alpha} \sup_{s \leq t} E \left( (f^N(X^N(s))) Z^N(s) \right)^2 \leq C E \left( \sup \sqrt{N} |Z^N(s)| \right)^2.
$$

Taking $\lim \sup N$ and applying Lemma 4.7 yields the seconds result.

Note that the above theorem does not assume $f$ is continuously differentiable. However, to state the result regarding the estimator variance for GT method, we still need to assume continuous differentiability on $f$ so that we can use Theorem 6.1.
Corollary 5.3. Suppose $f$ in (3.7) is continuous differentiable, then for given $t > 0$, the estimator variance of GT method is asymptotically $O(N^{2\alpha+1})$ uniformly on $[0, t]$.

Next, we will explore the variance of the centered Girsanov transformation approach.

Theorem 5.4. In addition to our running assumptions, we assume that $f$ in (3.7) is continuously differentiable. Then for each $t > 0$,

$$
\sup_{s \leq t} \mathbb{E} \left( (f^N(X^N(s)) - \mathbb{E} f^N(X^N(s))) Z^N(s) \right)^2 = O(N^{2\alpha}).
$$

Proof. Write

$$
\mathbb{E} \left( \left| \frac{f^N(X^N(s))}{N^\alpha} - \mathbb{E} \left( \frac{f^N(X^N(s))}{N^\alpha} \right) \right|^2 (Z^N(s))^2 \right)
\leq 2 \mathbb{E} \left( \left| \frac{f^N(X^N(s))}{N^\alpha} - f(X(s)) \right|^2 (Z^N(s))^2 \right)
+ 2 \mathbb{E} \left( \left| f(X(s)) - \mathbb{E} \left( \frac{f^N(X^N(s))}{N^\alpha} \right) \right|^2 (Z^N(s))^2 \right)
\leq 2 \mathbb{E} \left( \left| \frac{f^N(X^N(s))}{N^\alpha} - f(X(s)) \right|^2 (Z^N(s))^2 \right)
+ 2 \mathbb{E} \left( \left| \frac{f^N(X^N(s))}{N^\alpha} - f(X(s)) \right|^2 (Z^N(s))^2 \right) \mathbb{E}(Z^N(s))^2,
$$

where the last inequality is true due to the fact that $f(X(t))$ is deterministic. Using similar argument as in the proof of Theorem 5.1, the first term on the right-hand side can be bounded by

$$
4L^2_K \mathbb{E} \left( |\sqrt{N}Z_N(s)| \right)^2 + 4C^2_K \mathbb{E} \left( |V_N(s)|\sqrt{N}|Z_N(s)| \right)^2.
$$

Similarly, the second term on the right-hand side can be bounded by

$$
4L^2_K \mathbb{E} \left( \sqrt{N}|Z_N(s)| \right)^2 + 4C^2_K \mathbb{E}|V_N(s)|^2 \mathbb{E} \left( \sqrt{N}|Z_N(s)| \right)^2.
$$

Both of the above terms are bounded in $N$ uniformly on $[0, t]$ by Lemma 4.7 and 4.8.

Combining this result with Theorem 5.1, the following corollary is immediate.

Corollary 5.5. For any given $t > 0$, the estimator variance of CGT method is asymptotically $O(N^{2\alpha})$ uniformly on $[0, t]$.

Theorem 5.6. Suppose $f^N$ satisfies (3.7) and $f$ is continuously differentiable. Then for each $t > 0$ and $h > 0$,

$$
\sup_{s \leq t} \text{Var} \left( X^N(s, c + h) - X^N(s, c) \right) = O(N^{2\alpha-1}).
$$

That is, the estimator variance of FD method is asymptotically $O(N^{2\alpha-1})$. 

Proof. Note that
\[ \text{Var} \left( X^N(s, c + h) - X^N(s, c) \right) \leq 2 \text{Var} \left( X^N(s, c + h) \right) + 2 \text{Var} \left( X^N(s, c) \right), \]
hence it is sufficient to show that \( \text{Var} \left( X^N(t, c) \right) = \mathcal{O}(N^{2\alpha - 1}). \) We write
\[ \frac{1}{N^{2\alpha - 1}} \text{Var} \left( X^N(s, c) \right) = NE \left( \frac{f^N(X^N(s, c))}{N^\alpha} - E \left( \frac{f^N(X^N(t, c))}{N^\alpha} \right) \right)^2. \]

One can estimate the right hand side by using the same argument as is in Theorem 5.4 to get an upper bound \( 8L^2_K + 8C^2_K E \left( |V_N(s)|^2 \right), \) which is bounded in \( N \) uniformly on \([0, t]\) by Lemma 4.8.

Remark: Based on Theorem 5.1, Corollary 5.3, Corollary 5.5 and Theorem 5.6, we may expect the RSEs of the GT, CGT and FD methods to scale as \( \mathcal{O}(N^{1/2}), \mathcal{O}(1) \) and \( \mathcal{O}(N^{-1/2}) \), respectively. Since in Theorem 5.1, we do not have an exact limit for the sensitivity itself, this conclusion is not rigorously proven. Nevertheless, our numerical results in the next section support this expectation.

6. Numerical examples. We illustrate the dependence of RSE of various sensitivity estimators on the system size \( N \) via numerical examples. When comparing the GT or CGT methods with FD or RPD methods, we must bear in mind that while GT and CGT do not have method parameters, the FD method has a perturbation parameter \( h \) and the RPD method has a window size parameter \( w \), making the comparison not straightforward. A proper practical comparison involves choosing parameters \( h \) and \( w \) to obtain an acceptable bias. We do not pursue such a detailed comparison here as we are focused solely on the dependence on system size \( N \).

We note that in the very large system size limit, the stochastic system behaves nearly deterministically and hence none of these stochastic sensitivity methods are needed; traditional ODE sensitivity methods would do. However, when the system size \( N \) is modestly large, say \( N = 100 \), the system may not be approximated by the ODE and our asymptotic analysis will be relevant in this regime. Our numerical results below show this.

6.1. Numerical example 1. The reversible isomerization model consists of two species \( S_1 \) and \( S_2 \) and involves the following two reactions:
\[ S_1 \overset{c_1}{\rightleftharpoons} S_2, \quad S_2 \overset{c_2}{\rightleftharpoons} S_1. \]  
(6.1)

In the model with system size \( N \), the intensity functions for processes \( R_1^N \) and \( R_2^N \) are
\[ a_1^N(X^N(t), c) = c_1 X_1^N(t), \]
\[ a_2^N(X^N(t), c) = c_2 X_2^N(t), \]
respectively. The stoichiometric vectors are \( \nu_1 = [-1, 1]^T \) and \( \nu_2 = [1, -1]^T \).

In this example, the expectation of the population of species at a fixed time \( t \) can be computed analytically:
\[ E[X_1^N(t)] = X_1^N(0) + \frac{1 - e^{-(c_1 + c_2)t}}{c_1 + c_2} (c_2 X_2^N(0) - c_1 X_1^N(0)), \]  
(6.2)
Efficiency of the Girsanov transformation approach

\[ E[X_2^N(t)] = X_2^N(0) + \frac{1 - e^{-(c_1+c_2)t}}{c_1+c_2}(c_2X_2^N(0) - c_1X_1^N(0)), \]  

where \( X_1^N(0) \) and \( X_2^N(0) \) are assumed to be deterministic. One can compute the exact sensitivities by differentiating (6.2) and (6.3) with respect to parameters. In the numerical tests considered here, we choose parameters \( c_1 = 0.3 \) and \( c_2 = 0.2 \) and the initial population \( X_1^N(0) = N \) and \( X_2^N(0) = N \), where \( N \) is the system size parameter. We set the terminal time \( T = 10 \) and compute the sensitivity for various system size parameter \( N = 1, 2, 5, 10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 \). We use four different methods here, namely GT, CGT, CRN and RPD. We note that by CRN we mean the common random number (one-sided) finite difference method in conjunction with Gillespie’s SSA ([16]). The perturbation parameter for CRN method is \( h = 0.01 \) for parameter \( c_1 = 0.3 \) and the window size parameter \( w = 1.0 \) for terminal time \( T = 10 \). The number of trajectories for simulation is \( N_{tr} = 10^6 \) for each system size \( N \).

The first output function we consider here is \( f^N(x) = x_1 \) for all \( N \), that is, we compute the sensitivity of \( E(X_1^N(T)) \) with respect to parameter \( c_1 \). Obviously, conditions in Assumption 5 are satisfied with \( \alpha = 1 \) and \( f(x) = x_1 \). We examine the growth of sensitivity of \( E(X_1^N(T)) \) with respect to \( c_1 \) in terms of \( N \) using \( 10^6 \) independent trajectories. The computed sensitivity is shown in Fig 6.1(a) and Fig 6.1(b) shows the loglog plot of RSE of all four methods.

\[ f^N(x) = x_2^1 \] for all \( N \), By (3.7), \( f(x) = x_1^2 \) and \( \alpha = 2 \) in Assumption 5. See Figure 6.2 for sensitivity and RSE. The third output function is \( f^N(x) = \sin(x_1/N) \) and so \( f(x) = \sin x_1 \). It can be seen that for this case, \( \alpha = 0 \) in Assumption 5. Plot for the numerical result is shown in Figure 6.3.

The last output function we consider here is the indicator function \( f^N(x) = 1_{\{x_1 \leq x_2\}} \), which does not satisfy the conditions in our theorems since \( f = 1_{\{x_1 \leq x_2\}} \) is not continuously differentiable. However, numerical tests still show similar behavior as indicated by our theorems. Note that the sensitivity approaches zero as \( N \) increases to \( \infty \) and hence RSE is not well defined for large \( N \). Instead, we plot the estimator variance against \( N \) in Figure 6.4(b).

Finally, the Table 6.1 summarizes the rate of growth (as a power of \( N \)) of the numerically estimated RSE for the different estimators considered above. The numerical
results are in agreement with the theory.

<table>
<thead>
<tr>
<th></th>
<th>$R_1$</th>
<th>$R_2$</th>
<th>$R_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT</td>
<td>0.4992</td>
<td>0.4895</td>
<td>0.5724</td>
</tr>
<tr>
<td>CGT</td>
<td>-0.0004</td>
<td>-0.0008</td>
<td>0.0009</td>
</tr>
<tr>
<td>CRN</td>
<td>-0.5156</td>
<td>-0.5160</td>
<td>-0.5162</td>
</tr>
<tr>
<td>RPD</td>
<td>-0.5005</td>
<td>-0.5000</td>
<td>-0.5000</td>
</tr>
</tbody>
</table>

6.2. Numerical example 2. As a second numerical example, let us consider the decaying-dimerizing model described in (6.4).

$$S_1 \xrightarrow{c_1} \emptyset, \quad 2S_1 \xrightarrow{c_2} S_2, \quad S_2 \xrightarrow{c_3} 2S_1, \quad S_2 \xrightarrow{c_4} S_3,$$

(6.4)

The stoichiometric vectors are $\nu_1 = [-1, 0, 0]^T$, $\nu_2 = [-2, 1, 0]^T$, $\nu_3 = [2, -1, 0]^T$. 

Fig. 6.2. Sensitivity of $E(X_N^1(T))^2$ with respect to $c_1$ at final time $T = 10$ for reversible isomerization model.

Fig. 6.3. Sensitivity of $E(\sin(X_N^1(T)/N))$ with respect to $c_1$ at final time $T = 10$ for reversible isomerization model.
and \( \nu_4 = [0, -1, 1]^T \). We set the initial population to be \( X_1^N(0) = 10N, X_2^N(0) = 0, X_3^N(0) = 0 \). Using the stochastic mass action form \([3.1]\), the intensity for processes \( R_1^N, R_2^N, R_3^N \) and \( R_4^N \) are

\[
a_1^N(X^N(t), c) = c_1 X_1^N(t),
\]

\[
a_2^N(X^N(t), c) = \frac{c_2}{2N} X_1^N(t)(X_1^N(t) - 1),
\]

\[
a_3^N(X^N(t), c) = c_3 X_2^N(t),
\]

\[
a_4^N(X^N(t), c) = c_4 X_2^N(t).
\]

We set the parameters as follows, \( c_1 = 1.0, c_2 = 0.002, c_3 = 0.5 \) and \( c_4 = 0.04 \). Note that the intensity for the second reaction is not linear, hence an analytical formula for the sensitivity is not attainable. We test the sensitivity and RSE for \( \mathbb{E}[f^N(X_1^N)] \) with respect to \( c_1 \). For the CRN method, we use one-sided finite difference scheme and perturb the parameter \( c_1 \) by \( h = 0.01 \). Note that RPD is not applicable for this example since the firing of the first reaction will prevent the second reaction to happen when the population of \( S_1 \) is 1 (see \([20]\)), therefore we only examine the efficiency of GT, CGT and CRN here. For each system size \( N \), the number of trajectories we used for simulation is \( N_{\text{tr}} = 10^6 \). Plots of the sensitivity and RSE are shown in Figure 6.5, 6.6 and 6.7 for \( \mathbb{E}(X_1^N(T)), \mathbb{E}(X_1^N(T))^2 \) and \( \mathbb{E}(\sin(X_1^N(T)/N)) \), respectively. The rate of growth (as a power of \( N \)) of the numerically estimated RSE are summarized in Table 6.2.

7. Conclusion. In this paper, we suggested a centered GT method which reduces the estimator variance by centering the desired quantity at its expected value. The main contribution of this paper is the system size analysis of the sensitivity as well as the variances of various sensitivity estimators, namely GT, CGT and FD. If we use the relative standard error (RSE) as a measure of efficiency of an estimator,
then our analysis suggests that the RSE of GT, CGT and FD are $\mathcal{O}(N^{1/2})$, $\mathcal{O}(1)$
Table 6.2

Observed slopes (via regression) for the loglog plots for RSE for decaying-dimerizing model, that is, $R_1$, $R_2$ and $R_3$ are the observed asymptotic order of the estimator RSE (as a power of $N$) for $E(X^N(T))$, $E(X^N(T))^2$ and $E(sin(X^N(T)/N))$, respectively.

<table>
<thead>
<tr>
<th></th>
<th>$R_1$</th>
<th>$R_2$</th>
<th>$R_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT</td>
<td>0.4689</td>
<td>0.4100</td>
<td>0.4737</td>
</tr>
<tr>
<td>CGT</td>
<td>-0.0040</td>
<td>-0.0257</td>
<td>-0.0008</td>
</tr>
<tr>
<td>CRN</td>
<td>-0.6022</td>
<td>-0.6068</td>
<td>-0.6009</td>
</tr>
</tbody>
</table>

and $O(N^{-1/2})$, respectively. We note that even for a modestly large system size of $N = 100$ (where the system is still noisy), the FD method is expected to be 100 times more efficient than the GT method. The numerical examples provided also illustrate this point.

In many practical systems some species are present in small numbers while others are present in large numbers, the classical scaling studied here does not capture this. The more general scaling proposed in [4, 13] could be used to study the efficiency of various sensitivity estimators in order to provide further insight.

Appendix A. Weak convergence for joint distribution. We give a theorem regarding the weak convergence of joint distribution.

Lemma A.1. Suppose that $\alpha_n$ and $\beta_n$ are random variables taking values on a metric space $(S, \rho)$. If $\alpha_n \Rightarrow \alpha$ and $\rho(\alpha_n, \beta_n) \Rightarrow 0$, then $\beta_n \Rightarrow \alpha$.

Proof. See Theorem 3.1 in [3].

Theorem A.2. Let $X_n$ and $Y_n$ be $\mathbb{R}^m$ valued and $\mathbb{R}^k$ valued sequences of random variables on the same sample space. Suppose $X_n$ converges to $X$ in probability where $X$ is deterministic and $Y_n \Rightarrow Y$. Then $(X_n, Y_n) \Rightarrow (X, Y)$ in $\mathbb{R}^{m+k}$.

Proof. Let $x \in \mathbb{R}^m$ be such that $X = x$ almost surely. First we show that $(X_n, Y_n) \Rightarrow (X, Y)$. If $f : \mathbb{R}^{m+k} \rightarrow \mathbb{R}$ is bounded and continuous then so is $g : \mathbb{R}^k \rightarrow \mathbb{R}$ defined by $g(y) = f(x, y)$. Since $Y_n \Rightarrow Y$ we have that

$$E(g(Y_n)) = E(f(X, Y_n)) \Rightarrow E(g(Y)) = E(f(X, Y)).$$

Now $\|(X_n, Y_n) - (X, Y)\| = \|X_n - X\|$ and since $X_n \rightarrow X$ in probability, $\|X_n - X\| \rightarrow 0$ in probability (implies convergence in distribution). Thus by Lemma A.1 we have that $(X_n, Y_n) \Rightarrow (X, Y)$.

Appendix B. Martingale functional central limit theorem.

Theorem B.1. (See [22]) Let $M^N$ be a sequence of $\mathbb{R}^m$-valued local martingales. Suppose the expectation of the maximum jump in $M^N$ is asymptotically negligible, that is, for each $t > 0$

$$\lim_{N \rightarrow \infty} \mathbb{E}\{\sup_{s \leq t} |M^N(s) - M^N(s^-)|\} = 0,$$

and, for each pair $1 \leq i, j \leq m$ and each $t > 0$,

$$[M^N, M^N_j](t) \rightarrow c_{i,j}(t),$$

where the convergence is in probability and $C(t) = \{c_{i,j}(t)\}$ is deterministic and continuous matrix-valued function. Then $M^N \Rightarrow M$ in $D^m$, where $M$ is Gaussian with independent increments having mean vector and covariance matrix

$$\mathbb{E}\{M(t)\} = 0 \quad \text{and} \quad \mathbb{E}\{M(t)M(t)^T\} = C(t).$$
Appendix C. Differentiating inside an integral.

**Theorem C.1.** (See [3]) Suppose $G(c, \omega)$ is a random variable for each $c$ in some interval of the real line. Let $c_0$ be a specific value of $c$. Suppose the following hold:

1. For a set of $\omega$ with probability one, $G(c, \omega)$ is differentiable with respect to $c$ at $c = c_0$.
2. There exists an interval $(c_l, c_u)$ containing $c_0$ (independent of $\omega$) on which $G(c, \omega)$ is Lipschitz (in $c$) for a set of $\omega$ with probability one, with constant $K$ which may depend on $\omega$. That is, for any $c_1, c_2$ in the interval $(c_l, c_u)$, the following holds:
   \[ |G(c_1, \omega) - G(c_2, \omega)| \leq K(\omega)|c_1 - c_2|. \]
3. $E(K)$ is finite.
4. $E(|G(c, \omega)|)$ is finite for all $c$ in $(c_l, c_u)$.

Then the following holds:

\[
\left. \frac{d}{dc} \right|_{c=c_0} E(G(c)) = E \left( \left. \frac{d}{dc} \right|_{c=c_0} G(c) \right).
\]

**REFERENCES**


