*Proceedings of the 2020 Winter Simulation Conference*
*K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.*

# OPTIMALLY TUNING FINITE-DIFFERENCE ESTIMATORS

Haidong Li

Department of Industrial Engineering and
Management
Peking University
5 Yiheyuan Road
Beijing 100871, P. R. CHINA

Henry Lam

Department of Industrial Engineering and
Operations Research
Columbia University
500 W. 120th Street
New York, NY 10027 USA

## ABSTRACT

We consider stochastic gradient estimation when only noisy function evaluations are available. Central finite-difference scheme is a common method in this setting, which involves generating samples under perturbed inputs. Though it is widely known how to select the perturbation size to achieve the optimal order of the error, exactly achieving the optimal first-order error, which we call asymptotic optimality, is considered much more challenging and not attempted in practice. In this paper, we provide evidence that designing asymptotically optimal estimator is practically possible. In particular, we propose a new two-stage scheme that first estimates the required parameter in the perturbation size, followed by running finite-difference based on the estimated parameter in the first stage. Both theory and numerical experiments demonstrate the optimality of the proposed estimator and the robustness over conventional finite-difference schemes based on ad hoc tuning.

## 1 INTRODUCTION

In this paper, we consider finite-difference stochastic gradient estimation (e.g., Glasserman 2013, Asmussen and Glynn 2007, Fu 2006, L'Ecuyer 1991), commonly used when only noisy simulation observations are available to evaluate the function value or model output. In stochastic optimization, such setting is known as black-box or zeroth-order (Ghadimi and Lan 2013, Nesterov and Spokoiny 2017). Finite-difference estimators are in contrast to unbiased derivative estimators, which include the infinitesimal perturbation analysis (Ho et al. 1983, Heidelberger et al. 1988), the likelihood ratio or the score function method (Glynn 1990, Rubinstein 1986, Reiman and Weiss 1989), measure-valued or weak differentiation (Heidergott and Vázquez-Abad 2008, Heidergott et al. 2010), and other variants such as the generalized likelihood ratio method (Peng et al. 2018).

Finite-difference estimators consist of generating samples under perturbed input parameters. This perturbation size is chosen with consideration of controlling the bias and variance that contributes to the overall mean squared error (MSE). As the perturbation size increases, bias increases while variance decreases (and vice versa). To minimize the MSE, the perturbation size is determined by balancing the magnitudes of the two error sources. It is widely known (Zazanis and Suri 1993, Fox and Glynn 1989) that, for twice continuously differentiable functions, the optimal perturbation size of the central finite-difference (CFD) scheme turns out to be of order $n^{-1/6}$ which leads to an optimal MSE order $n^{-2/3}$, where $n$ refers to the number of differencing pairs in the simulation. On the other hand, in forward or backward finite-difference scheme, the optimal perturbation size is of order $n^{-1/4}$ and the corresponding optimal MSE order deteriorates to $n^{-1/2}$. Note that although the optimal order of error, in terms of $n$, is widely known and achievable, it is considered much more challenging to obtain an estimator that achieves the exact optimal first-order MSE (i.e., the "constant" in front of the $n$). This is because it requires additional model information such as higher-order derivatives and the noise variance, which are unknown a priori. Indeed, as far as we know, there have been no mainstream estimators that attempt to optimize the MSE at this level of accuracy.

In this paper, we provide some evidence that it is, in fact, practically possible to obtain finite-difference estimators that guarantee to exactly optimize the first-order MSE. We call such estimator *asymptotically optimal*. More precisely, we propose a new, conceptually simple, two-stage scheme to obtain such an estimator. This scheme first estimates the parameters required to tune the perturbation size that optimizes the first-order MSE (the "Estimation" phase). Then, at the second stage, we use the estimated parameters to obtain a nearly optimal perturbation size (the "Minimization" phase) and run standard finite-difference (we will focus on CFD due to its superiority over other variants, though our framework can be readily generalized). The key reason for the implementability of this approach is that the parameter estimation that allows a nearly optimal CFD only requires a loose accuracy, so that few samples need to be allocated to the Estimation phase. We materialize the above intuition and show that this two-stage scheme, which we call Estimation-Minimization Central Finite-Difference (EM-CFD), is nearly asymptotically optimal (where "nearly" means that asymptotic optimality can be achieved when the allocation to the Estimation phase is relatively negligible). Compared to conventional CFD using the right order of perturbation size, but with the underlying constant chosen in an ad hoc fashion, our estimator is more robust as it performs consistently close to an "oracle" CFD that assumes knowledge of unknown model parameters. We support our theory with empirical results.

The rest of the paper is organized as follows. In Section 2, we introduce the setting of CFD and the challenge in achieving asymptotic optimality. Section 3 proposes a two-stage scheme to optimize the performance of CFD. Sections 4 and 5 discuss the asymptotic properties of parameter estimation and the overall MSE, respectively. Section 6 presents numerical results, and Section 7 concludes the paper and outlines future directions.

## 2 SETTING AND MOTIVATION

In this paper, we focus our discussions on the single-dimensional case. Let $f(\cdot):\mathbb{R}\to\mathbb{R}$ be a performance measure of interest, where we have access to an unbiased estimate $\hat{f}(x)$ for any chosen $x\in\mathbb{R}$. In other words, $\hat{f}(\cdot)$ is a family of random variables indexed by $x$ such that $\mathbb{E}[\hat{f}(x)]=f(x)$ and $Var(\hat{f}(x))=\sigma^2(x)$ for any $x\in\mathbb{R}$. Suppose we do not apply common random numbers (CRNs) in generating $\hat{f}(x)$, and thus $\hat{f}(x)$'s are assumed to be independent across different points $x$. We would like to estimate the first-order derivative $f^{(1)}(x_0)$ where $x_0\in\mathbb{R}$ is the point of interest.

In estimating $f^{(1)}(x_0)$, the CFD scheme elicits the output

$$Y(\delta)=\frac{\hat{f}(x_0+\delta)-\hat{f}(x_0-\delta)}{2\delta},$$

where $\delta>0$ is the perturbation size. Suppose that $f(x)$ is thrice continuously differentiable with non-zero third-order derivative $f^{(3)}(x_0)$, we have as $\delta\to 0$

$$Y(\delta)=f^{(1)}(x_0)+(B\delta^2+o(\delta^2))+\frac{\epsilon(\delta)}{\delta},$$

where $B=f^{(3)}(x_0)/6$ and $\epsilon(\delta)\in\mathbb{R}$ is a random variable such that $\mathbb{E}[\epsilon(\delta)]=0$ and $Var(\epsilon(\delta))=\eta^2(\delta)$. Suppose we do not apply common random numbers (CRNs) in generating $\hat{f}(x_0+\delta)$ and $\hat{f}(x_0-\delta)$, and that $Var(\hat{f}(x_0\pm\delta))\to Var(\hat{f}(x_0))$ as $\delta\to 0$. Then $\eta^2(\delta)\to\sigma^2(x_0)/2$ as $\delta\to 0$. Given the capability to output independent runs of $Y_i(\delta),\ i=1,\dots,n$, the CFD scheme obtains an estimate of $f^{(1)}(x_0)$ by taking the sample average, i.e., $\widehat{\theta}=(1/n)\sum_{i=1}^n Y_i(\delta)$.

The MSE of $\widehat{\theta}$ can be expressed as

$$\begin{aligned}\text{MSE} &= \mathbb{E}[(\widehat{\theta} - f^{(1)}(x_0))^2] \\ &= (\mathbb{E}[\widehat{\theta} - f^{(1)}(x_0)])^2 + Var(\widehat{\theta} - f^{(1)}(x_0)) \\ &= (B + o(1))^2\delta^4 + (\sigma^2(x_0)/2 + o(1))/(n\delta^2),\end{aligned}$$

where $o(1)$ means a term that goes to zero as $\delta$ goes to zero. Consider choosing the optimal $\delta$, in relation to $n$, in order to minimize the MSE. First, it is well-known and easy to see that, in order to obtain the optimal order of the MSE, one should balance the squared bias term $B^2\delta^4$ and the variance term $(\sigma^2(x_0)/2)/(n\delta^2)$ to the same order in terms of $n$. This amounts to setting $\delta$ to be order $n^{-1/6}$ (since otherwise by perturbing the order of $\delta$ either one of the two terms would increase), which leads to an optimal MSE order $n^{-2/3}$.

Next, to get the optimal $\delta$ more precisely, note that the MSE depends on the priori unknown parameters $B$ and $\sigma^2(x_0)$. Note that by Holder's inequality

$$B^2\delta^4 + (\sigma^2(x_0)/2)/(n\delta^2) \geq 3\left(B^2\sigma^4(x_0)/(16n^2)\right)^{1/3}$$

and equality holds if $\delta = \left(\sigma^2(x_0)/(4nB^2)\right)^{1/6}$. That is, the MSE with an exactly optimal first-order term is

$$\text{MSE}_{\text{opt}} = 3\left(B^2\sigma^4(x_0)/(16n^2)\right)^{1/3}(1 + o(1)) \tag{1}$$

with the dominating order $n^{-2/3}$ and the frontal constant $3(B^2\sigma^4(x_0)/16)^{1/3}$. We call a scheme *asymptotically optimal* if it achieves (1).

Thus, though it is easy to find the optimal order of $\delta$ by setting $\delta = \alpha n^{-1/6}$ for some $\alpha > 0$ as in the conventional CFD scheme, exactly achieving the optimal first-order MSE, or asymptotic optimality, requires $\alpha = \left(\sigma^2(x_0)/(4B^2)\right)^{1/6}$ and hence knowing the model information $B$ and $\sigma^2(x_0)$. In the literature, extracting this information is often viewed as challenging, in the sense that doing so is not easier than estimating the gradient directly. Our main contribution is to propose a scheme that extracts and utilizes this information efficiently. Our key insight is that, in order to tune $\alpha$ close to optimal, it suffices to obtain only moderately accurate estimates of $B$ and $\sigma^2(x_0)$. Thus, one do not need to devote too much computational effort into this task, in order to achieve (1). This makes our procedure readily implementable.

## 3 OVERVIEW OF THE TWO-STAGE PROCEDURE

We propose a new two-stage scheme, which we call Estimation-Minimization Central Finite-Difference (EM-CFD), to obtain an asymptotically optimal estimator. The first stage is the Estimation stage (E-stage). At the E-stage, we allocate $n_1 = \lfloor \lambda n \rfloor,\ 0 < \lambda < 1$ samples, each with a possibly different perturbation size, to estimate $B$ and $\sigma^2(x_0)$. The detailed estimation method will be discussed in Section 4. The second stage is the Minimization stage (M-stage). Here, we plug the estimated parameters $\widehat{B}$ and $\widehat{\sigma}^2(x_0)$ into the optimal perturbation size $\delta = \left(\sigma^2(x_0)/(4nB^2)\right)^{1/6}$. Then allocate $n_2 = \lceil (1-\lambda)n \rceil$ samples and run standard CFD to estimate $f^{(1)}(x_0)$. The full implementation of EM-CFD is shown in Algorithm 1.

## 4 ESTIMATING MODEL PARAMETERS

We present in detail the E-stage, namely on the estimation of $B$ and $\sigma^2(x_0)$. We use linear regression. Since $B$ is related to the third order derivative of $f(x)$, we take the Taylor expansion of $f(x)$ to justify the estimation accuracy of $B$. Suppose that $f(x)$ is five-times continuously differentiable with non-zero finite $f^{(5)}(x_0)$ (the non-zero assumption can be dropped, but we put this to streamline our results), we have as $\delta \to 0$

$$\hat{f}(x_0 + \delta) - \hat{f}(x_0 - \delta) = f^{(1)}(x_0)2\delta + B2\delta^3 + (D2\delta^5 + o(\delta^5)) + \epsilon(\delta)$$

```
Algorithm 1: Estimation-Minimization Central Finite-Difference
    The total number of samples is $n$, the allocation ratio for the E-stage is $\lambda > 0$;
    E-stage: Allocate $n_1 = \lfloor \lambda n \rfloor$ samples whose perturbation sizes $\delta_i,\ i = 1, \ldots, n_1$, are i.i.d.
      generated from a distribution $\mathcal{P}$. Then estimate $B$ and $\sigma^2(x_0)$ by $\widehat{B}$ and $\widehat{\sigma}^2(x_0)$.
    M-stage: Allocate $n_2 = \lceil (1-\lambda)n \rceil$ samples with perturbation size $\delta = \left(\widehat{\sigma}^2(x_0)/(4n_2\widehat{B}^2)\right)^{1/6}$.
      Then estimate $f^{(1)}(x_0)$ by
        $\widehat{\theta}_{\text{EM}} = \frac{1}{n_2}\sum_{i=1}^{n_2} \frac{\hat{f}_i(x_0+\delta) - \hat{f}_i(x_0-\delta)}{2\delta}.$
    return the gradient estimate $\widehat{\theta}_{\text{EM}}$.
```

where $B = f^{(3)}(x_0)/6$, $D = f^{(5)}(x_0)/120$, and $\epsilon(\delta) \in \mathbb{R}$ is a random variable such that $\mathbb{E}[\epsilon(\delta)] = 0$ and $Var(\epsilon(\delta)) \to 2\sigma^2(x_0)$ as $\delta \to 0$. Let

$$\boldsymbol{y} = \left[\hat{f}(x_0+\delta_1) - \hat{f}(x_0-\delta_1), \ldots, \hat{f}(x_0+\delta_{n_1}) - \hat{f}(x_0-\delta_{n_1})\right]',$$

$$\boldsymbol{X} = \begin{bmatrix} 2\delta_1 & \cdots & 2\delta_{n_1} \\ 2\delta_1^3 & \cdots & 2\delta_{n_1}^3 \end{bmatrix}',$$

$$\boldsymbol{\beta} = \left[f^{(1)}(x_0), B\right]',$$

$$\boldsymbol{\mathcal{E}} = \left[D2\delta_1^5 + \epsilon(\delta_1), \ldots, D2\delta_{n_1}^5 + \epsilon(\delta_{n_1})\right]',$$

where $\delta_1, \ldots, \delta_{n_1}$ are the perturbation sizes of samples at the E-stage. Then, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\mathcal{E}}$, and we can estimate $\boldsymbol{\beta}$ given by $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Specifically, we have

$$\widehat{B} = \frac{\sum_{i=1}^{n_1}\delta_i^2\sum_{i=1}^{n_1}\delta_i^3 y_i - \sum_{i=1}^{n_1}\delta_i^4\sum_{i=1}^{n_1}\delta_i y_i}{2(\sum_{i=1}^{n_1}\delta_i^2\sum_{i=1}^{n_1}\delta_i^6 - (\sum_{i=1}^{n_1}\delta_i^4)^2)} \tag{2}$$

where $y_i$ represents the $i$-th component in $\boldsymbol{y}$.

Correspondingly, we have

$$\widehat{B} - B = \frac{\sum_{i=1}^{n_1}\delta_i^2\sum_{i=1}^{n_1}\delta_i^3\mathcal{E}_i - \sum_{i=1}^{n_1}\delta_i^4\sum_{i=1}^{n_1}\delta_i\mathcal{E}_i}{2(\sum_{i=1}^{n_1}\delta_i^2\sum_{i=1}^{n_1}\delta_i^6 - (\sum_{i=1}^{n_1}\delta_i^4)^2)}.$$

Without loss of generality, suppose $\delta_i$ is i.i.d. generated from a distribution with $\mathbb{E}[\delta_0] = 0$ and $Var[\delta_0] = \Sigma_0$, and $\Sigma_0 \to 0$ as $n_1 \to +\infty$. Additionally, any order moment of $\delta_i$ is assumed to be finite, which allows us to apply the Central Limit Theorem. Then we have

$$\sqrt{n_1}\left(\begin{bmatrix} (1/n_1)\sum_{i=1}^{n_1}\delta_i\mathcal{E}_i \\ (1/n_1)\sum_{i=1}^{n_1}\delta_i^3\mathcal{E}_i \\ (1/n_1)\sum_{i=1}^{n_1}\delta_i^2 \\ (1/n_1)\sum_{i=1}^{n_1}\delta_i^4 \\ (1/n_1)\sum_{i=1}^{n_1}\delta_i^6 \end{bmatrix} - \begin{bmatrix} 2D\mathbb{E}[\delta_0^6] \\ 2D\mathbb{E}[\delta_0^8] \\ \mathbb{E}[\delta_0^2] \\ \mathbb{E}[\delta_0^4] \\ \mathbb{E}[\delta_0^6] \end{bmatrix}\right) \xrightarrow[n_1 \to +\infty]{d}$$

$$N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4D^2Cov(\delta_0^6,\delta_0^6) + 2\sigma^2(x_0)\mathbb{E}[\delta_0^2] & 4D^2Cov(\delta_0^6,\delta_0^8) + 2\sigma^2(x_0)\mathbb{E}[\delta_0^4] & 2DCov(\delta_0^6,\delta_0^2) & 2DCov(\delta_0^6,\delta_0^4) & 2DCov(\delta_0^6,\delta_0^6) \\ 4D^2Cov(\delta_0^8,\delta_0^6) + 2\sigma^2(x_0)\mathbb{E}[\delta_0^4] & 4D^2Cov(\delta_0^8,\delta_0^8) + 2\sigma^2(x_0)\mathbb{E}[\delta_0^6] & 2DCov(\delta_0^8,\delta_0^2) & 2DCov(\delta_0^8,\delta_0^4) & 2DCov(\delta_0^8,\delta_0^6) \\ 2DCov(\delta_0^2,\delta_0^6) & 2DCov(\delta_0^2,\delta_0^8) & Cov(\delta_0^2,\delta_0^2) & Cov(\delta_0^2,\delta_0^4) & Cov(\delta_0^2,\delta_0^6) \\ 2DCov(\delta_0^4,\delta_0^6) & 2DCov(\delta_0^4,\delta_0^8) & Cov(\delta_0^4,\delta_0^2) & Cov(\delta_0^4,\delta_0^4) & Cov(\delta_0^4,\delta_0^6) \\ 2DCov(\delta_0^6,\delta_0^6) & 2DCov(\delta_0^6,\delta_0^8) & Cov(\delta_0^6,\delta_0^2) & Cov(\delta_0^6,\delta_0^4) & Cov(\delta_0^6,\delta_0^6) \end{bmatrix}\right).$$

With the multivariate delta method, we obtain

$$\sqrt{n_1}\left((\widehat{B}-B)-\frac{D(\mathbb{E}[\delta_0^2]\mathbb{E}[\delta_0^8]-\mathbb{E}[\delta_0^4]\mathbb{E}[\delta_0^6])}{\mathbb{E}[\delta_0^2]\mathbb{E}[\delta_0^6]-\mathbb{E}[\delta_0^4]\mathbb{E}[\delta_0^4]}\right)\xrightarrow[n_1\to+\infty]{d} N(0,\frac{\sigma^2(x_0)\mathbb{E}[\delta_0^2]}{2(\mathbb{E}[\delta_0^2]\mathbb{E}[\delta_0^6]-\mathbb{E}[\delta_0^4]\mathbb{E}[\delta_0^4])})$$

where we use $\Sigma_0 \to 0$ to argue the negligibility of higher-order terms.

The following theorem gives the expression for the MSE of $\widehat{B}$ when using $\delta_i$ that is randomly drawn from a normal distribution in an i.i.d. fashion. It shows that the MSE vanishes to zero as $n_1 \to +\infty$. In other words, the estimate $\widehat{B}$ is consistent.

**Theorem 1** Suppose $f(\cdot)$ is five-times continuously differentiable, and $\hat{f}(\cdot)$ has a finite second moment at any point with $Var(\hat{f}(x_0 \pm \delta)) \to Var(\hat{f}(x_0))$ as $\delta \to 0$. If we randomly draw $\delta_1,\ldots,\delta_{n_1}$ from $N(0,\Sigma_0)$, and consider $\widehat{B}$ in (2), we have

$$\mathbb{E}[(\widehat{B}-B)^2]=\left[(10D\Sigma_0)^2+\frac{\sigma^2(x_0)}{12n_1\Sigma_0^3}\right](1+o(1)),$$

where $o(1)$ means a term that goes to zero as $n_1 \to +\infty$. Moreover, for any $\Sigma_0 = \Theta(n_1^k)$, $-1/3 < k < 0$,

$$\lim_{n_1\to+\infty}\mathbb{E}[(\widehat{B}-B)^2]=0.$$

*Proof.* For $\delta_i \sim N(0,\Sigma_0),\ i=1,\ldots,n_1$,

$$\begin{aligned}\mathbb{E}[(\widehat{B}-B)^2] &= \left(\frac{D(1+o(1))(\mathbb{E}[\delta_0^2]\mathbb{E}[\delta_0^8]-\mathbb{E}[\delta_0^4]\mathbb{E}[\delta_0^6])}{\mathbb{E}[\delta_0^2]\mathbb{E}[\delta_0^6]-\mathbb{E}[\delta_0^4]\mathbb{E}[\delta_0^4]}\right)^2+\frac{\sigma^2(x_0)(1+o(1))\mathbb{E}[\delta_0^2]}{2n_1(\mathbb{E}[\delta_0^2]\mathbb{E}[\delta_0^6]-\mathbb{E}[\delta_0^4]\mathbb{E}[\delta_0^4])}\\ &= \left[(10D\Sigma_0)^2+\frac{\sigma^2(x_0)}{12n_1\Sigma_0^3}\right](1+o(1)).\end{aligned}$$

When $\Sigma_0 = \Theta(n_1^k)$, $-1/3 < k < 0$, we have $\lim_{n_1\to+\infty}(10D\Sigma_0)^2 = 0$ and $\lim_{n_1\to+\infty}\frac{\sigma^2(x_0)}{12n_1\Sigma_0^3} = 0$. Therefore, $\lim_{n_1\to+\infty}\mathbb{E}[(\widehat{B}-B)^2]=0$. □

Note that we can generalize the above to non-normal mean-zero $\delta_i$, though normal distribution is a natural choice. Moreover, we can show further that choosing $\Sigma_0 = \Theta(n_1^{-1/5})$ leads to an optimal order $n_1^{-2/5}$ of $\mathbb{E}[(\widehat{B}-B)^2]$. Therefore, at the E-stage, we randomly take perturbation size $\delta_i = \alpha_i n_1^{-1/10}$ where $\alpha_i$ is i.i.d. generated from a fixed normal distribution.

As for $\sigma^2(x_0)$, we use the sample variance of outputs $\left(\hat{f}(x_0+\delta_i)-\hat{f}(x_0-\delta_i)\right)$ to estimate it, i.e.,

$$\widehat{\sigma}^2(x_0)=\frac{1}{2}\frac{\|\boldsymbol{y}-\bar{y}\mathbb{1}_{n_1\times 1}\|_2^2}{n_1-1} \tag{3}$$

where $\bar{y}=\sum_{i=1}^{n_1}y_i/n_1$ and $\mathbb{1}_{n_1\times 1}$ is a column vector with all elements equal to one. The following theorem provides the consistency of $\widehat{\sigma}^2(x_0)$. This holds as long as $\Sigma_0 \to 0$ as $n_1 \to \infty$, and is not limited to the case that $\delta_i = \alpha_i n_1^{-1/10}$.

**Theorem 2** Suppose $\hat{f}(\cdot)$ has a finite second moment at any point with $Var(\hat{f}(x_0 \pm \delta)) \to Var(\hat{f}(x_0))$ as $\delta \to 0$. Suppose we randomly draw $\delta_1,\ldots,\delta_{n_1}$ from $N(0,\Sigma_0)$, with $\Sigma_0 \to 0$ as $n_1 \to \infty$. Consider $\widehat{\sigma}^2(x_0)$ in (3). We have

$$\lim_{n_1\to+\infty}\widehat{\sigma}^2(x_0)=\sigma^2(x_0),\ a.s.$$

*Proof.* This follows quite immediately from the consistency of sample variance. To give more details, since $\Sigma_0 \to 0$ as $n_1 \to +\infty$, we have $\lim\limits_{n_1 \to +\infty} \delta_i = 0,\ a.s.$ for any $i$. For each $n_1 \in \mathbb{N}^+$,

$$\widehat{\sigma}^2(x_0) = \frac{n_1}{2(n_1-1)} \left( \frac{1}{n_1} \sum_{i=1}^{n_1} y_i^2 - (\frac{1}{n_1} \sum_{i=1}^{n_1} y_i)^2 \right).$$

Combining $\lim\limits_{n_1 \to +\infty} \delta_i = 0,\ a.s.$ and the strong Law of Large Numbers, we have

$$\lim_{n_1 \to +\infty} \frac{1}{n_1} \sum_{i=1}^{n_1} y_i^2 = \mathbb{E}[\epsilon^2(0)],\ a.s.$$

and

$$\lim_{n_1 \to +\infty} \frac{1}{n_1} \sum_{i=1}^{n_1} y_i = \mathbb{E}[\epsilon(0)],\ a.s.$$

Therefore,

$$\lim_{n_1 \to +\infty} \widehat{\sigma}^2(x_0) = \frac{1}{2} \left( \mathbb{E}[\epsilon^2(0)] - (\mathbb{E}[\epsilon(0)])^2 \right) = \frac{1}{2} Var(\epsilon(0)) = \sigma^2(x_0),\ a.s.$$

□

## 5 OPTIMIZING FINITE-DIFFERENCE PERFORMANCE

At the E-stage, we estimate unknown model parameters $B$ and $\sigma^2(x_0)$ by (2) and (3). Then, at the M-stage, we use these parameter estimates to obtain a nearly optimal perturbation size and estimate $f^{(1)}(x_0)$ by standard CFD. The error in parameter estimation will propagate to the MSE of CFD. The following theorem provides the expression for the MSE of the resulting gradient estimator from EM-CFD. It also shows that this MSE closely matches the optimal in the first order, in turn implying that EM-CFD is nearly asymptotically optimal, when the allocation assigned to the E-stage is relatively small.

**Theorem 3** Suppose $f(\cdot)$ is five-times continuously differentiable, and $\hat{f}(\cdot)$ has a finite second moment at any point with $Var(\hat{f}(x_0 \pm \delta)) \to Var(\hat{f}(x_0))$ as $\delta \to 0$. Suppose we randomly draw $\delta_1, \ldots, \delta_{n_1}$ from $N(0, \Sigma_0)$, with $\Sigma_0 = \Theta(n_1^k)$, $-1/3 < k < 0$. Consider EM-CFD in Algorithm 1. The MSE of $\widehat{\theta}_{\text{EM}}$ is

$$\begin{aligned} \text{MSE}_{\text{EM}} &= \mathbb{E}[(\widehat{\theta}_{\text{EM}} - f^{(1)}(x_0))^2] \\ &= \left( \left( \mathbb{E}\left[ B \left( \widehat{\sigma}^2(x_0)/(4\widehat{B}^2) \right)^{1/3} n_2^{-1/3} \right] \right)^2 + \mathbb{E}\left[ \frac{\sigma^2(x_0)}{2n_2^{2/3} \left( \widehat{\sigma}^2(x_0)/(4\widehat{B}^2) \right)^{1/3}} \right] \right. \\ &\quad \left. + Var\left( B \left( \widehat{\sigma}^2(x_0)/(4\widehat{B}^2) \right)^{1/3} n_2^{-1/3} \right) \right)(1+o(1)), \end{aligned}$$

which is a function of $B$ and $\sigma^2(x_0)$. Moreover,

$$\lim_{\lambda \to 0} \lim_{n \to +\infty} n^{2/3} \left| \text{MSE}_{\text{EM}} - \text{MSE}_{\text{opt}} \right| = 0, \tag{4}$$

where $\text{MSE}_{\text{opt}}$ is defined in (1).

*Proof.* The M-stage uses the standard CFD scheme, and thus conditional on any given $\delta$ we have

$$\sqrt{n_2}\left(\widehat{\theta}_{\text{EM}} - (f^{(1)}(x_0) + B\delta^2)\right) \xrightarrow[n_2\to+\infty]{d} N\left(0, \frac{\sigma^2(x_0)}{2\delta^2}\right).$$

We use $\delta = \left(\widehat{\sigma}^2(x_0)/(4\widehat{B}^2)\right)^{1/6} n_2^{-1/6}$ where both $\widehat{B}$ and $\widehat{\sigma}^2(x_0)$ are random variables estimated from the E-stage. Therefore, for the overall EM-CFD scheme, we have

$$\mathbb{E}[\widehat{\theta}_{\text{EM}}] = \mathbb{E}\left[\mathbb{E}[\widehat{\theta}_{\text{EM}}|\widehat{B}, \widehat{\sigma}^2(x_0)]\right] = f^{(1)}(x_0) + \mathbb{E}\left[B\left(\widehat{\sigma}^2(x_0)/(4\widehat{B}^2)\right)^{1/3} n_2^{-1/3} + o(\delta^2)\right]$$

and

$$\begin{aligned}
& Var(\widehat{\theta}_{\text{EM}}) \\
= \; & \mathbb{E}\left[Var(\widehat{\theta}_{\text{EM}}|\widehat{B}, \widehat{\sigma}^2(x_0))\right] + Var\left(\mathbb{E}[\widehat{\theta}_{\text{EM}}|\widehat{B}, \widehat{\sigma}^2(x_0)]\right) \\
= \; & \mathbb{E}\left[\frac{\sigma^2(x_0)}{2n_2^{2/3}\left(\widehat{\sigma}^2(x_0)/(4\widehat{B}^2)\right)^{1/3}} + o(\delta^4)\right] + Var\left(B\left(\widehat{\sigma}^2(x_0)/(4\widehat{B}^2)\right)^{1/3} n_2^{-1/3} + o(\delta^2)\right).
\end{aligned}$$

Here $|o(\delta^2)/\delta^2| \to 0 \; a.s.$ and $|o(\delta^4)/\delta^4| \to 0 \; a.s.$ when $\delta \to 0 \; a.s.$ Then we obtain

$$\begin{aligned}
\text{MSE}_{\text{EM}} &= \mathbb{E}[(\widehat{\theta}_{\text{EM}} - f^{(1)}(x_0))^2] \\
&= \left(\mathbb{E}[\widehat{\theta}_{\text{EM}} - f^{(1)}(x_0)]\right)^2 + Var\left(\widehat{\theta}_{\text{EM}} - f^{(1)}(x_0)\right) \\
&= \left(\mathbb{E}\left[B\left(\widehat{\sigma}^2(x_0)/(4\widehat{B}^2)\right)^{1/3} n_2^{-1/3} + o(\delta^2)\right]\right)^2 + \mathbb{E}\left[\frac{\sigma^2(x_0)}{2n_2^{2/3}\left(\widehat{\sigma}^2(x_0)/(4\widehat{B}^2)\right)^{1/3}} + o(\delta^4)\right] \\
&\quad + Var\left(B\left(\widehat{\sigma}^2(x_0)/(4\widehat{B}^2)\right)^{1/3} n_2^{-1/3} + o(\delta^2)\right).
\end{aligned}$$

We have $\lim\limits_{n_2\to+\infty} \mathbb{E}(o(\delta^2)/\delta^2) = 0$, $\lim\limits_{n_2\to+\infty} \mathbb{E}(o(\delta^4)/\delta^4) = 0$, and $\lim\limits_{n_2\to+\infty} Var(o(\delta^2)/\delta^2) = 0$. With the results in Theorem 1 and 2, we have, asymptotically, $\text{MSE}_{\text{EM}}$ satisfies

$$\lim_{n\to+\infty} n^{2/3}\left|\text{MSE}_{\text{EM}} - 3\left(\frac{\sigma^2(x_0)B}{4n}\right)^{2/3}\right| = 3\left(\frac{\sigma^2(x_0)B}{4(1-\lambda)}\right)^{2/3} - 3\left(\frac{\sigma^2(x_0)B}{4}\right)^{2/3}.$$

Further,

$$\lim_{\lambda\to 0}\lim_{n\to+\infty} n^{2/3}\left|\text{MSE}_{\text{EM}} - 3\left(\frac{\sigma^2(x_0)B}{4n}\right)^{2/3}\right| = 0.$$

Recall that $\text{MSE}_{\text{opt}} = 3\left(\frac{\sigma^2(x_0)B}{4n}\right)^{2/3}(1 + o(1))$, which concludes the theorem. □

## 6 NUMERICAL RESULTS

In this section, we conduct numerical experiments to test the performance of the proposed EM-CFD scheme. We consider three examples. Example 1 aims to estimate the first-order derivative of $f(x) = k\sin(x)$ at the point $x_0 = 0$, where $k$ is an a priori unknown parameter. Observations at $\delta$ follow a normal distribution, i.e., $\hat{f}(\delta) \sim N(f(\delta), 1)$. Therefore in this example, we have $B = -k/6$ and $\sigma^2(x_0) = 1$ but they are unknown for EM-CFD scheme. Example 2 concerns the derivative estimation of a polynomial function $\hat{f}(x) = 1 - 6x + 36x^2 - 53x^3 + 22x^5 + N(0, 0.05)$ at different points $x$. We have $B = 220x^2 - 53$ and $\sigma^2(x) = 0.05$ accordingly. Example 3 considers a generic M/M/1 queueing system. Observations are the averaged system time of the first 10 customers, and the derivative of their expectation with respect to the arrival rate is of interest to us. The true derivative is 0.0946 when both arrival and service rates are 4.

The proposed EM-CFD scheme is compared with an "oracle" CFD that uses the optimal perturbation size assuming knowledge of $B$ and $\sigma^2(x_0)$ (CFD-opt), and the conventional CFD scheme that uses an arbitrary $\alpha$:

- CFD-opt: Both $B$ and $\sigma^2(x_0)$ are known for this scheme. Then it outputs $n$ independent runs of $Y_i(\delta) = (\hat{f}(x_0 + \delta) - \hat{f}(x_0 - \delta))/(2\delta)$ where $\delta = \left(\sigma^2(x_0)/(4B^2)\right)^{1/6} n^{-1/6}$, and estimates $f^{(1)}(x_0)$ by $\widehat{\theta} = (1/n)\sum_{i=1}^{n} Y_i(\delta)$.
- CFD: This scheme uses $\alpha n^{-1/6}$ for some arbitrarily chosen $\alpha$. We do it by arbitrarily choosing some value ($\widetilde{B}$ and $\widetilde{\sigma}^2(x_0)$) for each unknown parameter, and set the perturbation size $\delta = \left(\widetilde{\sigma}^2(x_0)/(4\widetilde{B}^2)\right)^{1/6} n^{-1/6}$. We take $\left(\widetilde{B}, \widetilde{\sigma}^2(x_0)\right) = (-100/6, 1)$ in example 1, $\left(\widetilde{B}, \widetilde{\sigma}^2(x_0)\right) = (5, 0.05)$ in example 2 while $\left(\widetilde{B}, \widetilde{\sigma}^2(x_0)\right) = (1/2, 1)$ in example 3. For each example, we output $n$ independent runs of $Y_i(\delta)$, and estimate $f^{(1)}(x_0)$ by $\widehat{\theta} = (1/n)\sum_{i=1}^{n} Y_i(\delta)$.

Additionally, in order to examine the performance of the proposed EM-CFD scheme under different allocation ratio $\lambda$, three settings are tested, namely $\lambda = 0.1$, $\lambda = 0.2$, and $\lambda = 0.3$. In all numerical experiments, the statistical efficiency of the estimation schemes is measured by the MSE estimated by 10,000 independent experimental replications. The MSE is reported as a function of the (unknown) model parameter $k$, the value of $x$, or the number of samples $n$ in each experiment.

In Figure 1, we can see that EM-CFD performs better than the conventional CFD when the model parameter $k$ is less than 60, which could be attributed to the reason that parameter estimation ($\widehat{B}$ and $\widehat{\sigma}^2(x_0)$) in EM-CFD is much closer to optimal than the arbitrary choice in CFD. As the model parameter $k$ increases, the arbitrary choice in CFD approaches the optimal one and is equal to the true value when $k = 100$. Of course, a priori we do not know this $k$, so that the arbitrary CFD can perform well or can perform poorly. In other words, EM-CFD is more robust. Moreover, the performance of EM-CFD with $\lambda = 0.1$ is still comparable to that of CFD-opt, showing that EM-CFD is close to having the best possible performance using any CFD for the whole considered range of $k$. Another observation is that the performance of EM-CFD gets better as the allocation ratio $\lambda$ decreases. Although allocating more samples in the E-stage can lead to a better parameter estimation, it would limit the sample size to estimate the gradient, ultimately resulting in a worse MSE. Such observation is in accord with the asymptotic property in (4).

Figure 2 further examines the performance of EM-CFD scheme with lower allocation ratio $\lambda$. Similar to results in Figure 1, EM-CFD remains more efficient than the conventional CFD, and CFD-opt is slightly better than EM-CFD. However, given a finite number of samples, the performance of EM-CFD becomes worse when decreasing the allocation ratio $\lambda$ from 0.01 to 0.005. It could be because the number of samples allocated to estimate the unknown parameters is too few. In other words, we need enough samples to guarantee the consistency of parameter estimation. The best balance in obtaining best overall performance seems to be obtained at $\lambda = 0.01$.

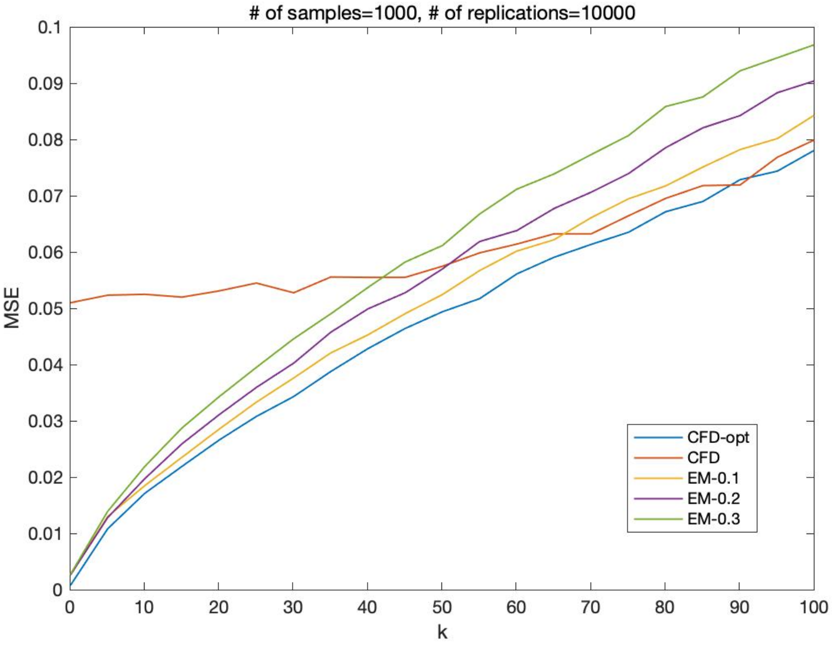

Figure 1: MSE of all tested CFD schemes in example 1.

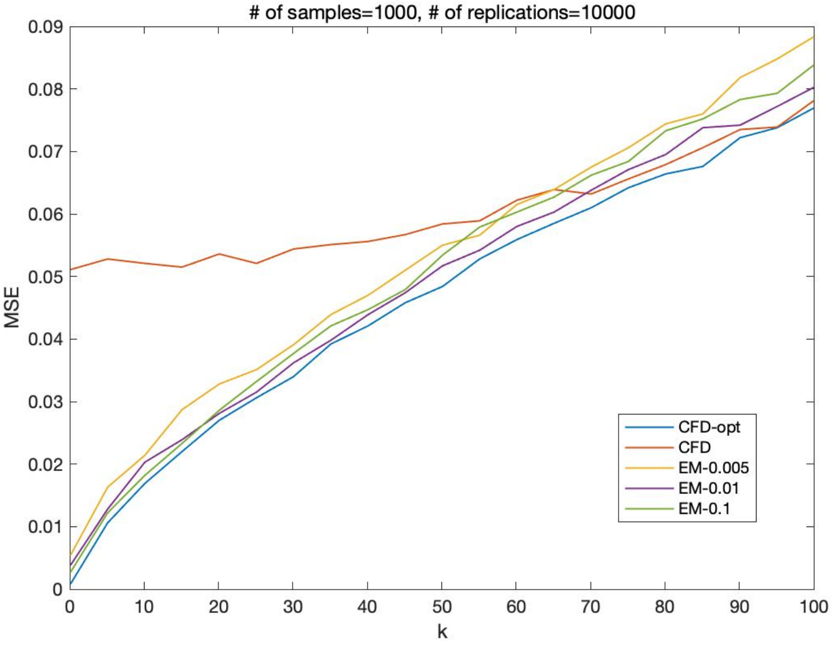

Figure 2: MSE of all tested CFD schemes in example 1.

Figure 3 also demonstrates the robustness of EM-CFD. The MSE of EM-CFD at $x \in [0, 1]$ is generally below $0.5$, and EM-CFD performs better than the conventional CFD when $x < 0.3$ or $x > 0.6$. In contrast, the conventional CFD can perform well when its arbitrary choice of $B$ is close to the true value, i.e., $x$ is around $0.5$. Unfortunately, the conventional CFD performs poorly at other times. At point $x = 1$, the MSE of the conventional CFD is greater than $2.5$, which is five times as large as that of EM-CFD.

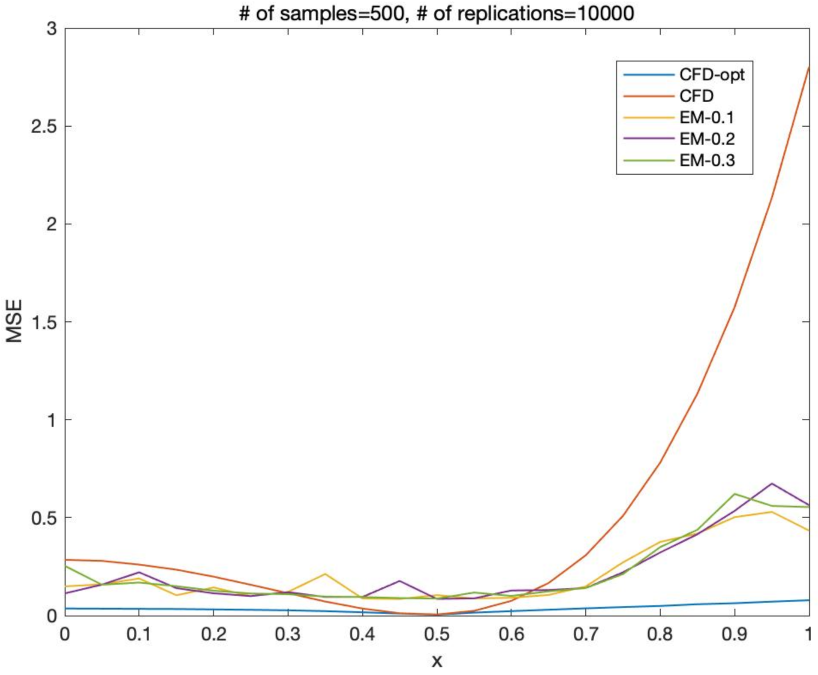

Figure 3: MSE of all tested CFD schemes in example 2.

Figure 4 illustrates the performance of CFD and EM-CFD in example 3. When $n = 1000$, we can see that the proposed EM-CFD achieves MSE $= 3.8 \times 10^{-4}$ whereas the conventional CFD scheme leads to MSE $= 7.7 \times 10^{-4}$. That is to say EM-CFD enhances the estimation efficiency by more than 50% compared to this arbitrarily tuned CFD. In addition, note that unlike examples 1 and 2, simulation variance $\sigma^2(x)$ varies with different points $x$ in this example, which can be viewed as a more challenging setting. The favorable results in Figure 4 thus demonstrate potential of EM-CFD to apply in more sophisticated and practical settings.

## 7 CONCLUSION

While it is well-known how to find perturbation sizes with the optimal order for finite-difference estimators in gradient estimation when only noisy function evaluations are available, finding the perturbation size that matches the exact first-order MSE, or what we called asymptotic optimality, is generally considered much more challenging as it relies on unknown model information. In this paper we derived and implemented an asymptotically optimal CFD, henceforth providing evidence that designing such estimators is practically possible. This estimator comprises a two-stage scheme, called EM-CFD, that first estimates the needed but unknown model parameters and then, based on these estimates, chooses a nearly optimal perturbation size. Its implementability hinges on the main insight that these model parameter estimates only need moderate accuracy in order to achieve optimality, thus allowing us to allocate few samples in the estimation stage. Besides theoretically proving near asymptotic optimality, we conducted some numerical studies to

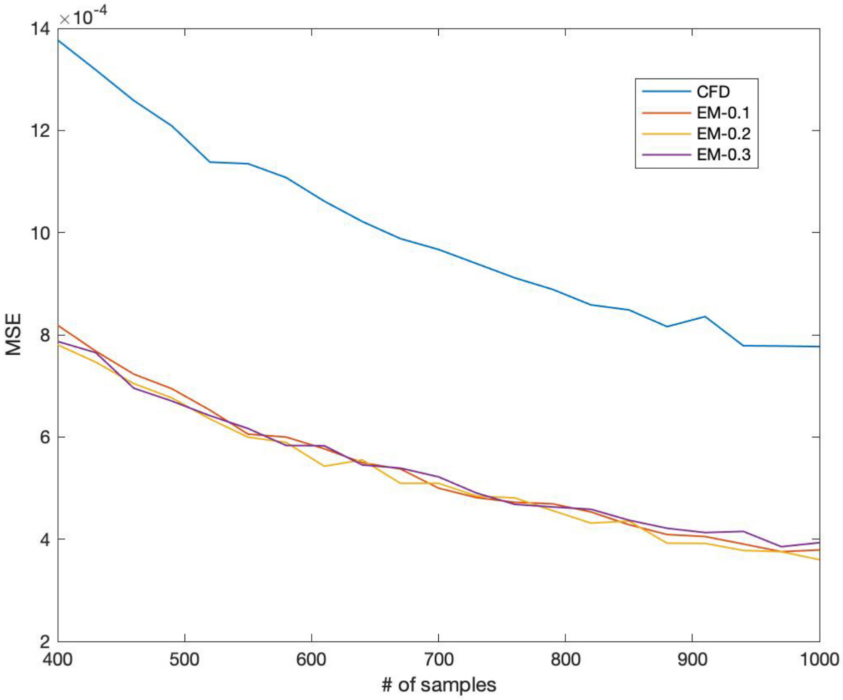

Figure 4: MSE of CFD and EM in example 3.

demonstrate that the proposed estimator works closely compared to an oracle benchmark and more robustly compared to CFD with arbitrarily chosen constant in the perturbation size despite the right order. In future work, we will investigate the multi-dimensional generalizations, investigate the higher-order convergence rate of the MSE that allows us to further enhance the choice of perturbation size, and conduct more extensive numerical studies.

## ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under CAREER CMMI-1834710 and IIS-1849280, and by the China Scholarship Council (CSC) that supports Haidong's visit at Columbia University.

## REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic simulation: algorithms and analysis*, Volume 57. Springer Science & Business Media.

Fox, B. L., and P. W. Glynn. 1989. "Replication schemes for limiting expectations". *Probability in the Engineering and Informational Sciences* 3(3):299–318.

Fu, M. C. 2006. "Gradient estimation". *Handbooks in operations research and management science* 13:575–616.

Ghadimi, S., and G. Lan. 2013. "Stochastic first-and zeroth-order methods for nonconvex stochastic programming". *SIAM Journal on Optimization* 23(4):2341–2368.

Glasserman, P. 2013. *Monte Carlo methods in financial engineering*, Volume 53. Springer Science & Business Media.

Glynn, P. W. 1990. "Likelihood ratio gradient estimation for stochastic systems". *Communications of the ACM* 33(10):75–84.

Heidelberger, P., X.-R. Cao, M. A. Zazanis, and R. Suri. 1988. “Convergence properties of infinitesimal perturbation analysis estimates”. *Management Science* 34(11):1281–1302.

Heidergott, B., and F. Vázquez-Abad. 2008. “Measure-valued differentiation for Markov chains”. *Journal of Optimization Theory and Applications* 136(2):187–209.

Heidergott, B., F. J. Vázquez-Abad, G. Pflug, and T. Farenhorst-Yuan. 2010. “Gradient estimation for discrete-event systems by measure-valued differentiation”. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 20(1):1–28.

Ho, Y.-C., X. Cao, and C. Cassandras. 1983. “Infinitesimal and finite perturbation analysis for queueing networks”. *Automatica* 19(4):439–445.

L’Ecuyer, P. 1991. “An overview of derivative estimation”. In *Winter Simulation Conference*, 207–217.

Nesterov, Y., and V. Spokoiny. 2017. “Random gradient-free minimization of convex functions”. *Foundations of Computational Mathematics* 17(2):527–566.

Peng, Y., M. C. Fu, J.-Q. Hu, and B. Heidergott. 2018. “A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters”. *Operations Research* 66(2):487–499.

Reiman, M. I., and A. Weiss. 1989. “Sensitivity analysis for simulations via likelihood ratios”. *Operations Research* 37(5):830–844.

Rubinstein, R. Y. 1986. “The score function approach for sensitivity analysis of computer simulation models”. *Mathematics and Computers in Simulation* 28(5):351–379.

Zazanis, M. A., and R. Suri. 1993. “Convergence rates of finite-difference sensitivity estimates for stochastic systems”. *Operations research* 41(4):694–703.

## AUTHOR BIOGRAPHIES

**HAIDONG LI** is a Ph.D. candidate in the Department of Industrial Engineering and Management, Peking University, Beijing, China. He received his B.S. Degree from the Department of Engineering Mechanics at Peking University. His research interests include simulation optimization and network analysis. He is a visiting scholar under the supervision of Prof. Henry Lam during 2019-2020. His email address is haidong.li@pku.edu.cn.

**HENRY LAM** is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. He received his Ph.D. degree in statistics from Harvard University in 2011, and was on the faculty of Boston University and the University of Michigan before joining Columbia in 2017. His research focuses on Monte Carlo simulation, uncertainty quantification, risk analysis, and stochastic and robust optimization. His email address is henry.lam@columbia.edu.