Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

MINIMAX EFFICIENT FINITE-DIFFERENCE GRADIENT ESTIMATORS

Henry Lam

Department of IEOR
Columbia University
500 W. 120th Street
New York, NY 10027 USA

Xuhui Zhang

School for the Gifted Young
University of Science and Technology of China
96 Jinzhai Road
Hefei, Anhui 230026 CHINA

ABSTRACT

We consider stochastic gradient estimation when only noisy measurements of the function are available. A standard approach is to use the finite difference method or its variants. While natural, it is open to our knowledge whether its statistical accuracy is the best possible. This paper argues that this is nearly so in a well-defined minimax sense. In particular, we show that central finite-difference is an optimal zeroth-order derivative estimator among the class of linear estimators, over the worst-case scenario among a wide class of twice differentiable functions. This optimality is achieved exactly for any finite sample in the single-dimensional case, and nearly (up to a multiplicative factor) in the multi-dimensional case. We also show that the same estimator is nearly optimal among the much larger class of all (nonlinear) estimators. We utilize elementary techniques for the linear minimax results and Le Cam's method for the general minimax counterparts.

1 INTRODUCTION

Stochastic gradient estimation is of central importance in simulation analysis and optimization. It concerns the estimation of gradients under noisy environments driven by data or Monte Carlo simulation runs. This problem arises as a key ingredient in sensitivity analysis and uncertainty quantification for simulation models, descent-based algorithms in stochastic optimization and machine learning, and other applications such as financial portfolio hedging. For an overview of stochastic gradient estimation and its applications, see, e.g., L'Ecuyer (1991), Fu (2006), Glasserman (2013) Chapter 7 and Asmussen and Glynn (2007) Chapter 7.

In this paper, we consider stochastic gradient estimation when only a noisy simulation oracle to evaluate the function value or model output is available. This corresponds to black-box settings in which it is costly, or even impossible, to utilize the underlying dynamics of a simulation model, or to distort the data collection mechanism in an experiment given the input. In stochastic optimization, such an oracle is also known as the zeroth-order (Ghadimi and Lan 2013; Nesterov and Spokoiny 2017). These gradient estimators are in contrast to unbiased estimators obtained from methods such as the infinitesimal perturbation analysis (Heidelberger et al. 1988; Ho et al. 1983), the likelihood ratio or score function method (Glynn 1990; Rubinstein 1986; Reiman and Weiss 1989), measure-valued differentiation (Heidergott et al. 2010) and other variants that require structural information on model dynamics.

Under the above setting, the most natural and common approach is to use the finite-difference (FD) method (Zazanis and Suri 1993; Fox and Glynn 1989; Frolov and Chentsov 1963). This entails simulating the function values at two neighboring inputs and using the first principle to approximate the derivative. The resulting estimator has a bias coming from this derivative approximation, on top of the variance coming from the function evaluation noise. This leads to a subcanonical overall mean squared error (MSE) and a need to rightly tune the perturbation size between the two input values. It is widely known that, for

twice continuously differentiable functions, the optimal attainable MSE for central finite-difference (CFD) schemes is of order $O(n^{-2/3})$, where n refers to the number of differencing pairs in the simulation. On the other hand, when one uses forward or backward finite-differences, the optimal MSE deteriorates to $O(n^{-1/2})$.

Although the optimal MSEs *within* the classes of FD schemes are well-known, a question arises whether these classes are optimal or better compared to other, possibly much larger, classes of gradient estimators. Our goal in this paper is to give a first such study on the optimality on a class level.

Our main results show that, under a general setting, CFD is nearly optimal among any possible gradient estimation schemes. This optimality is in a minimax sense. Namely, we consider the MSE of any gradient estimator over a collection of twice differentiable functions with unknown function characteristics (e.g., function value and higher-order gradients). Subject to this uncertainty of the function, we consider the minimizer of the worst-case MSE over this function collection, giving rise to what we call the minimax risk. Among the class of linear estimators, we show that, in the one-dimensional case, CFD exactly achieves the minimax risk, whereas in multi-dimensional case it achieves so up to a multiplicative factor that depends sublinearly on the input dimension. Furthermore, we show that, among the much larger class of all nonlinear estimators, CFD remains to be nearly minimax optimal up to a multiplicative factor that is polynomial in the dimension.

In terms of methodological contributions, we derive our linear minimax results by using a new elementary proof. We derive our general minimax results via Le Cam's method (Tsybakov 2009) with the imposition of an adversarially chosen hypothesis test. Lastly, we also demonstrate that, without extra knowledge on the gradient, randomized schemes such as simultaneous perturbation will lead to unbounded worst-case risks in general, due to the interaction between the gradient magnitude and the additional variance coming from the random perturbation. This indicates that less conservative analyses along this line would require more information on the magnitude of the gradient of interest.

Our work is related to, and contrasted with, the derivative estimation in nonparametric regression (e.g., Fan 1993; Fan et al. 1997), which focuses on the estimation of the conditional expectations and their derivatives given input values, a similar setting as ours. However, in these studies the data and in particular the available input values are often assumed given a priori. In contrast, in stochastic gradient estimation, one often has the capability to select the input points at which the function evaluation is conducted. This therefore endows more flexibility than nonparametric regression and, correspondingly, translates to superior minimax rates in our setting. For example, Fan et al. (1997) has established a minimax risk of order $O(n^{-4/7})$ for nonparametric derivative estimation, which is slower than our $O(n^{-2/3})$. Finally, we note that other works (Dai et al. 2016; De Brabanter et al. 2013; Wang and Lin 2015) have studied derivative estimation uniformly well over regular or equi-spaced input design points. Moreover, these papers consider asymptotic risks as the number of input points grow, in contrast to the finite-sample results in this work.

The remainder of the paper is as follows. Section 2 focuses on linear minimax risk and the corresponding optimal or nearly optimal estimators. Section 3 focuses on general risks and estimators. Section 4 concludes the paper and discusses future directions. Due to space limit, some results are shown without proofs. Additional results and full proofs will be presented in a journal version of this work.

2 LINEAR MINIMAX RISK AND OPTIMAL ESTIMATORS

In this section we focus on the class of linear estimators. Section 2.1 first presents the single-dimensional case, whereas Section 2.2 generalizes it to higher-dimensional counterparts. In both cases, we will derive the approximate minimax risk and show that CFD is a nearly optimal estimator. Furthermore, in the single-dimensional case, these results can be made exact for any finite sample.

2.1 Single-Dimensional Case

We first introduce our setting. Let $f(\cdot) : \mathbb{R} \to \mathbb{R}$ be a performance measure of interest, where we have access to an unbiased estimate $Y(x)$ for any chosen $x \in \mathbb{R}$. In other words, $Y(\cdot)$ is a family of random variables indexed by x such that $E[Y(x)] = f(x)$ and $Var(Y(x)) = \sigma^2(x)$ for any $x \in \mathbb{R}$. Suppose x_0 is the point of interest. Our goal is to estimate the derivative $f^{'}(x_0)$.

Given simulation budget $n \geq 1$, we can simulate independently at input design points $x_0 + \delta_j$, $j = 1, \cdots, n$, with δ_j of our choice, giving outputs $Y_j(x_0 + \delta_j)$'s. We consider the class of linear estimators in the form

$$L_n = \sum_{j=1}^{n} w_j Y_j(x_0 + \delta_j), \tag{1}$$

where w_j are the linear coefficients or weights. Note that for even budget n the CFD scheme

$$\bar{L}_n = \frac{1}{n/2} \sum_{j=1}^{n/2} \frac{Y_{2j-1}(x_0 + \delta) - Y_{2j}(x_0 - \delta)}{2\delta} \tag{2}$$

is an example of linear estimators where $\delta_j = (-1)^{j+1}\delta$ and $w_j = \frac{(-1)^{j+1}}{n\delta}$, for a perturbation size δ.

We aim to study the optimality within the class of all linear estimators in the form (1), and in particular investigate the role of CFD. We use the MSE as a performance criterion, which depends on the a priori unknown function f. Our premise is a minimax framework that seeks for schemes to minimize the worst-case MSE, among a suitable wide enough class of function f and simulation noise. More precisely, we consider

$$\begin{aligned} \mathscr{A} \quad = \quad & \{f(\cdot) : f^{(2)}(x_0) \text{ exists}, |f^{(2)}(x_0)| \leq b \text{ and} \\ & \left| f(x) - f(x_0) - f^{'}(x_0)(x - x_0) - \frac{f^{(2)}(x_0)}{2}(x - x_0)^2 \right| \leq \frac{a}{6}|x - x_0|^3\} \end{aligned} \tag{3}$$

and

$$\mathscr{B} = \{\sigma^2(\cdot) : \sigma^2(x) \leq d\},$$

where $a, b, d > 0$ are assumed given. In fact, as we will see, the parameter b does not play a role in our deductions.

Roughly speaking, $\mathscr{A}$ contains all twice differentiable functions whose second-order derivative is absolutely bounded by b and third-order derivative is absolutely bounded by a. This characterization is not exact, however, since the Taylor-series type expansion in (3) applies only to the point x_0, and thus $\mathscr{A}$ is more general than the aforementioned characterization. The reason for proposing this class, instead of other similar ones, is that $\mathscr{A}$ allows us to obtain a very accurate minimax analysis and derivation of optimal estimators.

We define the linear minimax L_2-risk as

$$R(n, \mathscr{A}, \mathscr{B}) = \inf_{\substack{\delta_j, j=1,\cdots,n \\ w_j, j=1,\cdots,n}} \sup_{f(\cdot)\in\mathscr{A}, \sigma^2(\cdot)\in\mathscr{B}} E\left(L_n - f^{'}(x_0)\right)^2,$$

which is the minimum worst-case MSE among all functions $f \in \mathscr{A}$ and noise levels in $\mathscr{B}$. The linear estimators are selected through the design points $x_0 + \delta_j$'s and linear coefficients w_j's.

The following theorem gives the exact expression for the minimax risk, and shows that a suitably calibrated CFD attains this risk level. In other words, CFD is the optimal linear estimator among the problem class specified by $(\mathscr{A}, \mathscr{B})$. The proof of this result involves only elementary inequalities.

Theorem 1 For any $n \geq 1$,

$$R(n, \mathscr{A}, \mathscr{B}) \geq \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3}.$$

Besides, if the budget n is even, the CFD estimator $\bar{L}_n$ in (2) with $\delta = (\frac{18d}{a^2})^{1/6}\frac{1}{n^{1/6}}$ satisfies

$$\sup_{f(\cdot)\in\mathscr{A},\sigma^2(\cdot)\in\mathscr{B}} E(\bar{L}_n - f^{'}(x_0))^2 = \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3}.$$

Thus the estimator $\bar{L}_n$ is the best linear estimator in the class of problems defined by $(\mathscr{A}, \mathscr{B})$. □

Proof. For any designs $\delta_j, j = 1, \cdots, n$ and linear coefficients $w_j, j = 1, \cdots, n$, supposing $f(\cdot) \in \mathscr{A}$ and $f(\cdot) \in C^3(\mathbb{R})$, we have, by Taylor's expansion,

$$f(x_0 + \delta_j) = f(x_0) + f^{'}(x_0)\delta_j + \frac{f^{(2)}(x_0)}{2}\delta_j^2 + \frac{f^{(3)}(x_0 + t_j\delta_j)}{6}\delta_j^3,$$

for any $j = 1, \cdots, n$, where $0 \leq t_j \leq 1$. Thus the bias of the estimator L_n is

$$EL_n - f^{'}(x_0) = f(x_0)\sum_{j=1}^n w_j + f^{'}(x_0)\left(\sum_{j=1}^n w_j\delta_j - 1\right) + \frac{f^{(2)}(x_0)}{2}\sum_{j=1}^n w_j\delta_j^2 + \sum_{j=1}^n \frac{f^{(3)}(x_0 + t_j\delta_j)}{6} w_j\delta_j^3.$$

On the other hand, the variance of the estimator L_n is

$$Var(L_n) = \sum_{j=1}^n w_j^2\sigma^2(x_0 + \delta_j).$$

If $\sum_{j=1}^n w_j \neq 0$, we consider the particular cases where $f^{'}(x) = f^{(2)}(x) = f^{(3)}(x) = 0$ for all x, and $f(x_0)$ arbitrary, concluding that

$$\sup_{f(\cdot)\in\mathscr{A}} (EL_n - f^{'}(x_0))^2 = \infty.$$

Therefore, for the purpose of deriving a lower bound, we can assume without loss of generality that $\sum_{j=1}^n w_j = 0$. Similarly we can assume $\sum_{j=1}^n w_j\delta_j - 1 = 0$. Furthermore, if $\delta_i = \delta_j$, we assume w.l.o.g. $w_i = w_j$ since it leads to smaller variance. Now consider $f(\cdot) \in \mathscr{A}$ such that $f(x_0 + \delta_j) = \frac{a}{6}|\delta_j|^3 \cdot sign(w_j)$, and $f(x) = 0$ otherwise. In such a case the MSE simplifies to

$$\left(\sum_{j=1}^n \frac{a}{6}|w_j\delta_j^3|\right)^2 + \sum_{j=1}^n w_j^2\sigma^2(x_0 + \delta_j).$$

Further considering the case $\sigma^2(x_0 + \delta_j) = d$, we get

$$\sup_{f(\cdot)\in\mathscr{A},\sigma^2(\cdot)\in\mathscr{B}} E(L_n - f^{'}(x_0))^2 \geq \frac{a^2}{36}\left(\sum_{j=1}^n |w_j\delta_j^3|\right)^2 + d\sum_{j=1}^n w_j^2.$$

Now since $\sum_{j=1}^n w_j\delta_j = 1$, by Hölder's inequality,

$$1 \leq \sum_{j=1}^n |w_j|^{1/3}|\delta_j||w_j|^{2/3} \leq \left(\sum_{j=1}^n |w_j\delta_j^3|\right)^{1/3}\left(\sum_{j=1}^n |w_j|\right)^{2/3}$$

and so

$$\left(\sum_{j=1}^{n}|w_j\delta_j^3|\right)^2 \geq \frac{1}{\left(\sum_{j=1}^{n}|w_j|\right)^4} \geq \frac{1}{n^2\left(\sum_{j=1}^{n}w_j^2\right)^2},$$

where we used Hölder's inequality $\left(\sum_{j=1}^{n}|w_j|\right)^2 \leq n\left(\sum_{j=1}^{n}w_j^2\right)$. Thus

$$\sup_{f(\cdot)\in\mathscr{A},\sigma^2(\cdot)\in\mathscr{B}} E\left(L_n - f^{'}(x_0)\right)^2 \geq \frac{a^2}{36}\frac{1}{n^2\left(\sum_{j=1}^{n}w_j^2\right)^2} + d\sum_{j=1}^{n}w_j^2 \geq \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3},$$

where the lower bound is achieved at

$$\sum_{j=1}^{n}w_j^2 = \left(\frac{a^2}{d}\right)^{1/3}\frac{1}{18^{1/3}n^{2/3}}.$$

Since δ_j, w_j are arbitrary, we conclude that

$$\inf_{\substack{\delta_j, j=1,\cdots,n\\ w_j, j=1,\cdots,n}} \sup_{f(\cdot)\in\mathscr{A},\sigma^2(\cdot)\in\mathscr{B}} E\left(L_n - f^{'}(x_0)\right)^2 \geq \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3}. \tag{4}$$

On the other hand, supposing the budget n is even, we see that the bias of the estimator $\bar{L}_n$ satisfies

$$\begin{aligned}
|E\bar{L}_n - f^{'}(x_0)| &= \left|\frac{f(x_0+\delta)-f(x_0-\delta)}{2\delta} - f^{'}(x_0)\right| = \left|\frac{f(x_0+\delta)-f(x_0-\delta)-2\delta f^{'}(x_0)}{2\delta}\right| \\
&= \frac{\left|\left(f(x_0+\delta)-f(x_0)-f^{'}(x_0)\delta-\frac{f^{(2)}(x_0)}{2}\delta^2\right)-\left(f(x_0-\delta)-f(x_0)+f^{'}(x_0)\delta-\frac{f^{(2)}(x_0)}{2}\delta^2\right)\right|}{2\delta} \\
&\leq \frac{\left|f(x_0+\delta)-f(x_0)-f^{'}(x_0)\delta-\frac{f^{(2)}(x_0)}{2}\delta^2\right| + \left|f(x_0-\delta)-f(x_0)+f^{'}(x_0)\delta-\frac{f^{(2)}(x_0)}{2}\delta^2\right|}{2\delta} \\
&\leq \frac{a}{6}\delta^2.
\end{aligned}$$

Also, the variance satisfies $Var(\bar{L}_n) \leq \frac{d}{n\delta^2}$. Thus the MSE satisfies

$$E\left(\bar{L}_n - f^{'}(x_0)\right)^2 \leq \frac{a^2}{36}\delta^4 + \frac{d}{n\delta^2} = \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3} \tag{5}$$

by plugging in $\delta = (\frac{18d}{a^2})^{1/6}\frac{1}{n^{1/6}}$. We note that the bound (5) holds for any $f(\cdot)\in\mathscr{A}$ and $\sigma^2(\cdot)\in\mathscr{B}$. Combining (4) and (5), we conclude the proof for Theorem 1. □

2.2 Multi-Dimensional Case

We now generalize to the multi-dimensional case. Consider a performance measure with multi-dimensional input, $f(\cdot):\mathbb{R}^p\to\mathbb{R}$ where $p\geq 2$. Analogous to the single-dimensional case, suppose $Y(\cdot)$ is an unbiased estimate of $f(\cdot)$. We would like to estimate $\nabla f(x_0)$ where $x_0\in\mathbb{R}^p$ is the point of interest. Consider

$$\begin{aligned}
\mathscr{A} \;=\; & \{f(\cdot): \nabla^2 f(x_0) \text{ exists}, \|\nabla^2 f(x_0)\|_2 \leq b \text{ and} \\
& \left|f(x) - f(x_0) - \nabla f(x_0)^T(x-x_0) - \frac{1}{2}(x-x_0)^T\nabla^2 f(x_0)(x-x_0)\right| \leq \frac{a}{6}\|x-x_0\|_2^3\}
\end{aligned}$$

and

$$\mathscr{B} = \{\sigma^2(\cdot) : \sigma^2(x) \le d\},$$

where $a, b, d > 0$, and $\|\cdot\|_2$ denotes the ℓ_2-norm. Given simulation budget $n \ge 1$, we simulate independently at design points $x_0 + \delta_j, j = 1, \cdots, n$, and form the vector-valued linear estimator

$$L_n^p = \sum_{j=1}^n w_j Y_j(x_0 + \delta_j),$$

where $w_j \in \mathbb{R}^p$ are the vector-valued linear coefficients.

Like before, we define the linear minimax L_2-risk as

$$R_p(n, \mathscr{A}, \mathscr{B}) = \inf_{\substack{\delta_j, j=1,\cdots,n \\ w_j, j=1,\cdots,n}} \sup_{f(\cdot)\in\mathscr{A}, \sigma^2(\cdot)\in\mathscr{B}} E\|L_n^p - \nabla f(x_0)\|_2^2.$$

The following theorem provides a lower estimate of R_p and shows that applying CFD on each of the p dimensions matches this lower estimate up to a multiplicative factor depending sublinearly on p. This implies in particular that multi-dimensional CFD is rate-optimal in the sample size n.

Theorem 2 For any $n \ge 1$,

$$R_p(n, \mathscr{A}, \mathscr{B}) \ge p^{4/3} \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3}. \tag{6}$$

Besides, if the budget n is a multiple of both p and 2, we allocate n/p budget and form the CFD estimator on each dimension. Denote the resulting estimator as $\bar{L}_n^p$. Then

$$\sup_{f(\cdot)\in\mathscr{A}, \sigma^2(\cdot)\in\mathscr{B}} E\|\bar{L}_n^p - \nabla f(x_0)\|_2^2 = p^{5/3} \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3}. \tag{7}$$

Thus the estimator $\bar{L}_n^p$ is optimal in the class of problems defined by $(\mathscr{A}, \mathscr{B})$ up to a sublinear multiplicative factor in p. Moreover, if we further restrict each coefficient w_j to have the same sign across components, i.e. $sign((w_j)_k) = sign((w_j)_l)$ for any $1 \le k, l \le p$, then we have

$$R_p(n, \mathscr{A}, \mathscr{B}) \ge p^{5/3} \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3}$$

for this restricted class of linear estimators, so that $\bar{L}_n^p$ is exactly optimal. □

We remark that the $p^{1/3}$ gap between (6) and (7) is due to a technical challenge that the $l_\infty \to l_2$ matrix norm does not admit an explicit expression. This challenge is bypassed if one restricts the same sign across all components in each coefficient, which recovers the minimax optimality of CFD.

Proof. Suppose $f(\cdot) \in \mathscr{A}, f(\cdot) \in C^3(\mathbb{R}^p)$. According to Taylor's expansion

$$f(x_0 + \delta_j) = f(x_0) + \nabla f(x_0)^T \delta_j + \frac{1}{2}\delta_j^T \nabla^2 f(x_0)\delta_j + \frac{1}{6}\sum_{k_1,k_2,k_3} \left(\nabla^3 f(x_0 + t_j\delta_j)\right)_{k_1k_2k_3} (\delta_j)_{k_1}(\delta_j)_{k_2}(\delta_j)_{k_3},$$

for any $j = 1, \cdots, n$, where $0 \le t_j \le 1$. Thus the bias of the estimator L_n^p satisfies

$$E(L_n^p)_i - (\nabla f(x_0))_i = f(x_0)\sum_{j=1}^n (w_j)_i + \nabla f(x_0)^T \left(\sum_{j=1}^n (w_j)_i \delta_j - e_i\right) + \sum_{j=1}^n \frac{1}{2}(w_j)_i \delta_j^T \nabla^2 f(x_0)\delta_j$$
$$+ \sum_{j=1}^n \frac{1}{6}(w_j)_i \sum_{k_1,k_2,k_3} \left(\nabla^3 f(x_0 + t_j\delta_j)\right)_{k_1k_2k_3} (\delta_j)_{k_1}(\delta_j)_{k_2}(\delta_j)_{k_3},$$

where e_i is the ith standard basis in $\mathbb{R}^p$. On the other hand, the variance of the estimator L_n^p is

$$E\|L_n^p - EL_n^p\|_2^2 = \sum_{i=1}^{p} Var((L_n^p)_i) = \sum_{i=1}^{p}\left(\sum_{j=1}^{n}(w_j)_i^2\sigma^2(x_0+\delta_j)\right).$$

If $\sum_{j=1}^{n}(w_j)_i \neq 0$, we consider the particular cases where $(\nabla f(x))_{k_1} = 0, (\nabla^2 f(x))_{k_1k_2} = 0, (\nabla^3 f(x))_{k_1k_2k_3} = 0$ for all x and k_1,k_2,k_3, and $f(x_0)$ arbitrary, concluding that

$$\sup_{f(\cdot)\in\mathscr{A}} (E(L_n^p)_i - (\nabla f(x_0))_i)^2 = \infty.$$

Thus, like in the proof for Theorem 1, for the purpose of deriving a lower bound, we can assume without loss of generality that $\sum_{j=1}^{n} w_j = 0$. Similarly we can assume $\sum_{j=1}^{n}(w_j)_i\delta_j - e_i = 0$. Furthermore, if $\delta_i = \delta_j$, we assume w.l.o.g. $w_i = w_j$ since it leads to smaller variance. Now consider $f(\cdot)\in\mathscr{A}$ such that $f(x_0+\delta_j) = \frac{a}{6}\|\delta_j\|_2^3 \cdot sign((w_j)_{i_0})$, and $f(x) = 0$ otherwise, where

$$i_0 = argmax_{1\le i\le p}\sum_{j=1}^{n}|(w_j)_i|\|\delta_j\|_2^3.$$

In such a case the MSE is bounded from below by

$$\left(\sum_{j=1}^{n}\frac{a}{6}|(w_j)_{i_0}|\|\delta_j\|_2^3\right)^2 + \sum_{i=1}^{p}\left(\sum_{j=1}^{n}(w_j)_i^2\sigma^2(x_0+\delta_j)\right).$$

Considering the case $\sigma^2(x_0+\delta_j) = d$, we get

$$\sup_{f(\cdot)\in\mathscr{A},\sigma^2(\cdot)\in\mathscr{B}} E\|L_n^p - \nabla f(x_0)\|_2^2 \ge \frac{a^2}{36}\left(\sum_{j=1}^{n}|(w_j)_{i_0}|\|\delta_j\|_2^3\right)^2 + d\sum_{j=1}^{n}\|w_j\|_2^2. \tag{8}$$

Now since $\sum_{j=1}^{n}(w_j)_i(\delta_j)_i = 1$, by Hölder's inequality,

$$p = \sum_{j=1}^{n} w_j^T\delta_j \le \sum_{j=1}^{n}\|w_j\|_2^{1/3}\|\delta_j\|_2\|w_j\|_2^{2/3} \le \left(\sum_{j=1}^{n}\|w_j\|_2\|\delta_j\|_2^3\right)^{1/3}\left(\sum_{j=1}^{n}\|w_j\|_2\right)^{2/3}.$$

Thus

$$\frac{p^6}{\left(\sum_{j=1}^{n}\|w_j\|_2\right)^4} \le \left(\sum_{j=1}^{n}\|w_j\|_2\|\delta_j\|_2^3\right)^2 \le \left(\sum_{j=1}^{n}\|w_j\|_1\|\delta_j\|_2^3\right)^2 \le p^2\left(\sum_{j=1}^{n}|(w_j)_{i_0}|\|\delta_j\|_2^3\right)^2,$$

where $\|\cdot\|_1$ denotes ℓ_1-norm. Since by Hölder's inequality,

$$\left(\sum_{j=1}^{n}\|w_j\|_2\right)^2 \le n\left(\sum_{j=1}^{n}\|w_j\|_2^2\right),$$

we have

$$\sup_{f(\cdot)\in\mathscr{A},\sigma^2(\cdot)\in\mathscr{B}} E\|L_n^p - \nabla f(x_0)\|_2^2 \ge \frac{a^2}{36}\frac{p^4}{n^2\left(\sum_{j=1}^{n}\|w_j\|_2^2\right)^2} + d\sum_{j=1}^{n}\|w_j\|_2^2 \ge p^{4/3}\left(\frac{3a^2d^2}{16}\right)^{1/3}n^{-2/3}, \tag{9}$$

where the lower bound is achieved at

$$\sum_{j=1}^{n} \|w_j\|_2^2 = \left(\frac{a^2}{d}\right)^{1/3} \frac{p^{4/3}}{18^{1/3} n^{2/3}}.$$

Since δ_j, w_j are arbitrary, we conclude that

$$\inf_{\substack{\delta_j, j=1,\cdots,n \\ w_j, j=1,\cdots,n}} \sup_{f(\cdot)\in\mathscr{A}, \sigma^2(\cdot)\in\mathscr{B}} E\|L_n^p - \nabla f(x_0)\|_2^2 \geq p^{4/3} \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3}.$$

Now supposing in addition that each w_j has the same sign across components, then instead of (8), we have the sharper lower bound

$$\frac{a^2}{36} \sum_{i=1}^{p} \left(\sum_{j=1}^{n} |(w_j)_i| \|\delta_j\|_2^3 \right)^2 + d \sum_{j=1}^{n} \|w_j\|_2^2.$$

Note that by Hölder's inequality

$$\left(\sum_{j=1}^{n} \|w_j\|_1 \|\delta_j\|_2^3 \right)^2 \leq p \sum_{i=1}^{p} \left(\sum_{j=1}^{n} |(w_j)_i| \|\delta_j\|_2^3 \right)^2.$$

Thus the $p^{4/3}$ factor in (9) can be improved to $p^{5/3}$. Finally, suppose the budget n is a multiple of p and 2. We allocate n/p budget to each dimension. Then for any $f(\cdot) \in \mathscr{A}$ and $\sigma^2(\cdot) \in \mathscr{B}$, we get

$$E\left((\bar{L}_n^p)_i - (\nabla f(x_0))_i\right)^2 \leq \left(\frac{3a^2d^2}{16}\right)^{1/3} \left(\frac{n}{p}\right)^{-2/3} = p^{2/3} \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3}.$$

Thus

$$E\|\bar{L}_n^p - \nabla f(x_0)\|_2^2 \leq p^{5/3} \left(\frac{3a^2d^2}{16}\right)^{1/3} n^{-2/3},$$

which completes our proof. □

The discussion above has focused on deterministic designs in which the perturbation sizes δ_j are fixed. In multi-dimensional gradient estimation, it is also common to use random perturbation in which a random vector in $\mathbb{R}^p$ is generated and FD is taken simultaneously for all dimensions by projecting the vector onto each direction. This leads to schemes such as simultaneous perturbation (Spall 1992) and Gaussian smoothing (Nesterov and Spokoiny 2017), and are frequently used in stochastic optimization. A question arises how these randomized schemes perform with respect to our presented risk criterion.

To proceed, let δ be a random vector in $\mathbb{R}^p$ where $\delta^i, i = 1, \cdots, p$ are i.i.d. symmetrically distributed about 0 and satisfy some additional properties (described in, e.g., Spall 1992, which will not concern us as we will see), and let $\phi(\delta) = (1/\delta^1, \cdots, 1/\delta^p)^T$. Other distributional choices of δ and the associated ϕ have also been suggested (e.g., Nesterov and Spokoiny 2017; Flaxman et al. 2005), for which our subsequent argument follows similarly. Suppose the simulation budget n is even. We choose a scaling parameter $h > 0$, then repeatedly and independently simulate $\delta_j \in \mathbb{R}^p$ and $Y_j(\cdot)$'s and output

$$S_n = \frac{1}{n/2} \sum_{j=1}^{n/2} \frac{Y_{2j-1}(x_0 + h\delta_j) - Y_{2j}(x_0 - h\delta_j)}{2h} \phi(\delta_j).$$

The following theorem shows that, without further assumptions on the magnitude of the first-order gradient, the L_2-risk of random perturbation can be arbitrarily large.

Theorem 3 For any $n \geq 2$ even,

$$\sup_{f(\cdot)\in\mathscr{A},\sigma^2(\cdot)\in\mathscr{B}} E\|S_n - \nabla f(x_0)\|_2^2 = \infty.$$

□

The proof is omitted due to space limit. The unboundedness of the worst-case L_2-risk in Theorem 3 is due to the interaction between the gradient of interest and the variance from the random perturbation. This hints that in general, to contain the worst-case L_2-risk for such schemes, extra knowledge on the magnitude of the gradient is needed.

3 GENERAL MINIMAX RISK

We now expand our analysis to consider estimators that are possibly nonlinear. We first present the single-dimensional case, followed by a brief discussion on the generalization to the multi-dimensional version. We derive tight approximation for the minimax risk and show that, in these expanded classes, the CFD estimators are still nearly optimal.

Adopting the notations from the previous sections, suppose the budget is n. We select the input design points $x_1, \cdots, x_n$ and for convenience let $Y_j = Y_j(x_j)$ be the independent unbiased noisy function evaluation of f at x_j with simulation variance $\sigma^2(x_j)$. We are interested in estimating $f^{'}$ at x_0, which w.l.o.g. we take as the origin 0. Denote $\widehat{\theta} = \widehat{\theta}(Y_1, \cdots, Y_n)$ as a generic estimator. We consider the class of problems specified by

$$\mathscr{A} = \{f(\cdot) : f^{(2)}(0) \text{ exists, } \left| f(x) - f(0) - f^{'}(0)x - \frac{f^{(2)}(0)}{2}x^2 \right| \leq \frac{a}{6}|x|^3 \text{ and } |f^{(2)}(0)| \leq b\}$$

and

$$\mathscr{B} = \{\sigma^2(\cdot) : \sigma^2(x) \leq d\},$$

where $a, b, d > 0$. Like in Section 2, the parameter b does not play a role subsequently. Define the minimax risk as

$$R(n, \mathscr{A}, \mathscr{B}) = \inf_{\substack{x_j, j=1,\cdots,n \\ \widehat{\theta}}} \sup_{f(\cdot)\in\mathscr{A},\sigma^2(\cdot)\in\mathscr{B}} E(\widehat{\theta} - f^{'}(0))^2.$$

Theorem 4 For any $n \geq 1$,

$$R(n, \mathscr{A}, \mathscr{B}) \geq \left(\frac{3^4(\log 2)^2 a^2 d^2}{2^{19}}\right)^{1/3} n^{-2/3}. \tag{10}$$

Consequently, the CFD estimator $\bar{L}_n$ in (2) is optimal up to a constant multiplicative factor. □

The last conclusion in Theorem 4 is a simple consequence from combining the risk estimate of $\bar{L}_n$ in Theorem 1 with (10). In contrast to the elementary proof for Theorem 1, here we use Le Cam's method (e.g., Tsybakov 2009) to estimate the minimax risk.

Proof. Let $f_1(x) = 0$ and

$$f_2(x) = \begin{cases} \varepsilon x - \frac{a}{6}x^3 & \text{if } |x| \leq \sqrt{\frac{6\varepsilon}{a}} \\ 0 & \text{if } |x| > \sqrt{\frac{6\varepsilon}{a}}, \end{cases}$$

where $\varepsilon > 0$. For convenience, denote $\bar{x} = \sqrt{\frac{6\varepsilon}{a}}$. It is clear that $f_1 \in \mathscr{A}$. For f_2, note that if $|x| \leq \bar{x}$, then

$$f_2(x) - f_2(0) - f_2^{'}(0)x - \frac{f_2^{(2)}(0)}{2}x^2 = f_2(x) - \varepsilon x = -\frac{a}{6}x^3.$$

On the other hand, if $x > \bar{x}$, then

$$\left| f_2(x) - f_2(0) - f_2'(0)x - \frac{f_2^{(2)}(0)}{2}x^2 \right| = |f_2(x) - \varepsilon x| \le |f_2(x) - f_2(\bar{x})| + |f_2(\bar{x}) - \varepsilon\bar{x}| + \varepsilon|x - \bar{x}|$$
$$\le 0 + \frac{a}{6}\bar{x}^3 + \frac{a}{6}\bar{x}^2(x - \bar{x}) \le \frac{a}{6}(\bar{x}^3 + x^3 - \bar{x}^3) \le \frac{a}{6}x^3.$$

Similarly for $x < -\bar{x}$. Thus $f_2 \in \mathscr{A}$. Now for any estimator $\widehat{\theta}$, we have

$$\varepsilon^2 = (f_1'(0) - f_2'(0))^2 \le 2(\widehat{\theta} - f_1'(0))^2 + 2(\widehat{\theta} - f_2'(0))^2.$$

Define test statistic ψ by

$$\psi(Y_1, \cdots, Y_n) = \begin{cases} 1 & \text{if } |\widehat{\theta} - f_2'(0)| \le |\widehat{\theta} - f_1'(0)| \\ 2 & \text{if } |\widehat{\theta} - f_2'(0)| > |\widehat{\theta} - f_1'(0)|. \end{cases}$$

Let $E_k, k = 1, 2$ denote the expectation (and P_k and p_k as the probability measure and density) under model

$$Y_j \sim f_k(x_j) + \eta_j, j = 1, \cdots, n,$$

where $\eta_j, j = 1, \cdots, n$ i.i.d. follows normal distribution with mean zero and variance d. We have

$$E_k(\widehat{\theta} - f_k'(0))^2 \ge E_k\left[(\widehat{\theta} - f_k'(0))^2 I(\psi = k)\right] \ge \frac{\varepsilon^2}{4} P_k(\psi = k).$$

Thus

$$\sup_{f \in \mathscr{A}, \sigma^2 \in \mathscr{B}} E(\widehat{\theta} - f'(0))^2 \ge \max_{k=1,2} E_k(\widehat{\theta} - f_k'(0))^2 \ge \frac{\varepsilon^2}{4} \frac{P_1(\psi = 1) + P_2(\psi = 2)}{2}.$$

Taking infimum over all possible estimators, we get

$$\inf_{\widehat{\theta}} \sup_{f \in \mathscr{A}, \sigma^2 \in \mathscr{B}} E(\widehat{\theta} - f'(0))^2 \ge \frac{\varepsilon^2}{8} \inf_{\psi} (P_1(\psi = 1) + P_2(\psi = 2)).$$

The right hand side is minimized by the Neyman-Pearson test, i.e.

$$\psi_0(y_1, \cdots, y_n) = \begin{cases} 1 & \text{if } p_2(y_1, \cdots, y_n) \ge p_1(y_1, \cdots, y_n) \\ 2 & \text{if } p_2(y_1, \cdots, y_n) < p_1(y_1, \cdots, y_n). \end{cases}$$

Thus by a standard relation (e.g. Lemma 2.6 in Tsybakov 2009)

$$\inf_{\widehat{\theta}} \sup_{f \in \mathscr{A}, \sigma^2 \in \mathscr{B}} E(\widehat{\theta} - f'(0))^2 \ge \frac{\varepsilon^2}{8} \int \min\{p_1(y_1, \cdots, y_n), p_2(y_1, \cdots, y_n)\} dy \ge \frac{\varepsilon^2}{16} e^{-KL(P_1, P_2)},$$

where $KL(P_1, P_2)$ denote the KL divergence between the distributions P_1, P_2. Now since $P_1 \sim \mathscr{N}(\mu_1, dI_{n\times n})$ and $P_2 \sim \mathscr{N}(\mu_2, dI_{n\times n})$, where $\mu_k = (f_k(x_1), \cdots, f_k(x_n)), k = 1, 2$, by direct computation,

$$KL(P_1, P_2) = \frac{1}{2d}(\mu_2 - \mu_1)^T(\mu_2 - \mu_1) = \frac{1}{2d}\|\mu_2\|_2^2.$$

Since $|f_2(x)| \le \frac{2\sqrt{2}}{3}\frac{\varepsilon^{3/2}}{\sqrt{a}}$, we get

$$\inf_{\widehat{\theta}} \sup_{f \in \mathscr{A}, \sigma^2 \in \mathscr{B}} E(\widehat{\theta} - f'(0))^2 \ge \frac{\varepsilon^2}{16} e^{-\frac{4n\varepsilon^3}{9ad}}.$$

Now take $\varepsilon = \left(\frac{9ad\log 2}{4n}\right)^{1/3}$, we get

$$\inf_{\widehat{\theta}} \sup_{f\in\mathscr{A},\sigma^2\in\mathscr{B}} E(\widehat{\theta} - f^{'}(0))^2 \geq \frac{\varepsilon^2}{32} = \left(\frac{3^4(\log 2)^2 a^2 d^2}{2^{19}}\right)^{1/3} n^{-2/3}.$$

We complete our proof by noting that the above bound holds for any design points $x_1, \cdots, x_n$. □

We comment that a similar result and proof hold for the multi-dimensional case, by defining $\mathscr{A}$ for functions with a multi-dimensional domain analogous to Section 2.2. In particular, the same bound for the L_2-risk as (10) can be shown to hold in this setting which, by combining with (7) in Theorem 2, shows that the multi-dimensional CFD estimator $\bar{L}_n^p$ is optimal among the general class of estimators up to a multiplicative constant and the factor $p^{5/3}$. Note that this discussed lower bound does not involve the dimension and is potentially loose. To get sharper bounds, one can analyze the "modulus of continuity" (Donoho 1994) that involves a constrained functional optimization problem, which we leave to future research.

4 CONCLUSION

In this paper we studied the minimax optimality of stochastic gradient estimators when only noisy function evaluations are available, with respect to the worst-case MSE among unknown twice differentiable functions. We derived the exact minimax risk for the class of linear estimators, and showed that CFD is optimal within this class, in the single-dimensional case. We extended the analysis to the multi-dimensional case to approximate the minimax risk and show the optimality of CFD up to multiplicative factors sublinear in the dimension. We also showed that, without further assumptions on the gradient magnitude, the worst-case risk of random perturbation schemes can be unbounded. Next we approximated the minimax risk over general class of (nonlinear) estimators and showed that CFD is still nearly optimal over this much larger estimator class. These approximations were shown up to a constant factor in the single-dimensional case and an additional factor depending polynomially on the dimension in the multi-dimensional case. In future work, we will investigate the use of additional a priori information on the considered function class (noting that the worst-case functions in our analysis were non-differentiable at input values away from the point of interest), and tightening our minimax estimates using more accurate alternate approaches.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1653339/1834710 and IIS-1849280.

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*, Volume 57. New York: Springer Science & Business Media.

Dai, W., T. Tong, and M. G. Genton. 2016. "Optimal Estimation of Derivatives in Nonparametric Regression". *The Journal of Machine Learning Research* 17(1):5700–5724.

De Brabanter, K., J. De Brabanter, B. De Moor, and I. Gijbels. 2013. "Derivative Estimation with Local Polynomial Fitting". *The Journal of Machine Learning Research* 14(1):281–301.

Donoho, D. L. 1994. "Statistical Estimation and Optimal Recovery". *The Annals of Statistics* 22(1):238–270.

Fan, J. 1993. "Local Linear Regression Smoothers and Their Minimax Efficiencies". *The Annals of Statistics* 21(1):196–216.

Fan, J., T. Gasser, I. Gijbels, M. Brockmann, and J. Engel. 1997. "Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency". *Annals of the Institute of Statistical Mathematics* 49(1):79–99.

Flaxman, A. D., A. T. Kalai, and H. B. McMahan. 2005. "Online Convex Optimization in the Bandit Setting: Gradient Descent without a Gradient". In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, 385–394. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

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

Frolov, A., and N. Chentsov. 1963. "On the Calculation of Definite Integrals Dependent on a Parameter by the Monte Carlo Method". *USSR Computational Mathematics and Mathematical Physics* 2(4):802–807.

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. New York: 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., 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* 20(1):5/1–5/28.

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 *Proceedings of the 1991 Winter Simulation Conference*, edited by B. Nelson, W. D. Kelton, and G. M. Clark, 207–217. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Nesterov, Y., and V. Spokoiny. 2017. "Random Gradient-Free Minimization of Convex Functions". *Foundations of Computational Mathematics* 17(2):527–566.

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.

Spall, J. C. 1992. "Multivariate Stochastic Approximation using a Simultaneous Perturbation Gradient Approximation". *IEEE transactions on automatic control* 37(3):332–341.

Tsybakov, A. B. 2009. *Introduction to Nonparametric Estimation*. New York: Springer Science & Business Media.

Wang, W. W., and L. Lin. 2015. "Derivative Estimation Based on Difference Sequence via Locally Weighted Least Squares Regression". *The Journal of Machine Learning Research* 16(1):2617–2641.

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

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. His research focuses on Monte Carlo simulation, uncertainty quantification, risk analysis, and stochastic and robust optimization. His email address is khl2114@columbia.edu.

XUHUI ZHANG is an undergraduate student in the School for the Gifted Young at University of Science and Technology of China, majoring in mathematics and applied mathematics. He was a summer intern under the supervision of Prof. Henry Lam in 2018. His email address is zxh1998@mail.ustc.edu.cn.