Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

ACHIEVING OPTIMAL BIAS-VARIANCE TRADEOFF IN ONLINE DERIVATIVE ESTIMATION

Thibault Duplay
Henry Lam
Xinyu Zhang

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

ABSTRACT

The finite-difference method has been commonly used in stochastic derivative estimation when an unbiased derivative estimator is unavailable or costly. The efficiency of this method relies on the choice of a perturbation parameter, which needs to be calibrated based on the number of simulation replications. We study the setting where such an a priori planning of simulation runs is difficult, which could arise due to the variability of runtime for complex simulation models or interruptions. We show how a simple recursive weighting scheme on simulation outputs can recover, in an online fashion, the optimal asymptotic bias-variance tradeoff achieved by the conventional scheme where the replication size is known in advance.

1 INTRODUCTION

Consider a performance measure $f(x) \in \mathbb{R}$ evaluated at a parameter $x \in \mathbb{R}$. We focus on the setting where given a value of x, we have access to a noisy, unbiased, observation of $f(x)$, i.e., $Y(x) = f(x) + \varepsilon(x)$, where $\varepsilon(x)$ is a random variable that satisfies $E[\varepsilon(x)|x] = 0$. We are interested in estimating $f'(x)$, the derivative of f with respect to a given x, using the above oracle. Derivative estimation of this sort is motivated from applications in sensitivity analysis, uncertainty quantification, and gradient-based optimization (see, e.g., the surveys Glasserman 2013; Asmussen and Glynn 2007; Fu 2015).

A conventional approach to the above problem is to use the finite-difference method. Namely, we select a perturbation parameter, called $\delta > 0$, and generate two simulation replications, one at $x+\delta$ and another at $x-\delta$, and output

$$\frac{Y(x+\delta) - Y(x-\delta)}{2\delta}. \tag{1}$$

This is known as the central finite-difference method. Another version is to simulate at $x+\delta$ and x, giving rise to

$$Z = \frac{Y(x+\delta) - Y(x)}{\delta} \tag{2}$$

which is known as the forward finite-difference. Similar idea arises as the backward finite-difference. Suppose we have a budget of R pairs of simulation replications. We will repeatedly generate independent copies of (1) or (2) above, and take their sample average.

Note that, depending on the smoothness of f, (1) and (2) incurs bias in estimating f'. It is widely known that under enough differentiability, by choosing $\delta = \Theta(1/R^{1/6})$, the central finite-difference estimator achieves an asymptotically optimal mean square error (MSE) of order $O(1/R^{2/3})$, while by choosing $\delta = \Theta(1/R^{1/4})$, the forward finite-difference estimator achieves an asymptotically optimal MSE $O(1/R^{1/2})$.

In other words, in order to achieve the optimal MSE, which involves making the best bias-variance tradeoff, one needs to calibrate δ according to R.

This paper investigates the situations when the total simulation budget R is unknown a priori. This situation can arise when planning simulation in advance is difficult due to the variability of runtime for complex simulation models, or when the simulation modeler wants to hedge against unexpected future interruptions. When this is the case, the conventional approach in choosing δ guided by the optimal MSE cannot be used directly. This issue is especially impactful if the simulation model is large-scale and consumes substantial computation resources.

Our main contribution is to study a simple recursive approach to update our derivative estimate "on the fly" that recovers the asymptotically optimal MSE as the simulation replication size increases. This can be viewed as an "online" derivative estimator, as opposed to the "offline" estimator in the conventional approach that needs to know R in advance. Our approach takes a weighted average of a new simulation output with the current derivative estimate. By carefully choosing the weighting sequence in terms of the current iteration number, we show that we can recover the asymptotically optimal MSE achieved by the offline estimator. Roughly speaking, our analysis comprises bounding separately the sequential bias and variance of the iterates to inform the overall MSE. This can be done for both the central and the forward/backward finite-difference schemes. In addition to theoretical analyses, we also provide numerical tests of our approach, and demonstrate how our approach avoids the suboptimal performance of the offline methods when R is not known in advance.

The recursion considered in our method resembles the widely used stochastic approximation (SA) (Kushner and Yin 2003) for stochastic root-finding and optimization, particularly the Kiefer-Wolfowitz scheme that does not use gradient information in the iterate. Compared to these problems tackled by SA, ours is in a sense simpler, as the derivative estimation problem is linear in nature (seen as a simple consequence of Taylor's expansion), and this in turn allows us to do more explicit analyses among various choices of our parameters. For general references on SA or stochastic gradient descent, see, e.g., Kushner and Yin (2003), Fu (1994), Pasupathy and Kim (2011). Our analysis follows in particular the methods in obtaining finite-time bounds commonly used in analyzing stochastic or convex optimization algorithms (e.g., Nemirovski et al. 2009).

Finally, in this work, we only consider the univariate case where x is a one-dimensional parameter, and will leave the multivariate generalization to a later study. We will also delegate the connections and integrations of our approach with other debiasing or variance reduction techniques such as randomized multi-level Monte Carlo (e.g., Blanchet and Glynn 2015; Blanchet et al. 2017; Blanchet and Glynn 2015) and common random numbers to future study.

In the remainder of this paper, we will present our algorithm (Section 2), the analysis (Section 3), and numerical demonstrations and comparisons with the conventional approaches (Section 4).

2 PROCEDURE

Given a point, say x_0, we are interested in estimating $\theta = f'(x_0)$ using copies of $Y(\cdot)$. We propose using an iterative approach to run our simulation-based estimation. Choose a step size sequence γ_n and a perturbation sequence δ_n, for $n = 1, 2, \ldots$. Given a current estimate θ_n, we update

$$\theta_{n+1} = (1-\gamma_n)\theta_n + \gamma_n \frac{Y_n(x_0+\delta_n) - Y_n(x_0-\delta_n)}{2\delta_n}$$

where $Y_n(x_0+\delta_n)$ and $Y_n(x_0-\delta_n)$ are independent simulation copies. We call this the iterative central finite-difference (ICFD). Alternately, we can use

$$\theta_{n+1} = (1-\gamma_n)\theta_n + \gamma_n \frac{Y_n(x_0+\delta_n) - Y_n(x_0)}{\delta_n}$$

giving rise to the iterative forward finite-difference (IFFD). Similar approach using $Y_n(x_0) - Y_n(x_0 - \delta_n)$ will lead to the backward finite-difference, which we skip to avoid redundancy in our discussion. We summarize ICFD and IFFD in the following algorithm. The only difference between ICFD and IFFD is in Step 4. The algorithm is understood to run until the budget is consumed or interruption occurs.

Algorithm 1 Iterative Central/Forward Finite-Difference for estimating $f'(x_0)$.

1: **Inputs:**
   simulation oracle to generate $Y(x)$ as an unbiased output for $f(x)$
   sequences $\{\gamma_n\}, \{\delta_n\}, n = 1, 2, \ldots$
   point of interest x_0
2: **Initialize:**
   $\theta_1 \leftarrow$ an arbitrary number, e.g., 0
   $n \leftarrow 1$
3: **repeat**
4: ICFD: $\theta_{n+1} \leftarrow (1-\gamma_n)\theta_n + \gamma_n \frac{Y_n(x_0+\delta_n) - Y_n(x_0-\delta_n)}{2\delta_n}$; IFFD: $\theta_{n+1} \leftarrow (1-\gamma_n)\theta_n + \gamma_n \frac{Y_n(x_0+\delta_n) - Y_n(x_0)}{\delta_n}$
5: $n \leftarrow n+1$
6: **until** budget depletion
7: **Output:** θ_n

3 THEORETICAL GUARANTEES AND ANALYSIS

We analyze the errors made by ICFD and IFFD. We first make the following smoothness assumptions:

Assumption 1 $f(x)$ is twice continuously differentiable in a neighborhood of x_0.

Assumption 2 $f(x)$ is three times continuously differentiable in a neighborhood of x_0.

Assumption 1 is used for IFFD, whereas Assumption 2 is used for ICFD. Next, we impose the following bound on the variance of the simulation:

Assumption 3 Simulation output $Y(x)$ has a bounded variance in the d-neighborhood of x_0, i.e. for some $M > 0$, $\sup_{y:|y-x_0|\le d} \mathrm{Var}(Y(y)) \le M$.

Our analysis will use the function form that $\gamma_n = c/n$ and $\delta_n = d/n^\alpha$ for n sufficiently large. The choice of γ_n is motivated from the standard step size used in SA, whereas the choice of δ_n is motivated from the perturbation size used commonly in the conventional finite-difference schemes.

The following shows the choices of parameters such that our procedure is consistent:

Theorem 1 (L_2-Consistency)
Set $\gamma_n = \min\{\frac{c}{n}, 1\}$ and $\delta_n = \frac{d}{n^\alpha}$, where $c, d, \alpha > 0$. For each case below, we assume that $x_0 \pm \delta_n$ for $n \ge 1$ is in the neighborhood of continuous differentiability of f.

1. Under Assumptions 1 and 3, and $0 < \alpha < \frac{1}{2}$, the iterates of IFFD, θ_n, converges to the true derivative θ in L_2 as $n \to \infty$, i.e., $E(\theta_n - \theta)^2 \to 0$ as $n \to \infty$.
2. Under Assumptions 2 and 3, and $0 < \alpha < \frac{1}{2}$, the iterates of ICFD, θ_n, converges to the true derivative θ in L_2 as $n \to \infty$, i.e., $E(\theta_n - \theta)^2 \to 0$ as $n \to \infty$.

Sketch of Proof. We first analyze IFFD. For simplicity, we denote $Z(\delta) = \frac{Y(x_0+\delta) - Y(x)}{\delta}$. Based on Assumptions 1 and 3, and by Taylor's expansion, $Z(\delta)$ has bias $(f^{(2)}(\xi)/2!)\delta$ where ξ is between x_0 and $x_0 + \delta$, for δ sufficiently small. This is absolutely bounded by $M'\delta$ for some $M' > 0$. The variance of $Z(\delta)$ is bounded by $2M/\delta^2$. Let $n_0 = \lceil c \rceil$ if $\lceil c \rceil \ne c$ and $c+1$ otherwise. Then $0 < 1 - \gamma_n < 1, n \ge n_0$. Denote

$V_n = \mathrm{Var}(\theta_n)$ and $b_n = |\mathrm{Bias}(\theta_n)|$. We have

$$V_{n+n_0} \le (1-\gamma_{n+n_0-1})^2 V_{n+n_0-1} + \gamma_{n+n_0-1}^2 \frac{2M}{\delta_{n+n_0-1}^2}$$

and

$$b_{n+n_0} \le (1-\gamma_{n+n_0-1}) b_{n+n_0-1} + \gamma_{n+n_0-1} M' \delta_{n+n_0-1}.$$

Through a recursive analysis, we see that

$$\begin{aligned}
V_{n+n_0-1} &\le (1-\gamma_{n+n_0-2})^2 V_{n+n_0-2} + \gamma_{n+n_0-2}^2 \frac{2M}{\delta_{n+n_0-2}^2} \\
&\le \prod_{i=1}^{n-1} (1-\gamma_{n_0+i-1})^2 V_{n_0} + \gamma_{n+n_0-2}^2 \frac{2M}{\delta_{n+n_0-2}^2} + \sum_{i=1}^{n-2} \prod_{k=i}^{n-2} (1-\gamma_{k+n_0})^2 \gamma_{i-1+n_0}^2 \frac{2M}{\delta_{i-1+n_0}^2} \\
&\le V_{n_0} e^{-2c \sum_{i=1}^{n-1} \frac{1}{n_0+i-1}} + \gamma_{n+n_0-2}^2 \frac{2M}{\delta_{n+n_0-2}^2} + \sum_{i=1}^{n-2} e^{-2c \sum_{k=i+1}^{n-1} \frac{1}{k+n_0-1}} \gamma_{n_0+i-1}^2 \frac{2M}{\delta_{i-1+n_0}^2} \\
&\le V_{n_0} \Big(\frac{n_0}{n+n_0-1}\Big)^{2c} + \gamma_{n+n_0-2}^2 \frac{2M}{\delta_{n+n_0-2}^2} + \sum_{i=1}^{n-2} \Big(\frac{i+n_0}{n+n_0-1}\Big)^{2c} \gamma_{i-1+n_0}^2 \frac{2M}{\delta_{i+n_0-1}^2} \\
&= V_{n_0} \Big(\frac{n_0}{n+n_0-1}\Big)^{2c} + \sum_{i=1}^{n-1} \Big(\frac{i+n_0}{n+n_0-1}\Big)^{2c} \gamma_{n_0+i-1}^2 \frac{2M}{\delta_{n_0+i-1}^2} \\
&= \frac{1}{(n+n_0-1)^{2c}} \Big(V_{n_0} n_0^{2c} + \frac{2Mc^2}{d^2} \sum_{i=1}^{n-1} (i+n_0)^{2c} (i+n_0-1)^{2\alpha-2} \Big).
\end{aligned}$$

Likewise,

$$\begin{aligned}
b_{n+n_0-1} &\le (1-\gamma_{n+n_0-2}) b_{n+n_0-2} + \gamma_{n+n_0-2} M' \delta_{n+n_0-2} \\
&\le \prod_{i=1}^{n-1} (1-\gamma_{i+n_0-1}) b_{n_0} + M' \gamma_{n+n_0-2} \delta_{n+n_0-2} + M' \sum_{i=1}^{n-2} \gamma_{i+n_0-1} \delta_{i+n_0-1} \prod_{k=i}^{n-2} (1-\gamma_{k+n_0}) \\
&\le b_{n_0} e^{-c \sum_{i=1}^{n-1} \frac{1}{i+n_0-1}} + M' \gamma_{n+n_0-2} \delta_{n+n_0-2} + M' \sum_{i=1}^{n-2} \gamma_{i+n_0-1} \delta_{i+n_0-1} e^{-c \sum_{k=i}^{n-2} \frac{1}{k+n_0}} \\
&\le b_{n_0} \Big(\frac{n_0}{n+n_0-1}\Big)^{c} + M' \gamma_{n+n_0-2} \delta_{n+n_0-2} + M' \sum_{i=1}^{n-2} \gamma_{i+n_0-1} \delta_{i+n_0-1} \Big(\frac{i+n_0}{n+n_0-1}\Big)^{c} \\
&= b_{n_0} \Big(\frac{n_0}{n+n_0-1}\Big)^{c} + \frac{M'}{(n+n_0-1)^c} \sum_{i=1}^{n-1} \gamma_{i+n_0-1} \delta_{i+n_0-1} (i+n_0)^c \\
&= \frac{1}{(n+n_0-1)^c} \Big[b_{n_0} n_0^c + M' cd \sum_{i=1}^{n-1} \frac{1}{(i+n_0-1)^{\alpha+1}} (i+n_0)^c \Big].
\end{aligned}$$

Therefore the MSE is bounded from above by

$$UB := \frac{1}{(n+n_0-1)^{2c}} \Big[V_{n_0} n_0^{2c} + \frac{2Mc^2}{d^2} \sum_{i=1}^{n-1} (i+n_0)^{2c} (n_0+i-1)^{2\alpha-2} \Big]$$

$$+\frac{1}{(n+n_0-1)^{2c}}\Big[b_{n_0}n_0^c+M'cd\sum_{i=1}^{n-1}\frac{1}{(n_0+i-1)^{\alpha+1}}(i+n_0)^c\Big]^2$$
$$=\frac{1}{(n+n_0-1)^{2c}}\Big[V_{n_0}n_0^{2c}+b_{n_0}^2n_0^{2c}+\frac{2Mc^2}{d^2}\sum_{i=1}^{n-1}(i+n_0)^{2c}(i+n_0-1)^{2\alpha-2}$$
$$+\Big(M'cd\sum_{i=1}^{n-1}\frac{(i+n_0)^c}{(n_0+i-1)^{\alpha+1}}\Big)^2+2b_{n_0}n_0^ccdM'\sum_{i=1}^{n-1}\frac{(i+n_0)^c}{(i+n_0-1)^{\alpha+1}}\Big].$$

Now letting $K_1:=\sum_{i=1}^{n-1}(i+n_0)^{2c}(i+n_0-1)^{2\alpha-2}$ and $K_2:=\sum_{i=1}^{n-1}\frac{(i+n_0)^c}{(i+n_0-1)^{\alpha+1}}$, we have

$$K_1=\sum_{i=1}^{n-1}(1+\frac{1}{i+n_0-1})^{2c}(i+n_0-1)^{2\alpha+2c-2}$$
$$\le(1+\frac{1}{n_0})^{2c}\sum_{i=1}^{n-1}(i+n_0-1)^{2\alpha+2c-2}$$
$$\le(1+\frac{1}{n_0})^{2c}\begin{cases}\frac{(n+n_0-1)^{2\alpha+2c-1}}{2\alpha+2c-1} & \text{if } 2\alpha+2c-2\ge 0\\ n_0^{2\alpha+2c-2}+\frac{(n+n_0-2)^{2c+2\alpha-1}}{2c+2\alpha-1} & \text{if } -1<2\alpha+2c-2<0\\ n_0^{2\alpha+2c-2}+\ln(\frac{n+n_0-2}{n_0}) & \text{if } 2\alpha+2c-2=-1\\ n_0^{2\alpha+2c-2}+\frac{n_0^{2\alpha+2c-1}}{1-2c-2\alpha} & \text{if } 2\alpha+2c-2<-1\end{cases}$$

$$K_2=\sum_{i=1}^{n-1}\frac{(i+n_0)^c}{i^{\alpha+1}}=\sum_{i=1}^{n-1}\frac{(1+\frac{1}{i+n_0-1})^c}{(i+n_0-1)^{\alpha+1-c}}$$
$$\le(1+\frac{1}{n_0})^c\sum_{i=1}^{n-1}(i+n_0-1)^{c-\alpha-1}$$
$$\le(1+\frac{1}{n_0})^c\begin{cases}\frac{(n+n_0-1)^{c-\alpha}}{c-\alpha} & \text{if } c-\alpha-1\ge 0\\ n_0^{c-\alpha-1}+\frac{(n+n_0-2)^{c-\alpha}}{c-\alpha} & \text{if } -1<c-\alpha-1<0\\ n_0^{c-\alpha-1}+\ln(\frac{n+n_0-2}{n_0}) & \text{if } c-\alpha-1=-1\\ n_0^{c-\alpha-1}+\frac{n_0^{c-\alpha}}{\alpha-c} & \text{if } c-\alpha-1<-1\end{cases}.$$

With the positivity of c and α and the above expressions, we obtain the following 13 scenarios in which C is some large constant:

$$UB\le\frac{C}{(n+n_0-1)^{2c}}\times$$

$$\begin{cases}1+(n+n_0-1)^{2\alpha+2c-1}+(n+n_0-1)^{2c-2\alpha}+(n+n_0-1)^{c-\alpha}\\ \quad\text{if } 2\alpha+2c-2\ge 0 \text{ and } c-\alpha-1\ge 0\\ 1+(n+n_0-1)^{2\alpha+2c-1}+(n+n_0-2)^{2c-2\alpha}+(n+n_0-2)^{c-\alpha}\\ \quad\text{if } 2\alpha+2c-2\ge 0 \text{ and } -1<c-\alpha-1<0\\ 1+(n+n_0-1)^{2\alpha+2c-1}+\ln(n+n_0-2)+\ln^2(n+n_0-2)\\ \quad\text{if } 2\alpha+2c-2\ge 0 \text{ and } c-\alpha-1=-1\\ 1+(n+n_0-1)^{2\alpha+2c-1}\\ \quad\text{if } 2\alpha+2c-2\ge 0 \text{ and } c-\alpha-1<-1\\ 1+(n+n_0-2)^{2c+2\alpha-1}+(n+n_0-2)^{c-\alpha}+(n+n_0-2)^{2c-2\alpha}\\ \quad\text{if } -1<2\alpha+2c-2<0 \text{ and } -1<c-\alpha-1<0\\ 1+(n+n_0-2)^{2c+2\alpha-1}+\ln(n+n_0-2)+\ln^2(n+n_0-2)\\ \quad\text{if } -1<2\alpha+2c-2<0 \text{ and } c-\alpha-1=-1\\ 1+(n+n_0-2)^{2c+2\alpha-1}\\ \quad\text{if } -1<2\alpha+2c-2<0 \text{ and } c-\alpha-1<-1\end{cases}\begin{cases}1+\ln(n+n_0-2)+(n+n_0-2)^{2c-2\alpha}+(n+n_0-2)^{c-\alpha}\\ \quad\text{if } 2\alpha+2c-2=-1 \text{ and } -1<c-\alpha-1<0\\ 1+\ln(n+n_0-2)+\ln^2(n+n_0-2)\\ \quad\text{if } 2\alpha+2c-2=-1 \text{ and } c-\alpha-1=-1\\ 1+\ln(n+n_0-2)\\ \quad\text{if } 2\alpha+2c-2=-1 \text{ and } c-\alpha-1<-1\\ 1+(n+n_0-2)^{2c-2\alpha}+(n+n_0-2)^{c-\alpha}\\ \quad\text{if } 2\alpha+2c-2<-1 \text{ and } -1<c-\alpha-1<0\\ 1+\ln^2(n+n_0-2)+\ln(n+n_0-2)\\ \quad\text{if } 2\alpha+2c-2<-1 \text{ and } c-\alpha-1=-1\\ 1\\ \quad\text{if } 2\alpha+2c-2<-1 \text{ and } c-\alpha-1<-1\end{cases}.$$

These can be simplified and summarized in Table 1:

Table 1: Upper bound of MSE for different scenarios under IFFD.

	$c-\alpha-1\geq 0$	$-1<c-\alpha-1<0$	$c-\alpha-1=-1$	$c-\alpha-1<-1$
$2\alpha+2c-2\geq 0$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}}+\frac{1}{n^{2\alpha}}+\frac{1}{n^{c+\alpha}})$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}}+\frac{1}{n^{2\alpha}}+\frac{1}{n^{c+\alpha}})$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}}+\frac{\ln^2(n)}{n^{2c}}+\frac{\ln(n)}{n^{2c}})$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}})$
$-1<2\alpha+2c-2<0$	n.a.	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}}+\frac{1}{n^{2\alpha}}+\frac{1}{n^{c+\alpha}})$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}}+\frac{\ln^2(n)}{n^{2c}}+\frac{\ln(n)}{n^{2c}})$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}})$
$2\alpha+2c-2=-1$	n.a.	$O(\frac{1}{n^{2c}}+\frac{\ln(n)}{n^{2c}}+\frac{1}{n^{2\alpha}}+\frac{1}{n^{c+\alpha}})$	$O(\frac{1}{n^{2c}}+\frac{\ln(n)}{n^{2c}}+\frac{\ln^2(n)}{n^{2c}}+\frac{\ln(n)}{n^{2c}})$	$O(\frac{1}{n^{2c}}+\frac{\ln(n)}{n^{2c}})$
$2\alpha+2c-2<-1$	n.a.	$O(\frac{1}{n^{2c}}+\frac{1}{n^{2\alpha}}+\frac{1}{n^{c+\alpha}})$	$O(\frac{1}{n^{2c}}+\frac{\ln^2(n)}{n^{2c}}+\frac{\ln(n)}{n^{2c}})$	$O(\frac{1}{n^{2c}})$

From Table 1 we can see that if $0<\alpha<1/2$, the MSE is $O(\frac{1}{n^{\Delta}})$ where $\Delta>0$, which implies it will converge to zero as the replication size increases.

Similarly, for ICFD, we denote $Z(\delta)=\frac{Y(x_0+\delta)-Y(x_0-\delta)}{2\delta}$. Based on Assumptions 2 and 3, and by Taylor's expansion, this estimator has bias $((f^{(3)}(\xi_1)+f^{(3)}(\xi_2))/3!)\delta^2$ where ξ_1 is between x_0 and $x_0+\delta$ and ξ_2 is between $x_0-\delta$ and x_0, for δ sufficiently small. This is absolutely bounded by $M'\delta^2$ for some $M'>0$. Moreover, the variance of $Z(\delta)$ is bounded by $M/(2\delta^2)$. So we have

$$V_{n+n_0}\leq(1-\gamma_{n+n_0-1})^2V_{n+n_0-1}+\gamma_{n+n_0-1}^2\frac{M}{2\delta_{n+n_0-1}^2}$$

and

$$b_{n+n_0}\leq(1-\gamma_{n+n_0-1})b_{n+n_0-1}+\gamma_{n+n_0-1}M'\delta_{n+n_0-1}^2.$$

Through a similar recursive analysis, we see that

$$V_n\leq\frac{1}{(n+n_0-1)^{2c}}\Big(V_{n_0}n_0^{2c}+\frac{Mc^2}{2d^2}\sum_{i=1}^{n-1}(n+n_0)^{2c}(n+n_0-1)^{2\alpha-2}\Big)$$

and

$$b_{n+n_0-1}\leq\frac{1}{(n+n_0-1)^{c}}\Big[b_{n_0}n_0^{c}+cd^2M'\sum_{i=1}^{n-1}\frac{1}{(n+n_0-1)^{2\alpha+1}}(n+n_0)^{c}\Big].$$

Therefore the MSE is bounded from above by

$$\begin{aligned}UB:=\frac{1}{(n+n_0-1)^{2c}}\Big[&V_{n_0}n_0^{2c}+b_{n_0}^2n_0^{2c}+\frac{Mc^2}{2d^2}\sum_{i=1}^{n-1}(n+n_0)^{2c}(n+n_0-1)^{2\alpha-2}\\&+\Big(cd^2M'\sum_{i=1}^{n-1}\frac{(n+n_0)^{c}}{(n+n_0-1)^{2\alpha+1}}\Big)^2+2b_{n_0}n_0^{c}cd^2M'\sum_{i=1}^{n-1}\frac{(n+n_0)^{c}}{(n+n_0-1)^{2\alpha+1}}\Big].\end{aligned}$$

Again, the upper bounds can be summarized in Table 2:

Table 2: Upper bound of MSE for different scenarios under ICFD.

	$c-2\alpha-1\geq 0$	$-1<c-2\alpha-1<0$	$c-2\alpha-1=-1$	$c-2\alpha-1<-1$
$2\alpha+2c-2\geq 0$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}}+\frac{1}{n^{4\alpha}}+\frac{1}{n^{c+2\alpha}})$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}}+\frac{1}{n^{4\alpha}}+\frac{1}{n^{c+2\alpha}})$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}}+\frac{\ln^2(n)}{n^{2c}}+\frac{\ln(n)}{n^{2c}})$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}})$
$-1<2\alpha+2c-2<0$	n.a.	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}}+\frac{1}{n^{4\alpha}}+\frac{1}{n^{c+2\alpha}})$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}}+\frac{\ln^2(n)}{n^{2c}}+\frac{\ln(n)}{n^{2c}})$	$O(\frac{1}{n^{2c}}+\frac{1}{n^{1-2\alpha}})$
$2\alpha+2c-2=-1$	n.a.	$O(\frac{1}{n^{2c}}+\frac{\ln(n)}{n^{2c}}+\frac{1}{n^{4\alpha}}+\frac{1}{n^{c+2\alpha}})$	$O(\frac{1}{n^{2c}}+\frac{\ln(n)}{n^{2c}}+\frac{\ln^2(n)}{n^{2c}}+\frac{\ln(n)}{n^{2c}})$	$O(\frac{1}{n^{2c}}+\frac{\ln(n)}{n^{2c}})$
$2\alpha+2c-2<-1$	n.a.	$O(\frac{1}{n^{2c}}+\frac{1}{n^{4\alpha}}+\frac{1}{n^{c+2\alpha}})$	$O(\frac{1}{n^{2c}}+\frac{\ln^2(n)}{n^{2c}}+\frac{\ln(n)}{n^{2c}})$	$O(\frac{1}{n^{2c}})$

From the table we can see that if $0<\alpha<1/2$, the MSE is $O(\frac{1}{n^{\Delta}})$ where $\Delta>0$, which implies it will converge to zero as the replication size increases.

□

Theorem 2 (Recovering optimal parameter choices and MSE in the offline setting)
Set $\gamma_n = \min\{\frac{c}{n}, 1\}$ and $\delta_n = \frac{d}{n^\alpha}$, where $c, d, \alpha > 0$. For each case below, we assume that $x_0 \pm \delta_n$ for $n \geq 1$ is in the neighborhood of continuous differentiability of f.

1. Under Assumptions 1 and 3, the MSE of IFFD is $O(\frac{1}{n^{1/2}})$ if $\alpha = \frac{1}{4}$ and $c > \frac{1}{4}$.
2. Under Assumptions 2 and 3, the MSE of ICFD is $O(\frac{1}{n^{2/3}})$ if $\alpha = \frac{1}{6}$ and $c > \frac{1}{3}$.

Sketch of Proof. We continue our proof for Theorem 1. For IFFD, we minimize the upper bound of the MSE for each scenario. Table 3 shows the respective optimal values and solutions:

Table 3: Minimum upper bound of MSE for different scenarios under IFFD.

	$c-\alpha-1 \geq 0$	$-1 < c-\alpha-1 < 0$	$c-\alpha-1 = -1$	$c-\alpha-1 < -1$
$2\alpha+2c-2 \geq 0$	$O(\frac{1}{n^{1/2}})$ with $\alpha = \frac{1}{4}, c \geq \frac{5}{4}$	$O(\frac{1}{n^{1/2}})$ with $\alpha = \frac{1}{4}, \frac{3}{4} \leq c < \frac{5}{4}$	$O(1)$ with $\alpha = \frac{1}{2}, c = \frac{1}{2}$	$O(n^\varepsilon)$ with $\alpha \to \frac{1}{2}, c \to \frac{1}{2}$
$-1 < 2\alpha+2c-2 < 0$	n.a.	$O(\frac{1}{n^{1/2}})$ with $\alpha = \frac{1}{4}, \frac{1}{4} < c < \frac{3}{4}$	$O(\frac{1}{n^{1/2-\varepsilon}})$ with $\alpha \to \frac{1}{4}, c = \alpha \to \frac{1}{4}$	$O(\frac{1}{n^{1/3-\varepsilon}})$ with $\alpha \to \frac{1}{4}, c \to \frac{1}{4}$
$2\alpha+2c-2 = -1$	n.a.	$O(\frac{\ln(n)}{n^{1/2-\varepsilon}})$ with $\alpha \to \frac{1}{4}, c = \frac{1}{2} - \alpha \to \frac{1}{4}$	$O(\frac{\ln^2(n)}{n^{1/2}})$ with $\alpha = \frac{1}{6}, c = \frac{1}{3}$	$O(\frac{\ln(n)}{n^{1/2-\varepsilon}})$ with $\alpha \to \frac{1}{4}, c = \frac{1}{2} - \alpha \to \frac{1}{4}$
$2\alpha+2c-2 < -1$	n.a.	$O(\frac{1}{n^{1/2-\varepsilon}})$ with $\alpha \to \frac{1}{4}, c \to \frac{1}{4}$	$O(\frac{\ln^2(n)}{n^{1/2-\varepsilon}})$ with $\alpha \to \frac{1}{4}, c = \alpha \to \frac{1}{4}$	$O(\frac{1}{n^{1/2-\varepsilon}})$ with $\alpha \to \frac{1}{4}, c \to \frac{1}{4}$

Here, ε is any small positive number. The symbol "$\to$" means the optimum is obtained by letting the left hand side to be very close to the right hand side.

Similarly, for ICFD, we minimize the upper bound of the MSE and obtain Table 4:

Table 4: Minimum upper bound of MSE for different scenarios under ICFD.

	$c-2\alpha-1 \geq 0$	$-1 < c-2\alpha-1 < 0$	$c-2\alpha-1 = -1$	$c-2\alpha-1 < -1$
$2\alpha+2c-2 \geq 0$	$O(\frac{1}{n^{2/3}})$ with $\alpha = \frac{1}{6}, c \geq \frac{4}{3}$	$O(\frac{1}{n^{2/3}})$ with $\alpha = \frac{1}{6}, \frac{5}{6} \leq c < \frac{4}{3}$	$O(\frac{1}{n^{1/3}})$ with $\alpha = \frac{1}{3}, c = \frac{2}{3}$	$O(\frac{1}{n^{1/3-\varepsilon}})$ with $\alpha \to \frac{1}{3}, c \to \frac{2}{3}$
$-1 < 2\alpha+2c-2 < 0$	n.a.	$O(\frac{1}{n^{2/3}})$ with $\alpha = \frac{1}{6}, \frac{1}{3} < c < \frac{5}{6}$	$O(\frac{\ln^2(n)}{n^{2/3-\varepsilon}})$ with $\alpha \to \frac{1}{6}, c = 2\alpha \to \frac{1}{6}$	$O(\frac{1}{n^{2/3-\varepsilon}})$ with $\alpha \to \frac{1}{6}, c \to \frac{1}{3}$
$2\alpha+2c-2 = -1$	n.a.	$O(\frac{\ln(n)}{n^{2/3-\varepsilon}})$ with $\alpha \to \frac{1}{6}, c = \frac{1}{2} - \alpha \to \frac{1}{3}$	$O(\frac{\ln^2(n)}{n^{2/3}})$ with $\alpha = \frac{1}{6}, c = \frac{1}{3}$	$O(\frac{\ln(n)}{n^{2/3-\varepsilon}})$ with $\alpha \to \frac{1}{6}, c = \frac{1}{2} - \alpha \to \frac{1}{3}$
$2\alpha+2c-2 < -1$	n.a.	$O(\frac{1}{n^{2/3-\varepsilon}})$ with $\alpha \to \frac{1}{6}, c \to \frac{1}{3}$	$O(\frac{\ln^2(n)}{n^{2/3-\varepsilon}})$ with $\alpha \to \frac{1}{6}, c = 2\alpha \to \frac{1}{3}$	$O(\frac{1}{n^{2/3-\varepsilon}})$ with $\alpha \to \frac{1}{6}, c \to \frac{1}{3}$

For the IFFD, if we choose $\alpha = \frac{1}{4}, c > \frac{1}{4}$, the MSE shrinks at order $O(\frac{1}{n^{1/2}})$ as n increases. For ICFD, if we choose $\alpha = \frac{1}{6}, c > \frac{1}{3}$, the MSE shrinks at order $O(\frac{1}{n^{2/3}})$ as n increases. This concludes our theorem. □

Note that Theorem 2 recovers the asymptotically optimal MSE for the forward and central finite-differences in the offline setting. This MSE is achieved by choosing the perturbation parameter with a decay against the iteration that corresponds precisely to the replication size in the offline setting (i.e., $\Theta(1/n^{1/4})$ and $\Theta(1/n^{1/6})$) and the standard step size of SA (i.e., $\Theta(1/n)$). Note that the choice of c that guarantees our various convergence behaviors does not depend on other constants. This is distinct from the SA literature where typically the c affects the convergence rate in a way dependent on the variances of the estimates.

4 NUMERICAL STUDY

We take a simple function $f(x) = \sin(x)$ for illustration and our simulation model is $Y(x) = \sin(x) + \varepsilon$ where $\varepsilon \sim N(0,1)$. Let $x_0 = 1$. We use both our iterative methods and the traditional methods for computing $f'(x) = \cos(x)$ using $Y(x)$. To approximate the MSE, in each test case we make 100 repetitions of implementation of each method. We recall that the traditional forward finite-difference (TFFD) estimator for $f'(x)$ is $\frac{\hat{f}_R(x+\delta)-\hat{f}_R(x)}{\delta}$ where $\hat{f}_R$ is an estimator using the average of R simulation replication pairs. Likewise, a traditional central finite-difference (TCFD) estimator for $f'(x)$ is $\frac{\hat{f}_R(x+\delta)-\hat{f}_R(x-\delta)}{2\delta}$. TFFD is known to achieve an optimal MSE of order $\Theta(1/R^{1/2})$ by choosing $\delta = d/R^{1/4}$ for some $d > 0$, and TCFD is known to have an optimal MSE of order $\Theta(1/R^{2/3})$ by choosing $\delta = d/R^{1/6}$ for some $d > 0$.

We do the following four experiments. In all experiments, we set $\gamma_n = \frac{c}{n}, \delta_n = \frac{d}{n^\alpha}$ for the iterative methods, and $\delta = \frac{d}{R^\alpha}$ for the traditional methods. To calculate MSE, We make the repetition number be 100.

4.1 Consistency

Figure 1 plots the trajectories of the derivative estimators using our iterative methods as the iteration goes in a single run. We set the total simulation budget as 2000 and let $c = d = 1$, $\alpha = \frac{1}{4}$ for IFFD and $\alpha = \frac{1}{6}$ for ICFD. In the figure, we see that both IFFD and ICFD exhibit a convergent sequence of estimates, validating the correctness of these methods. Moreover, ICFD shows a faster convergent trend as its trajectory is closer to the true derivative value (horizontal red line), as predicted by the convergence rates we obtained in Section 3. To further investigate the convergence rates of the iterative methods and compare with the traditional ones, Table 5 and Figure 2 shows the MSEs under the four methods with $c = d = 1$ and $\alpha = 1/6$. The total iteration number or replication size is varied among $40, 100, 200, 500, 1000, 5000, 10000$. We see that the iterative methods have consistently smaller MSEs compared to the traditional methods, when comparing within the forward approach and the central approach. While this superiority could be example-dependent, it could also hint that our iterative methods generate better constants in front of the convergence rates. Also, the central approaches have smaller MSEs than the forward approaches, again as predicted.

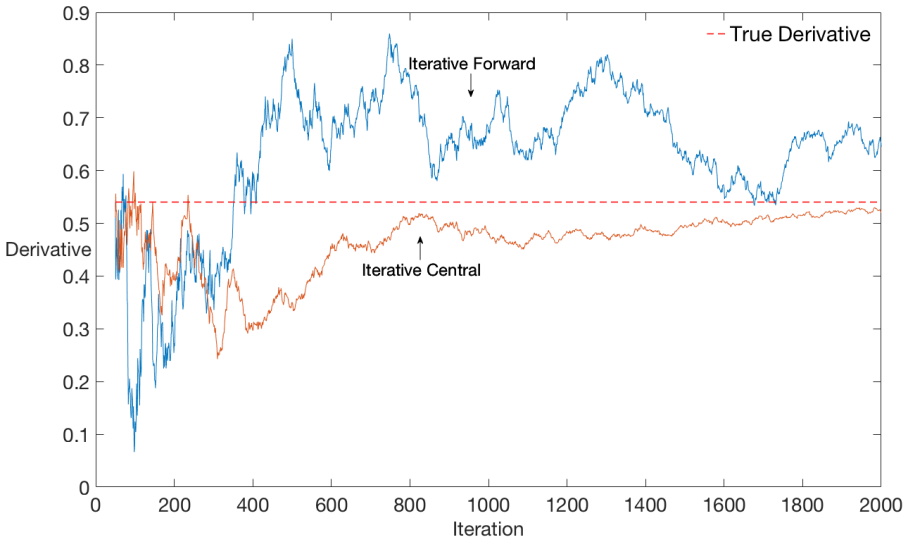


Figure 1: Derivative against iteration.

Table 5: MSE against different replication size R.

R	Iterative Forward	Iterative Central	Traditional Forward	Traditional Central
40	2.83×10^{-1}	2.57×10^{-2}	4.568×10^{-1}	2.96×10^{-2}
100	1.67×10^{-1}	1.13×10^{-2}	2.345×10^{-1}	2.54×10^{-2}
200	1.15×10^{-1}	6.76×10^{-3}	1.774×10^{-1}	1.53×10^{-2}
500	8.38×10^{-2}	4.57×10^{-3}	7.68×10^{-2}	9.20×10^{-3}
1000	5.05×10^{-2}	2.33×10^{-3}	6.73×10^{-2}	4.43×10^{-3}
5000	2.29×10^{-2}	1.16×10^{-3}	2.94×10^{-2}	1.28×10^{-3}
10000	1.78×10^{-2}	5.41×10^{-4}	2.22×10^{-2}	1.11×10^{-3}

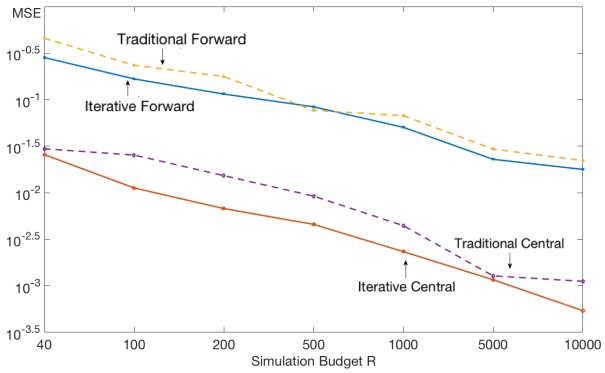


Figure 2: MSE against different replication size R.

To investigate the decay rates of the MSEs, we regress log(MSE) against $\log R$, i.e., we set up a linear model $\log(\text{MSE}) = \beta_0 + \beta_1 \log R$. The estimated coefficients for $\log R$ are $-0.504, -0.656, -0.543, -0.660$ respectively for IFFD, ICFD, TFFD and TCFD, which are very close to the predicted values $\frac{1}{2}, \frac{2}{3}, \frac{1}{2}, \frac{2}{3}$ respectively. The coefficients of determination are $0.996, 0.988, 0.985, 0.978$, indicating a strong linear dependence between log(MSE) and $\log R$.

4.2 Optimality

We analyze the MSEs under different choices of α for the four considered methods. We still let $c = d = 1$. We set the total iteration or replication size as 5×10^4, and vary α. Table 6 and Figure 3 show that the optimal choice of α, i.e., giving the smallest MSE, is $1/6$ for the central methods. The optimal choice of α is $1/4$ for iterative forward, but is a little off to $1/6$ for traditional forward. The iterative methods generally perform better than traditional. Specifically, under the same finite-difference scheme, the MSE from the iterative method is either smaller than or very close to that from the traditional method, among all cases of α even when the latter is not chosen optimally.

Table 6: MSE against α for different methods.

α	Iterative Forward	Iterative Central	Traditional Forward	Traditional Central
$1/200$	2.03×10^{-1}	6.16×10^{-3}	2.00×10^{-1}	5.97×10^{-3}
$1/20$	7.92×10^{-2}	1.12×10^{-3}	7.09×10^{-2}	9.52×10^{-4}
$1/6$	9.54×10^{-3}	2.26×10^{-4}	6.82×10^{-3}	3.45×10^{-4}
$1/4$	7.68×10^{-3}	1.34×10^{-3}	9.16×10^{-3}	2.39×10^{-3}
$1/3$	3.97×10^{-2}	8.13×10^{-3}	6.02×10^{-2}	1.43×10^{-2}
$5/12$	1.56×10^{-1}	3.74×10^{-2}	3.07×10^{-1}	9.22×10^{-2}
$1/2$	1.22×10^{0}	2.93×10^{-1}	1.54×10^{0}	4.39×10^{-1}
1	305×10^{2}	104×10^{2}	105×10^{3}	231×10^{2}

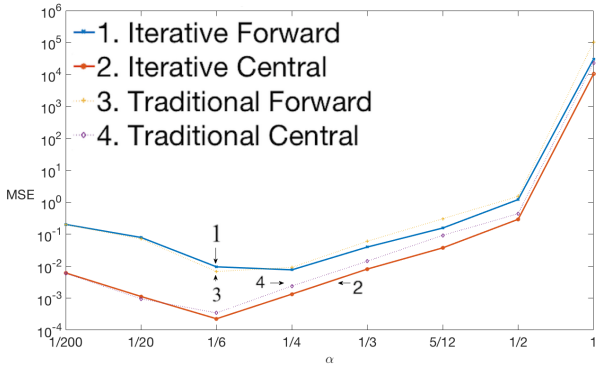


Figure 3: MSE against α for different methods.

4.3 Stability

We highlight the benefit of our iterative methods over the traditional methods when R is not known in advance. Set $d = 1$. Given a budget $R = 5000$, TFFD will set δ to be 0.119, and TCFD will set δ to be 0.242. However, when the actual R is different from 5000, these choices of δ are not optimal. To mimic this situation, we consider the case where the budgeted R is 5000, while the actual R is 100 or 2000, so that we are over-budgeted. We also consider the case where the actual R is 7500 or 10000, so that we are under-budgeted. Table 7 and Figure 4 show the MSEs when the δ for the traditional methods are set as if $R = 5000$, but the actual R can be different. We see that the MSEs from the traditional methods are significantly bigger compared with our iterative methods, in the case that the actual R deviates from 5000. This is because in such situations the δ cannot be chosen optimally in the traditional methods, whereas our iterative methods overcome this issue and maintain the optimal MSE rates.

Table 7: MSE against actual R.

R	Iterative Forward	Traditional Forward, budgeted $R = 5000$	Iterative Central	Traditional Central, budgeted $R = 5000$
100	1.65×10^{-1}	1.62×10^{0}	1.94×10^{-2}	8.56×10^{-2}
2000	2.50×10^{-2}	7.27×10^{-2}	2.07×10^{-3}	4.89×10^{-3}
5000	2.03×10^{-2}	3.53×10^{-2}	1.39×10^{-3}	1.56×10^{-3}
7500	1.96×10^{-2}	2.19×10^{-2}	9.49×10^{-4}	1.24×10^{-3}
10000	1.75×10^{-2}	1.98×10^{-2}	6.14×10^{-4}	7.53×10^{-4}

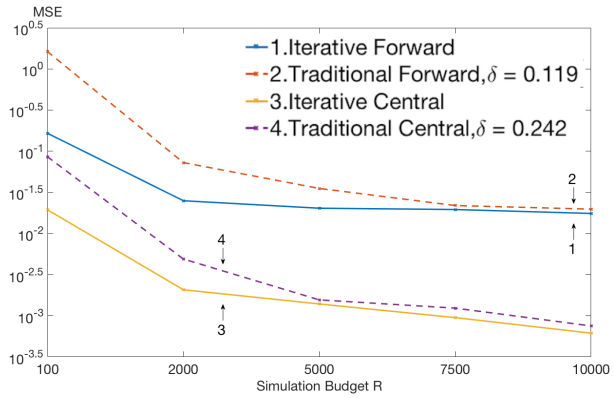
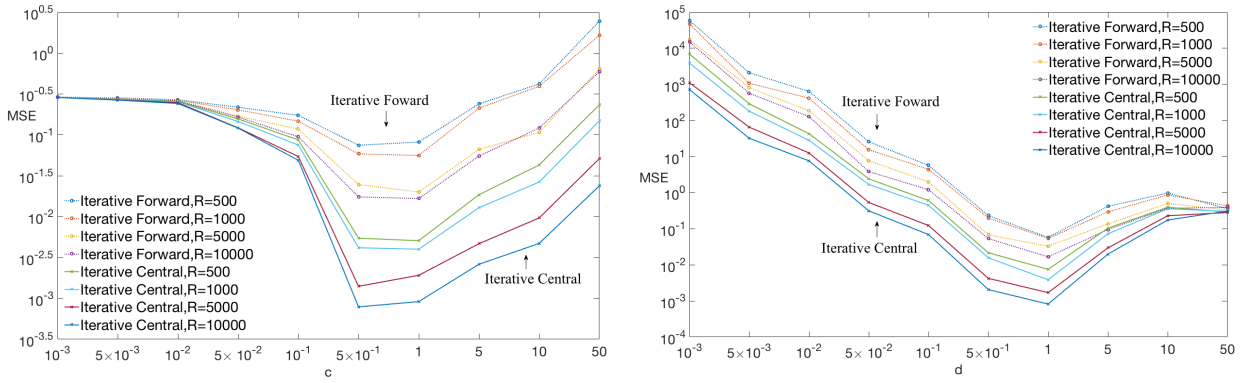


Figure 4: MSE against actual R.

4.4 Sensitivities to c and d

We investigate the sensitivities of our performances with respect to the choices of c and d. To choose these parameters optimally, we need additional information, such as the higher order derivatives of f, which is generally unavailable. We choose α optimally for each method (i.e., $1/4$ for forward and $1/6$ for central) and let the replication size vary among $500, 1000, 5000, 10000$. Figures 5a and 5b show that, as expected, moderate values of c and d give us the best MSEs. To compare the sensitivities of our iterative methods with the traditional ones, we also run the experiments on different d for the latter (note that there is no c there). Figure 6 shows that the sensitivities are quite comparable, and the iterative methods seem to consistently perform better. In this example, it appears that the optimal c is around 0.5 across the iterative methods, whereas the optimal d is around 1 for both traditional and iterative methods.



(a) MSE against different choices of c.

(b) MSE against different choices of d.

Figure 5: Sensitivities to c and d.

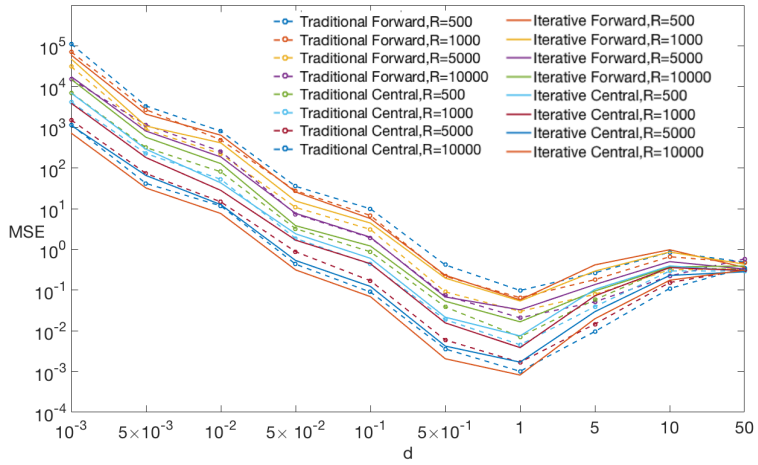


Figure 6: Sensitivity to d among iterative and traditional methods.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CMMI-1542020 and CAREER CMMI-1653339/1834710.

REFERENCES

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

Blanchet, J., D. Goldfarb, G. Iyengar, F. Li, and C. Zhou. 2017. "Unbiased Simulation for Optimizing Stochastic Function Compositions". *arXiv preprint arXiv:1711.07564*.

Blanchet, J. H., and P. W. Glynn. 2015. "Unbiased Monte Carlo for Optimization and Functions of Expectations via Multi-Level Randomization". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz et al., 3656–3667. Piscataway, New Jersey: IEEE.

Fu, M. C. 1994. "Optimization via Simulation: A Review". *Annals of Operations Research* 53(1):199–247.

Fu, M. C. 2015. "Stochastic Gradient Estimation". In *Handbook of Simulation Optimization*, 105–147. Springer.

Glasserman, P. 2013. *Monte Carlo Methods in Financial Engineering*, Volume 53. Springer Science & Business Media.

Kushner, H., and G. G. Yin. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*, Volume 35. Springer Science & Business Media.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro. 2009. "Robust Stochastic Approximation Approach to Stochastic Programming". *SIAM Journal on Optimization* 19(4):1574–1609.

Pasupathy, R., and S. Kim. 2011. "The Stochastic Root-Finding Problem: Overview, Solutions, and Open Questions". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 21(3):19:1–19:23.

AUTHOR BIOGRAPHIES

THIBAULT DUPLAY is a Master student in the Department of Industrial Engineering and Operations Research at Columbia University. He received his Diplôme d'Ingénieur degree (equivalent to a Master's Degree) in Mathematical Statistics and Probabiltiy at Ensae ParisTech in 2018. His email address is td2553@columbia.edu.

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 at 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, risk and uncertainty quantification, and stochastic optimization. His email address is khl2114@columbia.edu.

XINYU ZHANG is a Ph.D. student in the Department of Industrial Engineering and Operations Research at Columbia University. He obtained his bachelor's degree in the University of Michigan, majoring in physics and applied mathematics. His research interests include stochastic optimization and extreme event analysis. His email address is zhang.xinyu@columbia.edu.