Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

BATCHING ON BIASED ESTIMATORS

Shengyi He
Henry Lam

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

ABSTRACT

Existing batching methods are designed to cancel the variability parameter but not the bias of estimators, and thus are applied typically in the setting of unbiased estimation. We provide a batching scheme that cancel out the bias and variability parameters of estimators simultaneously, yielding asymptotically exact confidence intervals for biased estimation problems. We apply our batching method to finite difference estimators. We extend our method to the multivariate case in constructing confidence regions. We validate our theory and analyze the effect of the number of batches through numerical examples.

1 INTRODUCTION

Batching methods are widely used in simulation analysis. The basic idea of batching methods is to divide the data into batches. To construct a confidence interval (CI), these methods utilize pivotal statistics that cancel out the variability parameter by judiciously combining the batch estimates. They thus construct CIs without knowing the values of these variability parameters. These schemes are especially useful for problems where the variance of the output is hard to compute, such as quantile (Nakayama 2014) whose variance estimation involves density estimation, and in serially dependent problems and steady-state estimation (Asmussen and Glynn 2007; Nakayama 2007).

The current literature on batching methods focus on the estimation variability, but not the bias. To construct CIs for biased estimators, however, one could consider a batching strategy for the bias as well. Biased estimation arise in many places in the simulation literature, a basic example being the class of finite-difference estimators, which is widely used in stochastic gradient estimation under black-box settings where the underlying dynamics of the simulation model is inaccessible (Lam et al. 2021). For finite-difference estimators, under the arguably optimal selection of the perturbation parameter, the asymptotic distribution of the (scaled) estimation error is not centered at 0 (Glynn 1989). Moreover, the center of this asymptotic distribution depends on model characteristics that are unknown a priori, so an explicit correction for the bias may not be straightforward. In this case, classical batching methods would fail because they do not account for this bias. A bootstrap correction for the bias will not work either, since the estimator based on the resampled data has the same amount of bias as the original data. An explicit correction for the bias using some type of bias estimators is discussed in Section 4.4 of Hall (1992) for kernel density estimators (KDE). However, the bias estimator requires the order and differentiability of the kernels, which make it hard to be applied to general biased estimators.

In this paper, we propose a batching method that aims to cancel out both the bias and variance parameters simultaneously, so that it can be applied to biased estimation problems. The way we handle the bias is to take batch estimates with different batch sizes and cancel out the bias by scaling and subtracting these batch estimates. The use of different batch sizes has some resemblance with subsampling techniques (Politis

et al. (1999); Bickel et al. (1997)) where the resample size is chosen smaller than the original size in a bootstrap, but the motivation in our case is different as we aim to cancel out the bias, instead of removing smoothness conditions in a full-size bootstrap or handling serial dependence. The variability parameter, on the other hand, can be handled in a similar way as the usual batching methods. Combining them, we construct a pivotal statistic with a limiting t-distribution, which yields a CI with asymptotically exact coverage. We also extend our method to the multivariate case where we construct ellipsoidal confidence regions. We apply our method to finite-difference estimators both in the univariate and multivariate cases, where the latter is based on simultaneous perturbation.

The rest of the paper is organized as follows. Section 2 introduces biased problems. Section 3 discusses the challenges of usual batching and bootstrap methods. Section 4 proposes our batching method for biased problems. Section 5 extends our method to the multivariate case. Section 6 gives numerical examples. Section 7 concludes the paper.

2 BIASED ESTIMATION PROBLEMS

Consider an unknown target quantity ψ. Given n i.i.d. data $X_1,\ldots,X_n$, we use an estimator $\hat{\psi}_n$ that depends on the data and a tuning parameter δ_n to estimate ψ. Here, we consider biased problems in which we cannot obtain unbiased samples easily.

A basic example of biased estimators is finite-difference estimators in stochastic gradient estimation. Suppose $f(\cdot)$ is a black-box function that can only be noisily evaluated via unbiased samples. To estimate the derivative of $f(\cdot)$ at a point x_0, we would use the finite-difference estimator given by

$$\text{Forward Finite Difference (FFD): } \hat{\psi}_{n,\text{FFD}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{f}_1^{(i)}(x_0+\delta_n) - \hat{f}_2^{(i)}(x_0)}{\delta_n}$$

or

$$\text{Central Finite Difference (CFD): } \hat{\psi}_{n,\text{CFD}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{f}_1^{(i)}(x_0+\delta_n) - \hat{f}_2^{(i)}(x_0-\delta_n)}{2\delta_n}$$

where $\hat{f}_1^{(i)}(\cdot)$ and $\hat{f}_2^{(i)}(\cdot), i = 1,2,\ldots,n$ are two independent noisy runs of f, and δ_n is a perturbation parameter.

For biased problems such as the finite-difference estimators depicted above, we typically have a central limit theorem (CLT)

$$n^{\beta}\left(\hat{\psi}_n - \psi\right) \Rightarrow N\left(\mu, \sigma^2\right) \tag{1}$$

for some β, when δ_n is appropriately chosen in terms of n (Fox and Glynn 1989). Here, unlike standard CLT, because of the bias we have a non-zero-mean normal limit and β could be smaller than the canonical rate of $1/2$. For example, for FFD and CFD, under enough smoothness conditions, we have that (Glynn 1989)

$$n^{1/4}\left(\hat{\psi}_{n,\text{FFD}} - f'(x_0)\right) \Rightarrow N\left(-\frac{\delta f''(x_0)}{2}, \frac{2\text{Var}\hat{f}_2(x_0)}{\delta^2}\right), \text{ if } n^{1/4}\delta_n \to \delta \tag{2}$$

by choosing δ_n of order $n^{-1/4}$ and

$$n^{1/3}\left(\hat{\psi}_{n,\text{CFD}} - f'(x_0)\right) \Rightarrow N\left(-\frac{\delta^2 f'''(x_0)}{3}, \frac{2\text{Var}\hat{f}_2(x_0)}{\delta^2}\right), \text{ if } n^{1/6}\delta_n \to \delta \tag{3}$$

by choosing δ_n of order $n^{-1/6}$. The convergence rates in (2) and (3) are optimal for the respective classes of estimators. That is, if we choose δ_n at different rates than depicted, then the convergence rates would not improve. On a high level, this is because the chosen orders of δ_n balance the bias and variance in the estimation, and any distortions would increase either bias or variance that leads to a larger magnitude of overall error.

Our goal is to construct a CI for ψ by using (1). In many problems such as finite differences as shown above, it is difficult to estimate σ^2 and μ and thus a direct use of the CLT in CI construction. Moreover, as we will show, classical batching methods and bootstrap methods also face challenges.

3 CHALLENGES IN STANDARD BATCHING AND BOOTSTRAP METHODS

There are several variants of batching methods depending on the construction of the point estimator and the scheme in aggregating different batches. The ideas of these methods are similar in the sense that they all aim to cancel out the variability of the point estimator via dividing by a quantity that represents the variability among the batch estimates. We discuss one of the variant in detail here and similar discussion will also hold for other variants of batching methods.

One common batching method constructs a level $1-\alpha$ CI as

$$\left(\bar{\psi} \pm t_{m-1,1-\alpha/2}\frac{S_{\text{batch}}}{\sqrt{m}}\right) \tag{4}$$

where

$$\bar{\psi} = \frac{1}{m}\sum_{i=1}^{m}\hat{\psi}_n^{(i)},$$

$$S_{\text{batch}}^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\psi}_n^{(i)} - \bar{\psi}\right)^2,$$

$\hat{\psi}_n^{(i)}$ are independent estimators of ψ, and $t_{m-1,1-\alpha/2}$ is the $1-\alpha$ quantile of t distribution with $m-1$ degrees of freedom. This CI is based on the observation that when $\mu = 0$ in (1), we have that

$$\frac{\sqrt{m}(\bar{\psi}-\psi)}{\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\psi}_n^{(i)} - \bar{\psi}\right)^2}} \Rightarrow t_{m-1} \tag{5}$$

as $n \to \infty$. In (5), the asymptotic distribution does not depend on σ since it is canceled out by taking the ratio. However, if μ deviates from 0, the asymptotic distribution of the above statistic would change since μ is not canceled out. Therefore, the CI given in (4) is not valid when $\mu \neq 0$.

Bootstrap methods also have issues due to the bias. Suppose we want to use bootstrap to approximate the distribution of $n^{\beta}(\hat{\psi}_n - \psi)$ where $\hat{\psi}_n$ is calculated based on i.i.d. data $X_1, \ldots, X_n$. The general principle of bootstrap suggests to use the distribution of

$$n^{\beta}(\psi_n^* - \hat{\psi}_n)$$

conditional on $X_1, \ldots, X_n$, where ψ_n^* is the estimate based on resampled data $\{X_1^*, \ldots, X_n^*\}$, sampled with replacement from $\{X_1, \ldots, X_n\}$. In the finite difference setting (say FFD), we have that

$$\hat{\psi}_{n,\text{FFD}} = \mathbb{E}_{\hat{P}}[X]$$

where

$$\hat{P} := \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$$

and X_i is the i-th estimator:

$$X_i = \frac{\hat{f}_1^{(i)}(x_0+\delta_n) - \hat{f}_2^{(i)}(x_0)}{\delta_n}.$$

Here δ_{X_i} is point mass at X_i. So the estimator based on resampled data is given by

$$\psi^*_{n,\mathrm{FFD}} = \mathbb{E}_{P^*}[X]$$

where

$$P^* := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i^*}.$$

From the linearity of $\psi^*_{n,\mathrm{FFD}}$ as a function of P^*, it is clear that

$$\mathbb{E}\left[n^\beta\left(\psi^*_{n,\mathrm{FFD}} - \hat{\psi}_{n,\mathrm{FFD}}\right)|\{X_1,\ldots,X_n\}\right] = 0.$$

But on the other hand, from (2) we know that (under some conditions, e.g., uniform integrability)

$$\mathbb{E}\left[n^\beta\left(\hat{\psi}_{n,\mathrm{FFD}} - \psi\right)\right] \to -\frac{\delta f''(x_0)}{2} \neq 0.$$

Comparing the above two displayed relations, we conclude that the distirbution of $n^\beta\left(\psi^*_{n,\mathrm{FFD}} - \hat{\psi}_{n,\mathrm{FFD}}\right)$ conditional on the data cannot approximate the true distribution of $n^\beta\left(\hat{\psi}_{n,\mathrm{FFD}} - \psi\right)$.

Essentially, the reason that the usual bootstrap cannot work is that any estimator with the same choice of δ_n will share the same amount of bias. Therefore, comparisons between the original estimator and the estimator based on resampled data will not give any information about the bias. In Section 4.4.2 of Hall (1992), a similar issue is discussed for KDE, which is a related type of biased problems whose amount of bias is determined by the bandwith. Hall (1992) proposes an explicit correction for the bias, but the proposed formula for the correction is involved and depends on properties of the problem, such as the order of the kernel and the differentiability of the kernel function, which makes it hard to implement.

The above challenges faced by the standard batching and bootstrap suggests that we look for ways to cancel out the bias and variance simultaneously. An idea is to construct batch estimators with different amount of bias so that information on the bias can be obtained. This motivates the batching method that we will describe next.

4 A BATCHING METHOD FOR BIASED ESTIMATION

We consider a modified batching approach that handles the bias and variability of biased estimation simultaneously. The proposed method is given in Algorithm 1. In Algorithm 1, we first construct a new quantity $\nu^{(j)}_{n,s} = n^\beta\hat{\psi}_n - s^\beta\hat{\psi}_s$ which is governed by a (scaled) unbiased CLT. To explain, notice that replacing n with s in (1), we also have that

$$s^\beta\left(\hat{\psi}_s - \psi\right) \Rightarrow N\left(\mu, \sigma^2\right) \tag{6}$$

as $s \to \infty$. By subtracting (1) by (6), we have that for independent $\hat{\psi}_n$ and $\hat{\psi}_s$,

$$\left(n^\beta\hat{\psi}_n - s^\beta\hat{\psi}_s\right) - \left(n^\beta - s^\beta\right)\psi \Rightarrow N\left(0, 2\sigma^2\right). \tag{7}$$

as $n \to \infty, s \to \infty$. Therefore, up to a scaling of $n^\beta - s^\beta$, $\nu^{(j)}_{n,s}$ is governed by a CLT with mean zero in the limit. The rest of the steps in Algorithm 1 handle the variance using the variability among $\{\nu^{(j)}_{n,s}\}_{j=1,2,\ldots,m}$ in a similar way as usual batching. We have the following theorem regarding the asymptotic exactness of the output of Algorithm 1.

Theorem 1 Suppose that (1) holds. Then in Algorithm 1, we have that for any $\alpha \in (0,1)$, and $\mathscr{I} = \mathscr{I}_\alpha, \mathscr{I}_{\alpha,\mathrm{lower}}$ or $\mathscr{I}_{\alpha,\mathrm{upper}}$,

$$P\left(\psi \in \mathscr{I}\right) \to 1-\alpha$$

as $n, s \to \infty$.

Algorithm 1 Batching for biased estimation

Require: nominal level $1-\alpha$, batch sizes $n \neq s$, number of batches m, the value of β in (1)

for $j = 1, 2, \ldots, m$ (independently) **do**

$\hat{\psi}_n^{(j)} \leftarrow$ an estimator based on n i.i.d. samples and tuning parameter δ_n

$\hat{\psi}_s^{(j)} \leftarrow$ an estimator based on another s i.i.d. samples and tuning parameter δ_s

$\nu_{n,s}^{(j)} \leftarrow n^\beta \hat{\psi}_n^{(j)} - s^\beta \hat{\psi}_s^{(j)}$

end for

$\bar{\nu} \leftarrow \frac{1}{m}\sum_{j=1}^m \nu_{n,s}^{(j)}$

$S^2 \leftarrow \frac{1}{m-1}\sum_{j=1}^m \left(\nu_{n,s}^{(j)} - \bar{\nu}\right)^2$

$\mathscr{I}_\alpha \leftarrow \left[\left(\bar{\nu} - t_{m-1,1-\alpha/2}\frac{S}{\sqrt{m}}\right)/\left(n^\beta - s^\beta\right), \left(\bar{\nu} + t_{m-1,1-\alpha/2}\frac{S}{\sqrt{m}}\right)/\left(n^\beta - s^\beta\right)\right]$

$\mathscr{I}_{\alpha,\text{lower}} \leftarrow \left(-\infty, \left(\bar{\nu} + t_{m-1,1-\alpha}\frac{S}{\sqrt{m}}\right)/\left(n^\beta - s^\beta\right)\right]$

$\mathscr{I}_{\alpha,\text{upper}} \leftarrow \left[\left(\bar{\nu} - t_{m-1,1-\alpha}\frac{S}{\sqrt{m}}\right)/\left(n^\beta - s^\beta\right), \infty\right)$

return Two sided CI $\mathscr{I}_{\alpha,\text{lower}}$ and one-sided CIs $\mathscr{I}_{\alpha,\text{lower}}, \mathscr{I}_{\alpha,\text{upper}}$.

Proof. From the construction of $\nu_{n,s}^{(j)}$ and the observation (7), we have that

$$\left(\nu_{n,s}^{(j)} - \left(n^\beta - s^\beta\right)\psi\right)_{j=1,\ldots,m} \Rightarrow (Z_j)_{j=1,\ldots,m} \tag{8}$$

where $Z_j \overset{i.i.d}{\sim} N\left(0, 2\sigma^2\right)$. Therefore, we have that

$$\frac{\bar{\nu} - \left(n^\beta - s^\beta\right)\psi}{S/\sqrt{m}} \Rightarrow \frac{\bar{Z}}{S_Z/\sqrt{m}} \overset{d}{=} t_{m-1} \tag{9}$$

where $\bar{Z} = (1/m)\sum_{j=1}^m Z_j$ and $S_Z^2 = (1/(m-1))\sum_{j=1}^m (Z_j - \bar{Z})^2$, and t_{m-1} is a t-distribution with degree of freedom $m-1$. The convergence in distribution follows from (8) and the continuous mapping theorem, and the equality in distribution $\overset{d}{=}$ follows from elementary properties of independent normal variables.

By inverting (9), we derive the desired result. □

The use of the batch size s that is smaller than n has some resemblance with subsampling techniques (Politis et al. (1999); Bickel et al. (1997)) where the resample size in a bootstrap is chosen smaller than the full data size. However, the motivation here is different from that literature. While they use subsampling to remove smoothness conditions required by full-size bootstraps or to handle serial dependence, we use different batch sizes in order to combine these batches in a way to eliminate the bias. Suppose further that $s = \rho n$, i.e., ρ is the subsample factor of s with respect to n. In this case, the two-sided CI in Algorithm 1 can be written as

$$\frac{1}{1-\rho^\alpha}\left(\bar{Y} \pm t_{m-1,1-\alpha/2}\frac{S_Y}{\sqrt{m}}\right)$$

where $Y_j = \psi_n^{(j)} - \rho^\alpha \psi_s^{(j)}$, and $\bar{Y} = (1/m)\sum_{j=1}^m Y_j$ and $S_Y^2 = (1/(m-1))\sum_{j=1}^m (Y_j - \bar{Y})^2$.

5 GENERALIZATION TO THE MULTIVARIATE CASE

The analysis in Section 2 can be extended to the multivariate case, where we want a confidence region for multiple biased estimators. Suppose that we are interested in a multidimensional unknown target quantity $\Psi \in \mathbb{R}^d$ and have biased estimators given by $\hat{\Psi}_n$ which depend on a tuning parameter δ_n. In this case, we will start from a multivariate analog of (1):

$$n^\beta(\hat{\Psi}_n - \Psi) \Rightarrow N(\mu, \Sigma) \tag{10}$$

where $\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}$.

For example, for the gradient estimation of a performance function w.r.t. multiple parameters, simultaneous perturbation finite difference (SPFD) can be employed (Spall 1992). Suppose $f(\cdot): \mathbb{R}^d \to \mathbb{R}$ can only be noisily evalutaed and we want to estimate the gradient of f at $x_0 \in \mathbb{R}^d$. Then we would take a random vector $h \in \mathbb{R}^d$ and then for each of $i = 1,2,\ldots,d$, an estimator for $\nabla f(x_0)_i$ is given by

$$\frac{\hat{f}_1(x_0+\delta_n h) - \hat{f}_2(x_0-\delta_n h)}{2\delta_n h_i} \tag{11}$$

Therefore, with n simulation runs, the estimator is given by

$$\hat{\Psi}_{n,\text{SPFD}} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\hat{f}_1^{(j)}(x_0+\delta_n h^{(j)}) - \hat{f}_2^{(j)}(x_0-\delta_n h^{(j)})}{2\delta_n h_i^{(j)}}\right)_{i=1,2,\ldots,d} \tag{12}$$

where $\hat{f}_1^{(j)}(\cdot)$ and $\hat{f}_2^{(j)}(\cdot), j = 1,2,\ldots,n$ are two independent noisy runs of f, $h^{(j)}, j = 1,2,\ldots,n$ are independent samples with the same distribution as h, and δ_n is a (deterministic) perturbation parameter. As discussed in Lam et al. (2019), if we assume that h has mean zero, the components of h are independent, each component of h has finite inverse second moment and smoothness conditions hold for f, then the bias and variance of the estimator given in (11) are of orders δ_n^2 and δ_n^{-2} respectively, which is the same as CFD. Therefore, when δ_n is of order $n^{-1/6}$, following the general result provided in Theorem 2 of Fox and Glynn (1989) and using Cramér–Wold device, we have that under regularity conditions,

$$n^{1/3}(\hat{\Psi}_{n,\text{SPFD}} - \nabla f(x_0)) \Rightarrow N(\mu,\Sigma)$$

for some $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$.

For the multivariate case, an analog of Algorithm 1 is provided in Algorithm 2. In Algorithm 2, the way we handle the bias and variance is similar to the univariate case, which yields a statistic with asymptotic Hotelling T^2 distribution.

Algorithm 2 Batching for biased estimation (the multivariate version)

Require: nominal level $1-\alpha$, batch sizes $n \neq s$, number of batches m, the value of β in (10), dimension d

for $j = 1,2,\ldots,m$ (independently) **do**

$\hat{\Psi}_n^{(j)} \leftarrow$ an estimator based on n i.i.d. samples and tuning parameter δ_n

$\hat{\Psi}_s^{(j)} \leftarrow$ an estimator based on another s i.i.d. samples and tuning parameter δ_s

$V_{n,s}^{(j)} \leftarrow n^\beta \hat{\Psi}_n^{(j)} - s^\beta \hat{\Psi}_s^{(j)}$

end for

$\bar{V} \leftarrow \frac{1}{m}\sum_{j=1}^m V_{n,s}^{(j)}$

$\mathbf{S} \leftarrow \frac{1}{m-1}\sum_{j=1}^m \left(V_{n,s}^{(j)} - \bar{V}\right)\left(V_{n,s}^{j} - \bar{V}\right)^\top$

$\mathscr{I}_{\alpha,d} \leftarrow \left\{y \in \mathbb{R}^d : m\left(\bar{V} - (n^\beta - s^\beta)y\right)^\top \mathbf{S}^{-1}\left(\bar{V} - (n^\beta - s^\beta)y\right) \le T^2_{d,m-1,1-\alpha}\right\}$ (Here, $T^2_{d,m-1,1-\alpha}$ the $1-\alpha$ quantile of Hotelling T^2 with dimension d and $m-1$ degrees of freedom.)

return Confidence region $\mathscr{I}_{\alpha,d}$

For Algorithm 2, we have that

Theorem 2 Suppose that (10) holds. Moreover, suppose that $m \ge d+1$. Then in Algorithm 2,

$$P(\Psi \in \mathscr{I}_{\alpha,d}) \to 1-\alpha$$

as $n, s \to \infty$.

Proof. Similar to the one-dimensional case, from (1) and

$$n^{\beta}(\hat{\Psi}_n - \Psi) \Rightarrow N(\mu, \Sigma),$$

we have that

$$\left(V_{n,s}^{(j)} - \left(n^{\beta} - s^{\beta}\Psi\right)\right) \Rightarrow (\mathbf{Z}_j)_{j=1,2,\ldots,m}$$

where $\mathbf{Z}_j \sim N(0, 2\Sigma)$ are independent for each $j = 1, 2, \ldots, m$. Therefore,

$$m\left(\bar{V} - (n^{\beta} - s^{\beta})y\right)^{\top} \mathbf{S}^{-1} \left(\bar{V} - (n^{\beta} - s^{\beta})y\right) \Rightarrow m\bar{\mathbf{Z}}^{\top}\mathbf{S}_{\mathbf{Z}}^{-1}\bar{\mathbf{Z}}$$

where $\bar{\mathbf{Z}} = \frac{1}{n}\sum_{j=1}^{m}\mathbf{Z}_j$ and

$$\mathbf{S}_{\mathbf{Z}} = \frac{1}{m-1}\sum_{j=1}^{m}\left(\mathbf{Z}_j - \bar{\mathbf{Z}}\right)\left(\mathbf{Z}_j - \bar{\mathbf{Z}}\right)^{\top}.$$

Note that $m\bar{\mathbf{Z}}^{\top}\mathbf{S}_{\mathbf{Z}}^{-1}\bar{\mathbf{Z}} \stackrel{d}{=} T^2_{d,m-1}$. This gives the desired result. □

6 NUMERICAL EXAMPLES

6.1 A Toy Example

To illustrate the effectiveness of the method, we study the problem of constructing a 95% CI for the gradient of

$$f(\theta) := \theta + \theta^2 + \sin\theta.$$

at $\theta = 0$. Suppose we can only generate estimators $\hat{f}(\theta) = f(\theta) + \varepsilon$ where the noise ε is a standard normal. We consider both FFD and CFD. For FFD, we set $\delta_n = n^{-1/4}$. For CFD, we set $\delta_n = n^{-1/6}$. We define the total simulation budget as $N = m(s+n)$. We let $\rho = 1/2$. We test $m = 2, 5, 10, 15, 20$ and $N = 900, 9000, 90000, 900000$. Note that once N and m are specified, s and n are determined accordingly by the constraint for the simulation budget and the choice of ρ. We replicate the experiments 1000 times to estimate the coverage probabilities and the half length of the CIs.

The results for FFD and CFD are shown in Table 1 and Table 2 respectively. We can see that for both FFD and CFD, the empirical coverages are close to the nominal level. Indeed, all of the empirical coverages reported in Tables 1 and 2 are within the range 95%±2%. The difference in their performances appears in the half length. For a fixed simulation budget, when m increases, the average half length decreases. For example, in Table 1, when the simulation budget is fixed as 900 and m increases from 2 to 30, the average half length reduces from 22.6 to 2.2. A similar observation also holds for CFD. This can also be theoretically justified. Indeed, from the form of the CI given in Algorithm 1, we can see that the half length is of order $(\sqrt{m}(n^{\beta} - s^{\beta}))^{-1}$. Since $\beta < 1/2$, the effect of increasing m supersedes the effect of decreasing n, s when $m(n+s)$ and s/n are fixed. The half length also decreases as the simulation budget increases, but at a slow rate. For example, for FFD, when the simulation budget increases from 900 to 900000 (10^3 times larger), the average half length (when $m = 10$) only reduces by 81% (1-0.6/3.2). The half length for CFD has a faster decrease rate than FFD (which reduces by 90% when the simulation budget increase from 900 to 900000), but the amount of decrease in half length is still small compared to the increase in the simulation budget. Essentially, this is due to the slow rate in the CLT result (1). From the above analysis, and also suggested by the order of the half length, we can see that increasing m or increasing m and the simulation budget simultaneously is more effective than just increasing the simulation budget, if we want to shorten the length of the CI. Indeed, when N and m increases proportionally such that n and s are fixed, the order of length given by $(\sqrt{m}(n^{\beta} - s^{\beta}))^{-1}$ will decrease at rate $N^{-1/2}$. On the other hand, if we fix m and increase N, the rate is only $N^{-\beta}$. Lastly, comparing CFD and FD, we find that the half length for CFD is significantly smaller than the half length of FFD. For example, when the simulation budget is 9000 and $m = 10$, the lengths for FFD and CFD are 1.8 and 0.4 respectively. This is due to the fact that CFD has a faster rate of convergence than FFD in the CLTs (2) and (3).

Table 1: Empirical coverages and average half lengths for FFD in the toy example. The standard errors are shown in parentheses.

simulation budget	m	empirical coverage	average half length
900	2	95.9% (0.6%)	22.6 (16.7)
	5	95.3% (0.7%)	4.4 (1.6)
	10	95.9% (0.6%)	3.2 (0.8)
	20	95.8% (0.6%)	2.5 (0.4)
	30	94.6% (0.7%)	2.2 (0.3)
9000	2	95.0% (0.7%)	12.7 (9.1)
	5	95.6% (0.6%)	2.5 (0.9)
	10	94.8% (0.7%)	1.8 (0.4)
	20	95.0% (0.7%)	1.4 (0.2)
	30	96.4% (0.6%)	1.2 (0.2)
90000	2	93.8% (0.8%)	7.0 (5.4)
	5	94.0% (0.8%)	1.4 (0.5)
	10	95.7% (0.6%)	1.0 (0.2)
	20	93.6% (0.8%)	0.8 (0.1)
	30	94.8% (0.7%)	0.7 (0.1)
900000	2	94.7% (0.7%)	3.9 (3.0)
	5	95.3% (0.7%)	0.8 (0.3)
	10	95.2% (0.7%)	0.6 (0.1)
	20	94.3% (0.7%)	0.4 (0.1)
	30	94.8% (0.7%)	0.4 (0.1)

6.2 Sensitivity Analysis for an M/M/1 Queue

We consider the example in Lam et al. (2019). For an M/M/1 queue, we define the performance function as the expected value of the average system time of the first 10 customers. The parameters are the arrival rate λ and the service rate μ. Our goal is to construct 95% CIs for the gradients of the performance w.r.t. λ and μ.

First, we construct 95% CIs for the two gradients separately. We focus on CFD with $\delta_n = n^{-1/6}$. As in the toy example, the total simulation budget is defined as $N = m(s+n)$ and $\rho = 1/2$. We test $m = 2,5,10,20,30$ and $N = 900,1800,3600$. We use the true values for the gradients given in Lam et al. (2019). The estimated empirical coverage probabilities and half length are reported in Table 3. From Table 3, we have similar observations as in the toy example: all of the empirical coverages are close to the nominal level, the half length of the CI decreases as simulation budget increases or as m increases, but for a fixed m, the rate that the half length decreases as the simulation budget increases is slow.

Next, we study the 95% confidence region for the whole gradient vector w.r.t. λ and μ. We use SPFD and choose the distribution of h as (h_1, h_2) where h_1 and h_2 are independent taking 1 and -1 with equal probability. We choose $\delta_n = n^{-1/6}$ as in CFD. We test $m = 5,10,20,30$ and $N = 900,1800,3600$ (note that we can not take $m = 2$ as in the one-dimensional case since we need $m \geq d+1$ in Theorem 2).

The empirical coverages and average areas of the confidence regions are reported in Table 4. Again, we observe that the empirical coverages are close to the nominal level. The average area decreases faster as the simulation budget increases or as m increases than the one-dimensional case. For example, when $m = 30$ and the simulation budget increases from 900 to 3600, the average area becomes 33% (0.015/0.045) of the original average area. On the other hand, in Table 3, the half length for the gradient w.r.t. μ becomes 61% (0.14/0.23) of the original for the same change of simulation budget. This is because the order of the area of the confidence region is the square of the order of the half lengths.

Table 2: Empirical coverages and average half lengths for CFD in the toy example. The standard errors are shown in parentheses.

simulation budget	m	empirical coverage	average half length
900	2	95.2% (0.7%)	5.18 (3.72)
	5	94.1% (0.7%)	1.15 (0.42)
	10	94.7% (0.7%)	0.86 (0.21)
	20	95.4% (0.7%)	0.72 (0.12)
	30	94.4% (0.7%)	0.66 (0.09)
9000	2	95.1% (0.7%)	2.50 (1.86)
	5	94.8% (0.7%)	0.53 (0.20)
	10	95.3% (0.7%)	0.40 (0.09)
	20	93.4% (0.8%)	0.33 (0.05)
	30	95.3% (0.7%)	0.31 (0.04)
90000	2	94.8% (0.7%)	1.11 (0.85)
	5	94.3% (0.7%)	0.25 (0.09)
	10	94.6% (0.7%)	0.19 (0.04)
	20	95.6% (0.6%)	0.15 (0.03)
	30	94.9% (0.7%)	0.14 (0.02)
900000	2	94.9% (0.7%)	0.52 (0.38)
	5	95.6% (0.6%)	0.12 (0.04)
	10	94.9% (0.7%)	0.09 (0.02)
	20	95.2% (0.7%)	0.07 (0.01)
	30	94.7% (0.7%)	0.07 (0.01)

Table 3: Empirical coverages and average half lengths for CFD in the M/M/1 queueing example. Columns with (λ), (μ) correspond to gradients w.r.t. λ and μ, respectively. The standard errors are shown in parentheses.

simulation budget	m	empirical coverage (λ)	average half length (λ)	empirical coverage (μ)	average half length (μ)
900	2	94.9% (0.7%)	1.67 (1.28)	94.8% (0.7%)	1.83 (1.33)
	5	95.4% (0.7%)	0.38 (0.13)	94.2% (0.7%)	0.39 (0.15)
	10	95.0% (0.7%)	0.28 (0.07)	93.2% (0.8%)	0.29 (0.07)
	20	94.0% (0.8%)	0.23 (0.04)	95.5% (0.7%)	0.25 (0.04)
	30	94.6% (0.7%)	0.22 (0.03)	94.8% (0.7%)	0.23 (0.03)
1800	2	94.2% (0.7%)	1.34 (1.05)	94.7% (0.7%)	1.40 (1.06)
	5	95.1% (0.7%)	0.30 (0.11)	96.1% (0.6%)	0.31 (0.11)
	10	95.0% (0.7%)	0.22 (0.05)	94.8% (0.7%)	0.24 (0.06)
	20	94.5% (0.7%)	0.19 (0.03)	93.6% (0.8%)	0.20 (0.03)
	30	94.9% (0.7%)	0.17 (0.02)	96.3% (0.6%)	0.18 (0.02)
3600	2	95.1% (0.7%)	1.07 (0.81)	95.3% (0.7%)	1.13 (0.86)
	5	94.4% (0.7%)	0.23 (0.08)	95.1% (0.7%)	0.25 (0.09)
	10	94.6% (0.7%)	0.18 (0.04)	96.2% (0.6%)	0.18 (0.04)
	20	95.1% (0.7%)	0.15 (0.02)	96.2% (0.6%)	0.15 (0.03)
	30	93.9% (0.8%)	0.14 (0.02)	95.1% (0.7%)	0.14 (0.02)

Table 4: Empirical coverages and average areas for SPFD in the M/M/1 queueing example. The standard errors are shown in parentheses.

simulation budget	m	empirical coverage	average area
900	5	94.7% (0.7%)	0.191 (0.252)
	10	95.4% (0.7%)	0.076 (0.059)
	20	95.1% (0.7%)	0.050 (0.024)
	30	94.7% (0.7%)	0.045 (0.019)
1800	5	95.0% (0.7%)	0.115 (0.136)
	10	95.1% (0.7%)	0.042 (0.029)
	20	94.6% (0.7%)	0.030 (0.015)
	30	94.2% (0.7%)	0.025 (0.010)
3600	5	94.7% (0.7%)	0.063 (0.070)
	10	96.2% (0.6%)	0.026 (0.018)
	20	94.4% (0.7%)	0.017 (0.008)
	30	93.9% (0.8%)	0.015 (0.006)

7 CONCLUSION

This paper provides a batching scheme for biased estimators which cancels the bias and variability of estimators simultaneously. We prove that the proposed batching scheme gives asymptotically exact CIs and can be generalized to the multivariate case. A toy example and an example on the sensitivity analysis for an M/M/1 queue are provided to support the theory.

ACKNOWLEDGMENTS

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

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Steady-State Simulation*, 96–125. New York, NY: Springer New York.

Bickel, P. J., F. Götze, and W. R. van Zwet. 1997. "Resampling Fewer Than n Observations: Gains, Losses, and Remedies for Losses". *Statistica Sinica* 7(1):1–31.

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

Glynn, P. 1989. "Optimization of Stochastic Systems via Simulation". In *1989 Winter Simulation Conference Proceedings*, edited by E. A. MacNair, K. J. Musselman, and P. Heidelberger, 90–105. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hall, P. 1992. *The Bootstrap and Edgeworth Expansion*. New York, NY: Springer New York.

Lam, H., H. Li, and X. Zhang. 2021. "Minimax Efficient Finite-Difference Stochastic Gradient Estimators Using Black-Box Function Evaluations". *Operations Research Letters* 49(1):40–47.

Lam, H., X. Zhang, and X. Zhang. 2022. "Enhanced Balancing of Bias-Variance Tradeoff in Stochastic Estimation: A Minimax Perspective". *Operations Research*. https://pubsonline.informs.org/doi/abs/10.1287/opre.2022.2319.

Nakayama, M. K. 2007. "Fixed-width Multiple-comparison Procedures Using Common Random Numbers for Steady-state Simulations". *European Journal of Operational Research* 182(3):1330–1349.

Nakayama, M. K. 2014, November. "Confidence Intervals for Quantiles Using Sectioning When Applying Variance-Reduction Techniques". *ACM Transactions on Modeling and Computer Simulation* 24(4):1–21.

Politis, D., D. Wolf, J. Romano, M. Wolf, P. Bickel, P. Diggle, and S. Fienberg. 1999. *Subsampling*. Springer Series in Statistics. New York, NY: Springer New York.

Spall, J. 1992. "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation". *IEEE Transactions on Automatic Control* 37(3):332–341.

AUTHOR BIOGRAPHIES

SHENGYI HE is a PhD student in the Department of Industrial Engineering and Operations Research at Columbia University. He received his B.S. degree in statistics from Peking University in 2019. His research interests include variance reduction and uncertainty quantification via stochastic and robust optimization. His email address is sh3972@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 from Harvard University in 2011. His research interests include efficient methodologies and statistical uncertainty quantification for Monte Carlo computation, predictive modeling and data-driven optimization. His email address is khl2114@columbia.edu. His website is http://www.columbia.edu/ khl2114/.