Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Enhanced Balancing of Bias-Variance Tradeoff in Stochastic Estimation: A Minimax Perspective

Henry Lam
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, khl2114@columbia.edu

Xinyu Zhang
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, zhang.xinyu@columbia.edu

Xuhui Zhang
School for the Gifted Young, University of Science and Technology of China, zxh1998@mail.ustc.edu.cn

Biased stochastic estimators, such as finite-differences for noisy gradient estimation, often contain parameters that need to be properly chosen to balance impacts from the bias and the variance. While the optimal order of these parameters in terms of the simulation budget can be readily established, the precise best values depend on model characteristics that are typically unknown in advance. We introduce a framework to construct new classes of estimators, based on judicious combinations of simulation runs on sequences of tuning parameter values, such that the estimators consistently outperform a given tuning parameter choice in the conventional approach, regardless of the unknown model characteristics. We argue the outperformance via what we call the asymptotic minimax risk ratio, obtained by minimizing the worst-case asymptotic ratio between the mean square errors of our estimators and the conventional one, where the worst case is over any possible values of the model unknowns. In particular, when the minimax ratio is less than 1, the calibrated estimator is guaranteed to perform better asymptotically. We identify this minimax ratio for general classes of weighted estimators, and the regimes where this ratio is less than 1. Moreover, we show that the best weighting scheme is characterized by a sum of two components with distinct decay rates. We explain how this arises from bias-variance balancing that combats the adversarial selection of the model constants, which can be analyzed via a tractable reformulation of a non-convex optimization problem.

Key words: bias-variance tradeoff, minimax analysis, stochastic estimation, finite difference, robust optimization

1. Introduction

This paper studies biased stochastic estimators which, in the simplest form, are expressed as follows. To estimate a target quantity of interest $\theta_0 \in \mathbb{R}$, we use Monte Carlo simulation where each simulation run outputs

$$\theta(\delta) = \theta_0 + b(\delta) + v(\delta) \tag{1}$$

Here $v(\delta)$ represents the noise of the simulation and satisfies $E[v(\delta)] = 0$, and $b(\delta)$ is the bias given by $E[\theta(\delta)] - \theta_0$. We obtain the final estimate by averaging n independent runs produced by (1):

$$\frac{1}{n}\sum_{j=1}^{n}\theta_j(\delta) \tag{2}$$

where $\theta_j(\cdot)$ denotes an independent run.

The simulation runs in (1) are specified by a parameter δ that typically impacts the bias and the variance in an antagonistic fashion. A common example is finite-difference schemes for black-box or zeroth-order noisy gradient estimation, in which δ is the perturbation size for the function input of interest. As δ increases, bias increases while variance decreases (and vice versa). To minimize the mean square error (MSE), the best choice of δ, in terms of the simulation budget, balances the magnitudes of the two error sources. In central finite-difference for instance, this optimal δ turns out to be of order $n^{-\frac{1}{6}}$, whereas in forward or backward finite-difference it is of order $n^{-\frac{1}{4}}$ (e.g., Glasserman (2013) Chapter 7; Asmussen and Glynn (2007) Chapter 7; Fu (2006); L'Ecuyer (1991)).

While the above tradeoff and the optimal order of δ in n is well understood in the literature, the precise best choice of δ depends on other, typically unknown, model characteristics (i.e., the "constants" inside $b(\delta)$ and $v(\delta)$). For example, choosing $\delta = dn^{-\frac{1}{6}}$ in a central finite-difference, and considering only the first-order error term, the best choice of d depends on third-order derivative information and the variance of the noise that are typically unavailable in advance.

Our goal in this paper is to develop a framework that enhances the standard estimator (2) regarding the choice of δ subject to the ambiguity of the model characteristics. A key idea we will undertake is to consider estimators beyond the form of naive sample average, in a way that reduces the impact of this uncertainty. Under this framework, we derive new estimators that consistently improve (2) at a given choice of δ, regardless of these unknowns. This improvement is in terms of the asymptotic MSE as the simulation budget increases. More specifically, we consider the asymptotic ratio between the MSEs of any proposed estimator and (2):

$$R = \limsup_{n\to\infty} \frac{\text{MSE of a proposed estimator}}{\text{MSE of the conventional estimator (2)}} \tag{3}$$

The proposed estimator can be parametrized by possibly many tuning parameters. The asymptotic ratio R thus contains these parameters, the unknown model characteristics, and the δ in (2). Regarding (2) and its δ as a "baseline", we calibrate the tuning parameters in the proposed estimator to minimize the worst-case asymptotic MSE ratio, where the worst case is over all possible model characteristics and choices of δ. On a high level, this can be expressed as

$$R^* = \min_{\substack{\text{calibration}\\\text{strategy}}} \max_{\substack{\text{model}\\\text{characteristics}},\delta} R \tag{4}$$

This minimized worst-case ratio R^* provides a performance guarantee on our calibrated proposed estimator relative to (2) – The MSE of our estimator is asymptotically at most R^* of (2) at the chosen δ, independent of any possible model specifications. In particular, if $R^* < 1$, our estimator is guaranteed to strictly improve over (2). For convenience, we call R^* the *asymptotic minimax risk ratio (AMRR).*

As our main contributions, we systematically identify the AMRR R^*, achieve $R^* < 1$, and construct a scheme that consistently outperforms the conventional choice (2), for the class of weighted estimators in the form

$$\sum_{j=1}^{n} w_j \theta_j\left(\delta_j\right) \tag{5}$$

where $\delta_j, j = 1,\ldots,n$ is a suitable sequence of tuning parameters, and $w_j, j = 1,\ldots,n$ is any weighting sequence. For example, when w_i's are the uniform weights, (5) is precisely the so-called recursive estimator introduced in Glynn and Whitt (1992). Our main results show that, in general, the optimal weighting scheme to obtain R^* is in the form

$$w_j = \frac{\lambda_1}{j^{\beta_1}} + \frac{\lambda_2}{j^{\beta_2}} \tag{6}$$

where $\beta_1, \beta_2 > 0$ are two distinct decay rates. The two coefficients λ_1, λ_2 depend on the budget n, in a way that none of the two terms in (6) is asymptotically negligible when used in the weighted estimator. This weighting scheme and an associated transformation from δ to $\{\delta_j\}_j$ give rise to an explicitly identifiable R^*. This reveals that, for instance, in the central finite-difference scheme, R^* is 0.67 when the multiplicative constants in δ and $\{\delta_j\}_j$ are the same. Since $R^* < 1$, the weighted estimator using (6) always outperforms (2) in terms of asymptotic MSE, independent of the unknown constants in $b(\delta)$ and $v(\delta)$. In contrast, the corresponding R^* is 1.08 when the weights are obtained via the recursive estimator or its immediate generalization, indicating that such a restriction on the weighting sequence could lead to subpar performance in the MSE.

Our main analyses build on the insight that, to maintain a low worst-case risk ratio, one typically must calibrate the proposed estimator such that it maintains the relative magnitudes of bias and variance in a similar manner as the conventional scheme (2). We will show that any distortion away from such a balancing allows an "adversary" to enlarge the risk ratio, thus leading to suboptimal outcomes. This balancing requirement generally leads to a non-convex constrained optimization problem which, upon a reformulation, reveals a tractable structure and solution to the minimax problem in (4).

Finally, we conduct experiments to test and compare our optimally weighted estimator with the recursive estimator and sample-average baseline. Our experimental results demonstrate that, when applied to various models, our optimally weighted estimator exhibits lower MSE than the baseline when the simulation budget is as low as 20, whereas the optimal recursive estimator has a slightly higher MSE than the baseline, which match our theoretical predictions. Moreover, we illustrate the potential of harnessing our optimal weighting scheme in obtaining faster convergence for black-box stochastic optimization, by incorporating it in the finite difference estimator at each iteration of a zeroth-order stochastic gradient descent algorithm (Kushner and Yin (2003)). On the other hand, we observe some (positive) deviations of our risk ratios from the AMRR, suggesting non-negligible finite-sample effects, especially when δ_j differ from δ by a large factor. Nonetheless, a thorough theoretical understanding of the finite-sample behavior of our weighting scheme is beyond the scope of this paper and will be left for the future.

The remainder of the paper is as follows. Section 2 first reviews some related works. Section 3 describes the problem settings and reviews some established results on biased estimation. Section 4 presents our minimax framework and investigation on a special class of estimators. Section 5 presents our main results and explains their implications on general weighted estimators and AMRR. Section 6 discusses how our results carry to multivariate settings. Section 7 reports our numerical experiments. Section 8 concludes the paper. All proofs and additional numerical results are provided in the Appendix.

2. Related Literature

Our study is related to several lines of work. The minimax formulation that we use to analyze and construct estimators resembles robust optimization (e.g., Ben-Tal et al. (2009), Bertsimas et al. (2011), Ben-Tal and Nemirovski (2002)) and robust control (e.g., Zhou and Doyle (1998)) that advocates decision-making against the worst-case scenario. Such ideas also have roots in game theory (Cesa-Bianchi and Lugosi (2006)). Related notions have also been used in online optimization, in which decision is made at each step under a noisily observed dynamical process (e.g., Flaxman et al. (2005), Shalev-Shwartz (2012), Hazan et al. (2016)). The performance in this literature is often measured by the regret that indicates the suboptimality of a decision relative to the best decision assuming complete information (see, e.g., Besbes and Zeevi (2009, 2011) for applications in revenue management). Instead of using an "oracle" best as the benchmark in our minimax formulation, we use the sample average as our benchmark, and focus on improving this conventional estimator by analyzing the risk ratio. In this regard, we note that a ratio formulation and a non-oracle-best benchmark has been used in Agrawal et al. (2012), but in a different context in quantifying the impact of correlation in mean estimation, and their benchmark is an independent distribution with the worst-case being evaluated over a class of dependent models. Ratios between MSEs also appear in Pasupathy (2010) in studying the tradeoff between error tolerance and sample size in so-called retrospective approximation, which is a technique for solving stochastic root-finding or optimization problems via imposing a sequence of sample average approximation problems.

A main application of our work is finite-difference stochastic gradient estimation (e.g., Glasserman (2013) Chapter 7; Asmussen and Glynn (2007) Chapter 7; Fu (2006); L'Ecuyer (1991)), typically used when there is only a noisy simulation oracle to evaluate the function value or model output. Variants of the finite-difference method include the central, forward and backward finite-differences, with different perturbation directions and orders of bias (Zazanis and Suri (1993), Fox and Glynn (1989)). In contrast to finite-differences are unbiased derivative estimators, which include the infinitesimal perturbation analysis or pathwise differentiation (Ho et al. (1983), Heidelberger et al. (1988)), the likelihood ratio or the score function method (Glynn (1990), Rubinstein (1986), Reiman and Weiss (1989)), measure-valued or weak differentiation (Heidergott and Vázquez-Abad (2008), Heidergott et al. (2010)), and other variants such as the push-out method (Rubinstein (1992), L'Ecuyer (1990)), conditional and smoothed perturbation analysis (Gong and Ho (1987), Hong (2009), Fu and Hu (1992), Glasserman and Gong (1990), Fu et al. (2009)) and the generalized likelihood ratio method (Peng et al. (2018)). In multivariate settings, Spall (1992, 1997) study simultaneous perturbation to estimate gradients used in SA, by randomly generating a perturbation direction vector and properly weighting with the perturbation sizes to control estimation bias. Nesterov and Spokoiny (2017) proposes Gaussian smoothing with a different adjustment and investigates finite-sample behaviors in related optimization. Flaxman et al. (2005) suggests uniform sampling. Our framework can be applied to these procedures, as will be discussed in Section 6.

The main skeleton of our proposed estimators uses a sequentialized choice of the tuning parameter, which appears in Glynn and Whitt (1992) in their discussion of subcanonical estimators. A generalization of this latter scheme, which appeared in Duplay et al. (2018) and discussed in Section

4, resembles the idea of stochastic approximation (SA) in stochastic optimization and root-finding that iteratively updates noisy estimates (Kushner and Yin (2003), Borkar (2009), Pasupathy and Kim (2011), Nemirovski et al. (2009), Polyak and Juditsky (1992), Ruppert (1988)). Our analyses there utilize the classical asymptotic techniques in Fabian (1968) and Chung (1954), and also Polyak and Juditsky (1992) in the averaging case.

Finally, we compare our work to multi-level Monte Carlo (Giles (2008)). This approach aims to reduce variance in simulation in the presence of a parameter selection like δ, by stratifying the simulation budget into different δ values. Of particular relevance is the randomized level selection (Rhee and Glynn (2015), Blanchet and Glynn (2015), Rychlik (1990), McLeish (2011)) that can turn biased estimators in the form of (1) into unbiased estimators with possibly canonical square-root convergences. The approach is generalized in Vihola (2018), which uses further stratification to obtain an expanded class of unbiased estimators with efficiency matching their biased counterparts, thus incurring negligible cost in the debiasing operation. Multi-level Monte Carlo and its debiased variants have been applied successfully in many stochastic problems including the simulation of stochastic differential equations and nonlinear functions of expectations. However, they require a probabilistic coupling between simulation runs at consecutive levels to exhibit statistical advantages. In contrast, the framework studied in this paper consists of *black-box* simulation where we assume no internal structure can be leveraged, thus ruling out the possibility of coupling simulation runs.

3. Background and Problem Setting

We elaborate our problem and notations in the introduction. We are interested in estimating $\theta_0 \in \mathbb{R}$. Given a tuning parameter $\delta \in \mathbb{R}_+$, we run Monte Carlo simulation where each run outputs (1) with $b(\delta) = B\delta^{q_1} + o(\delta^{q_1})$ as $\delta \to 0$, $v(\delta) = \frac{\varepsilon(\delta)}{\delta^{q_2}}$, and $q_1, q_2 > 0$. We assume that:

Assumption 1. *We have*

1. $B \in \mathbb{R}$ *is a non-zero constant.*
2. $\varepsilon(\delta) \in \mathbb{R}$ *is a random variable such that* $E\varepsilon(\delta) = 0$ *and* $\sigma^2(\delta) = Var(\varepsilon(\delta)) \to \sigma^2 > 0$ *as* $\delta \to 0$.

The above assumptions dictate that the order of the bias $b(\delta)$ is δ^{q_1}, while the order of the variance is δ^{-2q_2}. The former is ensured by the first assumption and the latter by the second one.

As an example, in estimating the derivative of a function $f(x)$ with unbiased noisy function evaluation, the central finite-difference (CFD) scheme elicits the output

$$\frac{\hat{f}(x+\delta)-\hat{f}(x-\delta)}{2\delta}$$

where $\hat{f}(\cdot)$ is an unbiased evaluation of $f(\cdot)$, and $\delta>0$ is the perturbation size. Given that f is thrice continuously differentiable with non-zero $f'''(x)$, the bias term has order $q_1=2$. Typically, the order of the variance is $q_2=1$. Suppose we do not apply common random numbers (CRN) in generating $\hat{f}(x+\delta)$ and $\hat{f}(x-\delta)$, and that $Var\left(\hat{f}(x\pm\delta)\right)\to Var\left(\hat{f}(x)\right)$ as $\delta\to 0$. Then $\sigma^2=\frac{1}{2}Var\left(\hat{f}(x)\right)$. Suppose we are able to apply CRN so that $Cov\left(\hat{f}(x+\delta),\hat{f}(x-\delta)\right)\to Var\left(\hat{f}(x)\right)$ as $\delta\to 0$. Then, under standard assumptions (such as those in equation (2.4) in Glynn (1989)), the order of the variance becomes $q_2=\frac{1}{2}$.

Similarly, the forward finite-difference (FFD) scheme elicits the output

$$\frac{\hat{f}(x+\delta)-\hat{f}(x)}{\delta}$$

Given that f is twice continuously differentiable with non-zero $f''(x)$, the bias term has order $q_1=1$. Analogous conditions on the noise as above guarantees that $q_2=1$ or $\frac{1}{2}$. The same discussion holds for the backward finite-difference (BFD) scheme.

Given the capability to output independent runs of (1), say $\theta_j(\delta)$, the conventional approach to obtain an estimate of θ_0 is to take their sample average. Denote this as $\bar{\theta}_n=\frac{1}{n}\sum_{j=1}^{n}\theta_j(\delta)$. The MSE of $\bar{\theta}_n$, denoted $\text{MSE}_0=E\left(\bar{\theta}_n-\theta_0\right)^2$, can be expressed as

$$\text{MSE}_0=\text{bias}^2+\text{variance}=B^2\delta^{2q_1}+\frac{\sigma^2}{n\delta^{2q_2}}+\text{higher-order terms} \tag{7}$$

Considering the first order term, the bias increases with δ and the variance decreases with δ. Minimizing the MSE requires balancing these two errors to the same order, namely by choosing $\delta=\delta_n=\Theta\left(n^{-\alpha}\right)$ where $\alpha=\frac{1}{2(q_1+q_2)}$, which solves the equation $-2\alpha q_1=-1+2\alpha q_2$. This leads to an optimal MSE order $n^{-\frac{q_1}{q_1+q_2}}$. For example, in CFD and under the conditions we discussed above without CRN, we have $\delta_n = \Theta\left(n^{-\frac{1}{6}}\right)$, leading to an optimal MSE order $n^{-\frac{2}{3}}$; in FFD or BFD we have $\delta_n = \Theta\left(n^{-\frac{1}{4}}\right)$, leading to an optimal MSE order $n^{-\frac{1}{2}}$.

In order to fully optimize the first-order MSE, including the coefficient, one needs to choose

$$\delta_n = \left(\frac{\sigma^2 q_2}{B^2 q_1}\right)^{\frac{1}{2(q_1+q_2)}} n^{-\frac{1}{2(q_1+q_2)}}$$

(e.g., by applying the first-order optimality condition on the leading terms in (7)). This gives an optimal first-order MSE

$$B^{\frac{2q_2}{q_1+q_2}} \sigma^{\frac{2q_1}{q_1+q_2}} \left(\left(\frac{q_2}{q_1}\right)^{\frac{q_1}{q_1+q_2}} + \left(\frac{q_1}{q_2}\right)^{\frac{q_2}{q_1+q_2}}\right) n^{-\frac{q_1}{q_1+q_2}} \tag{8}$$

The above choice of δ_n depends on the "constants" in the bias and variance terms, namely B and σ^2. While q_1 and q_2 are often obtainable, constants like B and σ^2 are unknown a priori and can affect the performance of the simulation estimator, despite choosing an optimal order on n in δ_n using the knowledge of q_1 and q_2. Suppose we choose $\delta_n = dn^{-\alpha}$ for some $d > 0$, where $\alpha = \frac{1}{2(q_1+q_2)}$ is optimally chosen. Then the first-order MSE is

$$\left(B^2 d^{2q_1} + \frac{\sigma^2}{d^{2q_2}}\right) n^{-\frac{q_1}{q_1+q_2}} \tag{9}$$

which can be arbitrarily suboptimal relative to the best coefficient in (8). Our goal in this paper is to improve on this suboptimality, by considering estimators beyond the conventional sample average that consistently outperforms the constant showing up in (9).

The following theorem, which follows straightforwardly from Fox and Glynn (1989), summarizes the above discussion on the optimal order of the MSE:

Theorem 1. *Under Assumption 1, suppose that* $\lim_{n\to\infty} \delta_n n^{\alpha} = d$, *where* $0 < d < \infty$, *the sample-average-based estimator* $\bar{\theta}_n$ *exhibits the asymptotic MSE*

$$E\left(\bar{\theta}_n - \theta_0\right)^2 = d^{2q_1} B^2 n^{-2\alpha q_1} + \frac{\sigma^2}{d^{2q_2}} n^{2\alpha q_2 - 1} + o\left(n^{-2\alpha q_1} + n^{2\alpha q_2 - 1}\right) \ \ \textit{as } n \to \infty$$

Choosing $\alpha = \frac{1}{2(q_1+q_2)}$ *achieves the optimal MSE order, and the asymptotic MSE is*

$$E\left(\bar{\theta}_n - \theta_0\right)^2 = \left(d^{2q_1} B^2 + \frac{\sigma^2}{d^{2q_2}}\right) n^{-\frac{q_1}{q_1+q_2}} + o\left(n^{-\frac{q_1}{q_1+q_2}}\right) \ \ \textit{as } n \to \infty$$

Lastly, we mention that, in practice, there are other considerations in obtaining good estimators, such as issues regarding the finiteness of the sample that can affect the accuracy of the asymptotic results. These considerations are beyond the scope of this work, which focuses mainly on a theoretical framework on improving the asymptotic constant.

4. A Minimax Comparison Framework

We introduce a framework to assess, and calibrate, estimators beyond the sample-average-based estimator $\bar{\theta}_n$. This framework compares the asymptotic MSEs using $\bar{\theta}_n$ as a baseline based on a minimax argument. Section 4.1 presents this framework, and Section 4.2 provides an initial study on a special type of estimators.

4.1. Asymptotic Risk Ratio

Consider an estimator $\hat{\theta}_n$ for θ_0 using n simulation runs in the form (1), where the tuning parameter δ in each run can be arbitrarily chosen. Our goal is to calibrate $\hat{\theta}_n$ that performs well, or outperforms, $\bar{\theta}_n$ in the first-order coefficient of the MSE, presuming that both $\hat{\theta}_n$ and $\bar{\theta}_n$ have the optimal order of errors. Let MSE_1 denote the MSE of $\hat{\theta}_n$ for convenience.

The estimator $\hat{\theta}_n$ can depend on other tuning parameters in addition to the δ in each run. We denote the collection of all the parameters that $\hat{\theta}_n$ involves as ν, so that $\hat{\theta}_n = \hat{\theta}_n(\nu)$. Correspondingly, MSE_1 also depends on ν.

We suppose knowledge on the order of the bias and noise, namely q_1 and q_2 in (1). However, we do not know the constants B and σ^2. To make the discussion more precise, for fixed $q_1, q_2 > 0$, we denote the class of simulation outputs

$$\begin{aligned} H = \{\theta(\cdot) : \theta(\delta) = \theta_0 + b(\delta) + v(\delta) \text{ such that} \\ b(\delta) = B\delta^{q_1} + o(\delta^{q_1}) \text{ and } v(\delta) = \frac{\varepsilon(\delta)}{\delta^{q_2}} \text{ where } Var(\epsilon(\delta)) \to \sigma^2, \text{ as } \delta \to 0, \\ \text{for arbitrary non-zero } B \text{ and positive } \sigma^2\} \end{aligned} \tag{10}$$

In other words, H is the set of outputs with bias of order δ^{q_1} and noise of order δ^{-q_2}, with arbitrary constants B, σ^2.

The MSE of $\bar{\theta}_n$, MSE_0, depends on $\theta(\cdot)$ evaluated at chosen δ. To highlight this dependence, we write $\mathrm{MSE}_0 = \mathrm{MSE}_0(\theta(\cdot), \delta)$, where we make implicit the dependence on n for ease of notation. Similarly, MSE_1 depends on $\theta(\cdot)$ and ν, so that $\mathrm{MSE}_1 = \mathrm{MSE}_1(\theta(\cdot), \nu)$, where we also make implicit the dependence on n for ease of notation. We consider the *asymptotic risk ratio*

$$R(\theta(\cdot), \nu, \delta) = \limsup_{n\to\infty} \frac{\mathrm{MSE}_1(\theta(\cdot), \nu)}{\mathrm{MSE}_0(\theta(\cdot), \delta)} \tag{11}$$

that measures the performance of θ_n relative to $\bar{\theta}_n$ as a baseline. Since we only know $\theta(\cdot)$ is in H but not its exact forms (i.e., the constants), we consider the worst-case scenario of R, and search for the best parameters in $\hat{\theta}_n$ that minimize this worst-case risk. Namely, we aim to solve

$$\min_{\nu} \max_{\theta(\cdot)\in H} R(\theta(\cdot), \nu, \delta) \tag{12}$$

Note that (12), and the best choice of ν, depend on the δ used in $\bar{\theta}_n$. We now take a further viewpoint that an arbitrary user may select any δ, and we look for a strategy to calibrate $\hat{\theta}_n$ that is guaranteed to perform well no matter how δ is chosen. To write this more explicitly, we let $\nu = \nu(\delta)$ be dependent on δ, and we search for the best collection of parameters that is potentially a function $\nu(\cdot)$ on δ:

$$R^* = \min_{\nu(\cdot)\in\Lambda} \max_{\theta(\cdot)\in H, \delta\in\mathbb{R}^+} R(\theta(\cdot), \nu(\cdot), \delta) \tag{13}$$

where Λ denotes a set of functions. This set Λ depends on the class of estimators $\hat{\theta}_n$ we use, which will be described in detail. Moreover, as we will see, (12) and (13) are closely related; in fact, under the settings we consider, solving either of them simultaneously solves another. In the following, we will focus on (13) and discuss the immediate implications on (12) where appropriate. We shall call R^* the asymptotic minimax risk ratio (AMRR).

4.2. An Initial Example: Recursive Estimators

For convenience, let us from now on set $\delta = d(n+n_0)^{-\alpha}$ as the tuning parameter in the sample-average-based estimator $\bar{\theta}_n$, where $\alpha = \frac{1}{2(q_1+q_2)}$ so that it achieves the optimal MSE order. The number n_0 can be any fixed integer to prevent δ from being too big at the early stage, and does not affect our asymptotic analyses.

To construct our proposed estimator $\hat{\theta}_n$, we will first use the idea of the recursive estimator studied in Section 5 of Glynn and Whitt (1992). At run j, we simulate $\theta_j(\delta_j)$, where $\delta_j = \tilde{d}(j+n_0)^{-\alpha}$ for some constant $\tilde{d}$, and α is the same as in $\bar{\theta}_n$, i.e., the parameter is chosen as if the current simulation run is the last one in the budget if a conventional sample-average-based estimator is used. The estimator in Glynn and Whitt (1992) uses the average of $\theta_j(\delta_j)$, namely $\frac{1}{n}\sum_{j=1}^{n}\theta_j(\delta_j)$. As shown in Glynn and Whitt (1992), this estimator exhibits the optimal MSE order like $\bar{\theta}_n$. Moreover, as they have also noted, this estimator admits a recursive representation $\hat{\theta}_n = \left(1-\frac{1}{n}\right)\hat{\theta}_{n-1} + \frac{1}{n}\theta_n(\delta_n)$, where each update depends only on the parameter indexed by the current run number, rather than the budget. Thus, the optimal MSE order is achieved in an “online” fashion as n increases, independent of the final budget.

The initial class of estimators that we will consider is a generalization of Glynn and Whitt (1992), which is also considered in Duplay et al. (2018). Specifically, we consider estimators defined via the recursion

$$\hat{\theta}_n^{rec} = (1-\gamma_n)\,\hat{\theta}_{n-1}^{rec} + \gamma_n\theta_n(\delta_n) \tag{14}$$

where $\delta_n = \tilde{d}(n+n_0)^{-\alpha}$ is defined as before and $\alpha > 0$, and γ_n is in the form $c(n+n_0)^{-\beta}$ for some $c > 0$ and $\beta > 0$. $\hat{\theta}_0^{rec}$ can be arbitrary. Moreover, we also consider averaging $\hat{\theta}_n^{rec}$ in the form

$$\hat{\theta}_n^{avg} = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_j^{rec} \tag{15}$$

which resembles the standard Polyak-Ruppert averaging in SA (Polyak and Juditsky (1992)).

Our first result is that, in terms of the AMRR, the class of estimators $\hat{\theta}_n^{rec}$ and $\hat{\theta}_n^{avg}$ are quite restrictive and cannot bring in much improvement over $\bar{\theta}_n$. To elicit this result, we begin with some consistency properties of $\hat{\theta}_n^{rec}$:

Proposition 1. *Under Assumption 1, we have:*

1. *If $\beta \leq 1$ and $\alpha < \frac{\beta}{2q_2}$, the estimator $\hat{\theta}_n^{rec}$ is L_2-consistent for θ_0, i.e.,*

$$\lim_{n\to\infty} E\left(\hat{\theta}_n^{rec} - \theta_0\right)^2 = 0$$

2. *If $\beta \leq 1$ and $\alpha \geq \frac{\beta}{2q_2}$, or if $\beta > 1$, the error of $\hat{\theta}_n^{rec}$ in estimating θ_0 is bounded away from zero in L_2-norm as $n \to \infty$, i.e.,*

$$\liminf_{n\to\infty} E\left(\hat{\theta}_n^{rec} - \theta_0\right)^2 > 0$$

Proposition 1 shows that $\hat{\theta}_n^{rec}$ estimates θ_0 sensibly only when $\beta \leq 1$ and $\alpha < \frac{\beta}{2q_2}$. We thus focus on this case subsequently. The following describes the convergence rate:

Theorem 2. *Under Assumption 1, the MSE of $\hat{\theta}_n^{rec}$ in estimating θ_0 behaves as follows:*

1. *For $\beta < 1$ and $\alpha < \frac{\beta}{2q_2}$,*

$$E\left(\hat{\theta}_n^{rec} - \theta_0\right)^2 = d^{2q_1}B^2 n^{-2q_1\alpha} + \frac{c\sigma^2}{2d^{2q_2}} n^{2q_2\alpha - \beta} + o\left(n^{-2q_1\alpha} + n^{2q_2\alpha-\beta}\right) \text{ as } n \to \infty$$

2. *For $\beta = 1$, $\alpha = \frac{1}{2(q_1+q_2)}$ and $c > \frac{q_1}{2(q_1+q_2)}$,*

$$E\left(\hat{\theta}_n^{rec} - \theta_0\right)^2 = \left(\left(\frac{cd^{q_1}}{c - \frac{q_1}{2(q_1+q_2)}}\right)^2 B^2 + \frac{c^2\sigma^2}{\left(2c - \frac{q_1}{q_1+q_2}\right) d^{2q_2}}\right) n^{-\frac{q_1}{q_1+q_2}} + o\left(n^{-\frac{q_1}{q_1+q_2}}\right) \text{ as } n \to \infty \tag{16}$$

3. *For $\beta = 1$, $\alpha = \frac{1}{2(q_1+q_2)}$ and $c \leq \frac{q_1}{2(q_1+q_2)}$, or for $\beta = 1$ and $\alpha \neq \frac{1}{2(q_1+q_2)}$,*

$$\limsup_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} E\left(\hat{\theta}_n^{rec} - \theta_0\right)^2 = \infty$$

The proofs of the above results, which are detailed in Appendix B, utilize the classical asymptotic techniques for recursive sequences in Fabian (1968) and a slight modification of Chung's lemma (i.e., Lemma 1 in Appendix B).

We now look at the AMRR for $\hat{\theta}_n^{rec}$. First, Theorem 2 shows that the choice $\beta = 1, \alpha = \frac{1}{2(q_1+q_2)}$ is the unique choice that gives rise to the optimal MSE order $n^{-\frac{q_1}{q_1+q_2}}$. Moreover, given this choice of α, we need $c > \frac{q_1}{2(q_1+q_2)}$, in addition to $\beta = 1$. We will focus on these configurations for $\hat{\theta}_n^{rec}$ that achieve the same MSE order as the conventional estimator $\bar{\theta}_n$ with the same α.

Suppose we set $\tilde{d} = d$, but allow the free selection of c within the range that gives rise to the optimal MSE order. We thus can write $\hat{\theta}_n^{rec} = \hat{\theta}_n^{rec}(\nu)$ where $\nu = (d, c)$ is the collection of all tuning parameters that $\hat{\theta}_n^{rec}$ depends on, defined via (14) with $\gamma_n = c\,(n+n_0)^{-1}$ where $c > \frac{q_1}{2(q_1+q_2)}$. The integer n_0 does not affect any asymptotic and can be taken as any given value. The following characterizes the AMRR and the configuration that attains it:

Theorem 3. *Under Assumption 1, let $MSE_1^{rec}\left(\theta\left(\cdot\right),d,c\right)$ be the MSE of $\hat{\theta}_n^{rec}\left(d,c\right)$, and*

$$R^{rec}\left(\theta\left(\cdot\right),d,c\right) = \limsup_{n\to\infty} \frac{MSE_1^{rec}\left(\theta\left(\cdot\right),d,c\right)}{MSE_0\left(\theta\left(\cdot\right),d\right)}$$

We have

$$\min_{c>\frac{q_1}{2(q_1+q_2)}} \max_{\theta(\cdot)\in H, d>0} R^{rec}\left(\theta\left(\cdot\right),d,c\right) = \frac{q_1^2}{16\left(q_1+q_2\right)^2} + \frac{q_1}{2\left(q_1+q_2\right)} + 1$$

which is attained by choosing $c = \frac{5q_1+4q_2}{2(q_1+q_2)}$.

Next, we provide more flexibility in the choice of $\tilde{d}$ in $\hat{\theta}_n^{rec}\left(\nu\right)$, where $\nu = \left(\tilde{d},c\right)$. In particular, rather than setting $\tilde{d}=d$, we allow $\tilde{d}$ to depend on d in any arbitrary fashion, i.e., $\tilde{d} = g\left(d\right)$ where $g\left(\cdot\right): \mathbb{R}_+ \to \mathbb{R}_+$ is any function. Let $\mathcal{F}$ be the space of any functions from $\mathbb{R}_+$ to $\mathbb{R}_+$. We have the following results on the AMRR of this enhanced scheme where. For convenience, we denote

$$G^* = \left(\frac{q_1+2q_2}{4\left(q_1+q_2\right)}\right)^{\frac{1}{2(q_1+q_2)}} \tag{17}$$

Theorem 4. *Under Assumption 1, let $MSE_1^{rec}\left(\theta\left(\cdot\right),\tilde{d},c\right)$ be the MSE of $\hat{\theta}_n^{rec}\left(\tilde{d},c\right)$, and*

$$R^{rec}\left(\theta\left(\cdot\right),d,\tilde{d},c\right) = \limsup_{n\to\infty} \frac{MSE_1^{rec}\left(\theta\left(\cdot\right),\tilde{d},c\right)}{MSE_0\left(\theta\left(\cdot\right),d\right)}$$

We have

$$\min_{g(\cdot)\in\mathcal{F}, c>\frac{q_1}{2(q_1+q_2)}} \max_{\theta(\cdot)\in H, d>0} R^{rec}\left(\theta\left(\cdot\right),d,g\left(d\right),c\right) = 2^{\frac{2q_2}{q_1+q_2}} \left(\frac{q_1+2q_2}{q_1+q_2}\right)^{-\frac{q_1+2q_2}{q_1+q_2}}$$

which is attained by choosing $g\left(d\right) = G^* d$ and $c=1$, where G^* is defined in (17).

We note that Theorem 4 indicates $c=1$ is optimal in this enhanced scheme, while the optimal $\tilde{d}$ is chosen as a constant factor G^* of d.

Next we look at $\hat{\theta}_n^{avg}$. It turns out that the AMRR depicted for $\hat{\theta}_n^{rec}$ in Theorem 4 applies also to $\hat{\theta}_n^{avg}$. To this end, we first state the MSE of $\hat{\theta}_n^{avg}$:

Theorem 5. *Under Assumption 1, the MSE of $\hat{\theta}_n^{avg}$ in estimating θ_0 behaves as follows:*

1. *For $\beta < 1$ and $\alpha \leq \frac{1}{2(q_1+q_2)}$,*

$$E\left(\hat{\theta}_n^{avg} - \theta_0\right)^2 = \left(\frac{d^{q_1}}{1-q_1\alpha}\right)^2 B^2 n^{-2q_1\alpha} + \frac{\sigma^2}{(1+2q_2\alpha)\, d^{2q_2}} n^{2q_2\alpha-1} + o\left(n^{-2q_1\alpha} + n^{2q_2\alpha-1}\right) \text{ as } n \to \infty \tag{18}$$

2. *For $\beta < 1$ and $\alpha > \frac{1}{2(q_1+q_2)}$,*

$$E\left(\hat{\theta}_n^{avg} - \theta_0\right)^2 = \frac{\sigma^2}{(1+2q_2\alpha)\, d^{2q_2}} n^{2q_2\alpha-1} + o\left(n^{2q_2\alpha-1}\right) \text{ as } n \to \infty$$

Comparing Theorem 5 with Theorem 2, we see that, when $\alpha = \frac{1}{2(q_1+q_2)}$, the first-order MSE of $\hat{\theta}_n^{avg}$ in the considered regime (in (18)) exactly equals that of $\hat{\theta}_n^{rec}$ (in (16)) when $c = 1$ and $\beta = 1$. Like before, $\alpha = \frac{1}{2(q_1+q_2)}$ is the unique choice that optimizes the MSE order for $\hat{\theta}_n^{avg}$. Thus, we will focus on this choice of α in $\hat{\theta}_n^{avg}$. Note that then $\hat{\theta}_n^{avg} = \hat{\theta}_n^{avg}(\nu)$ where $\nu = \left(\tilde{d}, c, \beta\right)$ is the collection of tuning parameters that $\hat{\theta}_n^{avg}$ depends on. This leads us to the following AMRR:

Theorem 6. *Under Assumption 1, let $MSE_1^{avg}\left(\theta(\cdot), \tilde{d}, c, \beta\right)$ be the MSE of $\hat{\theta}_n^{avg} = \hat{\theta}_n^{avg}\left(\tilde{d}, c, \beta\right)$. Let*

$$R^{avg}\left(\theta(\cdot), d, \tilde{d}, c, \beta\right) = \limsup_{n\to\infty} \frac{MSE_1^{avg}\left(\theta(\cdot), \tilde{d}, c, \beta\right)}{MSE_0\left(\theta(\cdot), d\right)}$$

We have

$$\min_{g(\cdot)\in\mathcal{F}, c>0, 0<\beta<1} \max_{\theta(\cdot)\in H, d>0} R^{avg}\left(\theta(\cdot), d, g(d), c, \beta\right) = 2^{\frac{2q_2}{q_1+q_2}} \left(\frac{q_1+2q_2}{q_1+q_2}\right)^{-\frac{q_1+2q_2}{q_1+q_2}}$$

which is attained by choosing $g(d) = G^* d$, and any $c > 0$ and $0 < \beta < 1$, where G^* is defined in (17).

The minimax ratios stated in Theorems 3, 4 and 6 remain the same, in a uniform fashion, when the parameter d in $\bar{\theta}_n$ is fixed instead of being chosen by an adversarial user. In other words, the minimax risk ratio of $\hat{\theta}_n^{rec}$ or $\hat{\theta}_n^{avg}$ compared to $\bar{\theta}_n$ would not improve with a finer calibration on the tuning parameters $\tilde{d}, c, \beta$ catered to each specific d. This is described in the following result:

Theorem 7. *We have the following:*

1. *Under the conditions and notations in Theorem 3, we have, for any fixed d,*

$$\min_{c>\frac{q_1}{2(q_1+q_2)}} \max_{\theta(\cdot)\in H} R^{rec}\left(\theta\left(\cdot\right),d,c\right)=\frac{q_1^2}{16\left(q_1+q_2\right)^2}+\frac{q_1}{2\left(q_1+q_2\right)}+1$$

which is attained by choosing $c=\frac{5q_1+4q_2}{2(q_1+q_2)}$.

2. *Under the conditions and notations in Theorem 4, we have, for any fixed d,*

$$\min_{\tilde{d}>0,c>\frac{q_1}{2(q_1+q_2)}} \max_{\theta(\cdot)\in H} R^{rec}\left(\theta\left(\cdot\right),d,\tilde{d},c\right)=2^{\frac{2q_2}{q_1+q_2}}\left(\frac{q_1+2q_2}{q_1+q_2}\right)^{-\frac{q_1+2q_2}{q_1+q_2}}$$

which is attained by choosing $\tilde{d}=G^{*}d$ *and* $c=1$*, where* G^{*} *is defined in* (17).

3. *Under the conditions and notations in Theorem 6, we have, for any fixed d,*

$$\min_{\tilde{d}>0,c>0,0<\beta<1} \max_{\theta(\cdot)\in H} R^{avg}\left(\theta\left(\cdot\right),d,\tilde{d},c,\beta\right)=2^{\frac{2q_2}{q_1+q_2}}\left(\frac{q_1+2q_2}{q_1+q_2}\right)^{-\frac{q_1+2q_2}{q_1+q_2}}$$

which is attained by choosing $\tilde{d}=G^{*}d$*, and any* $c>0$ *and* $0<\beta<1$*, where* G^{*} *is defined in* (17).

Theorem 7 is consistent with Theorems 3, 4 and 6 in that the optimal strategies to calibrate the $\tilde{d}$ in $\hat{\theta}_n^{rec}$ and $\hat{\theta}_n^{avg}$ remain as a constant scaling on d, regardless of what the specific value of d is.

To get a numerical sense of the above results, Tables 1 and 2 show the AMRR and optimal configurations of $\hat{\theta}_n^{rec}$ and $\hat{\theta}_n^{avg}$. Table 1 illustrates the scenario $q_1=2$ and $q_2=1$ (the CFD case without CRN). Restricting $\tilde{d}=d$ in $\hat{\theta}_n^{rec}$ (i.e., Theorem 3), the AMRR is 1.38, attained by setting $c=2.33$ in $\hat{\theta}_n^{rec}$. In contrary, if we allow $\tilde{d}$ to arbitrarily depend on d (i.e., Theorem 4), the AMRR is reduced to 1.08, attained by setting $g\left(d\right)=0.83d$, and $c=1$ in $\hat{\theta}_n^{rec}$. Similarly, the AMRR for $\hat{\theta}_n^{avg}$ (i.e., Theorem 6) is also 1.08, attained again by setting $g\left(d\right)=0.83d$ but now with any $c>0$ and $0<\beta<1$.

Analogously, Table 2 illustrates the scenario $q_1=1$ and $q_2=1$ (the FFD and BFD cases without CRN). If we restrict $\tilde{d}=d$ in $\hat{\theta}_n^{rec}$ (i.e., Theorem 3), the AMRR becomes 1.27, attained by setting $c=2.25$ in $\hat{\theta}_n^{rec}$. In contrary, if we allow $\tilde{d}$ to arbitrarily depend on d (i.e., Theorems 4 and 6), the AMRR is 1.09, attained by setting $g\left(d\right)=0.78d$, and $c=1$ in $\hat{\theta}_n^{rec}$ or $c>0,0<\beta<1$ in $\hat{\theta}_n^{avg}$.

Note that, in all cases considered above, the AMRR is greater than 1, implying that without knowledge on the model characteristics, the estimators $\hat{\theta}_n^{rec}$ and $\hat{\theta}_n^{avg}$ can have a higher MSE than the baseline $\bar{\theta}_n$ asymptotically.

	$\hat{\theta}_n^{rec}$ (d unadjusted)	$\hat{\theta}_n^{rec}$ (d optimized)	$\hat{\theta}_n^{avg}$
AMRR	1.38	1.08	1.08
Optimal Configuration	$c = 2.33,\ \beta = 1$	$\tilde{d} = 0.83d, c = 1, \beta = 1$	$\tilde{d} = 0.83d, c > 0, 0 < \beta < 1$

Table 1 AMRR and optimal configurations for the case $q_1 = 2, q_2 = 1$

	$\hat{\theta}_n^{rec}$ (d unadjusted)	$\hat{\theta}_n^{rec}$ (d optimized)	$\hat{\theta}_n^{avg}$
AMRR	1.27	1.09	1.09
Optimal Configuration	$c = 2.25, \beta = 1$	$\tilde{d} = 0.78d, c = 1, \beta = 1$	$\tilde{d} = 0.78d, c > 0, 0 < \beta < 1$

Table 2 AMRR and optimal configurations for the case $q_1 = 1, q_2 = 1$

4.3. Maintaining Bias-Variance Balance

We provide an intuitive explanation on the minimax results in Section 4.2. More specifically, we demonstrate that a key argument to obtain the minimax calibration strategy of a proposed class of estimators is to balance bias and variance in a similar manner as the baseline estimator, in terms of the factors multiplying the unknown first-order constants B and σ^2. This insight is general and will be helpful in optimally calibrating wider classes of estimators, such as the general weighted estimators presented in the next section.

To explain, let us recall the notation in (11) that in general, the asymptotic risk ratio between a proposed estimator with parameter ν and a baseline estimator (where we hide its parameter for now) can be expressed as

$$R(\theta(\cdot), \nu) = \limsup_{n\to\infty} \frac{\mathrm{MSE}_1(\theta(\cdot), \nu)}{\mathrm{MSE}_0(\theta(\cdot))}$$

Suppose that both estimators have the same MSE order, which is obtained optimally by balancing the orders of the bias and variance. Then the limit in the above expression becomes

$$R(\theta(\cdot), \nu) = \frac{\mathrm{bias}_1(\nu)^2 + \mathrm{var}_1(\nu)}{\mathrm{bias}_0^2 + \mathrm{var}_0} \tag{19}$$

where $\mathrm{bias}_1(\nu)$ and $\mathrm{var}_1(\nu)$ refer to the first-order coefficient in the bias and variance terms of the proposed estimator, and similarly bias_0 and var_0 refer to the corresponding quantities of the baseline estimator. Furthermore, with the model constants B and σ^2, we can further write (19) as

$$R(\theta(\cdot), \nu) = \frac{C_1^{bias}(\nu) B^2 + C_1^{var}(\nu)\sigma^2}{C_0^{bias} B^2 + C_0^{var}\sigma^2}$$

where $C_1^{bias}(\nu)$ and $C_1^{var}(\nu)$ are the coefficients in front of B^2 and σ^2 in the first-order MSE of the proposed estimator, and C_0^{bias} and C_0^{var} are the corresponding quantities of the baseline estimator.

Now, given these coefficients, an adversary who attempts to maximize $R(\theta(\cdot),\nu)$ would select either an arbitrarily big B^2 or σ^2 , depending on which ratio $C_1^{bias}(\nu)/C_0^{bias}$ or $C_1^{var}(\nu)/C_0^{var}$ is larger respectively, which leads to a worst-case ratio $\max\{C_1^{bias}(\nu)/C_0^{bias}, C_1^{var}(\nu)/C_0^{var}\}$. This forces the minimizer to calibrate ν such that the two ratios are exactly the same, i.e., we choose ν such that

$$\frac{C_1^{bias}(\nu)}{C_0^{bias}} = \frac{C_1^{var}(\nu)}{C_0^{var}} = S \tag{20}$$

for some constant S. With this observation, the solution to solve for AMRR can be formulated as minimizing S subject to the constraint (20), namely

$$\min_{\nu} \quad S \quad \text{subject to} \quad \frac{C_1^{bias}(\nu)}{C_0^{bias}} = \frac{C_1^{var}(\nu)}{C_0^{var}} = S \tag{21}$$

which gives the AMRR R^*, and an optimal solution for (21) is the minimax calibration for the proposed estimator. This line of analysis applies similarly when the baseline estimator contains its own tuning parameter δ, and that the proposed estimator is calibrated in a way dependent on δ (either in formulation (12) or (13)).

Now let us consider $\hat{\theta}_n^{rec}$ in Theorem 3. From Theorems 1 and 2, since we assume both the parameters of $\bar{\theta}_n$ and $\hat{\theta}_n^{rec}$ are chosen to exhibit the optimal MSE order, we can write

$$\begin{aligned}
R^{rec}(\theta(\cdot),d,c) &= \limsup_{n\to\infty} \frac{\mathrm{MSE}_1^{rec}(\theta(\cdot),d,c)}{\mathrm{MSE}_0(\theta(\cdot),d)} \\
&= \limsup_{n\to\infty} \frac{\left(\left(\frac{cd^{q_1}}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2 B^2 + \frac{c^2\sigma^2}{2d^{2q_2}\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\right) n^{-\frac{q_1}{q_1+q_2}} + o\left(n^{-\frac{q_1}{q_1+q_2}}\right)}{\left(d^{2q_1}B^2 + \frac{\sigma^2}{d^{2q_2}}\right) n^{-\frac{q_1}{q_1+q_2}} + o\left(n^{-\frac{q_1}{q_1+q_2}}\right)} \\
&= \frac{\left(\frac{cd^{q_1}}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2 B^2 + \frac{c^2}{2d^{2q_2}\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\sigma^2}{d^{2q_1}B^2 + \frac{1}{d^{2q_2}}\sigma^2}
\end{aligned}$$

We set

$$\frac{\left(\frac{cd^{q_1}}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2}{d^{2q_1}} = \frac{\frac{c^2}{2d^{2q_2}\left(c-\frac{q_1}{2(q_1+q_2)}\right)}}{\frac{1}{d^{2q_2}}}$$

and notice that d can be all cancelled out, giving

$$\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2=\frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}$$

which upon solving leads to $c=\frac{5q_1+4q_2}{2(q_1+q_2)}$ and both sides of the equation being $\frac{q_1^2}{16(q_1+q_2)^2}+\frac{q_1}{2(q_1+q_2)}+1$, thus giving the corresponding result in Theorem 3. Note that, since d is cancelled out in the above derivation, the same result in Theorem 7 holds immediately for the setting of any fixed d.

For $\hat{\theta}_n^{rec}$ in Theorem 4, we can write

$$R^{rec}\left(\theta(\cdot),d,\tilde{d},c\right)=\frac{\left(\frac{c\tilde{d}^{q_1}}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2B^2+\frac{c^2}{2\tilde{d}^{2q_2}\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\sigma^2}{d^{2q_1}B^2+\frac{1}{d^{2q_2}}\sigma^2}$$

and we set

$$\frac{\left(\frac{c\tilde{d}^{q_1}}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2}{d^{2q_1}}=\frac{\frac{c^2}{2\tilde{d}^{2q_2}\left(c-\frac{q_1}{2(q_1+q_2)}\right)}}{\frac{1}{d^{2q_2}}} \tag{22}$$

However, the d is not cancelled out here. Nonetheless, we can rewrite (22) in terms of the ratio $\frac{\tilde{d}}{d}$, as

$$\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2\left(\frac{\tilde{d}}{d}\right)^{2q_1}=\frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\frac{1}{\left(\frac{\tilde{d}}{d}\right)^{2q_2}}$$

Optimizing jointly over c and $\eta=\frac{\tilde{d}}{d}$ gives $c=1$ and $\eta=\left(\frac{q_1+2q_2}{4(q_1+q_2)}\right)^{\frac{1}{2(q_1+q_2)}}$, and the value on both sides of the equation is $2^{\frac{2q_2}{q_1+q_2}}\left(\frac{q_1+2q_2}{q_1+q_2}\right)^{-\frac{q_1+2q_2}{q_1+q_2}}$. This shows the result for $\hat{\theta}_n^{rec}$ in Theorem 4. Moreover, note that regardless of whether d is chosen by the adversary or fixed in advance, we choose $\tilde{d}$ as ηd, and thus we also show the corresponding results in Theorem 7. Appendix B further details the above arguments.

5. General Weighted Estimators

We now consider a substantially more general class of estimators than $\hat{\theta}_n^{rec}$ and $\hat{\theta}_n^{avg}$. Namely, given we generate $\theta_j(\delta_j), j=1,\ldots,n$ where $\delta_j=\tilde{d}(j+n_0)^{-\alpha}$ with the optimally chosen $\alpha=\frac{1}{2(q_1+q_2)}$ and n_0 is any fixed integer, we consider

$$\hat{\theta}_n^{gen}=\sum_{j=1}^{n}w_{j,n}\theta_j(\delta_j) \tag{23}$$

where $w^{(n)} = (w_{j,n})_{j=1,\ldots,n}$ is any weighting sequence.

In the following, we will first present our main result on the AMRR of (23) relative to $\bar{\theta}_n$ with $\delta = d\,(n+n_0)^{-\alpha}$, and the associated characterization of the optimal weighting scheme as a sum of two distinct decaying components (Section 5.1). Then we will describe the key developments of the result that relies on analyzing a non-convex constrained optimization (Section 5.2).

5.1. Optimal Weighted Estimators and Two-Decay Characterization

The estimator $\hat{\theta}_n^{gen}$ in (23) contains the tuning parameter $\tilde{d}$ and the weighting sequence $w^{(n)}$. While $\tilde{d}$ is chosen independent of n in the asymptotic (as it appears in the asymptotic risk ratio that is independent of n), the sequence $\{w^{(n)}\}_{n=1,2,\ldots}$ is a triangular array of $w_{j,n}$ as $n \to \infty$. For convenience, we denote $W = \{w^{(n)}\}_{n=1,2,\ldots}$ as this array. We write $\mathrm{MSE}_1^{gen}\left(\theta(\cdot), \tilde{d}, w^{(n)}\right)$ as the MSE of $\hat{\theta}_n^{gen} = \hat{\theta}_n^{gen}(\nu)$, where $\nu = \left(\tilde{d}, w^{(n)}\right)$ is the collection of tuning parameters that $\hat{\theta}_n^{gen}$ depends on, and recall $\mathrm{MSE}_0(\theta(\cdot), d)$ as the MSE of the baseline estimator $\bar{\theta}_n = \bar{\theta}_n(d)$. We define

$$R^{gen}\left(\theta(\cdot), d, \tilde{d}, W\right) = \limsup_{n\to\infty} \frac{\mathrm{MSE}_1^{gen}\left(\theta(\cdot), \tilde{d}, w^{(n)}\right)}{\mathrm{MSE}_0(\theta(\cdot), d)} \tag{24}$$

as the asymptotic risk ratio between $\hat{\theta}_n^{gen}$ and $\bar{\theta}_n$.

Moreover, we impose a condition on the magnitude of $\tilde{d}$ relative to d. In particular, we restrict $\tilde{d}$ to be at most Kd for some constant $K > 0$. We consider calibration of $\tilde{d}$ as a function $g(\cdot)$ on d. This is equivalent to requiring $g(d) \le Kd$ for any d, for a maximal inflation factor $K > 0$. Denote

$$\mathcal{F}_K = \{g(\cdot) : g(d) \le Kd\}$$

$\mathcal{W}$ as the space of any triangular array, and H as in (10). We consider the AMRR

$$\min_{g(\cdot)\in\mathcal{F}_K, W\in\mathcal{W}} \max_{\theta(\cdot)\in H, d>0} R^{gen}(\theta(\cdot), d, g(d), W)$$

We have the following identification of the AMRR and the characterization of optimal calibration:

Theorem 8. *Under Assumption 1, we have the following:*

1. *The AMRR of $\hat{\theta}_n^{gen}$ satisfies*

$$\min_{g(\cdot)\in\mathcal{F}_K, W\in\mathcal{W}} \max_{\theta(\cdot)\in H, d>0} R^{gen}\left(\theta\left(\cdot\right), d, g\left(d\right), W\right) = \frac{q_1}{q_1+q_2}\frac{1}{K^{2q_2}} \tag{25}$$

2. *The weights $W^* = \left(w_{j,n}^*\right)_{\substack{j=1,\dots,n\\ n=1,2,\dots}}$ that achieve* (25) *is given by*

$$w_{j,n}^* = \frac{\lambda_1^*}{\left(j+n_0\right)^{\frac{q_1+2q_2}{2(q_1+q_2)}}} + \frac{\lambda_2^*}{\left(j+n_0\right)^{\frac{q_2}{q_1+q_2}}}$$

where λ_1^*, λ_2^* are solved by

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \end{bmatrix} \begin{bmatrix} a^* \\ 1 \end{bmatrix} \tag{26}$$

and a^* is an optimal solution to

$$\min_{a:\left(K^{2(q_1+q_2)}-\xi_{11}\right)a^2-2\xi_{12}a-\xi_{22}\geq 0} |a|^{\frac{2q_2}{q_1+q_2}}\left(\xi_{11}a^2+2\xi_{12}a+\xi_{22}\right)^{\frac{q_1}{q_1+q_2}} \tag{27}$$

where

$$\begin{bmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \end{bmatrix} = \begin{bmatrix} \phi\left(1\right) & \phi\left(\frac{q_1+2q_2}{2(q_1+q_2)}\right) \\ \phi\left(\frac{q_1+2q_2}{2(q_1+q_2)}\right) & \phi\left(\frac{q_2}{q_1+q_2}\right) \end{bmatrix}^{-1}$$

and $\phi\left(\kappa\right) = \sum_{j=1}^n \left(j+n_0\right)^{-\kappa}$. Moreover, $g\left(\cdot\right)$ is defined by $g\left(d\right) = Kd$.

Next, we also note the same result if we fix d in the baseline estimator $\bar{\theta}_n$, uniformly for any d:

Corollary 1. *Under the conditions and notations in Theorem 8, we have, for any fixed d,*

$$\min_{\substack{g(\cdot)\in\mathcal{F}_K\\ W\in\mathcal{W}}} \max_{\theta(\cdot)\in H} R^{gen}\left(\theta\left(\cdot\right), d, g\left(d\right), W\right) = \frac{q_1}{q_1+q_2}\frac{1}{K^{2q_2}}$$

which is attained by the weights $W^* = \left(w_{j,n}^*\right)_{\substack{j=1,\dots,n\\ n=1,2,\dots}}$ and setting $g\left(d\right) = Kd$ that achieve the AMRR in part 2 of Theorem 8.

Before we discuss some implications of the results above, we point out that the condition $g\left(d\right) \leq Kd$ is imposed to combat the hidden finite-sample impact of our asymptotic calculations. More precisely, the limiting value in (24) can incur two approximation errors in practice: First, while we have focused on the asymptotic first-order terms in the biases and variances, the second-order terms can play a role. Second, even assuming there are no second-order terms, the finiteness of sample size could still result in a discrepancy between the worst-case ratio and AMRR. As K gets larger, the tuning parameter $g(d)$ in our proposed estimator could get proportionately larger, thus strengthening the finite-sample effects, meaning that more budget n is needed to observe our asymptotic gain. This strengthening is likewise two-fold: First, the parameters δ_j's are larger, thus increasing the second-order effect. Second, the discrepancy between the first-order terms and AMRR also increases. Thus, while theoretically the AMRR gradually decays to zero as $K \to \infty$, such an interpretation should be cautioned with care. In our experiments (Section 7), we will see that simply choosing $K = 1$ gives numerical results largely coinciding with our theoretical calculations under reasonable budget (e.g., $n = 20$), while the results when $K = 3$ or 4 could deviate from the theoretical AMRR unless more sample size is used.

We discuss several implications of Theorem 8. First, the optimal weighting sequence $w^*_{j,n}$ comprises two components, each with a different decay rate, i.e., $\frac{q_1+2q_2}{2(q_1+q_2)}$ and $\frac{q_2}{q_1+q_2}$ respectively. The coefficients in these decays, namely λ^*_1 and λ^*_2, depend on n that is solved via a linear system of equations, which ensures that neither of the two components in $w^*_{j,n}$ is asymptotically negligible.

To illustrate the latter point, we demonstrate the asymptotic behaviors of λ^*_1, λ^*_2, which are revealed by first understanding the behavior of a^* and using (26). Note that $\phi(\kappa) \sim \frac{1}{1-\kappa} n^{1-\kappa}$ for $\kappa < 1$ and $\sim \log n$ for $\kappa = 1$, where $a_n \sim b_n$ represents asymptotic equivalence between two sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, i.e., $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$. Thus, the matrix

$$\begin{aligned}
\begin{bmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \end{bmatrix} &= \begin{bmatrix} \phi(1) & \phi\left(\frac{q_1+2q_2}{2(q_1+q_2)}\right) \\ \phi\left(\frac{q_1+2q_2}{2(q_1+q_2)}\right) & \phi\left(\frac{q_2}{q_1+q_2}\right) \end{bmatrix}^{-1} \\
&\sim \begin{bmatrix} \log n & \frac{2(q_1+q_2)}{q_1} n^{\frac{q_1}{2(q_1+q_2)}} \\ \frac{2(q_1+q_2)}{q_1} n^{\frac{q_1}{2(q_1+q_2)}} & \frac{q_1+q_2}{q_1} n^{\frac{q_1}{q_1+q_2}} \end{bmatrix}^{-1} \\
&= \frac{1}{\frac{q_1+q_2}{q_1} n^{\frac{q_1}{q_1+q_2}} \log n - \frac{4(q_1+q_2)^2}{q_1^2} n^{\frac{q_1}{q_1+q_2}}} \begin{bmatrix} \frac{q_1+q_2}{q_1} n^{\frac{q_1}{q_1+q_2}} & -\frac{2(q_1+q_2)}{q_1} n^{\frac{q_1}{2(q_1+q_2)}} \\ -\frac{2(q_1+q_2)}{q_1} n^{\frac{q_1}{2(q_1+q_2)}} & \log n \end{bmatrix}
\end{aligned} \tag{28}$$

where the asymptotic equivalence "$\sim$" is on every entry of the matrix.

Now, conjecturing that a^* is of order $n^{-\frac{q_1}{2(q_1+q_2)}}$, we write $a=\tilde{a}n^{-\frac{q_1}{2(q_1+q_2)}}$. By plugging in (28), we have

$$\begin{aligned}\xi_{11}a^2+2\xi_{12}a+\xi_{22} &= \begin{bmatrix}\frac{\tilde{a}}{n^{\frac{q_1}{2(q_1+q_2)}}} & 1\end{bmatrix}\begin{bmatrix}\xi_{11} & \xi_{12}\\ \xi_{21} & \xi_{22}\end{bmatrix}\begin{bmatrix}\frac{\tilde{a}}{n^{\frac{q_1}{2(q_1+q_2)}}}\\ 1\end{bmatrix}\\ &\sim \frac{1}{n^{\frac{q_1}{q_1+q_2}}}\begin{bmatrix}\tilde{a} & 1\end{bmatrix}\begin{bmatrix}0 & 0\\ 0 & \frac{q_1}{q_1+q_2}\end{bmatrix}\begin{bmatrix}\tilde{a}\\ 1\end{bmatrix}\\ &= \frac{q_1}{q_1+q_2}\frac{1}{n^{\frac{q_1}{q_1+q_2}}}\end{aligned}$$

Thus, as $n\to\infty$, an "asymptotic" version of (27), when multiplying the objective value by $n^{\frac{q_1}{q_1+q_2}}$, becomes

$$\min_{\tilde{a}:K^{2(q_1+q_2)}\tilde{a}^2\geq\frac{q_1}{q_1+q_2}} |\tilde{a}|^{\frac{2q_2}{q_1+q_2}}\left(\frac{q_1}{q_1+q_2}\right)^{\frac{q_1}{q_1+q_2}}$$

which gives $|\tilde{a}|=\sqrt{\frac{q_1}{q_1+q_2}}\frac{1}{K^{q_1+q_2}}$. This implies that

$$a^*\sim\sqrt{\frac{q_1}{q_1+q_2}}\frac{1}{K^{q_1+q_2}}n^{-\frac{q_1}{2(q_1+q_2)}} \tag{29}$$

Thus, putting (28) and (29) into (26), we obtain that

$$\lambda_1^*\sim\left(\sqrt{\frac{q_1}{q_1+q_2}}\frac{1}{K^{q_1+q_2}}-2\right)\frac{1}{n^{\frac{q_1}{2(q_1+q_2)}}\log n} \tag{30}$$

and

$$\lambda_2^*\sim\frac{q_1}{q_1+q_2}n^{-\frac{q_1}{q_1+q_2}} \tag{31}$$

We can now see that both terms in $w_{j,n}^*$, namely $\frac{\lambda_1^*}{(j+n_0)^{\frac{q_1+2q_2}{2(q_1+q_2)}}}$ and $\frac{\lambda_2^*}{(j+n_0)^{\frac{q_2}{q_1+q_2}}}$, contribute to the first-order bias. Note that the first-order bias is of order $\sum_{j=1}^n w_{j,n}\delta_j^{q_1}$, where $\delta_j=\tilde{d}(j+n_0)^{-\alpha}$ and $\alpha=\frac{1}{2(q_1+q_2)}$. Thus, using (30), the bias contribution from the first component in $w_{j,n}^*$ gives rise to an order

$$\begin{aligned}\frac{1}{n^{\frac{q_1}{2(q_1+q_2)}}\log n}\sum_{j=1}^n\frac{1}{(j+n_0)^{\frac{q_1+2q_2}{2(q_1+q_2)}}}\frac{1}{(j+n_0)^{\frac{q_1}{2(q_1+q_2)}}} &= \frac{1}{n^{\frac{q_1}{2(q_1+q_2)}}\log n}\sum_{j=1}^n\frac{1}{j+n_0}\\ &= \Theta\left(n^{-\frac{q_1}{2(q_1+q_2)}}\right)\end{aligned} \tag{32}$$

On the other hand, using (31), the bias contribution from the second component in $w_{j,n}^*$ gives rise to an order

$$\frac{1}{n^{\frac{q_1}{q_1+q_2}}}\sum_{j=1}^{n}\frac{1}{(j+n_0)^{\frac{q_2}{q_1+q_2}}}\frac{1}{(j+n_0)^{\frac{q_1}{2(q_1+q_2)}}}=\frac{1}{n^{\frac{q_1}{q_1+q_2}}}\sum_{j=1}^{n}\frac{1}{(j+n_0)^{\frac{q_1+2q_2}{2(q_1+q_2)}}}$$
$$=\Theta\left(n^{-\frac{q_1}{2(q_1+q_2)}}\right)$$

which is the same order as (32). Thus both terms in $w_{j,n}^*$ contribute significantly to the first-order bias term.

Similarly, the first-order variance is of order $\sum_{j=1}^n w_{j,n}^2\delta_j^{-2q_2}$. Using (30), the contribution from the first component in $w_{j,n}^*$ gives rise to an order

$$\frac{1}{n^{\frac{q_1}{q_1+q_2}}(\log n)^2}\sum_{j=1}^{n}\frac{1}{(j+n_0)^{\frac{q_1+2q_2}{q_1+q_2}}}(j+n_0)^{\frac{q_2}{q_1+q_2}}=\frac{1}{n^{\frac{q_1}{q_1+q_2}}(\log n)^2}\sum_{j=1}^{n}\frac{1}{j+n_0}$$
$$=\Theta\left(\frac{1}{n^{\frac{q_1}{q_1+q_2}}\log n}\right) \tag{33}$$

and, using (31), the contribution from the second component gives rise to an order

$$\frac{1}{n^{\frac{2q_1}{q_1+q_2}}}\sum_{j=1}^{n}\frac{1}{(j+n_0)^{\frac{2q_2}{q_1+q_2}}}(j+n_0)^{\frac{q_2}{q_1+q_2}}=\frac{1}{n^{\frac{2q_1}{q_1+q_2}}}\sum_{j=1}^{n}\frac{1}{(j+n_0)^{\frac{q_2}{q_1+q_2}}}$$
$$=\Theta\left(n^{-\frac{q_1}{q_1+q_2}}\right)$$

which has an order larger than (33) by a logarithmic factor. Thus, considering also the cross term between the two components in $w_{j,n}^*$ in the expansion of the variance, the first-order variance is of order $n^{-\frac{q_1}{q_1+q_2}}$, which is the same as the squared bias.

Next we present some basic numerical values of the AMRR. Table 3 shows the values of the AMRR for various maximal inflation factor K when $q_1=2$ and $q_2=1$ (the CFD case without CRN). The AMRR is non-increasing in K, as advocated in Theorem 8 and attributed to more optimizing power for the proposed estimator in the asymptotic limit as K increases (however, as discussed before, we should be cautious about finite-sample distortions). The critical threshold of K above which $\hat{\theta}_n^{gen}$ is guaranteed to improve over $\bar{\theta}_n$ is $K=\sqrt{\frac{2}{3}}=0.82$. In particular, when

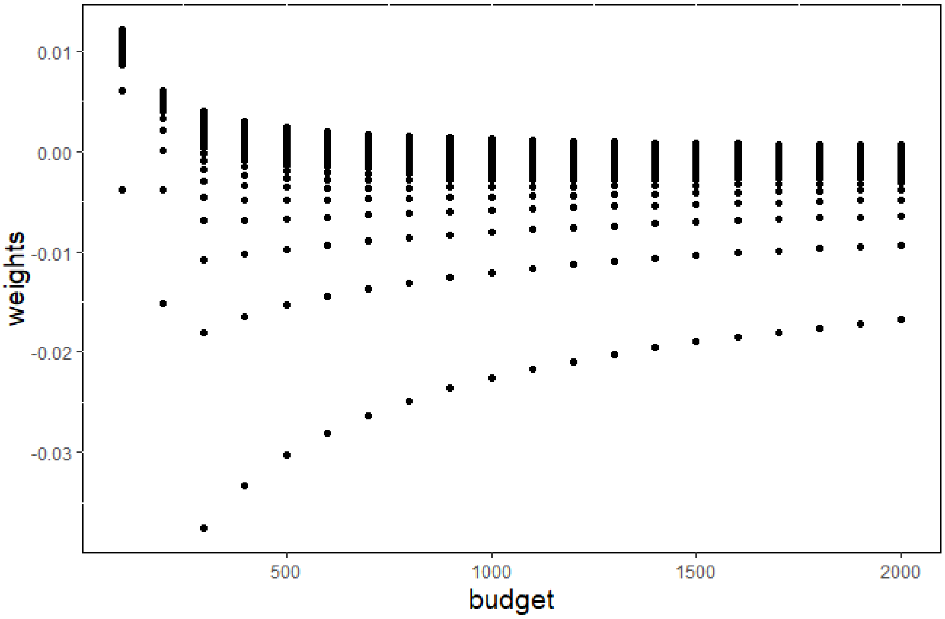


Figure 1 Distribution of weights, with $K=1$, and budget n from 100 to 2000, when $q_1=2, q_2=1$

$K=1$ (we only allow choosing $\tilde{d}$ as large as d at most), we have the AMRR equal to $\frac{2}{3}$, which is strictly less than 1. In other words, no matter what are the values of the model unknowns, the optimized calibration of $\hat{\theta}_n^{gen}$, in particular the two-decay weights $\{w_{j,n}^*\}_{j=1,\dots,n}$ and setting $\tilde{d}=d$, would achieve a better MSE than $\bar{\theta}_n$ asymptotically.

Figures 1 and 2 show the behaviors of the optimal weights for $K=1$. Figure 1 shows that in general the weights range across positive and negative numbers, with higher concentration around 0 as the budget increases. Figure 2 shows that, against the simulation run index, the weight starts from the most negative and gradually increases to the positive region. Lastly, Table 4 shows the AMRR when $q_1=1, q_2=1$ (the FFD and BFD cases without CRN) as a comparison. The AMRR in this case has the same decay rate and is smaller than that for $q_1=2, q_2=1$ across all K.

K	0.5	0.82	1.0	2.0	3.0	4.0
AMRR	2.67	1.00	0.67	0.17	0.07	0.04

Table 3 AMRR for general weighted estimators, against K, when $q_1=2, q_2=1$

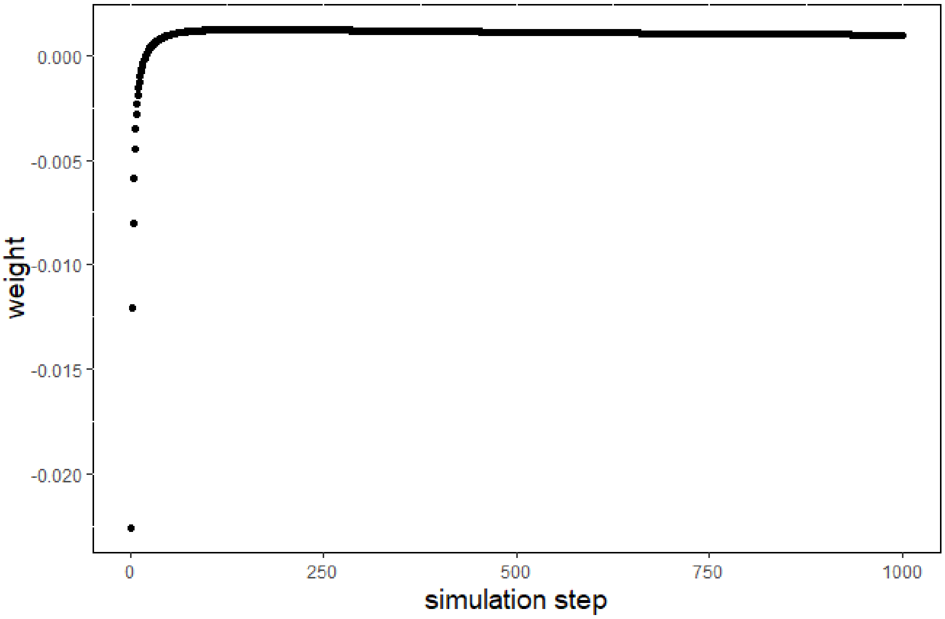


Figure 2 Distribution of weights against simulation step, with $K = 1$, and budget $n = 1000$, when $q_1 = 2, q_2 = 1$

K	0.5	0.71	1.0	2.0	3.0	4.0
AMRR	2.00	1.00	0.50	0.13	0.06	0.03

Table 4 AMRR for general weighted estimators, against K, when $q_1 = 1, q_2 = 1$

5.2. Constrained Optimization for Bias-Variance Balancing

We explain intuitively the key arguments that lead to the optimal two-decay weights $w^*_{j,n}$ and the identification of the AMRR in the form depicted in Theorem 8. We first note that to avoid arbitrarily large value of R^{gen}, the sequence $w_{j,n}$ must sum up to 1 (up to a vanishing error), since otherwise the scenario where $\theta(\cdot)$ has no bias and noise but θ_0 is arbitrarily big will blow up R^{gen}.

Thus, for simplicity let us assume that $\sum_{j=1}^n w_{j,n} = 1$. Also, for convenience, we shorthand w_j as $w_{j,n}$, and w as $w^{(n)}$ when no confusion arises. Moreover, without loss of generality, here we assume $n_0 = 0$ for notational convenience. Considering the bias and variance of $\sum_{j=1}^n w_j\theta_j(\delta_j)$, we can write

$$\begin{aligned}\mathrm{MSE}_1^{gen}\left(\theta(\cdot), \tilde{d}, w\right) &= \left(\sum_{j=1}^n w_j b(\delta_j)\right)^2 + \sum_{j=1}^n w_j^2 Var\left(v(\delta_j)\right)\\ &= \left(\sum_{j=1}^n w_j\left(B\frac{\tilde{d}^{q_1}}{j^{\alpha q_1}} + o\left(\frac{1}{j^{\alpha q_1}}\right)\right)\right)^2 + \sum_{j=1}^n w_j^2 \frac{\sigma^2(1+o(1))\, j^{2\alpha q_2}}{\tilde{d}^{2q_2}}\end{aligned}$$

$$= \left(B\tilde{d}^{q_1} \sum_{j=1}^{n} \frac{w_j}{j^{\alpha q_1}} \right)^2 + \frac{\sigma^2}{\tilde{d}^{2q_2}} \sum_{j=1}^{n} j^{2\alpha q_2} w_j^2 + \text{error} \tag{34}$$

Recall the discussion in Section 4.3. To control the adversary from increasing R^{gen}, we attempt to maintain the relative balance of bias and variance in a similar manner as the baseline. More specifically, presuming that $\hat{\theta}_n^{gen}$ exhibits the optimal MSE order $n^{-\frac{q_1}{q_1+q_2}}$, we keep the ratios of the coefficients in front of B^2 and σ^2 of the first-order MSE terms, between $\hat{\theta}_n^{gen}$ and $\bar{\theta}_n$, to be the same. The coefficient of the squared bias term is roughly

$$n^{\frac{q_1}{q_1+q_2}} \left(\tilde{d}^{q_1} \sum_{j=1}^{n} \frac{w_j}{j^{\alpha q_1}} \right)^2$$

while the coefficient of the variance term is roughly

$$n^{\frac{q_1}{q_1+q_2}} \frac{1}{\tilde{d}^{2q_2}} \sum_{j=1}^{n} j^{2\alpha q_2} w_j^2$$

Thus, similar to (20), we would like to ensure

$$n^{\frac{q_1}{q_1+q_2}} \left(\left(\frac{\tilde{d}}{d} \right)^{q_1} \sum_{j=1}^{n} \frac{w_j}{j^{\alpha q_1}} \right)^2 = n^{\frac{q_1}{q_1+q_2}} \frac{1}{\left(\frac{\tilde{d}}{d} \right)^{2q_2}} \sum_{j=1}^{n} j^{2\alpha q_2} w_j^2 \tag{35}$$

Denoting $\eta = \frac{\tilde{d}}{d}$, and dropping $n^{\frac{q_1}{q_1+q_2}}$ on both sides of (35), we consider the optimization problem

$$\begin{aligned} \min\nolimits_{w,\eta} \quad & S \\ \text{subject to } & S = \left(\eta^{q_1} \textstyle\sum_{j=1}^{n} \frac{w_j}{j^{\alpha q_1}} \right)^2 = \frac{1}{\eta^{2q_2}} \textstyle\sum_{j=1}^{n} j^{2\alpha q_2} w_j^2 \\ & \eta \le K \\ & \textstyle\sum_{j=1}^{n} w_j = 1 \end{aligned} \tag{36}$$

Note that the first constraint is the bias-variance-balancing condition as in (21). The second and third constraints capture the inflation condition $g(\cdot) \in \mathcal{F}_K$ and $\sum_{j=1}^{n} w_j = 1$. Denote the optimal value of (36) as S_n^*. Then roughly speaking, the AMRR would be $\lim_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} S_n^*$. The associated optimal solution w, η turns out to dominate any other possibilities, in particular those obtained by allowing any of the bias and variance terms dominate another.

In the rest of this subsection, we will explain how (36) leads to the two-decay representation of $w_{j,n}^*$, and leave other details to Appendix C. Note that (36) is non-convex. However, we can reformulate it into a convex program together with a simple one-dimensional line search over a region that consists of at most two intervals.

To this end, first notice that from the first constraint in (36), we have

$$\eta = \left(\frac{\sum_{j=1}^n j^{2\alpha q_2} w_j^2}{\left(\sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right)^2} \right)^{\frac{1}{2(q_1+q_2)}} \tag{37}$$

so that the second constraint is equivalent to

$$\sum_{j=1}^n j^{2\alpha q_2} w_j^2 \le K^{2(q_1+q_2)} \left(\sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right)^2$$

Moreover, by plugging in (37) to either expression of S in the first constraint of (36), the objective function becomes

$$\left| \sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right|^{\frac{2q_2}{q_1+q_2}} \left(\sum_{j=1}^n j^{2\alpha q_2} w_j^2 \right)^{\frac{q_1}{q_1+q_2}}$$

Therefore, (36) can be rewritten as

$$\begin{aligned} \min_w \quad & \left| \sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right|^{\frac{2q_2}{q_1+q_2}} \left(\sum_{j=1}^n j^{2\alpha q_2} w_j^2 \right)^{\frac{q_1}{q_1+q_2}} \\ \text{subject to } & \sum_{j=1}^n j^{2\alpha q_2} w_j^2 \le K^{2(q_1+q_2)} \left(\sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right)^2 \\ & \sum_{j=1}^n w_j = 1 \end{aligned} \tag{38}$$

To reduce (38) into a more tractable form, we introduce the variable $a = \sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}}$, and write (38) as

$$\begin{aligned} \min_{w,a} \quad & |a|^{\frac{2q_2}{q_1+q_2}} \left(\sum_{j=1}^n j^{2\alpha q_2} w_j^2 \right)^{\frac{q_1}{q_1+q_2}} \\ \text{subject to } & \sum_{j=1}^n j^{2\alpha q_2} w_j^2 \le K^{2(q_1+q_2)} a^2 \\ & \sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} = a \\ & \sum_{j=1}^n w_j = 1 \end{aligned} \tag{39}$$

Now we decompose the minimization in (39) into two layers, first minimizing w given a, and then minimizing a. This way, (39) can be rewritten as

$$\min_a |a|^{\frac{2q_2}{q_1+q_2}} Z_n^*(a)^{\frac{2q_1}{q_1+q_2}} \tag{40}$$

where

$$\begin{aligned} Z_n^*(a) = \min_w \qquad & \left(\sum_{j=1}^n j^{2\alpha q_2} w_j^2\right)^{\frac{1}{2}} \\ \text{subject to } & \sum_{j=1}^n j^{2\alpha q_2} w_j^2 \le K^{2(q_1+q_2)} a^2 \\ & \sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} = a \\ & \sum_{j=1}^n w_j = 1 \end{aligned} \tag{41}$$

Note that (41) is a quadratic program. We write it in a simpler form as

$$\begin{aligned} \min_w \qquad & \|\Sigma^{\frac{1}{2}} w\| \\ \text{subject to } & \|\Sigma^{\frac{1}{2}} w\|^2 \le K^{2(q_1+q_2)} a^2 \\ & \mu^\top w = a \\ & \mathbb{1}^\top w = 1 \end{aligned} \tag{42}$$

where $\Sigma = \operatorname{diag}\left(j^{2\alpha q_2}\right)_{j=1,\ldots,n} \in \mathbb{R}^{n\times n}$, $\mu = \left(j^{-\alpha q_1}\right)_{j=1,\ldots,n} \in \mathbb{R}^n$, $\mathbb{1} = (1)_j \in \mathbb{R}^n$, and $\|\cdot\|$ is the L_2-norm. We can further separate out the first constraint in (42) and consider the rest of the optimization. To this end, denote

$$\begin{aligned} \tilde{Z}_n^*(a) = \min_w \qquad & \|\Sigma^{\frac{1}{2}} w\| \\ \text{subject to } & \mu^\top w = a \\ & \mathbb{1}^\top w = 1 \end{aligned} \tag{43}$$

If $\tilde{Z}_n^*(a) \le K^{2(q_1+q_2)} a^2$, this means the w that solves (43) is a feasible solution to (42) and, since it is optimal without the first constraint, it then must be optimal too for the entire optimization in (42). Moreover, in this case $Z_n^*(a) = \tilde{Z}_n^*(a)$. Otherwise, if $\tilde{Z}_n^*(a) > K^{2(q_1+q_2)} a^2$, then there is no w that can satisfy the first constraint in (42) simultaneously with the second and third constraints, and thus (42) is infeasible. Therefore, we have

$$Z_n^*(a) = \begin{cases} \tilde{Z}_n^*(a) & \text{if } \tilde{Z}_n^*(a)^2 \le K^{2(q_1+q_2)} a^2 \\ \infty & \text{otherwise} \end{cases} \tag{44}$$

Putting in (44), optimization problem (40) becomes

$$\min_{a:\tilde{Z}_n^*(a)^2 \le K^{2(q_1+q_2)} a^2} |a|^{\frac{2q_2}{q_1+q_2}} \tilde{Z}_n^*(a)^{\frac{2q_1}{q_1+q_2}} \tag{45}$$

Thus, our strategy to solve (36) is to first solve for an optimal solution $w^*(a) = \left(w_j^*(a^*)\right)_{j=1,\dots,n}$ to (43) and obtain $\tilde{Z}_n^*(a)$, and then conduct a line search for a in (45). An optimal calibration configuration is given by the weighting sequence $w^*(a^*)$, where a^* is an optimal solution to (45), and η^*, where

$$\eta^* = \left(\frac{\sum_{j=1}^n j^{2\alpha q_2} w_j^*(a^*)^2}{\left(\sum_{j=1}^n \frac{w_j^*(a^*)}{j^{\alpha q_1}}\right)^2}\right)^{\frac{1}{2(q_1+q_2)}} \tag{46}$$

by using (37).

The two-decay characterization of the weighting sequence arises from the solution to (43). To illustrate, consider the Lagrangian

$$\|\Sigma^{\frac{1}{2}} w\| - \lambda_1\left(\mu^\top w - a\right) - \lambda_2\left(\mathbb{1}^\top w - 1\right)$$

Differentiating with respect to w and equating to 0, we get

$$\frac{\Sigma w}{\|\Sigma^{\frac{1}{2}} w\|} - \lambda_1 \mu - \lambda_2 \mathbb{1} = 0$$

which gives

$$w = \Sigma^{-1}\left(\lambda_1 \mu + \lambda_2 \mathbb{1}\right) = \lambda_1 \Sigma^{-1}\mu + \lambda_2 \Sigma^{-1}\mathbb{1}$$

for some λ_1, λ_2 (scaled by $\|\Sigma^{\frac{1}{2}} w\|$ compared to the ones displayed before). Note that this is equivalent to

$$w_j = \frac{\lambda_1}{j^{\alpha(q_1+2q_2)}} + \frac{\lambda_2}{j^{2\alpha q_2}} \tag{47}$$

for $j = 1, \dots, n$. This is precisely the form of $w_{j,n}^*$ in Theorem 8. By identifying λ_1 and λ_2 using the constraints in (43), and writing out η^* and $\tilde{Z}_n^*(a)$, we arrive at the depicted choices of w and $g(\cdot)$ in the theorem. The remainder of the argument comprises an analysis to show that no other choices of w and $g(\cdot)$ can give a better asymptotic minimax ratio, via comparing with an alternate optimization problem and demonstrating that the residual error induced by $w_{j,n}^*$ and η^* in (34) is indeed of higher order. Appendix C shows the details.

6. Multivariate Generalizations

All results we have presented apply to the multivariate version of (1). For convenience, we adopt the notations there. We are interested in estimating $\boldsymbol{\theta}_0 \in \mathbb{R}^p$. Given a tuning parameter $\delta \in \mathbb{R}_+$, we can run Monte Carlo simulation where each simulation run outputs

$$\boldsymbol{\theta}(\delta) = \boldsymbol{\theta}_0 + \mathbf{b}(\delta) + \mathbf{v}(\delta) \tag{48}$$

with $\mathbf{b}(\delta) = \mathbf{B}\delta^{q_1} + o(\delta^{q_1})$ as $\delta \to 0$, $\mathbf{v}(\delta) = \frac{\boldsymbol{\varepsilon}(\delta)}{\delta^{q_2}}$, and $q_1, q_2 > 0$. We assume that:

Assumption 2. *We have*

1. $\mathbf{B} \in \mathbb{R}^p$ *is a non-zero constant vector.*
2. $\boldsymbol{\varepsilon}(\delta) \in \mathbb{R}^p$ *is a family of random vectors such that* $E\boldsymbol{\varepsilon}(\delta) = 0$ *and* $\lim_{\delta \to 0} Cov(\boldsymbol{\varepsilon}(\delta)) = \Sigma$ *for some positive semidefinite matrix* Σ *with* $tr(\Sigma) > 0$.

The constructions of the considered estimators are generalized in a natural manner. Namely, the sample-average-based estimator $\bar{\boldsymbol{\theta}}_n$ is obtained by taking the average of n vectors of $\boldsymbol{\theta}(\delta)$. The recursive estimator (14) is obtained in a vectorized form, where the step size $\gamma_n \in \mathbb{R}_+$ is still in the form $c(n+n_0)^{-\beta}$ and $\delta_n = \tilde{d}(n+n_0)^{-\alpha}$. Similar vectorization holds for the averaging estimator (15). Lastly, the general weighted estimator in (23) can also be defined in a vectorized form, with $\{w_{j,n}\}_{j=1,\ldots,n,\ n=1,2,\ldots}$ still a triangular array of weights.

To gauge the error of an estimator $\hat{\boldsymbol{\theta}}_n$, we use the MSE given by $E\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2$. Note that we can decompose this into bias and variance in L_2, namely $\|E\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 + tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n\right)\right)$. With this definition of MSE, the asymptotic risk ratios (11) and (13) can be similarly defined. Then all the results in Sections 3, 4 and 5 hold with only cosmetic changes. Appendices A and B show the multivariate version of the theorems and proofs in Sections 3 and 4, while it will be clear from the developments in Appendix C that the multivariate analog of Theorem 8 follows from its proof directly (essentially, by replacing B^2 with $\|\mathbf{B}\|^2$ and σ^2 with $\mathrm{tr}(\Sigma)$).

Multivariate estimators in the form (48) arise in, for example, zeroth order gradient estimator using simultaneous perturbation (Spall (1992)). To estimate $\nabla f(\mathbf{x})$, a sample output would involve first simulating a random vector, say $\mathbf{h} = (h_i)_{i=1,\ldots,p} \in \mathbb{R}^p$, then generating two unbiased simulation runs $\hat{f}(\mathbf{x}+\delta\mathbf{h})$ and $\hat{f}(\mathbf{x}-\delta\mathbf{h})$, and finally outputting, for each direction i,

$$\frac{\hat{f}(\mathbf{x}+\delta\mathbf{h}) - \hat{f}(\mathbf{x}-\delta\mathbf{h})}{2\delta h_i} \tag{49}$$

where $\delta > 0$ is the perturbation size. This scheme satisfies (48) with $q_1 = 2, q_2 = 1$ by choosing h to have mean-zero, independent components with finite inverse second moments, and under enough smoothness conditions on f. One can also use several variants of (49) to obtain similar conclusions, for example the one-sided version $\frac{1}{\delta h_i}\hat{f}(\mathbf{x}+\delta\mathbf{h})$ (Spall (1997)), or $\frac{1}{\delta}\hat{f}(\mathbf{x}+\delta\mathbf{h})\,h_i$ by choosing $\mathbf{h}$ to satisfy other types of conditions, as in Gaussian smoothing (Nesterov and Spokoiny (2017)) or uniform sampling (Flaxman et al. (2005)).

Moreover, one important application of the above multivariate estimators concerns input uncertainty quantification (e.g., Barton (2012), Henderson (2003), Chick (2006), Song et al. (2014), Lam (2016)). In particular, a common estimation target in this problem is the output variance of a simulation experiment that is contributed from the statistical noises of the input models calibrated from external data sources, which is typically expressed in the form $\nabla\psi(\mathbf{x})^\top\Lambda\nabla\psi(\mathbf{x})$ where Λ is the sampling covariance of the estimates of the input parameter vector $\mathbf{x} \in \mathbb{R}^p$, $\nabla\psi(\mathbf{x})$ is the gradient of the simulation performance measure with respect to $\mathbf{x}$, and $^\top$ denotes transpose. Thus, this is in the form of $G(\boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0 = \nabla\psi(\mathbf{x})$ and $G(\boldsymbol{\theta}_0) = \boldsymbol{\theta}_0^\top\Lambda\boldsymbol{\theta}_0$. Our results applies to estimate $G(\boldsymbol{\theta}_0)$ with a plug-in of $\boldsymbol{\theta}_0$ and a standard application of the delta method to control the inherited error.

7. Numerical Results

We conduct a set of experiments to test the theoretical results derived in this paper. We consider several variants of an $M/M/1$ queueing system and target performance measures. Specifically, we have the following:

- *Case 1: Critically loaded queue and transient performance measure.* We set the arrival and service rates to be both 4, so that the system is critically loaded. The queue is initially empty. We consider a transient performance measure of the expected averaged system time of the first 10 customers, and are interested in the gradient of this quantity with respect to the arrival and service rates. Here, the true derivatives with respect to these rates are 0.0946 and −0.2501 respectively, which are calculated by the likelihood ratio / score function method (e.g., Glynn (1990), Rubinstein (1986), Reiman and Weiss (1989)) with 1 million simulation repetitions.

- *Case 2: Non-critically loaded queue and transient performance measure.* We set the arrival rate to be 3 and the service rate to be 5, so that the system is not critically loaded. The queue is initially empty. We consider the same performance measure and target gradient as the setting above. Here, the true derivatives with respect to the arrival and service rates are 0.0676 and −0.1136 respectively.
- *Case 3: Non-critically loaded queue and steady-state performance measure.* We set the arrival rate to be 3 and the service rate to be 5, so that the system is not critically loaded. The queue is initially empty. We consider a steady-state performance measure of the expected averaged system time of the first 1000 customers, and are interested in the gradient of this quantity with respect to the arrival and service rates. The true derivatives with respect to these rates are 0.2746 and −0.2440 respectively.

In our experiments we assume these systems or performance measures can be simulated only through black box, i.e., we cannot introduce effective coupling among simulation runs that allows one to use unbiased derivative estimators or multilevel Monte Carlo (however, we use unbiased derivative estimator, via the likelihood ratio / score function method, to obtain the ground truth in order to calculate MSEs). For each system and target performance measure above, we consider two settings. The first setting uses CFD to estimate the derivative with respect to the arrival rate. The second setting uses simultaneous perturbation (described in Section 6), with the perturbation vector $\mathbf{h}$ having each entry being independent symmetric variable on ±1, to estimate the gradient with respect to the arrival and service rates simultaneously. In each setting, we consider three estimators: 1) The conventional sample-average-based estimator $\bar{\theta}_n$; 2) the recursive estimator $\hat{\theta}_n^{rec}$; and 3) the general weighted estimator $\hat{\theta}_n^{gen}$. In $\bar{\theta}_n$, we set $\delta = d(n+n_0)^{-\frac{1}{6}}$ where $d=1$. In $\hat{\theta}_n^{rec}$, we set $c=1$, $\delta_j = \tilde{d}(j+n_0)^{-\frac{1}{6}}$ for the j-th simulation run, where $\tilde{d} = 3^{-\frac{1}{6}}d = 0.83d$. For $\hat{\theta}_n^{gen}$, we set $\delta_j = \tilde{d}(j+n_0)^{-\frac{1}{6}}$ where $\tilde{d} = \eta^* d$, and use weights $w_{j,n}^*$, with $\eta^* = K$ and $w_{j,n}^*$ both chosen according to Theorem 8. Moreover, we consider values of K ranging from 0.5 to 4 among different settings, which correspond to the values shown in Table 3. For each experimental setting, we consider simulation run-lengths n varying between 20 and 1000, with n_0 fixed to be 5. We repeat the simulation for 1000 times to estimate the empirical MSEs. Moreover, we output the 95% confidence intervals for the risk ratios, which are obtained by a standard application of the delta method.

Tables 5-10 summarize the results for the derivative estimation with respect to the arrival rate and the two-dimensional gradient estimation with respect to both the arrival and the service rates, respectively for Cases 1-3 above. Note that in interpreting these tables, one should focus on the risk ratios instead of the absolute magnitude of the derivatives. This is because we can always artificially inflate or deflate the values by simply multiplying the considered performance measures by a scalar. Thus, an appropriate measurement of the estimation error should be the relative error, namely MSE/(true value), where the denominator is canceled out in the risk ratio calculation. We see that, across all estimation settings in Cases 1 and 2 (Tables 5-8), the empirical risk ratios between the recursive estimator $\hat{\theta}_n^{rec}$ and the baseline $\bar{\theta}_n$ are stably around 0.96 ($n=700$ in Table 6) to 1.21 ($n=1000$ in Table 6) when the budget n is at least 100, while they range from 0.81 ($n=20$ in Table 5) to 1.03 ($n=50$ in Table 7) when n is 20 or 50. These behave quite consistently with the theoretical prediction of 1.08. For Case 3, the risk ratios can range from 0.27 to 6.36 ($n=20$ in Tables 10 and 9) for small n, but as n increases towards 1000 the ratio appears to converge to roughly 1.0-1.2 when estimating the derivative with respect to the arrival rate (Table 9), and to roughly 0.8-1.0 when estimating the gradient with respect to both arrival and service rates (Table 10). These results also appear to match our AMRR prediction of 1.08.

Next we discuss the general weighted estimator $\hat{\theta}_n^{gen}$. Its risk ratios vary with K. We first consider the transient measures in Cases 1 and 2. When $K=0.5$, the risk ratios across all estimation settings lie around 2.34 ($n=700$ in Table 6) to 3.38 ($n=100,140$ in Table 5), which is roughly around, though could be higher than, the theoretical AMRR value of 2.67. When $K = 0.82$, the ratios are around 0.93 ($n = 500$ in Table 7) to 1.31 ($n = 100$ in Table 5), which become closer to the theoretical AMRR value of 1. When $K = 1$, the ratios are around 0.68 ($n = 900$ in Table 7) to 0.89 ($n = 140$ in Table 5), again becoming closer to the theoretical AMRR value of 0.67. When $K = 2$, we see that in the derivative estimation of arrival rate, the ratios are around 0.15 ($n = 120$ in Table 5) to 0.21 ($n = 50$ in Table 7), which match the theoretical AMRR value of 0.17. However, for the gradient estimation with respect to both rates, the ratios increase to around 0.26 ($n = 1000$ in Table 8) to 0.45 ($n = 20$ in Table 6), indicating (positive) deviation away from the theoretical AMRR. Furthermore, when $K = 3$ or 4, the risk ratios appear a lot less stable, taking values as low as 0.04 ($n = 50$ in Table 5) and as high as 8.59 ($n = 100$ in Table 6). For the steady-state performance measure in Case 3 (Tables 9-10), we see that the trends for $K = 0.5$ to 1 behave roughly similar to Cases 1 and 2, but with higher variability in general. When $K = 0.5$, the risk ratios range between 1.32 ($n = 280$ in Table 9) and 2.74 ($n = 220$ in Table 10), which roughly match the theoretical AMRR of 2.67. However, when $K = 0.82$, the risk ratios take values as low as 0.69 ($n = 280$ in Table 9) and as high as 1.57 ($n = 120$ in Table 10), indicating more deviations away from the theoretical AMRR of 1 than Cases 1 and 2. When $K = 1$, the ratios range from 0.26 ($n = 20$ in Table 10) to 0.81 ($n = 1000$ in Table 9), again indicating more fluctuations away from the theoretical AMRR of 0.67 than Cases 1 and 2 (though the ratios are still lower than 1). The deviations from the theoretical AMRR suggest our asymptotic characterization in some of these cases is not accurate enough to capture the statistical behavior under the considered budget. As described in Section 5.1, these deviations can be attributed to two approximation errors, first the impact of the second-order terms in the biases and variances of estimators, and second, the finiteness of sample size for the first-order terms even if there are no second-order terms in the considered MSE ratio. As K gets larger, these finite-sample effects appear to strengthen and deem the need of a larger sample size to observe the asymptotic behavior. Moreover, though we have fixed $d = 1$ in this set of experiments, the choice of d also appears to have a finite-sample effect, which we illustrate with an additional numerical experiment in Appendix D. Precisely understanding these effects requires analyzing the finite-sample properties of the worst-case risk ratios, which is beyond the scope of this paper but would constitute important future work. Nonetheless, from our experiments, we see that simply choosing $K = 1$ seems to be robust across all of our considered settings.

We present additional experimental results in Appendix E to illustrate how our optimal weighting scheme can be potentially incorporated in zeroth-order stochastic gradient descent or SA to obtain faster convergence for black-box stochastic optimization. We also compare our scheme with some benchmarks. Like the finite-sample investigation, a full study on more efficient stochastic optimization based on the present framework will be left for future work.

8. Conclusion

We have studied a framework to construct new estimators that, in situations where simulation runs are biased for a target estimation quantity, consistently outperform baseline estimators as the sample averages of the simulation runs with a chosen tuning parameter. One challenge in choosing the latter lies in the often lack of knowledge on the model characteristics that affect the bias-variance tradeoff. To mitigate the adversarial impact of this ambiguity, we propose a minimax analysis on the asymptotic risk ratio that compares the mean square errors between proposed estimators and the baseline. In particular, we identify the asymptotic minimax risk ratio (AMRR) and the optimal configurations for recursive estimators and their standard averaging versions. We show that, in typical cases, the AMRR for these estimators are not small enough to justify any outperformance against the standard baseline. We then consider a more general class of weighted estimators, and identify the AMRR that can be significantly reduced to a level that implies that the resulting optimal estimator asymptotically outperforms the baseline, regardless of any realizations of the unknown model characteristics. Moreover, we provide an explicit characterizations of the optimal weights in a two-decay-rate form, and argue how this arises from a balancing of bias-variance that matches the baseline in order to control an adversarial enlargement of the risk ratio.

n	$\bar{\theta}_n$	$\hat{\theta}_n^{rec}$	$\hat{\theta}_n^{gen}, K = 0.5$	$\hat{\theta}_n^{gen}, K = 0.82$
20	1.08E-2	7.72E-3 (81 ± 11%)	3.52E-2 (325 ± 43%)	1.26E-2 (116 ± 15%)
50	5.95E-3	5.63E-3 (95 ± 12%)	1.78E-2 (300 ± 39%)	6.48E-3 (109 ± 14%)
100	3.32E-3	3.55E-3 (107 ± 14%)	1.12E-2 (338 ± 42%)	4.36E-3 (131 ± 17%)
120	3.33E-3	3.43E-3 (103 ± 12%)	8.93E-3 (268 ± 34%)	3.72E-3 (111 ± 13%)
140	2.66E-3	2.90E-3 (109 ± 14%)	9.02E-3 (338 ± 39%)	3.47E-3 (130 ± 16%)
160	2.63E-3	2.57E-3 (97 ± 12%)	7.65E-3 (291 ± 37%)	2.92E-3 (111 ± 13%)
180	2.36E-3	2.48E-3 (105 ± 13%)	6.94E-3 (294 ± 35%)	2.46E-3 (104 ± 12%)
200	2.20E-3	2.38E-3 (108 ± 13%)	6.69E-3 (305 ± 37%)	2.37E-3 (108 ± 14%)
220	2.11E-3	2.24E-3 (106 ± 13%)	6.07E-3 (288 ± 35%)	2.17E-3 (103 ± 13%)
240	2.01E-3	2.13E-3 (106 ± 13%)	5.99E-3 (298 ± 37%)	2.10E-3 (104 ± 14%)
260	1.87E-3	2.10E-3 (112 ± 14%)	5.17E-3 (276 ± 35%)	2.24E-3 (120 ± 15%)
280	1.87E-3	1.92E-3 (97 ± 12%)	5.15E-3 (276 ± 34%)	1.81E-3 (97 ± 12%)
300	1.64E-3	1.80E-3 (110 ± 13%)	4.78E-3 (291 ± 35%)	1.75E-3 (107 ± 13%)
400	1.38E-3	1.54E-3 (112 ± 13%)	4.19E-3 (303 ± 37%)	1.55E-3 (112 ± 12%)
500	1.22E-3	1.28E-3 (105 ± 13%)	3.66E-3 (299 ± 36%)	1.22E-3 (100 ± 13%)
600	1.04E-3	1.18E-3 (113 ± 14%)	3.03E-3 (292 ± 35%)	1.11E-3 (107 ± 13%)
700	9.24E-4	9.83E-4 (106 ± 13%)	2.56E-3 (276 ± 35%)	9.40E-4 (102 ± 13%)
800	8.61E-4	9.14E-4 (106 ± 13%)	2.60E-3 (302 ± 37%)	9.86E-4 (115 ± 14%)
900	8.23E-4	8.54E-4 (104 ± 13%)	2.30E-3 (280 ± 36%)	9.07E-4 (110 ± 14%)
1000	7.17E-4	8.12E-4 (113 ± 14%)	2.03E-3 (283 ± 36%)	7.66E-4 (107 ± 13%)

n	$\hat{\theta}_n^{gen}, K = 1$	$\hat{\theta}_n^{gen}, K = 2$	$\hat{\theta}_n^{gen}, K = 3$	$\hat{\theta}_n^{gen}, K = 4$
20	8.63E-3 (80 ± 10%)	2.13E-3 (20 ± 3%)	9.20E-4 (8 ± 1%)	5.25E-4 (5 ± 1%)
50	4.67E-3 (79 ± 10%)	1.03E-3 (17 ± 2%)	4.36E-4 (7 ± 1%)	2.56E-4 (4 ± 1%)
100	2.79E-3 (84 ± 10%)	6.76E-4 (20 ± 2%)	2.77E-4 (8 ± 1%)	1.99E-3 (60 ± 8%)
120	2.44E-3 (73 ± 9%)	5.16E-4 (15 ± 2%)	2.25E-4 (7 ± 1%)	1.53E-3 (46 ± 6%)
140	2.36E-3 (89 ± 11%)	5.31E-4 (20 ± 3%)	2.11E-4 (8 ± 1%)	1.32E-3 (49 ± 6%)
160	1.97E-3 (75 ± 9%)	4.56E-4 (17 ± 2%)	2.16E-3 (82 ± 10%)	1.16E-3 (42 ± 6%)
180	1.66E-3 (70 ± 8%)	4.51E-4 (19 ± 2%)	2.12E-3 (89 ± 11%)	9.88E-4 (39 ± 6%)
200	1.67E-3 (76 ± 9%)	4.14E-4 (19 ± 2%)	1.78E-3 (81 ± 10%)	8.55E-4 (33 ± 5%)
220	1.57E-3 (74 ± 9%)	3.49E-4 (17 ± 2%)	1.50E-3 (71 ± 9%)	6.99E-4 (32 ± 5%)
240	1.48E-3 (74 ± 9%)	3.50E-4 (17 ± 2%)	1.37E-3 (68 ± 9%)	6.48E-4 (33 ± 4%)
260	1.40E-3 (75 ± 9%)	3.02E-4 (16 ± 2%)	1.26E-3 (67 ± 9%)	6.26E-4 (31 ± 4%)
280	1.37E-3 (73 ± 9%)	3.28E-4 (18 ± 2%)	1.13E-3 (61 ± 7%)	5.81E-4 (34 ± 4%)
300	1.21E-3 (74 ± 9%)	3.21E-4 (20 ± 2%)	1.05E-3 (64 ± 7%)	5.55E-4 (28 ± 4%)
400	1.03E-3 (75 ± 9%)	2.62E-4 (19 ± 2%)	7.89E-4 (57 ± 7%)	3.82E-4 (26 ± 4%)
500	9.31E-4 (76 ± 9%)	2.16E-4 (18 ± 2%)	6.60E-4 (54 ± 6%)	3.19E-4 (25 ± 3%)
600	7.99E-4 (77 ± 9%)	2.00E-4 (19 ± 2%)	4.98E-4 (48 ± 6%)	2.59E-4 (24 ± 3%)
700	7.48E-4 (81 ± 10%)	1.68E-4 (18 ± 2%)	4.02E-4 (43 ± 5%)	2.26E-4 (24 ± 3%)
800	6.36E-4 (74 ± 9%)	1.58E-4 (18 ± 2%)	3.74E-4 (43 ± 5%)	1.91E-4 (22 ± 3%)
900	6.18E-4 (75 ± 10%)	1.57E-4 (19 ± 2%)	3.36E-4 (41 ± 5%)	1.79E-4 (22 ± 2%)
1000	5.76E-4 (80 ± 10%)	1.36E-4 (19 ± 2%)	2.83E-4 (39 ± 5%)	1.42E-4 (20 ± 2%)

Table 5 Empirical MSEs among estimators for the derivative with respect to the arrival rate for Case 1. Bracketed numbers represent the 95% confidence intervals (CIs) for the risk ratios between the considered estimators and the baseline $\bar{\theta}_n$.

Our work opens the door to multiple lines of expansion, in terms of both the formulating framework and the techniques. First is the finite-sample counterpart of our analyses that aims to more accurately capture the second-order effect of the bias-variance balance. Second, our framework can be used to find better estimators for problems where simulation runtime is significantly affected by the tuning parameters, in addition to bias and variance. Third, the statistical inference and construction of confidence intervals/regions of our weighted estimators, which involves analyzing central limit behaviors and the proper design of data-driven schemes like sectioning, are also of

n	$\bar{\theta}_n$	$\hat{\theta}_n^{rec}$	$\hat{\theta}_n^{gen}, K=0.5$	$\hat{\theta}_n^{gen}, K=0.82$	$\hat{\theta}_n^{gen}, K=1$	$\hat{\theta}_n^{gen}, K=2$	$\hat{\theta}_n^{gen}, K=3$	$\hat{\theta}_n^{gen}, K=4$
20	2.75E-2	2.32E-2 (85 ± 7%)	7.99E-2 (291 ± 27%)	3.56E-2 (130 ± 12%)	2.42E-2 (88 ± 8%)	1.25E-2 (45 ± 4%)	1.91E-2 (70 ± 6%)	5.61E-2 (204 ± 16%)
50	1.42E-2	1.34E-2 (95 ± 8%)	3.94E-2 (278 ± 25%)	1.68E-2 (119 ± 11%)	1.15E-2 (81 ± 7%)	5.85E-3 (41 ± 4%)	8.85E-3 (62 ± 6%)	2.82E-2 (199 ± 15%)
100	8.71E-3	8.81E-3 (101 ± 9%)	2.19E-2 (251 ± 22%)	9.52E-3 (109 ± 9%)	6.52E-3 (75 ± 7%)	3.04E-3 (35 ± 3%)	4.78E-3 (54 ± 5%)	7.49E-2 (859 ± 85%)
120	7.48E-3	7.53E-3 (101 ± 9%)	2.01E-2 (268 ± 23%)	8.85E-3 (118 ± 10%)	5.62E-3 (75 ± 7%)	2.57E-3 (34 ± 3%)	4.44E-3 (59 ± 5%)	5.41E-2 (723 ± 71%)
140	6.43E-3	6.59E-3 (102 ± 9%)	1.81E-2 (280 ± 24%)	7.46E-3 (116 ± 10%)	5.21E-3 (81 ± 7%)	2.38E-3 (37 ± 3%)	3.88E-3 (60 ± 5%)	4.14E-2 (644 ± 61%)
160	5.78E-3	6.02E-3 (104 ± 9%)	1.51E-2 (261 ± 22%)	6.75E-3 (117 ± 10%)	4.58E-3 (79 ± 7%)	2.18E-3 (38 ± 3%)	2.44E-2 (423 ± 40%)	3.47E-2 (601 ± 58%)
180	5.38E-3	5.66E-3 (105 ± 9%)	1.52E-2 (282 ± 25%)	6.33E-3 (103 ± 10%)	4.28E-3 (80 ± 7%)	1.89E-3 (35 ± 3%)	2.05E-2 (381 ± 37%)	2.90E-2 (539 ± 54%)
200	5.27E-3	5.28E-3 (100 ± 9%)	1.37E-2 (261 ± 23%)	5.45E-3 (108 ± 9%)	3.93E-3 (74 ± 6%)	1.73E-3 (33 ± 3%)	1.70E-2 (322 ± 30%)	2.68E-2 (509 ± 49%)
220	4.62E-3	5.00E-3 (108 ± 10%)	1.21E-2 (262 ± 24%)	5.00E-3 (108 ± 10%)	3.62E-3 (78 ± 7%)	1.50E-3 (33 ± 3%)	1.52E-2 (330 ± 32%)	2.40E-2 (520 ± 51%)
240	4.33E-3	4.64E-3 (107 ± 9%)	1.20E-2 (276 ± 24%)	4.69E-3 (107 ± 9%)	3.53E-3 (81 ± 7%)	1.56E-3 (36 ± 3%)	1.39E-2 (321 ± 32%)	2.13E-2 (491 ± 47%)
260	4.03E-3	4.40E-3 (109 ± 9%)	1.09E-2 (270 ± 23%)	4.31E-3 (107 ± 9%)	3.06E-3 (76 ± 6%)	1.48E-3 (37 ± 3%)	1.20E-2 (296 ± 28%)	1.88E-2 (467 ± 45%)
280	3.93E-3	4.17E-3 (106 ± 9%)	1.07E-2 (272 ± 23%)	4.25E-3 (108 ± 9%)	2.96E-3 (75 ± 7%)	1.40E-3 (36 ± 3%)	1.03E-2 (261 ± 23%)	1.80E-2 (456 ± 43%)
300	3.75E-3	4.04E-3 (108 ± 9%)	1.04E-2 (277 ± 24%)	3.82E-3 (102 ± 9%)	2.94E-3 (78 ± 7%)	1.30E-3 (35 ± 3%)	1.02E-2 (272 ± 24%)	1.56E-2 (416 ± 40%)
400	3.21E-3	3.46E-3 (108 ± 9%)	8.52E-3 (265 ± 23%)	3.22E-3 (100 ± 9%)	2.42E-3 (75 ± 7%)	1.00E-3 (31 ± 3%)	7.18E-3 (224 ± 22%)	1.17E-2 (363 ± 36%)
500	2.65E-3	2.71E-3 (102 ± 9%)	6.68E-3 (252 ± 22%)	2.83E-3 (106 ± 10%)	2.04E-3 (77 ± 7%)	8.31E-4 (31 ± 3%)	5.23E-3 (197 ± 19%)	8.77E-3 (331 ± 31%)
600	2.39E-3	2.44E-3 (102 ± 9%)	6.28E-3 (263 ± 23%)	2.42E-3 (101 ± 9%)	1.86E-3 (78 ± 7%)	6.99E-4 (29 ± 3%)	4.43E-3 (186 ± 18%)	6.90E-3 (289 ± 27%)
700	2.23E-3	2.13E-3 (96 ± 8%)	5.20E-3 (234 ± 20%)	2.26E-3 (102 ± 9%)	1.58E-3 (71 ± 6%)	6.56E-4 (29 ± 3%)	3.43E-3 (154 ± 14%)	6.19E-3 (278 ± 26%)
800	2.01E-3	2.08E-3 (103 ± 9%)	5.15E-3 (256 ± 22%)	1.97E-3 (98 ± 9%)	1.52E-3 (76 ± 6%)	5.97E-4 (30 ± 3%)	3.16E-3 (157 ± 14%)	4.89E-3 (243 ± 23%)
900	1.73E-3	1.88E-3 (109 ± 10%)	4.57E-3 (264 ± 23%)	1.85E-3 (107 ± 9%)	1.35E-3 (78 ± 7%)	5.40E-4 (31 ± 3%)	2.64E-3 (152 ± 14%)	4.27E-3 (247 ± 23%)
1000	1.55E-3	1.88E-3 (121 ± 11%)	4.45E-3 (287 ± 25%)	1.79E-3 (115 ± 10%)	1.28E-3 (82 ± 7%)	5.13E-4 (33 ± 3%)	2.28E-3 (147 ± 14%)	4.07E-3 (262 ± 25%)

Table 6 Empirical MSEs among estimators for the gradient with respect to the arrival and service rates for Case 1. Bracketed numbers represent the 95% CIs for the risk ratios between the considered estimators and the baseline $\bar{\theta}_n$.

interest. Lastly, we plan to expand the study on using our enhanced estimators in stochastic black-box optimization where the gradients in a descent algorithm are estimated via finite differences or zeroth-order schemes.

Acknowledgments

We gratefully acknowledge support from the National Science Foundation under grants CMMI-1542020, CMMI-1523453 and CAREER CMMI-1653339/1834710.

n	$\bar{\theta}_n$	$\hat{\theta}_n^{rec}$	$\hat{\theta}_n^{gen}, K=0.5$	$\hat{\theta}_n^{gen}, K=0.82$
20	4.02E-3	3.29E-3 (82 ± 11%)	1.33E-2 (331 ± 42%)	4.85E-3 (121 ± 15%)
50	1.94E-3	1.99E-3 (103 ± 13%)	5.57E-3 (287 ± 37%)	2.21E-3 (114 ± 15%)
100	1.24E-3	1.28E-3 (102 ± 13%)	3.63E-3 (292 ± 37%)	1.43E-3 (114 ± 15%)
120	1.06E-3	1.10E-3 (104 ± 13%)	2.94E-3 (278 ± 35%)	1.25E-3 (118 ± 14%)
140	9.45E-4	1.06E-3 (118 ± 14%)	2.88E-3 (305 ± 38%)	1.20E-3 (126 ± 16%)
160	8.41E-4	9.16E-4 (109 ± 13%)	2.31E-3 (274 ± 40%)	1.01E-3 (120 ± 14%)
180	7.58E-4	9.11E-4 (120 ± 15%)	2.31E-3 (305 ± 35%)	9.08E-4 (120 ± 15%)
200	7.44E-4	8.51E-4 (114 ± 14%)	2.20E-3 (296 ± 35%)	7.91E-4 (106 ± 13%)
220	7.15E-4	7.72E-4 (108 ± 14%)	1.99E-3 (278 ± 36%)	8.00E-4 (112 ± 15%)
240	6.99E-4	7.19E-4 (103 ± 13%)	2.02E-3 (288 ± 36%)	7.35E-4 (105 ± 13%)
260	6.57E-4	7.09E-4 (108 ± 14%)	1.91E-3 (291 ± 36%)	6.61E-4 (101 ± 13%)
280	6.00E-4	6.76E-4 (113 ± 14%)	1.75E-3 (292 ± 36%)	6.90E-4 (115 ± 14%)
300	5.82E-4	6.52E-4 (112 ± 14%)	1.67E-3 (288 ± 34%)	6.46E-4 (111 ± 14%)
400	4.75E-4	5.57E-4 (117 ± 14%)	1.39E-3 (293 ± 36%)	5.14E-4 (108 ± 13%)
500	4.37E-4	4.51E-4 (103 ± 13%)	1.17E-3 (267 ± 36%)	4.07E-4 (93 ± 13%)
600	3.60E-4	3.90E-4 (109 ± 13%)	1.07E-3 (297 ± 36%)	3.83E-4 (107 ± 13%)
700	3.20E-4	3.56E-4 (111 ± 14%)	8.62E-4 (270 ± 32%)	3.77E-4 (118 ± 14%)
800	2.83E-4	3.31E-4 (117 ± 14%)	8.29E-4 (293 ± 37%)	3.14E-4 (111 ± 13%)
900	2.78E-4	3.08E-4 (111 ± 14%)	7.27E-4 (262 ± 33%)	3.03E-4 (109 ± 14%)
1000	2.54E-4	2.84E-4 (112 ± 13%)	7.36E-4 (289 ± 35%)	2.77E-4 (109 ± 13%)

n	$\hat{\theta}_n^{gen}, K=1$	$\hat{\theta}_n^{gen}, K=2$	$\hat{\theta}_n^{gen}, K=3$	$\hat{\theta}_n^{gen}, K=4$
20	3.06E-3 (76 ± 10%)	7.83E-4 (19 ± 3%)	3.76E-4 (9 ± 1%)	2.71E-4 (7 ± 1%)
50	1.51E-3 (78 ± 10%)	4.03E-4 (21 ± 3%)	1.86E-4 (10 ± 1%)	1.46E-4 (8 ± 1%)
100	8.92E-4 (72 ± 9%)	2.22E-4 (18 ± 2%)	1.12E-4 (9 ± 1%)	8.92E-4 (72 ± 9%)
120	8.29E-4 (78 ± 9%)	2.09E-4 (20 ± 2%)	9.90E-5 (9 ± 1%)	6.87E-4 (65 ± 8%)
140	8.01E-4 (84 ± 10%)	1.93E-4 (20 ± 3%)	9.58E-5 (10 ± 1%)	5.44E-4 (58 ± 7%)
160	6.72E-4 (80 ± 10%)	1.63E-4 (19 ± 2%)	8.86E-4 (105 ± 13%)	5.17E-4 (62 ± 8%)
180	5.88E-4 (78 ± 10%)	1.55E-4 (20 ± 3%)	7.34E-4 (97 ± 13%)	4.18E-4 (55 ± 7%)
200	5.64E-4 (76 ± 9%)	1.42E-4 (19 ± 2%)	6.64E-4 (89 ± 11%)	3.41E-4 (46 ± 6%)
220	5.05E-4 (71 ± 9%)	1.37E-4 (19 ± 2%)	5.78E-4 (81 ± 10%)	3.31E-4 (46 ± 6%)
240	4.85E-4 (70 ± 8%)	1.31E-4 (19 ± 2%)	5.04E-4 (72 ± 9%)	3.16E-4 (45 ± 6%)
260	4.55E-4 (70 ± 8%)	1.17E-4 (18 ± 2%)	5.09E-4 (78 ± 10%)	2.60E-4 (40 ± 5%)
280	4.59E-4 (76 ± 10%)	1.14E-4 (19 ± 2%)	4.70E-4 (78 ± 10%)	2.39E-4 (40 ± 5%)
300	4.24E-4 (73 ± 9%)	1.05E-4 (18 ± 2%)	4.51E-4 (78 ± 10%)	2.27E-4 (39 ± 5%)
400	3.55E-4 (75 ± 9%)	9.06E-5 (19 ± 2%)	2.90E-4 (61 ± 10%)	1.63E-4 (34 ± 4%)
500	3.20E-4 (73 ± 9%)	7.58E-5 (17 ± 2%)	2.27E-4 (52 ± 8%)	1.22E-4 (28 ± 4%)
600	2.77E-4 (77 ± 10%)	6.54E-5 (18 ± 2%)	1.81E-4 (50 ± 6%)	1.09E-4 (30 ± 4%)
700	2.39E-4 (75 ± 9%)	5.87E-5 (18 ± 2%)	1.68E-4 (52 ± 6%)	8.84E-5 (28 ± 3%)
800	2.39E-4 (84 ± 10%)	5.21E-5 (18 ± 2%)	1.44E-4 (50 ± 7%)	7.75E-5 (27 ± 3%)
900	1.88E-4 (68 ± 9%)	5.37E-5 (19 ± 2%)	1.30E-4 (47 ± 6%)	6.66E-5 (24 ± 3%)
1000	1.82E-4 (71 ± 9%)	5.21E-5 (20 ± 3%)	1.19E-4 (47 ± 6%)	6.45E-5 (25 ± 3%)

Table 7 Empirical MSEs among estimators for the derivative with respect to the arrival rate for Case 2. Bracketed numbers represent the 95% CIs for the risk ratios between the considered estimators and the baseline $\bar{\theta}_n$.

References

Agrawal S, Ding Y, Saberi A, Ye Y (2012) Price of correlations in stochastic optimization. *Operations Research* 60(1):150–162.

Asmussen S, Glynn PW (2007) *Stochastic simulation: algorithms and analysis*, volume 57 (Springer Science & Business Media).

Barton RR (2012) Input uncertainty in outout analysis. *Proceedings of the Winter Simulation Conference*, 6 (IEEE).

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust optimization*, volume 28 (Princeton University Press).

n	$\bar{\theta}_n$	$\hat{\theta}_n^{rec}$	$\hat{\theta}_n^{gen}, K=0.5$	$\hat{\theta}_n^{gen}, K=0.82$
20	9.28E-3	7.62E-3 (82 ± 8%)	2.60E-2 (280 ± 28%)	1.11E-2 (119 ± 12%)
50	4.81E-3	4.62E-3 (96 ± 9%)	1.31E-2 (273 ± 25%)	5.37E-3 (112 ± 10%)
100	2.88E-3	2.88E-3 (100 ± 9%)	7.68E-3 (266 ± 24%)	3.13E-3 (109 ± 10%)
120	2.49E-3	2.48E-3 (100 ± 9%)	6.35E-3 (255 ± 23%)	2.87E-3 (115 ± 11%)
140	2.09E-3	2.21E-3 (106 ± 9%)	5.66E-3 (271 ± 25%)	2.72E-3 (130 ± 12%)
160	2.00E-3	2.25E-3 (113 ± 10%)	5.45E-3 (273 ± 24%)	2.21E-3 (111 ± 10%)
180	1.73E-3	1.83E-3 (106 ± 9%)	5.03E-3 (292 ± 25%)	1.92E-3 (105 ± 10%)
200	1.74E-3	1.78E-3 (102 ± 10%)	4.60E-3 (265 ± 24%)	1.82E-3 (105 ± 10%)
220	1.59E-3	1.76E-3 (111 ± 10%)	4.26E-3 (268 ± 24%)	1.79E-3 (107 ± 9%)
240	1.53E-3	1.50E-3 (98 ± 9%)	4.16E-3 (272 ± 24%)	1.69E-3 (110 ± 10%)
260	1.38E-3	1.47E-3 (106 ± 9%)	3.79E-3 (274 ± 25%)	1.53E-3 (111 ± 10%)
280	1.34E-3	1.37E-3 (102 ± 9%)	3.31E-3 (246 ± 21%)	1.36E-3 (101 ± 9%)
300	1.25E-3	1.28E-3 (102 ± 9%)	3.29E-3 (262 ± 23%)	1.43E-3 (114 ± 10%)
400	1.03E-3	1.08E-3 (106 ± 9%)	2.75E-3 (268 ± 22%)	1.13E-3 (111 ± 9%)
500	8.83E-4	9.89E-4 (112 ± 10%)	2.47E-3 (280 ± 25%)	9.73E-4 (110 ± 10%)
600	7.61E-4	9.00E-4 (118 ± 10%)	2.13E-3 (280 ± 24%)	8.31E-4 (109 ± 9%)
700	7.21E-4	7.46E-4 (105 ± 9%)	1.93E-3 (267 ± 24%)	7.33E-4 (102 ± 9%)
800	6.51E-4	6.89E-4 (106 ± 9%)	1.75E-3 (269 ± 23%)	7.26E-4 (112 ± 10%)
900	6.11E-4	6.27E-4 (102 ± 9%)	1.70E-3 (277 ± 25%)	6.28E-4 (103 ± 9%)
1000	5.62E-4	5.77E-4 (103 ± 9%)	1.48E-3 (263 ± 23%)	5.93E-4 (106 ± 9%)

n	$\hat{\theta}_n^{gen}, K=1$	$\hat{\theta}_n^{gen}, K=2$	$\hat{\theta}_n^{gen}, K=3$	$\hat{\theta}_n^{gen}, K=4$
20	7.63E-3 (82 ± 8%)	3.57E-3 (38 ± 4%)	4.32E-3 (47 ± 5%)	8.18E-3 (88 ± 9%)
50	3.59E-3 (75 ± 7%)	1.69E-3 (35 ± 3%)	2.01E-3 (42 ± 4%)	4.16E-3 (86 ± 7%)
100	2.13E-3 (74 ± 7%)	9.09E-4 (32 ± 3%)	1.17E-3 (40 ± 4%)	1.22E-2 (424 ± 44%)
120	1.97E-3 (79 ± 7%)	8.12E-4 (33 ± 3%)	1.04E-3 (42 ± 4%)	9.08E-3 (365 ± 40%)
140	1.76E-3 (84 ± 7%)	7.25E-4 (35 ± 3%)	8.60E-4 (41 ± 4%)	7.31E-3 (350 ± 35%)
160	1.44E-3 (72 ± 7%)	6.86E-4 (34 ± 3%)	5.64E-3 (282 ± 27%)	6.16E-3 (309 ± 33%)
180	1.31E-3 (76 ± 7%)	5.42E-4 (31 ± 3%)	5.04E-3 (292 ± 29%)	5.52E-3 (320 ± 36%)
200	1.32E-3 (76 ± 7%)	5.48E-4 (32 ± 3%)	4.20E-3 (242 ± 25%)	4.22E-3 (243 ± 26%)
220	1.22E-3 (77 ± 7%)	4.80E-4 (30 ± 3%)	3.85E-3 (242 ± 24%)	4.29E-3 (270 ± 28%)
240	1.10E-3 (72 ± 6%)	4.92E-4 (32 ± 3%)	3.20E-3 (209 ± 21%)	3.52E-3 (230 ± 23%)
260	1.09E-3 (79 ± 7%)	4.21E-4 (30 ± 3%)	3.11E-3 (224 ± 22%)	3.26E-3 (236 ± 24%)
280	9.89E-4 (74 ± 7%)	4.21E-4 (32 ± 3%)	2.90E-3 (215 ± 21%)	3.11E-3 (232 ± 25%)
300	1.01E-3 (80 ± 7%)	3.82E-4 (31 ± 3%)	2.79E-3 (223 ± 22%)	2.61E-3 (209 ± 22%)
400	7.94E-4 (77 ± 7%)	2.96E-4 (29 ± 2%)	1.87E-3 (182 ± 17%)	1.95E-3 (190 ± 21%)
500	6.85E-4 (78 ± 7%)	2.61E-4 (30 ± 3%)	1.46E-3 (165 ± 17%)	1.37E-3 (156 ± 16%)
600	6.00E-4 (79 ± 7%)	2.19E-4 (29 ± 3%)	1.06E-3 (139 ± 14%)	1.05E-3 (138 ± 15%)
700	5.42E-4 (75 ± 7%)	1.99E-4 (28 ± 3%)	9.34E-5 (130 ± 13%)	9.70E-3 (135 ± 14%)
800	4.82E-4 (74 ± 6%)	1.78E-4 (27 ± 2%)	7.96E-4 (122 ± 11%)	7.96E-3 (122 ± 12%)
900	4.66E-4 (76 ± 7%)	1.63E-4 (27 ± 2%)	7.23E-4 (118 ± 11%)	7.37E-3 (121 ± 12%)
1000	4.22E-4 (75 ± 7%)	1.45E-4 (26 ± 2%)	5.93E-4 (106 ± 10%)	6.13E-3 (109 ± 11%)

Table 8 Empirical MSEs among estimators for the gradient with respect to the arrival and service rates for Case 2. Bracketed numbers represent the 95% CIs for the risk ratios between the considered estimators and the baseline $\bar{\theta}_n$.

Ben-Tal A, Nemirovski A (2002) Robust optimization–methodology and applications. *Mathematical Programming* 92(3):453–480.

Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. *SIAM review* 53(3):464–501.

Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research* 57(6):1407–1420.

n	$\bar{\theta}_n$	$\hat{\theta}_n^{rec}$	$\hat{\theta}_n^{gen}, K=0.5$	$\hat{\theta}_n^{gen}, K=0.82$	$\hat{\theta}_n^{gen}, K=1$
20	7.27E-4	4.63E-3 (636 ± 178%)	1.85E-3 (254 ± 87%)	9.77E-4 (134 ± 62%)	5.11E-4 (70 ± 27%)
50	4.68E-4	1.40E-3 (299 ± 95%)	9.85E-4 (210 ± 76%)	4.86E-4 (104 ± 37%)	2.77E-4 (59 ± 24%)
100	3.63E-4	9.56E-4 (264 ± 75%)	7.78E-4 (214 ± 64%)	3.62E-4 (100 ± 33%)	2.35E-4 (65 ± 20%)
120	4.38E-4	8.05E-4 (184 ± 46%)	6.58E-4 (150 ± 45%)	4.41E-4 (101 ± 32%)	2.45E-4 (56 ± 17%)
140	3.93E-4	8.22E-4 (209 ± 51%)	7.45E-4 (189 ± 55%)	4.04E-4 (103 ± 37%)	2.27E-4 (58 ± 17%)
160	3.85E-4	7.80E-4 (202 ± 51%)	7.22E-4 (187 ± 47%)	3.50E-4 (91 ± 25%)	2.32E-4 (60 ± 18%)
180	4.09E-4	6.84E-4 (167 ± 38%)	8.02E-4 (196 ± 52%)	4.08E-4 (100 ± 25%)	2.07E-4 (51 ± 14%)
200	4.33E-4	6.93E-4 (160 ± 33%)	7.32E-4 (169 ± 40%)	3.06E-4 (71 ± 16%)	2.47E-4 (57 ± 15%)
220	4.02E-4	5.79E-4 (144 ± 34%)	7.72E-4 (192 ± 45%)	3.51E-4 (87 ± 22%)	2.43E-4 (61 ± 17%)
240	3.83E-4	6,66E-4 (174 ± 35%)	6.95E-4 (181 ± 45%)	3.42E-4 (89 ± 21%)	2.16E-4 (56 ± 13%)
260	4.51E-4	5.68E-4 (126 ± 27%)	7.71E-4 (171 ± 36%)	3.84E-4 (85 ± 17%)	2.16E-4 (48 ± 12%)
280	5.27E-4	5.68E-4 (108 ± 21%)	6.94E-4 (132 ± 28%)	3.66E-4 (69 ± 16%)	2.48E-4 (47 ± 11%)
300	3.83E-4	6.62E-4 (173 ± 35%)	9.11E-4 (237 ± 52%)	4.49E-4 (117 ± 26%)	2.56E-4 (67 ± 16%)
400	4.71E-4	5.49E-4 (117 ± 19%)	7.40E-4 (157 ± 32%)	4.90E-4 (103 ± 19%)	2.90E-4 (61 ± 11%)
500	4.47E-4	5.73E-4 (128 ± 21%)	7.26E-4 (163 ± 29%)	4.59E-4 (103 ± 17%)	3.28E-4 (73 ± 12%)
600	4.67E-4	5.59E-4 (120 ± 19%)	7.13E-4 (153 ± 26%)	4.50E-4 (96 ± 15%)	3.64E-4 (78 ± 13%)
700	4.60E-4	5.66E-4 (123 ± 17%)	7.67E-4 (167 ± 27%)	4.44E-4 (96 ± 14%)	3.42E-4 (74 ± 11%)
800	4.96E-4	5.82E-4 (117 ± 16%)	6.81E-4 (137 ± 23%)	4.96E-4 (100 ± 15%)	3.85E-4 (78 ± 11%)
900	5.16E-4	5.47E-4 (106 ± 13%)	6.99E-4 (135 ± 20%)	4.82E-4 (93 ± 13%)	3.91E-4 (76 ± 10%)
1000	5.34E-4	6.10E-4 (114 ± 15%)	7.09E-4 (133 ± 20%)	4.92E-4 (92 ± 12%)	4.34E-4 (81 ± 11%)

Table 9 Empirical MSEs among estimators for the derivative with respect to the arrival rate for Case 3. Bracketed numbers represent the 95% CIs for the risk ratios between the considered estimators and the baseline $\bar{\theta}_n$.

n	$\bar{\theta}_n$	$\hat{\theta}_n^{rec}$	$\hat{\theta}_n^{gen}, K=0.5$	$\hat{\theta}_n^{gen}, K=0.82$	$\hat{\theta}_n^{gen}, K=1$
20	4.18E-2	1.14E-2 (27 ± 10%)	7.54E-2 (180 ± 51%)	3.46E-2 (83 ± 25%)	1.09E-2 (26 ± 8%)
50	1.53E-2	7.59E-3 (50 ± 16%)	3.66E-2 (239 ± 69%)	1.48E-2 (97 ± 31%)	4.71E-3 (31 ± 10%)
100	9.71E-3	5.20E-3 (54 ± 17%)	2.12E-2 (218 ± 55%)	1.33E-2 (137 ± 39%)	3.48E-3 (36 ± 11%)
120	6.53E-3	4.88E-3 (74 ± 22%)	1.77E-2 (270 ± 69%)	1.02E-2 (157 ± 48%)	2.22E-3 (34 ± 10%)
140	6.66E-3	3.54E-3 (53 ± 15%)	1.77E-2 (266 ± 65%)	8.24E-3 (124 ± 34%)	2.23E-3 (32 ± 9%)
160	5.01E-3	3.53E-3 (71 ± 18%)	1.25E-2 (249 ± 55%)	5.92E-3 (118 ± 30%)	1.97E-3 (39 ± 11%)
180	4.85E-3	3.07E-3 (63 ± 18%)	1.25E-2 (258 ± 58%)	5.75E-3 (119 ± 28%)	1.93E-3 (40 ± 10%)
200	4.41E-3	3.42E-3 (78 ± 22%)	9.84E-3 (223 ± 50%)	5.93E-3 (135 ± 36%)	1.93E-3 (44 ± 11%)
220	3.62E-3	3.00E-3 (83 ± 21%)	9.93E-3 (274 ± 60%)	5.09E-3 (140 ± 39%)	1.86E-3 (51 ± 13%)
240	3.31E-3	2.42E-3 (73 ± 17%)	8.78E-3 (265 ± 54%)	4.91E-3 (149 ± 34%)	1.62E-3 (49 ± 11%)
260	3.47E-3	2.53E-3 (73 ± 17%)	9.04E-3 (261 ± 54%)	3.94E-3 (114 ± 27%)	1.62E-3 (47 ± 11%)
280	3.18E-3	2.29E-3 (72 ± 17%)	7.72E-3 (243 ± 47%)	3.31E-3 (104 ± 23%)	1.71E-3 (54 ± 13%)
300	2.99E-3	2.17E-3 (72 ± 17%)	7.54E-3 (252 ± 52%)	3.15E-3 (105 ± 24%)	1.68E-3 (56 ± 13%)
400	2.25E-3	2.04E-3 (91 ± 18%)	5.48E-3 (244 ± 46%)	2.44E-3 (108 ± 24%)	1.12E-3 (51 ± 11%)
500	2.14E-3	1.80E-3 (84 ± 17%)	4.91E-3 (230 ± 43%)	2.03E-3 (95 ± 19%)	9.77E-3 (46 ± 10%)
600	1.62E-3	1.31E-3 (81 ± 15%)	3.83E-3 (236 ± 43%)	1.78E-3 (110 ± 22%)	9.28E-3 (57 ± 11%)
700	1.50E-3	1.29E-3 (86 ± 15%)	3.48E-3 (232 ± 36%)	1.43E-3 (95 ± 17%)	1.04E-3 (69 ± 13%)
800	1.28E-3	1.29E-3 (100 ± 18%)	2.91E-3 (227 ± 36%)	1.59E-3 (124 ± 22%)	8.32E-3 (65 ± 11%)
900	1.20E-3	1.07E-3 (89 ± 14%)	2.27E-3 (189 ± 29%)	1.37E-3 (114 ± 18%)	8.32E-3 (70 ± 12%)
1000	1.08E-3	1.07E-3 (99 ± 16%)	2.13E-3 (197 ± 30%)	1.26E-3 (116 ± 20%)	7.74E-3 (72 ± 12%)

Table 10 Empirical MSEs among estimators for the gradient with respect to the arrival and service rates for Case 3. Bracketed numbers represent the 95% CIs for the risk ratios between the considered estimators and the baseline $\bar{\theta}_n$.

Besbes O, Zeevi A (2011) On the minimax complexity of pricing in a changing environment. *Operations Research* 59(1):66–79.

Blanchet JH, Glynn PW (2015) Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. *Proceedings of the Winter Simulation Conference*, 3656–3667 (IEEE).

Borkar VS (2009) *Stochastic approximation: a dynamical systems viewpoint*, volume 48 (Springer).

Cesa-Bianchi N, Lugosi G (2006) *Prediction, learning, and games* (Cambridge University Press).

Chick SE (2006) Bayesian ideas and discrete event simulation: why, what and how. *Proceedings of the Winter Simulation Conference*, 96–105 (IEEE).

Chung KL (1954) On a stochastic approximation method. *The Annals of Mathematical Statistics* 463–483.

Duplay D, Lam H, Zhang X (2018) Achieving optimal bias-variance tradeoff in online derivative estimation. *Proceedings of the Winter Simulation Conference* (IEEE).

Fabian V (1968) On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics* 39(4):1327–1332.

Flaxman AD, Kalai AT, McMahan HB (2005) Online convex optimization in the bandit setting: gradient descent without a gradient. *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, 385–394 (Society for Industrial and Applied Mathematics).

Fox BL, Glynn PW (1989) Replication schemes for limiting expectations. *Probability in the Engineering and Informational Sciences* 3(3):299–318.

Fu MC (2006) Gradient estimation. *Handbooks in operations research and management science* 13:575–616.

Fu MC, Hong LJ, Hu JQ (2009) Conditional Monte Carlo estimation of quantile sensitivities. *Management Science* 55(12):2019–2027.

Fu MC, Hu JQ (1992) Extensions and generalizations of smoothed perturbation analysis in a generalized semi-Markov process framework. *IEEE Transactions on Automatic Control* 37(10):1483–1500.

Giles MB (2008) Multilevel Monte Carlo path simulation. *Operations Research* 56(3):607–617.

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

Glasserman P, Gong WB (1990) Smoothed perturbation analysis for a class of discrete-event systems. *IEEE Transactions on Automatic Control* 35(11):1218–1230.

Glynn PW (1989) Optimization of stochastic systems via simulation. *Proceedings of the Winter Simulation Conference*, 90105 (IEEE).

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

Glynn PW, Whitt W (1992) The asymptotic efficiency of simulation estimators. *Operations Research* 40(3):505–520.

Gong WB, Ho YC (1987) Smoothed (conditional) perturbation analysis of discrete event dynamical systems. *IEEE Transactions on Automatic Control* 32(10):858–866.

Hazan E, et al. (2016) Introduction to online convex optimization. *Foundations and Trends® in Optimization* 2(3-4):157–325.

Heidelberger P, Cao XR, Zazanis MA, Suri R (1988) Convergence properties of infinitesimal perturbation analysis estimates. *Management Science* 34(11):1281–1302.

Heidergott B, Pflug G, Farenhorst-Yuan T, et al. (2010) Gradient estimation for discrete-event systems by measure-valued differentiation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 20(1):5.

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

Henderson SG (2003) Input model uncertainty: Why do we care and what should we do about it? *Proceedings of the Winter Simulation Conference*, 90–100 (IEEE).

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

Hong LJ (2009) Estimating quantile sensitivities. *Operations Research* 57(1):118–130.

Kushner H, Yin GG (2003) *Stochastic approximation and recursive algorithms and applications*, volume 35 (Springer Science & Business Media).

Lam H (2016) Advanced tutorial: Input uncertainty and robust analysis in stochastic simulation. *Proceedings of the Winter Simulation Conference*, 178–192 (IEEE).

L'Ecuyer P (1990) A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science* 36(11):1364–1383.

L'Ecuyer P (1991) An overview of derivative estimation. *Proceedings of the Winter Simulation Conference*, 207–217 (IEEE).

McLeish D (2011) A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods and Applications* 17(4):301–315.

Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19(4):1574–1609.

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

Pasupathy R (2010) On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. *Operations Research* 58(4-part-1):889–901.

Pasupathy R, Kim S (2011) The stochastic root-finding problem: Overview, solutions, and open questions. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 21(3):19.

Peng Y, Fu MC, Hu JQ, Heidergott B (2018) A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters. *Operations Research* 66(2):487–499.

Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization* 30(4):838–855.

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

Rhee Ch, Glynn PW (2015) Unbiased estimation with square root convergence for SDE models. *Operations Research* 63(5):1026–1043.

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

Rubinstein RY (1992) Sensitivity analysis of discrete event systems by the push out method. *Annals of Operations Research* 39(1):229–250.

Ruppert D (1988) Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Rychlik T (1990) Unbiased nonparametric estimation of the derivative of the mean. *Statistics & probability letters* 10(4):329–333.

Shalev-Shwartz S (2012) Online learning and online convex optimization. *Foundations and Trends® in Machine Learning* 4(2):107–194.

Song E, Nelson BL, Pegden CD (2014) Advanced tutorial: Input uncertainty quantification. *Proceedings of the Winter Simulation Conference*, 162–176 (IEEE).

Spall JC (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control* 37(3):332–341.

Spall JC (1997) A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica* 33(1):109–112.

Vihola M (2018) Unbiased estimators and multilevel Monte Carlo. *Operations Research* 66(2):448–462.

Zazanis MA, Suri R (1993) Convergence rates of finite-difference sensitivity estimates for stochastic systems. *Operations Research* 41(4):694–703.

Zhou K, Doyle JC (1998) *Essentials of robust control*, volume 104 (Prentice hall Upper Saddle River, NJ).

Author Biographies

Henry Lam is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. His research interests include Monte Carlo methods, uncertainty quantification, data-driven optimization and rare-event analysis. His works have been recognized by several venues such as the NSF CAREER Award, JP Morgan Chase Faculty Research Award and Adobe Faculty Research Award. Henry serves on the editorial boards of Operations Research, INFORMS Journal on Computing, Applied Probability Journals, Stochastic Models, Manufacturing and Service Operations Management, and Queueing Systems, and as the Area Editor in Stochastic Models and Data Science in Operations Research Letters. His current email address is `henry.lam@columbia.edu`.

Xinyu Zhang obtained his Ph.D. in the Department of Industrial Engineering and Operations Research at Columbia University in December 2021 under the supervision of Henry Lam. His current position is Software Engineer at Amazon. His current email address is `xz2691@columbia.edu`.

Xuhui Zhang is currently a PhD student in the Department of Management Science and Engineering at Stanford University. He graduated from the University of Science and Technology of China and was a summer intern under the supervision of Prof. Henry Lam in 2018. His current email address is `xzhang98@stanford.edu`.

Appendix A: Proofs for Section 3

We will prove a multivariate version of Theorem 1.

THEOREM 9. *Under Assumption 2, suppose that* $\lim_{n\to\infty}\delta_n n^\alpha = d$, *where* $0<d<\infty$ *the sample-average-based estimator* $\bar{\boldsymbol{\theta}}_n$ *exhibits the asymptotic MSE*

$$E\|\bar{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2 = d^{2q_1}\|\mathbf{B}\|^2 n^{-2\alpha q_1}+\frac{tr\left(\Sigma\right)}{d^{2q_2}}n^{2\alpha q_2-1}+o\left(n^{-2\alpha q_1}+n^{2\alpha q_2-1}\right) \text{ as } n\to\infty$$

Choosing $\alpha=\frac{1}{2(q_1+q_2)}$ *achieves the optimal MSE order, and the asymptotic MSE is*

$$E\|\bar{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2 = \left(d^{2q_1}\|\mathbf{B}\|^2+\frac{tr\left(\Sigma\right)}{d^{2q_2}}\right)n^{-\frac{q_1}{q_1+q_2}}+o\left(n^{-\frac{q_1}{q_1+q_2}}\right) \text{ as } n\to\infty$$

Proof of theorem 9. By the bias-variance decomposition, we have

$$\begin{aligned}E\|\bar{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2 &= \|E\bar{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2+tr\left(Cov\left(\bar{\boldsymbol{\theta}}_n\right)\right)\\ &=\|\mathbf{b}\left(\delta_n\right)\|^2+\frac{1}{n}tr\left(Cov\left(\mathbf{v}\left(\delta_n\right)\right)\right)\\ &=\|\mathbf{B}\|^2\delta_n^{2q_1}+o\left(\delta_n^{2q_1}\right)+\frac{1}{n}\frac{tr\left(\Sigma\right)+o\left(1\right)}{\delta_n^{2q_2}}\end{aligned}$$

Setting $\delta_n=\frac{d+o(1)}{n^\alpha}$, we obtain

$$\begin{aligned}E\|\bar{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\|^2 &= \|\mathbf{B}\|^2\frac{\left(d+o\left(1\right)\right)^{2q_1}}{n^{2\alpha q_1}}+o\left(n^{-2\alpha q_1}\right)+\frac{tr\left(\Sigma\right)+o\left(1\right)}{\left(d+o\left(1\right)\right)^{2q_2}}n^{2\alpha q_2-1}\\ &=\left(\|\mathbf{B}\|^2 d^{2q_1}+o\left(1\right)\right)n^{-2\alpha q_1}+\left(\frac{tr\left(\Sigma\right)}{d^{2q_2}}+o\left(1\right)\right)n^{2\alpha q_2-1}\end{aligned}$$

To achieve the optimal MSE order, we solve $-2\alpha q_1=2\alpha q_2-1$. Thus $\alpha=\frac{1}{2(q_1+q_2)}$ and the optimal order is $n^{-\frac{q_1}{q_1+q_2}}$.

Proof of Theorem 1. The proof follows immediately by considering dimension 1 in Theorem 9.
□

Appendix B: Proofs for Section 4.2

We provide and prove multivariate versions of the results, from which the ones in Section 4.2 follow immediately.

Frequently used in the subsequent proofs is the following result adapted from Lemma 4.2, a version of Chung's Lemma, in Fabian (1967):

LEMMA 1 **(Chung's Lemma)**. *For* v_n, c_n, b_n *real numbers, and* $0<\alpha\leq 1$, *suppose* $\lim_{n\to\infty}c_n = c>0$, *and consider the iteration*

$$v_{n+1}=\left(1-\frac{c_n}{n^\alpha}\right)v_n+\frac{b_n}{n^\alpha} \tag{50}$$

If $b_n\to 0$, *then* $v_n\to 0$; *if* $b_n\to b>0$, *then* $v_n\to\frac{b}{c}$; *and if* $b_n\to\infty$, *then* $v_n\to\infty$.

Proof of Lemma 1. Our version of Chung's lemma is different in appearance from Lemma 4.2 in Fabian (1968), and thus we repeat the proof here. First, if $b_n \to b$ where $b \geq 0$ is a real number, then for given $0 < \epsilon < c$, we can choose n_1 sufficiently large such that, for all $n \geq n_1$, we have $\frac{c_n}{n^\alpha} < 1$, $b_n < b + \epsilon$ and $c - \epsilon < c_n < c + \epsilon$. Now let $n \geq n_1$. If $v_n \geq \frac{b+2\epsilon}{c-\epsilon}$, then from the iteration (50)

$$v_{n+1} \leq v_n - \frac{b+2\epsilon}{c-\epsilon}(c-\epsilon)\frac{1}{n^\alpha} + (b+\epsilon)\frac{1}{n^\alpha} \leq v_n - \frac{\epsilon}{n^\alpha}$$

On the other hand, if $v_n \leq \frac{b+2\epsilon}{c-\epsilon}$, then since the right hand side of the iteration (50) is an increasing function of v_n, we have

$$v_{n+1} \leq \frac{b+2\epsilon}{c-\epsilon} - \frac{b+2\epsilon}{c-\epsilon}(c-\epsilon)\frac{1}{n^\alpha} + (b+\epsilon)\frac{1}{n^\alpha} \leq \frac{b+2\epsilon}{c-\epsilon}$$

Combined with the fact that $\sum\limits_{n=1}^{\infty} \frac{1}{n^\alpha}$ diverges, we have $\limsup_{n\to\infty} v_n \leq \frac{b+2\epsilon}{c-\epsilon}$. Since ϵ is arbitrary, we get

$$\limsup_{n\to\infty} v_n \leq \frac{b}{c} \tag{51}$$

If $b = 0$, $v_{n+1} \geq v_n + \frac{\epsilon}{n^\alpha}$ for $v_n \leq -\frac{2\epsilon}{c-\epsilon}$ and $v_{n+1} \geq -\frac{2\epsilon}{c-\epsilon}$ for $v_n \geq -\frac{2\epsilon}{c-\epsilon}$. Therefore we have $\liminf\limits_{n\to\infty} v_n \geq 0$ and $\limsup\limits_{n\to\infty} v_n \leq 0$. We conclude that $\lim_{n\to\infty} v_n = 0$. By the same analysis, if $b_n \to b > 0$, where b possibly take the value of ∞, we would have

$$\liminf_{n\to\infty} v_n \geq \frac{b}{c} \tag{52}$$

Thus if $b = \infty$, we conclude that $\lim_{n\to\infty} v_n \to \infty$, and if $0 < b < \infty$, combining (51) and (52), we get $\lim_{n\to\infty} v_n = \frac{b}{c}$. □

We now consider multivariate versions of our results and their proofs:

Proposition 2. *Under Assumption 2, we have:*

1. *If $\beta \leq 1$ and $\alpha < \frac{\beta}{2q_2}$, the estimator $\hat{\boldsymbol{\theta}}_n^{rec}$ is L_2-consistent for $\boldsymbol{\theta}_0$, i.e.,*

$$\lim_{n\to\infty} E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 = 0$$

2. *If $\beta \leq 1$ and $\alpha \geq \frac{\beta}{2q_2}$, or if $\beta > 1$, the error of $\hat{\boldsymbol{\theta}}_n^{rec}$ in estimating $\boldsymbol{\theta}_0$ is bounded away from zero in L_2 norm as $n \to \infty$, i.e.,*

$$\liminf_{n\to\infty} E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 > 0$$

Proof of Proposition 2. We first prove the proposition for $\beta \leq 1$. From the recursion

$$\hat{\boldsymbol{\theta}}_n^{rec} = (1-\gamma_n)\hat{\boldsymbol{\theta}}_{n-1}^{rec} + \gamma_n \boldsymbol{\theta}(\delta_n) \tag{53}$$

we have

$$E\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0 = (1-\gamma_n)\left(E\hat{\boldsymbol{\theta}}_{n-1}^{rec} - \boldsymbol{\theta}_0\right) + \gamma_n\left(E\boldsymbol{\theta}(\delta_n) - \boldsymbol{\theta}_0\right)$$

Since $E\boldsymbol{\theta}\left(\delta_n\right)-\boldsymbol{\theta}_0=\mathbf{b}\left(\delta_n\right)\to 0$ as $n\to\infty$, we have $E\hat{\boldsymbol{\theta}}_n^{rec}-\boldsymbol{\theta}_0\to 0$ by Chung's lemma. Note that $E\|\hat{\boldsymbol{\theta}}_n^{rec}-\boldsymbol{\theta}_0\|^2=\|E\hat{\boldsymbol{\theta}}_n^{rec}-\boldsymbol{\theta}_0\|^2+tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right)$. Thus the convergence will depend on the variance term. Taking covariance of (53), by independence we have

$$Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)=\left(1-\gamma_n\right)^2 Cov\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right)+\gamma_n^2 Cov\left(\boldsymbol{\theta}\left(\delta_n\right)\right) \tag{54}$$

Since

$$Cov\left(\boldsymbol{\theta}\left(\delta_n\right)\right)=\frac{1}{\delta_n^{2q_2}}Cov\left(\boldsymbol{\varepsilon}\left(\delta_n\right)\right)$$

we have

$$\lim_{n\to\infty}\frac{Cov\left(\boldsymbol{\theta}\left(\delta_n\right)\right)}{n^{2q_2\alpha}}=\frac{\Sigma}{d^{2q_2}}$$

We now rewrite the iteration (54) as

$$tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right)=\left(1-\left(2+o\left(1\right)\right)\gamma_n\right)tr\left(Cov\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right)\right)+\gamma_n s_n$$

where $s_n=c\frac{tr(\Sigma)}{d^{2q_2}}n^{2q_2\alpha-\beta}+o\left(n^{2q_2\alpha-\beta}\right)$. We note that $\lim_{n\to\infty}s_n=\infty$ if $\alpha>\frac{\beta}{2q_2}$, $\lim_{n\to\infty}s_n=c\frac{tr(\Sigma)}{d^{2q_2}}>0$ if $\alpha=\frac{\beta}{2q_2}$, and $\lim_{n\to\infty}s_n=0$ if $\alpha<\frac{\beta}{2q_2}$. Thus by Chung's lemma

$$\lim_{n\to\infty}tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right)\to\infty \text{ if } \alpha>\frac{\beta}{2q_2}$$

$$\lim_{n\to\infty}tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right)=c\frac{tr\left(\Sigma\right)}{2d^{2q_2}} \text{ if } \alpha=\frac{\beta}{2q_2}$$

and

$$\lim_{n\to\infty}tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right)=0 \text{ if } \alpha<\frac{\beta}{2q_2}$$

This completes the proof for $\beta\le 1$.

Next consider $\beta>1$, we now argue that choosing $\gamma_n=cn^{-\beta}$ does not lead to convergence. We note that $\hat{\boldsymbol{\theta}}_n^{rec}$ is a linear combination of $\hat{\boldsymbol{\theta}}_0^{rec},\boldsymbol{\theta}_i\left(\delta_i\right),i=1,\cdots,n$, i.e.

$$\hat{\boldsymbol{\theta}}_n^{rec}=a_0\hat{\boldsymbol{\theta}}_0^{rec}+\sum_{i=1}^n a_i\boldsymbol{\theta}_i\left(\delta_i\right)$$

where $a_0=\prod_{j=1}^n\left(1-\gamma_j\right)$ and $a_i=\gamma_i\prod_{j=i+1}^n\left(1-\gamma_j\right)$. Since $\sum_{n=1}^\infty\gamma_n=\sum_{n=1}^\infty\frac{c}{n^\beta}<\infty$, by the relation between infinite product and infinite sum, we get

$$\lim_{n\to\infty}a_i \text{ exists and is positive for any } i$$

Since by independence

$$tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right)=a_0^2 tr\left(Cov\left(\hat{\boldsymbol{\theta}}_0^{rec}\right)\right)+\sum_{i=1}^n a_i^2 tr\left(Cov\left(\boldsymbol{\theta}\left(\delta_i\right)\right)\right)$$

we have that

$$\liminf_{n\to\infty}tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right)>0$$

□

Theorem 10. *Under Assumption 2, the MSE of $\hat{\boldsymbol{\theta}}_n^{rec}$ in estimating $\boldsymbol{\theta}_0$ behaves as follows:*

1. *For $\beta < 1$ and $\alpha < \frac{\beta}{2q_2}$,*

$$E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 = d^{2q_1}\|\mathbf{B}\|^2 n^{-2q_1\alpha} + \frac{c}{2d^{2q_2}} tr\left(\Sigma\right) n^{2q_2\alpha - \beta} + o\left(n^{-2q_1\alpha} + n^{2q_2\alpha-\beta}\right) \;\; as\ n \to \infty$$

2. *For $\beta = 1$, $\alpha = \frac{1}{2(q_1+q_2)}$ and $c > \frac{q_1}{2(q_1+q_2)}$,*

$$E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 = \left(\left(\frac{cd^{q_1}}{c - \frac{q_1}{2(q_1+q_2)}}\right)^2 \|\mathbf{B}\|^2 + \frac{c^2}{\left(2c - \frac{q_1}{q_1+q_2}\right)d^{2q_2}} tr\left(\Sigma\right)\right) n^{-\frac{q_1}{q_1+q_2}} + o\left(n^{-\frac{q_1}{q_1+q_2}}\right) \;\; as\ n \to \infty$$

3. *For $\beta = 1$, $\alpha = \frac{1}{2(q_1+q_2)}$ and $c \le \frac{q_1}{2(q_1+q_2)}$, or for $\beta = 1$ and $\alpha \neq \frac{1}{2(q_1+q_2)}$,*

$$\limsup_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 = \infty$$

Proof of Theorem 10. Taking expectation of (53) and rearranging terms, we have

$$E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) = \left(1-\gamma_n\right)E\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec} - \boldsymbol{\theta}_0\right) + \gamma_n\left(E\boldsymbol{\theta}\left(\delta_n\right) - \boldsymbol{\theta}_0\right) = \left(1-\gamma_n\right)E\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec} - \boldsymbol{\theta}_0\right) + \gamma_n\left(\mathbf{B}\delta_n^{q_1} + o\left(\delta_n^{q_1}\right)\right) \tag{55}$$

If $\gamma_n = \frac{c}{n}$ and $\alpha \le \frac{1}{2(q_1+q_2)}$, we multiply (55) by $n^{q_1\alpha}$ to get

$$\begin{aligned} n^{q_1\alpha}E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) &= \left(\frac{n}{n-1}\right)^{q_1\alpha}\left(1-\frac{c}{n}\right)\left(n-1\right)^{q_1\alpha}E\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec} - \boldsymbol{\theta}_0\right) + \frac{c}{n}\left(\mathbf{B}d^{q_1} + o\left(1\right)\right) \\ &= \left(1 - \frac{c - q_1\alpha + o\left(1\right)}{n}\right)\left(n-1\right)^{q_1\alpha}E\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec} - \boldsymbol{\theta}_0\right) + \frac{c}{n}\left(\mathbf{B}d^{q_1} + o\left(1\right)\right) \end{aligned}$$

For $c > q_1\alpha$, by Chung's lemma, $\lim_{n\to\infty} n^{q_1\alpha}E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) = \frac{cd^{q_1}}{c-q_1\alpha}\mathbf{B}$. Thus

$$E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) = \frac{cd^{q_1}}{c-q_1\alpha}\mathbf{B}n^{-q_1\alpha} + o\left(n^{-q_1\alpha}\right)$$

If $\gamma_n = c/n$ and $\alpha > \frac{1}{2(q_1+q_2)}$, we multiply (55) by $n^{1/2-q_2\alpha}$ to get

$$n^{1/2-q_2\alpha}E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) = \left(1 - \frac{c - 1/2 + q_2\alpha + o\left(1\right)}{n}\right)\left(n-1\right)^{1/2-q_2\alpha}E\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec} - \boldsymbol{\theta}_0\right) + o\left(\frac{1}{n}\right)$$

For $c > 1/2 - q_2\alpha$, by Chung's lemma, $\lim_{n\to\infty} n^{1/2-q_2\alpha}E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) = 0$. Thus

$$E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) = o\left(n^{q_2\alpha - 1/2}\right)$$

Similarly, if $\gamma_n = \frac{c}{n^\beta}, \beta < 1$, we multiply (55) by $n^{q_1\alpha}$ to get

$$n^{q_1\alpha}E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) = \left(1 - \frac{c + o\left(1\right)}{n^\beta}\right)\left(n-1\right)^{q_1\alpha}E\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec} - \boldsymbol{\theta}_0\right) + \frac{c}{n^\beta}\left(\mathbf{B}d^{q_1} + o\left(1\right)\right)$$

For $c > 0$, by Chung's lemma, $\lim_{n\to\infty} n^{q_1\alpha}E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) = \mathbf{B}d^{q_1}$. Thus

$$E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) = \mathbf{B}d^{q_1}n^{-q_1\alpha} + o\left(n^{-q_1\alpha}\right) \tag{56}$$

Next we take covariance of (53) and by independence,

$$
\begin{aligned}
Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right) &= (1-\gamma_n)^2 Cov\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right) + \gamma_n^2 Cov\left(\boldsymbol{\theta}\left(\delta_n\right)\right) \\
&= (1-\gamma_n)^2 Cov\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right) + \gamma_n^2 \frac{Cov\left(\boldsymbol{\varepsilon}\left(\delta_n\right)\right)}{\delta_n^{2q_2}} \\
&= (1-\gamma_n)^2 Cov\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right) + \gamma_n^2 n^{2q_2\alpha} \frac{\Sigma + o(1)}{d^{2q_2}} \qquad (57)
\end{aligned}
$$

If $\gamma_n = \frac{c}{n}$ and $\alpha \geq \frac{1}{2(q_1+q_2)}$, we multiply (57) by $n^{1-2q_2\alpha}$ and take trace to get

$$
\begin{aligned}
n^{1-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right) &= \left(\frac{n}{n-1}\right)^{1-2q_2\alpha} \left(1-\frac{c}{n}\right)^2 (n-1)^{1-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right)\right) + \frac{c^2}{n} \frac{tr(\Sigma)+o(1)}{d^{2q_2}} \\
&= \left(1 - \frac{2c+2q_2\alpha-1+o(1)}{n}\right) (n-1)^{1-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right)\right) + \frac{c^2}{n} \frac{tr(\Sigma)+o(1)}{d^{2q_2}} \qquad (58)
\end{aligned}
$$

For $c > 1/2 - q_2\alpha$, by Chung's lemma, $\lim_{n\to\infty} n^{1-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right) = \frac{c^2 tr(\Sigma)}{(2c+2q_2\alpha-1)d^{2q_2}}$. Thus

$$
tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right) = \frac{c^2 tr(\Sigma)}{(2c+2q_2\alpha-1)\, d^{2q_2}} n^{2q_2\alpha-1} + o\left(n^{2q_2\alpha-1}\right)
$$

Similarly, if $\gamma_n = \frac{c}{n^\beta}, \beta < 1$, we multiply (57) by $n^{\beta-2q_2\alpha}$ and take trace to get

$$
\begin{aligned}
n^{\beta-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right) &= \left(\frac{n}{n-1}\right)^{\beta-2q_2\alpha} \left(1-\frac{c}{n^\beta}\right)^2 (n-1)^{\beta-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right)\right) + \frac{c^2}{n^\beta} \frac{tr(\Sigma)+o(1)}{d^{2q_2}} \\
&= \left(1 - \frac{2c+o(1)}{n^\beta}\right) (n-1)^{\beta-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right)\right) + \frac{c^2}{n^\beta} \frac{tr(\Sigma)+o(1)}{d^{2q_2}}
\end{aligned}
$$

For $c > 0$, by Chung's lemma, $\lim_{n\to\infty} n^{\beta-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right) = \frac{ctr(\Sigma)}{2d^{2q_2}}$. Thus

$$
tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right) = \frac{ctr(\Sigma)}{2d^{2q_2}} n^{2q_2\alpha-\beta} + o\left(n^{2q_2\alpha-\beta}\right)
$$

In conclusion, if $\gamma_n = \frac{c}{n}$, $\alpha = \frac{1}{2(q_1+q_2)}$ and $c > \frac{q_1}{2(q_1+q_2)}$, then

$$
\begin{aligned}
E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 &= \|E\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 + tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right) \\
&= \left(\frac{cd^{q_1}}{c-q_1\alpha}\right)^2 \|\mathbf{B}\|^2 n^{-2q_1\alpha} + o\left(n^{-2q_1\alpha}\right) + \frac{c^2 tr(\Sigma)}{(2c+2q_2\alpha-1)\, d^{2q_2}} n^{2q_2\alpha-1} + o\left(n^{2q_2\alpha-1}\right) \\
&= \left(\left(\frac{cd^{q_1}}{c-q_1/\left(2\left(q_1+q_2\right)\right)}\right)^2 \|\mathbf{B}\|^2 + \frac{c^2}{\left(2c-q_1/\left(q_1+q_2\right)\right) d^{2q_2}} tr(\Sigma)\right) n^{-q_1/(q_1+q_2)} + o\left(n^{-q_1/(q_1+q_2)}\right)
\end{aligned}
$$

If $\gamma_n = \frac{c}{n}$, $\alpha > \frac{1}{2(q_1+q_2)}$ and $c > 1/2 - q_2\alpha$, then

$$
E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 = \frac{c^2 tr(\Sigma)}{(2c+2q_2\alpha-1)\, d^{2q_2}} n^{2q_2\alpha-1} + o\left(n^{2q_2\alpha-1}\right) \qquad (59)
$$

Similarly, if $\gamma_n = \frac{c}{n^\beta}, \beta < 1$ and $c > 0$, then

$$
E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 = d^{2q_1}\|\mathbf{B}\|^2 n^{-2q_1\alpha} + o\left(n^{-2q_1\alpha}\right) + \frac{c}{2d^{2q_2}} tr(\Sigma)\, n^{2q_2\alpha-\beta} + o\left(n^{2q_2\alpha-\beta}\right)
$$

This completes the proof for part 1 and part 2 of the theorem.

Next we prove part 3 of the theorem. If $\alpha > \frac{1}{2(q_1+q_2)}$ and $c > 1/2 - q_2\alpha$, we note from (59) that

$$\lim_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 = \infty$$

If $\alpha \geq \frac{1}{2(q_1+q_2)}$ and $c \leq 1/2 - q_2\alpha$, and supposing that the sequence $n^{1-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right)$ is bounded, then from (58) we have that

$$n^{1-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right) \geq (n-1)^{1-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right)\right) + \frac{C_1 + o(1)}{n}$$

for some $C_1 > 0$, for all large enough n. Since $\sum_{n=1}^{\infty} 1/n = \infty$, we get

$$n^{1-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right) \to \infty \text{ as } n \to \infty$$

which is a contradiction. Thus

$$\limsup_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 \geq \limsup_{n\to\infty} n^{1-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{rec}\right)\right) = \infty$$

If $\alpha < \frac{1}{2(q_1+q_2)}$, and supposing that the sequence $n^{\frac{q_1}{2(q_1+q_2)}} E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right)$ is bounded, we multiply (55) by $n^{\frac{q_1}{2(q_1+q_2)}}$ to get

$$\begin{aligned}
& n^{\frac{q_1}{2(q_1+q_2)}} E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) \\
&= \left(1 - \frac{c - q_1/\left(2\left(q_1+q_2\right)\right) + o(1)}{n}\right)(n-1)^{\frac{q_1}{2(q_1+q_2)}} E\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec} - \boldsymbol{\theta}_0\right) + \frac{c}{n^{1-q_1\left(\frac{1}{2(q_1+q_2)}-\alpha\right)}}\left(\mathbf{B}d^{q_1} + o(1)\right) \\
&= (n-1)^{\frac{q_1}{2(q_1+q_2)}} E\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec} - \boldsymbol{\theta}_0\right) + \frac{c\mathbf{B}d^{q_1} + o(1)}{n^{1-q_1\left(\frac{1}{2(q_1+q_2)}-\alpha\right)}}
\end{aligned}$$

Since $\sum_{n=1}^{\infty} 1/n^{1-q_1\left(\frac{1}{2(q_1+q_2)}-\alpha\right)} = \infty$, we get

$$n^{\frac{q_1}{2(q_1+q_2)}} E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) \to \infty \text{ as } n \to \infty$$

which is a contradiction. Thus

$$\limsup_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} E\|\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 \geq \limsup_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} \|E\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\|^2 = \infty$$

This completes the proof for part 3 of the theorem. □

Proof of Theorem 2. This follows immediately from Theorem 10 by setting the dimension to 1. □

Proof of Theorem 3. We have

$$R^{rec}\left(\theta\left(\cdot\right), d, c\right) = \frac{\left(\frac{cd^{q_1}}{c - \frac{q_1}{2(q_1+q_2)}}\right)^2 B^2 + \frac{c^2}{2d^{2q_2}\left(c - \frac{q_1}{2(q_1+q_2)}\right)}\sigma^2}{d^{2q_1}B^2 + \frac{1}{d^{2q_2}}\sigma^2}$$

For any d, B and σ^2, we have

$$R^{rec}\left(\theta\left(\cdot\right), d, c\right) \le \max\left\{\frac{\left(\frac{cd^{q_1}}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2}{d^{2q_1}}, \frac{\frac{c^2}{2d^{2q_2}\left(c-\frac{q_1}{2(q_1+q_2)}\right)}}{\frac{1}{d^{2q_2}}}\right\} = \max\left\{\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2, \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\right\}$$

Note that the right hand side above is approachable by choosing B or σ^2 to be arbitrarily big. Therefore

$$\max_{\theta(\cdot)\in H, d>0} R^{rec}\left(\theta\left(\cdot\right), d, c\right) = \max\left\{\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2, \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\right\} \tag{60}$$

Now suppose that

$$\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2 > \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}$$

which is equivalent to $c < \frac{5q_1+4q_2}{2(q_1+q_2)}$. Since the function $\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2$ is monotonically decreasing in the region $\frac{q_1}{2(q_1+q_2)} < c < \frac{5q_1+4q_2}{2(q_1+q_2)}$, we have

$$\max\left\{\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2, \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\right\} = \left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2 \ge \left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2\Bigg|_{c=\frac{5q_1+4q_2}{2(q_1+q_2)}}$$

Similarly, suppose that

$$\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2 < \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}$$

which is equivalent to $c > \frac{5q_1+4q_2}{2(q_1+q_2)}$. Since the function $\frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}$ is monotonically increasing in the region $c > \frac{5q_1+4q_2}{2(q_1+q_2)}$, we have

$$\max\left\{\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2, \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\right\} = \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)} \ge \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\Bigg|_{c=\frac{5q_1+4q_2}{2(q_1+q_2)}}$$

Thus the minimization of (60) gives us $c = \frac{5q_1+4q_2}{2(q_1+q_2)}$, which solves

$$\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2 = \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}$$

and we note that both sides of this equation is $\frac{q_1^2}{16(q_1+q_2)^2} + \frac{q_1}{2(q_1+q_2)} + 1$. □

Proof of Theorem 4. We have

$$R^{rec}\left(\theta\left(\cdot\right), d, \tilde{d}, c\right) = \frac{\left(\frac{c\tilde{d}^{q_1}}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2 B^2 + \frac{c^2}{2\tilde{d}^{2q_2}\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\sigma^2}{d^{2q_1}B^2 + \frac{1}{d^{2q_2}}\sigma^2}$$

For any $d, \tilde{d}, B$ and σ^2, we have

$$R^{rec}\left(\theta\left(\cdot\right), d, \tilde{d}, c\right) \leq \max\left\{\frac{\left(\frac{c\tilde{d}^{q_1}}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2}{d^{2q_1}}, \frac{\frac{c^2}{2\tilde{d}^{2q_2}\left(c-\frac{q_1}{2(q_1+q_2)}\right)}}{\frac{1}{d^{2q_2}}}\right\}$$

$$= \max\left\{\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2\left(\frac{\tilde{d}}{d}\right)^{2q_1}, \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\frac{1}{\left(\frac{\tilde{d}}{d}\right)^{2q_2}}\right\}$$

Note that the right hand side above is approachable by choosing B or σ^2 to be arbitrarily big. Therefore

$$\max_{\theta(\cdot)\in H, d>0} R^{rec}\left(\theta\left(\cdot\right), d, \tilde{d}, c\right) = \max_{d>0}\max\left\{\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2\eta\left(d\right)^{2q_1}, \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\frac{1}{\eta\left(d\right)^{2q_2}}\right\} \quad (61)$$

where we let $\eta\left(d\right) = \frac{\tilde{d}}{d}$, and note that $\tilde{d} = g\left(d\right)$ is also a function of d. We minimize the right hand side of (61) via minimizing

$$\max\left\{\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2\eta\left(d\right)^{2q_1}, \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\frac{1}{\eta\left(d\right)^{2q_2}}\right\} \quad (62)$$

for each d. With d fixed arbitrarily, first, for any c, since both of the expressions in (62) are monotonic in $\eta\left(d\right)$, we need

$$\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2\eta\left(d\right)^{2q_1} = \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\frac{1}{\eta\left(d\right)^{2q_2}}$$

which upon solving leads to

$$\eta\left(d\right) = \left(\frac{c-\frac{q_1}{2(q_1+q_2)}}{2}\right)^{1/(2(q_1+q_2))}$$

Thus (62) becomes

$$\max\left\{\left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2\eta\left(d\right)^{2q_1}, \frac{c^2}{2\left(c-\frac{q_1}{2(q_1+q_2)}\right)}\frac{1}{\eta\left(d\right)^{2q_2}}\right\} = \left(\frac{c}{c-\frac{q_1}{2(q_1+q_2)}}\right)^2\eta\left(d\right)^{2q_1} = \frac{1}{2^{q_1/(q_1+q_2)}}\frac{c^2}{\left(c-\frac{q_1}{2(q_1+q_2)}\right)^{\frac{q_1+2q_2}{q_1+q_2}}}$$

We then optimize c over the region $c > \frac{q_1}{2(q_1+q_2)}$, i.e,

$$c = \arg\min_{c>\frac{q_1}{2(q_1+q_2)}}\frac{1}{2^{q_1/(q_1+q_2)}}\frac{c^2}{\left(c-\frac{q_1}{2(q_1+q_2)}\right)^{\frac{q_1+2q_2}{q_1+q_2}}} = 1$$

This gives $\eta\left(d\right) = \left(\frac{q_1+2q_2}{4(q_1+q_2)}\right)^{\frac{1}{2(q_1+q2)}}$ and (62) is $2^{\frac{2q_2}{q_1+q_2}}\left(\frac{q_1+2q_2}{q_1+q_2}\right)^{-\frac{q_1+2q_2}{q_1+q_2}}$. We note that the optimal $c, \eta\left(d\right)$ are independent of d, and therefore the value of (61) is also $2^{\frac{2q_2}{q_1+q_2}}\left(\frac{q_1+2q_2}{q_1+q_2}\right)^{-\frac{q_1+2q_2}{q_1+q_2}}$. □

Next, we consider the uniform-averaging scheme:

THEOREM 11. *Under Assumption 2, the MSE of $\hat{\boldsymbol{\theta}}_n^{avg}$ in estimating $\boldsymbol{\theta}_0$ behaves as follows:*

1. *For $\beta < 1$ and $\alpha \leq \frac{1}{2(q_1+q_2)}$,*

$$E\|\hat{\boldsymbol{\theta}}_n^{avg} - \boldsymbol{\theta}_0\|^2 = \left(\frac{d^{q_1}}{1-q_1\alpha}\right)^2 \|\mathbf{B}\|^2 n^{-2q_1\alpha} + \frac{1}{(1+2q_2\alpha)\, d^{2q_2}} tr\left(\Sigma\right) n^{2q_2\alpha-1} + o\left(n^{-2q_1\alpha} + n^{2q_2\alpha-1}\right) \ \ as\ n \to \infty$$

2. *For $\beta < 1$ and $\alpha > \frac{1}{2(q_1+q_2)}$,*

$$E\|\hat{\boldsymbol{\theta}}_n^{avg} - \boldsymbol{\theta}_0\|^2 = \frac{1}{(1+2q_2\alpha)\, d^{2q_2}} tr\left(\Sigma\right) n^{2q_2\alpha-1} + o\left(n^{2q_2\alpha-1}\right) \ \ as\ n \to \infty$$

Proof of Theorem 11. We first analyze $E\hat{\boldsymbol{\theta}}_n^{avg} - \boldsymbol{\theta}_0$, For $0 < \alpha \leq \frac{1}{2(q_1+q_2)}$, since $-1 < -q_1\alpha < 0$, we have that

$$\int_1^{n+1} s^{-q_1\alpha} ds \leq \sum_{i=1}^n i^{-q_1\alpha} \leq \int_0^n s^{-q_1\alpha} ds$$

Thus

$$\sum_{i=1}^n i^{-q_1\alpha} = \int_0^n s^{-q_1\alpha} ds + o\left(\int_0^n s^{-q_1\alpha} ds\right) = \frac{n^{1-q_1\alpha}}{1-q_1\alpha} + o\left(n^{1-q_1\alpha}\right)$$

and

$$\frac{1}{n}\sum_{i=1}^n i^{-q_1\alpha} = \frac{1}{1-q_1\alpha} n^{-q_1\alpha} + o\left(n^{-q_1\alpha}\right)$$

From (56) we have $E\left(\hat{\boldsymbol{\theta}}_n^{rec} - \boldsymbol{\theta}_0\right) = \mathbf{B} d^{q_1} n^{-q_1\alpha} + o\left(n^{-q_1\alpha}\right)$. Thus

$$E\hat{\boldsymbol{\theta}}_n^{avg} - \boldsymbol{\theta}_0 = \frac{1}{n}\sum_{i=1}^n E\left(\hat{\boldsymbol{\theta}}_i^{rec} - \boldsymbol{\theta}_0\right) = \frac{1}{n}\sum_{i=1}^n \left(\mathbf{B} d^{q_1} i^{-q_1\alpha} + o\left(i^{-q_1\alpha}\right)\right) = \frac{d^{q_1}}{1-q_1\alpha} \mathbf{B} n^{-q_1\alpha} + o\left(n^{-q_1\alpha}\right)$$

For $\alpha > \frac{1}{2(q_1+q_2)}$, by a similar analysis we get

$$E\hat{\boldsymbol{\theta}}_n^{avg} - \boldsymbol{\theta}_0 = \frac{1}{n}\sum_{i=1}^n E\left(\hat{\boldsymbol{\theta}}_i^{rec} - \boldsymbol{\theta}_0\right) = \begin{cases} O\left(\frac{1}{n^{q_1\alpha}}\right) & \text{if } -q_1\alpha > -1 \\ O\left(\frac{log(n)}{n}\right) & \text{if } -q_1\alpha = -1 \\ O\left(\frac{1}{n}\right) & \text{if } -q_1\alpha < -1 \end{cases}$$

Since $1/2 - q_2\alpha < 1$ and $1/2 - q_2\alpha < q_1\alpha$, we have

$$n^{1/2-q_2\alpha} E\left(\hat{\boldsymbol{\theta}}_n^{avg} - \boldsymbol{\theta}_0\right) = o\left(1\right)$$

We then analyze $tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{avg}\right)\right)$. Rewrite the iteration (53) as

$$\hat{\boldsymbol{\theta}}_n^{rec} - E\hat{\boldsymbol{\theta}}_n^{rec} = \left(1-\gamma_n\right)\left(\hat{\boldsymbol{\theta}}_{n-1}^{rec} - E\hat{\boldsymbol{\theta}}_{n-1}^{rec}\right) + \gamma_n\left(\boldsymbol{\theta}\left(\delta_n\right) - E\boldsymbol{\theta}\left(\delta_n\right)\right)$$

Let $\boldsymbol{U}_n = \hat{\boldsymbol{\theta}}_n^{rec} - E\hat{\boldsymbol{\theta}}_n^{rec}$. Thus

$$\boldsymbol{U}_n = \left(1-\gamma_n\right)\boldsymbol{U}_{n-1} + \gamma_n \mathbf{v}\left(\delta_n\right)$$

Following Polyak and Juditsky (1992), we can write

$$\boldsymbol{U}_n = \prod_{i=1}^{n} (1-\gamma_i)\boldsymbol{U}_0 + \sum_{i=1}^{n} \left(\prod_{j=i+1}^{n} (1-\gamma_j) \right) \gamma_i \mathbf{v}(\delta_i)$$

Thus $\hat{\boldsymbol{\theta}}_n^{avg} - E\hat{\boldsymbol{\theta}}_n^{avg}$ can be written as

$$\begin{aligned}\hat{\boldsymbol{\theta}}_n^{avg} - E\hat{\boldsymbol{\theta}}_n^{avg} &= \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{U}_k \\ &= \frac{1}{n}\sum_{k=1}^{n}\prod_{i=1}^{k}(1-\gamma_i)\boldsymbol{U}_0 + \frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{k}\left(\prod_{j=i+1}^{k}(1-\gamma_j)\right)\gamma_i\mathbf{v}(\delta_i) \\ &= \frac{1}{n}\sum_{k=1}^{n}\prod_{i=1}^{k}(1-\gamma_i)\boldsymbol{U}_0 + \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=i}^{n}\prod_{j=i+1}^{k}(1-\gamma_j)\right)\gamma_i\mathbf{v}(\delta_i)\end{aligned}$$

Let

$$p_n = \sum_{k=1}^{n}\prod_{i=1}^{k}(1-\gamma_i)$$

$$q_n^i = \gamma_i \sum_{k=i}^{n}\prod_{j=i+1}^{k}(1-\gamma_j)$$

and $w_n^i = q_n^i - 1$. Then

$$\hat{\boldsymbol{\theta}}_n^{avg} - E\hat{\boldsymbol{\theta}}_n^{avg} = \frac{p_n}{n}\boldsymbol{U}_0 + \frac{1}{n}\sum_{i=1}^{n}\mathbf{v}(\delta_i) + \frac{1}{n}\sum_{i=1}^{n} w_n^i \mathbf{v}(\delta_i) \tag{63}$$

From Lemma 1 and Lemma 2 in Polyak and Juditsky (1992), we have that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}|w_n^i| = 0, \text{ and } |w_n^i| \le C_1, \; |p_n| \le C_1, \text{ for some } C_1 > 0$$

Multiplying (63) by $n^{1/2-q_2\alpha}$, we have

$$n^{1/2-q_2\alpha}\left(\hat{\boldsymbol{\theta}}_n^{avg} - E\hat{\boldsymbol{\theta}}_n^{avg}\right) = \frac{p_n}{n^{1/2+q_2\alpha}}\boldsymbol{U}_0 + \frac{1}{n^{1/2+q_2\alpha}}\sum_{i=1}^{n}\mathbf{v}(\delta_i) + \frac{1}{n^{1/2+q_2\alpha}}\sum_{i=1}^{n} w_n^i\mathbf{v}(\delta_i)$$

Since p_n is bounded, $E\|\frac{p_n}{n^{1/2+q_2\alpha}}\boldsymbol{U}_0\|^2 = o(1)$. Besides, by independence,

$$E\|\frac{1}{n^{1/2+q_2\alpha}}\sum_{i=1}^{n} w_n^i\mathbf{v}(\delta_i)\|^2 = \frac{1}{n^{1+2q_2\alpha}}\sum_{i=1}^{n}\left(w_n^i\right)^2 E\|\mathbf{v}(\delta_i)\|^2 \le \frac{C_2}{n^{1+2q_2\alpha}}\sum_{i=1}^{n}|w_n^i| i^{2q_2\alpha} \le \frac{C_2}{n}\sum_{i=1}^{n}|w_n^i|$$

for some $C_2 > 0$. Therefore, $E\|\frac{1}{n^{1/2+q_2\alpha}}\sum_{i=1}^{n} w_n^i\mathbf{v}(\delta_i)\|^2 = o(1)$. Thus

$$\begin{aligned}n^{1-2q_2\alpha} tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{avg}\right)\right) &= \frac{1}{n^{1+2q_2\alpha}}\sum_{i=1}^{n} tr\left(Cov\left(\mathbf{v}(\delta_i)\right)\right) + o(1) \\ &= \frac{1}{n^{1+2q_2\alpha}}\sum_{i=1}^{n} i^{2q_2\alpha}\frac{tr(\Sigma) + o(1)}{d^{2q_2}} + o(1) \\ &= \frac{tr(\Sigma)}{(1+2q_2\alpha)d^{2q_2}} + o(1)\end{aligned}$$

In conclusion, for $\alpha \leq \frac{1}{2(q_1+q_2)}$, we have

$$E\|\hat{\boldsymbol{\theta}}_n^{avg} - \boldsymbol{\theta}_0\|^2 = \|E\hat{\boldsymbol{\theta}}_n^{avg} - \boldsymbol{\theta}_0\|^2 + tr\left(Cov\left(\hat{\boldsymbol{\theta}}_n^{avg}\right)\right)$$

$$= \left(\frac{d^{q_1}}{1-q_1\alpha}\right)^2 \|\mathbf{B}\|^2 n^{-2q_1\alpha} + \frac{tr\left(\Sigma\right)}{\left(1+2q_2\alpha\right)d^{2q_2}} n^{2q_2\alpha-1} + o\left(n^{-2q_1\alpha} + n^{2q_2\alpha-1}\right)$$

and for $\alpha > \frac{1}{2(q_1+q_2)}$, we have

$$E\|\hat{\boldsymbol{\theta}}_n^{avg} - \boldsymbol{\theta}_0\|^2 = \frac{tr\left(\Sigma\right)}{\left(1+2q_2\alpha\right)d^{2q_2}} n^{2q_2\alpha-1} + o\left(n^{2q_2\alpha-1}\right)$$

□

Proof of Theorem 5. This follows immediately from Theorem 11 by setting the dimension to 1. □

Proof of Theorem 6. The proof follows exactly that of Theorem 4 and setting the dimension to 1, by noting the equivalence between the MSE expressions in Theorem 11 and Theorem 10 with $c=1$, $\beta=1$ and $\alpha = \frac{1}{2(q_1+q_2)}$. □

Proof of Theorem 7. This follows immediately by noting that the proofs for Theorems 3 and 4 apply exactly the same when d is fixed. □

Appendix C: Proofs in Section 5.2

We prove Theorem 8. Note that part of the proof has been sketched in Section 5.2, and for clarity we will have slight amount of repetition to make this proof self-contained.

Proof of Theorem 8. Let $\alpha = \frac{1}{2(q_1+q_2)}$. For convenience, we skip the second subscript of $w_{j,n}$ and write w_j, and denote $w = \left(w_j\right)_{j=1,\ldots,n}$, when no confusion arises. We also assume $n_0 = 0$ without loss of generality.

First, we argue that $\sum_{j=1}^n w_j \to 1$. Suppose not, then there exists a subsequence n_k such that $\left|\sum_{j=1}^{n_k} w_j - 1\right| > \epsilon_0$ for some $\epsilon_0 > 0$. Assume without loss of generality that $\sum_{j=1}^{n_k} w_j - 1 > \epsilon_0$. Moreover, suppose the sequence

$$\sum_{j=1}^{n_k} w_j \left(B\frac{g\left(d\right)^{q_1}}{j^{\alpha q_1}} + o\left(\frac{1}{j^{\alpha q_1}}\right)\right) \tag{64}$$

is bounded. We can choose a sufficiently large θ_0 such that

$$\liminf_{k\to\infty} \left(\left(\sum_{j=1}^{n_k} w_j - 1\right)\theta_0 + \sum_{j=1}^{n_k} w_j \left(B\frac{g\left(d\right)^{q_1}}{j^{\alpha q_1}} + o\left(\frac{1}{j^{\alpha q_1}}\right)\right)\right)^2 > 0$$

On the other hand, suppose (64) is unbounded. Then we can choose $\theta_0 = 0$ so that

$$\limsup_{k\to\infty} \left(\left(\sum_{j=1}^{n_k} w_j - 1\right)\theta_0 + \sum_{j=1}^{n_k} w_j \left(B\frac{g\left(d\right)^{q_1}}{j^{\alpha q_1}} + o\left(\frac{1}{j^{\alpha q_1}}\right)\right)\right)^2 = \infty$$

Therefore, either way we would have $R_{n_k} \to \infty$.

Now, we consider a particular scheme $w, g(\cdot)$ such that $\sum_j w_j = 1$ and $g(d) = \eta d$ for some $\eta > 0$. Then

$$\begin{aligned}\mathrm{MSE}_1 &= \left(Bd^{q_1}\eta^{q_1}\sum_{j=1}^{n}\frac{w_j(1+o(1))}{j^{\alpha q_1}}\right)^2 + \frac{\sigma^2}{d^{2q_2}\eta^{2q_2}}\sum_{j=1}^{n} j^{2\alpha q_2} w_j^2 (1+o(1)) \\ &= \left(Bd^{q_1}\eta^{q_1}\sum_{j=1}^{n}\frac{w_j}{j^{\alpha q_1}}\right)^2 + \frac{\sigma^2}{d^{2q_2}\eta^{2q_2}}\sum_{j=1}^{n} j^{2\alpha q_2} w_j^2 + \varepsilon_n \qquad (65)\end{aligned}$$

where ε_n is an error term.

As described in Section 5.2, we consider optimization problem (36) to obtain w, η that minimizes (65) (or (34)) asymptotically. We call S_n^* the optimal value of (36). We will show that

$$\max_{\theta(\cdot)\in H, d>0} R^{gen}(\theta(\cdot), d, g(d), W) = \lim_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} S_n^*$$

is the asymptotic minimax risk ratio we seek for, and consequently the solution w, η to (36) is the optimal configuration. In the following, we first obtain a characterization of the solution to (36), and then verify that the solution also ensures the error term ε_n is negligible. Then we argue that no other configurations, namely $w, g(\cdot)$ such that $\sum_j w_j \to 1$ and $g(\cdot) \in \mathcal{F}_K$ that can give a better risk ratio. Although the solution η to (36) may depend on n, we will demonstrate that η converges to a positive number as $n \to \infty$, and it will be clear that substituting η with its limit will not affect the asymptotic risk ratio.

To solve (36), we follow the derivation in Section 5.2 starting from (37) to obtain the optimal weights in (47), with λ_1, λ_2 as the Lagrange multipliers of the two constraints in (43) evaluated at the solution a^* of (45), and the optimal η^* in (46).

Now, for convenience, we write

$$w = \begin{bmatrix}\Sigma^{-1}\mu & \Sigma^{-1}\mathbb{1}\end{bmatrix}\begin{bmatrix}\lambda_1 \\ \lambda_2\end{bmatrix} \qquad (66)$$

so that

$$\mu^\top w = \begin{bmatrix}\mu^\top\Sigma^{-1}\mu & \mu^\top\Sigma^{-1}\mathbb{1}\end{bmatrix}\begin{bmatrix}\lambda_1 \\ \lambda_2\end{bmatrix}$$

and

$$\mathbb{1}^\top w = \begin{bmatrix}\mathbb{1}^\top\Sigma^{-1}\mu & \mathbb{1}^\top\Sigma^{-1}\mathbb{1}\end{bmatrix}\begin{bmatrix}\lambda_1 \\ \lambda_2\end{bmatrix}$$

Setting $\mu^\top w = a$ and $\mathbb{1}^\top w = 1$, we get

$$\begin{bmatrix}\lambda_1 \\ \lambda_2\end{bmatrix} = \begin{bmatrix}\mu^\top\Sigma^{-1}\mu & \mu^\top\Sigma^{-1}\mathbb{1} \\ \mathbb{1}^\top\Sigma^{-1}\mu & \mathbb{1}^\top\Sigma^{-1}\mathbb{1}\end{bmatrix}^{-1}\begin{bmatrix}a \\ 1\end{bmatrix}$$

Let $\phi(\kappa) = \sum_{j=1}^{n} 1/j^{\kappa}$. We can write this as

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \phi(\alpha(2q_1 + 2q_2)) & \phi(\alpha(q_1 + 2q_2)) \\ \phi(\alpha(q_1 + 2q_2)) & \phi(2\alpha q_2) \end{bmatrix}^{-1} \begin{bmatrix} a \\ 1 \end{bmatrix}$$
$$= \begin{bmatrix} \phi(1) & \phi\left(\frac{q_1+2q_2}{2(q_1+q_2)}\right) \\ \phi\left(\frac{q_1+2q_2}{2(q_1+q_2)}\right) & \phi\left(\frac{q_2}{q_1+q_2}\right) \end{bmatrix}^{-1} \begin{bmatrix} a \\ 1 \end{bmatrix} \tag{67}$$

From (66), we can represent the optimal weights as

$$w^*(a) = \begin{bmatrix} \Sigma^{-1}\mu & \Sigma^{-1}\mathbb{1} \end{bmatrix} \begin{bmatrix} \mu^\top\Sigma^{-1}\mu & \mu^\top\Sigma^{-1}\mathbb{1} \\ \mathbb{1}^\top\Sigma^{-1}\mu & \mathbb{1}^\top\Sigma^{-1}\mathbb{1} \end{bmatrix}^{-1} \begin{bmatrix} a \\ 1 \end{bmatrix}$$

and write

$$\begin{aligned} \tilde{Z}_n^*(a)^2 &= \|\Sigma^{1/2} w^*(a)\|^2 \\ &= [a \;\; 1] \begin{bmatrix} \mu^\top\Sigma^{-1}\mu & \mu^\top\Sigma^{-1}\mathbb{1} \\ \mathbb{1}^\top\Sigma^{-1}\mu & \mathbb{1}^\top\Sigma^{-1}\mathbb{1} \end{bmatrix}^{-1} \begin{bmatrix} \mu'\Sigma^{-1} \\ \mathbb{1}^\top\Sigma^{-1} \end{bmatrix} \Sigma \begin{bmatrix} \Sigma^{-1}\mu & \Sigma^{-1}\mathbb{1} \end{bmatrix} \begin{bmatrix} \mu^\top\Sigma^{-1}\mu & \mu^\top\Sigma^{-1}\mathbb{1} \\ \mathbb{1}^\top\Sigma^{-1}\mu & \mathbb{1}^\top\Sigma^{-1}\mathbb{1} \end{bmatrix}^{-1} \begin{bmatrix} a \\ 1 \end{bmatrix} \\ &= [a \;\; 1] \begin{bmatrix} \mu^\top\Sigma^{-1}\mu & \mu^\top\Sigma^{-1}\mathbb{1} \\ \mathbb{1}^\top\Sigma^{-1}\mu & \mathbb{1}^\top\Sigma^{-1}\mathbb{1} \end{bmatrix}^{-1} \begin{bmatrix} \mu^\top\Sigma^{-1}\mu & \mu^\top\Sigma^{-1}\mathbb{1} \\ \mathbb{1}^\top\Sigma^{-1}\mu & \mathbb{1}^\top\Sigma^{-1}\mathbb{1} \end{bmatrix} \begin{bmatrix} \mu^\top\Sigma^{-1}\mu & \mu^\top\Sigma^{-1}\mathbb{1} \\ \mathbb{1}^\top\Sigma^{-1}\mu & \mathbb{1}^\top\Sigma^{-1}\mathbb{1} \end{bmatrix}^{-1} \begin{bmatrix} a \\ 1 \end{bmatrix} \\ &= [a \;\; 1] \, \Xi \begin{bmatrix} a \\ 1 \end{bmatrix} \end{aligned}$$

where

$$\Xi = \begin{bmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \end{bmatrix} = \begin{bmatrix} \mu^\top\Sigma^{-1}\mu & \mu^\top\Sigma^{-1}\mathbb{1} \\ \mathbb{1}^\top\Sigma^{-1}\mu & \mathbb{1}^\top\Sigma^{-1}\mathbb{1} \end{bmatrix}^{-1} = \begin{bmatrix} \phi(1) & \phi\left(\frac{q_1+2q_2}{2(q_1+q_2)}\right) \\ \phi\left(\frac{q_1+2q_2}{2(q_1+q_2)}\right) & \phi\left(\frac{q_2}{q_1+q_2}\right) \end{bmatrix}^{-1} \tag{68}$$

Thus, (45) can be written as

$$\min_{a:\left(K^{2(q_1+q_2)} - \xi_{11}\right)a^2 - 2\xi_{12}a - \xi_{22} \geq 0} |a|^{\frac{2q_2}{q_1+q_2}} \left(\xi_{11}a^2 + 2\xi_{12}a + \xi_{22}\right)^{\frac{q_1}{q_1+q_2}} \tag{69}$$

We now find the asymptotic limit of (36) scaled by $n^{\frac{q_1}{q_1+q_2}}$. First, we write a as $\tilde{a}n^{-\frac{q_1}{2(q_1+q_2)}}$. Then, reparametrizing by $\tilde{a}$ and denoting $\bar{Z}_n^*(\tilde{a}) = \tilde{Z}_n^*\left(\tilde{a}n^{-\frac{q_1}{2(q_1+q_2)}}\right)$, we have

$$\bar{Z}_n^*(\tilde{a})^2 = \begin{bmatrix} \frac{\tilde{a}}{n^{\frac{q_1}{2(q_1+q_2)}}} & 1 \end{bmatrix} \Xi \begin{bmatrix} \frac{\tilde{a}}{n^{\frac{q_1}{2(q_1+q_2)}}} \\ 1 \end{bmatrix}$$

Note that $\phi(1) \sim \log n$ and $\phi(\kappa) \sim \frac{1}{1-\kappa} n^{1-\kappa}$ for $\kappa < 1$ as $n \to \infty$. Thus,

$$\begin{aligned} & n^{\frac{q_1}{q_1+q_2}} \bar{Z}_n^*(\tilde{a})^2 \\ &= n^{\frac{q_1}{q_1+q_2}} \begin{bmatrix} \frac{\tilde{a}}{n^{\frac{q_1}{2(q_1+q_2)}}} & 1 \end{bmatrix} \begin{bmatrix} (1+o(1))\log n & \frac{2(q_1+q_2)(1+o(1))}{q_1} n^{\frac{q_1}{2(q_1+q_2)}} \\ \frac{2(q_1+q_2)(1+o(1))}{q_1} n^{\frac{q_1}{2(q_1+q_2)}} & \frac{(q_1+q_2)(1+o(1))}{q_1} n^{\frac{q_1}{q_1+q_2}} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\tilde{a}}{n^{\frac{q_1}{2(q_1+q_2)}}} \\ 1 \end{bmatrix} \\ &= n^{\frac{q_1}{q_1+q_2}} \begin{bmatrix} \frac{\tilde{a}}{n^{\frac{q_1}{2(q_1+q_2)}}} & 1 \end{bmatrix} \frac{\begin{bmatrix} \frac{(q_1+q_2)(1+o(1))}{q_1} n^{\frac{q_1}{q_1+q_2}} & -\frac{2(q_1+q_2)(1+o(1))}{q_1} n^{\frac{q_1}{2(q_1+q_2)}} \\ -\frac{2(q_1+q_2)(1+o(1))}{q_1} n^{\frac{q_1}{2(q_1+q_2)}} & (1+o(1))\log n \end{bmatrix}}{\frac{(q_1+q_2)}{q_1} n^{\frac{q_1}{q_1+q_2}} \log n (1+o(1)) - \frac{4(q_1+q_2)^2}{q_1^2} n^{\frac{q_1}{q_1+q_2}} (1+o(1))} \begin{bmatrix} \frac{\tilde{a}}{n^{\frac{q_1}{2(q_1+q_2)}}} \\ 1 \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
&= [\tilde{a} \quad 1] \frac{\begin{bmatrix} \frac{(q_1+q_2)(1+o(1))}{q_1} & -\frac{2(q_1+q_2)(1+o(1))}{q_1} \\ -\frac{2(q_1+q_2)(1+o(1))}{q_1} & (1+o(1))\log n \end{bmatrix}}{\frac{q_1+q_2}{q_1}\log n\,(1+o(1)) - \frac{4(q_1+q_2)^2}{q_1^2}(1+o(1))} \begin{bmatrix} \tilde{a} \\ 1 \end{bmatrix} \qquad (70) \\
&= [\tilde{a} \quad 1] \left(\tilde{\Xi} + o(1)\right) \begin{bmatrix} \tilde{a} \\ 1 \end{bmatrix}
\end{aligned}
$$

where

$$\tilde{\Xi} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{q_1}{q_1+q_2} \end{bmatrix}$$

Rewriting (45) in terms of $\tilde{a}$, we have that (45), when multiplying its objective value by $n^{\frac{q_1}{q_1+q_2}}$, becomes

$$\min_{\tilde{a}: n^{\frac{q_1}{q_1+q_2}} \bar{Z}_n^*(\tilde{a})^2 \le K^{2(q_1+q_2)}\tilde{a}^2} |\tilde{a}|^{\frac{2q_2}{q_1+q_2}} \left(n^{\frac{q_1}{q_1+q_2}} \bar{Z}_n^*(\tilde{a})^2\right)^{\frac{q_1}{q_1+q_2}} \qquad (71)$$

We consider an asymptotic version of (71) given by

$$\min_{\tilde{a}: \frac{q_1}{q_1+q_2} \le K^{2(q_1+q_2)}\tilde{a}^2} |\tilde{a}|^{\frac{2q_2}{q_1+q_2}} \left(\frac{q_1}{q_1+q_2}\right)^{\frac{q_1}{q_1+q_2}} = \frac{q_1}{q_1+q_2}\frac{1}{K^{2q_2}} \qquad (72)$$

We now argue that the absolute value of an optimal solution to (71), denoted $\tilde{a}_n^*$, converges to $\sqrt{\frac{q_1}{q_1+q_2}}\frac{1}{K^{q_1+q_2}}$, from which it follows immediately that the value of (71) converges to $\frac{q_1}{q_1+q_2}\frac{1}{K^{2q_2}}$, as $n \to \infty$. Suppose that $\left||\tilde{a}_{n_k}^*| - \sqrt{\frac{q_1}{q_1+q_2}}\frac{1}{K^{q_1+q_2}}\right| > \epsilon$, for some $\epsilon > 0$ and subsequence $n_k \to \infty$. If for infinitely many k it holds that $|\tilde{a}_{n_k}^*| < \sqrt{\frac{q_1}{q_1+q_2}}\frac{1}{K^{q_1+q_2}} - \epsilon$, then $\tilde{a}_{n_k}^*$ is excluded from the feasible region of (71), namely

$$\tilde{a}_{n_k}^* \notin \left\{\tilde{a} : n_k^{q_1/(q_1+q_2)} \bar{Z}_{n_k}^*(\tilde{a})^2 \le K^{2(q_1+q_2)}\tilde{a}^2\right\} \qquad (73)$$

infinitely often, which is a contradiction by the definition of $\tilde{a}_n^*$. Therefore we have $|\tilde{a}_{n_k}^*| > \sqrt{\frac{q_1}{q_1+q_2}}\frac{1}{K^{q_1+q_2}} + \epsilon$ for all k sufficiently large. Next, from (70), we have that $n^{\frac{q_1}{q_1+q_2}} \bar{Z}_n^*(\tilde{a})^2$ is bounded from below uniformly over $\tilde{a}$:

$$
\begin{aligned}
\min_{\tilde{a}} n^{\frac{q_1}{q_1+q_2}} \bar{Z}_n^*(\tilde{a})^2 &= \min_{\tilde{a}} [\tilde{a} \quad 1] \frac{\begin{bmatrix} \frac{(q_1+q_2)(1+o(1))}{q_1} & -\frac{2(q_1+q_2)(1+o(1))}{q_1} \\ -\frac{2(q_1+q_2)(1+o(1))}{q_1} & (1+o(1))\log n \end{bmatrix}}{\frac{q_1+q_2}{q_1}\log n\,(1+o(1)) - \frac{4(q_1+q_2)^2}{q_1^2}(1+o(1))} \begin{bmatrix} \tilde{a} \\ 1 \end{bmatrix} \\
&= \frac{(1+o(1))\left(\log n - \frac{4(q_1+q_2)}{q_1}\right)}{\frac{q_1+q_2}{q_1}\log n\,(1+o(1)) - \frac{4(q_1+q_2)^2}{q_1^2}(1+o(1))} \\
&= \frac{q_1}{q_1+q_2}(1+o(1))
\end{aligned}
$$

where in the second equality we have used the property for the minimum of a quadratic function. Suppose that $|\tilde{a}_{n_k}^*|$ is unbounded, then

$$\limsup_{k\to\infty} |\tilde{a}_{n_k}^*|^{\frac{2q_2}{q_1+q_2}} \left(n_k^{\frac{q_1}{q_1+q_2}} \bar{Z}_{n_k}^*\left(\tilde{a}_{n_k}^*\right)^2\right)^{\frac{q_1}{q_1+q_2}} = \infty$$

which is again a contradiction. Thus we are left with the case where $|\tilde{a}^*_{n_k}| > \sqrt{\frac{q_1}{q_1+q_2}\frac{1}{K^{q_1+q_2}}} + \epsilon$ and $|\tilde{a}^*_{n_k}|$ is bounded. Note that since $|\tilde{a}^*_{n_k}|$ is bounded we have

$$\left| n_k^{\frac{q_1}{q_1+q_2}} \bar{Z}^*_{n_k}\left(\tilde{a}^*_{n_k}\right)^2 - \frac{q_1}{q_1+q_2} \right| = o(1)$$

Thus

$$|\tilde{a}^*_{n_k}|^{\frac{2q_2}{q_1+q_2}} \left(n_k^{\frac{q_1}{q_1+q_2}} \bar{Z}^*_{n_k}\left(\tilde{a}^*_{n_k}\right)^2 \right)^{\frac{q_1}{q_1+q_2}} \geq \left(\sqrt{\frac{q_1}{q_1+q_2}\frac{1}{K^{q_1+q_2}}} + \epsilon \right)^{\frac{2q_2}{q_1+q_2}} \left(\frac{q_1}{q_1+q_2} + o(1) \right)^{\frac{q_1}{q_1+q_2}} \tag{74}$$

On the other hand, since the feasible region to (71) admits $\tilde{a}$ such that $\tilde{a} = \sqrt{\frac{q_1}{q_1+q_2}\frac{1}{K^{q_1+q_2}}} + o(1)$, we have for such $\tilde{a}$

$$|\tilde{a}|^{\frac{2q_2}{q_1+q_2}} \left(n_k^{\frac{q_1}{q_1+q_2}} \bar{Z}^*_{n_k}(\tilde{a})^2 \right)^{\frac{q_1}{q_1+q_2}} = \frac{q_1}{q_1+q_2}\frac{1}{K^{2q_2}} + o(1)$$

Comparing the above equation to (74), we again have a contradiction. Thus we have shown that the absolute value of a solution $\tilde{a}^*_n$ to (71) converges to $\sqrt{\frac{q_1}{q_1+q_2}\frac{1}{K^{q_1+q_2}}}$. Besides, we have

$$\eta^* = \left(\frac{\tilde{Z}^*_n(a^*)^2}{a^{*2}} \right)^{1/(2(q_1+q_2))} \to \left(\frac{\frac{q_1}{q_1+q_2}}{\frac{q_1}{q_1+q_2}\frac{1}{K^{2(q_1+q_2)}}} \right)^{1/(2(q_1+q_2))} = K \tag{75}$$

We now show that the error term in (65) is asymptotically negligible, which is true if

$$\sum_{j=1}^n \frac{w^*_j(1+o(1))}{j^{\alpha q_1}} = \sum_{j=1}^n \frac{w^*_j}{j^{\alpha q_1}} + o\left(\sum_{j=1}^n \frac{w^*_j}{j^{\alpha q_1}} \right) \tag{76}$$

and

$$\sum_{j=1}^n j^{2\alpha q_2} w_j^{*2}(1+o(1)) = \sum_{j=1}^n j^{2\alpha q_2} w_j^{*2} + o\left(\sum_{j=1}^n j^{2\alpha q_2} w_j^{*2} \right) \tag{77}$$

For (76), let $\gamma = \left(o\left(\frac{1}{j^{\alpha q_1}} \right) \right)_{j=1,\cdots,n} \in \mathbb{R}^n$. We first show that $\gamma^\top \Sigma^{-1}\mu = o(\mu^\top \Sigma^{-1}\mu)$. For any $\epsilon > 0$, by the definition of γ we have that $|\gamma_j| \leq \frac{\epsilon}{2}\mu_j$ for all $j > j_0$, for some $j_0 = j_0(\epsilon)$. Thus for all $n > j_0$

$$\gamma^\top \Sigma^{-1}\mu = \sum_{j=1}^n \gamma_j \Sigma^{-1}_{jj}\mu_j = \sum_{j=1}^{j_0} \gamma_j \Sigma^{-1}_{jj}\mu_j + \sum_{j=j_0+1}^n \gamma_j \Sigma^{-1}_{jj}\mu_j$$

where Σ^{-1}_{jj} denote the jth diagonal element of Σ^{-1}. Since $\mu^\top \Sigma^{-1}\mu \to \infty$ as $n \to \infty$, we have for all n large enough

$$\begin{aligned} \left|\gamma^\top \Sigma^{-1}\mu\right| &\leq \left| \sum_{j=1}^{j_0} \gamma_j \Sigma^{-1}_{jj}\mu_j \right| + \sum_{j=j_0+1}^n |\gamma_j| \Sigma^{-1}_{jj}\mu_j \\ &\leq \frac{\epsilon}{2}\mu^\top \Sigma^{-1}\mu + \frac{\epsilon}{2}\sum_{j=j_0+1}^n \mu_j \Sigma^{-1}_{jj}\mu_j \\ &\leq \epsilon \mu^\top \Sigma^{-1}\mu \end{aligned}$$

Thus $\gamma^\top\Sigma^{-1}\mu = o\left(\mu^\top\Sigma^{-1}\mu\right)$. Similarly we can show that $\gamma^\top\Sigma^{-1}\mathbb{1} = o\left(\mu^\top\Sigma^{-1}\mathbb{1}\right)$. We note that

$$\begin{aligned}
&\sum_{j=1}^{n} w_j^* o\left(\frac{1}{j^{\alpha q_1}}\right) \\
&= \left[\gamma^\top\Sigma^{-1}\mu \quad \gamma^\top\Sigma^{-1}\mathbb{1}\right] \begin{bmatrix} \phi(1) & \phi\left(\frac{q_1+2q_2}{2(q_1+q_2)}\right) \\ \phi\left(\frac{q_1+2q_2}{2(q_1+q_2)}\right) & \phi\left(\frac{q_2}{q_1+q_2}\right) \end{bmatrix}^{-1} \begin{bmatrix} a^* \\ 1 \end{bmatrix} \\
&= \left[o(\log n) \quad o\left(n^{\frac{q_1}{2(q_1+q_2)}}\right)\right] \frac{\begin{bmatrix} \frac{(q_1+q_2)(1+o(1))}{q_1} n^{\frac{q_1}{q_1+q_2}} & -\frac{2(q_1+q_2)(1+o(1))}{q_1} n^{\frac{q_1}{2(q_1+q_2)}} \\ -\frac{2(q_1+q_2)(1+o(1))}{q_1} n^{\frac{q_1}{2(q_1+q_2)}} & (1+o(1))\log n \end{bmatrix}}{\frac{(q_1+q_2)}{q_1} n^{\frac{q_1}{q_1+q_2}} \log n (1+o(1)) - \frac{4(q_1+q_2)^2}{q_1^2} n^{\frac{q_1}{q_1+q_2}} (1+o(1))} \begin{bmatrix} O\left(n^{\frac{q_1}{2(q_1+q_2)}}\right) \\ 1 \end{bmatrix} \\
&= \left[o(\log n) \quad o\left(n^{\frac{q_1}{2(q_1+q_2)}}\right)\right] \frac{\begin{bmatrix} O\left(n^{\frac{q_1}{2(q_1+q_2)}}\right) \\ O(\log n) \end{bmatrix}}{\frac{(q_1+q_2)}{q_1} n^{\frac{q_1}{q_1+q_2}} \log n (1+o(1)) - \frac{4(q_1+q_2)^2}{q_1^2} n^{\frac{q_1}{q_1+q_2}} (1+o(1))} \\
&= \frac{o\left(n^{\frac{q_1}{2(q_1+q_2)}} \log n\right)}{\frac{(q_1+q_2)}{q_1} n^{\frac{q_1}{q_1+q_2}} \log n (1+o(1)) - \frac{4(q_1+q_2)^2}{q_1^2} n^{\frac{q_1}{q_1+q_2}} (1+o(1))} \\
&= o\left(n^{-\frac{q_1}{2(q_1+q_2)}}\right) \\
&= o\left(\sum_{j=1}^{n} \frac{w_j^*}{j^{\alpha q_1}}\right)
\end{aligned}$$

where we have used the expression for w^*. For (77), since

$$n^{\frac{q_1}{q_1+q_2}} \sum_{j=1}^{n} \left(w_j^*\right)^2 o\left(j^{2\alpha q_2}\right) \to 0$$

we also have that

$$\sum_{j=1}^{n} \left(w_j^*\right)^2 j^{2\alpha q_2} o(1) = o\left(\sum_{j=1}^{n} \left(w_j^*\right)^2 j^{2\alpha q_2}\right)$$

Next, to show that no other choices of $W, g(\cdot)$ can asymptotically dominate $w^*(a^*)$ and $g(\cdot)$ where $g(d) = Kd$ obtained above, we consider a configuration of w, η obtained by solving w in

$$\begin{array}{ll} \min_w & Q = \frac{1}{K^{2q_2}} \sum_{j=1}^{n} j^{2\alpha q_2} w_j^2 \\ \text{subject to} & \frac{1}{K^{2q_2}} \sum_{j=1}^{n} j^{2\alpha q_2} w_j^2 > K^{2q_1} \left(\sum_{j=1}^{n} \frac{w_j}{j^{\alpha q_1}}\right)^2 \\ & \sum_{j=1}^{n} w_j = 1 \end{array} \tag{78}$$

and choosing $\eta = K$. Let Q_n^* the optimal value of (78). We first solve (78) and show that it does not give a smaller optimal value than (36) asymptotically. Consider

$$\begin{array}{rll} \tilde{L}_n(a) = & \min_w & \|\Sigma^{1/2}w\| \\ & \text{subject to} & \|\Sigma^{1/2}w\|^2 > K^{2(q_1+q_2)}a^2 \\ & & \mu^\top w = a \\ & & \mathbb{1}^\top w = 1 \end{array} \tag{79}$$

For any a, if the optimal solution to (43) satisfies

$$\tilde{Z}_n^*(a)^2 > K^{2(q_1+q_2)}a^2$$

then the minimum in definition (79) is attainable and $\tilde{L}_n(a) = \tilde{Z}_n^*(a)$. Otherwise, the minimum is possibly unattainable and $\tilde{L}_n(a) \geq K^{2(q_1+q_2)}a^2$. Let $a = \tilde{a}n^{-\frac{q_1}{2(q_1+q_2)}}$. Reparametrizing by $\tilde{a}$, we denote $\bar{L}_n(\tilde{a}) = \tilde{L}_n\left(\tilde{a}n^{-\frac{q_1}{2(q_1+q_2)}}\right)$. Multiplying the objective value of (78) by $n^{\frac{q_1}{q_1+q_2}}$, we have

$$n^{\frac{q_1}{q_1+q_2}}Q_n^* = n^{\frac{q_1}{q_1+q_2}}\inf_{\tilde{a}}\bar{L}_n(\tilde{a})^2\frac{1}{K^{2q_2}}$$

regardless of whether the minimum in (78) is attainable. Suppose that $n^{\frac{q_1}{q_1+q_2}}\bar{Z}_n^*(\tilde{a})^2 > K^{2(q_1+q_2)}\tilde{a}^2$. From (70) we have that $\tilde{a}$ is asymptotically bounded. Thus for some $o(1)$ uniform over such $\tilde{a}$, we have

$$n^{\frac{q_1}{q_1+q_2}}\bar{L}_n(\tilde{a})^2\frac{1}{K^{2q_2}} = n^{\frac{q_1}{q_1+q_2}}\bar{Z}_n^*(\tilde{a})^2\frac{1}{K^{2q_2}} \geq \frac{q_1}{q_1+q_2}(1+o(1))\frac{1}{K^{2q_2}}$$

On the other hand, suppose that $n^{\frac{q_1}{q_1+q_2}}\bar{Z}_n^*(\tilde{a})^2 \leq K^{2(q_1+q_2)}\tilde{a}^2$. Then

$$\begin{aligned}
n^{\frac{q_1}{q_1+q_2}}\bar{L}_n(\tilde{a})^2\frac{1}{K^{2q_2}} &\geq K^{2(q_1+q_2)}\tilde{a}^2\frac{1}{K^{2q_2}} \\
&\geq \left(K^{2(q_1+q_2)}\tilde{a}^2\right)^{\frac{q_2}{q_1+q_2}}\left(n^{\frac{q_1}{q_1+q_2}}\bar{Z}_n^*(\tilde{a})^2\right)^{\frac{q_1}{q_1+q_2}}\frac{1}{K^{2q_2}} \\
&\geq \min_{\tilde{a}:n^{\frac{q_1}{q_1+q_2}}\bar{Z}_n^*(\tilde{a})^2\leq K^{2(q_1+q_2)}\tilde{a}^2}|\tilde{a}|^{\frac{2q_2}{q_1+q_2}}\left(n^{\frac{q_1}{q_1+q_2}}\bar{Z}_n^*(\tilde{a})^2\right)^{\frac{q_1}{q_1+q_2}} \\
&\geq \frac{q_1}{q_1+q_2}\frac{1}{K^{2q_2}}(1+o(1))
\end{aligned}$$

for some $o(1)$ independent of $\tilde{a}$. Therefore, we have

$$\liminf_{n\to\infty} n^{\frac{q_1}{q_1+q_2}}Q_n^* \geq \lim_{n\to\infty} n^{\frac{q_1}{q_1+q_2}}S_n^*$$

Using (72) we identify the AMRR in the first part of the theorem. Using (47), (67), (68), (69) and (75) we identify the solution in the second part of the theorem.

It remains to argue that no other configurations $w, g(\cdot)$ such that $\sum_j^n w_j \to 1$ and $g(\cdot) \in \mathcal{F}_K$ that can give a better risk ratio. We first note that we can solve the variant of optimization (36)

$$\begin{array}{ll}
\min_{w,\eta} & T \\
\text{subject to} & T = \left(\eta^{q_1}\sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}}\right)^2 = \frac{1}{\eta^{2q_2}}\sum_{j=1}^n j^{2\alpha q_2}w_j^2 \\
& \eta \leq K \\
& \sum_{j=1}^n w_j = 1 + o(1)
\end{array} \qquad (80)$$

via solving (45) like before, but this time with the constraint $\mathbb{1}^\top w = 1$ in (43) replaced by $\mathbb{1}'w = 1 + o(1)$. This additional $o(1)$ term can be seen, by following the arguments above, to eventually be absorbed with no effect on the analysis. This gives an optimal solution T_n^* such that $\lim_{n\to\infty} T_n^*/S_n^* = 1$. Similarly, the variant of optimization (78)

$$\begin{array}{ll} \min_w & P = \frac{1}{K^{2q_2}} \sum_{j=1}^n j^{2\alpha q_2} w_j^2 \\ \text{subject to} & \frac{1}{K^{2q_2}} \sum_{j=1}^n j^{2\alpha q_2} w_j^2 > K^{2q_1} \left(\sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right)^2 \\ & \sum_{j=1}^n w_j = 1 + o(1) \end{array} \tag{81}$$

gives an optimal value P_n^* such that $\liminf_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} P_n^* \geq \lim_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} T_n^* = \frac{q_1}{q_1+q_2} \frac{1}{K^{2q_2}}$.

We aim to find $\theta(\cdot) \in H$ and $d > 0$, such that

$$R^{gen}(\theta(\cdot), d, g(d), W) \geq \frac{q_1}{q_1+q_2} \frac{1}{K^{2q_2}}$$

We will consider $\theta(\cdot) \in H$ with $\theta_0 = 0$ and without the higher order terms in the asymptotic expansion, i.e. $b(\delta) = B\delta^{q_1}$ for some $B \neq 0$ and $v(\delta) = \frac{\epsilon(\delta)}{\delta^{q_2}}$ such that $Var(\epsilon(\delta)) = \sigma^2 > 0$. In this case

$$\text{MSE}_1 = \left(B d^{q_1} \left(\frac{g(d)}{d} \right)^{q_1} \sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right)^2 + \frac{\sigma^2}{d^{2q_2}} \left(\frac{d}{g(d)} \right)^{2q_2} \sum_{j=1}^n j^{2\alpha q_2} w_j^2$$

For any $W, g(\cdot)$, we note that two cases can arise:

1. For all large enough n, either

$$\left(\frac{g(d)}{d} \right)^{2q_1} \left(\sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right)^2 = \left(\frac{d}{g(d)} \right)^{2q_2} \sum_{j=1}^n j^{2\alpha q_2} w_j^2$$

or

$$\left(\frac{g(d)}{d} \right)^{2q_1} \left(\sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right)^2 \neq \left(\frac{d}{g(d)} \right)^{2q_2} \sum_{j=1}^n j^{2\alpha q_2} w_j^2$$

but there exists $\eta \leq K$, such that

$$\eta^{2q_1} \left(\sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right)^2 = \frac{1}{\eta^{2q_2}} \sum_{j=1}^n j^{2\alpha q_2} w_j^2$$

2. There exists a subsequence n_k such that

$$K^{2q_1} \left(\sum_{j=1}^{n_k} \frac{w_j}{j^{\alpha q_1}} \right)^2 < \frac{1}{K^{2q_2}} \sum_{j=1}^{n_k} j^{2\alpha q_2} w_j^2$$

For case 1, by the definition of T_n^* we have

$$\max\{ \left(\frac{g(d)}{d} \right)^{2q_1} \left(\sum_{j=1}^n \frac{w_j}{j^{\alpha q_1}} \right)^2, \left(\frac{d}{g(d)} \right)^{2q_2} \sum_{j=1}^n j^{2\alpha q_2} w_j^2 g\} \geq T_n^*$$

Thus

$$\begin{aligned}
&\max_{\theta(\cdot)\in H, d>0} R^{gen}\left(\theta\left(\cdot\right), d, g\left(d\right), W\right)\\
&\geq \max_{B\neq 0,\sigma^2>0,d>0} \limsup_{n\to\infty} \frac{\left(Bd^{q_1}\left(\frac{g(d)}{d}\right)^{q_1}\sum_{j=1}^{n}\frac{w_j}{j^{\alpha q_1}}\right)^2+\frac{\sigma^2}{d^{2q_2}}\left(\frac{d}{g(d)}\right)^{2q_2}\sum_{j=1}^{n} j^{2\alpha q_2}w_j^2}{\frac{1}{n^{\frac{q_1}{q_1+q_2}}}\left(B^2d^{2q_1}+\frac{\sigma^2}{d^{2q_2}}\right)+o\left(\frac{1}{n^{\frac{q_1}{q_1+q_2}}}\right)}\\
&\geq \lim_{n\to\infty} n^{\frac{q_1}{q_1+q_2}} T_n^*\\
&\geq \frac{q_1}{q_1+q_2}\frac{1}{K^{2q_2}}
\end{aligned}$$

For case 2, we have

$$\left(\frac{g\left(d\right)}{d}\right)^{2q_1}\left(\sum_{j=1}^{n_k}\frac{w_j}{j^{\alpha q_1}}\right)^2 \leq K^{2q_1}\left(\sum_{j=1}^{n_k}\frac{w_j}{j^{\alpha q_1}}\right)^2 < \frac{1}{K^{2q_2}}\sum_{j=1}^{n_k} j^{2\alpha q_2}w_j^2 \leq \left(\frac{d}{g\left(d\right)}\right)^{2q_2}\sum_{j=1}^{n_k} j^{2\alpha q_2}w_j^2$$

Thus by the definition of P_n^*

$$\begin{aligned}
&\max_{\theta(\cdot)\in H, d>0} R^{gen}\left(\theta\left(\cdot\right), d, g\left(d\right), W\right)\\
&\geq \max_{B\neq 0,\sigma^2>0,d>0} \limsup_{k\to\infty} \frac{\left(Bd^{q_1}\left(\frac{g(d)}{d}\right)^{q_1}\sum_{j=1}^{n_k}\frac{w_j}{j^{\alpha q_1}}\right)^2+\frac{\sigma^2}{d^{2q_2}}\left(\frac{d}{g(d)}\right)^{2q_2}\sum_{j=1}^{n_k} j^{2\alpha q_2}w_j^2}{\frac{1}{n_k^{\frac{q_1}{q_1+q_2}}}\left(B^2d^{2q_1}+\frac{\sigma^2}{d^{2q_2}}\right)+o\left(\frac{1}{n_k^{\frac{q_1}{q_1+q_2}}}\right)}\\
&\geq \max_{B\neq 0,\sigma^2>0,d>0} \limsup_{k\to\infty} n_k^{\frac{q_1}{q_1+q_2}}\frac{1}{K^{2q_2}}\sum_{j=1}^{n_k} j^{2\alpha q_2}w_j^2 \frac{B^2d^{2q_1}\frac{\left(\left(\frac{g(d)}{d}\right)^{q_1}\sum_{j=1}^{n_k}\frac{w_j}{j^{\alpha q_1}}\right)^2}{\frac{1}{K^{2q_2}}\sum_{j=1}^{n_k} j^{2\alpha q_2}w_j^2}+\frac{\sigma^2}{d^{2q_2}}}{\left(B^2d^{2q_1}+\frac{\sigma^2}{d^{2q_2}}\right)+o\left(1\right)}\\
&\geq \limsup_{k\to\infty} n_k^{\frac{q_1}{q_1+q_2}}\frac{1}{K^{2q_2}}\sum_{j=1}^{n_k} j^{2\alpha q_2}w_j^2 \quad \text{(by considering } B \text{ arbitrarily close to 0)}\\
&\geq \limsup_{k\to\infty} n_k^{\frac{q_1}{q_1+q_2}} P_{n_k}^*\\
&\geq \frac{q_1}{q_1+q_2}\frac{1}{K^{2q_2}}
\end{aligned}$$

□

Proof of Corollary 1. This follows immediately by noting that the proof of Theorem 8 applies exactly the same when d is fixed. □

Appendix D: Finite-sample effect of the choice of d

We investigate the effect of the choice of d on the finite-sample risk ratios between our proposed estimator $\hat{\theta}_n^{gen}$ and the sample-average-based estimator $\bar{\theta}_n$ via a simple numerical experiment. Consider estimating

$$\theta_0 = \frac{d}{dx}E\left[1\{Z>x\}\right]|_{x=0} = \frac{d}{dx}P(Z>x)|_{x=0} = -\phi(x)|_{x=0} = -\phi(0),$$

where Z is a standard normal random variable and $\phi(\cdot)$ is the standard normal density. We can evaluate the simulation outputs where, for $\delta > 0$,

$$\theta(\delta) = \frac{1\{Z > \delta\} - 1\{Z > 0\}}{\delta}.$$

In the form of (1), the bias $b(\delta)$ is given by

$$\begin{aligned} b(\delta) = E\left[\theta(\delta)\right] - \theta_0 &= \frac{P(Z > \delta) - P(Z > 0)}{\delta} + \phi(0) = \frac{-\int_0^\delta \phi(x)dx}{\delta} + \phi(0) \\ &= -\frac{\phi(0)\delta + \frac{1}{2}\phi^{'}(0)\delta^2 + \frac{1}{6}\phi^{''}(0)\delta^3 + o(\delta^3)}{\delta} + \phi(0) = -\frac{1}{6}\phi^{''}(0)\delta^2 + o(\delta^2) \end{aligned}$$

since $\phi^{'}(0) = 0$. Similarly, the noise $v(\delta)$ is given by $v(\delta) = \theta(\delta) - E[\theta(\delta)]$ so that

$$Var(v(\delta)) = \frac{Var(1\{Z > \delta\} - 1\{Z > 0\})}{\delta^2} = \frac{Var(R)}{\delta^2},$$

where R takes value -1 with probability $P(0 < Z \leq \delta)$ and 0 with probability $1 - P(0 < Z \leq \delta)$. Therefore

$$Var(R) = P(0 < Z \leq \delta)(1 - P(0 < Z \leq \delta)) = \phi(0)\delta + o(\delta).$$

Therefore, the constants in the set H (in (10)) are $B = -\frac{1}{6}\phi^{''}(0), \sigma^2 = \phi(0), q_1 = 2, q_2 = \frac{1}{2}$. Table 11 shows the AMRR obtained from Corollary 1.

K	0.5	0.8	1.0	2.0	3.0	4.0
AMRR	1.60	1.00	0.80	0.40	0.27	0.20

Table 11 AMRR for general weighted estimators, against K, when $q_1 = 2, q_2 = \frac{1}{2}$

We fix a budget $n = 20, n_0 = 5$, choose values of K ranging from 0.5 to 4.0, and choose values of d among $\{0.5, 1, d^\star, 5\}$, where $d^\star$ is the optimal choice for the sample-average-based estimator $\bar{\theta}_n$, which can be calculated by

$$d^\star = \left(\frac{\sigma^2 q_2}{B^2 q_1}\right)^{\frac{1}{2(q_1+q_2)}} \approx 1.86.$$

For each configuration of K and d, we repeat the simulation for 1000 times to estimate the empirical MSEs of $\bar{\theta}_n$ and $\hat{\theta}_n^{gen}$. Moreover, we output the 95% confidence intervals for the risk ratios, which are obtained by a standard application of the delta method. Table 12 summarizes the finite-sample risk ratios together with the confidence intervals. We see that the choice of d indeed affects the finite-sample performance of our proposed estimator. When $d = 0.5$, the finite-sample risk ratios closely match with the theoretical AMRR for all values of K considered. As d increases, we see that the finite-sample risk ratios still roughly match the theoretical AMRR for K ranging from

0.5 to 1, but begin to deviate away from the theoretical AMRR for K ranging from 2 to 4. The deviation is largest for $d = 1.86$, which is intuitively unsurprising as this value of d is optimal for the conventional estimator $\bar{\theta}_n$. When $d = 5$, we see more fluctuations away from the theoretical AMRR for K ranging from 0.5 to 1. Although some deviations are in our favor, this indicates a significant finite-sample effect for such a large value of d.

d	$K = 0.5$	$K = 0.8$	$K = 1$	$K = 2$	$K = 3$	$K = 4$
0.5	$182 \pm 24\%$	$101 \pm 14\%$	$89 \pm 11\%$	$37 \pm 5\%$	$23 \pm 3\%$	$21 \pm 2\%$
1	$229 \pm 30\%$	$107 \pm 13\%$	$91 \pm 11\%$	$51 \pm 6\%$	$71 \pm 7\%$	$129 \pm 11\%$
1.86	$186 \pm 23\%$	$103 \pm 11\%$	$98 \pm 11\%$	$228 \pm 19\%$	$416 \pm 33\%$	$559 \pm 46\%$
5	$37 \pm 3\%$	$87 \pm 3\%$	$112 \pm 4\%$	$217 \pm 6\%$	$255 \pm 6\%$	$272 \pm 7\%$

Table 12 Finite-sample risk ratios with 95% CIs between $\hat{\theta}_n^{gen}$ and $\bar{\theta}_n$.

Appendix E: Further numerical illustrations in the application of stochastic optimization

We further study an incorporation of our enhanced estimators in zeroth-order stochastic gradient descent or SA for black-box stochastic optimization problem. Here, the gradients in the descent algorithm are estimated via finite differences. We consider the minimization of $f(\theta) = \mathbb{E}[F(\theta, \xi)]$, the expected value of a function depending on the random variable ξ and use the iterative algorithm shown in (82) to obtain solutions, where the gradient is approximated using some weighted finite difference (83):

$$\theta_{i+1} = \theta_i - \alpha_i \widehat{\nabla f(\theta_i)}, i = 1, \ldots, I \tag{82}$$

$$\widehat{\nabla f(\theta_i)} = \sum_{j=1}^{n_{\text{sim}}} w_j \frac{F(\theta_i + \delta_{i,j}, \xi_{i,j}) - F(\theta_i - \delta_{i,j}, \xi'_{i,j})}{2\delta_{i,j}}. \tag{83}$$

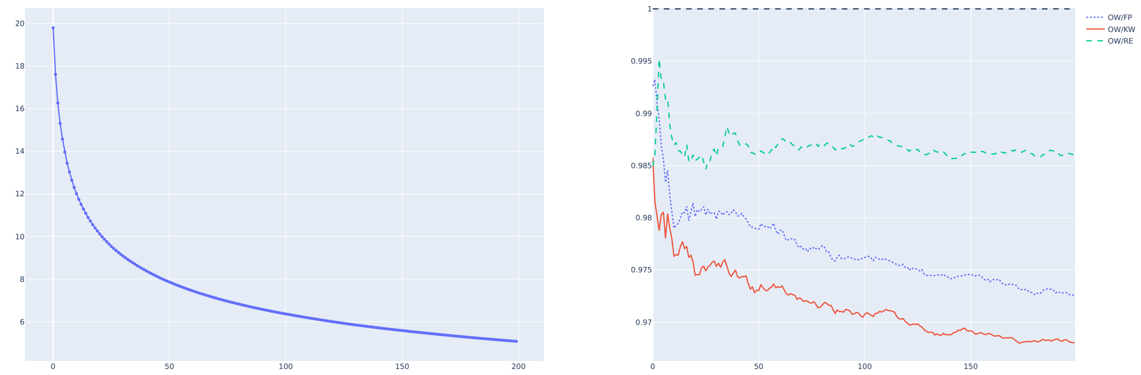
where $\alpha_1, \alpha_2, \ldots$ is a sequence of positive step sizes with i as the iteration index, and I is the total iteration number. $\{\xi_{i,j}\}_{j=1}^{n_{\text{sim}}}$ and $\{\xi'_{i,j}\}_{j=1}^{n_{\text{sim}}}$ are independent realizations of the random variable ξ, w_j is the weight coefficient and $\delta_{i,j}$ is the perturbation size for each finite difference. Four variants of finite difference are studied in our experiment:

1. (FP): Fixed perturbation, with $\delta_{i,j} = \frac{d}{(n_0 + n_{\text{sim}})^{1/6}}$ and $w_j = \frac{1}{n_{\text{sim}}}$;
2. (KW): Kiefer-Wolfowitz, with $\delta_{i,j} = \frac{d}{(n_0 + i)^{1/6}}$ and $w_j = \frac{1}{n_{\text{sim}}}$;
3. (RE): Recursive estimator, with $\delta_{i,j} = \frac{\tilde{d}}{(n_0 + j)^{1/6}}$, $\tilde{d} = 0.83d$ and $w_j = \frac{1}{n_{\text{sim}}}$;
4. (OW): Optimal weighting, with $\delta_{i,j} = \frac{\tilde{d}}{(n_0 + j)^{1/6}}$, $\tilde{d} = Kd$ and $w_j = w_j^{\text{opt}}$ given by Theorem 8.

We consider an objective function F expressed as a sum of deterministic and stochastic parts, i.e., $F(\theta,\xi)=F_1(\theta)+F_2(\xi)$. Two experimental sets are considered with different deterministic and stochastic parts. In the first experimental set, the deterministic part $F_1(\theta)$ is selected from one of $|\theta|,|\theta|^{1.1},|\theta|^{1.2},|\theta|^{1.3},|\theta|^{1.4}$, and the stochastic part $F_2(\xi)=\xi\sim N(0,\sigma^2)$ varies by the choice of σ to be among 1 to 10. In the second set, $F_1(\theta)$ is selected from one of $(2-\cos\theta)^1-1,(2-\cos\theta)^{1.1}-1,(2-\cos\theta)^{1.2}-1,(2-\cos\theta)^{1.3}-1,(2-\cos\theta)^{1.4}-1$, and $F_2(\xi)=\xi\sim N(0,\sigma^2)$ with σ chosen among 5 to 10. In both experimental sets, the true optimum is 0. Note that F_1 in the first set may not be (higher-order) differentiable at the optimum 0, and thus does not guarantee convergence of the standard Kiefer-Wolfowitz algorithm (Kiefer and Wolfowitz (1952)). However, in our implementation we focus on the early stage of the iterations where the updated solutions are still far away from the true optimum, and the easy form of F_1 facilitates the understanding on the impacts of curvature on the performances of the tested algorithms. On the other hand, F_1 in the second experimental set is infinitely differentiable everywhere in $\mathbb{R}$ and we would see similar experimental observations as the first set regarding early-iteration behaviors.

We first focus on the first experimental set. There is a combination of 50 ($=5\times10$) possible objective functions to be considered in total. We vary the simulation run-length per iteration n_{sim} to be among $100,150,200$, and step size α_i among $0.2,0.2\cdot i^{-\frac{6}{7}},0.2\cdot i^{-\frac{7}{8}},0.2\cdot i^{-1}$. Also we set $I=n_{\text{sim}}$. This is because we would like the order of $\delta_{i,j}$ among the four methods to be close to each other at the end of the iteration, in order to avoid significant impacts on the solution quality due to huge discrepancies in the magnitude of $\delta_{i,j}$ (note that KW has decreasing $\delta_{i,j}$ in i while the other three methods do not). We set $d=1,K=1,n_0=50,\theta_1=5$. Our performance metric is the MSE of the obtained solution at the last iteration, i.e., θ_{I+1} against the ground-truth optimal solution. We repeat the experiment 200 times to estimate the empirical MSEs.

Tables 13-16 summarize the results, where Table 13 shows the performance for fixed step size and Tables 14, 15 and 16 show those for varying step sizes $0.2\cdot i^{-\frac{6}{7}}$, $0.2\cdot i^{-\frac{7}{8}}$ and $0.2\cdot i^{-1}$ respectively. The varying step sizes all satisfy the typical requirement needed for convergence of the Kiefer-Wolfowitz algorithm. However, as noted earlier, in this experimental set differentiability at the optimum does not hold, while in our next experimental set differentiability would hold so that the algorithm asymptotically converges. Moreover, note that, as in Section 7, to interpret these tables, one should focus on the ratios of MSEs instead of the absolute magnitude of the MSE. This is because one can always arbitrarily inflate or deflate MSEs by simple scalar multiplication adjustments. Thus, an appropriate measurement of the estimation error is the ratio between the MSEs of two different algorithms (i.e., in the form $\text{MSE}_1/\text{MSE}_2$ where 1 and 2 represent some algorithms). Figure 3 gives one typical numerical trajectory when setting the objective function as $|\theta|^{1.4}+\xi,\xi\sim N(0,10^2)$, and $n_{\text{sim}}=200$ and $\alpha_i=0.2\cdot i^{-1}$. In this setting, the empirical MSE given by OW gradually shrinks as shown in Figure 3a, and the effectiveness of OW is further demonstrated by the MSE ratio curves in Figure 3b, where the ratios are well below the value 1 (dashed horizontal line) throughout the 200 update steps. Note that the solution at the 200-th step is still quite far away from the true optimum, hinted by the MSE being noticeably far away from 0 (as mentioned before, we focus on the early stage of iterations in this experimental set).



(a) MSE via OW against iteration number

(b) MSE ratio against iteration number

Figure 3 Performance trajectories for $F(\theta,\xi)=|\theta|^{1.4}+\xi,\xi\sim N(0,10^2)$ and $n_{\text{sim}}=200,\alpha_i=0.2\cdot i^{-1}$

We make several observations from Tables 13-16. First, throughout all the experiment settings, the empirical MSE ratios ($\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{FP}}}/\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{KW}}}/\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{RE}}}$) are all strictly less than 1. These ratios can reach as low as 0.66, 0.72 and 0.64 (in Table 13) respectively, implying a closer solution to the ground-truth optimal in using OW than other benchmark methods given the same total number of update steps. Note that, for more statistical preciseness, we can construct the 95% confidence intervals for these empirical ratios, where most of them would be seen to still lie strictly less than 1, but for ease of presentation we do not show these intervals in the tables.

Another noticeable pattern is observed along the choice of the stochastic part. The empirical MSE ratios all decrease as the variance parameter σ increases, indicating that OW gives a relatively smaller MSE when the function evaluation bears a higher variance. For example, for $\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{KW}}}$, the ratio monotonically decreases from 0.87 to 0.78 as σ increases from 1 to 10 under the setting of $F_1(\theta)=|\theta|^{1.4},\alpha_k=0.2$ and $n_{\text{sim}}=100$ in Table 13. Similar patterns can be found across different choices of n_{sim} and update step size. These observations hint that the bias-variance balancing using OW is more effective when the objective function evaluation is subject to a higher noise.

We notice that a larger n_{sim} may not necessarily lead to a greater improvement in term of MSE ratios. This could be due to several factors, including the step-size-sensitive behavior of the descent algorithm and the approximation errors in the bias-variance balancing analysis for finite-difference estimators mentioned in Sections 5.1 and 7. In addition, compared to Tables 14, 15 and 16 where we set varying update sizes, greater improvements are observed under the scenario of fixed step size shown in Table 13. This could be because during the early iterations accurate estimation of gradients obtained by OW has a larger influence on the solution update under the fixed-step scheme, while such influence is smaller for the other three schemes as their update sizes shrink along the iterations.

In the second experimental set, we again vary the simulation run-length per iteration n_{sim} to be among $100, 150, 200$. We vary the step size α_i among $0.1 \cdot i^{-6/7},\ 0.1 \cdot i^{-7/8},\ 0.1 \cdot i^{-1}$, and again set $I = n_{\text{sim}}$. We set $d = 1,\ K = 1,\ n_0 = 50,\ \theta_1 = 2$. We use the same criterion of MSE as in the first experimental set, and repeat the experiment 200 times to estimate the empirical MSEs.

Results are given in Tables 17-19. We observe similar patterns as in the first experimental set. OW obtains the smallest empirical MSE among the four methods in all cases. Moreover, when the objective function is subject to a larger variance of noise, i.e., larger σ, a larger outperformance of OW over other methods is observed as exhibited by smaller MSE ratios. Here, like in the first experimental set, we have focused on the early iterations and the solutions obtained are still far away from the optimum. Thus, the outperformance of OW can be attributed to a faster decay of the objective function brought by more accurate gradient estimation in the early stage of the descent algorithm.

Appendix References

Fabian V (1967) Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics* 191–200.

Fabian V (1968) On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics* 39(4):1327–1332.

Kiefer J, Wolfowitz J (1952) Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics* 23(3):462 – 466.

Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization* 30(4):838–855.

n_{sim}	σ	$\lvert\theta\rvert^{1}$	$\lvert\theta\rvert^{1.1}$	$\lvert\theta\rvert^{1.2}$	$\lvert\theta\rvert^{1.3}$	$\lvert\theta\rvert^{1.4}$
100	1	0.818/0.885/0.773	0.815/0.882/0.776	0.812/0.878/0.777	0.807/0.873/0.778	0.802/0.868/0.777
	2	0.818/0.885/0.773	0.815/0.882/0.776	0.811/0.877/0.777	0.806/0.873/0.777	0.801/0.867/0.777
	3	0.818/0.885/0.773	0.814/0.881/0.775	0.810/0.876/0.776	0.804/0.871/0.777	0.799/0.866/0.776
	4	0.817/0.884/0.773	0.812/0.879/0.774	0.806/0.873/0.775	0.800/0.868/0.775	0.795/0.862/0.774
	5	0.798/0.874/0.769	0.795/0.869/0.770	0.791/0.864/0.771	0.787/0.859/0.771	0.783/0.854/0.770
	6	0.765/0.848/0.754	0.766/0.847/0.758	0.767/0.845/0.760	0.767/0.843/0.762	0.766/0.840/0.763
	7	0.735/0.813/0.739	0.740/0.818/0.745	0.744/0.821/0.749	0.748/0.823/0.753	0.750/0.824/0.755
	8	0.710/0.774/0.715	0.720/0.785/0.727	0.727/0.794/0.735	0.733/0.801/0.742	0.738/0.807/0.746
	9	0.685/0.741/0.684	0.699/0.756/0.705	0.710/0.769/0.719	0.719/0.780/0.729	0.727/0.790/0.737
	10	0.663/0.715/0.655	0.682/0.733/0.682	0.695/0.748/0.703	0.707/0.762/0.717	0.717/0.775/0.727
150	1	0.774/0.820/0.972	0.769/0.814/0.973	0.764/0.809/0.974	0.759/0.804/0.974	0.755/0.800/0.973
	2	0.774/0.820/0.972	0.769/0.814/0.973	0.764/0.809/0.974	0.759/0.804/0.974	0.755/0.800/0.973
	3	0.774/0.820/0.972	0.769/0.814/0.973	0.763/0.809/0.973	0.759/0.804/0.973	0.754/0.800/0.972
	4	0.774/0.820/0.972	0.768/0.814/0.973	0.763/0.808/0.973	0.758/0.804/0.972	0.754/0.799/0.971
	5	0.772/0.819/0.970	0.766/0.812/0.971	0.761/0.807/0.971	0.756/0.802/0.970	0.752/0.798/0.969
	6	0.765/0.816/0.964	0.760/0.809/0.966	0.755/0.804/0.966	0.751/0.800/0.967	0.748/0.796/0.966
	7	0.757/0.810/0.968	0.753/0.805/0.967	0.750/0.800/0.967	0.747/0.796/0.966	0.744/0.793/0.966
	8	0.751/0.802/0.972	0.748/0.799/0.970	0.746/0.795/0.969	0.743/0.793/0.967	0.741/0.790/0.965
	9	0.746/0.797/0.963	0.741/0.796/0.964	0.739/0.794/0.965	0.738/0.792/0.965	0.737/0.790/0.963
	10	0.746/0.798/0.952	0.741/0.801/0.957	0.737/0.799/0.959	0.735/0.795/0.959	0.735/0.792/0.960
200	1	0.821/0.852/0.746	0.816/0.847/0.747	0.811/0.841/0.747	0.807/0.835/0.747	0.803/0.828/0.745
	2	0.821/0.852/0.746	0.816/0.847/0.747	0.811/0.841/0.747	0.807/0.834/0.747	0.803/0.827/0.745
	3	0.821/0.852/0.746	0.816/0.846/0.747	0.811/0.840/0.747	0.807/0.833/0.746	0.803/0.825/0.745
	4	0.821/0.850/0.744	0.816/0.843/0.745	0.811/0.836/0.745	0.807/0.829/0.744	0.803/0.822/0.743
	5	0.821/0.837/0.735	0.816/0.831/0.738	0.811/0.825/0.739	0.806/0.819/0.740	0.802/0.813/0.739
	6	0.816/0.822/0.724	0.811/0.817/0.728	0.807/0.812/0.731	0.803/0.807/0.733	0.800/0.802/0.734
	7	0.813/0.792/0.709	0.808/0.792/0.716	0.804/0.790/0.721	0.801/0.789/0.724	0.798/0.787/0.727
	8	0.801/0.768/0.685	0.799/0.771/0.699	0.798/0.772/0.708	0.796/0.773/0.715	0.794/0.774/0.719
	9	0.789/0.742/0.659	0.790/0.748/0.678	0.791/0.754/0.692	0.790/0.758/0.703	0.790/0.761/0.711
	10	0.781/0.718/0.638	0.783/0.727/0.660	0.785/0.735/0.678	0.787/0.743/0.692	0.787/0.749/0.703

Table 13 Ratios of empirical MSEs for our first experimental set for $\alpha_i = 0.2$. 200 replications to estimate empirical MSEs. Each entry reads $\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{FP}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{KW}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{RE}}}$.

n_{sim}	σ	$\lvert\theta\rvert^{1}$	$\lvert\theta\rvert^{1.1}$	$\lvert\theta\rvert^{1.2}$	$\lvert\theta\rvert^{1.3}$	$\lvert\theta\rvert^{1.4}$
100	1	1.000/0.997/0.998	1.000/0.997/0.998	1.000/0.997/0.998	1.000/0.996/0.998	1.000/0.996/0.997
	2	1.000/0.994/0.996	1.000/0.994/0.996	1.000/0.993/0.996	0.999/0.992/0.995	0.999/0.991/0.995
	3	0.999/0.992/0.995	0.999/0.991/0.994	0.999/0.990/0.994	0.999/0.989/0.993	0.999/0.988/0.992
	4	0.999/0.989/0.993	0.999/0.988/0.992	0.999/0.987/0.992	0.999/0.986/0.991	0.999/0.984/0.990
	5	0.999/0.987/0.991	0.999/0.986/0.990	0.999/0.984/0.990	0.999/0.983/0.989	0.998/0.981/0.987
	6	0.999/0.984/0.989	0.999/0.983/0.989	0.999/0.982/0.988	0.999/0.980/0.987	0.998/0.978/0.985
	7	0.999/0.982/0.988	0.999/0.981/0.987	0.999/0.979/0.986	0.998/0.977/0.985	0.998/0.975/0.983
	8	0.999/0.980/0.986	0.999/0.978/0.985	0.999/0.977/0.984	0.998/0.975/0.983	0.998/0.972/0.981
	9	0.999/0.978/0.985	0.999/0.976/0.984	0.999/0.975/0.982	0.998/0.972/0.981	0.998/0.970/0.979
	10	0.999/0.976/0.983	0.999/0.974/0.982	0.999/0.972/0.981	0.999/0.970/0.979	0.998/0.968/0.977
150	1	0.999/0.998/0.999	0.999/0.998/0.999	0.999/0.998/0.999	0.999/0.997/0.999	0.999/0.997/0.999
	2	0.998/0.996/0.999	0.998/0.996/0.999	0.998/0.996/0.998	0.998/0.995/0.998	0.998/0.994/0.998
	3	0.997/0.995/0.998	0.997/0.994/0.998	0.997/0.993/0.998	0.997/0.993/0.997	0.996/0.992/0.997
	4	0.996/0.993/0.997	0.996/0.992/0.997	0.996/0.991/0.997	0.995/0.990/0.996	0.994/0.989/0.996
	5	0.995/0.991/0.996	0.995/0.990/0.996	0.994/0.989/0.996	0.993/0.988/0.995	0.992/0.987/0.995
	6	0.994/0.989/0.996	0.994/0.988/0.995	0.993/0.987/0.995	0.991/0.986/0.994	0.990/0.984/0.993
	7	0.993/0.988/0.995	0.992/0.987/0.994	0.991/0.985/0.994	0.989/0.984/0.993	0.987/0.982/0.992
	8	0.992/0.986/0.994	0.991/0.985/0.993	0.989/0.983/0.992	0.987/0.981/0.991	0.985/0.979/0.990
	9	0.990/0.984/0.993	0.989/0.983/0.992	0.987/0.981/0.991	0.985/0.979/0.990	0.982/0.977/0.989
	10	0.989/0.983/0.992	0.987/0.981/0.991	0.985/0.980/0.990	0.982/0.977/0.989	0.979/0.975/0.987
200	1	0.998/0.996/0.999	0.997/0.996/0.999	0.997/0.995/0.999	0.997/0.995/0.998	0.997/0.994/0.998
	2	0.995/0.993/0.998	0.995/0.992/0.997	0.994/0.991/0.997	0.993/0.990/0.997	0.993/0.988/0.997
	3	0.992/0.989/0.996	0.992/0.988/0.996	0.991/0.987/0.996	0.989/0.985/0.995	0.988/0.983/0.995
	4	0.990/0.986/0.995	0.989/0.985/0.995	0.987/0.983/0.994	0.985/0.981/0.994	0.983/0.978/0.993
	5	0.987/0.983/0.994	0.986/0.981/0.993	0.984/0.979/0.993	0.981/0.976/0.992	0.978/0.973/0.992
	6	0.984/0.980/0.993	0.982/0.978/0.992	0.980/0.975/0.992	0.977/0.972/0.991	0.973/0.969/0.990
	7	0.981/0.977/0.992	0.979/0.974/0.991	0.976/0.972/0.990	0.972/0.969/0.989	0.967/0.965/0.989
	8	0.978/0.974/0.991	0.976/0.971/0.990	0.972/0.968/0.989	0.968/0.965/0.988	0.962/0.961/0.987
	9	0.975/0.971/0.989	0.972/0.968/0.989	0.968/0.965/0.988	0.963/0.962/0.987	0.956/0.958/0.986
	10	0.972/0.968/0.988	0.969/0.965/0.987	0.964/0.962/0.986	0.958/0.958/0.985	0.950/0.954/0.984

Table 14 Ratios of empirical MSEs for our first experimental set for $\alpha_i = 0.2 \cdot i^{-6/7}$. 200 replications to estimate empirical MSEs. Each entry reads $\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{FP}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{KW}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{RE}}}$.

n_{sim}	σ	$\lvert\theta\rvert^{1}$	$\lvert\theta\rvert^{1.1}$	$\lvert\theta\rvert^{1.2}$	$\lvert\theta\rvert^{1.3}$	$\lvert\theta\rvert^{1.4}$
100	1	1.000/0.997/0.998	1.000/0.997/0.998	1.000/0.997/0.998	1.000/0.996/0.998	1.000/0.996/0.997
	2	1.000/0.995/0.996	1.000/0.994/0.996	1.000/0.994/0.996	1.000/0.993/0.995	1.000/0.992/0.995
	3	1.000/0.992/0.995	1.000/0.991/0.994	0.999/0.991/0.994	0.999/0.990/0.993	0.999/0.988/0.993
	4	0.999/0.989/0.993	0.999/0.989/0.993	0.999/0.988/0.992	0.999/0.986/0.991	0.999/0.985/0.990
	5	0.999/0.987/0.991	0.999/0.986/0.991	0.999/0.985/0.990	0.999/0.984/0.989	0.999/0.982/0.988
	6	0.999/0.985/0.990	0.999/0.984/0.989	0.999/0.982/0.988	0.999/0.981/0.987	0.999/0.979/0.986
	7	0.999/0.983/0.988	0.999/0.981/0.987	0.999/0.980/0.986	0.999/0.978/0.985	0.998/0.976/0.984
	8	0.999/0.980/0.987	0.999/0.979/0.986	0.999/0.978/0.985	0.999/0.976/0.983	0.998/0.974/0.982
	9	0.999/0.978/0.985	0.999/0.977/0.984	0.999/0.975/0.983	0.999/0.974/0.981	0.998/0.971/0.980
	10	0.999/0.977/0.984	0.999/0.975/0.983	0.999/0.973/0.981	0.999/0.972/0.980	0.999/0.969/0.978
150	1	0.999/0.998/0.999	0.999/0.998/0.999	0.999/0.998/0.999	0.999/0.998/0.999	0.999/0.997/0.999
	2	0.998/0.997/0.999	0.998/0.996/0.999	0.998/0.996/0.998	0.998/0.995/0.998	0.998/0.995/0.998
	3	0.998/0.995/0.998	0.997/0.994/0.998	0.997/0.994/0.998	0.997/0.993/0.997	0.997/0.992/0.997
	4	0.997/0.993/0.997	0.996/0.992/0.997	0.996/0.992/0.997	0.995/0.991/0.996	0.995/0.990/0.996
	5	0.996/0.991/0.996	0.995/0.991/0.996	0.995/0.990/0.996	0.994/0.989/0.995	0.993/0.987/0.995
	6	0.994/0.990/0.996	0.994/0.989/0.995	0.993/0.988/0.995	0.992/0.987/0.994	0.991/0.985/0.994
	7	0.993/0.988/0.995	0.993/0.987/0.994	0.992/0.986/0.994	0.990/0.984/0.993	0.989/0.983/0.992
	8	0.992/0.987/0.994	0.991/0.985/0.993	0.990/0.984/0.993	0.988/0.982/0.992	0.986/0.980/0.991
	9	0.991/0.985/0.993	0.990/0.984/0.992	0.988/0.982/0.991	0.986/0.980/0.990	0.983/0.978/0.989
	10	0.989/0.984/0.992	0.988/0.982/0.991	0.986/0.981/0.990	0.984/0.979/0.989	0.981/0.976/0.988
200	1	0.998/0.997/0.999	0.998/0.996/0.999	0.997/0.996/0.999	0.997/0.995/0.998	0.997/0.994/0.998
	2	0.995/0.993/0.998	0.995/0.992/0.997	0.994/0.992/0.997	0.994/0.990/0.997	0.993/0.989/0.997
	3	0.993/0.990/0.996	0.992/0.989/0.996	0.991/0.988/0.996	0.990/0.986/0.995	0.989/0.984/0.995
	4	0.990/0.987/0.995	0.989/0.985/0.995	0.988/0.984/0.994	0.986/0.982/0.994	0.985/0.980/0.993
	5	0.988/0.983/0.994	0.986/0.982/0.994	0.985/0.980/0.993	0.983/0.978/0.992	0.980/0.975/0.992
	6	0.985/0.980/0.993	0.983/0.979/0.992	0.981/0.977/0.992	0.979/0.974/0.991	0.975/0.971/0.990
	7	0.982/0.977/0.992	0.980/0.976/0.991	0.978/0.973/0.990	0.975/0.970/0.990	0.970/0.967/0.989
	8	0.980/0.975/0.991	0.977/0.973/0.990	0.974/0.970/0.989	0.970/0.967/0.988	0.965/0.963/0.987
	9	0.977/0.972/0.990	0.974/0.970/0.989	0.971/0.967/0.988	0.966/0.964/0.987	0.960/0.960/0.986
	10	0.974/0.969/0.988	0.971/0.967/0.988	0.967/0.964/0.987	0.962/0.961/0.986	0.955/0.957/0.985

Table 15 Ratios of empirical MSEs for our first experimental set for $\alpha_i = 0.2 \cdot i^{-7/8}$. 200 replications to estimate empirical MSEs. Each entry reads $\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{FP}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{KW}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{RE}}}$.

n_{sim}	σ	$\lvert\theta\rvert^1$	$\lvert\theta\rvert^{1.1}$	$\lvert\theta\rvert^{1.2}$	$\lvert\theta\rvert^{1.3}$	$\lvert\theta\rvert^{1.4}$
100	1	1.000/0.997/0.998	1.000/0.997/0.998	1.000/0.997/0.998	1.000/0.997/0.998	1.000/0.996/0.998
	2	0.999/0.995/0.997	0.999/0.994/0.997	0.999/0.994/0.996	1.000/0.994/0.996	1.000/0.993/0.996
	3	0.999/0.992/0.995	0.999/0.992/0.995	0.999/0.991/0.995	0.999/0.990/0.994	0.999/0.990/0.994
	4	0.999/0.989/0.993	0.999/0.989/0.993	0.999/0.988/0.993	0.999/0.987/0.992	0.999/0.986/0.992
	5	0.999/0.987/0.992	0.999/0.986/0.991	0.999/0.985/0.991	0.999/0.984/0.990	0.999/0.983/0.990
	6	0.998/0.984/0.990	0.998/0.984/0.990	0.998/0.983/0.989	0.999/0.982/0.988	0.999/0.980/0.988
	7	0.998/0.982/0.988	0.998/0.981/0.988	0.998/0.980/0.987	0.998/0.979/0.986	0.998/0.977/0.986
	8	0.998/0.980/0.987	0.998/0.979/0.986	0.998/0.977/0.985	0.998/0.976/0.984	0.998/0.975/0.984
	9	0.998/0.977/0.985	0.998/0.976/0.984	0.998/0.975/0.983	0.998/0.973/0.983	0.998/0.972/0.982
	10	0.997/0.975/0.983	0.997/0.974/0.983	0.997/0.972/0.982	0.998/0.971/0.981	0.998/0.969/0.980
150	1	0.999/0.999/0.999	0.999/0.998/0.999	0.999/0.998/0.999	0.999/0.998/0.999	0.999/0.998/0.999
	2	0.999/0.997/0.999	0.999/0.997/0.998	0.999/0.997/0.998	0.999/0.996/0.998	0.999/0.996/0.998
	3	0.998/0.996/0.998	0.998/0.995/0.998	0.998/0.995/0.998	0.998/0.995/0.997	0.998/0.994/0.997
	4	0.997/0.994/0.997	0.997/0.994/0.997	0.997/0.994/0.997	0.997/0.993/0.996	0.997/0.993/0.996
	5	0.997/0.993/0.996	0.996/0.993/0.996	0.996/0.992/0.996	0.996/0.991/0.995	0.995/0.991/0.995
	6	0.996/0.992/0.995	0.995/0.991/0.995	0.995/0.991/0.995	0.995/0.990/0.994	0.994/0.989/0.994
	7	0.995/0.990/0.995	0.995/0.990/0.994	0.994/0.989/0.994	0.994/0.988/0.993	0.993/0.987/0.993
	8	0.994/0.989/0.994	0.994/0.988/0.993	0.993/0.988/0.993	0.992/0.987/0.992	0.992/0.986/0.992
	9	0.993/0.988/0.993	0.993/0.987/0.992	0.992/0.986/0.992	0.991/0.985/0.991	0.990/0.984/0.991
	10	0.992/0.986/0.992	0.992/0.986/0.991	0.991/0.985/0.991	0.990/0.984/0.990	0.989/0.983/0.990
200	1	0.998/0.997/0.999	0.998/0.997/0.999	0.998/0.997/0.999	0.998/0.996/0.999	0.998/0.996/0.999
	2	0.997/0.994/0.998	0.996/0.994/0.998	0.996/0.994/0.997	0.996/0.993/0.997	0.996/0.993/0.997
	3	0.995/0.992/0.997	0.994/0.991/0.996	0.994/0.991/0.996	0.994/0.990/0.996	0.993/0.989/0.996
	4	0.993/0.989/0.995	0.992/0.989/0.995	0.992/0.988/0.995	0.991/0.987/0.995	0.990/0.986/0.994
	5	0.991/0.987/0.994	0.990/0.986/0.994	0.990/0.985/0.994	0.989/0.984/0.993	0.988/0.982/0.993
	6	0.989/0.984/0.993	0.988/0.983/0.993	0.987/0.982/0.992	0.986/0.981/0.992	0.985/0.979/0.992
	7	0.987/0.982/0.992	0.986/0.981/0.992	0.985/0.979/0.991	0.984/0.978/0.991	0.982/0.976/0.990
	8	0.985/0.979/0.991	0.984/0.978/0.991	0.982/0.977/0.990	0.981/0.975/0.989	0.979/0.973/0.989
	9	0.983/0.977/0.990	0.982/0.976/0.989	0.980/0.974/0.989	0.978/0.973/0.988	0.976/0.971/0.987
	10	0.981/0.975/0.989	0.979/0.973/0.988	0.977/0.972/0.988	0.975/0.970/0.987	0.973/0.968/0.986

Table 16 Ratios of empirical MSEs for our first experimental set for $\alpha_i = 0.2 \cdot i^{-1}$. 200 replications to estimate empirical MSEs. Each entry reads $\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{FP}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{KW}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{RE}}}$.

n_{sim}	σ	$(2-\cos\theta)^{1}-1$	$(2-\cos\theta)^{1.1}-1$	$(2-\cos\theta)^{1.2}-1$	$(2-\cos\theta)^{1.3}-1$	$(2-\cos\theta)^{1.4}-1$
100	5	0.964/0.979/0.969	0.964/0.977/0.968	0.964/0.976/0.967	0.965/0.975/0.967	0.966/0.974/0.967
	6	0.956/0.976/0.962	0.955/0.974/0.961	0.955/0.973/0.960	0.956/0.972/0.960	0.956/0.971/0.960
	7	0.947/0.973/0.956	0.946/0.972/0.954	0.946/0.970/0.953	0.946/0.969/0.953	0.947/0.968/0.953
	8	0.938/0.970/0.949	0.937/0.969/0.948	0.936/0.968/0.946	0.937/0.967/0.946	0.937/0.966/0.946
	9	0.930/0.968/0.943	0.928/0.967/0.941	0.927/0.966/0.939	0.927/0.965/0.939	0.928/0.964/0.939
	10	0.921/0.966/0.936	0.919/0.965/0.934	0.917/0.964/0.932	0.917/0.963/0.931	0.918/0.962/0.932
150	5	0.988/0.968/0.998	0.990/0.966/0.998	0.992/0.965/0.998	0.993/0.963/0.998	0.995/0.962/0.999
	6	0.984/0.964/0.997	0.986/0.962/0.998	0.988/0.960/0.998	0.990/0.959/0.998	0.991/0.958/0.999
	7	0.980/0.959/0.997	0.982/0.957/0.997	0.984/0.955/0.998	0.986/0.954/0.998	0.988/0.953/0.998
	8	0.976/0.954/0.997	0.978/0.952/0.997	0.980/0.950/0.997	0.982/0.949/0.998	0.984/0.949/0.998
	9	0.972/0.950/0.997	0.974/0.948/0.997	0.975/0.946/0.997	0.978/0.945/0.998	0.980/0.944/0.998
	10	0.968/0.945/0.996	0.969/0.943/0.997	0.971/0.942/0.997	0.974/0.941/0.998	0.976/0.940/0.998
200	5	0.988/0.989/0.993	0.989/0.986/0.993	0.991/0.984/0.994	0.991/0.982/0.994	0.992/0.979/0.995
	6	0.984/0.988/0.992	0.985/0.986/0.992	0.986/0.984/0.992	0.987/0.981/0.993	0.987/0.979/0.994
	7	0.979/0.988/0.990	0.980/0.986/0.990	0.981/0.983/0.990	0.982/0.981/0.991	0.982/0.978/0.992
	8	0.975/0.987/0.988	0.975/0.985/0.988	0.976/0.983/0.989	0.977/0.980/0.990	0.977/0.978/0.991
	9	0.970/0.987/0.987	0.971/0.985/0.987	0.971/0.983/0.987	0.972/0.980/0.988	0.972/0.977/0.989
	10	0.965/0.987/0.985	0.966/0.985/0.985	0.966/0.983/0.985	0.967/0.980/0.986	0.967/0.977/0.988

Table 17 Ratios of empirical MSEs for our second experimental set for $\alpha_i = 0.1 \cdot i^{-6/7}$. 200 replications to estimate empirical MSEs. Each entry reads $\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{FP}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{KW}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{RE}}}$.

n_{sim}	σ	$(2-\cos\theta)^1-1$	$(2-\cos\theta)^{1.1}-1$	$(2-\cos\theta)^{1.2}-1$	$(2-\cos\theta)^{1.3}-1$	$(2-\cos\theta)^{1.4}-1$
100	5	0.965/0.979/0.970	0.965/0.978/0.969	0.965/0.976/0.968	0.966/0.975/0.968	0.967/0.974/0.968
	6	0.957/0.976/0.963	0.957/0.975/0.962	0.957/0.973/0.961	0.957/0.972/0.961	0.958/0.971/0.961
	7	0.949/0.973/0.957	0.948/0.972/0.956	0.948/0.971/0.955	0.948/0.970/0.954	0.949/0.969/0.954
	8	0.941/0.971/0.951	0.939/0.969/0.949	0.939/0.968/0.948	0.939/0.967/0.947	0.940/0.966/0.948
	9	0.932/0.968/0.945	0.930/0.967/0.943	0.930/0.966/0.941	0.929/0.965/0.940	0.930/0.964/0.941
	10	0.924/0.966/0.938	0.922/0.965/0.936	0.920/0.964/0.934	0.920/0.963/0.933	0.920/0.962/0.933
150	5	0.988/0.969/0.998	0.989/0.967/0.998	0.991/0.966/0.998	0.993/0.964/0.998	0.994/0.963/0.999
	6	0.984/0.965/0.997	0.985/0.963/0.997	0.987/0.961/0.998	0.989/0.960/0.998	0.991/0.959/0.998
	7	0.980/0.960/0.997	0.981/0.958/0.997	0.983/0.956/0.997	0.985/0.955/0.998	0.987/0.954/0.998
	8	0.976/0.956/0.996	0.977/0.953/0.997	0.979/0.952/0.997	0.981/0.950/0.998	0.983/0.950/0.998
	9	0.972/0.951/0.996	0.973/0.949/0.997	0.975/0.947/0.997	0.977/0.946/0.997	0.979/0.945/0.998
	10	0.968/0.947/0.996	0.969/0.945/0.996	0.971/0.943/0.997	0.973/0.942/0.997	0.976/0.941/0.998
200	5	0.989/0.989/0.993	0.990/0.987/0.993	0.991/0.985/0.993	0.992/0.982/0.994	0.992/0.980/0.995
	6	0.985/0.989/0.992	0.986/0.987/0.992	0.987/0.984/0.992	0.987/0.982/0.992	0.988/0.980/0.993
	7	0.980/0.988/0.990	0.981/0.986/0.990	0.982/0.984/0.990	0.983/0.981/0.991	0.983/0.979/0.992
	8	0.976/0.988/0.988	0.976/0.986/0.988	0.977/0.984/0.989	0.978/0.981/0.989	0.979/0.979/0.990
	9	0.971/0.988/0.987	0.972/0.986/0.987	0.973/0.984/0.987	0.973/0.981/0.988	0.974/0.978/0.989
	10	0.967/0.988/0.985	0.967/0.986/0.985	0.968/0.984/0.985	0.968/0.981/0.986	0.969/0.978/0.988

Table 18 Ratios of empirical MSEs for our second experimental set for $\alpha_i = 0.1 \cdot i^{-7/8}$. 200 replications to estimate empirical MSEs. Each entry reads $\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{FP}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{KW}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{RE}}}$.

n_{sim}	σ	$(2-\cos\theta)^1-1$	$(2-\cos\theta)^{1.1}-1$	$(2-\cos\theta)^{1.2}-1$	$(2-\cos\theta)^{1.3}-1$	$(2-\cos\theta)^{1.4}-1$
100	5	0.972/0.982/0.975	0.972/0.980/0.974	0.972/0.979/0.973	0.972/0.978/0.973	0.973/0.977/0.973
	6	0.966/0.979/0.970	0.965/0.978/0.969	0.965/0.976/0.968	0.965/0.975/0.967	0.966/0.974/0.967
	7	0.959/0.976/0.965	0.958/0.975/0.963	0.958/0.974/0.962	0.958/0.972/0.962	0.958/0.971/0.961
	8	0.953/0.974/0.959	0.952/0.972/0.958	0.951/0.971/0.957	0.951/0.970/0.956	0.951/0.969/0.955
	9	0.946/0.971/0.954	0.945/0.970/0.952	0.944/0.969/0.951	0.943/0.967/0.950	0.943/0.966/0.949
	10	0.940/0.969/0.949	0.938/0.968/0.947	0.937/0.966/0.945	0.936/0.965/0.944	0.936/0.964/0.943
150	5	0.987/0.975/0.997	0.987/0.973/0.997	0.988/0.971/0.997	0.990/0.970/0.997	0.991/0.968/0.997
	6	0.983/0.971/0.996	0.984/0.969/0.996	0.984/0.967/0.997	0.986/0.966/0.997	0.987/0.964/0.997
	7	0.979/0.967/0.996	0.980/0.965/0.996	0.981/0.964/0.996	0.982/0.962/0.996	0.983/0.960/0.996
	8	0.976/0.964/0.995	0.976/0.962/0.995	0.977/0.960/0.996	0.978/0.958/0.996	0.979/0.956/0.996
	9	0.972/0.960/0.995	0.972/0.958/0.995	0.973/0.956/0.995	0.974/0.954/0.995	0.976/0.953/0.996
	10	0.968/0.956/0.994	0.968/0.954/0.995	0.969/0.952/0.995	0.970/0.951/0.995	0.972/0.949/0.995
200	5	0.991/0.992/0.993	0.992/0.990/0.993	0.993/0.988/0.993	0.994/0.987/0.993	0.995/0.985/0.993
	6	0.988/0.992/0.991	0.989/0.990/0.991	0.989/0.988/0.991	0.990/0.986/0.991	0.991/0.984/0.992
	7	0.985/0.991/0.990	0.985/0.990/0.990	0.986/0.988/0.990	0.987/0.986/0.990	0.988/0.984/0.990
	8	0.982/0.991/0.988	0.982/0.989/0.988	0.983/0.988/0.988	0.983/0.986/0.988	0.984/0.984/0.989
	9	0.978/0.991/0.987	0.979/0.989/0.987	0.979/0.988/0.987	0.980/0.986/0.987	0.981/0.984/0.987
	10	0.975/0.991/0.985	0.975/0.989/0.985	0.975/0.987/0.985	0.976/0.986/0.985	0.977/0.984/0.986

Table 19 Ratios of empirical MSEs for our second experimental set for $\alpha_i = 0.1 \cdot i^{-1}$. 200 replications to estimate empirical MSEs. Each entry reads $\frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{FP}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{KW}}} / \frac{\text{MSE}_{\text{OW}}}{\text{MSE}_{\text{RE}}}$.