Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

CONTROL VARIATES BEYOND MEAN: VARIANCE REDUCTION FOR NONLINEAR STATISTICAL QUANTITIES

Henry Lam[1], and Zitong Wang[1]
[1]Industrial Eng. and Operations Research, Columbia University
New York, USA

ABSTRACT

Control variate (CV) is a powerful Monte Carlo variance reduction technique by injecting mean information about an auxiliary related variable to the simulation output. Partly due to this way of leveraging information, CV has been mostly studied in the context of mean estimation. In this paper, we study CV for general nonlinear quantities such as conditional value-at-risk and stochastic optimization. While we can extend the tools from mean estimation, a challenge in nonlinear generalizations is the proper calibration of the CV coefficient, which deviates from standard linear regression estimators and requires influence function or resampling. As a key contribution, we offer a general methodology that bypasses this challenge by harnessing a weighted representation of CV and interchanging weights between the empirical distribution and the nonlinear functional. We provide theoretical results in the form of central limit theorems to illustrate our performance gains and numerically demonstrate them with several experiments.

1 INTRODUCTION

CV is a powerful variance reduction method in Monte Carlo estimation (Yang and Nelson 1989; L'Ecuyer and Buist 2006; Glynn and Whitt 1989; Sun et al. 2023; Kim and Henderson 2007). Its main idea is to utilize auxiliary variables that are correlated with the target simulation outputs, and by suitably combining the auxiliary and target outputs, we obtain estimators that have smaller variance than naive Monte Carlo that uses the target outputs alone. More specifically, when running simulation, we can often obtain some side covariates or detailed dynamics of the simulated trajectories, and the CV method leverages our information on these auxiliary outputs to enhance estimation efficiency. This approach, while attractively simple and intuitive, has limitations nonetheless. It utilizes a linear-regression-based formulation and relatedly requires the mean information of the auxiliary variables. As such, its main applicability has been confined largely to mean estimation problems. On the other hand, many simulation tasks concern nonlinear statistical quantities, such as quantiles, conditional value-at-risk (CVaR), and stochastic optimization problems. Specifically, the map from the underlying distribution to the quantity of interest is nonlinear in these cases. However, existing CV frameworks, while efficient in mean estimations, have not been adopted to more general nonlinear functional settings.

This work studies a general approach to construct a CV estimator for nonlinear statistical quantities. To give a sense of this problem, note that many nonlinear problems, even though they are not exactly linear, are locally linearizable in the sense of being approximatable by a mean as sample size grows. The idea resembles common Taylor series, in which the gradients in our context are captured via the so-called influence function of the statistical quantity. From this view, it appears that we can borrow tools from linear CV in the locally linearizable setting. While this is indeed the case conceptually, a core challenge arises in calibrating the optimal coefficient in the CV formula. This latter coefficient bridges the CV with the target simulation output, and needs to be properly selected via the simulation data. In the case of mean estimation, the optimal coefficient can be calibrated readily via a linear regression calculation. In the nonlinear setting, however, the optimal coefficient requires the influence function, which needs to be

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

either calculated in closed form, or requires resampling approaches such as the bootstrap that substantially increases the computation overhead.

In view of the above challenge, our main contribution is to build a methodology to construct CV estimators for nonlinear quantities that bypass the bottleneck with respect to both influence function knowledge and resampling need. Our key idea leverages a weighted representation of the classical CV estimator, to our knowledge originated in Hesterberg (1996) and Hesterberg and Nelson (1998) in the context of quantile estimation. These works consider quantile estimation via inverting the cumulative distribution function (cdf). They apply CV on the cdf as the estimation target before taking its inverse and, since the cdf is a function, they observe that an efficient approach to apply CV across the function input values is to reformulate the CV estimator as a weighted Monte Carlo average, where the weights depend only on the CV and hence avoid recomputing a new CV estimator at each input value of the cdf. Their works essentially transform the quantile problem into a mean estimation problem where classical CV can be applied. Our main idea in this paper is inspired from this approach, but we make it substantially more general in the framework of local linearization. In particular, we show that, thanks to linearization, the weights can be interchanged between the nonlinear functional and the empirical distribution, so that by simply evaluating the nonlinear functional on a weighted empirical distribution, we achieve the same effect as a CV estimator with a nearly optimally calibrated coefficient, without the aforementioned overhead. Importantly, this approach does not apply only to the inversion operation on the cdf, but any "smooth" functional on the cdf.

In terms of theoretical results, we establish central limit theorems (CLTs) for our CV estimator, and demonstrate our ability to obtain asymptotically optimal coefficient without the computation overhead. We showcase our approach in estimating risk measures and solutions and optimal values of stochastic optimization problems. Moreover, in the latter setting, we further develop an asymptotic characterization of the optimality gap to illustrate our efficiency gain in terms of the optimality of the estimated solution via our CV. We illustrate our effectiveness through several canonical examples, including newsvendor problems and linear regression with correlated input features, showing that our CV estimators achieve substantial variance reduction particularly in low signal-to-noise settings.

1.1 Related Works

Our work is related to several aspects of CV. First regards the calibration of the coefficient in the CV formula, which is rather limited to our best knowledge. This is potentially attributed to the fact that, for mean estimation, this calibration can be readily conducted using direct empirical plug-in (Glasserman 2004; Asmussen and Glynn 2007). Beyond this, our conceptual foundation comes mainly from Hesterberg (1996), who studied CV and importance sampling when bootstrapping nonlinear statistics, and the work Hesterberg and Nelson (1998) shortly after. In particular, these works observed that a CV estimator for mean estimation using the plug-in estimate of the optimal coefficient can be written in the form of a weighted estimator, where the weights depend only on the CV and not the target statistic itself. From this observation, their approaches first convert nonlinear problems into estimating a cdf and apply CV on the latter as a mean estimation. Our approach further leverages their idea to combine with local linearization, and importantly, uses it to bypass the knowledge influence function and computation overhead introduced by resampling approaches.

Our work is related to, but also should be significantly contrasted with, Glynn and Whitt (1989) that investigated nonlinear CV. The nonlinearity that they address regards how to combine target simulation outputs with CVs, i.e., nonlinearly versus linearly combining them. This nonlinearity is different from ours, which is in the target quantity and concerns its relation with the underlying probability distribution. In terms of insights, Glynn and Whitt (1989) proved that nonlinear CV offers no asymptotic advantage over linear ones under suitable regularity conditions. This result supports the general choice to focus on linear CVs, even in the context of nonlinear statistics that we consider.

A significant line of CV research focuses on the construction of good CVs using auxiliary information. Kim and Henderson (2007) considered a parameterized family of CVs, and developed an optimization problem to search for good candidates. Tsai et al. (2023) developed a framework that adaptively selects CVs in ranking-and-selection problems. Glynn and Whitt (1989) developed CVs for queueing problems by leveraging Little's law, which they call an indirect estimator. Viewing their indirect estimator as a nonlinear CV, they further prove that nonlinear CVs are asymptotically equivalent to linear ones in terms of efficiency under standard assumptions. In the Markov Chain setting, Henderson and Glynn (2002) constructed CV by approximating martingales, which arise from the solution to Poisson's equation. Building on this idea, Henderson and Simon (2004) further developed an adaptive scheme that achieves better CV as the simulation proceeds. More recently, machine learning-based methods have been developed for constructing CVs, including neural, regression-based, and kernel-based approaches (Blanchet et al. 2023; Müller et al. 2020; Oates et al. 2017; Liu et al. 2017; Portier and Segers 2019). These works have different focuses from our goal to study the application of CV to nonlinear statistical quantities, but their ways of constructing good CVs would likely continue to be applicable to our nonlinear settings.

CVs have also been used in stochastic optimization algorithms (Johnson and Zhang 2013; Reddi et al. 2016; Wang et al. 2013; Fang et al. 2018). The seminal stochastic variance-reduced method, proposed by Johnson and Zhang (2013), uses periodically computed full-batch gradients as CVs to reduce the variance of gradient estimation and thus accelerate convergence. Wang et al. (2013) introduced a CV approach based on estimated low-order moments to reduce the variance of stochastic gradients, improving convergence speed and stability in both convex and non-convex settings. However, these methods primarily focus on reducing the variance of gradient estimators, which are still targeting the mean of the underlying distribution. In particular, while our work addresses variance reduction for stochastic optimization as an example, we view the target objective as a nonlinear functional that is fundamentally different from these works.

Finally, recent developments in prescriptive analytics or contextual optimization also make use of auxiliary covariates for improving decision-making under uncertainty (Elmachtoub and Grigas 2022; Elmachtoub et al. 2023). Bertsimas and Kallus (2020) proposed a general framework for learning data-driven decision rules that minimize expected cost conditional on observed features, bridging predictive modeling and optimization. Srivastava et al. (2021) studied contextual stochastic optimization problems where side information is used to inform decisions, proposing a regularized Nadaraya-Watson approach to estimate conditional expectations and optimize decisions accordingly. While our work similarly leverages auxiliary information, our focus is on variance reduction rather than better learning approaches. These two ideas can be naturally combined, as we will illustrate in our contextual optimization example.

2 CONTROL VARIATES ESTIMATOR FOR MEAN ESTIMATION

Suppose we aim to estimate the statistical functional $\varphi(P)$ of a random variable Y with unknown distribution P, where an auxiliary random variable X can be generated alongside. We assume that the expectation μ_X of X is known. To begin with, we first consider $\varphi(P)$ as the mean of Y, i.e., $\varphi(P) = \mathbb{E}[Y]$. For a generic random variable U, when $U_1,\dots,U_n$ are *i.i.d.* samples from its distribution, we denote the sample mean by $\bar{U}$.

Given *i.i.d.* observations $\{(Y_i,X_i)\}_{i=1}^n$, let P_n denote the corresponding empirical marginal distribution of Y. The vanilla Monte Carlo estimator for $\varphi(P)$ is

$$\varphi(P_n) = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

The CV estimators exploit the information contained in X to construct an unbiased estimator with reduced variance compared to the vanilla Monte Carlo estimator. For clarity, we call X the CV and, for a fixed $\beta \in \mathbb{R}$, the CV estimator associated with β is defined as

$$\varphi_\beta(P_n) = \bar{Y} - \beta(\bar{X} - \mu_X). \tag{1}$$

Note that we have abused notation slightly above (and throughout the rest of this paper too) that $\varphi_\beta(P_n)$ depends also on $\bar{X}$, which is suppressed to lighten notation when there is no confusion. A straightforward calculation deduces that the optimal choice β^* minimizing the variance of $\varphi_\beta(P_n)$ is

$$\beta^* = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}.$$

This β^* can be viewed as the idealized coefficient in the linear regression of Y against $X-\mu_X$, and the CV estimator outputs precisely the intercept. With β^*, we obtain the optimal variance reduction ratio over the vanilla Monte Carlo estimator

$$\mathrm{Var}(\varphi_{\beta^*}(P_n))/\mathrm{Var}(\varphi(P_n)) = 1-\rho^2, \quad \rho = \mathrm{Corr}(X,Y) \tag{2}$$

That is, when X and Y are more correlated (either positively or negatively), more variance can be reduced. In the extreme case where X and Y are perfectly aligned, the CV estimator completely eliminates the randomness and outputs μ_Y. In practice, β^* is typically unknown, and an empirical approximation $\hat{\beta}$ is used:

$$\hat{\beta} = \frac{\sum_i (X_i-\bar{X})(Y_i-\bar{Y})}{\sum_j (X_j-\bar{X})^2},$$

which achieves asymptotically the same variance reduction ratio as in Eq. (2).

We also point out that using X directly as the CV may not always be efficient. By Eq. (2), the variance reduction ratio depends on the correlation coefficient between data Y and the CV. When the dependence between X and Y is nonlinear, a naive application of X as the CV for $\mathbb{E}[Y]$ may be ineffective. For example, suppose we want to estimate the mean of $Y=\frac{1}{1+X^2}$, where X follows a uniform distribution between $[0,1]$. Although Y is fully determined by X, the correlation $\mathrm{Corr}(X,Y)$ is zero, leading to no variance reduction by Eq. (2). However, if we instead use $f(X)=X^2$ as the CV, the optimal variance reduction rate becomes approximately 70%, which is a significant reduction. This observation motivates the use of a transformation $f(X)$ as the CV instead of X itself. By Eq. (2), the optimal choice of f is the function of X that maximizes ρ. Specifically, the optimal CV function $f^*(X)$ is $f^*(X)=\mathbb{E}[Y|X]$. However, this optimal choice is often intractable in practice. There is a broad line of works discussing how to construct good CVs. When the distribution of X is known, one can construct good approximations to $\mathbb{E}[Y|X]$ using regression techniques, basis expansions (e.g., polynomials, splines), or other methods that exploit known structure in the problem. The problem has been widely studied (e.g., see Kim and Henderson 2007; Tsai et al. 2023; Glynn and Whitt 1989; Blanchet et al. 2023; Müller et al. 2020; Oates et al. 2017; Liu et al. 2017; Portier and Segers 2019), but as our focus is on the use of CVs for nonlinear statistics, we omit further discussion of these strategies.

We denote the CV estimator using the CV function f and parameter β by

$$\varphi_{\beta,f}(P_n) = \bar{Y} - \beta(\bar{f}-\mu_f), \text{ where } \mu_f = \mathbb{E}[f(X)].$$

This notation will be used when we establish results for CV for general statistics.

Finally, note that we have contained our discussion in the case where we only use one CV. Our discussion and approach extend naturally to using multiple CVs, but this extension is conceptually straightforward and for simplicity we focus on the univariate case in this paper.

3 CHALLENGES AND REMEDY IN NONLINEAR GENERALIZATIONS

Suppose now the target quantity $\varphi(P)$ is nonlinear in P, such as a quantile, a risk measure, or a solution or optimal value to a stochastic optimization problem. A natural application of CVs suggests modifying the estimator $\varphi(P_n)$ by subtracting a linear combination of the CVs:

$$\varphi_{\beta,f}(P_n) = \varphi(P_n) - \beta(\bar{f}-\mu_f), \tag{3}$$

where the CV $f(X)$ is pre-specified and β is the constant coefficient as in the mean estimation case.

While the above appears similar to the usage of CV in mean estimation, we consider and argue that the calibration of the optimal coefficient β needs more substantial handling. For mean estimation, the optimal coefficient is given by the linear regression coefficient, which primarily depends on the covariance between Y and $f(X)$ and is readily estimable by a plug-in estimator. In contrast, for a general nonlinear φ, the optimal β^* depends on the covariance between the influence function $IF_\varphi(Y;P)$ and $f(X)$. Specifically,

$$\beta^* = \mathrm{Var}(f(X))^{-1}\mathrm{Cov}(IF_\varphi(Y;P), f(X)). \tag{4}$$

Here, the influence function can be defined in terms of the Gateaux derivative of φ with respect to P (Hampel et al. 2011). This leads to a challenge because, for many problems, the influence function $IF_\varphi(Y;P)$ can be difficult to compute or estimate. In such cases, we may consider resampling methods like the bootstrap, and using the fact that $\mathrm{Cov}(\varphi(P), f(X)) \approx \mathrm{Cov}(IF_\varphi(Y;P), f(X))/n$, but these methods can introduce substantial computational overhead and additional noises due to the Monte Carlo runs in the resampling.

To overcome this bottleneck, we take inspiration from the weighted Monte Carlo view of CVs, to our knowledge first studied in Hesterberg (1996) and Hesterberg and Nelson (1998). Rather than relying on direct calibration of β that depends on a covariance with the influence function, we exploit the fact that classical CV estimators can be reformulated as weighted averages. In the case of mean estimation, such a weighted average is exactly equivalent to the original CV formula in Eq. (1). In the nonlinear case, however, note that the target quantity is not a mean to begin with, and there is no exact definition of weighting. Nonetheless, many nonlinear quantities are locally linearizable, so that we can indeed put weights on the linearized target. However, this local linearization would contain the influence function which gives rise to its appearance in the optimal β^* in Eq. (4). Our key insight is that, by weighting the empirical distribution first and then evaluating φ on it, we achieve an estimator that well approximates the weighted local linearization, and in turn Eq. (3) with an optimally calibrated coefficient. Importantly, these weights depend only on the auxiliary variable X but not the influence function, thus effectively bypassing the aforementioned bottleneck.

To describe our approach concretely, we construct a weighted empirical distribution

$$\tilde{P}_n(\cdot) = (1/n)\sum_{i=1}^{n} W_i I(Y_i \in \cdot), \text{ where } W_i = \frac{1}{n} + \frac{(\bar{f} - f(X_i))(\bar{f} - \mu_f)}{\sum_j (f(X_j) - \bar{f})^2}. \tag{5}$$

The weights W_i depend solely on f and not on Y, making them accessible even when the target statistic depends on Y in a more complicated way. Our new CV estimator for nonlinear target $\varphi(P)$ is

$$\varphi_{CV}(P_n) = \varphi(\tilde{P}_n). \tag{6}$$

Note that computing $\varphi_{CV}(P_n)$ does not require influence function nor resampling approaches. Instead, we only need to evaluate φ using a weighted empirical distribution.

We note that the weights W_i in Eq. (5) are not guaranteed to be nonnegative in finite sample, though they are nonnegative asymptotically. The negative weights in finite sample could create computational challenges sometimes. For example, in the stochastic optimization setting, as we will see later, negative weights lead to a nonconvex problem. Therefore, in practice, a modification to enforce nonnegativity might be necessary.

We provide an intuitive explanation of our CV estimator. First, in the mean estimation case, the standard CV estimator associated with $\hat{\beta}$ can be written as

$$\begin{aligned}
\varphi_{\hat{\beta},f}(P_n) &= \varphi(P_n) - \hat{\beta}(\bar{f}-\mu_f) \\
&= \bar{Y} - \left(\frac{\sum_{i=1}^n (f(X_i)-\bar{f})(Y_i-\bar{Y})}{\sum_{j=1}^n (f(X_j)-\bar{f})^2}\right)(\bar{f}-\mu_f) \\
&= \bar{Y} - \sum_{i=1}^n \left(\frac{(f(X_i)-\bar{f})(\bar{f}-\mu_f)}{\sum_{j=1}^n (f(X_j)-\bar{f})^2}\right)(Y_i-\bar{Y}) \\
&= \varphi(\tilde{P}_n).
\end{aligned} \tag{7}$$

This motivates our definition for φ_{CV} as in Eq. (6). Nonetheless, when φ is nonlinear, Eq. (7) no longer holds. However, if φ is Hadamard differentiable at P, we have

$$\varphi_{\hat{\beta},f}(P_n) = \varphi(P) + \frac{1}{n}\sum_{i=1}^n W_i IF_\varphi(Y_i;P) + Rem(P_n - P),$$

$$\varphi_{CV}(P_n) = \varphi(\tilde{P}_n) = \varphi(P) + \frac{1}{n}\sum_{i=1}^n W_i IF_\varphi(Y_i;P) + Rem(\tilde{P}_n - P).$$

where $IF_\varphi(\cdot;P)$ is the influence function. $Rem(P_n - P)$ and $Rem(\tilde{P}_n - P)$ are second-order remainder terms. As a result, although no longer equal, $\varphi_{CV}(P_n)$ is still close to the optimal CV $\varphi_{\hat{\beta},f}(P_n)$.

4 THEORETICAL GUARANTEES FOR OUR CONTROL VARIATE ESTIMATOR

We provide formal justifications to our CV estimator for nonlinear statistical quantities. Our main result is a central limit theorem of φ_{CV}, indicating that it achieves the optimal variance reduction ratio attainable for the given f.

Theorem 1 Suppose that φ is Hadamard differentiable at P and that the class of influence functions $\{IF_\varphi(Y;P)\}$ is Donsker. Then,

$$\sqrt{n}(\varphi_{CV}(P_n) - \varphi(P)) \xrightarrow{d} \mathcal{N}(0, \sigma_{CV}^2), \tag{8}$$

where $\sigma_{CV}^2 = (1 - \mathrm{Corr}(IF_\varphi(Y;P), f(X))^2)\sigma^2$, and $\sigma^2 = \mathrm{Var}(IF_\varphi(Y;P))$ is the asymptotic variance of the vanilla Monte Carlo estimator.

We briefly clarify key concepts involved in Theorem 1. Hadamard differentiability ensures that the function φ admits a linear approximation locally. The influence function $\{IF_\varphi(Y;P)\}$ serves as the functional gradient of φ evaluated at P, measuring the local sensitivity of φ with respect to changes of P along the direction of a point mass at Y. The Donsker condition guarantees that the empirical process formed by the influence functions converges weakly to a Gaussian process, which enables asymptotic normality of the estimator. These assumptions are standard. For background, see (Van der Vaart and Wellner 1996) and (Kosorok 2008).

Theorem 1 implies that our CV estimator gives rise to an asymptotic variance that is equal to the one when coefficient β is chosen optimally, and thus revealing that we assimilate the auxiliary information in f efficiently. To establish Theorem 1, we first consider the general CV for nonlinear statistical quantities in Eq. (3) with a fixed, possibly suboptimal parameter β.

Lemma 1. *Let β be a fixed CV parameter. Under the same assumption as in Theorem 1, the following holds*

$$\sqrt{n}(\varphi_{\beta,f}(P_n) - \varphi(P)) \xrightarrow{d} \mathcal{N}(0, \sigma_\beta^2), \tag{9}$$

where the asymptotic variance is

$$\sigma_\beta^2 = \mathrm{Var}(IF_\varphi(Y;P)) - 2\beta\mathrm{Cov}(IF_\varphi(Y;P), f(X)) + \beta^2\mathrm{Var}(f(X)). \tag{10}$$

The optimal CV coefficient minimizing σ_β^2 is

$$\beta^* = \mathrm{Var}(f(X))^{-1}\mathrm{Cov}(IF_\varphi(Y;P), f(X)). \tag{11}$$

The corresponding minimal variance for the given f satisfies

$$0 < \sigma_{\beta^*}^2 = (1 - \mathrm{Corr}(IF_\varphi(Y;P), f(X))^2)\sigma^2 \leq \sigma^2, \tag{12}$$

where $\sigma^2 = \mathrm{Var}(IF_\varphi(Y;P))$.

The preceding lemma presents the relation of the asymptotic variance of CV estimator in Eq. (3) in relation to the coefficient β. In particular, when β is well-calibrated, the CV estimator achieves the highest possible asymptotic variance reduction ratio for the given CV function f. The following lemma further formalizes this result for estimators that statistically consistently estimate β^*.

Lemma 2. *Let β_n be any consistent estimator of β^* such that $\sqrt{n}(\beta_n - \beta^*) = o_P(1)$. Then*

$$\sqrt{n}\left(\varphi_{\beta_n,f}(P_n) - \varphi(P)\right) \xrightarrow{d} \mathscr{N}(0, \sigma_{\beta^*}^2), \tag{13}$$

where $\sigma_{\beta^*}^2$, defined in Eq. (12)*, is the minimal asymptotic variance achieved by using the optimal β^*.*

By Lemma 2, estimating β^* does not affect the asymptotic efficiency of the CV estimator. In particular, $\varphi_{\hat{\beta},f}(P_n)$ gives the optimal variance reduction ratio. Importantly, since the difference between $\varphi_{\hat{\beta},f}(P_n)$ and $\varphi_{CV}(P_n)$ are asymptotically negligible, Theorem 1 follows directly from Lemma 2.

5 FURTHER DISCUSSIONS

5.1 Misspecified Control Variates

Suppose we only know an approximation $\mu_f + \delta_f$ of μ_f, where δ_f is deterministic or random. Given δ_f, the optimal CV coefficient is given through the following least-squares problem:

$$\begin{aligned}\hat{\beta}_\delta &:= \arg\min_\beta \mathbb{E}\left[\left((\bar{f} - \mu_f - \delta_f)\beta - (\varphi(P_n) - \varphi(P))\right)^2\right] \\ &= \left(\mathrm{Var}(\bar{f}) + \delta_f^2\right)^{-1}\left(\mathrm{Cov}(\bar{f}, \varphi(P_n)) - \delta_f\mathbb{E}[\varphi(P_n) - \varphi(P)]\right).\end{aligned}$$

If δ_f is random and potentially correlated with $\bar{f}$, then similar calculations yield:

$$\hat{\beta}_\delta = M^{-1}\mathrm{Var}(\bar{f})\beta^* - M^{-1}\mathbb{E}[(\varphi(P_n) - \varphi(P))\delta_f], \quad M = \mathrm{Var}(\bar{f}) - 2\mathrm{Cov}(\bar{f}, \delta_f) + \mathbb{E}[\delta_f^2] \tag{14}$$

Eq. (14) shows that when the CV mean is misspecified by an amount δ_f, the estimated coefficient $\hat{\beta}_\delta$ deviates from the optimal β^*. In particular, the first term scales β^* by a factor depending on the relative size of the misspecification variance δ_f^2 compared to the variance of $\bar{f}$. When δ_f^2 is small relative to $\mathrm{Var}(\bar{f})$, the effect of the first term is negligible, and $\hat{\beta}_\delta$ remains close to β^*. The second term introduces an additional bias that depends on the correlation between δ_f and $\varphi(P_n)$. Thus, small or uncorrelated misspecifications in μ_f have limited impact on the CV effectiveness, but large or systematically biased misspecifications can downgrade the performance.

5.2 Sample Average Approximation and Optimality Gap

A prime example of nonlinear statistical quantity is stochastic optimization, which encompasses a range of statistics of interest. For example, quantile estimation can be viewed as a special case of using sample average approximation (SAA) as an estimator of a solution to a stochastic optimization problem. Consider

$$\varphi(P) = \arg\min_{\theta} \mathbb{E}_{X\sim P}[h(X,\theta)],$$

where h is a loss function. In the optimization context, a natural performance measure of a given solution is the optimality gap:

$$\mathscr{G}(\theta) = \mathbb{E}_{X\sim P}[h(X,\theta)] - \mathbb{E}_{X\sim P}[h(X,\theta^*)].$$

Assume that $H(\theta) = \mathbb{E}[h(X,\theta)]$ is twice continuously differentiable, we have the following expansion

$$\mathscr{G}(\theta) = \frac{1}{2}(\theta-\theta^*)^2\nabla^2 H(\theta^*) + o(|\theta-\theta^*|^2). \tag{15}$$

From Eq. (15), by the CLT for $\varphi_{CV}(P_n)$, we can derive the following:

Corollary 2 Assume that $\hat{\theta}_n = \varphi_{CV}(P_n)$ is consistent and satisfies $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathscr{N}(0,\sigma_{CV})$. Suppose that $H(\theta)$ is twice continuously differentiable at θ^*. Then the scaled optimality gap satisfies:

$$n\mathscr{G}(\hat{\theta}_n) \xrightarrow{d} \frac{1}{2}U^2\nabla^2 H(\theta^*), \text{ where } U \sim \mathscr{N}(0,\sigma_{CV}). \tag{16}$$

6 EXAMPLES

6.1 Contextual Optimization

In contextual optimization, the decision-maker aims to select an optimal action or policy based on contextual information (features or covariates) observed prior to making decisions. Formally, the goal is to minimize a cost that explicitly depends on context X, typically represented as optimizing an objective of the form $\mathbb{E}[h(Y,\theta(X))]$, where $\theta(X)$ is the decision policy depending on the observed context X and Y is the uncertainty. For an observed context x, the estimator of $\theta(x)$ can be estimated by solving

$$\hat{\theta}_n(x) \in \arg\min_{z\in Z} \sum_{i=1}^{n} w_i(x;\{X_j\}_{j=1}^n) h(Y_i,z),$$

where $w_i(x;\{X_j\}_{j=1}^n)$ are data-driven weights depending on the observed covariates $\{X_j\}_{j=1}^n$. The above formulation for $\hat{\theta}_n(x)$ can also be viewed as SAA by considering $g(z,x;\{X_j\}_{j=1}^n,Y_i) = nw_i(x;\{X_j\}_{j=1}^n)h(Y_i,z)$. Let f be the CV function specified in advance. We can apply CV by considering

$$\hat{\theta}_n^{CV}(x) \in \arg\min_{z\in Z} \sum_{i=1}^{n} w_i^{CV}(\{X_j\}_{j=1}^n) g(z,x;\{X_j\}_{j=1}^n,Y_i),$$

where $w_i^{CV}(\{X_j\}_{j=1}^n) = \frac{1}{n} + \frac{(\bar{f}-f(X_i))(\bar{f}-\mu_f)}{\sum_j(\bar{f}-f(X_j))^2}$ is obtained via our discussion in the previous sections.

6.2 Linear Regression

Linear regression defines a nonlinear statistic when viewed as a mapping from the data distribution to the estimated coefficient. Specifically, suppose we want to regress $Y \in \mathbb{R}$ against $X \in \mathbb{R}^p$. Also, suppose there is another random vector $C \in \mathbb{R}^q$ that can be sampled along with X and is correlated with X and Y. We consider two ways of regression for Y. One is standard linear regression using either X or (X,C) as predictors. The other is our CV approach, where we regress Y on X but reweight observations based on C. The two approaches are good under different scenarios. Regressing Y against (X,C) requires X and C to be as orthogonal as possible, while the CV approach requires C to be aligned with the influence function of mean squared loss evaluated at β^*.

Suppose we observe *i.i.d.* samples $\{X_i,Y_i,C_i\}$. Let $\mathbf{X}\in\mathbb{R}^{n\times p}$ and $\mathbf{C}\in\mathbb{R}^{n\times q}$ be matrices with rows X_i and C_i respectively. And let $\mathbf{Y}\in\mathbb{R}^n$ be the vector consisting of Y_i. In this case, the CV solves the following reweighted SAA:

$$\min_{\theta}\sum_{i=1}^{n} W_i(X_i^\top\theta - Y_i)^2, \tag{17}$$

where

$$W_i = \frac{1}{n} + (\bar{C}-C_i)^\top(\sum_j (C_j-\bar{C})(C_j-\bar{C})^\top)^{-1}(\bar{C}-\mu_C)\in\mathbb{R}.$$

As a result, if W_i's are positive, Eq. (17) has a unique minimizer

$$\hat{\theta}_n^{cv} = (\mathbf{X}^\top\Lambda_w\mathbf{X})^{-1}\mathbf{X}^\top\Lambda_w\mathbf{Y}, \quad \Lambda_w = diag([W_1,\ldots,W_n]).$$

In contrast, regressing Y against generic data matrix $\mathbf{D}$ gives

$$\hat{\theta}_n = (\mathbf{D}^\top\mathbf{D})^{-1}\mathbf{D}^\top\mathbf{Y},$$

where this $\mathbf{D}$ can be $\mathbf{X}$ or $[\mathbf{X}^\top,\mathbf{C}^\top]^\top$. Later, we will see our CV approach outperform standard linear regression under certain situations.

7 EXPERIMENTS

7.1 Newsvendor Problem

We consider the classical one-dimensional newsvendor setting, where the decision variable $\theta\in\mathbb{R}_+$ represents the order quantity and Y is a random demand. The cost function is given by

$$h(Y,\theta) = p_1\theta + p_2(Y-\theta)^+,$$

where $(a)^+ = \max(a,0)$ and $p_1 < p_2$. The first term models the linear ordering cost, while the second term penalizes unmet demand at a higher unit penalty p_2. The goal is to minimize the expected cost $\mathbb{E}[h(Y,\theta)]$ over θ. This optimization problem is equivalent to a quantile estimation task, where the optimal solution corresponds to the $\frac{p_2-p_1}{p_2}$ quantile of the distribution of Y. The distribution of demand Y often depends on observable covariates X, such as weather, time, or external signals (e.g., news). This dependence turns the newsvendor problem into a contextual optimization problem, where the optimal order quantity θ varies with the covariates.

Specifically, in this experiment, we set the unit ordering cost $p_1 = 2$ and unit shortage penalty $p_2 = 3$. We generate $n = 100$ data points, where each sample consists of a covariate $X\sim\text{Poisson}(\lambda_x)$ with known mean $\lambda_x = 1$, and a demand variable $Y\sim\text{Poisson}(\lambda_y)$ with (unknown) mean $\lambda_y = 2$. The CV X is constructed to be correlated with Y using a Gaussian Copula (Durante and Sempi 2010) with correlations chosen from the set {0.5, 0.65, 0.8, 0.9, 0.99}. For each correlation level, we repeat the estimation procedure $N_{\text{repeat}} = 500$ times to evaluate the bias and variance of the estimators.

Figure 1 compares the performance of CV estimators in both non-contextual and contextual stochastic optimization settings. In the left panel, we compare the standard SAA estimator with our CV estimator under varying levels of correlation between the target variable and the CV. The right panel shows a similar comparison under a nonparametric regression (NW) framework. In both cases, the CV estimators achieve lower bias and variance compared to their baselines, with the improvement becoming more pronounced as the correlation increases.

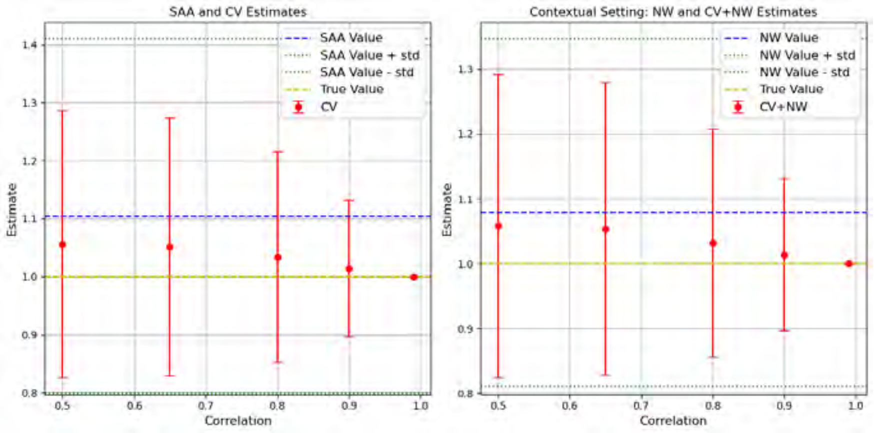

Figure 1: Newsvendor Problem. Comparison of estimation results with and without CVs under varying correlations. The left panel corresponds to the non-contextual setting, and the right panel shows the contextual stochastic optimization settings. Red dots denote the CV-based estimates; red bars show ±1 standard deviation computed over 500 independent trials. Blue dashed lines show baseline mean estimates, and green dotted lines show ±1 standard deviation. The ground truth is shown in yellow.

7.2 Linear Regression

We construct a synthetic latent factor model that simulates structured dependence between the primary features, auxiliary covariates, and the target variables. Each sample is associated with a latent vector $Z \sim \mathcal{N}(0, I_d)$, which controls both the primary features $X \in \mathbb{R}^p$ and a set of auxiliary features $\tilde{C} \in \mathbb{R}^q$ through linear projections. The projection matrices $A_X \in \mathbb{R}^{p \times d}$ and $A_C \in \mathbb{R}^{q \times d}$ are constructed as a combination of shared identity components and independent Gaussian noise. Relationship between X and $\tilde{C}$ is controlled by $\gamma \in [0,1]$.

$$A_X = \gamma \times I_{p \times d} + (1-\gamma) \times \mathcal{N}(0,1)^{p \times d}, \quad A_C = \gamma \times I_{q \times d} + (1-\gamma) \times \mathcal{N}(0,1)^{q \times d}.$$

Here $I_{p \times d}$ denotes the first p rows of the $d \times d$ identity matrix. The primary and auxiliary features are then generated as:

$$X = ZA_X^\top + \sigma_X \times \mathcal{N}(0,1)^p, \quad \tilde{C} = ZA_C^\top + \sigma_Y \times \mathcal{N}(0,1)^q.$$

An independent noise term $\varepsilon_y \sim \mathcal{N}(0, \sigma^2)$ is sampled for each sample. The response variable is given by:

$$Y = X\theta_X^* + \tilde{C}\theta_C^* + \varepsilon_y,$$

where $\theta_X^* \in \mathbb{R}^p$ and $\theta_C^* \in \mathbb{R}^q$ are the true underlying coefficients.

We do not observe $\tilde{C}$ directly. Instead, we observe a noisy proxy C, constructed as:

$$C = \tilde{C} + \varepsilon_y X.$$

This formulation introduces a structured correlation between the auxiliary features, the primary features, and the label noise. It mimics real-world situations where auxiliary information (e.g., diagnostic tests, sensor data, etc.) may be influenced by both latent factors and components of the primary feature due to shared measurement pipelines or feedback loops. While synthetic, this construction enables control over the signal-to-noise ratio and feature correlation structure, allowing us to evaluate the algorithm in the presence of structured noise and information leakage.

Unless otherwise specified, we set $p = q = 5$, $d = 10$, $\gamma = 0.95$, and $\sigma_x = \sigma_Y = 0.5$, with $\sigma = 10$ for label noise. The true parameters are $\theta_X^* = [5,6,7,8,9]^\top$, and $\theta_C^* = [1,2,3,4,5]^\top$.. We use $n_{train} = 50$ and $n_{test} = 100$, and average results over 500 independent trials.

In Figure 2, we compare three regression methods: (i) a baseline using both X and C, (ii) a model using X alone, and (iii) our proposed CV estimator. The left panel shows that the CV method consistently achieves lower test mean squared error (MSE), especially as the output noise σ increases. These indicate the robust outperformance of our CV method in low signal-to-noise regimes. The right panel shows the variance of estimated coefficient for X across trials, where CV method again outperforms both baselines.

To further explore the effect of dependence between C and X, we vary the projection alignment parameter γ. As shown in Figure 3, increasing γ improves all methods, as it reduces conditional variance in Y. Nevertheless, the CV method consistently yields superior performance across the full range of γ, indicating its effectiveness under strong feature correlation.

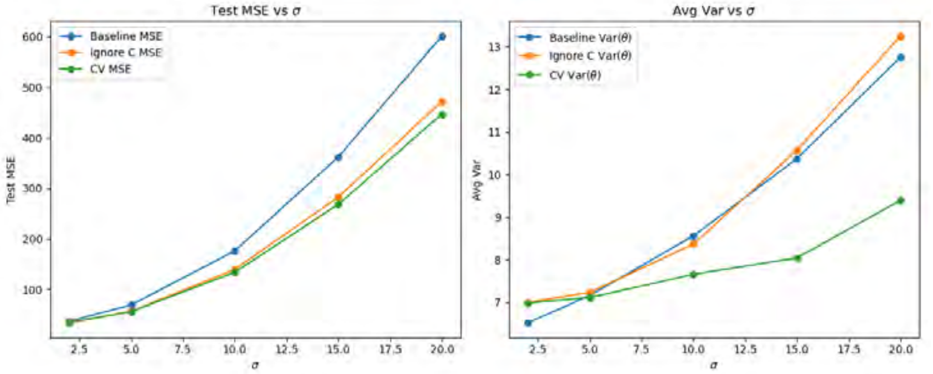

Figure 2: Linear Regression. Test MSE and variance of θ_X-estimation v.s. standard deviation of ε_y ranging from 2 to 20.

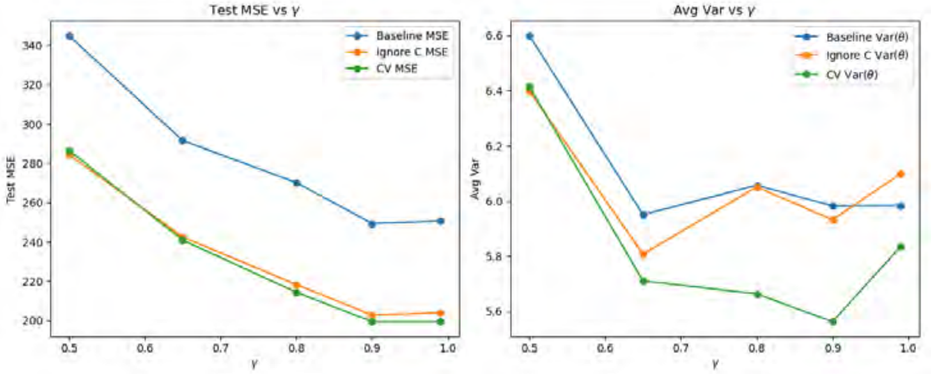

Figure 3: Linear Regression. Test MSE and variance of θ_X-estimation under different level of dependence.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, Laboratory for AI-Powered Financial Technologies, and the Columbia Innovation Hub Award.

REFERENCES

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

Bertsimas, D., and N. Kallus. 2020. "From Predictive to Prescriptive Analytics". *Management Science* 66(3):1025–1044.

Blanchet, J., H. Chen, Y. Lu, and L. Ying. 2023. "When Can Regression-Adjusted Control Variate Help? Rare Events, Sobolev Embedding and Minimax Optimality". In *NeurIPS 2023*. Poster, December 3–9, 2023, Vancouver, Canada, 36566–36578.

Durante, F., and C. Sempi. 2010. "Copula Theory: An Introduction". In *Copula Theory and Its Applications*, 3–31. Berlin, Heidelberg: Springer Berlin Heidelberg.

Elmachtoub, A. N., and P. Grigas. 2022. "Smart "Predict, then Optimize"". *Management Science* 68(1):9–26.

Elmachtoub, A. N., H. Lam, H. Zhang, and Y. Zhao. 2023. "Estimate-then-Optimize versus Integrated-Estimation-Optimization versus Sample Average Approximation: A Stochastic Dominance Perspective". *arXiv preprint arXiv:2304.06833*.

Fang, C., C. J. Li, Z. Lin, and T. Zhang. 2018. "Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator". In *NeurIPS 2018*. December 2-8, 2018, Montréal, Canada.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*, Volume 53. Springer.

Glynn, P. W., and W. Whitt. 1989. "Indirect Estimation via L= λ W". *Operations Research* 37(1):82–103.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 2011. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

Henderson, S. G., and P. W. Glynn. 2002. "Approximating Martingales for Variance Reduction in Markov Process Simulation". *Mathematics of Operations Research* 27(2):253–271.

Henderson, S. G., and B. Simon. 2004. "Adaptive Simulation using Perfect Control Variates". *Journal of applied probability* 41(3):859–876 https://doi.org/DOI:10.1239/jap/1091543430.

Hesterberg, T. 1996. "Control Variates and Importance Sampling for Efficient Bootstrap Simulations". *Statistics and Computing* 6(2):147–157.

Hesterberg, T., and B. L. Nelson. 1998. "Control Variates for Probability and Quantile Estimation". *Management Science* 44(9):1295–1312.

Johnson, R., and T. Zhang. 2013. "Accelerating Stochastic Gradient Descent using Predictive Variance Reduction". *NeurIPS 2013*:315–323 https://doi.org/10.5555/2999611.2999647. December 5–10, 2013, Lake Tahoe, Nevada, USA.

Kim, S., and S. G. Henderson. 2007. "Non-linear Control Variates for Regenerative Steady-State Simulation". In *2007 Winter Simulation Conference (WSC)*, 430–438 https://doi.org/10.1109/WSC.2007.4419632.

Kosorok, M. R. 2008. *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

L'Ecuyer, P., and E. Buist. 2006. "Variance Reduction in the Simulation of Call Centers". In *2006 Winter Simulation Conference (WSC)*, 604–613 https://doi.org/10.1109/WSC.2006.323136.

Liu, H., Y. Feng, Y. Mao, D. Zhou, J. Peng, and Q. Liu. 2017. "Action-Depedent Control Variates for Policy Optimization via Stein's Identity". *arXiv preprint arXiv:1710.11198*.

Müller, T., F. Rousselle, A. Keller, and J. Novák. 2020. "Neural Control Variates". *ACM Transactions on Graphics (TOG)* 39(6):1–19.

Oates, C. J., M. Girolami, and N. Chopin. 2017. "Control Functionals for Monte Carlo Integration". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79(3):695–718.

Portier, F., and J. Segers. 2019. "Monte Carlo Integration with a Growing Number of Control Variates". *Journal of Applied Probability* 56(4):1168–1186.

Reddi, S. J., A. Hefny, S. Sra, B. Póczos, and A. Smola. 2016. "Stochastic Variance Reduction for Nonconvex Optimization". In *NeurIPS 2016*, 3140–3148. December 5-10, 2016, Barcelona, Spain.

Srivastava, P. R., Y. Wang, G. A. Hanasusanto, and C. P. Ho. 2021. "On Data-Driven Prescriptive Analytics with Side Information: A Regularized Nadaraya-Watson Approach". *arXiv preprint arXiv:2110.04855*.

Sun, Z., C. J. Oates, and F.-X. Briol. 2023. "Meta-learning Control Variates: Variance Reduction with Limited Data". *arXiv preprint arXiv:2303.04756*.

Tsai, S. C., J. Luo, G. Jiang, and W. C. Y. and. 2023. "Adaptive Fully Sequential Selection Procedures with Linear and Nonlinear Control Variates". *IISE Transactions* 55(6):561–573 https://doi.org/10.1080/24725854.2022.2076178.

Van der Vaart, A. W., and J. A. Wellner. 1996. *Weak Convergence and Empirical Processes*. Springer.

Wang, C., X. Chen, A. J. Smola, and E. P. Xing. 2013. "Variance Reduction for Stochastic Gradient Optimization". In *NeurIPS 2013*, 181–189. December 5–10, 2013, Lake Tahoe, Nevada, USA.

Yang, W. N., and B. Nelson. 1989. "Optimization using Common Random Numbers, Control Variates and Multiple Comparisons with the Best". In *1989 Winter Simulation Conference (WSC)*, 444–449 https://doi.org/10.1109/WSC.1989.718711.

AUTHOR BIOGRAPHIES

HENRY LAM is an Associate Professor in the Department of Industrial Engineering and Operations Research at Columbia University. His email address is henry.lam@columbia.edu and his website is http://www.columbia.edu/~khl2114/.

ZITONG WANG is a Ph.D. student of Industrial Engineering and Operations Research at Columbia University. His primary research interests are stochastic optimization and uncertainty quantification. His email address is zw2690@columbia.edu.