A GENERALIZED FRAMEWORK FOR APPROXIMATE CONTROL VARIATES

ALEX A. GORODETSKY*†, GIANLUCA GERACI*‡, MIKE ELDRED‡, AND JOHN D. JAKEMAN‡

Abstract. We describe and analyze a Monte Carlo (MC) sampling framework for accelerating the estimation of statistics of computationally expensive simulation models using an ensemble of models with lower cost. Our approach uses control variates, with unknown means that must be estimated from data, to reduce the variance in statistical estimators relative to MC. Our framework unifies existing multi-level, multi-index, and multi-fidelity MC algorithms and leads to new and more efficient sampling schemes. Our results indicate that the variance reduction achieved by existing algorithms that explicitly or implicitly estimate control means, such as multilevel MC and multifidelity MC, is limited to that of a single linear control variate with known mean regardless of the number of control variates. We show how to circumvent this limitation and derive a new family of schemes that make full use of all available information sources. In particular, we demonstrate that a significant gap can exist, of orders of magnitude in some cases, between the variance reduction achievable by current schemes and our generalized schemes. We also present initial sample allocation approaches for exploiting this gap, which yield the greatest benefit when augmenting the high-fidelity model evaluations is impractical because, for instance, they arise from a legacy database. Several analytic examples and two PDE problems (viscous Burgers and steady state diffusion) are considered to demonstrate the methodology.

1. Introduction. The numerical evaluation of integrals is a foundational aspect of mathematics, with impact on diverse areas such as finance, uncertainty quantification, stochastic programming, and many others. Monte Carlo (MC) sampling is arguably the most robust means of estimating such integrals and can be easily applied to arbitrary integration domains and measures. The MC estimate of an integral is unbiased, and its rate of convergence is independent of the number of variables and smoothness of the integrand.

However, when the integrand consists of a function that is expensive to evaluate, for instance a high-fidelity computational simulation, obtaining a MC estimate with even moderate accuracy can be infeasible because the variance of a MC estimator is proportional to the ratio of the variance of the true integral and the number of samples used. Control variate (CV) techniques have a long history of enhancing MC estimates of integrals [18, 20, 19, 14]. They reduce the number of samples to compute an integral by introducing additional functions, which are correlated with the integrand, to reduce the variance of its estimate.

The use of CV methods has recently seen a resurgence, driven by the field of uncertainty quantification (UQ). In UQ, the need to evaluate statistics of computationally expensive high-fidelity models is of paramount importance, and multiple information sources, arising from multiple simulation models, are often available to accelerate the convergence of statistics [22, 11, 27, 2, 6]. Such models can be used to accelerate algorithms for both forward UQ problem and inverse UQ [1]. A set of models can result from simulating different sets of equations (i.e., the multifidelity case of differing model forms) and from varying temporal and spatial discretizations (i.e., the multilevel case of differing numerical resolutions for the same set of equations). Multiple conceptual dimensions can exist within a modeling hierarchy, leading to the generalized notion of multi-index MC methods [12], although it can be important to distinguish dimensions which define a convergent sequence (i.e., multilevel) from those that do not (e.g., multifidelity). In the most general setting, the model ensemble could include reduced-order models [28], dimension-reduction or surrogate models [23] (e.g., active subspace approximations), and even data from physical experiments [17]. Finally, both multi-physics and multi-scale simulations contribute additional combinatorial richness to the modeling ensemble.

Traditional CV methods [18] require explicit knowledge of the statistics (for instance the expected value) of their approximate information sources, however these estimates are frequently unavailable *a priori* in the UQ simulation-based context. Consequently, CV methods must be modified to balance the computational cost of evaluating lower fidelity models and the reduction in error they each provide. Existing strategies for using these CV techniques in the UQ context have therefore focused on slightly modified frameworks. For instance, in the case where additional information sources arise from a hierarchy of discretization levels, multilevel and multiindex Monte Carlo (MLMC) [8, 12] use the additional information sources through the estimation of statistics of discrepancy terms. When the variance of these discrepancy decays, significant variance reduction can be achieved. In more general contexts, heuristic estimators have been developed and analyzed [24, 26]. These estimators use additional samples of low-fidelity information sources to ob-

*The first two authors contributed equally to the concepts in this paper

†Department of Aerospace Engineering, University of Michigan, Ann Arbor, MI (goroda@umich.edu)

‡Optimization and Uncertainty Quantification Department, Sandia National Laboratories, Albuquerque, NM, 87185-1318 (ggeraci,mseldre,jdjakem@sandia.gov)

tain approximations of their missing statistical information. Finally, more flexible multilevel-multifidelity estimators that combine these notions have also been investigated [3, 5, 6].

Within these multilevel, multifidelity, and multilevel-multifidelity UQ strategies, an optimal sample allocation is defined across different model instances for the case where a recursive model dependency structure has been defined *a priori*. In this paper, we present a framework to leverage control variate techniques in the context where both the control-variate means and the model dependency structure are unknown. We demonstrate how existing algorithms, e.g., multilevel Monte Carlo (MLMC) [9] and multifidelity Monte Carlo (MFMC) [21, 26], are suboptimal realizations of this framework in that they do not extract all possible value from the correlations among information sources.

There are two mechanisms by which control variates can be used to reduce the variance of a MC estimator: (1) increasing the number of high fidelity evaluations and (2) leveraging the correlations of the low-fidelity models to the greatest extent possible. An optimal way to reduce the variance uses both of these mechanisms to allocate samples amongst all available information sources to reduce the variance per unit cost. In this paper, we provide a framework to exploit the second pathway for variance reduction in a more optimal way. Furthermore, we then embed this framework into an optimal sample allocation scheme. In particular, we extend the control variate theory cited above to the case of unknown control-variate means and make three primary contributions:

1. New theoretical results demonstrating the sub-optimality (in term of maximum possible variance reduction for a fixed high-fidelity number of evaluations) of MLMC and MFMC.
   (a) Within the context of MLMC, we derive a novel estimator that, for fixed resources, is provably more accurate than the standard MLMC estimator.
   (b) Within the context of MFMC, we derive a drop-in replacement ("ACV-MF") that can address this sub-optimality.
2. A framework for new control variate algorithms that, for estimated control means, converges toward a level of variance reduction that can be orders of magnitude greater than the maximum level of variance reduction achievable by MLMC and MFMC.
3. Initial sample allocation strategies for optimal variance reduction given a fixed computational cost.

A significant gap can exist, of orders of magnitude in some cases, between the variance reduction achievable by current schemes and our generalized schemes. The initial sample allocation strategies we present for exploiting this gap are most effective when the number of high-fidelity samples that can be generated is fixed or the cost of the high fidelity model is significantly more expensive than the low fidelity models. This situation occurs in practice when the high fidelity model is expensive to run and requires significant computational resources. In this scenario, our estimators can achieve greater variance reduction by changing the allocation and pairings of samples for each of the approximate models used as control variates. Central to these observations, we prove the following two results:

- MLMC and MFMC are recursive estimators, and as a result, asymptotically approach the variance reduction that is possible using only a single control variate with known mean (Theorems 2.4 and 2.7), regardless of the amount of data.
- An approximate control variate scheme can be devised to converge to the optimal linear control variate in the limit of infinite data (Theorem 3.6).

The remainder of the paper is structured as follows: in Section 2, we describe a unifying framework for multi-model control variate approaches and then identify existing multilevel and multifidelity estimators within this general formalism; in Section 3, we develop a set of new approximate control variate estimators; and, in Section 4, we demonstrate our approaches for several prototypical applications in fluid dynamics.

2. A unifying approximate control variate framework. In this section we provide background on control variate methods and present a generalized framework for estimating integrals using control variates with unknown mean. Then we show how existing methodologies (MLMC and MFMC) are suboptimal instances of this general framework.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\underline{Z} = (Z_1, \ldots, Z_d) \subset \mathbb{R}^d$, be a vector-valued random variable defined on this space. Let $Q : \mathbb{R}^d \to \mathcal{Q} \subset \mathbb{R}$ denote a mapping from $\underline{Z}$ to a scalar valued output, we will call this mapping a "high-fidelity model" because it will represent the process whose statistics we desire to estimate. In particular, our goal is to generate an estimate of $\mu = \mathbb{E}[Q]$ using a set of samples $\boldsymbol{z} = (\underline{z}^{(1)}, \ldots, \underline{z}^{(N)})$ of the input random variables and the corresponding evaluations of the function at those

samples $q^{(i)} = Q(\underline{z}^{(i)})$. The MC estimate of the mean

$$\hat{Q} = N^{-1} \sum_{i=1}^{N} q^{(i)} \tag{2.1}$$

is unbiased, that is $\mathbb{E}[\hat{Q}] = \mathbb{E}[Q]$, and conveges a.s. For Q with bounded variance, the Central Limit Theorem implies that the error in the estimate becomes normally distributed with variance $N^{-1}\mathbb{V}[Q]$, as $N \to \infty$.

2.1. Traditional control variate estimation. The MC-based[1] linear CV algorithm seeks a new estimator $\hat{Q}^{\mathrm{CV}}$ that has smaller variance than $\hat{Q}$ while still requiring only N evaluations of Q. The algorithm works by first introducing an additional random variable Q_1 with known mean μ_1. Then it requires computing an estimator $\hat{Q}_1$ of μ_1 using the same samples that were used for $\hat{Q}$. Finally, these quantities are assembled into the following estimator

$$\hat{Q}^{\mathrm{CV}} = \hat{Q} + \alpha \left(\hat{Q}_1 - \mu_1 \right), \tag{2.2}$$

for some scalar α[2]. As we will see shortly, the variance of $\hat{Q}^{\mathrm{CV}}$ is strongly influenced by the correlation between $\hat{Q}$ and $\hat{Q}_1$, and greater variance reduction is achieved for higher correlation. To this end we will call $\hat{Q}_1$ the CV correlated mean estimator (CME), and we will refer to μ_1 as the control variate mean (CVM). This approach can be extended to M information sources $\left(Q_i : \mathbb{R}^d \to \mathcal{Q}_i \subset \mathbb{R}\right)_{i=1}^{M}$, see e.g., [18, 20, 19], using

$$\hat{Q}^{\mathrm{CV}} = \hat{Q} + \sum_{i=1}^{M} \alpha_i \left(\hat{Q}_i - \mu_i \right), \tag{2.3}$$

where the estimator now uses a vector of CV weights $\underline{\alpha} \equiv (\alpha_1, \ldots, \alpha_M)$[3].

The effectiveness of the control variate algorithm is measured by the *variance reduction ratio* ratio between the variances of $\hat{Q}^{\mathrm{CV}}$ and $\hat{Q}$:

$$\textbf{Variance Reduction Ratio:} \quad \gamma^{\mathrm{CV}}(\underline{\alpha}) \equiv \frac{\mathbb{V}\mathrm{ar}\left[\hat{Q}^{\mathrm{CV}}(\underline{\alpha})\right]}{\mathbb{V}\mathrm{ar}\left[\hat{Q}\right]}. \tag{2.4}$$

The optimal CV (OCV) uses weights that minimize this variance:

$$\underline{\alpha}^* = \arg\min_{\underline{\alpha}} \gamma^{\mathrm{CV}}(\underline{\alpha}) = \arg\min_{\underline{\alpha}} \mathbb{V}\mathrm{ar}\left[\hat{Q}^{\mathrm{CV}}(\underline{\alpha})\right]. \tag{2.5}$$

Following [20], let $\boldsymbol{C} \in \mathbb{R}^{M \times M}$ denote the covariance matrix among Q_i and $\boldsymbol{c} \in \mathbb{R}^M$ denote the vector of covariances between Q and each Q_i. The solution to Equation (2.5) is $\underline{\alpha}^* = \boldsymbol{C}^{-1}\boldsymbol{c}$. If we further define $\bar{\boldsymbol{c}} = \boldsymbol{c}/\mathbb{V}\mathrm{ar}^{1/2}[Q] = \left[\rho_1 \mathbb{V}\mathrm{ar}^{1/2}[Q_1], \ldots, \rho_M \mathbb{V}\mathrm{ar}^{1/2}[Q_M]\right]^{\mathrm{T}}$, where ρ_i is the Pearson correlation coefficient between Q and Q_i, then the variance reduction corresponding to these weights becomes

$$\gamma^{\mathrm{CV}}(\underline{\alpha}^*) = 1 - R^2_{\mathrm{CV}}, \quad 0 \le R^2_{\mathrm{CV}} \le 1, \tag{2.6}$$

where $R^2_{\mathrm{CV}} = \bar{\boldsymbol{c}}^T \boldsymbol{C}^{-1} \bar{\boldsymbol{c}}$. When $R^2_{\mathrm{CV}} = 0$, then we have no advantage over MC, and when $R^2_{\mathrm{CV}} = 1$, we have the maximum variance reduction possible. When there is only a single approximate information source, this quantity is equivalent to $R^2_{\mathrm{CV}} = \rho_1^2$.

[1] Control-variate-based variance reduction algorithms can certainly be performed with other sampling-based and non-sampling-based estimators. In this paper we focus on the commonly Monte Carlo estimators.

[2] In the rest of the paper we drop the term "linear." It is assumed that any scheme that is discussed is linear. For reference, nonlinear control variates are those that may have polynomial corrections of the form $\hat{Q}^{\mathrm{CV}} = \hat{Q} + \sum_{i=1}^{p} \alpha_i \left(\hat{Q}_1 - \mu_1 \right)^i$

[3] CV methods can estimate multiple quantities of interest, e.g. [29, 31]. For simplicity we only consider a scalar Q.

2.2. Approximate control variate estimation. Traditional control variate estimation assumes that the means μ_i are known. In our motivating problem of multifidelity uncertainty quantification, these means are not known, but rather must be estimated from lower fidelity simulations. To this end, let $\boldsymbol{z}$ denote a set of samples used to evaluate the high-fidelity function Q. Furthermore, let $\boldsymbol{z}_i$ denote an ordered set of N_i samples partitioned into two ordered subsets $\boldsymbol{z}_i^1 \subset \boldsymbol{z}_i$ and $\boldsymbol{z}_i^2 \subset \boldsymbol{z}_i$ such that $\boldsymbol{z}_i^1 \cup \boldsymbol{z}_i^2 = \boldsymbol{z}_i$. Note that $\boldsymbol{z}_i^1$ and $\boldsymbol{z}_i^2$ are not required to be disjoint, i.e., in many strategies they overlap so that $\boldsymbol{z}_i^1 \cap \boldsymbol{z}_i^2 \neq \emptyset$. We will construct and analyze approximate CV estimators of the following form

$$\tilde{Q}(\underline{\alpha}, \underline{\boldsymbol{z}}) = \hat{Q}(\boldsymbol{z}) + \sum_{i=1}^{M} \alpha_i \left(\hat{Q}_i\left(\boldsymbol{z}_i^1\right) - \hat{\mu}_i\left(\boldsymbol{z}_i^2\right) \right) = \hat{Q}(\boldsymbol{z}) + \sum_{i=1}^{M} \alpha_i \Delta_i(\boldsymbol{z}_i) = \hat{Q} + \underline{\alpha}^T \underline{\Delta}, \tag{2.7}$$

where $\underline{\Delta} = (\Delta_1, \ldots, \Delta_M)$, $\Delta_i(\boldsymbol{z}_i) = \hat{Q}_i\left(\boldsymbol{z}_i^1\right) - \hat{\mu}_i\left(\boldsymbol{z}_i^2\right)$ and we use $\underline{\boldsymbol{z}} \equiv (\boldsymbol{z}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_M)$ to denote the input values to the high-fidelity model $\boldsymbol{z}$ and each CV model $\boldsymbol{z}_i, i = 1, \ldots, M$. Here we use additional samples to estimate each mean μ_i, and this estimate is denoted by $\hat{\mu}_i$ and called the estimated control variate mean (ECVM). *All of the estimators in the literature, and those that we derive in this paper, are differentiated by how the samples $\boldsymbol{z}_i$ are related among the CVs and how the subsets $\boldsymbol{z}_i^1$ and $\boldsymbol{z}_i^2$ are defined within each CV.* It will be useful to consider ordered sets of samples such that $\boldsymbol{z}_i = [\boldsymbol{z}_i^1, \boldsymbol{z}_i^2] = [\underline{z}_i^{1(1)}, \ldots, \underline{z}_i^{1(L_1)}, \underline{z}_i^{2(1)}, \ldots, \underline{z}_i^{2(L_2)}]$. In this case, if $\boldsymbol{z}_i^1$ and $\boldsymbol{z}_i^2$ intersect then $\underline{z}_i^{1(j)} = \underline{z}_i^{2(k)}$ for some $1 \leq j \leq L_1$ and $1 \leq k \leq L_2$.

Because we must now use extra samples to estimate μ_i, we will have to account for the additional cost this imposes. Let N denote the number of available realizations of Q and

$$N_i = \lfloor r_i N \rfloor$$

denote the total number of available evaluations of the i-th CV Q_i for scaling factor $r_i \in \mathbb{R}_+$. Let $w_i < 1$ denote the ratio between the cost of a single realization of control variates Q_i and the cost of obtaining a realization of the high-fidelity Q. Then the cost of the CV estimator is

$$w_{\text{cv}} = N + \sum_{i=1}^{M} N_i w_i. \tag{2.8}$$

For fixed $\underline{\alpha}$, the estimator $\tilde{Q}$ is unbiased when each component, $\hat{Q}$, $\hat{Q}_i$, and $\hat{\mu}_i$ for $i = 1, \ldots, M$ is unbiased. The variance of $\tilde{Q}$ is given in the following proposition.

PROPOSITION 2.1 (Variance of the approximate CV estimator). *The variance of the approximate CV estimator* (2.7) *for fixed $\underline{\alpha}$ is*

$$\mathbb{V}\text{ar}\left[\tilde{Q}\right] = \mathbb{V}\text{ar}\left[\hat{Q}\right]\left(1 + \underline{\alpha}^T \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]}{\mathbb{V}\text{ar}\left[\hat{Q}\right]}\underline{\alpha} + 2\underline{\alpha}^T \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]}{\mathbb{V}\text{ar}\left[\hat{Q}\right]}\right), \quad \textit{where } \underline{\Delta} = (\Delta_1, \ldots, \Delta_M). \tag{2.9}$$

Proof. This result follows directly from the the rules regarding the variance of MC estimators, the variance of scaled random variables, and the variance of sums of random variables. We have

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\tilde{Q}(\underline{\alpha})\right] &= \mathbb{V}\text{ar}\left[\hat{Q}(\boldsymbol{z}) + \sum_{i=1}^{M} \alpha_i \Delta_i(\boldsymbol{z}_i)\right] = \mathbb{V}\text{ar}\left[\hat{Q}\right] + \underline{\alpha}^T \mathbb{C}\text{ov}[\underline{\Delta}, \underline{\Delta}]\underline{\alpha} + 2\underline{\alpha}^T \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right] \\
&= \mathbb{V}\text{ar}[\hat{Q}]\left(1 + \underline{\alpha}^T \frac{\mathbb{C}\text{ov}[\underline{\Delta}, \underline{\Delta}]}{\mathbb{V}\text{ar}[\hat{Q}]}\underline{\alpha} + 2\underline{\alpha}^T \frac{\mathbb{C}\text{ov}[\underline{\Delta}, \hat{Q}]}{\mathbb{V}\text{ar}[\hat{Q}]}\right),
\end{aligned}$$

where the middle term on the first line is the definition, the last term follows from the variance of sums of random variables, and the second line factors out the baseline estimator variance. ∎

This result means that the variance ratio of the approximate control variate (ACV) estimator is

$$\gamma^{\text{ACV}}(\underline{\alpha}) = 1 + \underline{\alpha}^T \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]}{\mathbb{V}\text{ar}\left[\hat{Q}\right]}\underline{\alpha} + 2\underline{\alpha}^T \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]}{\mathbb{V}\text{ar}\left[\hat{Q}\right]}. \tag{2.10}$$

The optimal approximate CV estimator consists of the $\underline{\alpha}$ which minimizes this ratio. The properties of this optimal estimator are given in the following proposition.

PROPOSITION 2.2 (Optimal approximate CV). *Assume* $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ *is positive definite and* $\mathbb{V}\text{ar}\left[\hat{Q}\right] > 0$. *The CV weight that provides the greatest variance reduction for the approximate CV Equation* (2.7) *is given by*

$$\underline{\alpha}^{\text{ACV}} = \arg\min_{\underline{\alpha}} \gamma^{\text{ACV}}(\underline{\alpha}) = -\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]^{-1} \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right] \tag{2.11}$$

with corresponding estimator variance

$$\mathbb{V}\text{ar}\left[\tilde{Q}(\underline{\alpha}^{\text{ACV}})\right] = \mathbb{V}\text{ar}\left[\hat{Q}\right]\left(1 - \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]^T \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]^{-1}}{\mathbb{V}\text{ar}\left[\hat{Q}\right]} \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]\right). \tag{2.12}$$

Proof. The variance reduction γ^{ACV} is a quadratic function of $\underline{\alpha}$ and so its extremum can be found by setting the gradient to zero

$$0 = \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]}{\mathbb{V}\text{ar}[\hat{Q}]}\underline{\alpha}^{\text{ACV}} + \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]}{\mathbb{V}\text{ar}\left[\hat{Q}\right]} \qquad \Longrightarrow \qquad \underline{\alpha}^{\text{ACV}} = -\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]^{-1} \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right].$$

If we substitute this weight into Equation (2.9), we obtain the stated result

$$\begin{aligned}\mathbb{V}\text{ar}\left[\underline{\alpha}^{\text{ACV}}\right] = \mathbb{V}\text{ar}\left[\hat{Q}\right]\Big(1 + \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]^T \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]^{-1}}{\mathbb{V}\text{ar}\left[\hat{Q}\right]} \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right] - \\ 2\,\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]^T \mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]^{-1} \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]}{\mathbb{V}\text{ar}\left[\hat{Q}\right]}\Big) \\ = \mathbb{V}\text{ar}\left[\hat{Q}\right]\left(1 - \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]^T \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]^{-1}}{\mathbb{V}\text{ar}\left[\hat{Q}\right]} \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]\right).\end{aligned}$$

∎

Proposition 2.2 provides a means to construct and theoretically compare different control variate estimators. In the following, we will use this tool to derive a number of existing estimators which use different sampling strategies to compute $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ and $\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]$.

2.3. Multilevel Monte Carlo (MLMC). The MLMC estimator [8] is inspired by ideas drawn from multigrid solvers for PDEs. It posits that if a stochastic process can be simulated using various resolution methods, then low resolution simulations can be used to obtain a reduction in the variance of the statistics of higher resolution simulations. An initial set of motivating applications included stochastic differential equations [7, 9, 13], but it has been extended also to elliptic problems [30] (among others), filtering problems [15] and a large number of other scientific and engineering applications [16, 6, 4, 10].

In this section, we provide several new contributions. First, we derive a novel expression for the variance reduction provided by MLMC that is a function of the correlations between each model. Second, we identify that MLMC is sub-optimal in that it cannot provide a greater variance reduction than the optimal control variate estimator with one control variate (OCV-1), regardless of how many resolutions or how much data is used. Third, we provide a simple enhancement to the standard MLMC assembly strategy to provide a provably greater variance reduction for virtually no additional cost.

Let $\{P_0, \ldots, P_L\}$ be a sequence of random variables, ordered bottom-up[4] from coarsest resolution at level 0 to finest resolution at level L. The MLMC estimator seeks the expectation of P_L by rewriting it as

[4]Here, we use P_i instead of Q_i because in this paper Q_i is always of "higher fidelity" than Q_j, for $i < j$, and MLMC traditionally reverses these notions. We use a different letter and then provide a mapping back to Q_i to reflect this discrepancy.

a telescopic sum $\mathbb{E}[P_L] = \mathbb{E}[P_0] + \sum_{\ell=1}^{L} \mathbb{E}[P_\ell - P_{\ell-1}]$. Then MLMC calls for (by definition) estimating the expectation using *independent* sample sets for each level ℓ, by summing the following sequence of estimators

$$\hat{P}_0 + \sum_{i=1}^{L} \left(\widehat{P_i - P_{i-1}} \right). \tag{2.13}$$

From the cancellation of terms, we see that this estimator is still an unbiased estimate of $\mathbb{E}[P_L]$.

To link this expression to Eq. (2.7), we make the following associations: the number of control variates is $M = L$, the finest resolution level is our quantity of interest $Q = P_L$, and the order of P_i is reversed to obtain our control variates $Q_i = P_{L-i}$ for $i = 1, \ldots, L$. Using this assignment it is clear that the the same sum can be written as

$$\hat{Q}^{\text{MLMC}}(\underline{\boldsymbol{z}}) = \left(\widehat{Q - Q_1} \right)(\boldsymbol{z}) + \sum_{i=1}^{M-1} \left(\widehat{Q_i - Q_{i+1}} \right)(\boldsymbol{z}_i^2) + \hat{Q}_M(\boldsymbol{z}_M^2) \tag{2.14}$$

where we have introduced independent sample sets $\boldsymbol{z}_i^2$ for each level with $\boldsymbol{z}_i^2 \cap \boldsymbol{z}_j^2 = \emptyset$ for $i \neq j$ and $\boldsymbol{z}_i^2 \cap \boldsymbol{z} = \emptyset$. From here it is clear that two estimators of each QoI Q_i are obtained; one estimator uses $\boldsymbol{z}_i^2$ and the other $\boldsymbol{z}_{i-1}^2$. Thus we can identify the ECVM of Q_i with the estimator that uses $\boldsymbol{z}_i^2$ since this is the set with more samples. Accounting for the sample set for the high fidelity model we can rewrite this equation as

$$\hat{Q}^{\text{MLMC}}(\underline{\boldsymbol{z}}) = \hat{Q}(\boldsymbol{z}) - \hat{Q}_1(\boldsymbol{z}) + \sum_{i=1}^{M-1} \left(\hat{\mu}_i(\boldsymbol{z}_i^2) - \hat{Q}_{i+1}(\boldsymbol{z}_i^2) \right) + \hat{\mu}_M(\boldsymbol{z}_M^2).$$

Finally, re-pairing terms to reflect the same level but different sample sets, we arrive at the form of the approximate CV in (2.7)

$$\hat{Q}^{\text{MLMC}}(\underline{\boldsymbol{z}}) = \hat{Q}(\boldsymbol{z}) + \sum_{i=1}^{M} (-1) \left(\hat{Q}_i(\boldsymbol{z}_i^1) - \hat{\mu}_i(\boldsymbol{z}_i^2) \right) \tag{2.15}$$

where we have set $\boldsymbol{z}_1^1 = \boldsymbol{z}$ and $\boldsymbol{z}_i^2 = \boldsymbol{z}_{i+1}^1$ for $i = 1, \ldots, M-1$ and $\alpha_i = -1$ for all i. It is important to note that the sampling strategy implies a recursive underpinning to the MLMC strategy. In this sense, MLMC can be derived by starting with a single Q_1 with unknown CVM. Then we introduce Q_i *as a control variate* for the ECVM. It is for this reason that the samples between $\hat{\mu}_i$ and $\hat{Q}_{i+1}$ are shared.

Given this partitioning, the receptive cardinality of $\boldsymbol{z}_i^2$ and $\boldsymbol{z}_i^1$ is $\bar{r}_i N$ and $\bar{r}_{i-1} N$, and

$$\Delta_i(\boldsymbol{z}_i) = \hat{Q}_i\left(\boldsymbol{z}_i^1\right) - \hat{\mu}_i\left(\boldsymbol{z}_i^2\right) = \frac{1}{\bar{r}_{i-1} N} \sum_{j=1}^{\bar{r}_{i-1} N} Q_i(\underline{\boldsymbol{z}}_i^{1(j)}) - \frac{1}{\bar{r}_i N} \sum_{j=1}^{\bar{r}_i N} Q_i(\underline{\boldsymbol{z}}_i^{2(j)}), \tag{2.16}$$

for $i = 1, \ldots, M$. Here the evaluation ratios for each model are $r_i = \bar{r}_i + \bar{r}_{i-1}$ and $\bar{r}_0 = 1$ because $\boldsymbol{z}_i^2 \cap \boldsymbol{z}_i^1 = \emptyset$. The sampling scheme is illustrated in Figure 2.1a for reference. The variance reduction of the MLMC estimator is given in the following proposition.

LEMMA 2.3 (Variance reduction of MLMC). *Let the MLMC estimator be defined as*

$$\hat{Q}^{\text{MLMC}}(\underline{\boldsymbol{\alpha}}, \underline{\boldsymbol{z}}) = \hat{Q}(\boldsymbol{z}) + \sum_{i=1}^{M} (-1) \left(\hat{Q}_i(\boldsymbol{z}_i^1) - \hat{\mu}_i(\boldsymbol{z}_i^2) \right), \tag{2.17}$$

with the sampling strategy: $\underline{\boldsymbol{z}} = (\boldsymbol{z}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_M)$. *Each ordered set of samples* $\boldsymbol{z}_i$ *is partitioned according to* $\boldsymbol{z}_i = \boldsymbol{z}_i^1 \cup \boldsymbol{z}_i^2$, $\boldsymbol{z}_i^1 \cap \boldsymbol{z}_i^2 = \emptyset$, $\boldsymbol{z}_1^1 = \boldsymbol{z}$, $\boldsymbol{z}_1^2 \neq \emptyset$ *and* $\boldsymbol{z}_i^2 = \boldsymbol{z}_{i+1}^1$ *for* $i = 1, \ldots, M-1$. *The variance of this estimator is*

$$\mathbb{V}\text{ar}\left[\hat{Q}^{\text{MLMC}}(\underline{\boldsymbol{\alpha}})\right] = \mathbb{V}\text{ar}\left[\hat{Q}\right]\left(1 - R^2_{\text{MLMC}}\right), \tag{2.18}$$

where

$$R^2_{\text{MLMC}} = -\sum_{i=1}^{M} \frac{\bar{r}_i + \bar{r}_{i-1}}{\bar{r}_i \bar{r}_{i-1}} \tau_i^2 + 2 \sum_{i=1}^{M-1} \frac{1}{\bar{r}_i} \rho_{i(i+1)} \tau_i \tau_{i+1} - 2\rho_1 \tau_1, \tag{2.19}$$

$\tau_i = \frac{\mathbb{V}\text{ar}^{1/2}[Q_i]}{\mathbb{V}\text{ar}^{1/2}[Q]}$ *is the ratio of the standard deviations,* ρ_{ij} *is the Pearson correlation coefficient between* (Q_i, Q_j)*, and* $\bar{r}_i$ *are defined recursively according to*

$$\bar{r}_0 = 1 \text{ and } \bar{r}_i = r_i - \bar{r}_{i-1}.$$

The proof is provided in Appendix A. MLMC provides a smaller variance reduction than the single optimal CV.

THEOREM 2.4 (Maximum variance reduction of MLMC). *The variance reduction of MLMC is bounded above by the optimal single CV i.e.,*

$$R^2_{\text{MLMC}} < \rho_1^2. \tag{2.20}$$

The proof is provided in Appendix B. This bound is a result of the sampling strategy used by MLMC, which does not allocate any of the samples used to evaluate the high fidelity model to the models Q_i for $i > 1$. Because Q_i and Q do not share samples there is no correlation between the estimators of Δ_i and $\hat{Q}$ for $i > 1$. Ignoring these correlations limits the maximum attainable variance reduction of the final estimator to the one corresponding to OCV-1; specifically, the control variates Q_i, for $i > 1$, are used to accelerate convergence of $\hat{Q}^{\text{MLMC}}$ to OCV-1 using model Q_1.

Note that the bound given in Theorem 2.4 *does not depend on the number of samples of each* Q_i. Even with an infinite number of samples of the CVs, the manner in which the samples are distributed restricts the variance reduction achieved. As far as we are aware, this sub-optimality has not been discussed previously in the literature, where the focus has been to show *some* variance reduction as compared with MC.

This sub-optimality property can have a significant effect in practice. We will show that significantly greater variance reduction can be obtained with essentially no additional computational expense by simply redistributing the sample allocation (even within the two components of Δ_i).

Finally, we briefly mention that multi-index Monte Carlo MIMC algorithm [12] is a multidimensional analogue of MLMC in that it considers a two- (or multi-) dimensional hierarchy of models P_{ij}. This estimator can also be written in the language of control variates; whereas MLMC introduces a single Q_i to reduce the variance in estimating the ECVM $\hat{\mu}_{i-1}$, MIMC introduces several control variates Q_j, Q_k to reduce the variance of the ECVM $\hat{\mu}_{i-1}$. As a result MIMC can be more efficient because it has the potential for greater variance reduction; however, it still suffers from a similar suboptimality, with respect to γ metric, because it still uses lower-fidelity models to only accelerate convergence of the μ_i of the immediate higher-fidelity models rather than of Q – in other words they are still recursive.

2.3.1. Weighted Multilevel Monte Carlo. A small modification of the MLMC algorithm can reduce its variance with no additional computational cost. This improvement is obtained by replacing the weights $\alpha_i = -1$ with the optimal weights $\underline{\alpha}^{\text{ACV}}$ given by Proposition 2.2. Since these weights guarantee that the estimator achieves the greatest variance reduction, using these weights will yields an estimator that employs the same sampling strategy as the standard MLMC algorithm (i.e., sampling in a sequence of model pairs), but has a lower variance (error).

Pseudo-code for constructing the optimal multi-level MC estimator is given in Algorithm 2.1. The values of $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ and $\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]$ are provided in Appendix A. Here we assume that the allocation of the samples $\underline{\boldsymbol{z}}$ is given a priori. For example a typical MLMC allocation strategy that attempts to balance physical prediction bias and statistical error can be used. Such an allocation may no longer be optimal. We discuss how to construct an optimal allocation strategy in Section 3.3.

2.4. Multifidelity Monte Carlo (MFMC). An alternative to MLMC, called Multifidelity Monte Carlo (MFMC), was recently proposed in [26, 22, 24, 27]. In this section, we provide several new results regarding this sampler. First, we derive a new variance reduction expression, an alternative to [26, Lemma 3.3], that relates the correlation of the CVs to the variance reduction. Second, we use this expression to

Algorithm 2.1 Weighted Multilevel Monte Carlo (W-MLMC)

Require: High fidelity model Q; M control variates $(Q_i)_{i=1}^M$; Samples for the highest fidelity model $\boldsymbol{z}$; Partitioned samples for each CV: $(\boldsymbol{z}_i = [\boldsymbol{z}_i^1, \boldsymbol{z}_i^2])_{i=1}^M$

Ensure: $\hat{Q}^{\mathrm{CV}}$: optimal MLMC estimate

$N = \mathrm{card}(\boldsymbol{z})$
$\bar{r}_0 = 1$
$\boldsymbol{A} = \mathrm{zeros}(M, M)$
for $i = 1, \dots, M-1$ **do**
  $\bar{r}_i = \mathrm{card}(\boldsymbol{z}_i) - \bar{r}_{i-1}$
  $v = \mathrm{EstimateVariance}(Q_i, \boldsymbol{z}_i)$ {Estimate the variance of the CV}
  $\boldsymbol{A}_{ii} = \frac{\bar{r}_i + \bar{r}_{i-1}}{\bar{r}_i \bar{r}_{i-1}} v$
  $c = \mathrm{EstimateCovariance}(Q_i, Q_{i+1}, \boldsymbol{z}_i^2, \boldsymbol{z}_{i+1}^1)$ {Estimate the covariance between control variates}
  $\boldsymbol{A}_{i(i+1)} = \boldsymbol{A}_{(i+1)i} = -\frac{1}{\bar{r}_i} c$
end for
$\bar{r}_M = \mathrm{card}(\boldsymbol{z}_M) - \bar{r}_{M-1}$
$v = \mathrm{EstimateVariance}(Q_M, \boldsymbol{z}_M)$
$\boldsymbol{A}_{MM} = \frac{\bar{r}_M + \bar{r}_{M-1}}{\bar{r}_M \bar{r}_{M-1}} v$
$\underline{\alpha} = -\left[\boldsymbol{A}^{-1}\right]_{:1} \mathbb{C}\mathrm{ov}\left[Q, Q_1\right]$, where $\left[\boldsymbol{A}^{-1}\right]_{:1}$ is the first column of the inverse matrix.
$\hat{Q}^{\mathrm{CV}} = \mathrm{AssembleSamples}(\underline{\alpha}, \boldsymbol{z}, \boldsymbol{z}_1, \dots, \boldsymbol{z}_M)$ {Assemble according to Equation (2.7)}

demonstrate that MFMC is a recursive estimator, like MLMC, and therefore is also sub-optimal. That is, the maximum variance reduction provided by this estimator is limited to that provided by the single optimal CV, regardless of how many samples are used to evaluate Q_i.

The MFMC estimator can be obtained from the following recursive procedure. Samples $\boldsymbol{z}$ are used to obtain $\hat{Q}$ and the CME ($\hat{Q}_1$) of Q_1. Then an enriched set of samples $\boldsymbol{z}_1$ are used for the ACVM ($\hat{\mu}_1$) of Q_1 to obtain

$$\tilde{Q}\left(\alpha, \boldsymbol{z}, \boldsymbol{z}_1\right) = \hat{Q}(\boldsymbol{z}) + \alpha(\hat{Q}_1(\boldsymbol{z}) - \hat{\mu}_1(\boldsymbol{z}_1)), \tag{2.21}$$

where we have chosen $\boldsymbol{z}_1^1 = \boldsymbol{z}$ and $\boldsymbol{z}_1^2 = \boldsymbol{z}_1$. In other words, $\boldsymbol{z}_1$ consists of the original $\boldsymbol{z}$ samples along with an additional set.

Now, let us introduce another CV to reduce the variance of the $\hat{\mu}_1$ according to

$$\begin{aligned}\tilde{Q}\left(\alpha, \tilde{\alpha}, \boldsymbol{z}, \boldsymbol{z}_1, \boldsymbol{z}_2\right) &= \hat{Q}(\boldsymbol{z}) + \alpha(\hat{Q}_1(\boldsymbol{z}_1^1) - \left(\hat{\mu}_1(\boldsymbol{z}_1) + \tilde{\alpha}\left(\hat{Q}_2(\boldsymbol{z}_1) - \hat{\mu}_2(\boldsymbol{z}_2)\right)\right) \\ &= \hat{Q}(\boldsymbol{z}) + \alpha(\hat{Q}_1(\boldsymbol{z}_1^1) - \hat{\mu}_1(\boldsymbol{z}_1)) - \alpha\tilde{\alpha}\left(\hat{Q}_2(\boldsymbol{z}_1) - \hat{\mu}_2(\boldsymbol{z}_2)\right)\end{aligned}$$

where we have again set $\boldsymbol{z}_2^1 = \boldsymbol{z}_1$ and $\boldsymbol{z}_2^2 = \boldsymbol{z}_2$.

Continuing in this recursive pattern and collapsing α-products, we obtain the MFMC estimator

$$\hat{Q}^{\mathrm{MFMC}}\left(\underline{\alpha}, \underline{\boldsymbol{z}}\right) = \hat{Q}(\boldsymbol{z}) + \sum_{i=1}^{M} \alpha_i \Delta_i(\boldsymbol{z}_i) \quad \text{such that} \quad \Delta_i(\boldsymbol{z}_i) = \hat{Q}_i\left(\boldsymbol{z}_i^1\right) - \hat{\mu}_i\left(\boldsymbol{z}_i\right), \tag{2.22}$$

and the sampling strategy partitions $\boldsymbol{z}_i$ into $\boldsymbol{z}_i^1$ and $\boldsymbol{z}_i^2$ according to

$$\boldsymbol{z}_i^1 = \boldsymbol{z}_{i-1} \quad \text{and} \quad \boldsymbol{z}_i^2 = \boldsymbol{z}_i.$$

for $i = 2, \dots, M$ and $\boldsymbol{z}_1^1 = \boldsymbol{z}$ and $\boldsymbol{z}_1^2 = \boldsymbol{z}_1$. The sampling scheme for the MFMC estimator is illustrated in Figure 2.1b for reference.

Unlike the MLMC estimator, this estimator has free weights that we can tune, and the optimal weights are given in the following Lemma.

LEMMA 2.5 (Optimal CV weights for MFMC). *The optimal weights for the MFMC estimator are*

$$\alpha_i^{\mathrm{MFMC}} = -\frac{\mathbb{C}\mathrm{ov}\left[Q, Q_i\right]}{\mathbb{V}\mathrm{ar}\left[Q_i\right]} \quad \textit{for} \quad i = 1, \dots, M. \tag{2.23}$$

The proof is provided in Appendix C, and provides an alternate derivation to the identical result in [26, Th. 2.4], but it does not require the same assumptions; in the prior work, these weights were derived using assumptions on model costs and correlation orderings. The next Lemma provides a new representation of the variance reduction of this estimator in terms of the correlations between model pairs.

LEMMA 2.6 (Variance reduction of the optimal MFMC). *Assume* $|\rho_1| > 0$ *and let* $r_0 = 1$. *The variance reduction of the optimal MFMC estimator is*

$$\mathbb{V}\mathrm{ar}\left[\underline{\alpha}^{\mathrm{MFMC}}\right] = \frac{\mathbb{V}\mathrm{ar}\left[Q\right]}{N}\left(1 - R^2_{\mathrm{MFMC}}\right) \tag{2.24}$$

where

$$R^2_{\mathrm{MFMC}} = \sum_{i=1}^{M} \frac{r_i - r_{i-1}}{r_i r_{i-1}}\rho_i^2 = \rho_1^2\left(\frac{r_1 - 1}{r_1} + \sum_{i=2}^{M} \frac{r_i - r_{i-1}}{r_i r_{i-1}}\frac{\rho_i^2}{\rho_1^2}\right). \tag{2.25}$$

The proof is provided in Appendix D. Under the assumption made in [26] that, without loss of generality (can be constructed by reordering), $|\rho_1| \geq |\rho_i|$ for $i = 2, \ldots, M$, we can use Lemma 2.6 to show that the MFMC estimator cannot perform better than the optimal single CV, regardless of how many samples are allocated to the control variate models.

THEOREM 2.7 (Maximum variance reduction of MFMC). *The variance reduction of MFMC is bounded above by the optimal single CV, i.e.,*

$$R^2_{\mathrm{MFMC}} < \rho_1^2. \tag{2.26}$$

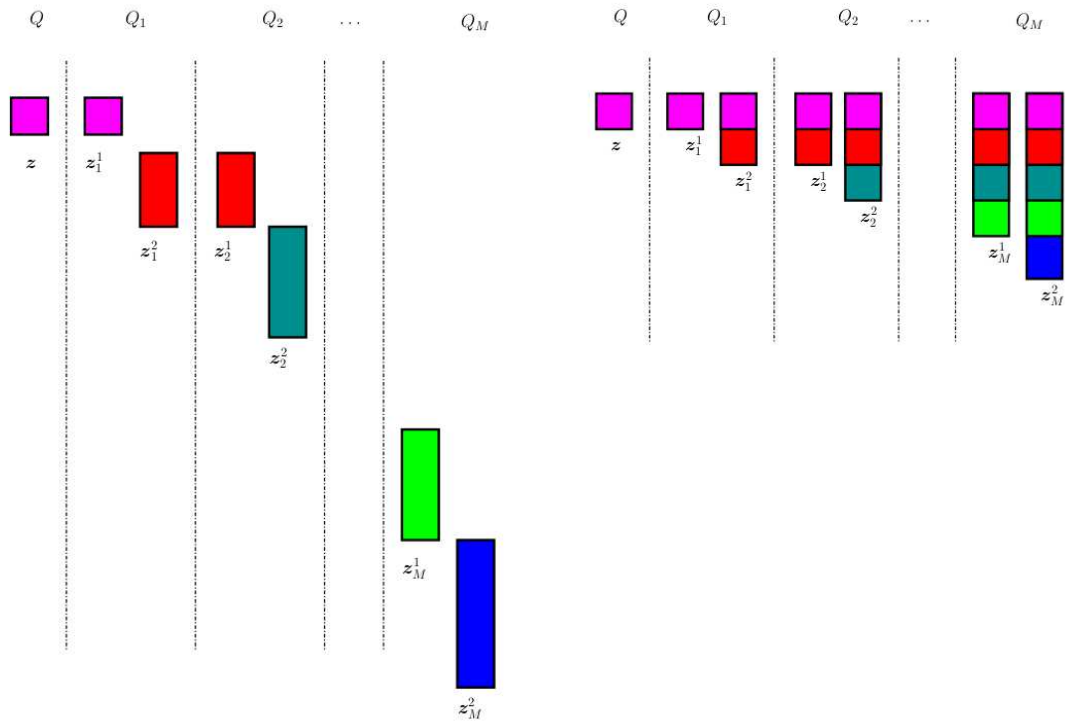
The proof is provided in Appendix E. The variance reduction R^2_{MFMC} of MFMC approaches ρ_1^2 in the limit of infinite data. Though the MFMC estimator is limited by the highest correlation ρ_1^2, it will be able to approach this limit more efficiently than a single low-fidelity CV because r_M is often larger than r_1, given the (assumed) decreasing computational cost in the sequence from the first to the last low-fidelity model M. Similarly to the MLMC case, the suboptimality of MFMC is due to the recursive nature of this estimator. Each successive CV Q_i is used to enhance convergence of $\hat{\mu}_{i-1}$.

2.5. Summary and example. Table 2.1 below summarizes the sample distribution and the variance reduction ratio of the state-of-the-art methods described above, and Figure 2.1 pictorially illustrate the existing variance reduction algorithms that do not require a known mean.

Algorithm	Relation between $\boldsymbol{z}$ and $\boldsymbol{z}_i$	$\boldsymbol{z}_i^1$	$\boldsymbol{z}_i^2$	Reduction ratio γ
OCV	$\boldsymbol{z}_i = \boldsymbol{z}$	$\boldsymbol{z}$	$\emptyset$	$1 - R^2_{\mathrm{CV}}$
OCV-1	$\boldsymbol{z}_i = \boldsymbol{z}$	$\boldsymbol{z}$	$\emptyset$	$1 - \rho_1^2$
MLMC [8]	$\boldsymbol{z}_1^1 = \boldsymbol{z}$, $\boldsymbol{z} \cap \boldsymbol{z}_i = \emptyset$ for $i > 1$	$\boldsymbol{z}_{i-1}^2$	$\boldsymbol{z}_i \setminus \boldsymbol{z}_i^1$	$1 - R^2_{\mathrm{MLMC}}$, $R^2_{\mathrm{MLMC}} < \rho_1^2$
MFMC [26]	$\boldsymbol{z}_i \supset \boldsymbol{z}$ for all i	$\boldsymbol{z}_{i-1}$	$\boldsymbol{z}_i$	$1 - R^2_{\mathrm{MFMC}}$, $R^2_{\mathrm{MFMC}} < \rho_1^2$

Table 2.1: Representations of various CV-type variance reduction estimators in the framework of Equation (2.7). The estimators OCV and OCV-1 refer to optimal CV estimator with known means, where OCV-1 only uses a single CV and OCV uses all M. For these estimators, no samples are required to estimate μ and therefore $\boldsymbol{z}_i^2$ is empty. As summarized in the final column, the greatest variance reduction possible with MLMC and MFMC is strictly less than or equal to the optimal CV using a single model, *regardless of the number of models or data used for MLMC and MFMC*, since Theorem 2.4 shows that $R^2_{\mathrm{MLMC}} < \rho_1^2$ and Theorem 2.7 shows that $R^2_{\mathrm{MFMC}} < \rho_1^2$.

To demonstrate these results we consider the following simple monomial example: let $Q(\omega) = \omega^5$ and $Q_i(\omega) = \omega^{5-i}$ for $i = 1, \ldots, 4$, where $\omega \sim \mathcal{U}(0, 1)$. The correlation matrix for this problem is given in Table 2.2. As we have not yet introduced a cost model, we first explore the performance of these methods through the lens of an assumed sample ratio $r_i = 2^{i+x}$ so that $\underline{r}(x) = [2 \times 2^x, 4 \times 2^x, 8 \times 2^x, 16 \times 2^x]$ for $x = 0, 1, \ldots, 29$. As a result, the sample allocations across i are prescribed by r_i, rather than computed



(a) MLMC sampling strategy.

(b) MFMC sampling strategy.

Figure 2.1: Comparison of MLMC and MFMC sampling structures.

	Q	Q_1	Q_2	Q_3	Q_4
Q	1	0.994995	0.975042	0.927132	0.820633
Q_1	0.994995	1	0.992172	0.958367	0.865941
Q_2	0.975042	0.992172	1	0.986021	0.916385
Q_3	0.927132	0.958367	0.986021	1	0.968153
Q_4	0.820633	0.865941	0.916385	0.968153	1

Table 2.2: Correlation matrix for monomial example computed with 10^5 samples.

from known w_i and estimated ρ_{ij}; this is the viewpoint taken in Figures 2.2, 3.2, and 3.3. We explore the introduction of w_i and resulting optimal allocation strategies later in Sections 3.3.1 and 4.1. These two views highlight two critical features for these algorithms: (a) the existence of a significant gap between OCV-1 and OCV, and (b) an effective ability to exploit this gap to achieve computational savings.

Figure 2.2 shows the variance reduction ratio γ for the described algorithms. The dotted horizontal lines provide baseline variance reduction ratios corresponding to MC ($\gamma = 1$), the single OCV (OCV-1), double OCV (OCV-2), triple OCV (OCV-3), and the optimal control variate that uses all the models (OCV). We clearly see that the variance reduction of the existing estimators is bounded by that provided by OCV-1, as established by Theorems 2.4 and 2.7 and summarized in Table 2.1.

To summarize, this example shows that *when the number of high fidelity samples is fixed*, the existing state-of-the-art estimators are sub-optimal. As such, in applications where it is infeasible to obtain more samples of the highest fidelity model, we can potentially achieve orders of magnitude greater variance reduction. We want to emphasize that we are aware of no existing methods in the form (2.7) that can converge to the OCV estimator (with known means) as the amount of data increases. This defines our goal for the

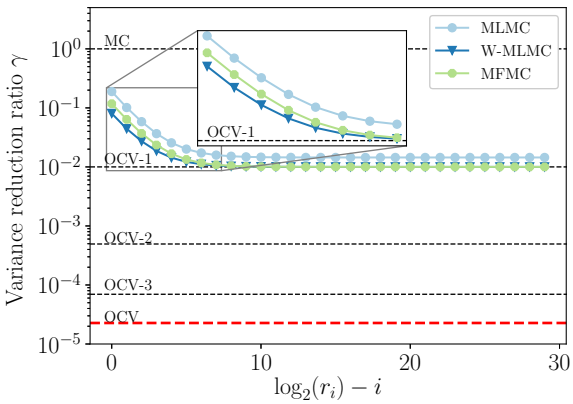


Figure 2.2: Variance reduction ratios for various estimators as a function of number of sample per level r_i and fixed number of high fidelity samples. Baselines are provided for one (OCV-1), two (OCV-2), three (OCV-3), and four (OCV) CV estimators. Both W-MLMC and MFMC converge to the OCV-1 baseline with W-MLMC performing slightly better. MLMC does not converge to the same baseline. None of the estimators converge to the optimal control variate shown in red.

discussion to follow: overcome the sub-optimality of the existing estimators by developing algorithms that can converge to the optimal control variate estimator (red OCV line in Figure 2.2).

3. Approximate CV. In this section we use Propositions 2.1 and 2.2 to derive approximate CV (ACV) estimators that converge to the known-mean optimal control variate (OCV) estimators with increasing data.

3.1. Two convergent estimators. The most straightforward way to guarantee convergence of an ACV estimator to an OCV estimator is to set $\boldsymbol{z}_i^1 = \boldsymbol{z}$ for each CV and then to use all available samples to estimate $\hat{\mu}_i$, i.e., $\boldsymbol{z}_i^2 = \boldsymbol{z}_i$. In other words, we use the *exact same input samples for $\hat{Q}_i$ as those used for $\hat{Q}$; any additional samples available for each control variate to evaluate $\hat{\mu}_i$.*

The formal definition of this estimator, which we coin approximate CV-IS (ACV-IS) where IS stands for "independent samples", is

DEFINITION 3.1 (ACV-IS). *Let $\boldsymbol{z}_i^1 = \boldsymbol{z}$, $\boldsymbol{z}_i^2 = \boldsymbol{z}_i$ and $(\boldsymbol{z}_i \setminus \boldsymbol{z}_i^1) \cap (\boldsymbol{z}_j \setminus \boldsymbol{z}_j^1) = \emptyset$ for $i \neq j$ and $i = 1, \ldots, M$. Then the estimator ACV-IS is defined as*

$$\hat{Q}^{\text{ACV-}IS}(\underline{\alpha}, \underline{\boldsymbol{z}}) = \hat{Q}(\boldsymbol{z}) + \sum_{i=1}^{M} \alpha_i \left(\hat{Q}_i(\boldsymbol{z}) - \hat{\mu}_i(\boldsymbol{z}_i) \right). \tag{3.1}$$

The estimator only requires shared evaluations for the N input samples that make up $\hat{Q}$ and each $\hat{Q}_i$. The rest of the samples $\boldsymbol{z}_i \setminus \boldsymbol{z}_i^1$ for each control variate are completely independent. As a result, an attractive feature of the sample distribution strategy for ACV-IS is that each control variate can be evaluated separately and in parallel. The sampling scheme for the ACV-IS estimator is illustrated in Figure 3.1a for reference.

The optimal control variate weights and variance reduction ratio for ACV-IS are now given.[5]

THEOREM 3.2 (Optimal CV-weights and variance reduction for ACV-IS). *The optimal CV weights and resulting estimator variance for the ACV-IS estimator are*

$$\underline{\alpha}^{\text{ACV-}IS} = -\left[\boldsymbol{C} \circ \boldsymbol{F}^{(IS)}\right]^{-1} \left[\text{diag}\left(\boldsymbol{F}^{(IS)}\right) \circ \boldsymbol{c}\right], \tag{3.2}$$

and

$$\mathbb{V}\text{ar}\left[\hat{Q}^{\text{ACV-}IS}(\underline{\alpha}^{\text{ACV-}IS})\right] = \frac{\mathbb{V}\text{ar}[Q]}{N}\left(1 - R^2_{\text{ACV-}IS}\right), \tag{3.3}$$

[5] In this paper, $\circ$ denotes a Hadamard, or elementwise, product.

where

$$R^2_{\text{ACV-}IS} = \left[\text{diag}\left(\boldsymbol{F}^{(IS)}\right) \circ \bar{\boldsymbol{c}}\right]^T \left[\boldsymbol{C} \circ \boldsymbol{F}^{(IS)}\right]^{-1} \left[\text{diag}\left(\boldsymbol{F}^{(IS)}\right) \circ \bar{\boldsymbol{c}}\right], \tag{3.4}$$

and $\boldsymbol{F}^{(IS)} \in \mathbb{R}^{M\times M}$ *such that*

$$\boldsymbol{F}^{(IS)}{}_{ij} = \begin{cases} \frac{r_i-1}{r_i}\frac{r_j-1}{r_j} & \text{if } i \neq j \\ \frac{r_i-1}{r_i} & \text{otherwise} \end{cases}. \tag{3.5}$$

The proof is provided in Appendix F.

Next we consider the case where the samples in the sets $\boldsymbol{z}_i \setminus \boldsymbol{z}_i^1$ are not independent among models, but rather use overlapping sets. We call this estimator ACV-MF.

Definition 3.3 (ACV-MF). *Let* $\boldsymbol{z}_i^1 = \boldsymbol{z}$, $\boldsymbol{z}_i^2 = \boldsymbol{z}_i$ *and* $\boldsymbol{z}_j^{(k)} = \boldsymbol{z}_i^{(k)}$ *for* $j > i$ *and* $k \leq \min(r_i, r_j)N$. *Then ACV-MF estimator is*

$$\hat{Q}^{\text{ACV-}MF}(\underline{\alpha}, \underline{\boldsymbol{z}}) = \hat{Q}(\boldsymbol{z}) + \sum_{i=1}^{M} \alpha_i \left(\hat{Q}_i(\boldsymbol{z}) - \hat{\mu}_i(\boldsymbol{z}_i)\right). \tag{3.6}$$

Note that this estimator can use an identical set of samples to the MFMC estimator and can thus be considered a drop-in replacement. *The only difference is that* $\hat{Q}_i$ *is evaluated using only the first* N *samples instead of* $r_{i-1}N$. Furthermore, using less samples for $\hat{Q}_i$ does not cause loss of accuracy because the control variates method does not require an accurate estimate of $\hat{Q}_i$ in terms of how close it is to μ_i, it requires an estimator $\hat{Q}_i$ that is *correlated to* $\hat{Q}$ and unbiased. The sampling scheme for the ACV-MF estimator is illustrated in Figure 3.1b for reference.

The optimal weights and resulting variance reduction of the ACV-MF estimator are now provided.

Theorem 3.4 (Optimal CV-weights and variance reduction for ACV-MF). *The optimal CV weights and resulting estimator variance for the ACV-MF estimator are*

$$\underline{\alpha}^{\text{ACV-}MF} = -\left[\boldsymbol{C} \circ \boldsymbol{F}^{(MF)}\right]^{-1} \left[\text{diag}\left(\boldsymbol{F}^{(MF)}\right) \circ \boldsymbol{c}\right], \tag{3.7}$$

and

$$\mathbb{V}\text{ar}\left[\hat{Q}^{\text{ACV-}MF}(\underline{\alpha}^{\text{ACV-}MF})\right] = \frac{\mathbb{V}\text{ar}[Q]}{N}\left(1 - R^2_{\text{ACV-}MF}\right), \tag{3.8}$$

where

$$R^2_{\text{ACV-}MF} = \left[\text{diag}\left(\boldsymbol{F}^{(MF)}\right) \circ \bar{\boldsymbol{c}}\right]^T \left[\boldsymbol{C} \circ \boldsymbol{F}^{(MF)}\right]^{-1} \left[\text{diag}\left(\boldsymbol{F}^{(MF)}\right) \circ \bar{\boldsymbol{c}}\right], \tag{3.9}$$

and $\boldsymbol{F}^{(MF)} \in \mathbb{R}^{M\times M}$ *such that*

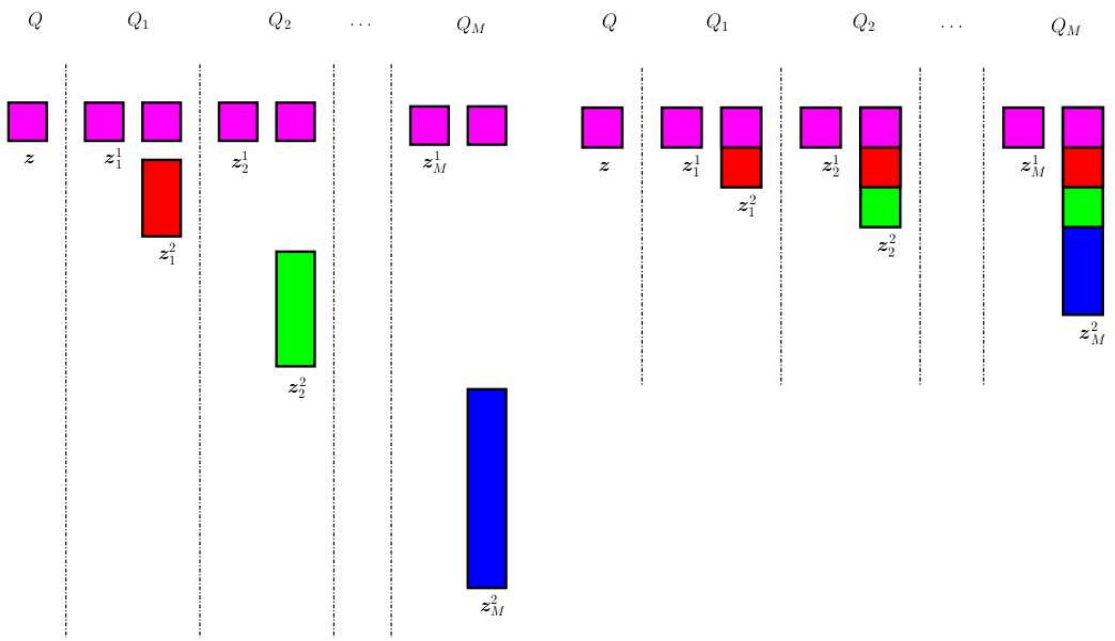
$$\boldsymbol{F}^{(MF)}{}_{ij} = \begin{cases} \frac{\min(r_i,r_j)-1}{\min(r_i,r_j)} & \text{if } i \neq j \\ \frac{r_i-1}{r_i} & \text{otherwise} \end{cases}. \tag{3.10}$$

The proof is provided in Appendix G. A visual illustration of thee above two schemes is given in Figure 3.1.

Interestingly, the form of the optimal estimators for ACV-IS and ACV-MF only differ in terms of the matrices $\boldsymbol{F}^{(IS)}$ and $\boldsymbol{F}^{(MF)}$. The way in which these matrices enter is algebraically identical, and they differ only in the $\boldsymbol{C} \circ \boldsymbol{G}$ term since their diagonals are identical. We analyze the conditions under which this algebraic form converges to the optimal CV.

Proposition 3.5 (Convergent estimators). *If an approximate CV estimator with* $\underline{r} = [r_1, \ldots, r_M]$ *yields an optimal variance reduction with*

$$R^2_{\boldsymbol{G}(\underline{r})} = \left[\text{diag}(\boldsymbol{G}) \circ \boldsymbol{c}\right]^T \frac{\left[\boldsymbol{C} \circ \boldsymbol{G}\right]^{-1}}{\mathbb{V}\text{ar}[Q]} \left[\text{diag}(\boldsymbol{G}) \circ \boldsymbol{c}\right], \tag{3.11}$$



(a) ACV-IS sampling strategy.

(b) ACV-MF sampling strategy.

Figure 3.1: Comparison of ACV-IS and ACV-MF sampling structures.

and $\boldsymbol{G}(\underline{r}) \to \mathbf{1}_{M\times M}$ *then the estimator converges to the optimal CV*

$$\lim_{\underline{r}\to\infty} R^2_{\boldsymbol{G}(\underline{r})} = R^2, \tag{3.12}$$

where $\underline{r} = [r_1, \ldots, r_M]$ *and* $\underline{r} \to \infty$ *means that* $r_i \to \infty$ *for* $i = 1, \ldots, M$..

The proof of this proposition is self evident since $\boldsymbol{c} \circ \mathbf{1}_M = \boldsymbol{c}$ and $\boldsymbol{C} \circ \mathbf{1}_{M\times M} = \boldsymbol{C}$.

THEOREM 3.6 (Convergence of ACV-IS and ACV-MF estimators to the optimal CV.). *The variance reduction of ACV-IS and ACV-MF converges to that of the optimal CV with increasing data.*

$$\lim_{\underline{r}\to\infty} R^2_{\text{ACV-}IS} = \lim_{\underline{r}\to\infty} R^2_{\text{ACV-}MF} = R^2. \tag{3.13}$$

The proof is provided in Appendix H.

We now return to the monomial example of Section 2.5. In Figure 3.2, we add the proposed estimators ACV-IS and ACV-MF. As expected, these estimators converge to the OCV baseline. They also converge with a similar rate and magnitude. The choice of estimator, in this case, may depend on which sampling strategy is more suitable for a given computing environment.

When the ratios r_i are small (left side of Figure 3.2), we are in a regime where the variance reduction is less than the single optimal CV. In this regime, the recursive estimators (MLMC, W-MLMC, and MFMC) perform better because their goal is to accelerate convergence to the single optimal CV. This suggests that a hybrid approach may be possible where, for a given set of r_i, we determine what is the "maximum" attainable variance reduction and then use recursion to accelerate the estimator to this optimum. Such a hybrid approach approach is described in the next section.

To summarize, the key features of ACV-IS and ACV-MF, are given in Table 3.1 and Figure 3.1.

3.2. Accelerating the approximate CV. Sections 2.3 and 2.4 show that a recursive CV estimator limits the maximum achievable variance reduction. However these strategies may be useful to accelerate convergence to a target CV level in cases where there is not enough data from higher-fidelity models to

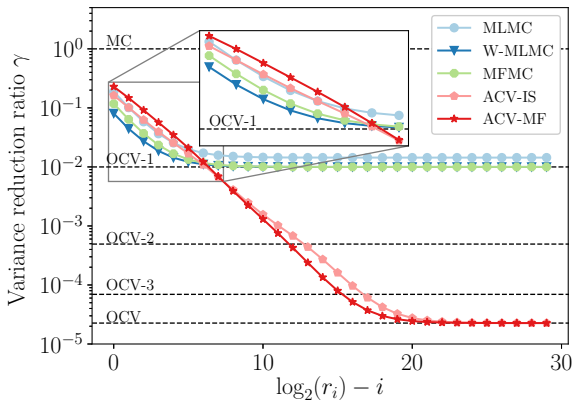


Figure 3.2: Comparison of variance reduction of ACV-IS and ACV-MF with MLMC, W-MLMC, and MFMC on the example from Section 2.5. ACV-IS and ACV-MF are the only ones to converge to OCV.

Algorithm	Relation between $\boldsymbol{z}$ and $\boldsymbol{z}_i$	$\boldsymbol{z}_i^1$	$\boldsymbol{z}_i^2$	Reduction Ratio γ
ACV-IS	$\boldsymbol{z} \cap \boldsymbol{z}_i = \boldsymbol{z}_i^1$, $\boldsymbol{z}_i^2 \cap \boldsymbol{z}_j^2 = \emptyset$ for $1 \leq i \neq j$	$\boldsymbol{z}$	$\boldsymbol{z}_i \setminus \boldsymbol{z}_i^1$	$1 - R^2_{\text{ACV-}IS}$
ACV-MF	$\boldsymbol{z} \cap (\boldsymbol{z}_i \setminus \boldsymbol{z}_i^1) = \emptyset$	$\boldsymbol{z}$	$\boldsymbol{z}_i$	$1 - R^2_{\text{ACV-}MF}$

Table 3.1: Summary of the two convergent estimators ACV-IS and ACV-MF. The main difference between the two estimators is that the sets $\boldsymbol{z}_i^2$ are independent between two distinct models for ACV-IS, whereas we have $\boldsymbol{z}_{i+1}^2 \cap \boldsymbol{z}_i^2 = \boldsymbol{z}_i^2$ for ACV-MF. ACV-MF is closely related to MFMC; the main difference is that only N samples are used for $\boldsymbol{z}_i^1$. The variance reduction of both ACV-IS and ACV-MF *in the limit of infinite samples of* Q_i is greater than or equal to ρ_1^2 ($R^2_{\text{ACV-}IS} \geq \rho_1^2$ and $R^2_{\text{ACV-}MF} \geq \rho_1^2$). For reference, the reverse is true for MLMC and MFMC.

achieve the targeted performance directly. This targeted level could be the single or two-level optimal CV (OCV-1 and OCV-2), or any other level up to OCV (all available CVs). The recursive techniques we have discussed all had sampling strategies that accelerated the convergence of $\hat{\mu}_i$ by CV Q_{i+1}, and as such was limited to convergence to OCV-1. Nevertheless, these algorithms converged to OCV-1 faster than either ACV-IS or ACV-MF did. In this section we seek to develop an algorithm to show how our framework can be used to create new schemes that accelerate to any given target. Our proposed algorithm is only one realization of a myriad of possible approaches. Our intention is to provide an example of how the framework can be used to achieve such an objective.

The scheme is conceptually very simple. We first partition all of the control variates into two groups; the first K variables form a K-level approximate control variate, and the last $M - K$ variables are used to reduce the variance of estimating μ_L for some $L \leq K$. The resulting estimator accelerates quickly to OCV-K, and L provides a degree of freedom for targeting a control variate level that contributes the greatest to the estimator variance.

DEFINITION 3.7 (ACV-KL: accelerated CV). *Let $K, L \leq M$ and $K \in \mathbb{Z}_+$ with $0 \leq L \leq K$. The ACV-KL estimator is given by the following expression and sampling scheme*

$$\hat{Q}^{\text{ACV-KL}}(\underline{\alpha}, \underline{\boldsymbol{z}}) = \hat{Q}(\boldsymbol{z}) + \sum_{i=1}^{K} \alpha_i \left(\hat{Q}_i(\boldsymbol{z}) - \hat{\mu}_i(\boldsymbol{z}_i) \right) + \sum_{i=K+1}^{M} \alpha_i \left(\hat{Q}_i(\boldsymbol{z}_L) - \hat{\mu}_i(\boldsymbol{z}_i) \right), \tag{3.14}$$

where we used that $\boldsymbol{z}_i^1 = \boldsymbol{z}$ for $i \leq K$ and $\boldsymbol{z}_i^1 = \boldsymbol{z}_L$ for $i > K$. Furthermore $\boldsymbol{z}_i^2 = \boldsymbol{z}_i$ for all i. The sets $\boldsymbol{z}_i \setminus \boldsymbol{z}_i^1$

can be chosen in several ways. Here we choose the same sampling strategy as ACV-MF:[6] $\boldsymbol{z}_i^1 = \boldsymbol{z}$, $\boldsymbol{z}_i^2 = \boldsymbol{z}_i$ *and* $\boldsymbol{z}_j^{(k)} = \boldsymbol{z}_i^{(k)}$ *for* $j > i$ *and* $k \leq \min(r_i, r_j)N$.

To identify the manner in which this CV estimator differs from previous recursive estimators, consider Eq. (3.14). The first two sums correspond to an ACV-MF estimator with K CVs and the last term adds a CV scheme to the ACV-MF estimator, i.e.

$$\hat{Q}^{\text{ACV-KL}}(\underline{\alpha}, \underline{\boldsymbol{z}}) = \hat{Q}_K^{\text{ACV-}MF}(\alpha_1, \ldots, \alpha_K, \boldsymbol{z}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_k) + \sum_{i=K+1}^{M} \alpha_i \left(\hat{Q}_i\left(\boldsymbol{z}_L\right) - \hat{\mu}_i\left(\boldsymbol{z}_i\right) \right) \tag{3.15}$$

The inclusion of the ACV-MF estimator enables the ACV-KL estimator to converge to the OCV estimator and the last term reduces the variance of $\hat{\mu}_L$, thereby accelerating convergence of the scheme. The optimal weights and variance reduction for the ACV-KL estimator are now provided.

THEOREM 3.8 (Optimal CV-weights and variance reduction for ACV-KL). *Assume* $r_i > r_L$ *for* $i > L$, *then the optimal weights for the ACV-KL control variate are*

$$\underline{\alpha}^{\text{ACV-KL}}(K, L) = -\left[\boldsymbol{C} \circ \boldsymbol{F}^{(K,L)}\right]^{-1} \left[\text{diag}\left(\boldsymbol{F}^{(K,L)}\right) \circ \boldsymbol{c}\right], \tag{3.16}$$

and the estimator variance is

$$\mathbb{V}\text{ar}\left[\hat{Q}^{\text{ACV-KL}}(\underline{\alpha}^{\text{ACV-KL}}(K, L))\right] = \frac{\mathbb{V}\text{ar}\left[Q\right]}{N}\left(1 - R^2_{\text{ACV-KL}}(K, L)\right), \tag{3.17}$$

where

$$R^2_{\text{ACV-KL}}(K, L) = \left[\text{diag}\left(\boldsymbol{F}^{(K,L)}\right) \circ \boldsymbol{c}\right]^T \frac{\left[\boldsymbol{C} \circ \boldsymbol{F}^{(K,L)}\right]^{-1}}{\mathbb{V}\text{ar}\left[Q\right]} \left[\text{diag}\left(\boldsymbol{F}^{(K,L)}\right) \circ \boldsymbol{c}\right], \tag{3.18}$$

and $\boldsymbol{F}^{(K,L)} \in \mathbb{R}^{M \times M}$ *such that*

$$\boldsymbol{F}_{ij}^{(K,L)} = \begin{cases} \frac{\min(r_i, r_j) - 1}{\min(r_i, r_j)} & \text{if } i, j \leq K \\ \frac{(r_i - r_L)(r_j - r_L) + r_L(\min(r_i, r_j) - r_L)}{r_i r_j r_L} & \text{if } i, j > K \\ \left[\frac{r_i - r_L}{r_i r_L}\right] & \text{if } L < i \leq K,\ j > K \\ \left[\frac{r_j - r_L}{r_j r_L}\right] & \text{if } L < j \leq K,\ i > K \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } i \neq j. \tag{3.19}$$

The diagonal elements are $\boldsymbol{F}_{ii}^{(K,L)} = \frac{r_i - 1}{r_i}$ *if* $i \leq K$ *and* $\boldsymbol{F}_{ii}^{(K,L)} = \frac{r_i - r_L}{r_i r_L}$ *otherwise.*

The proof, provided in Appendix I, is constructive in that it provides an explicit expression for the variance reduction and resulting estimator. Algorithm 3.1 provides pseudocode that summarizes this procedure. The algorithm requires (an estimate of) the covariances $\boldsymbol{C}$ and $\boldsymbol{c}$ and the variance of Q. Using these quantities, the algorithm provides an (approximate) optimal control variate weight and associated variance reduction for any given sample sizes defined by the ratios r_i and the parameters K and L. Note that, for this algorithm, $\boldsymbol{F}^{(K,L)}$ does not converge to $\boldsymbol{1}_{M \times M}$ as the number of samples goes to infinity, unless $K = L = M$ for which ACV-KL becomes ACV-MF. Furthermore, since this ACV-KL approach generalizes the ACV-MF sampling strategy, it is also a drop-in replacement of MFMC.

The convergence of the ACV-KL estimator for various (K, L) parameters is shown in Figure 3.3. The plot highlights that the ACV-MF estimator can be accelerated to any target baseline level, outperforming the baseline ACV-MF algorithm. Furthermore, the results demonstrate that it is possible for ACV-KL to achieve similar performance to the fully recursive algorithm in the low r_i region.

[6] Note that the ACV-IS sampling strategy for $\boldsymbol{z}_i^2$ could also have been chosen, but we do not analyze this approach here. Our aim is to demonstrate a basic framework for deriving approximate CV estimators, and many combinations are possible. We have chosen representative realizations of the framework to convey the main concepts.

Algorithm 3.1 Approximate Control Variate (ACV-KL) weights and variance reduction

Require: $\boldsymbol{C} \in \mathbb{R}^{M\times M}$: estimate of covariance amongst control variates; $\boldsymbol{c} \in \mathbb{R}^{M}$: estimate of covariance between Q and each CV; V: estimate of the variance of the Q; $(r_i)_{i=1}^{M}$: ratio of the number of evaluations of Q_i to Q; K and L: algorithm parameters where where $K \leq M$ and $L \leq K$.

Ensure: $\underline{\alpha}, R_{\text{ACV-KL}}$: optimal weights and estimated variance reduction for the ACV-KL

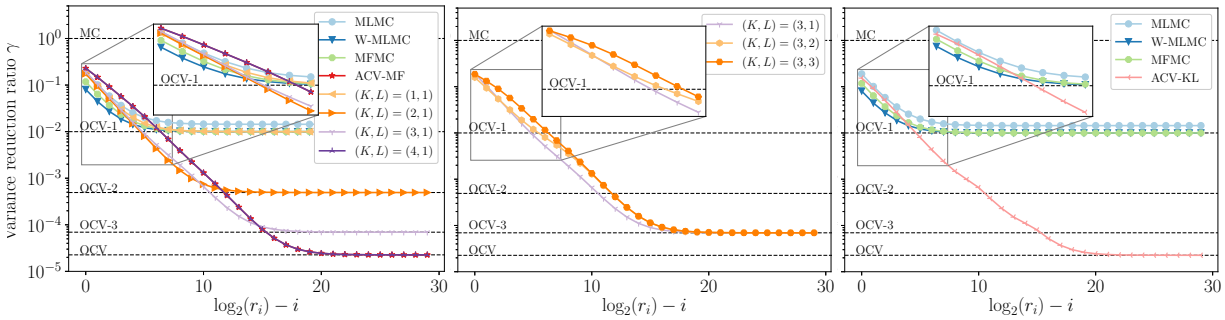
$\boldsymbol{A} = \text{zeros}(M, M)$
$\boldsymbol{b} = \text{zeros}(M)$
for $i = 1, \ldots, M$ **do**
  if $i \leq K$ **then**
    $\boldsymbol{A}_{ii} = \frac{r_i - 1}{r_i}\boldsymbol{C}_{ii}$
  else
    $\boldsymbol{A}_{ii} = \frac{r_i - r_L}{r_i r_L}\boldsymbol{C}_{ii}$
  end if
  $\boldsymbol{b}_i = \boldsymbol{A}_{ii}\boldsymbol{c}_i$
  for $j = i + 1, \ldots, M$ **do**
    if $i \leq K$ and $j \leq K$ **then**
      $\boldsymbol{A}_{ij} = \left[\frac{\min(r_i, r_j) - 1}{\min(r_i, r_j)}\right]\boldsymbol{C}_{ij}$
    else if $i > K$ and $j > K$ **then**
      $\boldsymbol{A}_{ij} = \left[\frac{(r_i - r_L)(r_j - r_L) + r_L(\min(r_i, r_j) - r_L)}{r_i r_j r_L}\right]\boldsymbol{C}_{ij}$
    else if $L < i \leq K$ and $j > K$ **then**
      $\boldsymbol{A}_{ij} = \left[\frac{r_i - r_L}{r_i r_L}\right]\boldsymbol{C}_{ij}$
    end if
    $\boldsymbol{A}_{ji} = \boldsymbol{A}_{ij}$
  end for
end for
$\underline{\alpha} = -\boldsymbol{A}^{-1}\boldsymbol{b}$
$R^2_{\text{ACV-KL}} = \boldsymbol{b}^T\boldsymbol{A}^{-1}\boldsymbol{b}/V$

The best choice of K and L is problem dependent; however, they can be estimated at negligible cost. Specifically, we can embed Algorithm 3.1 inside an outer loop that, for a given evaluation strategy, searches over all combinations of parameters K and L to minimize the variance. This approach would essentially follow the lowest-variance line of the (K, L) options shown in the combined left and middle panels of Figure 3.3. Figure 3.3c shows the performance of the ACV-KL estimator that chooses the best (K, L) combination. As a result of these discrete choices, it has "kinks" as increasing sample sizes lead to different combinations.

3.3. Sample allocation. As mentioned previously, there are two ways to achieve variance reduction within a control variate approach: (1) increasing the number of high fidelity evaluations and (2) leveraging the correlations of the low-fidelity models to the greatest extent possible. Thus far, we have emphasized the second approach, and in this section, we consider the overall distribution of the samples.

3.3.1. Optimization setup. We will seek to minimize the estimator variance subject to a constraint on the total cost. The form of the MLMC and MFMC estimators enable analytic closed-form solutions to similar minimization problems, and their optimal allocation strategies can be found in [8] and [26], respectively. In our case the W-MLMC, ACV-MF, and ACV-KL estimators do not yield analytic solutions, as far as we are aware, because of the complex inversion of $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$. In this paper, we rely on optimization approaches.

Let $J_{ACV}(N, \underline{r}) = \left(1 - R^2_{ACV}\right)\frac{\mathbb{C}\text{ov}[Q]}{N}$ denote the objective function for some approximate control variate, where the expressions for R^2_{ACV} are dependent on the ACV type. We seek to minimize this objective subject



(a) Varying K with $L=1$ (b) Varying L fixed $K=3$ (c) Best combination of (K,L).

Figure 3.3: Accelerated convergence to target levels by the ACV-KL class of estimators for various (K,L) combinations on the monomial example. The kinks in optimal (K,L) this plot are indicative of transitions between optimal (K,L) combinations

to an inequality constraint on the cost and linear and bound constraints on the sampling design parameters:

$$\min_{N,\underline{r},K,L} \log(J_{ACV}(N,\underline{r},K,L)) \quad \text{subject to}$$

$$N\left(w+\sum_{i=1}^{M} w_i r_i\right) \le C, \quad N \ge 1, \quad r_1 \ge 1, \text{ and } r_i \ge r_{i-1} \quad \text{for } i=2,\ldots,M \tag{3.20}$$

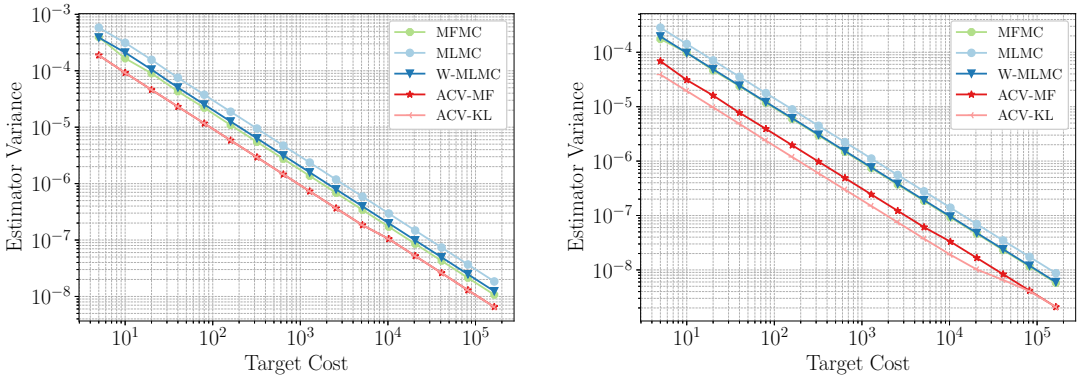
where we recall that w and w_i denote the cost of obtaining a sample of Q and Q_i. Note that for this problem to be well posed we require $C \ge w+\sum_{i=1}^{M} w_i$, i.e., that the cost allowable is larger than that which corresponds to evaluating the quantity of interest and each control variates a single time.

In terms of implementation, we use a local gradient-based optimization procedure, in particular the interior point method SLSQP, to optimize over N and r_i. We use the automatic differentiation tool of the python PyTorch[7] library to obtain gradients of the objective with respect to the optimization variables. A couple of comments are in order:

1. We minimize the log of the cost function since we have empirically found that it yields a more well-conditioned objective function without affecting the solution.
2. The covariances among the QoI and the control variates are not known, rather they are estimated from pilot samples; these pilot samples are considered offline and their expense is not included in the cost constraint.
3. We enforce $r_i \ge r_{i-1}$, which is only necessary for the W-MLMC ACV. The ACV-MF and ACV-KL estimators do not strictly require this constraint; however, in practice we have found that it yields consistent solutions.
4. This is a mixed-integer nonlinear program (MINLP) where N is a non-categorical integer variable, $r_i N$ are derived integer quantities, and, in the case of ACV-KL, (K,L) are categorical integer variables. We address the categorical variables through enumeration, i.e., by optimizing for each possible (K,L) pair and choosing the estimator with the best objective function. We address the non-categorical variable by simple relaxation, i.e. by optimizing for continuous N and r_i and rounding the resulting sample counts for each model to the nearest integer, ignoring any induced infeasibility in the cost constraint.

We leave the development of additional optimization formulations for future work. Our goal here is to obtain a baseline sample allocation procedure for use in studying the estimators.

[7] https://pytorch.org



(a) $w = 1$ and $w_i = 10^{-i}$ for $i = 1, 2, 3, 4$.

(b) $w = 1$ and $w_i = 10^{-i-1}$ for $i = 1, 2, 3, 4$.

Figure 4.1: Variance reduction for the motivating example from Section 2.5 under an optimal allocation strategy for each estimator. MLMC performs slightly worse than the two other recursive estimators. W-MLMC and MFMC perform virtually identically. The non-recursive ACV-KL estimator achieves significantly greater variance reduction. Right panel indicates greater reduction for greater cost difference.

4. Numerical Experiments. Now we consider several numerical experiments to demonstrate the results of the theory.

4.1. Model problem. First, we again consider the monomial example from Section 2.5. Recall, that the efficiency of our methods depend on the particular problem only through the correlation matrix and the costs. The correlation matrix in Table 2.2 is reasonable for what might be observed in practice. However, the costs have thus far been unspecified. Here we investigate the effects of various cost prescriptions.

In Figure 4.1, we plot the actual variance reduction for two different costs prescriptions. Both have $w = 1$, but they differ in the gap between the QoI and the first control variate; in the left panel we have $w_i = 10^{-i}$ and in the right panel we have $w_i = 10^{-i-1}$ for $i = 1, 2, 3, 4$. These results are based on solving (3.20) to select the number of samples assigned to each model. We see that all of the recursive estimators MLMC, W-MLMC, and MFMC perform virtually identically, while ACV-KL provides greater variance reduction.

The performance of all the sampling algorithms discussed in this paper are dependent on the relative cost of the models used. In Figure 4.2, we compare the variance reduction of W-MLMC and ACV-KL, when they are applied to the three monomial models, as we vary the costs of the two control variates. We constrain our search space to the case where the second control variate is less expensive to evaluate than the first, i.e., $w_2 < w_1$. Our goal is to demonstrate the scales at which the ACV-KL approach is able to provide significant variance reduction over recursive estimators. We only consider W-MLMC for this test because the results for other estimators are very similar (e.g., see Figure 4.1). Figure 4.2 demonstrates that ACV-KL is able to achieve significant performance gains over the recursive approaches when the cost of the first control variate is at least 100 times lower than the truth model ($w_1 < 0.01$). Above this threshold, MLMC performs better. In virtually all cases, ACV-KL performs better than regular Monte Carlo.

4.2. A parametric model problem. We now introduce a parametric model problem that enables us to quantify the performance of the different algorithms under several scenarios. We consider three two-dimensional functions; the first describes the high-fidelity quantity of interest and the next two serve as control variates.

$$
\begin{aligned}
Q &= A\left(\cos\theta\, x^5 + \sin\theta\, y^5\right), \\
Q_1 &= A_1\left(\cos\theta_1\, x^3 + \sin\theta_1\, y^3\right), \\
Q_2 &= A_2\left(\cos\theta_2\, x + \sin\theta_2\, y\right),
\end{aligned}
\tag{4.1}
$$

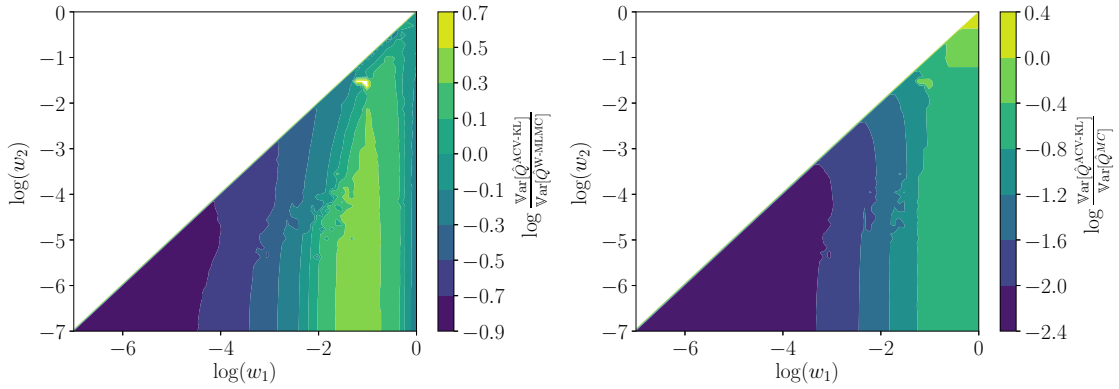


Figure 4.2: Ratio of variance reduction achieved by ACV-KL compared to W-MLMC (left) and Monte Carlo (right) in the case of three total models ($Q(z) = z^3$, $Q_1(z) = z^2$, and $Q_2(z) = z$) as a function of the costs (w_1, w_2) of evaluating the control variates. The cost of the qoi is fixed to $w = 1$.

	Q	Q_1	Q_2
Q	1	$AA_1/9\,(\sin\theta\sin\theta_1 + \cos\theta\cos\theta_1)$	$AA_2/7\,(\sin\theta\sin\theta_2 + \cos\theta\cos\theta_2)$
Q_1	*sym*	1	$A_1A_2/5\,(\sin\theta_1\sin\theta_2 + \cos\theta_1\cos\theta_2)$
Q_2	*sym*	*sym*	1

Table 4.1: Correlation/covariance matrix for problem given by Equation (4.1) with $A = \sqrt{11}$, $A_1 = \sqrt{7}$, and $A_2 = \sqrt{3}$.

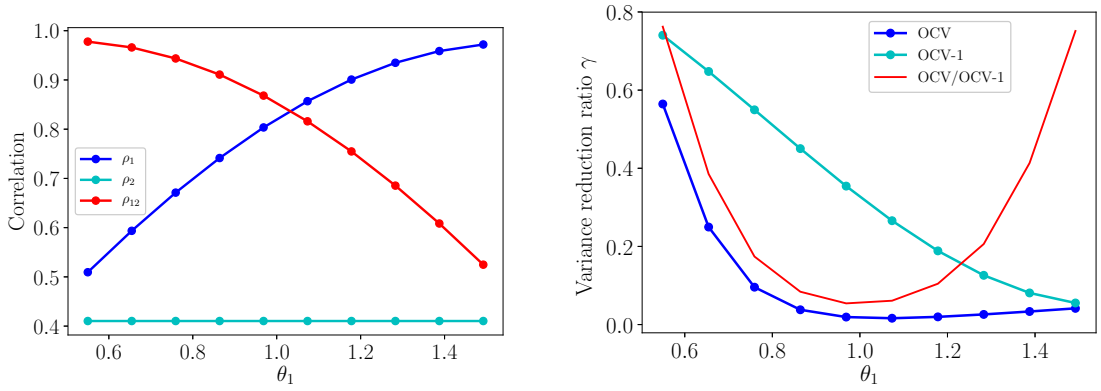
where $x, y \sim \mathcal{U}(-1, 1)$ and all A and θ coefficients are real. We choose to set $A = \sqrt{11}$, $A_1 = \sqrt{7}$ and $A_2 = \sqrt{3}$ to obtain unitary variance for each model. This choice of unit variance for each model reduces the number of degrees of freedom in the problem parameterization since the correlation and covariance matrices are identical. Specifically, the analytic correlation/covariance matrix is given in Table 4.1. To further reduce the number of degrees of freedom, we fix $\theta = \pi/2$ and $\theta_2 = \pi/6$ and let θ_1 vary uniformly in the bounds $\theta_2 < \theta_1 < \theta$. Each particular value of θ_1 induces a different correlation ρ_1 between Q_1 and Q and a different correlation ρ_{12} between Q_1 and Q_2, whereas the correlation ρ_2 between Q_2 and Q remains fixed. These correlations are reported for different settings of θ_1 in Figure 4.3a.

First we demonstrate that the variance reduction ratio of the control variate that considers both models OCV is larger than the one obtained using only a single control variate OCV-1. This ratio of the estimator variance of OCV-1 and OCV is reported in Figure 4.3b. In Figure 4.3b we also report the variance reduction obtained by OCV and OCV-1 with respect to MC. The greatest gap between OCV and OCV-1 occurs when θ_1 is approximately $\pi/3$ (the middle of its range).

Next we consider the effect of different cost relationships among the three models. For these purposes, we assign a relative cost of 1 for Q, $1/w$ for Q_1 and $1/w^2$ for Q_2. For an equivalent total cost of 100 runs of Q, we consider the effects of w on the performance of each estimator. These results are given in Figure 4.4.

Several interesting features are present in these results. First, the ACV-KL estimator generally outperforms all other estimators across the range of cost ratios enumerated in plots (a) through (e), and across the range of θ_1 from $\pi/6$ to $\pi/2$. The ACV-KL is nearly identical to ACV-MF since large deviations in (K, L) cannot occur when there are only 2 control variates. In most of the scenarios, these two estimators also outperforms OCV-1. Second, we see a gradual convergence of the recursive estimators to OCV-1 and optimal estimators (ACV-IS, ACV-MF, ACV-KL) to OCV as w increases. Qualitatively, the relationship of the performance of these two groups of estimators, with θ_1, is very similar to their OCV-1 and OCV counterparts. In particular MFMC, MLMC, and W-MLMC, seem to decay almost linearly, just like OCV-1; and ACV-IS, ACV-MF, and ACV-KL seem to plateau with a similar shape to OCV.

Next, we see that for small cost ratios (e.g., plot (a)), MLMC can perform worse than MC. The other



(a) Correlations

(b) Variance reduction ratios

Figure 4.3: Correlation (left) and variance reduction gap (right) for the parametric problem of Equation (4.1) with $\theta = \pi/2$, $\theta_2 = \pi/6$ and $\theta_2 < \theta_1 < \theta$.

estimators also approach MC for these small cost ratios when θ_1 is small. As expected, the correction of MLMC to obtain W-MLMC ensures that the W-MLMC estimator does not behave worse than MC. In virtually all cases, W-MLMC performs almost identically to MFMC, and this will continue in the PDE examples to follow. Finally, we see that the greatest advantage of our proposed estimators correspond to the cases of θ_1 for which there is the greatest gap between OCV-1 and OCV. As this gap shrinks, our advantage decays.

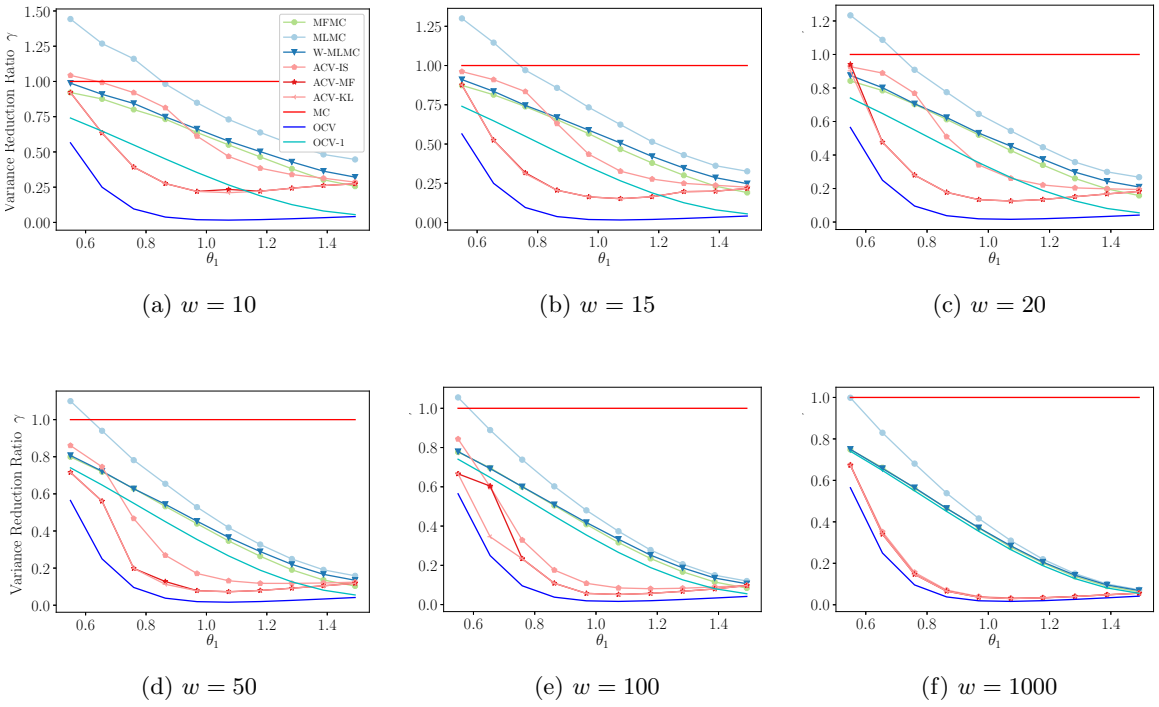
In Figure 4.5, we directly compare the ACV-KL estimator, representing the pursuit of OCV performance, with the MFMC estimator, representing the pursuit of OCV-1 performance, over the range of θ_1 for an expanded range of w values. For the central portion of this plot, the ACV-KL estimator outperforms MFMC. MFMC gains an advantage for the more extreme values of θ_1, and these are the same regions for which the OCV-1 and OCV gap is small according to Figure 4.3b.

4.3. Burgers equation. In our next example, we consider the viscous Burger equation utilizing a definition similar to that described by Peherstorfer and colleagues [25]. Define $\Omega = [-1, 1] \subset \mathbb{R}$ as the spatial domain, $T \in \mathbb{R}$ as the end time, and $u : \Omega \times [0, T] \to \mathbb{R}$ as the time-dependent solution of

$$\frac{\partial u(x,t)}{\partial t} + u\frac{\partial u(x,t)}{\partial x} - \nu\frac{\partial^2 u(x,t)}{\partial x^2} = 0, \quad u(-1,t) = 1 + Z, \quad u(1,t) = -1 \quad t \in [0,T],\ Z \in \Omega.$$

We treat Z as a random variable whose distribution is uniform in $[10^{-2}, 10^{-1}]$, and we use a viscosity $\nu = 0.2$. Our quantity of interest is $Q[u] = u(0)$, and we obtain a sequence of models discretized on grids of sizes $(512, 256, 128, 64, 32)$, with the finest discretization corresponding to our high-fidelity truth model. This discretization hierarchy leads to costs $w = 1$ and $w_i = 2^{-i}$ for $i = 1, \ldots, M$. 1000 samples of each model were used to evaluate the statistics and to provide the optimal variance reduction for the single and full control variates. Using these samples, the correlation among models is shown in Figure 4.7a. Recall that all of our methods work directly with this correlation matrix, and the variances of each model. In other words, no other aspects of the original problem matter with respect to the variance reduction.

In Figure 4.6a we show that, for this problem, when there is a possibility of increasing the number high fidelity samples then all of the control variates (aside from MLMC) perform similarly. However, when there is no additional capability to evaluate the high fidelity model, then we can achieve a significant advantage. Interestingly, the un-weighted MLMC estimator actually performs worse than standard MC. In this problem, the variance does not immediately decay, which induces greater need for computing optimal weights in order to manage this non-monotonicity. Moreover the cost scales only linearly with the levels and the slow variance decay cannot be efficiently compensated. This scenario demonstrates how the simple introduction of the weights for MLMC may significantly improve the performance of the plain algorithm whenever the cost



(a) $w = 10$ (b) $w = 15$ (c) $w = 20$

(d) $w = 50$ (e) $w = 100$ (f) $w = 1000$

Figure 4.4: Variance reduction for cost ratios of $[1, 1/w, 1/w^2]$ for Q, Q_1, and Q_2.

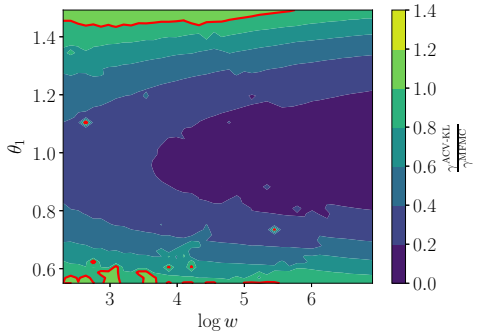
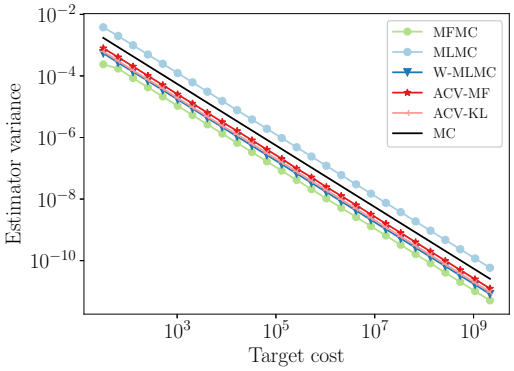


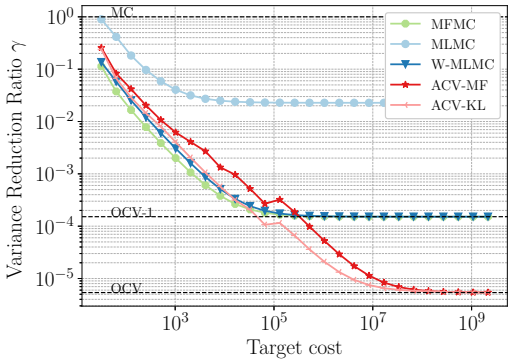
Figure 4.5: Comparison of the variance reduction between ACV-KL and MFMC over a range of θ_1 and w. ACV-KL outperforms MFMC over a wide range of settings. The red line indicates the level-1 contour where the estimators have equal performance.

separation between the models is not sufficient to make it competitive with respect to MC. The performance is similar amongst the optimally weighted CV schemes because it seems that limiting the variance reduction to a single control variate does not hurt because one can also increase the number of high-fidelity evaluations. To this end, we next study what happens where we cannot increase the number of high-fidelity evaluations.

In Figure 4.6b we demonstrate the variance reduction achieved by each estimator for a fixed number of high fidelity samples. Here we assume only 10 high fidelity samples are available and optimize the number of samples, using (3.20), assigned to each control variate. Our ACV estimators are able to achieve more than four orders of magnitude improvement over the recursive schemes. Remarkably, this indicates that we



(a) Except for MLMC, all ACV perform similarly.



(b) ACV-MF and ACV-KL achieve an order of magnitude greater variance reduction.

Figure 4.6: Comparison of estimators for viscous Burgers under optimal sample allocation strategies when the number of high fidelity samples can be increased(left) and it is fixed (right).

$$\begin{bmatrix} 1. & 0.9999 & 0.9977 & 0.9591 & 0.8540 \\ 0.9999 & 1. & 0.9983 & 0.9598 & 0.8528 \\ 0.9977 & 0.9983 & 1. & 0.9713 & 0.8656 \\ 0.9591 & 0.9598 & 0.9713 & 1. & 0.9400 \\ 0.8540 & 0.8528 & 0.8656 & 0.9400 & 1. \end{bmatrix}$$

(a) Correlation between models

Target cost	2^5	2^6	2^7	2^8	2^9	2^{28}
K	3	4	3	3	4	5
L	1	2	1	1	1	5

(b) The optimal values reach $(4, 1)$ quickly and converge to the full ACV-MF given by $(5, 5)$ after a cost of 2^{28}, or 2^{28} equivalent high fidelity evaluations.

Figure 4.7: Correlation between viscous Burgers models (left) and chosen (K, L) pairs during the fixed-high high-fidelity samples run (right).

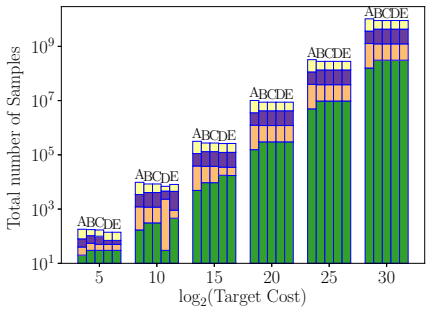
can reduce the variance of the estimator quite significantly *with no additional evaluations of the high fidelity model*. Such performance would be highly desirable in the scenario where high fidelity model evaluations were obtained on supercomputers with significant resources, and additional evaluations are unattainable. For the new algorithms we propose, ACV-KL significantly outperforms ACV-MF in the regime in the middle of the graph. Nevertheless, with enough data, these two estimators converge to the same performance, as expected. In Table 4.7b we show the optimal choices of (K, L) chosen at several target costs throughout the run.

Note that the classical MLMC scheme does not even converge to the optimal single CV estimator OCV-1. However, with the simple modification proposed in Section 2.3, our optimally-weighted MLMC scheme does converge to OCV-1 and moreover exhibits a convergence profile that is now competitive with MFMC.

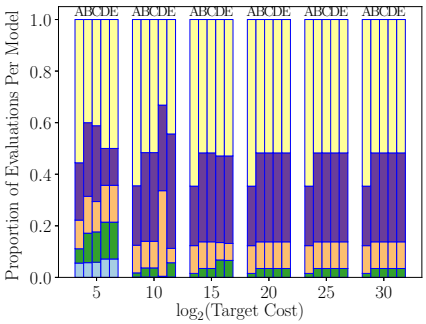
The sample distributions for certain target costs are shown in Figure 4.8. The distributions across levels for all of the estimators aside from MFMC are virtually identical as the target cost increases. However, at lower target costs $(2^5 - 2^{10})$, there is some variability in the sample distributions which eventually stabilizes.

4.4. Steady-state diffusion. Consider the steady-state diffusion equation in one spatial dimension

$$-\frac{d}{dx}\left[k(x, \underline{Z})\frac{du}{dx}(x, \underline{Z})\right] = 10, \quad u(0) = 0, \quad u(1) = 0, \quad (x, \underline{Z}) \in (0, 1) \times \Omega \tag{4.2}$$



(a) Cumulative sum (log scale) of evaluations per level



(b) Distribution of total samples per level.

Figure 4.8: Sample allocation for several target costs for the viscous Burgers problem. Each group of bars refers to MFMC (A), MLMC (B), W-MLMC (C), ACV-MF (D), and ACV-KL (E). The evaluations for each control variate are depicted in the colors green, salmon, purple, and yellow, in order of finest to coarsest discretization. The evaluation of the QoI is in blue. Note that the left panel uses a log scale in order to better visualize small sample sets (the green samples appear dominant at left, but are actually a small fraction of the total in the right panel).

where the random diffusivity k is a random field represented by the Karhunen-Loéve expansion expansion (KLE)

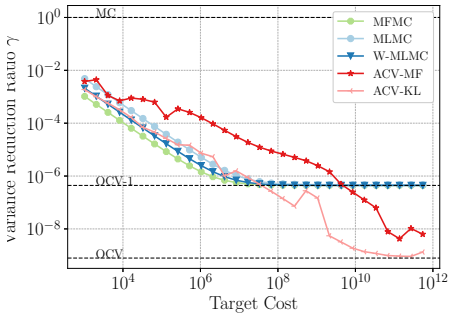
$$k(x,\underline{Z}) = \exp\left\{\bar{k}(x) + \sigma\sum_{k=1}^{d}\sqrt{\lambda_k}\phi_k(x)Z_k\right\} \qquad C(x_1,x_2) = \exp\left[-\frac{|x_1-x_2|}{l}\right], \tag{4.3}$$

where $\sigma = 10$. Here $\{\lambda_k\}_{k=1}^d$ and $\{\phi_k(x)\}_{k=1}^d$ are, respectively, the eigenvalues and eigenfunctions of the covariance kernel with correlation length set to $l = \frac{1}{100}$. The mean of the field is $\bar{k}(x) = \log(1 + \frac{1}{2}\cos(\frac{\pi}{2}x))$ and the KLE is truncated at $d = 100$ terms with each variable assumed to be an independently and identically distributed uniform random variable $Z_k \in [-1, 1]$, $k = 1, \ldots, d$.

We wish to determine the expected value of the quantity $Q(\underline{Z}) = u(x = \frac{1}{2}, \underline{Z})$. The diffusion equation (4.2) cannot be solved analytically, so we solve the equation numerically using central finite differences. Let Q be obtained from the numerical solution which uses 321 mesh points and let $h_{1,\alpha} = h_{1,0}2^{-\alpha}$ denote the mesh sizes in the spatial directions x used to compute each of the control variates $Q_i, i = 1\ldots5$, where $h_{1,0}$=0.1. The number of points in a spatial mesh is $n_{h_{1,i}} = h_{1,\alpha}^{-1} + 1$. Solving (4.2) requires solving the structured finite-difference linear system. We assume that the cost of solving this system grows linearly with total number of degrees of freedom $n_{h_{1,i}}$. The relative computational cost of, and the correlation (computed with 1000 samples)) between, each model approximation is given in Figure 4.9b.

In this problem we saw (results not shown for brevity) that optimizing over both the number of high fidelity samples and the ratios leads to similarly performing estimators. Instead, we show again what happens when we fix the number of high fidelity samples. The results are shown in Figure 4.9a. We assume only ten high fidelity samples are available and optimize the number of samples, using (3.20), assigned to each control variate. Again our ACV estimators are able to achieve at least two orders of magnitude improvement over the recursive schemes in the limit of computational resources. The distribution of the number of samples found by the optimizer is shown in Figure 4.10. The distributions across levels for all of the estimators, aside from MFMC, are virtually identical as the target cost increases. However, at lower target costs ($2^{10} - 2^{15}$), there is again some variability in the sample distributions that eventually stabilizes.

5. Conclusion. We have presented a framework that unifies existing control variate and multi-level Monte Carlo sampling methods. In doing so, we have proven that the structure of existing recursive estimators is fundamentally limited in the amount of information that is extracted from multifidelity data sources. Regardless of the number of control variates these recursive schemes use, the estimators they produce can



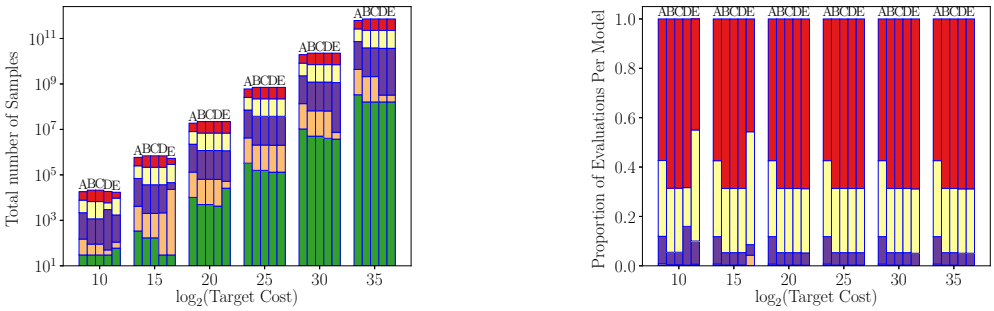
(a) Achievable variance reduction.

w_5	w_4	w_3	w_2	w_1	w
0.034	0.065	0.128	0.252	0.501	1

$$\begin{bmatrix} 1. & 0.9999 & 0.9999 & 0.9997 & 0.9569 & 0.7692 \\ 0.9999 & 1. & 0.9999 & 0.9997 & 0.9570 & 0.7691 \\ 0.9999 & 0.9999 & 1. & 0.9997 & 0.9572 & 0.7690 \\ 0.9997 & 0.9997 & 0.9997 & 1. & 0.9575 & 0.7677 \\ 0.9569 & 0.9570 & 0.9572 & 0.9575 & 1. & 0.6695 \\ 0.7692 & 0.7691 & 0.7690 & 0.7677 & 0.6695 & 1. \end{bmatrix}$$

(b) Relative computational cost and correlation matrix

Figure 4.9: Comparison of estimators (left) and relative computational cost and correlation matrix (right) for the diffusion problem.



(a) Cumulative sum (logscale) of evaluations per level

(b) Distribution of total samples per level

Figure 4.10: Sample allocation for several target costs for the steady state diffusion problem. Each group of bars refers to MFMC (A), MLMC (B), W-MLMC (C), ACV-MF (D), and ACV-KL (E). The evaluations for each control variate are depicted in the colors green, salmon, purple, yellow, red in order of finest to coarsest discretization. As for Figure 4.8, note that the left panel uses a log scale to better visualize small sample sets that are a very small fraction of the total and essentially disappear without log scale in the right panel.

only converge to the performance of the optimal control variate that uses a single model with known mean. This sub-optimal behavior can limit the efficiency and accuracy of these estimators, especially in applications where the ability to obtain additional samples of the high-fidelity models is restricted.

To overcome this sub-optimal behavior, we proposed new estimators that leverage all existing correlations among information sources. In other words, the formulations we propose break through the sub-optimality barrier of existing recursive approaches and converge to the optimal multi-model control variate. In a number of numerical examples, our proposed methods produce estimators that, in the computational resources for the control variates, have variance levels that are several orders of magnitude smaller than that of existing estimators.

In terms of numerical results, our monomial test problem shows strong benefit with significant variance reduction both in the case of increasing numbers of high-fidelity evaluations and with fixed-numbers of high fidelity evaluations. Although the monomial test problem is quite simplified, the correlation matrix that it produces is representative of what may be encountered in practice. Since our algorithm only requires access to this correlation (or covariance) matrix, this indicates that there are potentially significant benefits to using

our proposed approach. In order to gain even more insight regarding the relation between the performance of the schemes, the correlation matrix and the cost ratio between models, we introduce a parametric model problem that allows to explore a large variety of scenarios. In particular, this example shows that ultimately the performance of our approach are related to the existence of a gap between OCV and OCV-1 and that this gap can be numerically exploited under certain cost conditions. As this gap shrinks, the advantage of the strategies proposed here with respect to the other state-of-the-art-techniques diminish. In the two highly-correlated multi-model cases (viscous Burgers with $\rho_i > 0.99$ and the steady-state diffusion problem), it is shown that recursive approaches are generally adequate if one can generate evaluations of the high fidelity model since the first control variate reduces the bulk of the variance. However, if the computational expense between models is great, we can still achieve orders of magnitude improved reduction in the variance.

Future work will focus on developing optimal sample allocation schemes for the approximate control variate estimators. We conjecture that different optimization formulations can lead to better exploitation of the variance gaps between the single and multiple control variate estimators. These optimal sample allocation schemes will most likely require sub-selecting groups of control variates that should be further evaluated and bypassing those whose uncertainty cannot be sufficiently reduced.

Within our framework, other estimators can also be developed. In particular, optimal groupings of control variates can be investigated. Our ACV-KL estimator uses a particular scheme, where the last $(M - K)$ control variates enhance the convergence of the Lth mean. Other groupings must exist that can potentially provide greater variance reduction, and the development of automated ways of determining such groupings is required.

6. Acknowledgements. The authors thank Dr. Laura Swiler and Dr. Tim Wildey from Sandia National Laboratories for their insightful comments and suggestions regarding an earlier version of this manuscript. This work was fully supported by the DARPA EQUiPS project and partially supported by the DOE SciDAC Advanced Scientific Computing Research (ASCR) program. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

REFERENCES

[1] J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. Control variates for stochastic gradient mcmc. *Statistics and Computing*, Aug 2018.

[2] A. Doostan, G. Geraci, and G. Iaccarino. A bi-fidelity approach for uncertainty quantification of heat transfer in a rectangular ribbed channel. In *ASME. Turbo Expo: Power for Land, Sea, and Air, Volume 2C: Turbomachinery*. ASME, 2016.

[3] H. Fairbanks, A. Doostan, C. Ketelsen, and G. Iaccarino. A low-rank control variate for multilevel monte carlo simulation of high-dimensional uncertain systems. *Journal of Computational Physics*, 341:121–139, 2017.

[4] C. M. Fleeter, G. Geraci, D. E. Schiavazzi, A. M. Kahn, M. S. Eldred, and A. L. Marsden. Multilevel multifidelity approaches for cardiovascular flow under uncertainty. In *Sandia Center for Computing Research Summer Proceedings 2017, A.D. Baczewski and M.L. Parks, eds.*, volume Technical Report SAND2018-2780O, pages 27–50. Sandia National Laboratories, 2018.

[5] G. Geraci, M. Eldred, and G. Iaccarino. A multifidelity control variate approach for the multilevel monte carlo technique. In *Center for Turbulence Research Annual Research Briefs*, pages 169–181. Center for Turbulence Research, Stanford University, 2015.

[6] G. Geraci, M. S. Eldred, and G. Iaccarino. A multifidelity multilevel monte carlo method for uncertainty propagation in aerospace applications. In *19th AIAA Non-Deterministic Approaches Conference*, page 1951, 2017.

[7] M. Giles and L. Szpruch. Multilevel monte carlo methods for applications in finance. In *Recent Developments in Computational Finance*. World Scientific / Imperial College Press, 2013.

[8] M. B. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

[9] M. B. Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015.

[10] M. B. Giles and T. Goda. Decision-making under uncertainty: using MLMC for efficient estimation of EVPPI. *Statistics and Computing*, Sep 2018.

[11] J. Goh, D. Bingham, J. P. Holloway, M. J. Grosskopf, C. C. Kuranz, and E. Rutter. Prediction and computer model calibration using outputs from multifidelity simulators. *Technometrics*, 55(4):501–512, 2013.

[12] A.-L. Haji-Ali, F. Nobile, and R. Tempone. Multi-index Monte Carlo: when sparsity meets sampling. *Numerische Mathematik*, 132(4):767–806, Apr 2016.

[13] A.-L. Haji-Ali and R. Tempone. Multilevel and multi-index monte carlo methods for the McKean–Vlasov equation.

Statistics and Computing, 28(4):923–935, Jul 2018.
[14] T. Hesterberg. Control variates and importance sampling for efficient bootstrap simulations. *Statistics and Computing*, 6(2):147–157, Jun 1996.
[15] A. Jasra, K. J. H. Law, and P. P. Osei. Multilevel particle filters for lévy-driven stochastic differential equations. *Statistics and Computing*, Oct 2018.
[16] L. Jofre, G. Geraci, H. Fairbanks, A. Doostan, and G. Iaccarino. Multi-fidelity uncertainty quantification ofirradiated particle-laden turbulence. In *Center for Turbulence Research Annual Research Briefs*, pages 21–34. Center for Turbulence Research, Stanford University, 2017.
[17] Y. Kuya, K. Takeda, X. Zhang, and A. I. J. Forrester. Multifidelity surrogate modeling of experimental and computational aerodynamic data sets. *AIAA journal*, 49(2):289–298, 2011.
[18] S. Lavenberg, T. Moeller, and P. Welch. *Statistical results on multiple control variables with application to variance reduction in queueing network simulation*. IBM Thomas J. Watson Research Division, 1978.
[19] S. S. Lavenberg, T. L. Moeller, and P. D. Welch. Statistical results on control variables with application to queueing network simulation. *Operations Research*, 30(1):182–202, 1982.
[20] S. S. Lavenberg and P. D. Welch. A perspective on the use of control variables to increase the efficiency of monte carlo simulations. *Management Science*, 27(3):322–335, 1981.
[21] L. W.-T. Ng. *Multidelity approaches for design under uncertainty*. PhD thesis, Massachusetts Institute of Technology, 2013.
[22] L. W.-T. Ng and M. Eldred. Multifidelity uncertainty quantification using non-intrusive polynomial chaos and stochastic collocation. In *53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA*, page 1852, 2012.
[23] A. S. Padron, J. J. Alonso, F. Palacios, M. F. Barone, and M. S. Eldred. Multi-fidelity uncertainty quantification: application to a vertical axis wind turbine under an extreme gust. In *15th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, page 3013, 2014.
[24] R. Pasupathy, B. W. Schmeiser, M. R. Taaffe, and J. Wang. Control-variate estimation using estimated control means. *IIE Transactions*, 44(5):381–385, 2012.
[25] B. Peherstorfer, T. Cui, Y. Marzouk, and K. Willcox. Multifidelity importance sampling. *Computer Methods in Applied Mechanics and Engineering*, 300:490–509, 2016.
[26] B. Peherstorfer, K. Willcox, and M. Gunzburger. Optimal model management for multifidelity Monte Carlo estimation. *SIAM Journal on Scientific Computing*, 38(5):A3163–A3194, 2016.
[27] B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *Preprint*, pages 1–57, 2016.
[28] O. Roderick, M. Anitescu, and Y. Peet. Proper orthogonal decompositions in multifidelity uncertainty quantification of complex simulation models. *International Journal of Computer Mathematics*, 91(4):748–769, 2014.
[29] R. Y. Rubinstein and R. Marcus. Efficiency of multivariate control variates in monte carlo simulation. *Operations Research*, 33(3):661–677, 1985.
[30] A. Teckentrup, R. Scheichl, M. Giles, and E. Ullmann. Further analysis of multilevel monte carlo methods for elliptic pdes with random coefficients. *Numerische Mathematik*, 125(3):569–600, 2013.
[31] S. Venkatraman and J. R. Wilson. The efficiency of control variates in multiresponse simulation. *Operations Research Letters*, 5(1):37 – 42, 1986.

Appendix A. Proof of Lemma 2.3.

Proof. Our proof uses Propositions 2.1 and requires computing $\mathbb{C}\text{ov}\left[\underline{\Delta},\underline{\Delta}\right]$ and $\mathbb{C}\text{ov}\left[\underline{\Delta},\hat{Q}\right]$ for the MLMC sampling strategy. We begin with $\mathbb{C}\text{ov}\left[\underline{\Delta},\underline{\Delta}\right]$. Equation (2.16) implies this matrix is tridiagonal because CVs i and $i+2$ are never evaluated at overlapping samples. The diagonal elements of this matrix are the variances of Δ_i and can be evaluated as follows

$$\begin{aligned}\mathbb{V}\text{ar}\left[\Delta_i\right] &= \mathbb{V}\text{ar}\left[\frac{1}{\bar{r}_{i-1}N}\sum_{j=1}^{\bar{r}_{i-1}N} Q_i(\boldsymbol{z}_i^{1(j)}) - \frac{1}{\bar{r}_iN}\sum_{j=1}^{\bar{r}_iN} Q_i(\boldsymbol{z}_i^{2(j)})\right]\\ &= \mathbb{V}\text{ar}\left[\frac{1}{\bar{r}_{i-1}N}\sum_{j=1}^{\bar{r}_{i-1}N} Q_i(\boldsymbol{z}_i^{1(j)})\right] + \mathbb{V}\text{ar}\left[\frac{1}{\bar{r}_iN}\sum_{j=1}^{\bar{r}_iN} Q_i(\boldsymbol{z}_i^{2(j)})\right]\\ &= \frac{1}{\bar{r}_{i-1}N}\mathbb{V}\text{ar}\left[Q_i\right] + \frac{1}{\bar{r}_iN}\mathbb{V}\text{ar}\left[Q_i\right] = \frac{\bar{r}_i+\bar{r}_{i-1}}{\bar{r}_i\bar{r}_{i-1}}\frac{\mathbb{V}\text{ar}\left[Q_i\right]}{N},\end{aligned}$$

for $i = 1,\dots,M$ where $\bar{r}_0 = 1$. The second line follows because the two sums are uncorrelated due to a lack

of shared samples, i.e., $\boldsymbol{z}_i^1 \cap \boldsymbol{z}_i^2 = \emptyset$. A similar calculation yields that the off-diagonals are

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left[\Delta_i, \Delta_{i+1}\right] &= \mathbb{C}\mathrm{ov}\Bigg(\frac{1}{\bar{r}_{i-1}N}\sum_{j=1}^{\bar{r}_{i-1}N} Q_i(\boldsymbol{z}_i^{1(j)}) - \frac{1}{\bar{r}_i N}\sum_{j=1}^{\bar{r}_i N} Q_i(\boldsymbol{z}_i^{2(j)}), \\
&\qquad\qquad \frac{1}{\bar{r}_i N}\sum_{j=1}^{\bar{r}_i N} Q_{i+1}(\boldsymbol{z}_{i+1}^{1(j)}) - \frac{1}{\bar{r}_{i+1}N}\sum_{j=1}^{\bar{r}_{i+1}N} Q_{i+1}(\boldsymbol{z}_{i+1}^{2(j)})\Bigg) \\
&= \mathbb{C}\mathrm{ov}\left[-\frac{1}{\bar{r}_i N}\sum_{j=1}^{\bar{r}_i N} Q_i(\boldsymbol{z}_i^{2(j)}), \frac{1}{\bar{r}_i N}\sum_{j=1}^{\bar{r}_i N} Q_{i+1}(\boldsymbol{z}_{i+1}^{1(j)})\right] \\
&= -\frac{1}{(\bar{r}_i N)^2}\sum_{j=1}^{\bar{r}_i N} \mathbb{C}\mathrm{ov}\left[Q_i(\boldsymbol{z}_i^{2(j)}), Q_{i+1}(\boldsymbol{z}_{i+1}^{1(j)})\right] \\
&= -\frac{1}{\bar{r}_i N}\mathbb{C}\mathrm{ov}(Q_i, Q_{i+1}),
\end{aligned}$$

for $i = 1, \ldots, M-1$ and where in the second equality we used the fact that $\boldsymbol{z}_i^2 = \boldsymbol{z}_{i+1}^1$. Symmetry requires the same values for $\mathbb{C}\mathrm{ov}\left[\Delta_{i+1}, \Delta_i\right]$.

If we introduce the ratio between the standard deviations $\tau_i = \mathbb{V}\mathrm{ar}^{1/2}\left[Q_i\right]/\mathbb{V}\mathrm{ar}^{1/2}\left[Q\right]$, then the product $\underline{\alpha}^{\mathrm{T}}\mathbb{C}\mathrm{ov}\left[\underline{\Delta}, \underline{\Delta}\right]\underline{\alpha}$, with $\alpha_i = -1$, becomes

$$\begin{aligned}
\underline{\alpha}^{\mathrm{T}}\frac{\mathbb{C}\mathrm{ov}\left[\underline{\Delta}, \underline{\Delta}\right]}{\mathbb{V}\mathrm{ar}\left[\hat{Q}\right]}\underline{\alpha} &= \frac{1}{N}\left[\sum_{i=1}^{M}\frac{\bar{r}_i + \bar{r}_{i-1}}{\bar{r}_i\bar{r}_{i-1}}\frac{\mathbb{V}\mathrm{ar}\left[Q_i\right]}{\mathbb{V}\mathrm{ar}\left[\hat{Q}\right]} - 2\sum_{i=1}^{M-1}\frac{1}{\bar{r}_i}\frac{\mathbb{C}\mathrm{ov}(Q_i, Q_{i+1})}{\mathbb{V}\mathrm{ar}\left[\hat{Q}\right]}\right] \\
&= \sum_{i=1}^{M}\frac{\bar{r}_i + \bar{r}_{i-1}}{\bar{r}_i\bar{r}_{i-1}}\tau_i^2 - 2\sum_{i=1}^{M-1}\frac{1}{\bar{r}_i}\rho_{i(i+1)}\tau_i\tau_{i+1},
\end{aligned}$$

where ρ_{ij} is the Pearson correlation coefficient between Q_i and Q_j.

The covariance vector $\mathbb{C}\mathrm{ov}\left[\underline{\Delta}, \hat{Q}\right]$ is equal to $[-\mathbb{C}\mathrm{ov}(Q, Q_1)/N, 0, \ldots, 0]$ because only the estimator for the first LF model Δ_1 shares samples with $\hat{Q}$. Therefore

$$\underline{\alpha}^{\mathrm{T}}\frac{\mathbb{C}\mathrm{ov}\left[\underline{\Delta}, \hat{Q}\right]}{\mathbb{V}\mathrm{ar}\left[\hat{Q}\right]} = \frac{\mathbb{C}\mathrm{ov}(Q, Q_i)}{\mathbb{V}\mathrm{ar}\left[Q\right]} = \frac{\mathbb{C}\mathrm{ov}\left[Q, Q_1\right]\mathbb{V}\mathrm{ar}\left[Q_1\right]^{1/2}}{\mathbb{V}\mathrm{ar}\left[Q\right]^{1/2}\mathbb{V}\mathrm{ar}\left[Q\right]^{1/2}\mathbb{V}\mathrm{ar}\left[Q_1\right]^{1/2}} = \rho_1\tau_1. \tag{A.1}$$

Using this equation within Proposition 2.1 we obtain our stated result

$$\mathbb{V}\mathrm{ar}\left[\hat{Q}^{\mathrm{MLMC}}\right] = \mathbb{V}\mathrm{ar}\left[\hat{Q}\right]\left(1 + \sum_{i=1}^{M}\frac{\bar{r}_i + \bar{r}_{i-1}}{\bar{r}_i\bar{r}_{i-1}}\tau_i^2 - 2\sum_{i=1}^{M-1}\frac{1}{\bar{r}_i}\rho_{i(i+1)}\tau_i\tau_{i+1} + 2\rho_1\tau_1\right). \tag{A.2}$$

□

Appendix B. Proof of Theorem 2.4.

Proof. We seek to find an upper bound for R^2_{MLMC}. Our approach is to show that R^2_{MLMC} increases monotonically with data, and that, in the limit of infinite data, it is less than ρ_1^2.

We begin by restating the definition of R^2_{MLMC}

$$R^2_{\mathrm{MLMC}} = -\sum_{i=1}^{M}\frac{\bar{r}_i + \bar{r}_{i-1}}{\bar{r}_i\bar{r}_{i-1}}\tau_i^2 + 2\sum_{i=1}^{M-1}\frac{1}{\bar{r}_i}\rho_{i(i+1)}\tau_i\tau_{i+1} - 2\rho_1\tau_1. \tag{B.1}$$

We need to show that both summations exhibit monotonic behavior. Consider the coefficients of the first summation

$$g(\bar{r}_i, \bar{r}_{i-1}) = \frac{\bar{r}_i + \bar{r}_{i-1}}{\bar{r}_i\bar{r}_{i-1}}. \tag{B.2}$$

We need to show that this function monotonically decreases as the data is increased:

$$g(\bar{r}_i + a, \bar{r}_{i-1} + a) < g(\bar{r}_i, \bar{r}_{i-1}), \quad a > 0, a \in \mathbb{R}. \tag{B.3}$$

in which case R^2_{MLMC} would increase with increasing data. Therefore, we have

$$g(\bar{r}_i + a, \bar{r}_{i-1} + a) = \frac{\bar{r}_i + \bar{r}_{i-1} + 2a}{(\bar{r}_i + a)(\bar{r}_{i-1} + a)} = \frac{\bar{r}_i + \bar{r}_{i-1} + 2a}{\bar{r}_i\bar{r}_{i-1} + a(\bar{r}_i + \bar{r}_{i-1}) + a^2} \tag{B.4}$$

This function is continuous over the domain, and we can show that it is monotonically decreasing with a by showing that its derivative with respect to a is negative:

$$\frac{\partial}{\partial a} g(\bar{r}_i + a, \bar{r}_{i-1} + a) = \frac{1}{D(a)^2}\left[2\left(\bar{r}_i\bar{r}_{i-1} + a(\bar{r}_i + \bar{r}_{i-1}) + a^2\right) - (\bar{r}_i + \bar{r}_{i-1} + 2a)(\bar{r}_i + \bar{r}_{i-1} + 2a)\right] \tag{B.5}$$

$$= \frac{1}{D(a)^2}\left[-2a^2 - 2a(\bar{r}_{i+1} + \bar{r}_i) + 2\bar{r}_i\bar{r}_{i-1} - (\bar{r}_i + \bar{r}_{i-1})^2\right] \tag{B.6}$$

$$= \frac{1}{D(a)^2}\left[-2a^2 - 2a(\bar{r}_{i+1} + \bar{r}_i) - \bar{r}_i^2 - \bar{r}_{i-1}^2\right] < 0 \tag{B.7}$$

where $D(a)$ is the denominator of Equation (B.4) whose square is positive. The last equality follows from the fact that each term is positive. Furthermore the denominator of Equation (B.4) is positive so that $\lim_{a\to\infty} g(\bar{r}_i + a, \bar{r}_{i-1} + a) = 0$.

The second summation has coefficients $h(\bar{r}_i) = 1/\bar{r}_i$. Clearly, we have $h(\bar{r}_i + a) < h(\bar{r}_i)$ and $\lim_{a\to\infty} h(\bar{r}_i + a) = 0$ for positive a. Together these imply

$$R^2_{\text{MLMC}}(0) = -\sum_{i=1}^{M} g(\bar{r}_i + 0, \bar{r}_{i-1} + 0)\tau_i^2 + 2\sum_{i=1}^{M-1} h(\bar{r}_i + 0)\rho_{i(i+1)}\tau_i\tau_{i+1} - \rho_1\tau_1 \tag{B.8}$$

$$< -g(\bar{r}_1 + a, \bar{r}_0)\tau_1^2 - \sum_{i=2}^{M} g(\bar{r}_i + a, \bar{r}_{i-1} + a)\tau_i^2 + 2\sum_{i=1}^{M-1} h(\bar{r}_i + a)\rho_{i(i+1)}\tau_i\tau_{i+1} - 2\rho_1\tau_1 \tag{B.9}$$

$$= R^2_{\text{MLMC}}(a) \quad \text{for all } a > 0, \tag{B.10}$$

and

$$\lim_{a\to\infty} R^2_{\text{MLMC}}(a) = -\tau_1^2 - 2\rho_1\tau_1 = -\tau_1(\tau_1 + 2\rho_1). \tag{B.11}$$

Note that we separated out $g(\bar{r}_1, \bar{r}_0)$ because $\bar{r}_0$ is a boundary condition that is fixed to $\bar{r}_0 = 1$. This limit is maximized when $\tau_1 = -\rho_1$.

$$\lim_{a\to\infty} R^2_{\text{MLMC}}(a) = \rho_1^2. \tag{B.12}$$

This bound and monotonicity leads to our stated result. ∎

Appendix C. Proof of Lemma 2.5.

Proof. Our result follows from computing $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ and $\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]$ and applying Proposition 2.2. For this estimator we have

$$\Delta_i(\boldsymbol{z}_i) = \frac{1}{r_{i-1}N}\sum_{j=1}^{r_{i-1}N} Q_i\left(\boldsymbol{z}_i^{1(j)}\right) - \frac{1}{r_iN}\sum_{j=1}^{r_iN} Q_i\left(\boldsymbol{z}_i^{2(j)}\right). \tag{C.1}$$

Next we partition this quantity into three sums that contain non-overlapping (independent) sets of samples. The first sum is partitioned according to

$$\frac{1}{r_{i-1}N}\sum_{j=1}^{r_{i-1}N} Q_i\left(\boldsymbol{z}_i^{1(j)}\right) = \frac{1}{r_{i-1}N}\left[\sum_{j=1}^{N} Q_i\left(\boldsymbol{z}_j^{1(j)}\right) + \sum_{j=N+1}^{r_{i-1}N} Q_i\left(\boldsymbol{z}_i^{1(j)}\right)\right].$$

The second sum is partitioned according to

$$\frac{1}{r_i N}\sum_{j=1}^{r_i N} Q_i\left(\boldsymbol{z}_i^{2(j)}\right) = \frac{1}{r_i N}\left[\sum_{j=1}^{N} Q_i\left(\boldsymbol{z}_i^{1(j)}\right) + \sum_{j=N+1}^{r_{i-1}N} Q_i\left(\boldsymbol{z}_i^{1(j)}\right) + \sum_{j=r_{i-1}N+1}^{r_i N} Q_i\left(\boldsymbol{z}_i^{2(j)}\right)\right]$$

Now we can rewrite Δ_i as

$$\Delta_i(\boldsymbol{z}_i) = \frac{r_i - r_{i-1}}{r_i r_{i-1}}\frac{1}{N}\sum_{j=1}^{N} Q_i\left(\boldsymbol{z}_i^{(j)}\right) + \frac{r_i - r_{i-1}}{r_i r_{i-1}}\frac{1}{N}\sum_{j=N+1}^{r_{i-1}N} Q_i\left(\boldsymbol{z}_i^{(j)}\right) - \frac{1}{r_i}\frac{1}{N}\sum_{j=r_{i-1}N+1}^{r_i N} Q_i\left(\boldsymbol{z}_i^{(j)}\right), \tag{C.2}$$

where we dropped the superscripts because it is no longer necessary to distinguish between $\boldsymbol{z}_i^1$ and $\boldsymbol{z}_i^2$.

Now we consider the diagonal of $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$. Since the sums are independent, the variance of the sums is the sum of the variances

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\Delta_i\right] &= \frac{(r_i - r_{i-1})^2}{r_i^2 r_{i-1}^2}\frac{\mathbb{V}\text{ar}\left[Q\right]}{N} + \frac{(r_i - r_{i-1})^2}{r_i^2 r_{i-1}^2}(r_{i-1} - 1)\frac{\mathbb{V}\text{ar}\left[Q\right]}{N} + \frac{r_i - r_{i-1}}{r_i^2}\frac{\mathbb{V}\text{ar}\left[Q\right]}{N} \\
&= r_{i-1}\frac{(r_i - r_{i-1})^2}{r_i^2 r_{i-1}^2}\frac{\mathbb{V}\text{ar}\left[Q\right]}{N} + \frac{r_i - r_{i-1}}{r_i^2}\frac{\mathbb{V}\text{ar}\left[Q\right]}{N} \\
&= \left[\frac{r_i^2 - 2r_i r_{i-1} + r_{i-1}^2}{r_i^2 r_{i-1}} + \frac{r_{i-1}}{r_{i-1}}\frac{r_i - r_{i-1}}{r_i^2}\right]\frac{\mathbb{V}\text{ar}\left[Q\right]}{N} = \frac{r_i\left(r_i - r_{i-1}\right)}{r_i^2 r_{i-1}}\frac{\mathbb{V}\text{ar}\left[Q\right]}{N} \\
&= \frac{r_i - r_{i-1}}{r_i r_{i-1}}\frac{\mathbb{V}\text{ar}\left[Q\right]}{N}.
\end{aligned}$$

Next we consider the off-diagonal terms of $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$. Without loss of generality, consider Δ_i and Δ_j for $j > i$. The nested structure of the sampling set means that $\boldsymbol{z}_j^{(k)} = \boldsymbol{z}_i^{(k)}$ for $k \leq r_i N$. Using this fact we can rewrite Equation (C.2) as

$$\Delta_i(\boldsymbol{z}_i) = \frac{r_i - r_{i-1}}{r_i r_{i-1}}\frac{1}{N}\sum_{k=1}^{r_{i-1}N} Q_i\left(\boldsymbol{z}_i^{(k)}\right) - \frac{1}{r_i}\frac{1}{N}\sum_{k=r_{i-1}N+1}^{r_i N} Q_i\left(\boldsymbol{z}_i^{(k)}\right) \tag{C.3}$$

for the ith CV

$$\begin{aligned}
\Delta_j(\boldsymbol{z}_j) = \frac{r_j - r_{j-1}}{r_j r_{j-1}}\frac{1}{N}\Bigg[&\sum_{k=1}^{r_{i-1}N} Q_j\left(\boldsymbol{z}_i^{(i)}\right) + \sum_{k=r_{i-1}N+1}^{r_i N} Q_j\left(\boldsymbol{z}_i^{k}\right) + \sum_{k=r_i N+1}^{r_{j-1}N} Q_j\left(\boldsymbol{z}_k^{k}\right)\Bigg] \\
&- \frac{1}{r_j}\frac{1}{N}\sum_{k=r_{j-1}N+1}^{r_j N} Q_j\left(\boldsymbol{z}_j^{(k)}\right),
\end{aligned}$$

for the jth. Only the first two sums in this last expression share samples with the Δ_i, and these sums correspond to the samples delegated to $\hat{Q}_j$. Therefore we have

$$\mathbb{C}\text{ov}\left[\Delta_i, \Delta_j\right] = \frac{(r_i - r_{i-1})(r_j - r_{j-1})}{r_i r_{i-1} r_j r_{j-1}} r_{i-1}\mathbb{C}\text{ov}\left[Q_i, Q_j\right] - \frac{r_j - r_{j-1}}{r_i r_j r_{j-1}}(r_i - r_{i-1})\mathbb{C}\text{ov}\left[Q_i, Q_j\right] = 0. \tag{C.4}$$

To summarize $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ is diagonal

$$\mathbb{C}\text{ov}\left[\Delta_i, \Delta_j\right] = \begin{cases} \dfrac{r_i - r_{i-1}}{r_i r_{i-1}}\dfrac{\mathbb{V}\text{ar}\left[Q_i\right]}{N} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}, \quad \text{where } r_0 = 1. \tag{C.5}$$

Next we turn to $\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]$. From Equation C.2 we see that samples that each CV shares with the high fidelity model are entirely contained in the first sum. Therefore $\mathbb{C}\text{ov}\left[\hat{Q}, \Delta_i\right]$ becomes

$$\mathbb{C}\text{ov}\left[\hat{Q}, \Delta_i\right] = \frac{r_i - r_{i-1}}{r_i r_{i-1}}\frac{\mathbb{C}\text{ov}\left[Q, Q_i\right]}{N}. \tag{C.6}$$

Using $\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]$ and $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ within Equation 2.12 of Proposition 2.2 yields our stated result

$$\underline{\alpha}^{\text{MFMC}} = -\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]^{-1} \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right] = -\left[\frac{\mathbb{C}\text{ov}\left[Q, Q_1\right]}{\mathbb{V}\text{ar}\left[Q_1\right]}, \ldots, \frac{\mathbb{C}\text{ov}\left[Q, Q_M\right]}{\mathbb{V}\text{ar}\left[Q_M\right]}\right]. \qquad \square$$

Appendix D. Proof of Lemma 2.6.

Proof. The variance reduction follows directly from Proposition 2.2 as

$$\mathbb{V}\text{ar}\left[\hat{Q}^{\text{MFMC}}(\underline{\alpha})\right] = \mathbb{V}\text{ar}\left[\hat{Q}\right]\left(1 - \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]^T \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]^{-1}}{\mathbb{V}\text{ar}\left[\hat{Q}\right]} \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]\right). \tag{D.1}$$

Equations (C.5) and (C.6) in the proof of Lemma 2.5 provide us with the expressions for $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ and $\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]$, respectively. Since this covariance is diagonal and easily invertible we obtain

$$R^2_{\text{MFMC}} = \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]^T \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]^{-1}}{\mathbb{V}\text{ar}\left[\hat{Q}\right]} \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right] = \sum_{i=1}^{M} \frac{r_i - r_{i-1}}{r_i r_{i-1}} \frac{\mathbb{C}\text{ov}\left[Q, Q_i\right]^2}{\mathbb{V}\text{ar}\left[Q_i\right]\mathbb{V}\text{ar}\left[Q\right]} = \sum_{i=1}^{M} \frac{r_i - r_{i-1}}{r_i r_{i-1}} \rho_i^2.$$

We pull out ρ_1 from this expression

$$R^2_{\text{MFMC}} = \frac{r_1 - 1}{r_1}\rho_1^2 + \rho_1^2 \sum_{i=2}^{M} \frac{r_i - r_{i-1}}{r_i r_{i-1}} \frac{\rho_i^2}{\rho_1^2} = \rho_1^2 \left(\frac{r_1 - 1}{r_1} + \sum_{i=2}^{M} \frac{r_i - r_{i-1}}{r_i r_{i-1}} \frac{\rho_i^2}{\rho_1^2}\right), \tag{D.2}$$

where we have used $r_0 = 1$, to enable a clear comparison with the single control variate. $\square$

Appendix E. Proof of Theorem 2.7.

Proof. The proof follows from the following straightforward algebraic manipulation of (2.25).

$$\begin{aligned} R^2_{\text{MFMC}} = \rho_1^2 \left(\frac{r_1 - 1}{r_1} + \sum_{i=2}^{M} \frac{r_i - r_{i-1}}{r_i r_{i-1}} \frac{\rho_i^2}{\rho_1^2}\right) &\leq \rho_1^2 \left(\frac{r_1 - 1}{r_1} + \sum_{i=2}^{M} \frac{r_i - r_{i-1}}{r_i r_{i-1}}\right) \\ &= \rho_1^2 \left(1 - \frac{1}{r_1} + \sum_{i=2}^{M} \frac{1}{r_{i-1}} - \frac{1}{r_i}\right) = \rho_1^2 \left(1 - \frac{1}{r_M}\right) \\ &< \rho_1^2. \end{aligned}$$ $\square$

Appendix F. Proof of Theorem 3.2.

For reference we recall the definition of $\boldsymbol{F}^{(IS)}$

$$\boldsymbol{F}^{(IS)}{}_{ij} = \begin{cases} \frac{r_i - 1}{r_i}\frac{r_j - 1}{r_j} & \text{if } i \neq j \\ \frac{r_i - 1}{r_i} & \text{otherwise} \end{cases}.$$

Proof. This proof again makes use of Propositions 2.1 and 2.2, which require the computation of $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ and $\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]$. We begin with $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ by noticing that $\Delta_i(\boldsymbol{z}_i) = \hat{Q}_i(\boldsymbol{z}) - \hat{\mu}_i(\boldsymbol{z}_i)$ and that for $i \neq j$, Δ_i is correlated with Δ_j only through the first N samples. This pattern will emerge for all the various quantities that we require and so we first split each Δ_i into two sums that have independent samples:

$$\begin{aligned} \Delta_i(\boldsymbol{z}_i) &= \frac{1}{N}\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}) - \frac{1}{r_i N}\sum_{k=1}^{r_i N} Q_i(\boldsymbol{z}_i^{(k)}) \\ &= \frac{1}{N}\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}) - \frac{1}{r_i N}\left[\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}) + \sum_{k=N+1}^{r_i N} Q_i(\boldsymbol{z}_i^{(k)})\right] \\ &= \frac{r_i - 1}{r_i N}\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}) - \frac{1}{r_i N}\sum_{k=N+1}^{r_i N} Q_i(\boldsymbol{z}_i^{(k)}). \end{aligned} \tag{F.1}$$

Now that these two sums are independent, the covariance between Δ_i and Δ_j for $i \neq j$ is due to the first summation:

$$\mathbb{C}\text{ov}\left[\Delta_i, \Delta_j\right] = \mathbb{C}\text{ov}\left[\frac{r_i - 1}{r_i N}\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}), \frac{r_j - 1}{r_j N}\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)})\right] = \frac{r_i - 1}{r_i}\frac{r_j - 1}{r_j}\frac{1}{N}\mathbb{C}\text{ov}\left[Q_i, Q_j\right].$$

Using the same splitting we can derive the diagonal terms (variances) as

$$\mathbb{V}\text{ar}\left[\Delta_i\right] = \frac{(r_i - 1)^2}{r_i^2 N}\mathbb{V}\text{ar}\left[Q_i\right] + \frac{r_i - 1}{r_i^2 N}\mathbb{V}\text{ar}\left[Q_i\right] = \left(\frac{(r_i - 1)^2}{r_i^2} + \frac{r_i - 1}{r_i^2}\right)\frac{\mathbb{V}\text{ar}\left[Q_i\right]}{N} = \frac{r_i - 1}{r_i}\frac{\mathbb{V}\text{ar}\left[Q_i\right]}{N}. \tag{F.2}$$

Ths result can be succinctly represented as

$$\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right] = \frac{1}{N}\left[\boldsymbol{C} \circ \boldsymbol{F}^{(IS)}\right], \tag{F.3}$$

where $\circ$ denotes the Hadamard (elementwise) product.

Now we consider the $\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]$ term. The covariance between $\hat{Q}$ and Δ_i is again a result of the first N samples of $\boldsymbol{z}_i$ all the samples used for $\hat{Q}_i$ and for the first N samples used for $\hat{\mu}_i$. The splitting (F.1) then yields

$$\mathbb{C}\text{ov}\left[\hat{Q}, \Delta_i\right] = \mathbb{C}\text{ov}\left[\frac{1}{N}\sum_{i=1}^{N} Q_i(\boldsymbol{z}^{(k)}), \frac{r_i - 1}{r_i N}\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)})\right] = \frac{r_i - 1}{r_i}\frac{\mathbb{C}\text{ov}\left[Q, Q_i\right]}{N}. \tag{F.4}$$

More succinctly we have

$$\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right] = \frac{1}{N}\left[\text{diag}\left(\boldsymbol{F}^{(IS)}\right) \circ \boldsymbol{c}\right]. \tag{F.5}$$

Using $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ and $\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]$ and Propositions 2.1 and 2.2, we obtain our stated result

$$\begin{aligned}
R^2_{\text{ACV-}IS} &= \mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]^T \frac{\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]^{-1}}{\mathbb{V}\text{ar}\left[\hat{Q}\right]}\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right], \\
&= \frac{1}{N}\left[\text{diag}\left(\boldsymbol{F}^{(IS)}\right) \circ \boldsymbol{c}\right]^T \frac{N}{\frac{\mathbb{V}\text{ar}[Q]}{N}}\left[\boldsymbol{C} \circ \boldsymbol{F}^{(IS)}\right]^{-1}\frac{1}{N}\left[\text{diag}\left(\boldsymbol{F}^{(IS)}\right) \circ \boldsymbol{c}\right], \\
&= \left[\text{diag}\left(\boldsymbol{F}^{(IS)}\right) \circ \bar{\boldsymbol{c}}\right]^T \left[\boldsymbol{C} \circ \boldsymbol{F}^{(IS)}\right]^{-1}\left[\text{diag}\left(\boldsymbol{F}^{(IS)}\right) \circ \bar{\boldsymbol{c}}\right].
\end{aligned}$$

∎

Appendix G. Proof of Theorem 3.4. For reference, recall the definition of $\boldsymbol{F}^{(MF)}$

$$\boldsymbol{F}^{(MF)}{}_{ij} = \begin{cases} \frac{\min(r_i, r_j) - 1}{\min(r_i, r_j)} & \text{if } i \neq j \\ \frac{r_i - 1}{r_i} & \text{otherwise} \end{cases}.$$

Proof. The difference between ACV-IS and ACV-MF is that the samples used to estimate $\hat{\mu}_i$ are reused for the first $\min(r_j N, r_i N)$ samples of $\hat{\mu}_j$ when $j \geq i$. As a result, we have different expressions for the off-diagonal terms of $\mathbb{C}\text{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$, but we retain the same expressions for the diagonal terms $\mathbb{V}\text{ar}\left[\Delta_i\right]$ and for $\mathbb{C}\text{ov}\left[\underline{\Delta}, \hat{Q}\right]$. Therefore, we need to derive a new expression for $\mathbb{C}\text{ov}\left[\Delta_i, \Delta_j\right]$ for this estimator, and then reuse the previous results within Propositions 2.1 and 2.2 to obtain our results.

We have two cases to consider, when $r_j \geq r_i$ and when $r_j < r_i$. Our derivation begins by splitting Δ_i and Δ_j into sums with independent samples and assuming that $r_j \geq r_i$. The split for Δ_i is the same as Equation (F.1) and is repeated here for convenience

$$\Delta_i(\boldsymbol{z}_i) = \frac{r_i - 1}{r_i N}\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}) - \frac{1}{r_i N}\sum_{k=N+1}^{r_i N} Q_i(\boldsymbol{z}_i^{(k)})$$

The splitting for Δ_j now includes the samples $\boldsymbol{z}_i$

$$\Delta_j(\boldsymbol{z}_j) = \frac{1}{N}\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) - \frac{1}{r_j N}\sum_{k=1}^{r_j N} Q_j(\boldsymbol{z}_j^{(k)})$$

$$= \frac{1}{N}\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) - \frac{1}{r_j N}\left[\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) + \sum_{k=N+1}^{r_i N} Q_j(\boldsymbol{z}_i^{(k)}) + \sum_{k=r_i N+1}^{r_j N} Q_j(\boldsymbol{z}_j^{(k)})\right] \tag{G.1}$$

$$= \frac{r_j - 1}{r_j N}\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) - \frac{1}{r_j N}\sum_{k=N+1}^{r_i N} Q_j(\boldsymbol{z}_i^{(k)}) - \frac{1}{r_j N}\sum_{k=r_i N+1}^{r_j N} Q_j(\boldsymbol{z}_j^{(k)}) \tag{G.2}$$

where for the third summation in the second equality we used the fact that $\boldsymbol{z}_j^{(k)} = \boldsymbol{z}_i^{(k)}$ for $k \le r_i N$. Now, we see that the covariance between Δ_i and Δ_j is due to the samples from the first two summations. This fact implies

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left[\Delta_i, \Delta_j\right] = \mathbb{C}\mathrm{ov}\Bigg[&\frac{r_i - 1}{r_i N}\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}) - \frac{1}{r_i N}\sum_{k=N+1}^{r_i N} Q_i(\boldsymbol{z}_i^{(k)}),\\ &\frac{r_j - 1}{r_j N}\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) - \frac{1}{r_j N}\sum_{k=N+1}^{r_i N} Q_j(\boldsymbol{z}_i^{(k)})\Bigg]\\ = \frac{r_i - 1}{r_i}\frac{r_j - 1}{r_j}&\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N} + \frac{(r_i - 1)}{r_i r_j}\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N}\\ = \left[\frac{r_i r_j - r_i - r_j + 1}{r_i r_j} + \frac{r_i - 1}{r_i r_j}\right]&\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N} = \frac{r_i - 1}{r_i}\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N}\end{aligned}$$

Next we consider the case where $r_i > r_j$. This requires a different splitting of Δ_j

$$\begin{aligned}\Delta_j(\boldsymbol{z}_j) &= \frac{1}{N}\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) - \frac{1}{r_j N}\sum_{k=1}^{r_j N} Q_j(\boldsymbol{z}_j^{(k)})\\ &= \frac{1}{N}\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) - \frac{1}{r_j N}\left[\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) + \sum_{k=N+1}^{r_j N} Q_j(\boldsymbol{z}_i^{(k)})\right]\\ &= \frac{r_j - 1}{r_j N}\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) - \frac{1}{r_j N}\sum_{k=N+1}^{r_j N} Q_j(\boldsymbol{z}_i^{(k)}),\end{aligned}$$

where the difference is that the first equation doesn't have an extra term to account for samples in $\boldsymbol{z}_j$ that are not in $\boldsymbol{z}_i$. Using the same reasoning as above we have

$$\mathbb{C}\mathrm{ov}\left[\Delta_i, \Delta_j\right] = \frac{r_i - 1}{r_i}\frac{r_j - 1}{r_j}\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N} + \frac{(r_j - 1)}{r_i r_j}\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N} = \frac{r_j - 1}{r_j}\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N}.$$

We combine these two cases into a single formula

$$\mathbb{C}\mathrm{ov}\left[\Delta_i, \Delta_j\right] = \frac{\min(r_i, r_j) - 1}{\min(r_i, r_j)}\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N}. \tag{G.3}$$

The expression for $\mathbb{V}\mathrm{ar}\left[\Delta_i\right]$ is identical to that of estimator ACV-IS (Equation (F.2)) because the same splitting strategy can be used. The expression for $\mathbb{C}\mathrm{ov}\left[\underline{\Delta}, \hat{Q}\right]$ is also identical to that of estimator ACV-IS (Equation (F.5)) for the same reason. Therefore, the stated result follows by using these expressions within Propositions 2.1 and 2.2 as before. ∎

Appendix H. Proof of Theorem 3.6.

Proof. Using the previous proposition we need to show that $\boldsymbol{F}^{(IS)} \to \mathbf{1}_{M\times M}$ and that $\boldsymbol{F}^{(MF)} \to \mathbf{1}_{M\times M}$. Since these two matrices share the diagonal entry, we begin with the diagonal

$$\lim_{r_i\to\infty} \boldsymbol{F}^{(IS)}{}_{ii} = \lim_{r_i\to\infty} \boldsymbol{F}^{(MF)}{}_{ii} = \lim_{r_i\to\infty} \frac{r_i - 1}{r_i} = 1.$$

Next we consider the off-diagonals of $\boldsymbol{F}^{(IS)}$:

$$\lim_{r_i,r_j\to\infty} \boldsymbol{F}^{(IS)}{}_{ij} = \lim_{r_i,r_j\to\infty} \frac{r_i - 1}{r_i}\frac{r_j - 1}{r_j} = 1$$

because each component converges to 1. Finally we consider the off-diagonals of $\boldsymbol{F}^{(MF)}$

$$\lim_{r_i,r_j\to\infty} \boldsymbol{F}^{(MF)}{}_{ij} = \lim_{r_i,r_j\to\infty} \frac{\min(r_i,r_j) - 1}{\min(r_i,r_j)} = 1$$

because the minimum value goes to infinity since both r_i and r_j go to infinity. ∎

Appendix I. Proof of Theorem 3.8.

Proof. Again we rely on Propositions 2.1 and 2.2. The separated nature of this estimator implies that the computation of $\mathbb{C}\mathrm{ov}\left[\underline{\Delta}, \hat{Q}\right]$ and $\mathbb{C}\mathrm{ov}\left[\underline{\Delta}, \underline{\Delta}\right]$ can be separated into three regimes: (1) $i,j \le K$; (2) when both $i,j > K$; and (3) either $i \le K$ and $j > K$, which is symmetric to $j \le K$ and $i > K$. The first two cases are the most straight forward because they essentially follow from ACV-MF.

Case 1: $i,j \le K$. This case is identical to ACV-MF for a CV estimator with K CVs. Therefore we have

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left[\hat{Q}, \Delta_i\right] &= \frac{r_i - 1}{r_i}\frac{\mathbb{C}\mathrm{ov}\left[Q, Q_i\right]}{N} \\
\mathbb{C}\mathrm{ov}\left[\Delta_i, \Delta_j\right] &= \frac{\min(r_i,r_j) - 1}{\min(r_i,r_j)}\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N}, \\
\mathbb{V}\mathrm{ar}\left[\Delta_i\right] &= \frac{r_i - 1}{r_i}\frac{\mathbb{V}\mathrm{ar}\left[Q_i\right]}{N}
\end{aligned}$$

Case 2: $i,j > K$. For this set of indices we generate a new set of splittings into the independent components amongst $\boldsymbol{z}$, $\boldsymbol{z}_L$, $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. If we assume $r_j > r_i$ for $j > i$ then we have

$$\begin{aligned}
\Delta_j(\boldsymbol{z}_j) &= \frac{1}{r_L N}\sum_{k=1}^{r_L N} Q_j(\boldsymbol{z}_L^{(k)}) - \frac{1}{r_j N}\sum_{k=1}^{r_j N} Q_j(\boldsymbol{z}_j^{(k)}) \\
&= \frac{1}{r_L N}\sum_{k=1}^{r_L N} Q_j(\boldsymbol{z}_L^{(k)}) - \frac{1}{r_j N}\left[\sum_{k=1}^{r_L N} Q_j(\boldsymbol{z}_L^{(k)}) + \sum_{k=r_L N+1}^{r_i N} Q_j(\boldsymbol{z}_i^{(k)}) + \sum_{k=r_i N+1}^{r_j N} Q_j(\boldsymbol{z}_j^{(k)})\right] \\
&= \frac{r_j - r_L}{r_L r_j N}\left[\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) + \sum_{k=N+1}^{r_L N} Q_j(\boldsymbol{z}_L^{(k)})\right] - \frac{1}{r_j N}\left[\sum_{k=r_L N+1}^{r_i N} Q_j(\boldsymbol{z}_i^{(k)}) + \sum_{k=r_i N+1}^{r_j N} Q_j(\boldsymbol{z}_j^{(k)})\right] \qquad \text{(I.1)}
\end{aligned}$$

If on the other hand, $r_j < r_i$ for $j > i$ then instead of two sums in the second component of Equation (I.1) we simply have one sum

$$\Delta_j(\boldsymbol{z}_j) = \frac{r_j - r_L}{r_L r_j N}\left[\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) + \sum_{k=N+1}^{r_L N} Q_j(\boldsymbol{z}_L^{(k)})\right] - \frac{1}{r_j N}\sum_{k=r_L N+1}^{r_j N} Q_j(\boldsymbol{z}_i^{(k)}), \qquad \text{(I.2)}$$

because in this case $\boldsymbol{z}_j$ would consist of the first $r_j N$ samples of $\boldsymbol{z}_i$.

Now we use Equations (I.1) and (I.2) to compute the required covariances. First we have

$$\mathbb{C}\mathrm{ov}\left[\hat{Q}, \Delta_i\right] = \frac{r_i - r_L}{r_i r_L}\frac{\mathbb{C}\mathrm{ov}\left[Q, Q_i\right]}{N},$$

where we have used the fact that the N shared samples used in Δ_i are contained in the first sum in Equation (I.1), and that this sum has the coefficient $\frac{r_i - r_L}{r_i r_L}\frac{1}{N}$.

Next we have the covariance between the control variates. These estimators share several groups of independent samples that we will first consider separately and then sum together. The first group of shared samples is $\boldsymbol{z}_L$ which consists of $r_L N$ samples. These samples correspond to the first two summations of Equation (I.1), and therefore their covariance becomes

$$\frac{r_i - r_L}{r_i}\frac{r_j - r_L}{r_j}\frac{1}{r_L}\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N}.$$

Next, these estimators share an additional $(\min(r_i, r_j) - r_L)N$. If $r_j > r_i$ for $j > i$ then these shared samples arise in the third sum of the splitting (I.1) and we obtain $\frac{1}{r_i r_j}(r_i - r_L)\frac{\mathbb{C}\mathrm{ov}[Q_i, Q_j]}{N}$, and if $r_j < r_i$ for $j > i$ then we have $\frac{1}{r_i r_j}(r_j - r_L)\frac{\mathbb{C}\mathrm{ov}[Q_i, Q_j]}{N}$. Together these imply for $j > i$ we have

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left[\Delta_i, \Delta_j\right] &= \left[\frac{r_i - r_L}{r_i}\frac{r_j - r_L}{r_j}\frac{1}{r_L} + \frac{1}{r_i r_j}(\min(r_i, r_j) - r_L)\right]\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N} \\
&= \left[\frac{(r_i - r_L)(r_j - r_L) + r_L(\min(r_i, r_j) - r_L)}{r_i r_j r_L}\right]\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N}
\end{aligned}$$

Finally, we have the diagonal component, which considers both components of Equation (I.1)

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[Q_i\right] &= \left[\frac{(r_i - r_L)^2}{r_i^2 r_L} + \frac{1}{r_i^2}(r_i - r_L)\right]\frac{\mathbb{V}\mathrm{ar}\left[Q_i\right]}{N} = \frac{r_i^2 - 2r_i r_L + r_L^2 + r_L r_i - r_L^2}{r_i^2 r_L}\frac{\mathbb{V}\mathrm{ar}\left[Q_i\right]}{N} \\
&= \frac{r_i - r_L}{r_i r_L}\frac{\mathbb{V}\mathrm{ar}\left[Q_i\right]}{N}.
\end{aligned}$$

Case 3: Now we consider the final case $i \leq K$, and $j > K$. The reverse follows from symmetry. This case itself can be broken into three subcases: (a) $i = L$, (b) $i < L$, and (c) $i > L$. First we consider case (3a). In this case, we have the splitting from Equation (F.1) and the one just derived above (I.1). As for Case 2, we compute and add together the covariances resulting from the sums of independent samples from each splitting to obtain

$$\mathbb{C}\mathrm{ov}\left[\Delta_L, \Delta_j\right] = \left[\frac{r_L - 1}{r_L}\frac{r_j - r_L}{r_L r_j} - \frac{1}{r_L}\frac{r_j - r_L}{r_L r_j}(r_L - 1)\right]\frac{\mathbb{C}\mathrm{ov}\left[Q_L, Q_j\right]}{N} = 0,$$

For case (3b) all of the samples in $\boldsymbol{z}_i$ are shared and we need to use a different splitting Δ_j to elucidate this fact. The relevant splitting for Δ_i is given by Equation (F.1), and the necessary splitting for Δ_j becomes

$$\begin{aligned}
\Delta_j(\boldsymbol{z}_j) &= \frac{1}{r_L N}\sum_{k=1}^{r_L N} Q_j(\boldsymbol{z}_L^{(k)}) - \frac{1}{r_j N}\sum_{k=1}^{r_j N} Q_j(\boldsymbol{z}_j^{(k)}) \\
&= \frac{1}{r_L N}\left[\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) + \sum_{k=N+1}^{r_i N} Q_j(\boldsymbol{z}_i^{(k)}) + \sum_{k=r_i N+1}^{r_L N} \boldsymbol{z}_L^{(k)}\right] - \\
&\qquad \frac{1}{r_j N}\left[\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) + \sum_{k=N+1}^{r_i N} Q_j(\boldsymbol{z}_i^{(k)}) + \sum_{k=r_i N+1}^{r_L N} Q_j(\boldsymbol{z}_L^{(k)}) + \sum_{k=r_L N+1}^{r_j N} Q_j(\boldsymbol{z}_j^{(k)})\right] \\
&= \frac{r_j - r_L}{r_j r_L N}\left[\sum_{k=1}^{N} Q_j(\boldsymbol{z}^{(k)}) + \sum_{k=N+1}^{r_i N} Q_j(\boldsymbol{z}_i^{(k)}) + \sum_{k=r_i N+1}^{r_L N} Q_j(\boldsymbol{z}_L^{(k)})\right] - \frac{1}{r_j N}\sum_{k=r_L N+1}^{r_j N} Q_j(\boldsymbol{z}_j^{(k)}).
\end{aligned} \tag{I.3}$$

Therefore we can collect the terms that share $\boldsymbol{z}$ and $\boldsymbol{z}_i$ to obtain

$$\mathbb{C}\mathrm{ov}\left[\Delta_i, \Delta_j\right] = \left[\frac{r_i - 1}{r_i}\frac{r_j - r_L}{r_j r_L} - \frac{1}{r_i}\frac{r_j - r_L}{r_j r_L}(r_i - 1)\right]\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N} = 0$$

Finally, we consider case (3c) where $L < i \leq K$ and $j > K$. We now need a new splitting for Δ_i

$$\begin{aligned}
\Delta_i(\boldsymbol{z}_i) &= \frac{1}{N}\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}) - \frac{1}{r_i N}\sum_{k=1}^{r_i N} Q_i(\boldsymbol{z}_i^{(k)}) \\
&= \frac{1}{N}\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}) - \frac{1}{r_i N}\left[\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}) + \sum_{k=N+1}^{r_L N} Q_i(\boldsymbol{z}_L^{(k)}) + \sum_{k=r_L N+1}^{r_i N} Q_i(\boldsymbol{z}_i^{(k)})\right] \\
&= \frac{r_i - 1}{r_i N}\sum_{k=1}^{N} Q_i(\boldsymbol{z}^{(k)}) - \frac{1}{r_i N}\left[\sum_{k=N+1}^{r_L N} Q_i(\boldsymbol{z}_L^{(k)}) + \sum_{k=r_L N+1}^{r_i N} Q_i(\boldsymbol{z}_i^{(k)})\right],
\end{aligned}$$

and we can use the splitting Equation (I.1) for Δ_j to obtain components from three summations with shared samples

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left[\Delta_i, \Delta_j\right] &= \left[\frac{r_i - 1}{r_i}\frac{r_j - r_L}{r_j r_L} - \frac{1}{r_i}\frac{r_j - r_L}{r_j r_L}(r_L - 1) + \frac{1}{r_i}\frac{1}{r_j}(r_i - r_L)\right]\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N} \\
&= \left[\frac{r_i r_j - r_i r_L - r_j + r_L - r_j r_L + r_j + r_L^2 - r_L + r_L r_i - r_L^2}{r_i r_j r_L}\right]\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N} \\
&= \left[\frac{r_i - r_L}{r_i r_L}\right]\frac{\mathbb{C}\mathrm{ov}\left[Q_i, Q_j\right]}{N}
\end{aligned}$$

To summarize, we have

$$\text{(I.4)}\quad \mathbb{C}\mathrm{ov}\left[\Delta_i, \Delta_j\right] = \begin{cases} \frac{\min(r_i, r_j) - 1}{\min(r_i, r_j)}\frac{\mathbb{C}\mathrm{ov}[Q_i, Q_j]}{N} & \text{if } i, j \leq K \\ \left[\frac{(r_i - r_L)(r_j - r_L) + r_L(\min(r_i, r_j) - r_L)}{r_i r_j r_L}\right]\frac{\mathbb{C}\mathrm{ov}[Q_i, Q_j]}{N} & \text{if } i, j > K \\ \left[\frac{r_i - r_L}{r_i r_L}\right]\frac{\mathbb{C}\mathrm{ov}[Q_i, Q_j]}{N} & \text{if } L < i \leq K,\ j > K \\ \left[\frac{r_j - r_L}{r_j r_L}\right]\frac{\mathbb{C}\mathrm{ov}[Q_i, Q_j]}{N} & \text{if } L < j \leq K,\ i > K \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } i \neq j.$$

The diagonal elements are $\mathbb{V}\mathrm{ar}\left[\Delta_i\right] = \frac{r_i - 1}{r_i}\frac{\mathbb{V}\mathrm{ar}[Q_i]}{N}$ if $i \leq K$ and $\mathbb{V}\mathrm{ar}\left[\Delta_i\right] = \frac{r_i - r_L}{r_i r_L}\frac{\mathbb{V}\mathrm{ar}[Q_i]}{N}$ otherwise. Furthermore, $\mathbb{C}\mathrm{ov}\left[\underline{\Delta}, \hat{Q}\right]$ is equal to the diagonal of this matrix. The stated result follows by using these expressions within Propositions 2.1 and 2.2 as before. ∎