On Multilevel Best Linear Unbiased Estimators*

Daniel Schaden† and Elisabeth Ullmann†

Abstract. We present a general variance reduction technique for the estimation of the expectation of a scalar-valued quantity of interest associated with a family of model evaluations. The key idea is to reformulate the estimation as a linear regression problem. We then show that the associated estimators are variance minimal within the class of linear unbiased estimators. By solving a sample allocation problem we further construct a variance minimal, linear, and unbiased estimator for a given computational budget. We compare our proposed estimator to other multilevel estimators such as multilevel Monte Carlo, multifidelity Monte Carlo, and approximate control variates. In addition, we provide a sharp lower bound for the variance of any linear unbiased multilevel estimator, and show that our estimator approaches this bound in the infinite data limit. The results are illustrated by numerical experiments where the underlying output quantity of interest is generated by an elliptic partial differential equation.

Key words. Uncertainty quantification, partial differential equation, variance reduction, control variates, multilevel Monte Carlo, multifidelity Monte Carlo

AMS subject classifications. 35R60, 62J05 65N30, 65C05

1. Introduction.

The estimation of the expectation of a scalar-valued output quantity of interest (QoI) is a building block in computational statistics and uncertainty quantification (UQ). The Monte Carlo (MC) estimator is a linear, unbiased and robust estimator for this task. Robustness means that the rate of convergence of MC is independent of the smoothness of the QoI and its underlying dimension. Unfortunately, many QoIs in modern applications are associated with models that involve partial differential equations (PDEs) and are expensive to handle. Typically, MC estimators require tens of thousands of model evaluations and are thus computationally infeasible in PDE-based applications.

In the last decade, *multilevel estimators* have been developed to address this problem and design estimators with significantly smaller computational complexity. Multilevel estimators rely on the idea of *variance reduction* by linearly combining model evaluations of different resolutions or *fidelities*. They work with the target high fidelity model and families of low fidelity models that are correlated with the high fidelity model. Arguably the most prominent example to date is the multilevel Monte Carlo (MLMC) estimator [7, 8]. MLMC for PDE-based models has been initiated by Cliffe et al. [6], and has since been very popular with many recent works showing that MLMC has a smaller computational complexity compared to Monte Carlo. See e.g. [4, 13, 14] for forward UQ calculations, [2, 3, 11] for inverse UQ problems, and [1, 20] for optimization under uncertainty.

By construction, MLMC is a linear, unbiased estimator; it relies on a well-known telescop-

*Submitted to the editors DATE.

Funding: The authors gratefully acknowledge the support by the Deutsche Forschungsgemeinschaft (DFG) through the International Research Training Group IGDK 1754 "Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures".

†Department of Mathematics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching b. München, Germany, (schaden@ma.tum.de,eullmann@ma.tum.de).

ing sum. This is, in fact, a simple approach to guarantee unbiasedness, but it is by no means the only option. Indeed, many other linear unbiased multilevel estimators have been devised in recent years. One example are multifidelity Monte Carlo (MFMC) estimators [15, 16, 17] which rely on the idea of multiple control variates (CVs). Recently, Gorodetsky et al. [10] introduced approximate control variate estimators (ACVs). However, the authors of [10] correctly point out, that neither the telescoping sum approach in MLMC nor the CV approach in MFMC guarantee a small or even minimal estimator variance. The work in [10] addresses this defect by designing various ACV estimators. However, this again does not necessarily give the largest variance reduction possible. We address this problem and introduce a linear, unbiased, multilevel estimator with *guaranteed* smallest variance independently of the number of model evaluations. In addition, we provide a sharp lower bound on the variance of *any* linear unbiased multilevel estimator in the limit of infinitely many low fidelity model evaluations.

Our estimator is well known in statistics under the name of best linear unbiased estimator (BLUE). We remark that Monte Carlo is in fact a BLUE. Unfortunately, being a BLUE alone does not guarantee a feasible computational complexity. To address this, MLMC combines high and low fidelity model evaluations; in fact, MLMC linearly combines Monte Carlo estimators. However, such a linear combination of BLUEs is not necessarily a BLUE as well. In our work we construct a BLUE and show how to achieve a target variance with minimal computational complexity.

The idea of our proposed multilevel estimator is simple. We assume that we are given a certain number of model evaluations (samples) of models with different fidelities. The model evaluations are treated as *observations* of an underlying unknown true parameter. We then construct the BLUE for the true parameter using the observations. In other words: the estimator fits the observations "best", i.e., with minimal variance given the linearity constraint. This problem is a generalized least-squares problem. It can also be considered as generalized linear model where the model error has mean zero and a covariance matrix which depends on the correlations of the high and low fidelity models.

The ACV-type estimators in [10] are constructed and analyzed by partitioning the input samples into two ordered subsets where each ordered subset is associated with a control variate or level. It can be shown that MLMC and MFMC also fit into the ACV framework. However, this point of view does not emphasize a property that is essential for variance reduction, namely, the correlation between models in a family. Instead of grouping the *input* samples we form model groups with respect to the *outputs*. We present a framework based on model groups which share the exact same samples as input and thus produce a correlated output. This differs from the ACV framework yet it is sufficient to study a variety of linear unbiased estimators such as MLMC, MFMC, and ACVs.

The main contributions of this work are as follows: (i) a general framework for multilevel estimators, including multilevel Monte Carlo, multifidelity Monte Carlo, and approximate control variates, (ii) a novel multilevel best linear unbiased estimator (MBLUE) which achieves the minimal variance possible for any given configuration of model evaluations (samples), (iii) a specific MBLUE estimator termed SAOB with optimal sample allocation given a fixed computational budget, and (iv) a sharp lower bound on the variance of any linear, unbiased multilevel estimator in the infinite low fidelity data limit.

The remainder of this work is structured as follows. In Section 2 we introduce our novel multilevel estimator and prove some essential properties of it. In Section 3 we discuss sample allocations, and construct a variance minimal estimator given a fixed computational budget. In Section 4 we study the maximal possible variance reduction for the MBLUE. In Section 5 we discuss connections of the MBLUE to classical estimators in the literature, such as multilevel and multifidelity Monte Carlo [8, 15], control variates [9] and approximate control variates [10]. It turns out that our MBLUE satisfies the exact same lower bound for the variance reduction as the ACV estimators in [10]. We reproduce the result on the variance reduction for the optimal control variate given in [10]. Moreover, we prove that the ACV-IS estimator introduced in [10] is a BLUE. In Section 6 we conduct numerical experiments to support the theoretical results. Section 7 offers concluding remarks.

2. Multilevel best linear unbiased estimator.

Let $Z_1, \dots, Z_L$ denote scalar-valued random variables. In our context these are typically output quantities of interest associated with a family of models. The models are indexed by a certain *level* or *fidelity*, ordered from the coarsest level $\ell = 1$ to the finest level $\ell = L$. We wish to construct an estimator for $\mathbb{E}[Z_L]$ using samples of $Z_1, \dots, Z_L$. The expectation μ_ℓ, variance σ_ℓ^2, covariance $c_{\ell,j}$ and Pearson correlation coefficient $\rho_{\ell,j}$ associated with $Z_1, \dots, Z_L$ are defined as

$$\begin{aligned} &\mu_\ell := \mathbb{E}[Z_\ell], &&\sigma_\ell^2 := \mathbb{E}[|Z_\ell - \mathbb{E}[Z_\ell]|^2], &&\ell \in \{1, \dots, L\}, \\ &c_{\ell,j} := \mathbb{E}[(Z_\ell - \mathbb{E}[Z_\ell])(Z_j - \mathbb{E}[Z_j])], &&\rho_{\ell,j} := \frac{c_{\ell,j}}{\sigma_\ell \sigma_j}, &&\ell, j \in \{1, \dots, L\}, \end{aligned}$$

respectively. The model covariance matrix $C := (c_{\ell,j})_{\ell,j=1}^L$ and the vector of mean values is $\mu := (\mu_1, \dots, \mu_L)^T$. We assume that all those quantities exist and are finite. Our goal is to construct an unbiased estimator $\widehat{\mu}_L$ for $\mathbb{E}[Z_L]$ such that the variance of $\widehat{\mu}_L$ is minimal.

2.1. Definition of the estimator.

Let $(S^k)_{k=1}^K$ be a collection of the $K := 2^L - 1$ different non-empty subsets of $\{1, \dots, L\}$, that is,

$$S^k \in 2^{\{1,\dots,L\}} \setminus \{\emptyset\}, \quad S^i \neq S^j \text{ for all } i \neq j. \tag{2.1}$$

In the context of multilevel estimators each model group S^k tells us which outputs $Z_1, \dots, Z_L$ are statistically coupled by using the exact same sample as model input. For every index $k \in \{1, \dots, K\}$ we define the vectors Z^k, μ^k, η^k and matrix C^k as follows,

$$Z^k := (Z_\ell)_{\ell \in S^k}, \quad \mu^k := (\mu_\ell)_{\ell \in S^k}, \quad \eta^k := Z^k - \mu^k, \quad C^k := (c_{\ell,j})_{\ell,j \in S^k} = \mathrm{Cov}(\eta^k, \eta^k). \tag{2.2}$$

Furthermore, we define the restriction matrix $R^k \in \mathbb{R}^{|S^k| \times L}$ such that it holds

$$R^k v = (v_\ell)_{\ell \in S^k} \quad \text{for all } v \in \mathbb{R}^L. \tag{2.3}$$

The prolongation matrix is then defined as $P^k := (R^k)^T$. Combining (2.2) and (2.3) gives

$$Z^k = R^k \mu + \eta^k, \qquad k = 1, \dots, K. \tag{2.4}$$

Note that in statistics, a relation such as (2.4) is known as *linear model* (see e.g. [18]), where Z^k is a vector of observations, and $R^k \mu$ contains the parameters to be estimated (here the expected values of a subset of outputs $Z_1, \dots, Z_L$ with indices in S^k). Finally, η^k in (2.4) is a mean-zero, additive noise vector with covariance matrix C^k that is used to model observation errors. The parameter estimation problem associated with the linear model in (2.4) is also known as *linear regression problem*.

We now assume that for every index k we have $m^k \in \mathbb{N}_0$ independent samples of the random vector Z^k. Furthermore, we assume that samples of Z^ℓ and Z^j are statistically independent for $\ell \neq j$. The key idea of our multilevel sampler is to assemble the linear models in (2.4) for every k and every sample into a large, block-stuctured linear model of the following form

$$Y = H\mu + \varepsilon, \tag{2.5}$$

where

$$\begin{aligned} Y &:= (Y^k)_{k=1}^K, & H &:= (H^k)_{k=1}^K, & \varepsilon &:= (\varepsilon^k)_{k=1}^K,\\ Y^k &:= (Z^k(\omega_i^k))_{i=1}^{m^k}, & H^k &:= (R^k)_{i=1}^{m^k}, & \varepsilon^k &:= (\eta^k(\omega_i^k))_{i=1}^{m_k}, \end{aligned}$$

and the samples ω_i^k are i.i.d. Note that each vector Y^k contains samples of the output vector Z^k and is thus associated with a linear model in (2.4). Before we continue we illustrate (2.4) and (2.5) by an example.

Example 2.1 (Linear model). Let $L = 3$ and enumerate the model groups of $\{1, 2, 3\}$

$$\begin{aligned} &S^1 = \{1\}, && S^2 = \{2\} && S^3 = \{3\} && S^4 = \{1, 2\},\\ &S^5 = \{1, 3\}, && S^6 = \{2, 3\}, && S^7 = \{1, 2, 3\}. \end{aligned}$$

We are interested in the model groups given by S^1, S^4 and S^6, where (2.4) reads

$$\begin{aligned} Z_1 &= \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + (Z_1 - \mu_1) = R^1\mu + \eta^1, && \text{for } S^1,\\ \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} Z_1 - \mu_1 \\ Z_2 - \mu_2 \end{pmatrix} = R^4\mu + \eta^4, && \text{for } S^4,\\ \begin{pmatrix} Z_2 \\ Z_3 \end{pmatrix} &= \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} Z_2 - \mu_2 \\ Z_3 - \mu_3 \end{pmatrix} = R^6\mu + \eta^6, && \text{for } S^6. \end{aligned}$$

Now let $m^1 = m^4 = 1, m^6 = 2$ and $m^k = 0$ if $k \notin \{1, 4, 6\}$. Then, the block linear model in (2.5) reads

$$\begin{pmatrix} Z_1(\omega_1^1) \\ Z_1(\omega_1^4) \\ Z_2(\omega_1^4) \\ Z_2(\omega_1^6) \\ Z_3(\omega_1^6) \\ Z_2(\omega_2^6) \\ Z_3(\omega_2^6) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} Z_1(\omega_1^1) - \mu_1 \\ Z_1(\omega_1^4) - \mu_1 \\ Z_2(\omega_1^4) - \mu_2 \\ Z_2(\omega_1^6) - \mu_2 \\ Z_3(\omega_1^6) - \mu_3 \\ Z_2(\omega_2^6) - \mu_2 \\ Z_3(\omega_2^6) - \mu_3 \end{pmatrix} = \begin{pmatrix} R^1 \\ R^4 \\ R^6 \\ R^6 \end{pmatrix} \mu + \begin{pmatrix} \eta^1(\omega_1^1) \\ \eta^4(\omega_1^4) \\ \eta^6(\omega_1^6) \\ \eta^6(\omega_2^6) \end{pmatrix},$$

where $\omega_1^1, \omega_1^4, \omega_1^6, \omega_2^6$ are i.i.d. samples of some random source. Let us put this example into perspective with the MLMC estimator [7], which is defined as follows:

$$\widehat{\mu}_3^{\mathrm{MLMC}} := \frac{1}{m^6}\sum_{i=1}^{m^6}(Z_3(\omega_i^6) - Z_2(\omega_i^6)) + \frac{1}{m^4}\sum_{i=1}^{m^4}(Z_2(\omega_i^4) - Z_1(\omega_i^4)) + \frac{1}{m^1}\sum_{i=1}^{m^1} Z_1(\omega_i^1).$$

This estimator linearly combines m^6 samples of the group S^6, m^4 samples of S^4 and m^1 samples of S^1, respectively. However, MLMC is not derived from the perspective of a linear model of the form (2.5).

It is easy to verify that the linear model in (2.5) satisfies the following properties.

Proposition 2.2. *Let $G^k := \operatorname{diag}((C^k)_{i=1}^{m^k})$. Then, there holds*

$$\mathbb{E}[\varepsilon] = 0, \qquad \operatorname{Cov}(\varepsilon, \varepsilon) = \operatorname{diag}((G^k)_{k=1}^K), \qquad \mathbb{E}[Y] = H\mu.$$

We now define the key components of our multilevel estimator, the matrix $\Psi \in \mathbb{R}^{L\times L}$ and the vector $y \in \mathbb{R}^L$, as follows,

$$\Psi := \sum_{k=1}^K m^k P^k (C^k)^{-1} R^k, \qquad y := \sum_{k=1}^K P^k (C^k)^{-1} \sum_{i=1}^{m^k} Z^k(\omega_i^k). \tag{2.6}$$

Finally, our (linear) estimator $\widehat{\mu}^{\mathrm{B}}$ is defined such that it satisfies the equation

$$\Psi \widehat{\mu}^{\mathrm{B}} = y. \tag{2.7}$$

It turns out that $\widehat{\mu}^{\mathrm{B}}$ is well defined if we evaluate every model at least once. Formally, we define the set U of evaluated models as

$$U := \left\{ \ell \in \{1, \dots, L\} \,\middle|\, \text{there exists a } k \text{ with } m^k > 0 \text{ and } \ell \in S^k \right\}. \tag{2.8}$$

We then have the following result.

Lemma 2.3. *Let the matrices C^k be positive definite for every k and let $U = \{1, \dots, L\}$. Then the matrix Ψ in (2.6) is positive definite and thus $\widehat{\mu}^{\mathrm{B}}$ in (2.7) is well defined.*

Proof. Since $(P^k)^T = R^k$ and each matrix C^k is positive definite by assumption, the matrix $P^k(C^k)^{-1}R^k$ is positive semi-definite and hence Ψ is also positive semi-definite. It remains to show that if $v^T \Psi v = 0$ for some $v \in \mathbb{R}^L$ then $v = 0$. Observe that

$$0 = v^T \Psi v = \sum_{k=1}^K m^k v^T P^k (C^k)^{-1} R^k v.$$

Hence for all k this implies $m^k v^T (R^k)^T (C^k)^{-1} R^k v = 0$. Now, if $m^k > 0$ and since C^k is positive definite by assumption, and since R^k is the restriction operator, it follows $v_\ell = 0$ for all $\ell \in S^k$. Because $U = \{1, \dots, L\}$ we finally conclude $v = 0$. ∎

The estimator $\widehat{\mu}^{\mathrm{B}}$ delivers estimates of every component μ_ℓ, $\ell = 1, \ldots, L$. However, we are typically only interested in an estimate of the expectation of the finest model, μ_L. To construct such a partial estimator, let

$$\widehat{\mu}^{\mathrm{B}}_\ell := e_\ell^T \widehat{\mu}^{\mathrm{B}}, \tag{2.9}$$

where e_ℓ denotes the ℓth unit vector in $\mathbb{R}^L$, $\ell = 1, \ldots, L$. More generally, for an arbitrary vector $\alpha \in \mathbb{R}^L \backslash \{0\}$ we want to state conditions such that the estimator

$$\widehat{\mu}^{\mathrm{B}}_\alpha := \alpha^T \widehat{\mu}^{\mathrm{B}} \tag{2.10}$$

is well defined. Intuitively, if $\alpha_\ell = 0$ for some ℓ, then the estimator $\widehat{\mu}^{\mathrm{B}}_\alpha$ should be well defined even if we do not evaluate the ℓth model. We make this intuition precise.

Lemma 2.4 (Partial model estimation). *Let the matrices C^k be positive definite for every k and assume that we do not want to estimate expectations of models we do not evaluate, that is, $\alpha_\ell = 0$ for all $\ell \notin U$. Then $\widehat{\mu}^{\mathrm{B}}_\alpha$ is well defined as the limit*

$$\widehat{\mu}^{\mathrm{B}}_\alpha = \lim_{\delta \to 0^+} \left[\alpha^T (\Psi + \delta I)^{-1} \right] y = \alpha_U^T \Psi_{U,U}^{-1} y_U.$$

Proof. The claims follows by using the block diagonal form of

$$\Psi + \delta I = \begin{pmatrix} \delta I_{U^c,U^c} & 0 \\ 0 & \Psi_{U,U} + \delta I_{U,U} \end{pmatrix}.$$

■

2.2. Properties of the estimator. In this section we show that – by construction – the estimators $\widehat{\mu}^{\mathrm{B}}$ and $\widehat{\mu}^{\mathrm{B}}_\alpha$ are best linear unbiased estimators (BLUEs) for μ and $\alpha^T\mu$, respectively. Recall that a linear estimator $\widehat{\mu} = AY$ is an unbiased estimator for μ, if it holds

$$\mu = \mathbb{E}[\widehat{\mu}] = A\mathbb{E}[Y] = AH\mu$$

for every possible value of μ. We use the Gauss–Markov–Aitken Theorem (see e.g. [18, Theorem 4.4]) to show that $\widehat{\mu}^{\mathrm{B}}$ is the linear unbiased estimator for μ with the smallest variance, or simply that $\widehat{\mu}^{\mathrm{B}}$ is the BLUE.

Theorem 2.5. *Let the assumptions of Lemma 2.3 be true. Then, $\widehat{\mu}^{\mathrm{B}}$ is the BLUE for μ and the covariance matrix of $\widehat{\mu}^{\mathrm{B}}$ is $\mathrm{Cov}(\widehat{\mu}^{\mathrm{B}}, \widehat{\mu}^{\mathrm{B}}) = \Psi^{-1}$.*

Proof. The Gauss–Markov–Aitken Theorem states that the BLUE $\widehat{\mu}$ for the parameter vector μ in (2.5) satisfies

$$(H^T \mathrm{Cov}(\varepsilon, \varepsilon)^{-1} H)\widehat{\mu} = H^T \mathrm{Cov}(\varepsilon, \varepsilon)^{-1} Y,$$

and that the covariance of $\widehat{\mu}$ is

$$\mathrm{Cov}(\widehat{\mu}, \widehat{\mu}) = (H^T \mathrm{Cov}(\varepsilon, \varepsilon)^{-1} H)^{-1}.$$

A straightforward computation using Proposition 2.2 shows that $H^T \mathrm{Cov}(\varepsilon, \varepsilon)^{-1} H = \Psi$, and $H^T \mathrm{Cov}(\varepsilon, \varepsilon)^{-1} Y = y$, and thus $\widehat{\mu} = \widehat{\mu}^{\mathrm{B}}$. ■

Theorem 2.5 tells us that $\widehat{\mu}^{\mathrm{B}}$ is the BLUE for the entire vector μ. The next goal is to show that $\widehat{\mu}^{\mathrm{B}}_\alpha$ is a BLUE for the vector $\alpha^T\mu$. We call an estimator $\widehat{\mu}_\alpha = \beta Y$ a linear unbiased estimator for $\alpha^T\mu$ if it holds

$$\alpha^T\mu = \mathbb{E}[\widehat{\mu}_\alpha] = \beta\mathbb{E}[Y] = \beta H\mu$$

for every μ. We now show that $\widehat{\mu}^{\mathrm{B}}_\alpha$ is the BLUE for $\alpha^T\mu$.

Theorem 2.6. *Let the assumptions of Lemma 2.3 be true and let $\alpha \in \mathbb{R}^L$. Then $\widehat{\mu}^{\mathrm{B}}_\alpha$ is the BLUE for $\alpha^T\mu$ with variance*

$$\mathrm{Var}(\alpha^T\widehat{\mu}^{\mathrm{B}}) = \alpha^T\Psi^{-1}\alpha. \tag{2.11}$$

Proof. The proof follows [12, Appendix A] where the result is referred to as Gauss–Markov Theorem. Clearly the estimator $\widehat{\mu}^{\mathrm{B}}_\alpha$ is unbiased and linear. Let $\widehat{\mu}_\alpha$ be another linear unbiased estimator such that for a suitable vector β it holds

$$\begin{aligned} \widehat{\mu}_\alpha &= \widehat{\mu}^B_\alpha + \beta^T Y, && \text{(Linearity)}, \\ 0 &= \alpha^T\mu - \mathbb{E}[\widehat{\mu}_\alpha] = \beta^T H\mu, && \text{(Unbiasedness)}. \end{aligned} \tag{2.12}$$

Since the unbiasedness is assumed for every μ, we conclude $\beta^T H = 0$. Now the variance of $\widehat{\mu}_\alpha$ satisfies

$$\mathbb{E}[(\widehat{\mu}_\alpha - \alpha^T\mu)^2] = \mathbb{E}[(\alpha^T(\widehat{\mu}^{\mathrm{B}} - \mu))^2] + \mathbb{E}[(\beta^T Y)^2] + 2\mathbb{E}[\alpha^T(\widehat{\mu}^{\mathrm{B}} - \mu)\beta^T Y]. \tag{2.13}$$

The last term on the right-hand side in (2.13) satisfies

$$\begin{aligned} &\alpha^T\mathbb{E}[(H^T\operatorname{Cov}(\varepsilon,\varepsilon)^{-1}H)^{-1}H^T\operatorname{Cov}(\varepsilon,\varepsilon)^{-1}Y - \mu)\beta^T Y] \\ &= \alpha^T\mathbb{E}[(H^T\operatorname{Cov}(\varepsilon,\varepsilon)^{-1}H)^{-1}H^T\operatorname{Cov}(\varepsilon,\varepsilon)^{-1}(H\mu+\varepsilon) - \mu)\beta^T(H\mu+\varepsilon)] \\ &= \alpha^T\mathbb{E}[(H^T\operatorname{Cov}(\varepsilon,\varepsilon)^{-1}H)^{-1}H^T\operatorname{Cov}(\varepsilon,\varepsilon)^{-1}\varepsilon\varepsilon^T\beta] \\ &= \alpha^T(H^T\operatorname{Cov}(\varepsilon,\varepsilon)^{-1}H)^{-1}H^T\operatorname{Cov}(\varepsilon,\varepsilon)^{-1}\operatorname{Cov}(\varepsilon,\varepsilon)\beta \\ &= 0, \end{aligned}$$

where we used the fact that $\beta^T H = H^T\beta = 0$ and $\mathbb{E}[\varepsilon\varepsilon^T] = \operatorname{Cov}(\varepsilon,\varepsilon)$. We arrive at

$$\mathbb{E}[(\widehat{\mu}_\alpha - \alpha^T\mu)^2] = \mathrm{Var}(\widehat{\mu}^{\mathrm{B}}_\alpha) + \mathbb{E}[(\beta^T Y)^2]. \tag{2.14}$$

Thus the choice $\beta = 0$ minimizes the variance of $\widehat{\mu}^{\mathrm{B}}_\alpha$.

We now show the uniqueness of the BLUE. We rearrange the vector Y in (2.12) such that $Y = (Z^{k_i}(\omega_i))_{i=1}^N$ where the random variables ω_i are i.i.d. Moreover, let $\beta = (\beta^i)_{i=1}^N$. To minimize the variance in (2.14), necessarily $\mathbb{E}[(\beta^T Y)^2] = 0$. Since $\mathbb{E}[\beta^T Y] = 0$ it follows

$$0 = \mathbb{E}[(\beta^T Y)^2] = \mathrm{Var}(\beta^T Y) = \sum_{i=1}^N \mathrm{Var}((\beta^i)^T Z^{k_i}(\omega_i)) = \sum_{i=1}^N (\beta^i)^T C^{k_i}\beta^i.$$

Since all matrices C^k are positive definite by assumption, we obtain $\beta = 0$. That the variance $\mathrm{Var}(\alpha^T\widehat{\mu}^{\mathrm{B}}) = \alpha^T\Psi^{-1}\alpha$ follows straightforwardly. ■

The special case that some models are not evaluated can be analyzed analogously.

Theorem 2.7 (Partial model estimation). *Let the assumptions of Lemma 2.4 be true. Then $\widehat{\mu}_\alpha^{\mathrm{B}}$ is the BLUE for $\alpha^T\mu$ with variance*

$$\mathrm{Var}(\widehat{\mu}_\alpha^{\mathrm{B}}) = \alpha_U^T \Psi_{U,U}^{-1} \alpha_U.$$

Proof. Lemma 2.4 shows $\alpha^T\widehat{\mu}^{\mathrm{B}} = \alpha_U^T\widehat{\mu}_U^{\mathrm{B}}$. We rename models in U such that $U = \{1,\dots,L'\}$. Theorem 2.6 now shows the result. ∎

Remark 2.8. We expect that the assumption of positive definite C^k, which we make throughout this section and the rest of this paper, can be dropped at the cost of technically more involved proofs. We then have to work with a generalized inverse similar to [18, Section 3.2], where this was done in the context of ordinary least squares. We refrain from doing so to clearly convey the main ideas of our method.

3. Sample allocation.

The estimators $\widehat{\mu}^{\mathrm{B}}$ and $\widehat{\mu}_\alpha^{\mathrm{B}}$ in Section 2 have the smallest variance possible regardless of the sample allocation among the model outputs $Z_1,\dots,Z_L$. Of course, using more samples will in general further decrease the variance of $\widehat{\mu}^{\mathrm{B}}$ and $\widehat{\mu}_\alpha^{\mathrm{B}}$. However, in practice each model output comes with a certain computational cost. Moreover, the cost for a model evaluation can vary substantially among the levels. In this section we construct an optimal sample allocation. We determine the model groups and the number of samples for each group such that the resulting BLUE has the smallest variance and the total cost of the estimator does not exceed a given budget.

3.1. Integer sample allocation problem.

We assume throughout this section that each evaluation of Z_ℓ has a fixed cost $w_\ell \in \mathbb{R}_+$, $\ell = 1,\dots,L$. The cost for a single evaluation of the vector Z^k in (2.2) is denoted by

$$W^k := \sum_{\ell\in S^k} w_\ell.$$

Recall that the estimator $\widehat{\mu}_\alpha^{\mathrm{B}}$ in (2.10) (and also $\widehat{\mu}^{\mathrm{B}}$ in (2.7)) is constructed by forming $K = 2^L - 1$ groups of models that share m^k samples each, $k = 1,\dots,K$. Hence our optimization problem involves the variables $m^1,\dots,m^K$, which we collect in a vector $m := (m^1,\dots,m^K)^T \in \mathbb{N}_0^K$. We define the cost functional J as variance of the BLUE using the sample allocation m,

$$J(m) := \mathrm{Var}(\widehat{\mu}_\alpha^{\mathrm{B}}(m)),$$

where the dependence of $\widehat{\mu}_\alpha^{\mathrm{B}}$ on m is made explicit. If $\widehat{\mu}_\alpha^{\mathrm{B}}(m)$ is not well defined according to Lemma 2.4, that is, we do not evaluate a model ℓ but $\alpha_\ell \neq 0$, we set $J(m) := +\infty$. The goal is now to minimize J given a maximal cost $p > 0$ for the estimator. In addition, we select a *coupling number* $\kappa \in \mathbb{N}$ which limits the number of models within a group S^k. The integer sample allocation problem is then given as follows:

$$\left\{\begin{array}{ll} \min_{m\in\mathbb{N}_0^K} \quad J(m) & \\ \sum_{k=1}^K m^k W^k \le p, & \\ m^k = 0, & \text{if } |S^k| > \kappa. \end{array}\right. \tag{3.1}$$

We now summarize the basic properties of this optimization problem.

Lemma 3.1. *Let the matrices C^k be positive definite for every k and assume that we can evaluate required models at least once. That is,*

$$\sum_{\{\ell\in\{1,\dots,L\}\,|\,\alpha_\ell\neq 0\}} w_\ell \le p.$$

Then there is at least one sample allocation m satisfying (3.1) with $J(m)\neq+\infty$. Furthermore, the set of feasible sample allocations is bounded and $J(m)>0$ for all m.

Proof. Consider the sample allocation m, where $m^k=1$ if $S^k=\{\ell\in\{1,\dots,L\}\,|\,\alpha_\ell\neq 0\}$ and $m^k=0$ otherwise. According to Theorem 2.7 this is a well defined BLUE and thus $J(m)\neq+\infty$. The cost constraint ensures that the set of feasible sample allocations is bounded. Let U denote the set of used models in (2.8). Since $\Psi_{U,U}$ is positive definite and $\alpha_U\neq 0$, Theorem 2.7 tells us that

$$J(m)=\mathrm{Var}(\widehat{\mu}^{\mathrm{B}}_\alpha(m))=\alpha_U^T\Psi_{U,U}^{-1}\alpha_U>0.$$

∎

Now let m_* be a minimizer of (3.1). We define a sample allocation optimal BLUE with coupling number κ, denoted by the superscript (SAOB,κ), as follows,

$$\widehat{\mu}^{\mathrm{SAOB},\kappa}_\alpha := \widehat{\mu}^{\mathrm{B}}_\alpha(m_*). \tag{3.2}$$

If no coupling restriction applies, i.e., $\kappa=+\infty$, we drop the superscript κ and simply denote the estimator by SAOB. We now show that SAOB is variance minimal under all linear unbiased estimators with costs not exceeding the budget p.

Theorem 3.2. *Let the matrices C^k be positive definite for every k and let $\widehat{\mu}_\alpha$ be a linear unbiased estimator for $\alpha^T\mu$ using only samples from models $1,\dots,L$ with total cost bounded by p. Then, it holds*

$$\mathrm{Var}(\widehat{\mu}_\alpha)\ge\mathrm{Var}(\widehat{\mu}^{\mathrm{SAOB}}_\alpha).$$

Proof. Let us denote the sample allocation of $\widehat{\mu}_\alpha$ with $m(\widehat{\mu}_\alpha)$. W.l.o.g. we assume that $\mathrm{Var}(\widehat{\mu}_\alpha)\neq+\infty$, otherwise there is nothing to show. This together with the unbiasedness and Theorem 2.7 gives $\mathrm{Var}(\widehat{\mu}_\alpha)\ge\mathrm{Var}(\widehat{\mu}^{\mathrm{B}}_\alpha(m(\widehat{\mu}_\alpha)))=J(m(\widehat{\mu}_\alpha))$. Finally, observe that $J(m(\widehat{\mu}_\alpha))\ge J(m_*)=\mathrm{Var}(\widehat{\mu}^{\mathrm{SAOB}}_\alpha)$, since both $m(\widehat{\mu}_\alpha)$ and m_* are feasible sample allocations in (3.1). This concludes the proof. ∎

3.2. Optimality conditions for the relaxed problem. Throughout the rest of this section we relax the integer constraint $m\in\mathbb{N}_0^K$ and work with $m\in\mathbb{R}^K_{\ge 0}$. Since we wish to minimize the variance of $\widehat{\mu}^{\mathrm{B}}_\alpha$, combining (2.6) and (2.11) gives the cost functional

$$J_\delta(m):=\alpha^T\left(\sum_{k=1}^K m^kP^k(C^k)^{-1}R^k+\delta I\right)^{-1}\alpha, \tag{3.3}$$

where $\delta > 0$ is fixed. For $\delta = 0$, if the BLUE is well defined, we have

$$J_0(m) = \mathrm{Var}(\widehat{\mu}_\alpha^{\mathrm{B}}(m)). \tag{3.4}$$

Remark 3.3 (Choice of δ). Adding the matrix δI in (3.3) ensures that the matrix inversion is well defined. Note that the matrix Ψ in (2.6) is positive-definite, if *all* models $Z_1, \dots, Z_L$ are used in the multilevel estimator, i.e., $U = \{1, \dots, L\}$ in (2.8) (see Lemma 2.3). In this case, we can work with $\delta = 0$. Otherwise, if $U \neq \{1, \dots, L\}$, we work with $\delta > 0$.

In summary, the relaxed sample allocation problem reads

$$\begin{cases} \min_{m \in \mathbb{R}^K_{\geq 0}} \quad J_\delta(m) \\ \sum_{k=1}^K m^k W^k = p, \\ m^k = 0, \qquad\qquad \text{if } |S^k| > \kappa. \end{cases} \tag{3.5}$$

Here the cost constraint is now an equality constraint, and any optimizer of this problem will satisfy this constraint with equality.

Next we derive some basic properties of the cost functional J_δ in (3.4). To this end we introduce the following notation. Let $C_{Q,Q} \in \mathbb{R}^{|Q| \times |Q|}$ denote the principal submatrix of the model covariance matrix C with row and column indices in the set Q. Moreover, let $\beta_Q \in \mathbb{R}^{|Q|}$ denote the subvector of $\beta \in \mathbb{R}^L$ with row indices in the set Q.

Lemma 3.4 (Properties of J_δ).

(i) For any $\delta, \lambda > 0$ we have $J_\delta(\lambda m) = J_{\delta/\lambda}(m)/\lambda$. This property also holds for $\delta = 0$ if J_0 is well defined.

(ii) Using more samples does not increase the variance, that is, J_δ is monotonically decreasing in each component of m for any $\delta \geq 0$.

(iii) More coupling among the model outputs does not increase the variance, that is, if $S^k \subseteq S^j$ then

$$J_\delta(m + \lambda e_k) \geq J_\delta(m + \lambda e_j), \qquad \text{for all } \lambda \geq 0, \quad \delta \geq 0.$$

Proof. Property (i) follows from the definition of J_δ in (3.3). Property (ii) and (iii) for integer values follows from the fact that using more observations in the linear regression problem (2.5) does not increase the variance of the BLUE. Formally, we have for independent events ω_i and suitable indices $k_i \in \{1, \dots, K\}$,

$$J_\delta(m) = \sum_{i=1}^N \mathrm{Var}((\beta^i)^T Z^{k_i}(\omega_i)) = \sum_{i=1}^N (\beta^i)^T C^{k_i} \beta^i,$$

where the vectors β^i are always chosen to minimize $J_\delta(m)$, and to satisfy a bias constraint. Adding another sample ω_i increases the number of degrees of freedom in this minimization problem by one. This in turn cannot increase J_δ, and thus (ii) is shown. We extend this result to non-integer samples by observing that we can replace one sample of Z^k by four independent samples of the modified model $2Z^k$, since this does not change the variance. Thus, every fractional increase ξ of m^k can be viewed as one independent additional observation of the model $\xi^{-1/2}Z^k$, which does not increase J_δ. Similarly, for (iii) with $Q := S^j \setminus S^k$ we have

$$(\beta^i)^T C^j \beta^i = (\beta^i_{S^k})^T C^k \beta^i_{S^k} + 2(\beta^i_{S^k})^T C_{S^k,Q}\beta^i_Q + (\beta^i_Q)^T C_{Q,Q}\beta^i_Q.$$

That is, the components of β^i_Q are additional degrees of freedom. ■

In addition, it is straightforward to verify the following: If there exists a model Z_ℓ that is not used, i.e., $\ell \notin U$, and if $\alpha_\ell \neq 0$, it holds

$$\lim_{\delta\to 0^+} J_\delta(m) = +\infty.$$

This tells us that $\alpha_\ell \neq 0$ implies $\ell \in U$ for sufficiently small δ for the optimal solution. Unless noted otherwise, all results in this section are stated for $\delta > 0$. We now show that (3.5) is a well posed *convex* minimization problem.

Theorem 3.5. *Let the matrices C^k be positive definite for every k. Then the cost functional J_δ in (3.3) is convex on the feasible set defined by the constraints in (3.5). Furthermore, the MC estimator with*

$$m^k = \begin{cases} p/W^k, & \text{for } k \text{ with } S^k = \{\ell \in \{1,\dots,L\} \colon \alpha_\ell \neq 0\}, \\ 0, & \text{otherwise.} \end{cases}$$

is feasible and (3.5) has at least one minimizer m_*.

Proof. The MC estimator satisfies the constraints in (3.5) and thus the feasible set is non-empty. To show the convexity, let m_1 and m_2 be two feasible allocations, and let $\lambda \in [0,1]$. Then the convex combination $m_\lambda := (1-\lambda)m_1 + \lambda m_2$ is again a feasible allocation.

Next we investigate the convexity of J_δ. Recall that for positive definite matrices X, Y the matrix $Z(\lambda) = (1-\lambda)X + \lambda Y$ is positive definite. In addition, it is easy to see that the function

$$\varphi(\lambda) = \alpha^T Z(\lambda)^{-1}\alpha$$

is a convex function in λ for any vector α of suitable length. Now, using the definition of Ψ in (2.6), it is easy to verify that it holds $\Psi(m_\lambda) = (1-\lambda)\Psi(m_1) + \lambda\Psi(m_2)$. Hence

$$\begin{aligned} J_\delta(m_\lambda) &= \alpha^T(\Psi(m_\lambda) + \delta I)^{-1}\alpha = \alpha^T((1-\lambda)(\Psi(m_1)+\delta I) + \lambda(\Psi(m_2)+\delta I))^{-1}\alpha \\ &= \alpha^T Z(\lambda)^{-1}\alpha = \varphi(\lambda) \end{aligned}$$

where $X = \Psi(m_1) + \delta I$ and $Y = \Psi(m_2) + \delta I$ are positive definite matrices. Hence J_δ is convex. Finally, the constraint

$$\sum_{k=1}^{K} m^k W^k = p$$

ensures that $m^k \leq c$ for all $k = 1, \dots, K$ for some constant c. This shows the existence of a minimizer m_*. ■

Note that the function J_δ is in general not strictly convex, since the mapping $\Psi = \Psi(m)$ is not necessarily injective. Importantly, Theorem 3.5 allows us to write down the KKT conditions for the optimization problem (3.5), since the constraints are linear (see [5, Section 5.5.3]). To this end, we introduce the Lagrange multipliers ξ^k, $k = 1, \dots, K$, and ξ^{cost}. The optimality conditions read

$$(3.6)\qquad \begin{aligned} \alpha^T(\Psi(m)+\delta I)^{-1}P^k(C^k)^{-1}R^k(\Psi(m)+\delta I)^{-1}\alpha &= \xi^{cost}W^k - \xi^k, && k = 1,\dots,K,\\ \sum_{k=1}^K m^k W^k &= p, &&\\ m^k \ge 0, \quad \xi^k \ge 0, \quad \xi^k m^k &= 0, && k = 1,\dots,K,\\ m^k &= 0, && \text{if } |S^k| > \kappa. \end{aligned}$$

3.3. Number of model groups. Observe that the number of unknowns in the optimization problem (3.5) is equal to $2^L - 1$ if $\kappa = +\infty$. Otherwise, it is of the order $\mathcal{O}(L^\kappa)$, i.e., exponential in the number of models L in any case. We now show that we can restrict the evaluation to at most L different groups. Formally, we define the set of active model groups

$$(3.7)\qquad A_{>0}(m) := \{k \in \{1,\dots,K\} \,|\, m^k > 0\}.$$

We can always find a suitable allocation m with $|A_{>0}(m)| \le L$.

Theorem 3.6. *Let the matrices C^k be positive definite for every k and let m be a feasible allocation of (3.5). Then there exists a feasible allocation m' with $|A_{>0}(m')| \le L$ satisfying*

$$J_\delta(m') \le J_\delta(m).$$

In particular, there exists a minimizer m_* of (3.5) with $|A_{>0}(m_*)| \le L$.

Proof. Let m be a feasible allocation such that w.l.o.g. $m^1, \dots, m^{L+1} > 0$. The basic idea of the proof is to find a direction t along which J_δ remains constant and the cost does not increase. We then show that the allocation $m + st$ evaluates at least one less model group if s is chosen suitably.

Since $m^1, \dots, m^{L+1} > 0$, by a dimension counting argument, there exists a direction $t \neq 0$, such that with $x = (\Psi(m) + \delta I)^{-1}\alpha$ it holds

$$(3.8)\qquad \sum_{\ell=1}^{L+1} t^\ell P^\ell (C^\ell)^{-1} R^\ell x = \sum_{\ell=1}^{L+1} t^\ell x^\ell = 0,$$

where we defined $x^\ell := P^\ell (C^\ell)^{-1} R^\ell x$. This is possible since $x^\ell \in \mathbb{R}^L$ and we define $t^k := 0$ if $k > L + 1$. W.l.o.g. we assume that the cost along t does not increase

$$(3.9)\qquad \sum_{\ell=1}^{L+1} t^\ell W^\ell \le 0.$$

Otherwise we change the direction of t by working with $-t$. Since $W^\ell > 0$ we conclude that there exists an index $\ell \in \{1, \dots, L+1\}$ with $t^\ell < 0$. Thus $s_{\max}$ is well defined, that is,

$$(3.10)\qquad s_{\max} := \max\{s \ge 0 \,|\, m^k + st^k \ge 0, \text{ for all } k \in \{1,\dots,K\}\} < +\infty.$$

Since $m^1, \dots, m^{L+1} > 0$ by assumption, we obtain $s_{\max} > 0$. Furthermore, the vector $m + s_{\max} t$ has at least one index $\ell \in \{1, \dots, L+1\}$ with $m^\ell + s_{\max} t^\ell = 0$, since s is maximized by (3.10). Together with $t^k = 0$ for $k > L+1$ we obtain

$$|A_{>0}(m + s_{\max} t)| \leq |A_{>0}(m)| - 1. \tag{3.11}$$

We use (3.8) and $t^k = 0$ for $k > L+1$ to conclude that

$$\begin{aligned} \alpha = (\Psi(m) + \delta I)x &= \sum_{k=1}^{K} m^k P^k (C^k)^{-1} R^k x + \delta x \\ &= \sum_{k=1}^{K} (m^k + s_{\max} t^k) P^k (C^k)^{-1} R^k x + \delta x = (\Psi(m + s_{\max} t) + \delta I)x. \end{aligned}$$

Hence the functional J_δ is constant along the direction t, meaning that

$$J_\delta(m) = \alpha^T (\Psi(m) + \delta I)^{-1} \alpha = \alpha^T x = \alpha^T (\Psi(m + s_{\max} t) + \delta I)^{-1} \alpha = J_\delta(m + s_{\max} t). \tag{3.12}$$

Here it is crucial to remark that $m + s_{\max} t \geq 0$ and thus $\Psi(m + s_{\max} t) + \delta I$ is invertible. We collect our findings in (3.12), (3.11) and (3.9):

$$\begin{aligned} J_\delta(m + s_{\max} t) &= J_\delta(m), \\ |A_{>0}(m + s_{\max} t)| &\leq |A_{>0}(m)| - 1, \\ \sum_{k=1}^{K} (m^k + s_{\max} t^k) W^k &\leq \sum_{k=1}^{K} m^k W^k = p. \end{aligned}$$

Note that the cost constraint in the last line above can be achieved with equality if we rescale $m + s_{\max} t$ to a larger value potentially decreasing the variance by Lemma 3.4 (*ii*).

In summary, starting from a feasible sample allocation m, we found a new feasible allocation $m + s_{\max} t$ that uses one model group less and does not increase J_δ. We can now repeat the process outlined in this proof with the sample allocation $m + s_{\max} t$ until we obtain a feasible sample allocation $\widetilde{m}$ where the initial assumptions fails, that is $\widetilde{m}^k > 0$ for at most L different values of k. ■

4. Lower bound for the variance.

In this section we derive a lower bound on the variance of $\widehat{\mu}_\alpha^{\mathrm{B}}$ in (2.10). To avoid the trivial lower bound equal to zero, we consider a specific sample allocation for all estimators in this section. We define the sets of models $Q, Q_\infty \subseteq \{1, \dots, L\}$ such that $Q \cup Q_\infty = \{1, \dots, L\}$ and $Q \not\subseteq Q_\infty$. Moreover, let $N, M \in \mathbb{N}$. We consider the sample allocation

$$m^k(Q, Q_\infty, N) := \begin{cases} N, & \text{if } S^k \subseteq Q_\infty, \\ M, & \text{if } S^k = Q, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for all } k = 1, \dots, K. \tag{4.1}$$

In (4.1) we distinguish models with indices in Q which are evaluated M-times, and models with indices in Q_∞ which are evaluated N-times. Our goal is to study the variance of $\widehat{\mu}_\alpha^{\mathrm{B}}$ in the limit $N \to +\infty$, denoted by

$$\gamma(\alpha, Q, Q_\infty) := \lim_{N\to+\infty} \operatorname{Var}(\widehat{\mu}_\alpha^{\mathrm{B}}(m(Q, Q_\infty, N))). \tag{4.2}$$

Note that this models a situation which is often encountered in practice. Models with indices in $Q \setminus Q_\infty$ are high fidelity, yet expensive, allowing only a fixed number M of evaluations. On the other hand, models with indices in Q_∞ are cheap to evaluate, and in the limit $N \to +\infty$ we assume that infinitely many evaluations are possible. If $Q \subseteq Q_\infty$ then $Q \cup Q_\infty = \{1, \ldots, L\}$ shows the trivial bound of zero variance, hence the restriction $Q \not\subseteq Q_\infty$ is assumed. This setup follows the analysis of Gorodetsky et al. [10]. Note that since $U = Q \cup Q_\infty = \{1, \ldots, L\}$ by assumption we can work with $\delta = 0$ in the estimator variance (3.3). The key observation in our analysis is the fact that the limit in (4.2) can be formulated in terms of a minimization problem.

Lemma 4.1 (Limit of $\operatorname{Var}(\widehat{\mu}_\alpha^{\mathrm{B}})$). *Let the matrices C^k be positive definite for every k. Then there holds*

$$\gamma(\alpha, Q, Q_\infty) = \frac{1}{M} \min_{\substack{\overline{\alpha} \in \mathbb{R}^L,\\ \overline{\alpha}_\ell = \alpha_\ell,\ \ell \notin Q_\infty.}} \overline{\alpha}_Q^T C_{Q,Q} \overline{\alpha}_Q. \tag{4.3}$$

Proof. We write down the estimator $\widehat{\mu}_\alpha^{\mathrm{B}} = \alpha^T \Psi^{-1} y$ using the events $(\omega_i)_{i=1}^M$ as follows,

$$\widehat{\mu}_\alpha^{\mathrm{B}}(m(Q, Q_\infty, N)) = \sum_{\ell \in Q} \beta_\ell \frac{1}{M} \sum_{i=1}^M Z_\ell(\omega_i) + r(N), \tag{4.4}$$

where r is a remainder term depending only on models in Q_∞ and $\beta \in \mathbb{R}^L$ is a suitably chosen vector. Because $\widehat{\mu}_\alpha^{\mathrm{B}}$ is unbiased and since r contains only models in Q_∞ we conclude that $\beta_\ell = \alpha_\ell$ for all $\ell \notin Q_\infty$. Note that the remainder r in (4.4) is statistically independent of the first term. Hence

$$\begin{aligned} \operatorname{Var}(\widehat{\mu}_\alpha^{\mathrm{B}}(m(Q, Q_\infty, N))) &= \operatorname{Var}\left(\sum_{\ell \in Q} \beta_\ell \frac{1}{M} \sum_{i=1}^M Z_\ell(\omega_i)\right) + \operatorname{Var}(r(N)) \\ &\geq \min_{\substack{\overline{\alpha} \in \mathbb{R}^L,\\ \overline{\alpha}_\ell = \alpha_\ell,\ \ell \notin Q_\infty.}} \operatorname{Var}\left(\sum_{\ell \in Q} \overline{\alpha}_\ell \frac{1}{M} \sum_{i=1}^M Z_\ell(\omega_i)\right), \end{aligned} \tag{4.5}$$

where we dropped the positive variance and minimize over some β_ℓ. This shows the lower bound for $\gamma(\alpha, Q, Q_\infty)$. To show the upper bound, let ω_i^1, ω_i^2, $i = 1, \ldots, N/2$, denote events occurring in the remainder r that are statistically independent of $(\omega_i)_{i=1}^M$, and that are also mutually statistically independent. Consider the following estimator:

$$\widehat{\mu}_\alpha := \sum_{\ell \in Q} \overline{\alpha}_\ell \frac{1}{M} \sum_{i=1}^M Z_\ell(\omega_i) + \sum_{\ell \in Q \cap Q_\infty} (\alpha_\ell - \overline{\alpha}_\ell) \frac{2}{N} \sum_{i=1}^{N/2} Z_\ell(\omega_i^1) + \sum_{\ell \in Q^c} \alpha_\ell \frac{2}{N} \sum_{i=1}^{N/2} Z_\ell(\omega_i^2),$$

where $\overline{\alpha}_\ell = \alpha_\ell$ for $\ell \notin Q_\infty$. Note that $\widehat{\mu}_\alpha$ is a linear and unbiased estimator for $\alpha^T\mu$. Indeed, by construction it holds

$$\mathbb{E}[\widehat{\mu}_\alpha] = \sum_{\ell\in Q} \overline{\alpha}_\ell \mu_\ell + \sum_{\ell\in Q\cap Q_\infty} (\alpha_\ell - \overline{\alpha}_\ell)\mu_\ell + \sum_{\ell \in Q^c} \alpha_\ell \mu_\ell,$$

and by considering the cases $Q\cap Q_\infty = \emptyset$ and $Q \cap Q_\infty \neq \emptyset$ it follows $\mathbb{E}[\widehat{\mu}_\alpha] = \alpha^T\mu$ in any case. Now, since $\widehat{\mu}^{\mathrm{B}}_\alpha$ is the BLUE using more samples, from Lemma 3.4 (ii) it follows

$$\begin{aligned}\operatorname{Var}(\widehat{\mu}^{\mathrm{B}}_\alpha(m(Q,Q_\infty,N))) \leq \operatorname{Var}(\widehat{\mu}_\alpha) = \frac{1}{M}\overline{\alpha}_Q^T C_{Q,Q}\overline{\alpha}_Q \\ + \frac{2}{N}(\alpha-\overline{\alpha})^T_{Q\cap Q_\infty} C_{Q\cap Q_\infty, Q\cap Q_\infty}(\alpha - \overline{\alpha})_{Q\cap Q_\infty} + \frac{2}{N}\alpha^T_{Q^c} C_{Q^c,Q^c}\alpha_{Q^c}.\end{aligned}$$

W.l.o.g. we may assume $\|\overline{\alpha}\| < c$ for sufficiently large c. Thus $\operatorname{Var}(\widehat{\mu}_\alpha)$ converges uniformly for $N \to +\infty$ w.r.t. $\overline{\alpha}$. This allows us to exchange the minimum and limit operator, arriving at

$$\lim_{N\to+\infty} \min_{\substack{\overline{\alpha}\in\mathbb{R}^L,\\ \overline{\alpha}_\ell = \alpha_\ell,\ \ell\notin Q_\infty.}} \operatorname{Var}(\widehat{\mu}_\alpha) = \min_{\substack{\overline{\alpha}\in\mathbb{R}^L,\\ \overline{\alpha}_\ell = \alpha_\ell,\ \ell\notin Q_\infty.}} \lim_{N\to+\infty} \operatorname{Var}(\widehat{\mu}_\alpha).$$

This shows the upper bound for $\gamma(\alpha, Q, Q_\infty)$ and concludes the proof. ■

Remark 4.2 (Schur complement). $\gamma(\alpha,Q,Q_\infty)$ solves the quadratic minimization problem in (4.3) under equality constraints. Hence it can equivalently be written in terms of a Schur complement of C. Indeed, with $V := Q\setminus Q_\infty$, and $W := Q\cap Q_\infty$ it holds

$$\gamma(\alpha,Q,Q_\infty) = \alpha_V^T(C_{V,V} - C_{V,W}C^{-1}_{W,W}C_{W,V})\alpha_V^T/M. \tag{4.6}$$

This is in fact the same expression obtained by Gorodetsky et al. [10, Proposition 2.2] for $\alpha = e_L$, $Q_\infty = \{1,\dots,L-1\}$, $Q = \{1,\dots,L\}$, $M = 1$ and $\operatorname{Var}(Z_L) = 1$.

We now derive a lower bound on the variance of a general linear unbiased estimator $\widehat{\mu}_\alpha$ for $\alpha^T\mu$ under some conditions on the sample allocation $m(\widehat{\mu}_\alpha)$. In particular, the bound holds for the BLUE $\widehat{\mu}^{\mathrm{B}}_\alpha$ with a suitable sample allocation.

Theorem 4.3 (Bound for $\operatorname{Var}(\widehat{\mu}_\alpha)$). *Let the matrices C^k be positive definite for every k and let $\widehat{\mu}_\alpha$ be a linear unbiased estimator for $\alpha^T\mu$ with the sample allocation $m(\widehat{\mu}_\alpha)$ such that*

$$m^k(\widehat{\mu}_\alpha) > 0 \quad \textit{implies } S^k \subseteq Q \textit{ or } S^k \subseteq Q_\infty. \tag{4.7}$$

Then, letting $M := \sum_{S^k\subseteq Q} m^k(\widehat{\mu}_\alpha)$, the estimator $\widehat{\mu}_\alpha$ satisfies

$$\operatorname{Var}(\widehat{\mu}_\alpha) \geq \gamma(\alpha,Q,Q_\infty).$$

Proof. It is sufficient to argue that for N large enough it holds

$$\operatorname{Var}(\widehat{\mu}_\alpha) \geq \operatorname{Var}(\widehat{\mu}^{\mathrm{B}}_\alpha(m(\widehat{\mu}_\alpha))) \geq \operatorname{Var}(\widehat{\mu}^{\mathrm{B}}_\alpha(m(Q,Q_\infty,N))). \tag{4.8}$$

The first inequality in (4.8) follows independently of N since $\widehat{\mu}_\alpha^{\mathrm{B}}(m(\widehat{\mu}_\alpha))$ is a BLUE with minimal variance and identical sample allocation. Now, we use Assumption (4.7) which tells us that all indices of non-trivial model groups in the estimator $\widehat{\mu}_\alpha$ are completely contained in either Q or Q_∞. First, we replace all samples in $m(\widehat{\mu}_\alpha)$ of the form $S^k \subseteq Q$ with M samples of the form $S^{k_Q} = Q$. By Lemma 3.4 (*iii*) this does not increase the variance of the BLUE. Finally, we replace the remaining samples with $S^k \subseteq Q_\infty$ by N samples, where

$$N := \sum_{S^k \subseteq Q_\infty} m^k(\widehat{\mu}_\alpha).$$

This yields the sample allocation $m(Q, Q_\infty, N)$. Again, Lemma 3.4 (*ii*) tells us that the variance of the BLUE does not increase. Hence, $\mathrm{Var}(\widehat{\mu}_\alpha) \geq \mathrm{Var}(\widehat{\mu}_\alpha^{\mathrm{B}}(m(Q, Q_\infty, N))) \to \gamma(\alpha, Q, Q_\infty)$ in the limit $N \to +\infty$. ∎

Now we formulate a corollary of Theorem 4.3 for the case $\alpha = e_L$ and $L \in Q$, that is, Z_L is a high fidelity model. In addition, we assume $Q_\infty = \{1, \ldots, L-1\}$.

Corollary 4.4 (Bound for $\mathrm{Var}(\widehat{\mu}_L)$). *Let the matrices C^k be positive definite for every k, let $\alpha = e_L$, $Q_\infty = \{1, \ldots, L-1\}$, and $Q \subseteq \{1, \ldots, L\}$. Let $\widehat{\mu}_L$ be a linear unbiased such that the assumptions of Theorem 4.3 are satisfied. Then it holds*

$$\mathrm{Var}(\widehat{\mu}_L) \geq \gamma(e_L, Q, Q_\infty) \geq \gamma(e_L, \{1, \ldots, L\}, Q_\infty) =: \gamma_{\min}. \tag{4.9}$$

Proof. The first inequality in (4.9) was proved in Theorem 4.3. The second inequality in (4.9) follows from Lemma 3.4 (*iii*) before proceeding to the limit $N \to +\infty$, since

$$J(m(Q, Q_\infty, N)) \geq J(m(\{1, \ldots, L\}, Q_\infty, N)).$$

∎

Remark 4.5. The restriction $Q \cup Q_\infty = \{1, \ldots, L\}$ can be removed if $\alpha_\ell = 0$ for all $\ell \notin Q \cup Q_\infty$. In this case however, to derive lower bounds for the variance, the estimator $\widehat{\mu}_\alpha$ cannot use models $\ell \notin Q \cup Q_\infty$. If this condition is satisfied, we exclude unused models and follow the same steps as in the proofs of Lemma 4.1, Theorem 4.3 and Corollary 4.4.

5. Comparison to other linear unbiased estimators.

In this section we discuss other estimators in the literature, focusing on linear and unbiased estimators. We will see that alternative multilevel estimators are in general not BLUEs. However, they can be cast into our framework in Subsection 2.1, where we form groups of model outputs $Z_1, \ldots, Z_L$ sharing the exact same random inputs.

First, we observe that the Monte Carlo (MC) estimator for the expectation μ_L of the model output Z_L is a BLUE. The MC estimator only evaluates the model group $S^1 = \{L\}$ using m^1 input samples. Thus, using (2.6), we obtain

$$\Psi_{L,L} = m^1(\sigma_L^2)^{-1}, \quad y_L = (\sigma_L^2)^{-1} \sum_{i=1}^{m^1} Z_L(\omega_i^1),$$

where we exclude the trivial case of Z_L having zero variance. Now we apply Lemma 2.4 to arrive at the familiar form

$$\widehat{\mu}_L^{\mathrm{MC}} := \widehat{\mu}_L^{\mathrm{B}} = \Psi_{L,L}^{-1} y_L = \frac{1}{m^1}\sum_{i=1}^{m^1} Z_L(\omega_i^1).$$

Since this estimator is a BLUE not using coarse models, the variance bound $\gamma(e_L, Q, Q_\infty)$ in is achieved with equality and Lemma 4.1 shows that

$$\mathrm{Var}(\widehat{\mu}_L^{\mathrm{MC}}) = \gamma(e_L, \{L\}, \{1,\dots,L-1\}) = \mathrm{Var}(Z_L)/M.$$

5.1. Multilevel Monte Carlo. Next, we consider the Multilevel Monte Carlo (MLMC) estimator in the works of Giles [7, 8]. For μ_L, the MLMC estimator is defined as

$$\widehat{\mu}_L^{\mathrm{MLMC}} := \sum_{\ell=1}^{L} \widehat{E}^{MC}(Z_\ell - Z_{\ell-1}) = \sum_{\ell=1}^{L}\frac{1}{n^\ell}\sum_{i=1}^{n^\ell}(Z_\ell(\omega_i^\ell) - Z_{\ell-1}(\omega_i^\ell)), \tag{5.1}$$

where $Z_0 := 0$. Here the differences are estimated with independent MC estimators each using n^ℓ samples. This is clearly a linear and unbiased estimator for μ_L for arbitrary values of μ.

Remark 5.1 (Sample allocation of MLMC). The MLMC estimator $\widehat{\mu}_L^{\mathrm{MLMC}}$ in (5.1) fits into our framework by defining the model groups $S^1 = \{1\}$, $S^2 = \{1,2\}, \dots, S^L = \{L-1, L\}$ with $m^1 = n^1$ evaluations of Z_1, $m^2 = n^2$ evaluations of Z_1 and Z_2 sharing n^2 input samples, etc., up to $m^L = n^L$ evaluations of Z_{L-1} and Z_L sharing n^L input samples.

Now we derive *two* lower bounds on the variance of the MLMC estimator in (5.1). The smaller bound can be obtained by combining [10, Lemma 2.3] and [10, Theorem 2.4], however, we will see in our numerical experiments that this lower bound is not sharp in general.

Corollary 5.2. *The variance of $\widehat{\mu}_L^{\mathrm{MLMC}}$ is bounded from below by*

$$\mathrm{Var}(\widehat{\mu}_L^{\mathrm{MLMC}}) \geq \left(\sigma_L^2 + \sigma_{L-1}^2 - 2c_{L,L-1}\right)/n^L \geq \sigma_L^2\left(1 - \rho_{L,L-1}^2\right)/n^L. \tag{5.2}$$

Proof. To obtain the expression after the first inequality sign in (5.2) we simply drop the variance terms associated with some low fidelity models,

$$\mathrm{Var}(\widehat{\mu}_L^{\mathrm{MLMC}}) = \sum_{\ell=1}^{L}\mathrm{Var}(Z_\ell - Z_{\ell-1})/n^\ell \geq \mathrm{Var}(Z_L - Z_{L-1})/n^L = (\sigma_L^2 + \sigma_{L-1}^2 - 2c_{L,L-1})/n^L.$$

To obtain the second bound in (5.2) we consider the estimator

$$\widehat{\mu} := \frac{1}{n^L}\sum_{i=1}^{n^L}\left(Z_L(\omega_i) - Z_{L-1}(\omega_i)\right)$$

for $\mu := \mathbb{E}[Z_L - Z_{L-1}]$. Now, using Theorem 4.3 and (4.6) with $\alpha = (-1,1)^T$, $Q = \{L-1, L\}$ and $Q_\infty = \{L-1\}$ we arrive at

$$\begin{aligned}\mathrm{Var}(Z_L - Z_{L-1})/n^L = \mathrm{Var}(\widehat{\mu}) \geq \gamma(\alpha, Q, Q_\infty) &= (\sigma_L^2 - c_{L,L-1}\sigma_{L-1}^{-2}c_{L,L-1})/n^L \\ &= \sigma_L^2(1 - \rho_{L,L-1}^2)/n^L.\end{aligned}$$ ∎

The MLMC estimator is in general not a BLUE, since it does not depend on the entries of the model covariance matrix C.

5.2. Control Variates. Multiple control variate (CV) estimators (see e.g. [9]) for μ_L which use the coarse models $Z_1, \dots, Z_{L-1}$ have the form

$$\widehat{\mu}_L^{\mathrm{CV}} := \frac{1}{M}\sum_{i=1}^{M} Z_L(\omega_i) + \sum_{\ell=1}^{L-1} \beta_\ell \left(\frac{1}{M}\sum_{i=1}^{M} Z_\ell(\omega_i) - \mu_\ell \right), \tag{5.3}$$

where we assume that the expected values $\mu_1, \dots, \mu_{L-1}$ are known. The estimator $\widehat{\mu}_L^{\mathrm{CV}}$ is clearly unbiased for every choice of the coefficients $(\beta_\ell)_{\ell=1}^{L-1}$. In addition, the coefficients are chosen to minimize the variance of $\widehat{\mu}_L^{\mathrm{CV}}$. Note that this does *not* necessarily imply that the CV estimator is a BLUE since we already prescribe a *specific* linear combination of models by the form in (5.3). However, we have the following result.

Corollary 5.3. *The multiple control variate estimator $\widehat{\mu}_L^{\mathrm{CV}}$ is a BLUE. Moreover,*

$$\mathrm{Var}(\widehat{\mu}_L^{\mathrm{CV}}) = \gamma_{\min}. \tag{5.4}$$

Proof. Introduce $\overline{\alpha} \in \mathbb{R}^L$. Then, the optimization problem for the coefficients reads

$$\min_{\beta \in \mathbb{R}^{L-1}} \mathrm{Var}\left(\widehat{\mu}_L^{\mathrm{CV}}\right) = \min_{\substack{\overline{\alpha} \in \mathbb{R}^L, \\ \overline{\alpha}_L = 1}} \mathrm{Var}\left(\overline{\alpha}_L \frac{1}{M}\sum_{i=1}^{M} Z_L(\omega_i) + \sum_{\ell=1}^{L-1} \overline{\alpha}_\ell \left(\frac{1}{M}\sum_{i=1}^{M} Z_\ell(\omega_i) - \mu_\ell \right)\right).$$

The minimium is equal to $\gamma_{\min} = \gamma(e_L, \{1, \dots, L\}, Q_\infty)$ in (4.3) with $Q_\infty = \{1, \dots, L-1\}$. We now show that the CV estimator is a BLUE. Observe that any linear unbiased estimator $\widehat{\mu}_L$, that is allowed to use the values $\mu_1, \dots, \mu_{L-1}$ similiarly to the CV estimator, satisfies

$$\widehat{\mu}_L = \sum_{i=1}^{M} (\beta^i)^T Z^{k_Q}(\omega_i) + \sum_{\ell=1}^{L-1} b_\ell \mu_\ell, \tag{5.5}$$

where the β^i and b_ℓ satisfy a bias constraint such that for $\ell \in \{1, \dots, L\}$ it holds

$$\sum_{i=1}^{N} \beta^i_{j(\ell)} + b_\ell = \alpha_\ell, \quad \beta^i_{j(\ell)} := e_\ell^T (P^{k_i} \beta^i). \tag{5.6}$$

Here the subscript $j(\ell)$ selects the component of β^i that is multiplied by $Z_\ell(\omega_i)$ in the scalar product $(\beta^i)^T Z^{k_Q}(\omega_i)$, and $\beta^i_{j(\ell)} = 0$ if $\ell \notin Q$. Notice that the variance of the second summand in (5.5) is equal to zero. This allows us to choose the b_ℓ such that the bias constraints (5.6) for $\ell \in \{1, \dots, L-1\}$ are always satisfied. Hence

$$\mathrm{Var}(\widehat{\mu}_L) \geq \min_{\substack{\overline{\alpha}^i \in \mathbb{R}^{|Q|}, \\ \sum_{i=1}^{M} \overline{\alpha}^i_{j(L)} = 1}} \mathrm{Var}\left(\sum_{i=1}^{M} (\overline{\alpha}^i)^T Z^{k_Q}(\omega_i) \right) = \min_{\substack{\overline{\alpha}^i \in \mathbb{R}^L, \\ \sum_{i=1}^{M} \overline{\alpha}^i_L = 1}} \sum_{i=1}^{M} (\overline{\alpha}^i_Q)^T C_{Q,Q} \overline{\alpha}^i_Q.$$

This is the same bound as in (4.3) for $Q_\infty = \{1, \dots, L-1\}$ except that we have potentially individual weights in front of every sample. This does not decrease the variance allowing us to use equal weights $\overline{\alpha}^i = \overline{\alpha}/M$ which is exactly the expression in (4.3). Therefore it holds

$$\operatorname{Var}(\widehat{\mu}_L) \geq \frac{1}{M} \min_{\substack{\overline{\alpha} \in \mathbb{R}^L, \\ \overline{\alpha}_L = 1}} \overline{\alpha}_Q^T C_{Q,Q} \overline{\alpha}_Q = \gamma(e_L, Q, Q_\infty) = \gamma_{\min} = \operatorname{Var}(\widehat{\mu}_L^{\mathrm{CV}}).$$

We conclude that every linear unbiased estimator $\widehat{\mu}_L$ satisfies $\operatorname{Var}(\widehat{\mu}_L) \geq \operatorname{Var}(\widehat{\mu}_L^{\mathrm{CV}})$ showing that the CV estimator is a BLUE. ■

Remark 5.4 (Sample allocation of CV estimator). The CV estimator in (5.3) has a sample allocation with a single model group $S^1 = \{1, \dots, L\}$ and $m^1 = M$ correlated evaluations of $Z_1, \dots, Z_L$, respectively.

5.3. Multifidelity Monte Carlo. Peherstorfer et al. [15, 16] introduce the Multifidelity Monte Carlo (MFMC) estimator based on multiple control variates as follows,

$$(5.7) \qquad \widehat{\mu}_L^{\mathrm{MFMC}} := \frac{1}{n^L} \sum_{i=1}^{n^L} Z_L(\omega_i) + \sum_{\ell=1}^{L-1} \beta_\ell \left(\frac{1}{n^\ell} \sum_{i=1}^{n^\ell} Z_\ell(\omega_i) - \frac{1}{n^{\ell+1}} \sum_{i=1}^{n^{\ell+1}} Z_\ell(\omega_i) \right).$$

The coefficients β_ℓ in (5.7) are chosen such that the variance of $\widehat{\mu}_L^{\mathrm{MFMC}}$ is minimized. Moreover, we assume that the number of samples satisfies $n^1 > \dots > n^L$. The MFMC estimator is linear and unbiased for μ_L.

Remark 5.5 (Sample allocation of MFMC). The MFMC estimator $\widehat{\mu}_L^{\mathrm{MFMC}}$ in (5.7) fits into our framework by the model groups $S^1 = \{1\}$, $S^2 = \{1, 2\}, \dots, S^L = \{1, 2, \dots, L\}$ with $m^1 = n^1 - n^2$ evaluations of Z_1, $m^2 = n^2 - n^3$ evaluations of Z_1 and Z_2 sharing m^2 input samples, etc., up to $m^L = n^L$ evaluations of $Z_1, \dots, Z_L$ sharing n^L input samples.

Theorem 5.6. *The MFMC estimator $\widehat{\mu}_L^{\mathrm{MFMC}}$ is a BLUE for $L = 2$.*

Proof. Observe that the BLUE is a linear combination of the vector y in (2.6). For $L = 2$ this reads

$$\begin{aligned} \widehat{\mu}_L^{\mathrm{B}} &= \alpha_1 \frac{1}{n^2} \sum_{i=1}^{n^2} Z_2(\omega_i) + \alpha_2 \frac{1}{n^2} \sum_{i=1}^{n^2} Z_1(\omega_i) + \alpha_3 \frac{1}{n^2} \sum_{i=n^2+1}^{n^1} Z_1(\omega_i) \\ &= \alpha_1 \frac{1}{n^2} \sum_{i=1}^{n^2} Z_2(\omega_i) + (\alpha_2 - \alpha_3) \frac{1}{n^2} \sum_{i=1}^{n^2} Z_1(\omega_i) + \alpha_3 \frac{n^1}{n^2} \frac{1}{n^1} \sum_{i=1}^{n^1} Z_1(\omega_i) \end{aligned}$$

for suitable coefficients $\alpha_1, \alpha_2, \alpha_3$. Since $\widehat{\mu}_L^{\mathrm{B}}$ is unbiased, we obtain

$$\alpha_1 = 1, \quad \alpha_2 - \alpha_3 = -\alpha_3 n^1 / n^2 =: -\beta_1,$$

which is exactly the expression in (5.7) for $L = 2$. Now since β_1 is chosen to minimize the variance of the estimator, $\widehat{\mu}_L^{\mathrm{MFMC}}$ is the BLUE. ■

For $L > 2$ the MFMC estimator is in general not the BLUE. The reason is that

$$\beta_\ell = \rho_{L,\ell}\sigma_L/\sigma_\ell$$

is the optimal choice [15, Theorem 3.4]. However, this does not depend on $\rho_{\ell,\ell'}$ for $\ell \neq L$ and $\ell' \neq L$. In contrast, the BLUE depends on $\rho_{\ell,\ell'}$; allowing us to also use these correlations for an increased variance reduction. This is also reflected in the next statement.

Corollary 5.7 ([10, Theorem 2.7]). *The variance of $\widehat{\mu}_L^{\text{MFMC}}$ is bounded from below by*

$$\text{Var}(\widehat{\mu}_L^{\text{MFMC}}) \geq \gamma(e_L, \{L-1, L\}, \{1, \dots, L-1\}) = \sigma_L^2(1-\rho_{L,L-1}^2)/n^L. \tag{5.8}$$

An intuitive explanation for this result can be obtained by the definition in (5.7). Let us take a look at the event ω_i with $i \leq n^\ell$, where we have evaluated the expression for every model

$$\frac{1}{n^L} Z_L(\omega_i) + \sum_{\ell=1}^{L-1} \beta_\ell \left(\frac{1}{n^\ell} Z_\ell(\omega_i) - \frac{1}{n^{\ell+1}} Z_\ell(\omega_i) \right).$$

Consider $\ell \neq L$. If n^ℓ is large and since β_ℓ does not depend on n^ℓ, the magnitude of the associated model evaluations is reduced and vanishes in the limit $n^\ell \to +\infty$. Thus, in the infinite data limit the MFMC estimator couples only model evaluations Z_L and Z_{L-1}. Corollary 4.4 with $Q = \{L-1, L\}$ then gives the lower bound $\gamma(e_L, \{L-1, L\}, \{1, \dots, L-1\})$ in (5.8).

Alternatively, we can use the result in Corollary 4.4 with $Q = \{1, \dots, L\}$ as suggested by the sample allocation of MFMC (see Remark 5.5). This gives the lower bound

$$\text{Var}(\widehat{\mu}_L^{\text{MFMC}}) \geq \gamma(e_L, \{1, \dots, L\}, \{1, \dots, L-1\}) = \gamma_{\min}.$$

Note that by the definition of γ in (4.3) it is possible that

$$\gamma(e_L, \{L-1, L\}, \{1, \dots, L-1\}) \gg \gamma(e_L, \{1, \dots, L\}, \{1, \dots, L-1\}).$$

This "gap" is closed by the estimators in the next section.

5.4. Approximate Control Variates. Gorodetsky et al. [10] introduce a general framework for the estimation with Approximate Control Variates (ACVs), and consider several ACV-type estimators in their work. The Approximate Control Variate Independent Samples (ACV-IS) estimator [10, Def. 3.1] is given as follows,

$$\widehat{\mu}_L^{\text{ACV-IS}} := \frac{1}{n^L} \sum_{i=1}^{n^L} Z_L(\omega_i^L) + \sum_{\ell=1}^{L-1} \beta_\ell \left(\frac{1}{n^L} \sum_{i=1}^{n^L} Z_\ell(\omega_i^L) - \frac{1}{n^\ell} \sum_{i=1}^{n^\ell} Z_\ell(\omega_i^\ell) \right) \tag{5.9}$$

with i.i.d. samples ω_i^ℓ. Again, the weights $(\beta_\ell)_{\ell=1}^{L-1}$ are chosen such that the variance of $\widehat{\mu}_L^{\text{ACV-IS}}$ is minimal. A closed-form expression for the optimal weights can be found in [10, Theorem 3.2].

Remark 5.8 (Sample allocation of ACV-IS). The model groups of the ACV-IS estimator are given by $S^1 = \{1\}$, $S^2 = \{2\}$, ..., $S^{L-1} = \{L-1\}$, and $S^L = \{1, \dots, L\}$ with $m^k = n^k$, $k = 1, \dots, L$. Thus, we use independent samples except for S^L, which couples every model $Z_1, \dots, Z_{L-1}$ with the high fidelity model Z_L.

Theorem 5.9. *The ACV-IS estimator* $\widehat{\mu}_L^{\text{ACV-IS}}$ *is a BLUE for every* L.

Proof. The proof is similar to the proof of Theorem 5.6. The BLUE with the ACV-IS sample allocation satisfies

$$\widehat{\mu}_L^{\text{B}} = \sum_{\ell=1}^{L} \alpha_\ell^L \frac{1}{n^L} \sum_{i=1}^{n^L} Z_\ell(\omega_i^L) + \sum_{\ell=1}^{L-1} \alpha_\ell^\ell \frac{1}{n^\ell} \sum_{i=1}^{n^\ell} Z_\ell(\omega_i^\ell)$$

for suitable coefficients $\alpha_1^1, \dots, \alpha_{L-1}^{L-1}, \alpha_1^L, \dots, \alpha_L^L$. The unbiasedness requires us to satisfy

$$\alpha_L^L = 1, \quad \alpha_\ell^L = -\alpha_\ell^\ell =: \beta_\ell, \quad \text{for all } \ell = 1, \dots, L-1.$$

Since the coefficients β_ℓ are chosen to minimize the variance of $\widehat{\mu}_L^{\text{B}}$, $\widehat{\mu}_L^{\text{ACV-IS}} = \widehat{\mu}_L^{\text{B}}$ follows. ∎

Gorodetsky et al. also introduce the ACV-MF estimator [10, Def. 3.3], defined by

$$\widehat{\mu}_L^{\text{ACV-MF}} := \frac{1}{n^L} \sum_{i=1}^{n^L} Z_L(\omega_i) + \sum_{\ell=1}^{L-1} \beta_\ell \left(\frac{1}{n^L} \sum_{i=1}^{n^L} Z_\ell(\omega_i) - \frac{1}{n^\ell} \sum_{i=1}^{n^\ell} Z_\ell(\omega_i) \right), \tag{5.10}$$

where the samples ω_i for different i are independent, and the coefficients $(\beta_\ell)_{\ell=1}^{L-1}$ are chosen to minimize the variance of $\widehat{\mu}_L^{\text{ACV-MF}}$. Following the same idea as in the proof of Theorem 5.6, it is easy to see that the estimator $\widehat{\mu}_L^{\text{ACV-MF}}$ is a BLUE for $L = 2$.

Proposition 5.10. *The ACV-MF estimator* $\widehat{\mu}_L^{\text{ACV-MF}}$ *is a BLUE for* $L = 2$.

Gorodetsky et al. further introduce the ACV-KL estimator in [10, Def. 3.7],

$$\begin{aligned} \widehat{\mu}_L^{\text{ACV-KL}} := \frac{1}{n^L} \sum_{i=1}^{n^L} Z_L(\omega_i) &+ \sum_{\ell=L^{\text{MF}}}^{L-1} \beta_\ell \left(\frac{1}{n^L} \sum_{i=1}^{n^L} Z_\ell(\omega_i) - \frac{1}{n^\ell} \sum_{i=1}^{n^\ell} Z_\ell(\omega_i) \right) \\ &+ \sum_{\ell=1}^{L^{\text{MF}}-1} \beta_\ell \left(\frac{1}{n^{L^{\text{red}}}} \sum_{i=1}^{n^{L^{\text{red}}}} Z_\ell(\omega_i) - \frac{1}{n^\ell} \sum_{i=1}^{n^\ell} Z_\ell(\omega_i) \right), \end{aligned} \tag{5.11}$$

where again ω_i are independent samples. The idea behind the estimator $\widehat{\mu}_L^{\text{ACV-KL}}$ is to use the ACV-MF estimator for the levels $\{L^{MF}, \dots, L\}$ and then reduce the variance of the estimation of $\widehat{\mu}_{L^{\text{red}}}$ using the third summand in (5.11). The sensible choice is thus $L^{\text{red}} \in \{L^{MF}, \dots, L\}$ and $L^{\text{MF}} \in \{1, \dots, L\}$. In particular, for $L^{\text{red}} = L$ the ACV-KL estimator is equal to the ACV-MF estimator.

The parameters $(\beta_\ell)_{\ell=1}^{L-1} \in \mathbb{R}^{L-1}$ and the integer values L^{red} and L^{MF} in (5.11) are chosen such that the variance of $\widehat{\mu}_L^{\text{ACV-KL}}$ is minimal. Observe that by Proposition 5.10 the ACV-MF estimator is a BLUE for $L = 2$. Moreover, for $L = 2$ the ACV-KL and ACV-MF estimator coincide. Hence we have the following result.

Proposition 5.11. *The ACV-KL estimator* $\widehat{\mu}_L^{\text{ACV-KL}}$ *is a BLUE for* $L = 2$.

Remark 5.12 (Sample allocation of ACV-MF and ACV-KL). Let us assume $n^1 > \cdots > n^L$. Then it can be shown that the model groups of ACV-MF and ACV-KL are identical to MFMC (see Remark 5.5). Moreover, the numbers of samples m^k for each model group S^k coincide with those of MFMC as well. The three estimators only differ in the way by which they linearly combine the samples to obtain an unbiased estimator.

Corollary 5.13. *The variance of the ACV-IS, ACV-MF and ACV-KL estimator reaches the bound* $\gamma(e_L, Q, Q_\infty) = \gamma_{\min}$ *for the sample allocation in* (4.1) *in the limit* $N \to +\infty$.

Proof. For ACV-IS and ACV-MF this is proven in [10, Theorem 3.6]. The claim for ACV-KL follows from the discussion in [10, Sec. 3.2]. ■

Finally, we remark that there are multiple choices to define the ACV-KL estimators by modifying the dependency structures of the samples (see [10, Sec. 3.2]). We, however, simply use [10, Def. 3.7].

5.5. Fully coupled BLUE. Remark 5.12 motivates us to define a *fully coupled* (FC) estimator with the same model groups as MFMC, ACV-MF and ACV-KL. We thus have

$$S^\ell_{\mathrm{FC}} = \{1, \dots, \ell\}, \quad \ell = 1, \dots, L, \tag{5.12}$$

with $m^\ell_{\mathrm{FC}} > 0$ if $\ell \leq L$, and $m^k_{\mathrm{FC}} = 0$ otherwise. Note that the numbers m^ℓ_{FC} are uniquely defined once the number of model evaluations $n^1 > \dots > n^L$ has been fixed. We then define the estimator $\widehat{\mu}^{\mathrm{FC}}_L$ as BLUE using the matrix Ψ and vector y in (2.6),

$$\widehat{\mu}^{\mathrm{FC}}_L := \widehat{\mu}^{\mathrm{B}}_L(m_{\mathrm{FC}}) = e^T_L \Psi(m_{\mathrm{FC}})^{-1} y(m_{\mathrm{FC}}). \tag{5.13}$$

According to Theorem 2.6, $\widehat{\mu}^{\mathrm{FC}}_L$ has an equally large or strictly smaller variance compared to the estimators MFMC, ACV-MF and ACV-KL. In particular, neither MFMC, ACV-MF or ACV-KL is in general a BLUE for $L > 2$. We refer to Subsection 6.1 for a numerical illustration of this point. For further illustration purposes we also define the FC, k estimator that simply starts at level $L - k + 1$ and thus only uses the model groups $S^1_{\mathrm{FC}} = \{L - k + 1\}, \dots, S^k_{\mathrm{FC}} = \{L - k + 1, \dots, L\}$. The estimator FC, k only couples the k models with the largest index. In particular, $\widehat{\mu}^{\mathrm{FC}}_L = \widehat{\mu}^{\mathrm{FC,L}}_L$.

5.6. Summary. We summarize the properties of all estimators discussed in this paper in Table 1. For each estimator we minimize the variance given some computational budget p. This requires us to solve an optimization problem of the form (3.5) to compute the optimal number of samples. The entries in the column “Optimization” state whether solving this problem is done analytically or numerically, and the column “DoF” gives the number of degrees of freedom in the optimization problem. The column “Solve with C” indicates whether solving a linear system with the model covariance matrix (or a matrix derived from it) is required. Here SAOB, k and FC, k only require solving a system with a $k \times k$ principal submatrix of C. Finally, the column “Variance bound” gives the (largest) lower bound on the variance for the estimator in the infinite data limit, that is, as the number of samples in $Q_\infty = \{1, \dots, L-1\}$ goes to infinity.

6. Numerical experiments.

In this section we want to numerically verify the main results of this paper. To this end we study two simple academic examples in Subsection 6.1–Subsection 6.2. A practically more relevant example is presented in Subsection 6.3 where we estimate the expectation of a QoI associated with an elliptic PDE with a random diffusion coefficient.

Estimator	BLUE	Solve with C	Optimization	DoF	Variance bound
MC	yes	no	analytic	1	$\gamma(e_L, \{L\}, Q_\infty)$
MLMC (5.1)	$L = 1$	no	analytic	L	$\mathrm{Var}(Z_L - Z_{L-1})/n^L$
MFMC (5.7)	$L \leq 2$	no	analytic	L	$\gamma(e_L, \{L-1, L\}, Q_\infty)$
ACV-IS (5.9)	yes	yes	numeric	L	$\gamma_{\min}$
ACV-MF (5.10)	$L \leq 2$	yes	numeric	L	$\gamma_{\min}$
ACV-KL (5.11)	$L \leq 2$	yes	numeric	$L+2$	$\gamma_{\min}$
FC, k (5.13)	yes	yes, $k \times k$	numeric	k	$\gamma(e_L, \{L-k+1, \dots, L\}, Q_\infty)$
SAOB, k (3.2)	yes	yes, $k \times k$	numeric	$\mathcal{O}(L^k)$	$\gamma(e_L, \{L-k+1, \dots, L\}, Q_\infty)$
SAOB (3.2)	yes	yes	numeric	$2^L - 1$	$\gamma_{\min}$

Table 1: Overview of linear unbiased estimators and their properties.

6.1. Monomial example.

This example is taken from [10, Sec. 2.5]. The model outputs are defined as

$$Z_\ell(\omega) = \omega^\ell, \quad \ell = 1, \dots, L,$$

for $L = 5$, where $\omega \sim U(0,1)$. We fix the total number of evaluations for $Z_1, \dots, Z_L$ as $n^\ell = 2^N 2^{L-\ell}$ for $\ell = 1, \dots, L-1$ and $n^L = 1$. Hence the total cost for each estimator is the same. We vary N to simulate the limit process $\lim_{N \to +\infty} \mathrm{Var}(\widehat{\mu}_L)$. Note that we estimated the required covariance matrix using 10^5 independent pilot samples. Let us now introduce some abbreviations for the variance bounds,

$$\gamma_\ell := \gamma(e_L, \{L-\ell+1, \dots, L\}, \{1, \dots, L-1\}), \quad \ell = 1, \dots, L,$$

where $\gamma_{\min} = \gamma_L$. The estimator variances together with the bounds are shown in Figure 1. We see that MLMC does not reach γ_2, whereas MFMC does reach γ_2, however, no further improvements are made as recorded in Table 1. Hence the bound γ_2 for the variance of MLMC is in general not sharp. This has been already observed in [10, Sec. 2.5]. Our experiments reveal that the variance of MLMC satisfies the sharper bound

$$\mathrm{Var}(\widehat{\mu}_L^{\mathrm{MLMC}}) \geq \left(\sigma_L^2 + \sigma_{L-1}^2 - 2c_{L,L-1}\right)/n^L$$

as proved in Corollary 5.2. We further observe that the novel BLUE estimators FC, k reach the respective bound γ_k and do not improve any further. Finally, the ACV estimators all reach the bound $\gamma_{\min} = \gamma_5$, albeit at a smaller pace than the BLUE FC, 5. The variance of all estimators is bounded by $\gamma_{\min}$ as predicted by Theorem 4.3. Since MFMC, ACV-MF and ACV-KL use the exact same sample allocation as FC, 5, we conclude that in general none of them are BLUEs.

6.2. Noisy monomial example.

The following example is a modification of the example in Subsection 6.1. We define the quantity of interest as before as $Z_L(\omega) := \omega^5$, $L = 6$, together with the models

$$Z_\ell(\omega, \xi) := \omega^{\ell-1} + \xi, \quad \ell = 1, \dots, 5,$$

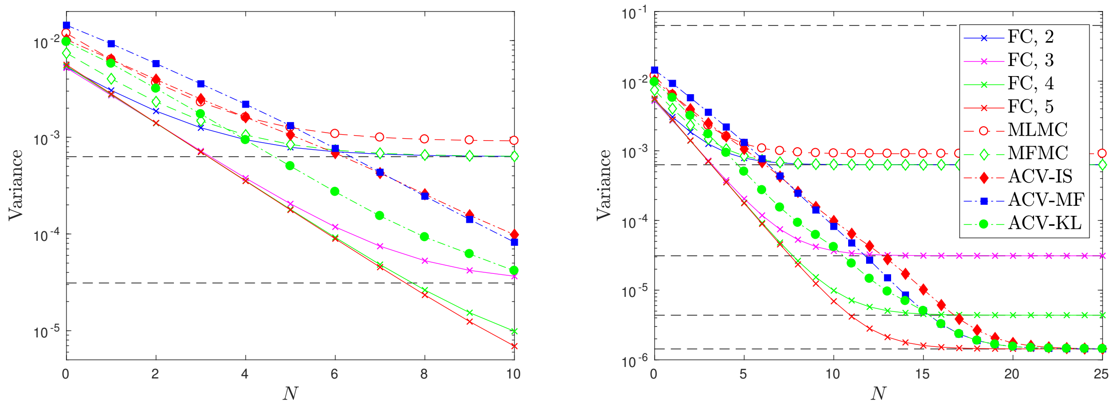

Figure 1: Monomial example: Estimator variances for different numbers of samples N. The minimally achievable variances $\gamma_1 > \cdots > \gamma_5 = \gamma_{\min}$ are drawn horizontally, dashed and black. The variance of the MC and FC, 1 estimator coincides with γ_1, the dashed line at the top of the image on the right-hand side. The image on the left-hand side is a zoom in for $N = 0, \ldots, 10$.

Model	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
Z_1	1.0000	0.9898	0.9891	0.9902	0.9913	0.0012
Z_2	sym	1.0000	0.9993	0.9983	0.9974	0.1182
Z_3	sym	sym	1.0000	0.9997	0.9991	0.1374
Z_4	sym	sym	sym	1.0000	0.9998	0.1374
Z_5	sym	sym	sym	sym	1.0000	0.1319
Z_6	sym	sym	sym	sym	sym	1.0000

Table 2: Sample correlation coefficients for the models in the noisy monomial example estimated with 10^5 samples.

where $\omega \sim U(0,1)$ and $\xi \sim N(0,2)$ are independent random variables. Here, the additional term ξ acts as noise and results in a small correlation of Z_L with $Z_1, \ldots, Z_{L-1}$ as recorded in Table 2. In fact, $Z_1 = \xi$ and Z_L are nearly uncorrelated. (The correlation coefficients in Table 2 have been estimated using 10^5 samples.)

The estimator variances are shown in Figure 2. We observe that the variance of MLMC is the largest among all estimators, in fact, it is nearly two orders of magnitude larger than the variance of Monte Carlo. The variance of MFMC is of the same order of magnitude as the variance of Monte Carlo. The ACV-type estimators have a much smaller variance for larger values of N, and approach the minimal variance possible, $\gamma_{\min}$, as predicted by the theory. The FC, 6 estimator approaches $\gamma_{\min}$ as well. However, in the preasymptotic regime for N small, the variance of the FC, 6 estimator is up to three orders of magnitude smaller compared to the ACV-type estimators.

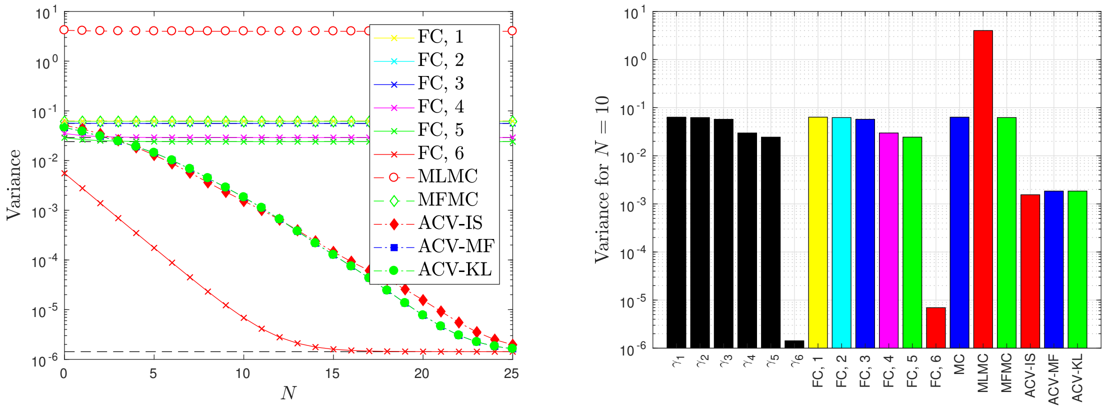

Figure 2: Noisy monomial example: Estimator variances for different numbers of samples N are shown in the left image. The minimally achievable variances $\gamma_1 > \cdots > \gamma_6$ are drawn horizontally, dashed and black. The right image shows the estimator variances for $N = 10$.

In this example, adding the model Z_1 to the components of the BLUE reduces the variance significantly despite the fact that Z_1 and Z_L are actually independent. Adding Z_1 decreases the variance the most, whereas adding $Z_2, \dots, Z_{L-1}$ – which have a larger correlation with the high fidelity model Z_L – only yields a small decrease of the variance. We further conclude that even if every correlation of $Z_1, \dots, Z_{L-1}$ with Z_L is small, we still obtain a significant variance reduction by using these models in the BLUE.

We conclude that BLUEs can yield a significant variance reduction if the quantity of interest is sufficiently well approximated by a linear combination of models where each model may capture a different type of randomness. This may be satisfied even if each model has only a small correlation with the high fidelity model.

6.3. Elliptic PDE with random diffusion coefficient. In this section we apply the SAOB to estimate the expected value of

$$Z(\omega) = \frac{1}{|D_{\text{obs}}|} \int_{D_{\text{obs}}} y(x,\omega)dx,$$

where $D_{\text{obs}} := \left(\frac{3}{4}, \frac{7}{8}\right) \times \left(\frac{7}{8}, 1\right) \subseteq D := (0,1)^2$. Here, y solves an elliptic PDE

$$\begin{aligned} -\operatorname{div}(a(x,\omega)\nabla y(x,\omega)) &= 1, && \text{for } x \in D, \\ y(x,\omega) &= 0, && \text{for } x \in \partial D. \end{aligned} \tag{6.1}$$

The random diffusion coefficient $a(x,\omega) = \exp(\kappa(x,\omega))$, where κ is a mean zero Gaussian random field with Whittle–Matérn covariance function [19] with smoothness parameter $\nu = 3/2$, variance $\sigma^2 = 2$ and correlation length $\rho = 0.1$. We discretize Z by using a uniform mesh refinement with standard linear finite elements (FEs) to obtain the models $Z_1, \dots, Z_L$ with $L = 6$. The data for the discretization is shown in Table 3. Table 4 shows the Pearson

Model	#Nodes	Mesh size	w_ℓ	$\mathrm{Var}(Z_\ell)$	$\mathrm{Bias}(Z_\ell)$
Z_1	81	0.1768	0.0016s	$8.4 \cdot 10^{-4}$	$7.7 \cdot 10^{-3}$
Z_2	289	0.0884	0.0021s	$2.1 \cdot 10^{-4}$	$3.5 \cdot 10^{-3}$
Z_3	1089	0.0442	0.0044s	$4.6 \cdot 10^{-3}$	$8.2 \cdot 10^{-4}$
Z_4	4225	0.0221	0.0148s	$6.0 \cdot 10^{-3}$	$2.0 \cdot 10^{-4}$
Z_5	16641	0.0110	0.0564s	$6.5 \cdot 10^{-3}$	$5.6 \cdot 10^{-5}$
Z_6	66049	0.0055	0.2443s	$6.7 \cdot 10^{-3}$	$1.7 \cdot 10^{-5}$

Table 3: PDE example: The column "#Nodes" lists the number of FE basis function and "Mesh size" gives the maximum diameter of the triangles in the mesh. w_ℓ denotes the expected time (in seconds) to compute a realization of Z_ℓ. The last two columns list the variance and bias.

Model	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
Z_1	1.0000	0.8781	0.7722	0.7229	0.7035	0.6957
Z_2	sym	1.0000	0.9719	0.9460	0.9343	0.9294
Z_3	sym	sym	1.0000	0.9952	0.9907	0.9885
Z_4	sym	sym	sym	1.0000	0.9992	0.9985
Z_5	sym	sym	sym	sym	1.0000	0.9999
Z_6	sym	sym	sym	sym	sym	1.0000

Table 4: PDE example: The Pearson correlation coefficient matrix $(\rho_{\ell j})_{\ell,j=1}^L$ estimated with 10^4 samples.

correlation coefficients between the models. The covariance matrix, correlation coefficients and the work per level w_ℓ were estimated with 10^4 samples. Note that we do not include this cost in our complexity analysis. We estimated the bias using Monte Carlo with the same samples as follows,

$$\mathrm{Bias}(Z_\ell) := |\mathbb{E}[Z_\ell] - \mathbb{E}[Z]| \approx |\widehat{\mu}_\ell^{\mathrm{MC}} - \widehat{\mu}_L^{\mathrm{MC}}|, \quad \text{for } \ell = 1, 2, 3, 5.$$

We set $\mathrm{Bias}(Z_4) := \mathrm{Bias}(Z_3)/4$, since Z_4 seemed to have a smaller bias than Z_5. We extrapolated the resulting values to obtain $\mathrm{Bias}(Z_L)$. We remark that this bias estimation is crude and can be improved, however, it is sufficient for our purposes, since we only compare estimators with the same bias. We want to obtain a Mean Square Error (MSE) of the estimator $\widehat{\mu}_L$ of at most ε^2, that is,

$$\mathbb{E}[(\widehat{\mu}_\ell - \mathbb{E}[Z])^2] = \mathrm{Bias}(Z_\ell)^2 + \mathrm{Var}(\widehat{\mu}_\ell) \leq \varepsilon^2,$$

such that the cost of $\widehat{\mu}_\ell$ is minimized. We choose the level ℓ such that $\mathrm{Bias}(Z_\ell)^2 \leq \varepsilon^2/2$ and afterwards ensure that

$$\mathrm{Var}(\widehat{\mu}_\ell) \leq \varepsilon^2/2. \tag{6.2}$$

We achieve this variance constraint with minimal cost by solving a sample allocation problem for every estimator, where we allow fractional samples. Formally, for a parameter vector x and an estimator $\widehat{\mu}_\ell(x)$ we solve a problem of the form

$$\min_x \operatorname{Var}(\widehat{\mu}_\ell(x)) \quad \text{such that } \operatorname{Cost}(\widehat{\mu}_\ell(x)) \leq p. \tag{6.3}$$

Then, we rescale the number of samples to achieve (6.2). For SAOB, k we solve the problem (3.5) with $\delta = 0$. For MC x is the number of evaluations of Z_ℓ. For MLMC we optimize the variance over $n^1, \dots, n^\ell$, and for ACV-KL we optimize over the parameters $n^1, \dots, n^\ell$, L^{MF} and L^{red}. For ACV-KL we follow a brute force approach and optimize over all feasible integer values of L^{MF} and L^{red}. We carry out the corresponding optimization also for ACV-IS and ACV-MF. We further remark that we optimize over the first used level, that is, if one of the estimators has a smaller variance starting at level 2 instead of level 1, then the first model is never evaluated. For MC, MLMC and MFMC there are analytic expressions for the number of samples available. For the remaining estimators we employ Matlab's `fmincon` function which uses an interior point algorithm, where we supply the gradient of the variance of the estimator. For SAOB, k we additionally supply the Hessian.

The computed cost allowing fractional samples is shown in Figure 3. We see that the SAOB achieves the target root mean square error (RMSE) with the smallest cost. For the smallest RMSE, where we have estimators with $\operatorname{Bias}(Z_L)$, MLMC is $\approx 52\%$ more expensive than SAOB. For MFMC this value is $\approx 71\%$, for SAOB, 2 it is $\approx 35\%$ and for SAOB, 3 it is $\approx 3\%$. The estimator SAOB, 4, which is not plotted, is only $\approx 0.7\%$ more expensive than SAOB. We clearly see that increasing the coupling number reduces the variance but the overall cost savings decrease.

We remark that we were not able to solve (6.3) for ACV-MF and ACV-KL for the two leftmost datapoints. After 10^4 iterations in `fmincon`, the method had not converged yet and we used the final (suboptimal) sample allocation. Hence, these two values in Figure 3 for ACV-MF and ACV-KL are only upper bounds for the variance. We think that the reason for the non-convergence of `fmincon` is the ill-conditioning of the model covariance matrix C.

We now focus on the data point with the smallest RMSE in Figure 3. The total number of evaluations of $Z_1, \dots, Z_L$ is shown in Figure 4. We see that the MC estimator uses only the high fidelity model, MLMC uses all models, and MFMC starts with model Z_3. Similarly, SAOB, 2 and SAOB, 3 do not use the coarsest model. A possible explanation for this is the fact that the mesh size associated with Z_1 is larger compared to the correlation length of the diffusion coefficient a. The SAOB uses all models. We conclude that estimators that use fewer evaluations of the expensive high fidelity model have smaller costs. Here SAOB has ≈ 270 high fidelity evaluations whereas SAOB, 3 has ≈ 1150, that is more than four times as many, however SAOB, 3 is only $\approx 3\%$ more expensive.

We now also comment on the model groups and coefficients β^k for the SAOB, k estimators which are shown in Figure 5. The terms β^k denote the coefficients in the linear combination of the models in the final estimator. For example, for SAOB with independent events ω_i^k it

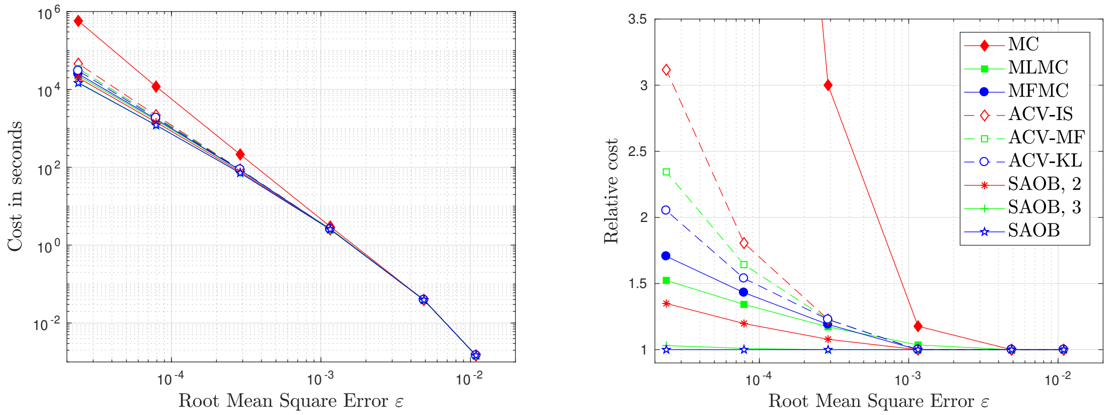

Figure 3: PDE example: Computed cost to achieve a certain RMSE for different estimators with fractional samples. The left image shows the absolute cost in seconds and the right image the relative cost w.r.t. the SAOB estimator.

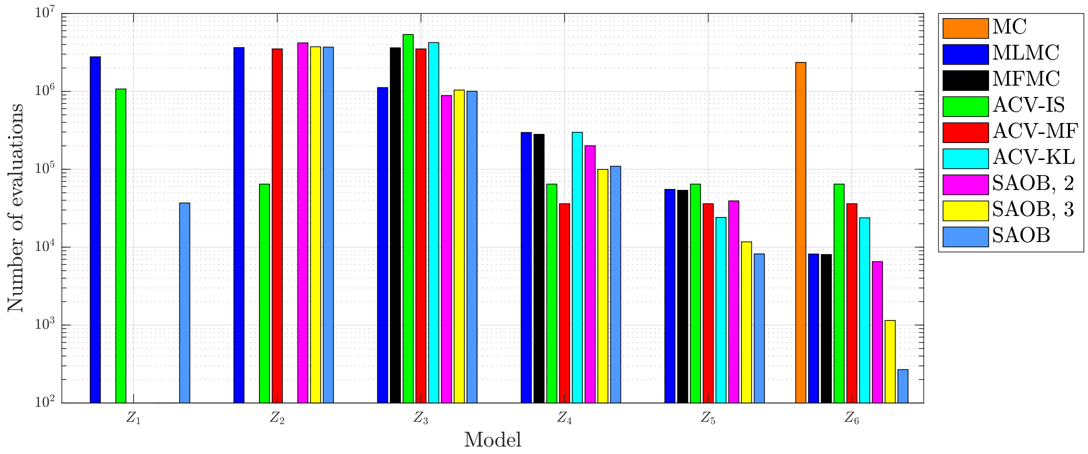

Figure 4: PDE example: Computed total number of evaluations for every model in logarithmic scale.

holds

$$\widehat{\mu}_L^{\text{SAOB}} = \sum_{k=1}^{6} \sum_{\ell \in S^k} \beta_\ell^k \frac{1}{m^k} \sum_{i=1}^{m^k} Z_\ell(\omega_i^k).$$

We can see in Figure 5 that for each estimator the row sum along Z_ℓ, $\ell = 1, \dots, L-1$, is equal to zero, and the row sum along Z_L is equal to one. This is consistent with the unbiasedness requirement. The optimal sample allocation obtained by solving (3.5) is actually unique for

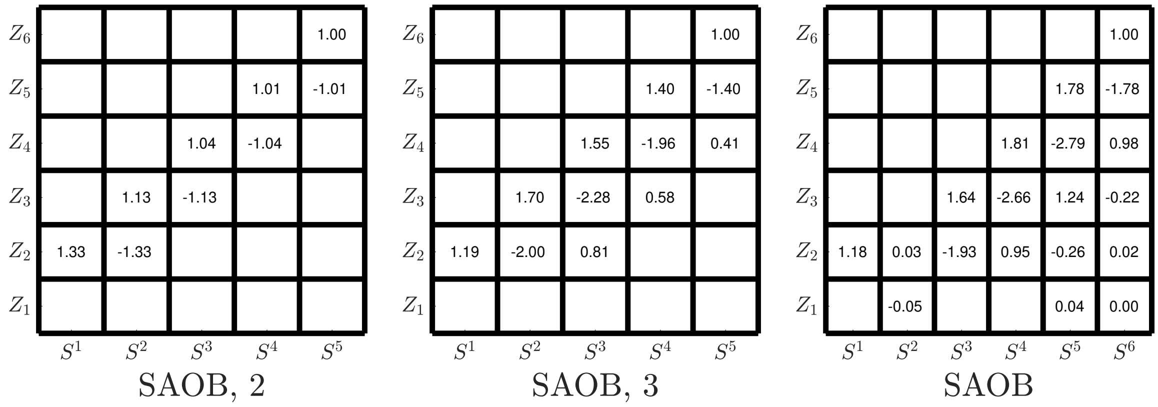

Figure 5: PDE example: Model groups for the SAOB, k estimators in the optimal sample allocation. Each column represents a model group. A non-empty square in the box means that the model is an element of the model group. For example, for SAOB we have $S^4 = \{2, 3, 4\}$ with $\beta_2^4 = 0.95$, $\beta_3^4 = -2.66$ and $\beta_4^4 = 1.81$.

all SAOB, k and results in six or less model groups S^k. In fact, SAOB, 2 uses the same model groups as MLMC but starts with model Z_2.

Let us comment on fractional samples, which we used up until now. For the SAOB, k estimators we round the number of samples to the next biggest integer. This way, the variance target (6.2) is still satisfied, however, with an increased cost. This additional cost was at most ≈ 0.6 seconds accounting for a relative increase of at most $\approx 0.004\%$. For this example, the small increase in cost is negligible and thus working with fractional instead of integer samples for the optimization problem (6.3) is justified. Ceiling the number of samples for the other estimators also increases their cost by a negligible margin.

In other examples, rounding up the number of samples may significantly increase the cost of the estimator. Then one has to explicitly deal with the integer constraints. To this end one could apply Integer Programming techniques like branch-and-bound, where we branch on the number of model evaluations. However, the efficiency of such methods is highly problem dependent, and further investigations are out of the scope of this paper.

We verify our implementation by re-computing all estimators for μ_4 where we apply the ceiling of the number of samples. We average those estimates over 100 independent runs. The sample mean for each estimator is ≈ 0.0157. We further compute the sample variance and compare it with the target $\tau := \mathrm{Bias}(Z_4)^2/2$. The variance for MC was $\approx 0.70\tau$, MLMC $\approx 0.99\tau$, MFMC $\approx 1.12\tau$, ACV-IS $\approx 0.89\tau$, ACV-MF $\approx 0.98\tau$, ACV-KL $\approx 0.80\tau$, SAOB, 2 $\approx 0.99\tau$, SAOB, 3 $\approx 0.99\tau$ and finally SAOB $\approx 0.92\tau$. Here the ACV-MF and ACV-KL estimators are actually identical, since for the target RMSE we obtain $L^{\mathrm{MF}} = L^{\mathrm{red}} = 4$. We thus conclude that our implementation yields consistent results.

Finally, we remark that for the smallest RMSE for SAOB we spent ≈ 6 seconds to compute the optimal sample allocation, which is a relative cost increase of $\approx 0.04\%$. In practice, we cannot neglect the cost to estimate the covariance matrix C and cost per level w_ℓ, which we did not include in our analysis and which has to be done in an offline step. We however expect that adaptive methods can be used to address this point.

7. Conclusions. We introduce and analyze a multilevel best linear unbiased estimator to approximate the expected value of a scalar-valued output quantity of interest. We show that this estimator is variance minimal independently of the number of model evaluations. We prove a sharp lower bound on the variance of any linear unbiased multilevel estimator and show that our proposed estimator approaches the exact same lower bound as the ACV-type estimators in the infinite low fidelity data limit. Moreover, we suggest an optimal sample allocation scheme that constructs the model groups such that a target estimator variance is achieved with a given computational budget. We demonstrate in numerical experiments that the multilevel BLUE can achieve a significant variance reduction for models that are nearly uncorrelated with the high fidelity model and that are polluted by noise. The BLUE handles such situations by linearly combining the model evaluations in an optimal way. This is in contrast to the multilevel estimators in the literature, such as MLMC and MFMC, where the linear combinations are (partially) fixed and cannot fully be adapted to the problem at hand.

We point out that our analysis is completely independent of the type of models in the multifidelity hierarchy. It can be expected that by making specific assumptions on the models, in particular, PDE-based outputs, we are able to show convergence rates of the multilevel BLUE with respect to e.g. mesh size parameters and analyze its complexity. This is the subject of ongoing work. Moreover, it would be desirable to eliminate the need to know the model covariance matrix, or at the very least, analyze the errors introduced to the BLUE by using an inexact, sample-based covariance matrix. We envision that adaptive approaches, possibly by using data assimilation techniques, such as the Kalman filter and ensemble Kalman filter, could be useful for this task. The ill-conditioning of the model covariance matrix for highly correlated models is also a problem that requires further attention.

Acknowledgements. The authors thank Michael Ulbrich for the finite element code that was used to solve the PDE and to sample from the mean zero Gaussian random field with Matern 3/2 covariance in Subsection 6.3. We further thank Jonas Latz for his helpful suggestions which improved the readability of this paper.

REFERENCES

[1] A. Ahmad Ali, E. Ullmann, and M. Hinze, *Multilevel Monte Carlo Analysis for Optimal Control of Elliptic PDEs with Random Coefficients*, SIAM/ASA Journal on Uncertainty Quantification, 5 (2017), pp. 466–492, https://doi.org/10.1137/16M109870X.

[2] A. Beskos, A. Jasra, K. Law, Y. Marzouk, and Y. Zhou, *Multilevel sequential Monte Carlo with dimension-independent likelihood-informed proposals*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 762–786, https://doi.org/10.1137/17M1120993.

[3] A. Beskos, A. Jasra, K. Law, R. Tempone, and Y. Zhou, *Multilevel sequential Monte Carlo samplers*, Stochastic Process. Appl., 127 (2017), pp. 1417–1440, https://doi.org/10.1016/j.spa.2016.08.004, https://doi.org/10.1016/j.spa.2016.08.004.

[4] C. Bierig and A. Chernov, *Convergence analysis of multilevel Monte Carlo variance estimators and application for random obstacle problems*, Numer. Math., 130 (2015), pp. 579–613, https://doi.org/10.1007/s00211-014-0676-3.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004, https://doi.org/10.1017/CBO9780511804441.
[6] K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup, *Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients*, Comput. Vis. Sci., 14 (2011), pp. 3–15, https://doi.org/10.1007/s00791-011-0160-x.
[7] M. B. Giles, *Multi-level Monte Carlo path simulation*, Operations Research, 56 (2008), pp. 607–617, https://doi.org/10.1287/opre.1070.0496.
[8] M. B. Giles, *Multilevel Monte Carlo methods*, Acta Numerica, 24 (2015), pp. 259–328, https://doi.org/10.1017/S096249291500001X.
[9] P. W. Glynn and R. Szechtman, *Some new perspectives on the method of control variates*, in Monte Carlo and quasi-Monte Carlo methods, 2000 (Hong Kong), Springer, Berlin, 2002, pp. 27–49, https://doi.org/10.1007/978-3-642-56046-0_3.
[10] A. A. Gorodetsky, G. Geraci, M. Eldred, and J. D. Jakeman, *A Generalized Framework for Approximate Control Variates*, Preprint, arXiv:1811.04988 (2018).
[11] H. k. Hoel, K. J. H. Law, and R. Tempone, *Multilevel ensemble Kalman filtering*, SIAM J. Numer. Anal., 54 (2016), pp. 1813–1839, https://doi.org/10.1137/15M100955X, https://doi.org/10.1137/15M100955X.
[12] J. Humpherys, P. Redd, and J. West, *A fresh look at the Kalman filter*, SIAM Rev., 54 (2012), pp. 801–823, https://doi.org/10.1137/100799666.
[13] R. Kornhuber and E. Youett, *Adaptive multilevel Monte Carlo methods for stochastic variational inequalities*, SIAM J. Numer. Anal., 56 (2018), pp. 1987–2007, https://doi.org/10.1137/16M1104986.
[14] S. Krumscheid and F. Nobile, *Multilevel Monte Carlo approximation of functions*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 1256–1293, https://doi.org/10.1137/17M1135566.
[15] B. Peherstorfer, K. Willcox, and M. Gunzburger, *Optimal Model Management for Multifidelity Monte Carlo Estimation*, SIAM Journal on Scientific Computing, 38 (2016), pp. A3163–A3194, https://doi.org/10.1137/15M1046472.
[16] B. Peherstorfer, K. Willcox, and M. Gunzburger, *Survey of multifidelity methods in uncertainty propagation, inference, and optimization*, SIAM Rev., 60 (2018), pp. 550–591, https://doi.org/10.1137/16M1082469.
[17] A. Quaglino, S. Pezzuto, and R. Krause, *High-dimensional and higher-order multifidelity Monte Carlo estimators*, Journal of Computational Physics, 388 (2019), pp. 300–315, https://doi.org/https://doi.org/10.1016/j.jcp.2019.03.026.
[18] C. R. Rao, H. Toutenburg, Shalabh, and C. Heumann, *Linear models and generalizations*, Springer Series in Statistics, Springer, Berlin, extended ed., 2008, https://doi.org/10.1007/978-3-540-74227-2. Least squares and alternatives, With contributions by Michael Schomaker.
[19] M. L. Stein, *Interpolation of spatial data*, Springer Series in Statistics, Springer-Verlag, New York, 1999, https://doi.org/10.1007/978-1-4612-1494-6. Some theory for Kriging.
[20] A. Van Barel and S. Vandewalle, *Robust optimization of PDEs with random coefficients using a multilevel Monte Carlo method*, SIAM/ASA J. Uncertain. Quantif., 7 (2019), pp. 174–202, https://doi.org/10.1137/17M1155892.