An approximate control variates approach to multifidelity distribution estimation*

Ruijian Han†, Boris Kramer‡, Dongjin Lee§, Akil Narayan¶, and Yiming Xu‖

Abstract. Forward simulation-based uncertainty quantification that studies the distribution of quantities of interest (QoI) is a crucial component for computationally robust engineering design and prediction. There is a large body of literature devoted to accurately assessing statistics of QoIs, and in particular, multilevel or multifidelity approaches are known to be effective, leveraging cost-accuracy tradeoffs between a given ensemble of models. However, effective algorithms that can estimate the full distribution of QoIs are still under active development. In this paper, we introduce a general multifidelity framework for estimating the cumulative distribution function (CDF) of a vector-valued QoI associated with a high-fidelity model under a budget constraint. Given a family of appropriate control variates obtained from lower-fidelity surrogates, our framework involves identifying the most cost-effective model subset and then using it to build an approximate control variates estimator for the target CDF. We instantiate the framework by constructing a family of control variates using intermediate linear approximators and rigorously analyze the corresponding algorithm. Our analysis reveals that the resulting CDF estimator is uniformly consistent and asymptotically optimal as the budget tends to infinity, with only mild moment and regularity assumptions on the joint distribution of QoIs. The approach provides a robust multifidelity CDF estimator that is adaptive to the available budget, does not require *a priori* knowledge of cross-model statistics or model hierarchy, and applies to multiple dimensions. We demonstrate the efficiency and robustness of the approach using test examples of parametric PDEs and stochastic differential equations including both academic instances and more challenging engineering problems.

Key words. control variates, distribution estimation, model selection, multifidelity, robustness

AMS subject classifications. 62J05, 62G30, 62F12, 62-08

1. Introduction.

Physical systems are often modeled with computational simulations or emulators, and as such, understanding the error in these constructed approximations is of utmost importance. One particular source of uncertainty in the output is due to the input uncertainty in these models, either through uncertainty in model parameters (which can be finite- or infinite-dimensional) or through modeled stochasticity in the system, e.g., systems driven with white noise processes. To make the resulting models trustworthy, it is crucial to quantify the resulting uncertainty in QoIs; that is, to estimate the QoI's distribution or some statistical summary of it. One popular approach for achieving this is through Monte Carlo (MC) simulation, which is easy to implement and provides robust results but has a slow con-

*Submitted to the editors on May 16, 2024.

Funding: RH is partially supported by the Hong Kong Research Grants Council (grant no. 14301821) and Start-up Fund for New Recruits, The Hong Kong Polytechnic University. AN is partially supported by NSF DMS-1848508 and AFOSR FA9550-20-1-0338. BK and DL are supported by the Ministry of Trade, Industry and Energy (MOTIE) and the Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program (grant no. P0019804, Digital twin based intelligent unmanned facility inspection solutions).

†Department of Applied Mathematics, The Hong Kong Polytechnic University

‡Department of Mechanical and Aerospace Engineering, University of California San Diego

§Department of Automotive Engineering, Hanyang University

¶Scientific Computing and Imaging Institute, and Department of Mathematics, University of Utah

‖Department of Statistics and Actuarial Science, University of Waterloo (yiming2358@gmail.com).

vergence rate. A typical MC procedure requires drawing a large number of samples or running repeated experiments, which is expensive given the increasing complexity of computational simulations.

To address this issue, methods based on multilevel [11, 12] and multifidelity modeling [29, 28, 30, 19, 27, 16, 18, 32, 9, 35, 33, 17, 8, 10] have been developed to estimate the statistics of QoIs associated with the (high-fidelity) model. The core idea behind multilevel/multifidelity methods lies in leveraging models of different accuracies and costs to improve computational efficiency. However, a major limitation of the existing literature is that it predominantly focuses on the estimation of the statistical mean of the QoIs (or other scalar-valued descriptive statistics such as quantiles or conditional expectations), providing only partial insight into the uncertainty of the QoIs. A more comprehensive understanding would require assessing, for example, higher-order statistics of the QoI, or even the entire distribution.

Existing methods to estimate CDFs in the multilevel and multifidelity setup have seen notable success [12, 23, 13, 21, 3, 36]. In [12], the authors proposed a multilevel approach to computing the CDFs of univariate random variables arising from stochastic differential equations and derived an upper bound for the cost in terms of the error. The methodology in [12] was further developed and applied in several subsequent works [23, 21, 13, 3]. In particular, [23] designed an *a posteriori* optimization strategy to calibrate the smoothing function and showed its superiority over MC in oil reservoir simulations; [21] generalized the ideas in [12] to approximate more general parametric expectations such as characteristic functions; [13] applied an adaptive approach for parameter selection that yields an improved cost bound; [3] provides a novel computable error estimator to enhance algorithm tuning. Despite the substantive contributions of these approaches, nearly all of them make relatively restrictive assumptions regarding model hierarchy (e.g., the model cost versus accuracy tradeoffs), and do not immediately extend to the general non-hierarchical multifidelity setup. For this more general multifidelity estimation of CDFs, the only work we are aware of is the adaptive explore-then-commit algorithm for distribution learning (AETC-d) [36]. However, the large-budget performance of AETC-d is restricted by its own set of statistical assumptions that are often too stringent to satisfy in practice. Moreover, the QoI in all the above references is assumed to be a scalar.

An outline of the paper is as follows. The remainder of this section lists our contributions, introduces overall notation, and summarizes the main theory and algorithmic advances. Sections 2 through 4 describe the necessary mathematical and statistical background for our method: Section 2 gives a brief overview of the control variates method; Section 3 introduces a multifidelity CDF estimation framework based on approximate control variates estimators. Section 4 provides a computational construction for the control variates through linear approximators. Section 5 develops our new meta algorithm (cvMDL) that accomplishes autonomous model selection together with an algorithmic correction to preserve the monotonicity of the resulting CDF estimators. The meta algorithm cvMDL itself does not specify how to compute the control variates: A specialization to using the linear approximations from Section 4 yields a computationally explicit algorithm that we study in detail, establishing both uniform consistency and budget-asymptotic optimality. Section 6 contains a detailed simulation study and showcases applications that use estimated CDFs to compute probabilistic risk metrics.

1.1. Contributions. The main goal of this article is to provide novel solutions that mitigate the deficiencies described above. We develop an efficient algorithm for estimating the CDF in a general non-hierarchical multifidelity approximation setting under computational budget constraints. The proposed method satisfies the following criteria: 1) it requires as input neither cross-model statistics nor model hierarchy; 2) it can provide distributional estimates for vector-valued QoIs, and 3) it is empirically robust and enjoys theoretical guarantees. Although our approach uses a similar meta algorithm as in [35, 36] (all borrowing ideas from the explore-then-commit algorithm in bandit learning [22]), it contains a substantial number of new ingredients that extend applicability and improve robustness. In more technical language, our contributions are twofold:

- We propose a control variates-based exploration-exploitation strategy for multifidelity CDF estimation under a budget constraint. The *exploration* step leverages statistical estimation to select a subset of low-fidelity models for the control variates construction, followed by the *exploitation* step that utilizes the learned information to build an approximate control variates estimator for the target CDF. This procedure is initialized with no *a priori* oracle information[1] about model relationships, in contrast to several methods that require such information as input. In addition, our estimator for the CDF applies to both scalar-valued and vector-valued QoI, which differentiates it from existing methods that apply only to scalar-valued QoI.
- Through examination of the average weighted-L^2 loss that balances errors in exploration and exploitation, we design a new meta algorithm, the control variates multifidelity distribution learning algorithm ("cvMDL", summarized in Figure 1 and detailed in Algorithm 5.2), that accomplishes model (subset) selection and CDF estimation. Using control variates constructed from linear approximators, we establish both uniform consistency and asymptotic optimality of the estimator produced by cvMDL as the budget approaches infinity (Theorem 5.7). Our analysis illustrates that the proposed procedure significantly ameliorates the restrictive model assumptions in [36].

A verbatim usage of our approaches produces a CDF estimator that enjoys the previously-mentioned theoretical guarantees but is not necessarily monotonic and hence may not be itself a distribution function. To mitigate this artifact, we utilize an empirical algorithmic correction that restores the monotonicity of the estimated CDFs and additionally makes its manipulation more computationally convenient (e.g. for extraction of quantiles and conditional expectations); see Algorithm 5.1. We observe that in some cases this empirical correction further reduces errors.

1.2. Notation. For $n \in \mathbb{N}$, let $\{1:n\} := \{1, \dots, n\}$. We use bold upper-case and lower-case letters to denote matrices and vectors, respectively. The Euclidean (ℓ^2) norm on a vector $\boldsymbol{v}$ is denoted $\|\boldsymbol{v}\|_2$. For a matrix $\boldsymbol{A}$, $\boldsymbol{A}^\top$ is the transpose and $\boldsymbol{A}^\dagger$ is the pseudoinverse; $\boldsymbol{A}^\dagger$ coincides with the regular inverse $\boldsymbol{A}^{-1}$ when $\boldsymbol{A}$ is invertible. The ith column of $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^{(i)}$. The Frobenius norm of $\boldsymbol{A}$ is denoted by $\|\boldsymbol{A}\|_F = (\sum_i \|\boldsymbol{A}^{(i)}\|_2^2)^{1/2}$. We use $\otimes$ to denote the tensor product operator. For a set $\mathcal{T} \subseteq \mathbb{R}^d$, we denote its interior as $\mathcal{T}^\circ$, and

[1] In this article, oracle information refers to model statistics that we treat as exact. These statistics may be exactly computed, but more often are approximations identified through simulations with a large enough computational expense so that the approximations are treated as ground truth.

$\mathbf{1}_{\mathcal{T}}(\boldsymbol{x}) := \mathbf{1}_{\{\boldsymbol{x}\in\mathcal{T}\}}$ as the indicator function on $\mathcal{T}$. For two vectors $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(d)})^\top$ and $\boldsymbol{y} = (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(d)})^\top$, we use $\vee$ and $\wedge$ to denote the componentwise *max* and *min* operators, respectively, i.e.,

$$\boldsymbol{x} \vee \boldsymbol{y} := \left(\max\{\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}\}, \ldots, \max\{\boldsymbol{x}^{(d)}, \boldsymbol{y}^{(d)}\}\right)^\top$$
$$\boldsymbol{x} \wedge \boldsymbol{y} := \left(\min\{\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}\}, \ldots, \min\{\boldsymbol{x}^{(d)}, \boldsymbol{y}^{(d)}\}\right)^\top.$$

Moreover, we say $\boldsymbol{x} \le \boldsymbol{y}$ if $\boldsymbol{x}^{(i)} \le \boldsymbol{y}^{(i)}$ for all $i \in \{1 : d\}$. We consider the QoIs from computational models as random variables that jointly lie in some common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For a random vector $X \in \mathbb{R}^d$, we let $F_X(\boldsymbol{x}) = \mathbb{P}(X^{(1)} \le \boldsymbol{x}^{(1)}, \ldots, X^{(d)} \le \boldsymbol{x}^{(d)})$ denote its CDF. For two sequences of random variables $\{a_m(\omega)\}$ and $\{b_m(\omega)\}$ where $\omega \in \Omega$ is a probabilistic event, we write $a_m(\omega) \lesssim b_m(\omega)$ if almost surely (a.s.), $a_m(\omega) \le \eta(\omega) b_m(\omega)$ for all $m \in \mathbb{N}$, where the constant $\eta(\omega)$ is independent of m. For convenience, we let

$$Y = (Y^{(1)}, \ldots, Y^{(d)})^\top \in \mathbb{R}^d \qquad X_i = (X_i^{(1)}, \ldots, X_i^{(d_i)})^\top \in \mathbb{R}^{d_i} \qquad i \in \{1 : n\}$$

denote the high-fidelity and the ith low-fidelity QoIs, respectively. Here $d, d_i \in \mathbb{N}$ are the corresponding dimensions of Y and X_i. There are n low-fidelity models in total. We use $\mathbb{E}[\cdot], \mathrm{Var}[\cdot]/\mathrm{Cov}[\cdot]$, and $\mathrm{Corr}[\cdot]$ to denote the expectation, variance/covariance, and correlation operators respectively. We use $\perp\!\!\!\perp$ to represent probabilistic independence.

1.3. Model assumptions.

We assume the sampling costs for Y and $X_1, \ldots, X_n$, denoted by positive numbers c_0 and $c_1, \ldots, c_n$, are deterministic and known. For $\mathcal{S} \subseteq \{1 : n\}$, let $c_{\mathcal{S}} = \sum_{i\in\mathcal{S}} c_i$, corresponding to the cost of sampling all (low-fidelity) models from subset $\mathcal{S}$. We let $B > 0$ be the total budget (deterministic and known) that is available to expend on sampling the models. Moreover, for $\mathcal{S} \subseteq \{1 : n\}$, we let $X_{\mathcal{S}} = (X_i^\top)_{i\in\mathcal{S}}^\top \in \mathbb{R}^{d_{\mathcal{S}}}$, where $d_{\mathcal{S}} = \sum_{i\in\mathcal{S}} d_i$, and $X_{\mathcal{S}^+} = (1, X_{\mathcal{S}}^\top)^\top$, where the latter is used when considering a linear model approximation with the intercept/bias term.

The central goal in the rest of the article is to develop a multifidelity estimator for $F_Y(\boldsymbol{x})$ through drawing samples of $(Y, X_{\{1:n\}})$ and of $X_{\mathcal{S}}$ for some adaptively-determined $\mathcal{S} \subseteq \{1 : n\}$, subject to the sampling cost not exceeding the total budget constraint $B > 0$. No other high-level assumptions are made. In other words, we assume only that Y is a known high-fidelity model; we do not assume any ordering/hierarchy in the models $X_{\{1:n\}}$, and we do not assume known statistics (e.g., correlations) between any models. While such generality is sufficient for algorithmic purposes, our theoretical guarantees require additional technical assumptions that are articulated in Subsection 5.4. These technical assumptions are mild regularity conditions, related to finite moments of random variables and CDF functional regularity.

The notation we have introduced is enough to present the overall cvMDL algorithm in the next section. The actual computations that make the algorithm practical, however, require more technical details which are provided in Section 2 through Section 5.

1.4. Summary of the algorithm.

The proposed cvMDL meta algorithm is shown in Figure 1. In summary, we first gather m full joint samples of $(Y, X_{\{1:n\}})$ through an *exploration phase* that identifies (i) how models are related, (ii) which model subset $\mathcal{S}$ optimally balances

cost versus accuracy, and (iii) whether more samples m are needed to certify a robust exploration or whether the choice of $\mathcal{S}$ is statistically robust enough to proceed with exploitation. Exploration is followed by the *exploitation phase*, where we exhaust the remaining computational budget to sample the optimal model subset $X_{\mathcal{S}}$. Exploitation corresponds to exercise of a particular approximate control variates estimator for F_Y. A more detailed description is as follows:

Exploration phase

- *Mininum exploration*: This step ensures that the number of exploration samples m is set large enough so that non-degenerate empirical statistics can be computed.
- *Analyze low-fidelity models*: We are interested in estimating the minimal loss associated with an estimated CDF that utilizes the model subset $\mathcal{S}$. For such a goal, this step identifies for each model subset $\mathcal{S}$ both an estimated number of optimal exploration samples $\widehat{m}^*_{\mathcal{S}}$ along with the corresponding loss function minimum $\widehat{L}(m \vee \widehat{m}^*_{\mathcal{S}}; m)$. The value $\widehat{L}(z; m)$ is an estimator with the currently-available m exploration samples and measures the estimated loss if we eventually use z exploration samples. When evaluating the minimum loss, we require the input $z \leftarrow m \vee \widehat{m}^*_{\mathcal{S}}$ since if $m > \widehat{m}^*_{\mathcal{S}}$ then the number of exploration samples should be m, and not $\widehat{m}^*_{\mathcal{S}}$ (we cannot take fewer exploration samples than already committed and we assume $\widehat{L}(z; m)$ is convex and has a unique minimizer). The definitions of $\widehat{L}$, $\widehat{m}^*_{\mathcal{S}}$, and $\widehat{\mathcal{S}}^*$ are given in (5.6) and (5.7).
- *Select optimal model*: The estimated optimal model subset $\widehat{\mathcal{S}}^*$ is computed by choosing the subset $\mathcal{S}$ with minimal loss from the previous step.
- *Continue exploration*: If the current number of exploration samples m is smaller than the estimated optimal number of samples $\widehat{m}^*_{\widehat{\mathcal{S}}^*}$ required for the optimal subset $\widehat{\mathcal{S}}^*$, then we continue exploration, with the precise number of additional exploration samples determined by the function $Q(\cdot, \cdot)$ that is defined in (5.12). If $m \geq \widehat{m}^*_{\widehat{\mathcal{S}}^*}$, then exploration terminates and we move to the exploitation phase.

Exploitation phase

- *Expend budget*: After exploration terminates and an "optimal" model subset $\widehat{\mathcal{S}}^*$ has been identified, we expend the remaining computational budget on sampling $X_{\widehat{\mathcal{S}}^*}$.
- *Estimate CDF*: Using the collected samples, we construct the CDF estimator $\widetilde{F}_{\widehat{\mathcal{S}}^*}$ for F_Y, which is defined in (5.10).

The precise details of how the loss function is computed and the CDF estimator is constructed is the topic of Section 5, with Sections 2 to 4 serving to make requisite mathematical and statistical definitions.

A more detailed version of the algorithm is given in Algorithm 5.2, which lists more explicit computational steps that must be taken. The coming sections are devoted to the theoretical construction of quantities in Figure 1; in particular Sections 2 and 3 provide a construction of a loss function that is the integral part of exploration decision-making.

2. Background: control variates.

We first introduce the control variates method, which is a standard approach for variance reduction in MC simulation. For a random variable X with bounded variance $\sigma_X^2 > 0$, the size-m MC estimator for $\mathbb{E}[X]$ based on i.i.d. data X_ℓ, $\widehat{x} = \sum_{\ell \in \{1:m\}} X_\ell / m$, is unbiased and has variance σ_X^2/m. Given a random vector $Z =$

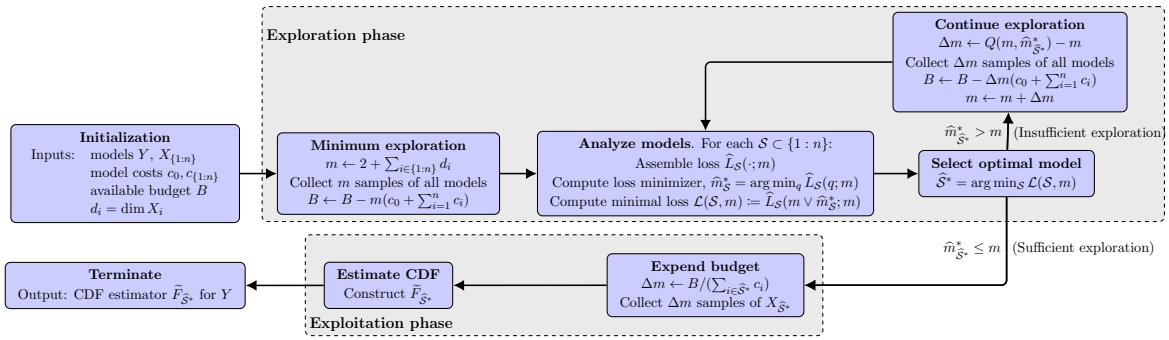


Figure 1: Flowchart illustration of the cvMDL algorithm. More details of the steps are discussed in Subsection 1.4. The full algorithm is presented in Algorithm 5.2.

$(Z^{(1)}, \ldots, Z^{(d)})^\top \in \mathbb{R}^d$ that lives in the same probability space as X, one may use joint i.i.d samples of (X, Z), i.e., $(X_\ell, Z_\ell^\top) = (X_\ell, Z_\ell^{(1)}, \ldots, Z_\ell^{(d)})$ for $\ell \in \{1:m\}$, to construct a control variates estimator $\widehat{x}_{\mathrm{cv}}$ for $\mathbb{E}[X]$:

$$\widehat{x}_{\mathrm{cv}} = \frac{1}{m} \sum_{\ell\in\{1:m\}} X_\ell - \frac{1}{m} \sum_{\ell\in\{1:m\}} (Z_\ell^\top \beta - \mathbb{E}[Z]^\top \beta),$$

where $\beta \in \mathbb{R}^d$ is some appropriately chosen vector and $\mathbb{E}[Z]$ is assumed known. The estimator $\widehat{x}_{\mathrm{cv}}$ is also unbiased and has variance

$$\sigma_{\mathrm{cv}}^2 = \mathrm{Var}[\widehat{x}_{\mathrm{cv}}] = \frac{\mathrm{Var}[X - Z^\top \beta]}{m}.$$

This variance is minimized when β is the least-squares coefficient for centered linear regression, i.e., for regressing $(X - \mathbb{E}[X])$ on $(Z - \mathbb{E}[Z])$,

(2.1)
$$\beta = \mathrm{Cov}[Z]^{-1}\mathrm{Cov}[Z, X] \implies \sigma_*^2 = \min_{\beta\in\mathbb{R}^d} \sigma_{\mathrm{cv}}^2 = \frac{(1-\rho^2)\sigma_X^2}{m}, \quad \rho = \mathrm{Corr}(X, Z^\top \mathrm{Cov}[Z]^{-1}\mathrm{Cov}[Z, X]).$$

When $|\rho| \approx 1$, the variance reduction is significant, in which case $Z^\top\beta$ accounts for most of $\mathrm{Var}[X]$.

When $\mathbb{E}[Z]$ is unknown, one may consider the following *approximate control variates* estimator that uses an independent size-N MC estimator in place of $\mathbb{E}[Z]$ using samples $\widetilde{Z}_j$:

(2.2)
$$\widehat{x}_{\mathrm{acv}} = \frac{1}{m} \sum_{\ell\in\{1:m\}} X_\ell - \frac{1}{m} \sum_{\ell\in\{1:m\}} \left(Z_\ell^\top \beta - \frac{1}{N} \sum_{j\in\{1:N\}} \widetilde{Z}_j^\top \beta \right) \quad (\ell, j) \in \{1:m\} \times \{1:N\},$$

where we assume $Z_\ell \perp\!\!\!\perp \widetilde{Z}_j$. Then this has variance

(2.3)
$$\sigma_\star^2 = \sigma_*^2 + \frac{\mathrm{Var}[Z^\top\beta]}{N} = \frac{(1-\rho^2)\sigma_X^2}{m} + \frac{\rho^2\sigma_X^2}{N}.$$

Construction of such approximate control variates estimators has been recently studied in the multifidelity estimation of first-order statistics [16, 35]. The terms $\mathrm{Cov}[Z]^{-1}$ and $\mathrm{Cov}[Z, X]$ may be estimated empirically at the cost of incurring higher-order trajectory-wise statistical errors in m [14, 26].

3. Variance reduction for CDF estimation. Control variates can be more generally applied to CDF estimation of nonlinear functions of random variables [15, 20]. For example, in risk management applications [15], the authors considered using the *delta-gamma* approximation[2] (i.e. the second-order Taylor expansion) of a loss function L at a given position $\boldsymbol{x}$ along random market move direction $\boldsymbol{\eta}$ as control variates to compute its quantiles. More precisely, one uses a quadratic function of $\boldsymbol{\eta}$ to approximate the loss at $\boldsymbol{x}$:

$$-(L(\boldsymbol{x}+\boldsymbol{\eta}) - L(\boldsymbol{x})) =: \ell(\boldsymbol{x}) \approx \widehat{\ell}(\boldsymbol{x}) := -\nabla L(\boldsymbol{x})^\top \boldsymbol{\eta} - \frac{1}{2}\boldsymbol{\eta}^\top \nabla^2 L(\boldsymbol{x})\boldsymbol{\eta}.$$

Fixing a scalar C, $\mathbf{1}_{\{\widehat{\ell}(\boldsymbol{x})\le C\}}$ can be used as a control variate for $\mathbf{1}_{\{\ell(\boldsymbol{x})\le C\}}$ to compute the latter's expectation, which in particular provides a way to compute CDFs. More advanced approximation techniques have been introduced in [20] to construct other control variates in the value-at-risk computation.

We apply a similar idea in the proposed multifidelity setup here. In our setup, a specific functional form may be computationally difficult to produce, and Taylor-like approximations can be inaccurate outside local regions. Our alternative strategy is to employ a global emulator for Y based on linear combinations of $X_{\{1:n\}}$, which can be effective when the correlation between these quantities is high. For example, this situation is often true when modeling parametric PDEs. In the rest of the section, we introduce a general multifidelity approach to estimate $F_Y(x)$ subject to a budget constraint.

3.1. Control variates for multifidelity CDF estimation. In developing the proposed method, we frequently resort to the simple observation that

$$F_Y(\boldsymbol{x}) = \mathbb{E}[\mathbf{1}_{\{Y\le\boldsymbol{x}\}}], \qquad \boldsymbol{x}\in\mathbb{R}^d.$$

If we fix $\mathcal{S}\subseteq\{1:n\}$, the control variate based on $X_{\mathcal{S}}$ that minimizes variance (and hence is optimal) is $\mathbb{E}[\mathbf{1}_{\{Y\le\boldsymbol{x}\}}|X_{\mathcal{S}}]$ [31]. This quantity requires the orthogonal projection of Y onto the sigma-field generated by $X_{\mathcal{S}}$, which is computationally intractable without special assumptions (e.g. joint normality). In order to approximate $\mathbb{E}[\mathbf{1}_{\{Y\le\boldsymbol{x}\}}|X_{\mathcal{S}}]$, we use $h(X_{\mathcal{S}};\boldsymbol{x})$ to denote a general $X_{\mathcal{S}}$-measurable function that serves as the control variates for $\mathbf{1}_{\{Y\le\boldsymbol{x}\}}$. We make a particular choice for h in Section 4.

Analogous to (2.2), we construct an approximate control variates estimator for $F_Y(\boldsymbol{x})$, where the m and N in (2.2) are related by the budget constraint (the cost of sampling Y and $X_{\mathcal{S}}$). Since different subsets $\mathcal{S}$ are considered simultaneously, we take a *uniform exploration policy* that first collects m i.i.d joint exploration samples of the *full* model for variance reduction and then commits the remainder of the budget to collect $N_{\mathcal{S}}$ samples of

[2]Here we refer to the "full" delta-gamma approximation. The more commonly used delta-gamma approximation in practice does not consider the second-order cross terms.

a selected model subset $\mathcal{S}$ of low-fidelity models to compute the control variates mean. The exploration samples and exploitation samples under a uniform exploration policy are denoted:

$$\text{(3.1)} \qquad \text{Exploration samples: } \{(X_{\text{epr},\ell,1}^\top, \ldots, X_{\text{epr},\ell,n}^\top, Y_{\text{epr},\ell}^\top)^\top\}_{\ell\in\{1:m\}} \subset \mathbb{R}^{d+\sum_{i=1}^n d_i}$$

$$\text{(3.2)} \qquad \text{Exploitation samples: } \{X_{\text{ept},j,\mathcal{S}}\}_{j\in\{1:N_\mathcal{S}\}},$$

where the subscripts "epr" and "ept" specify the stage where a sample is used. The parameters m and $N_\mathcal{S}$ are related by the budget constraint:

$$\text{(3.3)} \qquad N_\mathcal{S} = \frac{B - c_{\text{epr}} m}{c_\mathcal{S}} \qquad\qquad c_{\text{epr}} = \sum_{i=0}^{n} c_i, \qquad\qquad c_\mathcal{S} = \sum_{i\in\mathcal{S}} c_i,$$

where we ignore integer rounding effects to simplify the discussion. The control variates estimator for $F_Y(\boldsymbol{x})$ based on $h(X_\mathcal{S};\boldsymbol{x})$ is

$$\text{(3.4)} \qquad \widehat{F}_\mathcal{S}(\boldsymbol{x}) = \frac{1}{m}\sum_{\ell\in\{1:m\}} \mathbf{1}_{\{Y_{\text{epr},\ell}\le\boldsymbol{x}\}} - \frac{1}{m}\sum_{\ell\in\{1:m\}} \alpha(\boldsymbol{x}) \left(h(X_{\text{epr},\ell,\mathcal{S}};\boldsymbol{x}) - \frac{1}{N_\mathcal{S}} \sum_{j\in\{1:N_\mathcal{S}\}} h(X_{\text{ept},j,\mathcal{S}};\boldsymbol{x}) \right),$$

where $\alpha(\boldsymbol{x})$ is the optimal scaling coefficient as in (2.1):

$$\text{(3.5)} \qquad \alpha(\boldsymbol{x}) = \text{Cov}[h(X_\mathcal{S};\boldsymbol{x})]^{-1}\text{Cov}[\mathbf{1}_{\{Y\le\boldsymbol{x}\}}, h(X_\mathcal{S};\boldsymbol{x})].$$

Note $\alpha(\boldsymbol{x})$ is undefined if $\text{Cov}[h(X_\mathcal{S};\boldsymbol{x})] = 0$. In this case, the value of the estimator $\widehat{F}_\mathcal{S}(\boldsymbol{x})$ does not depend on $\alpha(\boldsymbol{x})$, and we set $\alpha(\boldsymbol{x})$ to 0 for convenience. The quantity $\widehat{F}_\mathcal{S}(\boldsymbol{x})$ is an unbiased estimator for $F_Y(\boldsymbol{x})$ with variance

$$\text{Var}[\widehat{F}_\mathcal{S}(\boldsymbol{x})] = \frac{(1-\rho_\mathcal{S}^2(\boldsymbol{x}))F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))}{m} + \frac{\rho_\mathcal{S}^2(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))}{N_\mathcal{S}},$$

where

$$\text{(3.6)} \qquad \rho_\mathcal{S}(\boldsymbol{x}) = \text{Corr}[\mathbf{1}_{\{Y\le\boldsymbol{x}\}}, h(X_\mathcal{S};\boldsymbol{x})].$$

3.2. A control variates loss function. To measure the overall accuracy of $\widehat{F}_\mathcal{S}(\boldsymbol{x})$, we introduce the loss $L_\mathcal{S}$ defined by the average $\omega(\boldsymbol{x})$-weighted L^2-norm square of $\widehat{F}_\mathcal{S}(\boldsymbol{x}) - F_Y(\boldsymbol{x})$:

$$\text{(3.7)} \qquad L_\mathcal{S} \coloneqq \mathbb{E}\left[\int_{\mathbb{R}^d} \omega(\boldsymbol{x})|\widehat{F}_\mathcal{S}(\boldsymbol{x}) - F_Y(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x}\right],$$

where $\omega(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}_{\ge 0}$ is a weight function. The $\omega(\boldsymbol{x})$-weighted L^2-norm square is related to other more widely used metrics on distributions, e.g., it reduces to the Cramér–von Mises distance when $\omega(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathrm{d}F_Y(\boldsymbol{x})$. To estimate $L_\mathcal{S}$, note that by Tonelli's theorem, we have,

$$\text{(3.8)} \qquad L_\mathcal{S} = \int_{\mathbb{R}^d} \omega(\boldsymbol{x})\text{Var}[\widehat{F}_\mathcal{S}(\boldsymbol{x})]\mathrm{d}\boldsymbol{x} = \frac{k_1(\mathcal{S})}{m} + \frac{k_2(\mathcal{S})}{B - c_{\text{epr}} m},$$

where

$$k_1(\mathcal{S}) = \int_{\mathbb{R}^d} \omega(\boldsymbol{x})(1-\rho^2_{\mathcal{S}}(\boldsymbol{x}))F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}$$

$$k_2(\mathcal{S}) = c_{\mathcal{S}} \int_{\mathbb{R}^d} \omega(\boldsymbol{x})\rho^2_{\mathcal{S}}(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}. \tag{3.9}$$

Since $k_1(\mathcal{S})$ and $k_2(\mathcal{S})$ are nonnegative, a sufficient and necessary condition for $k_1(\mathcal{S})$ and $k_2(\mathcal{S})$ being well-defined (i.e. finite) is

$$k_1(\mathcal{S}) + c_{\mathcal{S}}^{-1} k_2(\mathcal{S}) = \int_{\mathbb{R}^d} \omega(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} < \infty, \tag{3.10}$$

but this need not hold for arbitrary choice of ω. For instance, when $\omega(\boldsymbol{x}) \equiv 1$, (3.10) is true when $d = 1$ if $\mathbb{E}[|Y|^{1+\delta}] < \infty$ for some $\delta > 0$. However, when $d \geq 2$, (3.10) is generally not true when the support for the distribution of Y is unbounded since $F_Y^{-1}([\varepsilon, 1-\varepsilon])$ may have infinite Lebesgue measure in $\mathbb{R}^d$ for some $\varepsilon > 0$. For such scenarios, requiring that $\omega(\boldsymbol{x})$ is integrable ensures (3.10), i.e.,

$$\int_{\mathbb{R}^d} \omega(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} < \int_{\mathbb{R}^d} \omega(\boldsymbol{x})\mathrm{d}\boldsymbol{x} < \infty.$$

Some typical choices for integrable $\omega(\boldsymbol{x})$ include $\omega(\boldsymbol{x}) = \mathbf{1}_{\mathcal{T}}$ where $\mathcal{T} \subset \mathbb{R}^d$ is a bounded domain of interest or $\omega(\boldsymbol{x})$ with reasonably fast decaying tails as $\|\boldsymbol{x}\|_2 \to \infty$. In the following discussion, we assume (3.10) holds (and later codify this as Assumption 5.5). We make different choices for ω in our numerical results of Section 6.

3.3. Exploration-exploitation trade-off. Equation (3.8) is similar to the exploration-exploitation loss trade-off that was originally formulated in [35], where k_1 and k_2 measure the errors committed by the exploration and the exploitation, respectively. Note that $L_{\mathcal{S}}$ is a strictly convex function for a valid exploration rate m, i.e., for $0 < m < B/c_{\mathrm{epr}}$, and achieves its unique minimum at $m^*_{\mathcal{S}}$ with corresponding minimum loss $L^*_{\mathcal{S}}$:

$$m^*_{\mathcal{S}} = \frac{B}{c_{\mathrm{epr}} + \sqrt{\frac{c_{\mathrm{epr}}k_2(\mathcal{S})}{k_1(\mathcal{S})}}}, \quad L^*_{\mathcal{S}} \coloneqq \min_{0<m<\frac{B}{c_{\mathrm{epr}}}} L_{\mathcal{S}}(m) = \frac{(\sqrt{c_{\mathrm{epr}}k_1(\mathcal{S})} + \sqrt{k_2(\mathcal{S})})^2}{B} =: \frac{\gamma_{\mathcal{S}}}{B}. \tag{3.11}$$

An optimal subset $\mathcal{S}$ is the one that minimizes the m-optimized loss value,

$$\mathcal{S}^* = \arg\min_{\mathcal{S}\subseteq\{1:n\}} \gamma_{\mathcal{S}}. \tag{3.12}$$

A uniform exploration policy is called *optimal* if it collects $m^*_{\mathcal{S}^*}$ joint samples for exploration and uses model $\mathcal{S}^*$ for exploitation. This is, in effect, a model selection procedure, as an optimal exploration policy selects the model subset that yields the smallest error via optimally balancing the trade-off between exploration and exploitation. In the following discussion, we assume $\mathcal{S}^*$ is unique.

As a benchmark to the procedure above (with oracle information), one can consider an empirical (ECDF) procedure that devotes the full budget to sampling the high-fidelity model Y, ignoring the lower-fidelity models. The following result relates the error between these two approaches.

Lemma 3.1. *With $L^*_{\mathcal{S}^*} = \gamma_{\mathcal{S}^*}/B$ the minimum error achieved by a uniform exploration policy as described above, and $c_{epr}\int_{\mathbb{R}^d}\omega(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))d\boldsymbol{x}/B$ the expected error achieved by an ECDF estimator for F_Y, then*

$$\frac{c_{epr}\int_{\mathbb{R}^d}\omega(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))d\boldsymbol{x}/B}{\gamma_{\mathcal{S}^*}/B} \geq \frac{1}{2\left(\frac{c_{\mathcal{S}}}{c_{epr}}+\mathbb{E}_Z[(1-\rho^2_{\mathcal{S}^*}(Z))]\right)} \geq \frac{1}{4}$$

where Z is a random variable with (unnormalized) density $\omega(\boldsymbol{z})F_Y(\boldsymbol{z})(1-F_Y(\boldsymbol{z}))$.

Proof. We have

$$\begin{aligned}\frac{c_{\mathrm{epr}}\int_{\mathbb{R}^d}\omega(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}/B}{\gamma_{\mathcal{S}^*}/B} &= \frac{c_{\mathrm{epr}}\int_{\mathbb{R}^d}\omega(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}{\left(\sqrt{c_{\mathrm{epr}}k_1(\mathcal{S}^*)}+\sqrt{k_2(\mathcal{S}^*)}\right)^2}\\ &\geq \frac{c_{\mathrm{epr}}\int_{\mathbb{R}^d}\omega(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}{2\left(c_{\mathrm{epr}}k_1(\mathcal{S}^*)+k_2(\mathcal{S}^*)\right)}\\ &\geq \frac{1}{2\left(\frac{c_{\mathcal{S}}}{c_{\mathrm{epr}}}+\mathbb{E}_{\boldsymbol{x}}[(1-\rho^2_{\mathcal{S}^*}(\boldsymbol{x}))]\right)} \geq \frac{1}{4},\end{aligned}$$

where the expectation $\mathbb{E}_{\boldsymbol{x}}[\cdot]$ is taken with respect to

$$\boldsymbol{x} \sim \frac{\omega(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}^d}\omega(\boldsymbol{z})F_Y(\boldsymbol{z})(1-F_Y(\boldsymbol{z}))\mathrm{d}\boldsymbol{z}},$$

and the last inequality follows by noting $c_{\mathcal{S}} \leq c_{\mathrm{epr}}$ and $0 \leq \mathbb{E}_{\boldsymbol{x}}[(1-\rho^2_{\mathcal{S}^*}(\boldsymbol{x}))] \leq 1$. ∎

Hence, the relative efficiency of a uniform exploration policy compared to the ECDF estimator is unconditionally bounded below by $1/4$, and hence the uniform exploration policy can at worst realize a loss value of 4 times a naive ECDF procedure. On the other hand, the relative efficiency is $\gg 1$ if both $c_{\mathcal{S}}/c_{\mathrm{epr}}$ and $\mathbb{E}_{\boldsymbol{x}}[(1-\rho^2_{\mathcal{S}^*}(\boldsymbol{x}))]$ are small. This happens, for instance, if $X_{\mathcal{S}^*}$ has a much smaller sampling cost than Y and $h(X_{\mathcal{S}};\boldsymbol{x})$ are "good" control variates for $\mathbf{1}_{\{Y\leq\boldsymbol{x}\}}$ uniformly for all $\boldsymbol{x}\in\mathbb{R}^d$, both of which are realistic occurrences in multifidelity applications.

4. Choosing control variates from linear approximations.

We propose a procedure for selecting the control variate h, which boils down to constructing approximations of $\mathbb{E}[\mathbf{1}_{\{Y\leq\boldsymbol{x}\}}|X_{\mathcal{S}}]$ that both retain high correlation with Y and are budget-friendly. While one may generate special forms for approximations in particular cases, our goal is a simple and generic choice that is useful for many practical applications.

Recall that $X_{\mathcal{S}^+} = (1, X_i^\top)^\top_{i\in\mathcal{S}} \in \mathbb{R}^{d_{\mathcal{S}}+1}$. For $i\in\{1:d\}$, let $\boldsymbol{\beta}^{(i)}_{\mathcal{S}^+}$ be the optimal linear projection coefficients for estimating the ith component of Y using $X_{\mathcal{S}^+}$:

$$\boldsymbol{\beta}^{(i)}_{\mathcal{S}^+} = (\mathbb{E}[X_{\mathcal{S}^+}X_{\mathcal{S}^+}^\top])^{-1}\mathrm{Cov}[X_{\mathcal{S}^+}, Y^{(i)}] \in \mathbb{R}^{d_{\mathcal{S}}+1}, \quad \boldsymbol{B}_{\mathcal{S}^+} = [\boldsymbol{\beta}^{(1)}_{\mathcal{S}^+},\cdots,\boldsymbol{\beta}^{(d)}_{\mathcal{S}^+}] \in \mathbb{R}^{(d_{\mathcal{S}}+1)\times d}, \tag{4.1}$$

The least squares approximation of Y using linear combinations of $X_{\mathcal{S}}$ and 1 is given by

$$H_{\mathcal{S}}(X_{\mathcal{S}}) := (X_{\mathcal{S}^+}^\top\boldsymbol{B}_{\mathcal{S}^+})^\top = \begin{bmatrix} X_{\mathcal{S}^+}^\top\boldsymbol{\beta}^{(1)}_{\mathcal{S}^+} \\ \vdots \\ X_{\mathcal{S}^+}^\top\boldsymbol{\beta}^{(d)}_{\mathcal{S}^+}\end{bmatrix} \in \mathbb{R}^d.$$

When all quantities are scalars, i.e., $d = d_1 = \cdots = d_n = 1$, one can directly manipulate $H_{\mathcal{S}}$ to estimate the statistics of Y [35, 36]. Such an approach is easy to implement and enjoys certain robustness for first-order statistics, but is more prone to model misspecification effects (e.g. expressibility of the linear model, noise assumption, etc.) when the whole distribution of Y is to be learned due to the limitation of linear approximation [36].

To address the issue, we take an additional nonlinear step beyond $H_{\mathcal{S}}$. In particular, we consider the following family of control variates that slice the estimator $H_{\mathcal{S}}$:

$$h(X_{\mathcal{S}}; \boldsymbol{x}) = \mathbf{1}_{\{H_{\mathcal{S}} \le \boldsymbol{x}\}}. \tag{4.2}$$

Intuitively, we may expect $\mathbf{1}_{\{Y \le \boldsymbol{x}\}}$ and $h(X_{\mathcal{S}}; \boldsymbol{x})$ to be correlated if $\mathbb{E}[\|Y - H_{\mathcal{S}}\|_2^2]$ is small. However, this may not be true for $\boldsymbol{x}$ approaching the tails of Y. For instance, assuming $d = 1$ and a standard joint Gaussian random vector (X, Y) with correlation ρ,

$$\lim_{x \to -\infty} \operatorname{Corr}(\mathbf{1}_{\{Y \le x\}}, \mathbf{1}_{\{X \le x\}}) = \lim_{x \to 0} \frac{C_{X,Y}(x, x)}{x} = \begin{cases} 1, & |\rho| = 1, \\ 0, & |\rho| < 1 \end{cases}$$

where $C_{X,Y}(x, y) = \mathbb{P}\left(\Phi^{-1}(X) \le x, \Phi^{-1}(Y) \le y\right)$ is the Gaussian copula and Φ^{-1} is the quantile of a standard normal distribution; see [24]. Hence, $\mathbf{1}_{\{X \le x\}}$ is not a good control variate for $\mathbf{1}_{\{Y \le x\}}$ when $|x| \to \infty$ unless $|\rho| = 1$, i.e., only if $X \propto Y$. Nevertheless, our experiments in Section 6 show that in practice $h(X_{\mathcal{S}}; \boldsymbol{x})$ provides a reasonable control variates choice for many scenarios in multifidelity simulations, and thus suggests that situations described above are less common for the applications of our interest. We discuss computational aspects of using (4.2) as control variates in Section 5.1.

Choosing h as in (4.2), the coefficient $\alpha(\boldsymbol{x})$ in (3.5) can be explicitly computed as

$$\alpha(\boldsymbol{x}) = \begin{cases} \frac{F_{Y \vee H_{\mathcal{S}}}(\boldsymbol{x}) - F_Y(\boldsymbol{x}) F_{H_{\mathcal{S}}}(\boldsymbol{x})}{F_{H_{\mathcal{S}}}(\boldsymbol{x})(1 - F_{H_{\mathcal{S}}}(\boldsymbol{x}))} & \boldsymbol{x} \in \operatorname{supp}(F_{H_{\mathcal{S}}}(\boldsymbol{x}))^\circ \\ 0 & \text{otherwise} \end{cases}. \tag{4.3}$$

One useful technical result is that $\alpha(\boldsymbol{x})$ is bounded.

Lemma 4.1. *Let $\alpha(\boldsymbol{x})$ be given as in (4.3). Then, $|\alpha(\boldsymbol{x})| \le 1$.*

Proof. It suffices to check that for $\boldsymbol{x} \in \operatorname{supp}(F_{H_{\mathcal{S}}}(\boldsymbol{x}))^\circ$, $\alpha(\boldsymbol{x}) \le 1$ and $-\alpha(\boldsymbol{x}) \le 1$ hold simutaneously:

$$\begin{aligned} \frac{F_{Y \vee H_{\mathcal{S}}}(\boldsymbol{x}) - F_Y(\boldsymbol{x}) F_{H_{\mathcal{S}}}(\boldsymbol{x})}{F_{H_{\mathcal{S}}}(\boldsymbol{x})(1 - F_{H_{\mathcal{S}}}(\boldsymbol{x}))} &\le \frac{F_Y(\boldsymbol{x}) \wedge F_{H_{\mathcal{S}}}(\boldsymbol{x}) - F_Y(\boldsymbol{x}) F_{H_{\mathcal{S}}}(\boldsymbol{x})}{F_{H_{\mathcal{S}}}(\boldsymbol{x})(1 - F_{H_{\mathcal{S}}}(\boldsymbol{x}))} \\ &= \frac{F_Y(\boldsymbol{x})}{F_{H_{\mathcal{S}}}(\boldsymbol{x})} \wedge \frac{1 - F_Y(\boldsymbol{x})}{1 - F_{H_{\mathcal{S}}}(\boldsymbol{x})} \le 1, \end{aligned}$$

and

$$\begin{aligned} \frac{F_Y(\boldsymbol{x}) F_{H_{\mathcal{S}}}(\boldsymbol{x}) - F_{Y \vee H_{\mathcal{S}}}(\boldsymbol{x})}{F_{H_{\mathcal{S}}}(\boldsymbol{x})(1 - F_{H_{\mathcal{S}}}(\boldsymbol{x}))} &\le \frac{F_Y(\boldsymbol{x}) F_{H_{\mathcal{S}}}(\boldsymbol{x}) - F_Y(\boldsymbol{x}) + 1 - F_{H_{\mathcal{S}}}(\boldsymbol{x})}{F_{H_{\mathcal{S}}}(\boldsymbol{x})(1 - F_{H_{\mathcal{S}}}(\boldsymbol{x}))} \wedge \frac{F_Y(\boldsymbol{x}) F_{H_{\mathcal{S}}}(\boldsymbol{x})}{F_{H_{\mathcal{S}}}(\boldsymbol{x})(1 - F_{H_{\mathcal{S}}}(\boldsymbol{x}))} \\ &\le \frac{1 - F_Y(\boldsymbol{x})}{F_{H_{\mathcal{S}}}(\boldsymbol{x})} \wedge \frac{F_Y(\boldsymbol{x})}{1 - F_{H_{\mathcal{S}}}(\boldsymbol{x})} \le 1. \end{aligned}$$

∎

5. Algorithms. We revisit the cvMDL algorithm in Figure 1: the loss function $L_{\mathcal{S}}$ in (3.8) is the desired loss function to optimize over but requires oracle statistics (i.e. $k_1(\mathcal{S})$ and $k_2(\mathcal{S})$). Thus, we replace it with an approximation $\widehat{L}_{\mathcal{S}}$ that we describe in this section. Additionally, the computations in the "Analyze Models" step are now more transparent: The oracle computations are given by (3.12) and (3.11). In a practical algorithmic setting, we replace these with consistent approximate computations, which is the topic of this section.

When using approximate quantities to compute $L_{\mathcal{S}}$, the explicit exploration-exploitation loss decomposition in (3.8) may no longer be true. Nevertheless, if the quantities we estimate are sufficiently accurate, then such a decomposition is expected to be approximately valid. Thus, in devising practical algorithms, we use the oracle loss form (3.8) (with estimated coefficients) instead of (3.7) as the criteria for model selection. We present in the numerical section some empirical evidence that such a replacement has little impact on model selection.

Since the proposed estimators change when new exploration samples are collected, the dependence on this number of exploration samples must be made explicit. For $\mathcal{S} \subseteq \{1:n\}$, we let $\widehat{L}_{\mathcal{S}}(\cdot;t)$ denote the estimated loss function $L_{\mathcal{S}}$ after having collected t exploration samples. We then let $\widehat{m}^*_{\mathcal{S}}$ be the corresponding estimator for the optimal exploration sample size $m^*_{\mathcal{S}}$. Summarizing this: the intuition behind the cvMDL algorithm is that we use currently collected exploration data (t samples) to find the estimated optimal model ($\widehat{\mathcal{S}}^*$) and the corresponding exploration rate ($\widehat{m}^*_{\widehat{\mathcal{S}}^*}$). Based on the value of $\widehat{m}^*_{\widehat{\mathcal{S}}^*}$ relative to t, we decide whether to continue to explore or to switch to exploitation.

5.1. Estimators for oracle quantities. In this section, we discuss how to estimate $L_{\mathcal{S}}$, $m^*_{\mathcal{S}}$, and $\alpha(\boldsymbol{x})$ from exploration data when instantiating cvMDL using the linear approximators as introduced in Section 4. The control variates $h(X_{\mathcal{S}};\boldsymbol{x}) = \mathbf{1}_{\{H_{\mathcal{S}}(X_{\mathcal{S}})\le \boldsymbol{x}\}}$ belong to a parametric family characterized by $\boldsymbol{\beta}^{(i)}_{\mathcal{S}^+}, i \in \{1:d\}$ from (4.1), which can be estimated from exploration data.

Recall from (3.1) that the ℓth exploration sample of all low-fidelity models in $\mathcal{S}$ is denoted by $X_{\mathrm{epr},\ell,\mathcal{S}}$. Similarly, we define $X_{\mathrm{epr},\ell,\mathcal{S}^+} := (1, X^\top_{\mathrm{epr},\ell,\mathcal{S}})^\top$. To estimate $\boldsymbol{\beta}^{(i)}_{\mathcal{S}^+}$, we use the least-squares estimator:

$$\widehat{\boldsymbol{\beta}}^{(i)}_{\mathcal{S}^+} = \boldsymbol{Z}^\dagger_{\mathcal{S}} \boldsymbol{Y}^{(i)} \qquad \boldsymbol{Z}_{\mathcal{S}} = \begin{bmatrix} X^\top_{\mathrm{epr},1,\mathcal{S}^+} \\ \vdots \\ X^\top_{\mathrm{epr},m,\mathcal{S}^+} \end{bmatrix} \in \mathbb{R}^{m\times(d_{\mathcal{S}}+1)} \quad \boldsymbol{Y}^{(i)} = \begin{bmatrix} Y^{(i)}_{\mathrm{epr},1} \\ \vdots \\ Y^{(i)}_{\mathrm{epr},m} \end{bmatrix} \in \mathbb{R}^m, \tag{5.1}$$

where $(X^\top_{\mathrm{epr},\ell,1}, \dots, Y^\top_{\mathrm{epr},\ell})^\top_{\ell\in\{1:m\}}$ are joint exploration samples, and the design matrix $\boldsymbol{Z}_{\mathcal{S}}$ is assumed to have full column rank[3].

For $\boldsymbol{x} \in \mathbb{R}^d$, $h(X_{\mathcal{S}};\boldsymbol{x})$ can be estimated as

$$\widehat{h}(X_{\mathcal{S}};\boldsymbol{x}) = \mathbf{1}_{\{\widehat{H}_{\mathcal{S}}(X_{\mathcal{S}})\le \boldsymbol{x}\}} \qquad \widehat{H}_{\mathcal{S}}(X_{\mathcal{S}}) \coloneqq \widehat{\boldsymbol{B}}^\top_{\mathcal{S}^+} X_{\mathcal{S}^+} =: \begin{bmatrix} X^\top_{\mathcal{S}^+}\widehat{\boldsymbol{\beta}}^{(1)}_{\mathcal{S}^+} \\ \vdots \\ X^\top_{\mathcal{S}^+}\widehat{\boldsymbol{\beta}}^{(d)}_{\mathcal{S}^+} \end{bmatrix}, \tag{5.2}$$

[3]This motivates the minimal exploration size condition in Algorithm 1, which is a neccesary condition for full rank here.

For ease of notation, we write $\widehat{H}_{\mathcal{S}}(X_{\mathcal{S}})$ and $\widehat{h}_{\mathcal{S}}(X_{\mathcal{S}};\boldsymbol{x})$ as $\widehat{H}_{\mathcal{S}}$ and $\widehat{h}_{\mathcal{S}}$ when $X_{\mathcal{S}^+}$ is generic and not necessarily related to the exploration and exploitation data. We introduce some additional notation for quantities involving both estimated coefficients and empirical CDFs using exploration data:

$$\widehat{F}_Y(\boldsymbol{x}) = \frac{1}{m}\sum_{\ell\in\{1:m\}} \mathbf{1}_{\{Y_{\mathrm{epr},\ell}\le\boldsymbol{x}\}}$$

$$\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) = \frac{1}{m}\sum_{\ell\in\{1:m\}} \widehat{h}(X_{\mathrm{epr},\ell,\mathcal{S}};\boldsymbol{x}) = \frac{1}{m}\sum_{\ell\in\{1:m\}} \mathbf{1}_{\{\widehat{H}_{\mathcal{S}}(X_{\mathrm{epr},\ell,\mathcal{S}})\le\boldsymbol{x}\}}$$

$$\widehat{F}_{Y\vee\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) = \frac{1}{m}\sum_{\ell\in\{1:m\}} \mathbf{1}_{\{Y_{\mathrm{epr},\ell}\vee\widehat{H}_{\mathcal{S}}(X_{\mathrm{epr},\ell,\mathcal{S}})\le\boldsymbol{x}\}}.$$

To compute the loss function approximation, we build approximations to k_1 and k_2 in (3.9), which requires us to compute $\rho^2_{\mathcal{S}}$ in (3.6). For this purpose, observe that

$$(1-\rho^2_{\mathcal{S}}(\boldsymbol{x}))F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x})) = \mathbb{E}[(\mathbf{1}_{\{Y\le\boldsymbol{x}\}} - F_Y(\boldsymbol{x})) - \alpha(\boldsymbol{x})(\mathbf{1}_{\{H_{\mathcal{S}}\le\boldsymbol{x}\}} - F_{H_{\mathcal{S}}}(\boldsymbol{x}))]^2,$$

where $\alpha(\boldsymbol{x})$ is defined in (4.3). The quantity in the expectation is the mean squared regression residual between two centered Bernoulli random variables $\mathbf{1}_{\{Y\le\boldsymbol{x}\}}$ and $\mathbf{1}_{\{H_{\mathcal{S}}\le\boldsymbol{x}\}}$. Thus, a natural estimator for $(1-\rho^2_{\mathcal{S}}(\boldsymbol{x}))F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))$ is to compute an empirical mean-squared difference between $\mathbf{1}_{\{Y\le x\}}$ and a regressor with covariates $\mathbf{1}_{\{\widehat{H}_{\mathcal{S}}\le x\}}$, which requires data for Y. Since we have (uncentered) data for Y on the exploration samples $Y_{\mathrm{epr},j}$ for $j\in\{1:m\}$, we can evaluate a regressor for Y with covariates $\widehat{H}_{\mathcal{S}}$ together with the intercept term on the exploration data sites. This results in the following estimator $\mathcal{K}_1$ for $(1-\rho^2_{\mathcal{S}}(\boldsymbol{x}))F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))$

$$\mathcal{K}_1(\boldsymbol{x}) = \frac{1}{m}\sum_{\ell\in\{1:m\}} (\mathbf{1}_{\{Y_{\mathrm{epr},\ell}\le\boldsymbol{x}\}} - r_\ell(\boldsymbol{x}))^2 \qquad \begin{bmatrix} r_1(\boldsymbol{x}) \\ \vdots \\ r_m(\boldsymbol{x}) \end{bmatrix} = \boldsymbol{W}_{\mathcal{S}}\boldsymbol{W}_{\mathcal{S}}^{\dagger} \begin{bmatrix} \mathbf{1}_{\{Y_{\mathrm{epr},1}\le\boldsymbol{x}\}} \\ \vdots \\ \mathbf{1}_{\{Y_{\mathrm{epr},m}\le\boldsymbol{x}\}} \end{bmatrix}, \tag{5.3}$$

where

$$\boldsymbol{W}_{\mathcal{S}} = \begin{bmatrix} 1 & \widehat{h}_{\mathcal{S}}(X_{\mathrm{epr},1,\mathcal{S}};\boldsymbol{x}) \\ \vdots & \vdots \\ 1 & \widehat{h}_{\mathcal{S}}(X_{\mathrm{epr},m,\mathcal{S}};\boldsymbol{x}) \end{bmatrix}.$$

This allows us to estimate $\rho^2_{\mathcal{S}}(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))$ as

$$\mathcal{K}_2(\boldsymbol{x}) = \widehat{F}_Y(\boldsymbol{x})(1-\widehat{F}_Y(\boldsymbol{x})) - \mathcal{K}_1(\boldsymbol{x}). \tag{5.4}$$

Consequently, we can estimate $k_1(\mathcal{S})$ and $k_2(\mathcal{S})$ as

$$\widehat{k}_1(\mathcal{S}) = \int_{\mathbb{R}^d} \omega(\boldsymbol{x})\mathcal{K}_1(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \qquad \widehat{k}_2(\mathcal{S}) = c_{\mathcal{S}}\int_{\mathbb{R}^d} \omega(\boldsymbol{x})\mathcal{K}_2(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{5.5}$$

The above estimators for $k_1(\mathcal{S})$ and $k_2(\mathcal{S})$ are positive and coincide with empirical estimators for these quantities whenever defined (see Appendix A.2.3), which is a crucial realization for our consistency results later. Plugging the above estimates into (3.8) and (3.11) yields estimates for $L_{\mathcal{S}}$ and $m^*_{\mathcal{S}}$:

$$\widehat{L}_{\mathcal{S}}(z;m) = \frac{\widehat{k}_1(\mathcal{S})}{z} + \frac{\widehat{k}_2(\mathcal{S})}{B - c_{\mathrm{epr}} z} \qquad \widehat{m}^*_{\mathcal{S}} = \frac{B}{c_{\mathrm{epr}} + \sqrt{\frac{c_{\mathrm{epr}}\widehat{k}_2(\mathcal{S})}{\widehat{k}_1(\mathcal{S})}}}. \tag{5.6}$$

Note that $\widehat{L}_{\mathcal{S}}(z;m)$ has a parameter m indicating the number of exploration samples used to compute $\widehat{k}_1(\mathcal{S})$ and $\widehat{k}_2(\mathcal{S})$, and a variable z denoting the exploration rate where to evaluate $\widehat{L}_{\mathcal{S}}$. We define $\widehat{\mathcal{S}}^*$ as the optimal model selected by this estimator,

$$\widehat{\mathcal{S}}^* = \operatorname*{arg\,min}_{\mathcal{S}\subseteq\{1:n\}} \widehat{L}_{\mathcal{S}}(\widehat{m}^*_{\mathcal{S}}; m), \tag{5.7}$$

which parallels the oracle choice (3.12). We have described all quantities needed to complete the exploration phase of Figure 1. What remains is to describe how the CDF estimator $\widetilde{F}_{\widehat{\mathcal{S}}^*}$ in the exploitation phase of Figure 1 is generated.

Our exploitation goal is to generate an estimator for (3.4), and so we need to estimate $\alpha(\boldsymbol{x})$:

$$\widehat{\alpha}(\boldsymbol{x}) = \frac{\widehat{F}_{Y\vee\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) - \widehat{F}_Y(\boldsymbol{x})\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x})}{\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x})(1 - \widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}))} \qquad \boldsymbol{x} \in \operatorname{supp}(\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}))^\circ, \tag{5.8}$$

and $\widehat{\alpha}(\boldsymbol{x}) = 0$ zero otherwise. By a similar reasoning as in Lemma 4.1, one has

$$|\widehat{\alpha}(\boldsymbol{x})| \le 1. \tag{5.9}$$

Finally, the *exploitation* estimator $\widetilde{F}_{\mathcal{S}}(\boldsymbol{x})$ for $F_Y(\boldsymbol{x})$ based on estimated parameters, utilizes $N_{\mathcal{S}}$ exploitation samples (i.e., exhausts the remaining budget B) and is given by,

$$\widetilde{F}_{\mathcal{S}}(\boldsymbol{x}) \coloneqq \widehat{F}_Y(\boldsymbol{x}) - \frac{1}{m}\sum_{\ell\in\{1:m\}} \left(\widehat{\alpha}(\boldsymbol{x})\widehat{h}_{\mathcal{S}}(X_{\mathrm{epr},\ell,\mathcal{S}};\boldsymbol{x}) - \frac{1}{N_{\mathcal{S}}}\sum_{j\in\{1:N_{\mathcal{S}}\}} \widehat{\alpha}(\boldsymbol{x})\widehat{h}_{\mathcal{S}}(X_{\mathrm{ept},\ell,\mathcal{S}};\boldsymbol{x}) \right), \tag{5.10}$$

where $\mathcal{S} = \widehat{\mathcal{S}}^*$ is the selected model based on $\widehat{k}_1(\mathcal{S})$ and $\widehat{k}_2(\mathcal{S})$. By inspection, we observe that $\widetilde{F}_{\mathcal{S}}(\boldsymbol{x})$ is a piecewise affine correction of $\widehat{F}_Y$, where the correction is based on the control variates $\widehat{h}_{\mathcal{S}}$.

Remark 5.1. *The estimator $\widehat{\alpha}(\boldsymbol{x})$ is undefined and manually set to zero for $\boldsymbol{x}$ outside the support of $\widehat{F}_{\widehat{H}_{\mathcal{S}}}$, as in that case the denominator vanishes. Alternatively, one can define $\widehat{\alpha}(\boldsymbol{x})$ for $\boldsymbol{x}$ outside the support of $\widehat{F}_{\widehat{H}_{\mathcal{S}}}$ as $\widehat{\alpha}(\boldsymbol{x}')$ for some $\boldsymbol{x}'$ inside the support of $\widehat{F}_{\widehat{H}_{\mathcal{S}}}$ that can be accurately estimated yet remains close to $\boldsymbol{x}$. To illustrate, consider $d = 1$. Assuming*

$\alpha(x)$ is a continuous function of x in supp$(F_{H_{\mathcal{S}}})$ and supp$(\widehat{F}_{\widehat{H}_{\mathcal{S}}}(x)) = [x_{\min}, x_{\max}]$ for some $x_{\min} < x_{\max}$, for $x \in$ supp$(\widehat{F}_{\widehat{H}_{\mathcal{S}}}(x))^c$, we may estimate $\alpha(x)$ outside $[x_{\min}, x_{\max}]$ as

$$\widehat{\alpha}(x) = \begin{cases} \frac{\widehat{F}_{Y\vee\widehat{H}_{\mathcal{S}}}(x(\tau))-\widehat{F}_Y(x(\tau))\tau}{\tau(1-\tau)} & x \le x_{\min} \\ \frac{\widehat{F}_{Y\vee\widehat{H}_{\mathcal{S}}}(x(1-\tau))-\widehat{F}_Y(x(1-\tau))(1-\tau)}{\tau(1-\tau)} & x \ge x_{\max}, \end{cases} \tag{5.11}$$

where $x(\tau)$ and $x(1-\tau)$ are the τ and $1-\tau$ quantiles of $\widehat{F}_{\widehat{H}_{\mathcal{S}}}$ for some small $\tau \in (0,1)$:

$$x(\tau) = \widehat{F}^{-1}_{\widehat{H}_{\mathcal{S}}}(\tau) \qquad\qquad x(1-\tau) = \widehat{F}^{-1}_{\widehat{H}_{\mathcal{S}}}(1-\tau).$$

This allows us to get nontrivial estimates of F_Y outside $[x_{\min}, x_{\max}]$, i.e. in the tail regime. When $d \ge 2$, one may generalize the ideas above by projecting the points in the tail regime to some bounded set in $\mathbb{R}^d$ that contains most of the measure in the domain.

5.2. Monotonicity of the exploitation CDF estimator. By construction, $\widetilde{F}_{\mathcal{S}}(\boldsymbol{x})$ is a piecewise constant function on some d-dimensional rectangular partition of $\mathbb{R}^d$, but is not necessarily a monotone nondecreasing function in each direction due to the fluctuations of estimators used in the construction. To address this issue, we introduce a dimension-wise recursive-sorting post-processing procedure on values in the range of $\widetilde{F}_{\mathcal{S}}$ to recover the desired monotonicity and ensure that we compute an actual distribution function. To represent $\widetilde{F}_{\mathcal{S}}(\boldsymbol{x})$ as a d-dimensional tensor, we introduce the index set $\boldsymbol{I} = \otimes_{i\in\{1:d\}}(z_{i,1}, \ldots, z_{i,M_i})$, $-\infty = z_{i,1} \le \cdots \le z_{i,M_i} = +\infty$, where $z_{i,j}$ is the jth order statistic of the projected partition points associated with $\widetilde{F}_{\mathcal{S}}(\boldsymbol{x})$, and M_i is the total number of such projected points. Using this notation, we express $\widetilde{F}_{\mathcal{S}}(\boldsymbol{x})$ as a tensor $\boldsymbol{T}$, where

$$\widetilde{F}_{\mathcal{S}}(\boldsymbol{x}) = \boldsymbol{T}_{z_{1,s_1},\cdots,z_{d,s_d}} \qquad\qquad \boldsymbol{x} \in \prod_{i\in\{1:d\}} [z_{i,s_i}, z_{i,s_i+1}).$$

The desired monotonicity in each dimension can be recovered by an alternating dimension-wise sorting of entries in $\boldsymbol{T}$ until convergence. The details are given in Algorithm 5.1.

An example when $d = 2$ is given below:

$$\begin{bmatrix} 0.7 & 0.4 & 0 \\ 0.3 & 0.5 & 0.2 \\ 1 & 0.8 & 0.6 \end{bmatrix} \xrightarrow{\text{sort rows}} \begin{bmatrix} 0 & 0.4 & 0.7 \\ 0.2 & 0.3 & 0.5 \\ 0.6 & 0.8 & 1 \end{bmatrix} \xrightarrow{\text{sort columns}} \begin{bmatrix} 0 & 0.3 & 0.5 \\ 0.2 & 0.4 & 0.7 \\ 0.6 & 0.8 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 0.7 & 0.4 & 0 \\ 0.3 & 0.5 & 0.2 \\ 1 & 0.8 & 0.6 \end{bmatrix} \xrightarrow{\text{sort columns}} \begin{bmatrix} 0.3 & 0.4 & 0 \\ 0.7 & 0.5 & 0.2 \\ 1 & 0.8 & 0.6 \end{bmatrix} \xrightarrow{\text{sort rows}} \begin{bmatrix} 0 & 0.3 & 0.4 \\ 0.2 & 0.5 & 0.7 \\ 0.6 & 0.8 & 1 \end{bmatrix}$$

As shown above, sorting ends up in some stationary point with desired monotonicity after a finite number of steps (see Theorem 5.2), but different orders of sorting may lead to different sorted CDF representations when $d \ge 2$. However, in our case, $\widetilde{F}_{\mathcal{S}}(\boldsymbol{x})$ is itself a perturbation of the CDF of Y, so the sorting procedure is often beneficial for stabilizing the algorithm. A more detailed empirical study on this is given in Section 6. The sorting procedure described converges (i.e., achieves monotonicity in the values of $\boldsymbol{T}$) in a finite number of iterations.

Algorithm 5.1 Alternating sorting.

Input: a tensor $\boldsymbol{T}$ that represents the estimated CDF $\widetilde{F}_{\mathcal{S}}(\boldsymbol{x})$
Output: a sorted tensor sorted($\boldsymbol{T}$) with desired monotonicity

Initialize sorted($\boldsymbol{T}$) as a all-zeros tensor with the same size as $\boldsymbol{T}$
while sorted($\boldsymbol{T}$)$\neq \boldsymbol{T}$ **do**
  sorted($\boldsymbol{T}$) $\leftarrow \boldsymbol{T}$
  for $i \in \{1:d\}$ **do**
    for $\boldsymbol{z} := (z_{1,s_1}, \ldots, z_{i-1,s_{i-1}}, z_{i+1,s_{i+1}}, \ldots, z_{d,s_d}) \in \otimes_{j\in\{1:d\}\setminus\{i\}}(z_{j,1}, \ldots, z_{j,M_i})$ **do**
      $\boldsymbol{T}[\boldsymbol{z},:] \leftarrow \text{sort}(\boldsymbol{T}[\boldsymbol{z},:])$, where $\boldsymbol{T}[\boldsymbol{z},:] := (\boldsymbol{T}_{z_{1,s_1},\ldots,z_{i,j},\ldots,z_{d,s_d}})_{j\in\{1:M_i\}}$
    end for
  end for
end while
Return sorted($\boldsymbol{T}$)

Theorem 5.2. *Assume that all the entries in $\boldsymbol{T}$ are distinct. The alternating sorting algorithm described in Algorithm 5.1 converges to a stationary point with desired monotonicity within a finite number of iterations.*

Proof. See Appendix A.1. ∎

5.3. Exploration sampling. We next describe the precise action taken when we decide to continue exploring. In particular, we need to define the function $Q(m, \widehat{m}_{\widehat{\mathcal{S}}^*})$ in Figure 1. When the current number m of exploration samples is smaller than the estimated optimal number of samples $\widehat{m}^*_{\widehat{\mathcal{S}}^*}$, the function Q determines how to increase m. A natural choice for Q is $Q(m, \widehat{m}^*_{\widehat{\mathcal{S}}^*}) = m+1$, i.e., simply increase by a single additional exploration sample. In practice, we observe that this behavior can be overly conservative and time-consuming when B is large. As an alternative, one could use a more aggressive strategy, say $Q(m, \widehat{m}^*_{\widehat{\mathcal{S}}^*}) = \frac{1}{2}\left(m + \widehat{m}^*_{\widehat{\mathcal{S}}^*}\right)$, which more quickly closes the gap between m and $\widehat{m}_{\widehat{\mathcal{S}}^*}$. However, there are situations when this is too aggressive. For example, if m is small (such as at initialization) then estimated quantities can be poor approximations, and in some cases $\widehat{m}_{\widehat{\mathcal{S}}^*}$ is significantly overestimated, and thus increasing m to $\frac{1}{2}\left(m + \widehat{m}^*_{\widehat{\mathcal{S}}^*}\right)$ can actually result in substantially overshooting the oracle value of $m^*_{\mathcal{S}^*}$. The probability of such an event is often positive and does not vanish as B increases.

As a compromise between these conservative and aggressive behaviors, we choose the following form:

$$Q(m, \widehat{m}^*_{\widehat{\mathcal{S}}^*}) = \begin{cases} 2m, & 1 \le m < \frac{\widehat{m}^*_{\widehat{\mathcal{S}}^*}}{2} \\ \frac{1}{2}\left(m + \widehat{m}^*_{\widehat{\mathcal{S}}^*}\right), & \frac{\widehat{m}^*_{\widehat{\mathcal{S}}^*}}{2} \le m < \widehat{m}^*_{\widehat{\mathcal{S}}^*} \end{cases}. \tag{5.12}$$

Since $\widehat{m}^*_{\widehat{\mathcal{S}}^*}$ is proportional to B, the above choice ensures that there is a sufficient amount of time for the algorithm to take exponential exploration whose growth manner is independent of the value of $\widehat{m}^*_{\widehat{\mathcal{S}}^*}$, which ensures both efficiency and accuracy of the algorithm. We note that neither the exponential rate two nor taking the average between m and $\widehat{m}^*_{\widehat{\mathcal{S}}^*}$ in (5.12) is

special, and can respectively be replaced with other rates greater than one and nonuniform averaging operations subject to appropriate modifications. Both the theoretical conclusions and numerical simulations in the subsequent sections assume that Q has the form above, but other reasonable choices for Q do not change the theoretical conclusions.

We have completed all technical descriptions of Figure 1. A more fleshed-out pseudocode version is given in Algorithm 5.2 that provides more details for every step. Next, we establish that the proposed algorithm enjoys optimality guarantees relative to model selection and budget allocation strategies produced by an oracle.

Algorithm 5.2 The detailed cvMDL algorithm.

Input: B: total budget, model costs $c_0, c_1, \ldots, c_n$
Output: an estimator for $F_Y(\boldsymbol{x})$

Initialize exploration = TRUE
Initialize $m = \sum_{i\in\{1:n\}} d_i + 2$
Generate m exploration samples of $(Y, X_{\{1:n\}})$
while exploration = TRUE **do**
  for $\mathcal{S} \subseteq \{1:n\}$ **do**
    Compute regression coefficients $\widehat{\boldsymbol{\beta}}^{(i)}_{\mathcal{S}^+}$, $i \in \{1:d\}$ from (5.1)
    Construct $\widehat{H}_{\mathcal{S}}(X_{\mathcal{S}})$ and $\widehat{h}(X_{\mathcal{S}}; \boldsymbol{x})$ from (5.2)
    Compute regression coefficients $r_j(\boldsymbol{x})$, $j \in \{1:m\}$ from (5.3)
    Construct $\mathcal{K}_1$ and $\mathcal{K}_2$ from (5.3) and (5.4), respectively
    Evaluate $\widehat{k}_1(\mathcal{S})$ and $\widehat{k}_2(\mathcal{S})$ using (5.5) and a quadrature rule on $\mathbb{R}^d$
    Compute $\widehat{m}^*_{\mathcal{S}}$ and $\widehat{L}_{\mathcal{S}}(\,\cdot\,; m)$ from (5.6)
    Compute the minimal expected loss $\widehat{L}_{\mathcal{S}}(m \vee \widehat{m}^*_{\mathcal{S}}; m)$ from (5.6)
  end for
  Choose $\widehat{\mathcal{S}}^* = \arg\min_{\mathcal{S}\subseteq\{1:n\}} \widehat{L}_{\mathcal{S}}(m \vee \widehat{m}^*_{\mathcal{S}}; m)$;
  if $m < \widehat{m}^*_{\widehat{\mathcal{S}}^*}$ **then**
    Generate $Q(m, \widehat{m}^*_{\widehat{\mathcal{S}}^*}) - m$ additional samples of $(Y, X_{\{1:n\}})$, where Q is given in (5.12)
    Increase m: $m \leftarrow Q(m, \widehat{m}^*_{\widehat{\mathcal{S}}^*})$
  else
    exploration = FALSE
  end if
end while
Generate $N_{\widehat{\mathcal{S}}^*}$ samples of $X_{\widehat{\mathcal{S}}^*}$, with $N_{\mathcal{S}}$ given in (3.3)
Construct $\widehat{\alpha}(\boldsymbol{x})$ for $\mathcal{S} \leftarrow \widehat{\mathcal{S}}^*$ using (5.8)
Generate $\widehat{\mathcal{S}}^*$ exploitation estimator $\widetilde{F}_{\widehat{\mathcal{S}}^*}$ using (5.10).

5.4. Model consistency and optimality. We now provide theoretical guarantees for Algorithm 5.2 (corresponding to the flowchart in Figure 1). In summary, we show that as the budget B tends to infinity, the model subset $\widehat{\mathcal{S}}^*$ chosen along with the number of exploration samples m taken in Algorithm 5.2, both converge to the oracle optimal model $\mathcal{S}^*$ and the

optimal number of exploration samples $m^*_{\mathcal{S}^*}$, respectively.

We need some technical assumptions in order to proceed with our results. Since we estimate quadratic moments, we require quadratic moments to exist. We also require that there are no pairs of low-fidelity QoIs that are perfectly correlated. These are codified in the following two assumptions.

Assumption 5.3. *The models $X_{\{1:n\}}$ and Y have bounded second moments:*

$$\mathbb{E}[\|X_{\{1:n\}}\|_2^2 + \mathbb{E}\|Y\|_2^2] < \infty. \tag{5.13}$$

Assumption 5.4. *The uncentered second moment matrix $\mathbb{E}[X_{\{1:n\}+}X_{\{1:n\}+}^\top]$ is invertible, where $X_{\{1:n\}+} = (1, X_{\{1:n\}}^\top)^\top$.*

Assumption 5.3 is the minimal moment condition on QoIs that we require to make oracle quantities well-defined. Random variables that violate Assumption 5.4 exhibit perfect multicollinearity and in practice are relatively rare. Assumption 5.4 being violated does not cause any conceptual breakdown of our procedure; the only consequence is that all the linear regression procedures suffer from a lack of identifiability of optimal covariates. While there are numerous standard procedures to remedy multicollinearity, such as covariate removal or regularization, violation of this assumption did not surface in our experiments, so we do not utilize any of these remedies.

The model selection procedure requires estimating the average ω-weighted L^2 norm. This requires us to make certain assumptions about ω.

Assumption 5.5. *The weight $\omega(\boldsymbol{x})$ is chosen so that either of the following conditions is true:*

(a) $\|\omega\|_{L^\infty(\mathbb{R}^d)} < \infty$ (e.g. $\omega(\boldsymbol{x}) \equiv 1$) and $d = 1$; or

(b) $\|\omega\|_{L^1(\mathbb{R}^d)} < \infty$.

The final more technical assumption we require involves some regularity on distribution functions. In particular, we show pointwise convergence in $\boldsymbol{x}$ of the estimator $\widehat{\alpha}(\boldsymbol{x})$ to the oracle parameter $\alpha(\boldsymbol{x})$, and to accomplish this we require bounds on the local variations of $F_{H_\mathcal{S}}$ and $F_{H_\mathcal{S}\vee Y}$ constructed in the model selection procedure. More technically, a sufficient assumption is a bounded local variations condition involving CDFs of certain d-dimensional sketches of $X_{\{1:n\}}$ and Y.

Assumption 5.6. *Define*

$$V(\boldsymbol{A}) := (X_{\mathcal{S}^+}^\top \boldsymbol{A})^\top \in \mathbb{R}^d \qquad \boldsymbol{A} := [\boldsymbol{A}^{(1)}, \cdots, \boldsymbol{A}^{(d)}] \in \mathbb{R}^{d_\mathcal{S} \times d},$$

and recall the optimal coefficient matrix $\boldsymbol{B}_{\mathcal{S}^+}$ in (4.1). We assume the CDFs of $V(\boldsymbol{A})$ and $V(\boldsymbol{A}) \vee Y$, denoted by $F_{V(\boldsymbol{A})}$ and $F_{V(\boldsymbol{A})\vee Y}$, are globally Lipschitz near $\boldsymbol{B}_{\mathcal{S}^+}$ for all $\mathcal{S}$. In particular, we assume that there exists $\varepsilon > 0$ such that

$$\max_{\mathcal{S}\subseteq\{1:n\}} \sup_{\boldsymbol{A}:\|\boldsymbol{A}-\boldsymbol{B}_{\mathcal{S}^+}\|_F \le \varepsilon} \left\{\|F_{V(\boldsymbol{A})}\|_{Lip} + \|F_{V(\boldsymbol{A})\vee Y}\|_{Lip}\right\} = C < \infty,$$

where $\|\cdot\|_{Lip}$ is the Lipschitz constant defined as

$$\|f\|_{Lip} = \sup_{\boldsymbol{x}\neq\boldsymbol{x}'} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{x}')|}{\|\boldsymbol{x} - \boldsymbol{x}'\|_2} \qquad f : \mathbb{R}^d \to \mathbb{R}.$$

This final assumption is less transparent than our previous ones. An unnecessarily stronger sufficient condition to ensure that Assumption 5.6 holds is to assume that both Y and all unit linear combinations of $X_{\{1:n\}}$ have uniformly bounded densities, and that every high-fidelity covariate $Y^{(i)}$ is correlated with every low-fidelity covariate $X_j^{(r)}$, i.e., $\min_{i,j,r} |\mathrm{Corr}(Y^{(i)}, X_j^{(r)})| > 0$. Alternatively, one could assume that the same bounded density condition, and the rather reasonable condition that the oracle regression coefficients $\boldsymbol{B}_{\mathcal{S}^+}$ select at *least* one non-deterministic covariate for every $\mathcal{S}$.

We can now present our main results regarding applying the cvMDL algorithm with $h(X_{\mathcal{S}}; \boldsymbol{x})$ constructed using linear approximations, with the corresponding loss function parameters estimated from (5.5) and (5.6). In particular, we have that the adaptive exploration rate $m(B)$ asymptotically matches the optimal (oracle) exploration rate $m^*_{\mathcal{S}^*}$ defined in Section 3.3, and the selected model $\mathcal{S}(B)$ converges to the optimal (oracle) model $\mathcal{S}^*$ as $B \to \infty$:

Theorem 5.7 (Uniform consistency and asymptotic optimality of cvMDL in Algorithm 5.2). *Let $h(X_{\mathcal{S}}; \boldsymbol{x})$ be defined in (4.2), i.e., we use the linear approximation estimators from Section 4, and assume the model parameters are estimated via (5.5) and (5.6). Then consider Algorithm 5.2 with an input budget B, and let*

- *$m(B) = \widehat{m}_{\widehat{\mathcal{S}}^*}$ be the number of exploration samples chosen by Algorithm 5.2,*
- *$\mathcal{S}(B) = \widehat{\mathcal{S}}^*$ be the model selected for exploitation,*
- *$\widetilde{F}(\boldsymbol{x}; B) = \widetilde{F}_{\widehat{\mathcal{S}}^*}(\boldsymbol{x})$ be the CDF estimator for F_Y.*

Under Assumptions 5.3, 5.4, 5.5, and 5.6, then with probability one,

$$\lim_{B\to\infty} \frac{m(B)}{m^*_{\mathcal{S}^*}} = 1, \tag{5.14a}$$

$$\lim_{B\to\infty} \mathcal{S}(B) = \mathcal{S}^*, \tag{5.14b}$$

$$\lim_{B\to\infty} \sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{F}(\boldsymbol{x}; B) - F_Y(\boldsymbol{x})| = 0, \tag{5.14c}$$

where $\mathcal{S}^*$ and $m^*_{\mathcal{S}^*}$ are the unique optimal (oracle) model choice and exploration sample size defined in Section 3.3.

The proof is given in Appendix A.4. The result (5.14c) should not come as a surprise since uniform consistency is generally true for empirical CDF estimators. Therefore, while (5.14a) and (5.14b) show that cvMDL in Algorithm 5.2 exhibits optimality (relative to an oracle) for the choice of exploration samples and sample allocation across models, (5.14c) is not evidence that the multifidelity estimator $\widetilde{F}(\boldsymbol{x}; B)$ is superior to the empirical CDF estimator that uses only the high-fidelity samples, although it confirms that $\widetilde{F}(\boldsymbol{x}; B)$ behaves as expected. The major difference that distinguishes $\widetilde{F}(\boldsymbol{x}; B)$ from a standard empirical CDF estimator is a constant term resulting from the mean ω-weighted L^2 convergence rate; see the discussion near the end of Subsection 3.3.

The statements in Theorem 5.7 and [36, Theorem 5.2] are similar, but in the former, the requisite assumptions are *much* weaker and the guarantees are stronger. In fact, for [36, Theorem 5.2] to hold, one must assume that $\mathbb{E}[Y|X_{\mathcal{S}}]$ is a linear function of $X_{\mathcal{S}}$ and $(Y - \mathbb{E}[Y|X_{\mathcal{S}}]) \perp\!\!\!\perp X_{\mathcal{S}}$ for all $\mathcal{S} \subseteq \{1:n\}$. However, none of these assumptions is needed in Theorem 5.7. Additionally, Theorem 5.7 ensures convergence for a *multivariate* distribution

function instead of the univariate convergence statements in [36, Theorem 5.2].

5.5. A brief view into proving Theorem 5.7. While we leave the technical parts of proving Theorem 5.7 to Appendix A.4, we can summarize the crucial intermediate results that allow the proof to succeed. The major results we need revolve around the consistency of various estimators as m and/or $N_{\mathcal{S}}$ approach infinity. The following two sets of results leverage the assumptions to conclude the consistency of intermediate computations in the algorithm.

The first collection of results shows that the finite-sample estimators for quantities computed in the exploration phase are consistent as the number of exploration samples m tends to infinity.

Lemma 5.8 (Asymptotic consistency of exploration estimators). *We have the following technical estimates and consistency results for all $\mathcal{S} \subseteq \{1:n\}$:*

(i) Under Assumptions 5.3 and 5.6, then with probability one,

$$\sup_{\|\boldsymbol{A}-\boldsymbol{B}_{\mathcal{S}^+}\|_F<\varepsilon} \sup_{\boldsymbol{x}\in\mathbb{R}^d} |F_{V(\boldsymbol{A})}(\boldsymbol{x}) - F_{V(\boldsymbol{B}_{\mathcal{S}^+})}(\boldsymbol{x})| \lesssim \|\boldsymbol{A}-\boldsymbol{B}_{\mathcal{S}^+}\|_F^{2/3} \lesssim \varepsilon^{2/3} \tag{5.15a}$$

$$\sup_{\|\boldsymbol{A}-\boldsymbol{B}_{\mathcal{S}^+}\|_F<\varepsilon} \sup_{\boldsymbol{x}\in\mathbb{R}^d} |F_{V(\boldsymbol{A})\vee Y}(\boldsymbol{x}) - F_{V(\boldsymbol{B}_{\mathcal{S}^+})\vee Y}(\boldsymbol{x})| \lesssim \|\boldsymbol{A}-\boldsymbol{B}_{\mathcal{S}^+}\|_F^{2/3} \lesssim \varepsilon^{2/3}. \tag{5.15b}$$

(ii) Under Assumptions 5.3 and 5.4, then with probability one,

$$\lim_{m\to\infty} \widehat{\boldsymbol{B}}_{\mathcal{S}^+} = \boldsymbol{B}_{\mathcal{S}^+}.$$

(iii) Under Assumptions 5.3, 5.4, and 5.6, then with probability one,

$$\lim_{m\to\infty} \sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) - F_{H_{\mathcal{S}}}(\boldsymbol{x})| = 0 \qquad \lim_{m\to\infty} \sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widehat{F}_{Y\vee\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) - F_{Y\vee H_{\mathcal{S}}}(\boldsymbol{x})| = 0.$$

(iv) Under Assumptions 5.3, 5.4, and 5.6, then almost surely as $m\to\infty$ we have that,

$$\mathcal{K}_1(\boldsymbol{x}) \to (1-\rho_{\mathcal{S}}^2(\boldsymbol{x}))F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x})) \quad \mathcal{K}_2(\boldsymbol{x}) \to \rho_{\mathcal{S}}^2(\boldsymbol{x})F_Y(\boldsymbol{x})(1-F_Y(\boldsymbol{x}))$$

for all $\boldsymbol{x}\in\mathbb{R}^d$.

(v) Under Assumptions 5.3, 5.4, 5.5, and 5.6, then with probability one, $\lim_{m\to\infty}\widehat{k}_1(\mathcal{S}) = k_1(\mathcal{S})$ and $\lim_{m\to\infty}\widehat{k}_2(\mathcal{S}) = k_2(\mathcal{S})$.

(vi) Under Assumptions 5.3, 5.4, and 5.6, for $\boldsymbol{x}\in(supp(F_{H_{\mathcal{S}}}))^\circ$, $\widehat{\alpha}(\boldsymbol{x})$ is a consistent estimator of $\alpha(\boldsymbol{x})$ almost surely, i.e., $\lim_{m\to\infty}\widehat{\alpha}(\boldsymbol{x}) = \alpha(\boldsymbol{x})$ for every $\boldsymbol{x}\in(supp(F_{H_{\mathcal{S}}}))^\circ$.

The proof is given in Appendix A.2.

Remark 5.9. *Note that $\widehat{\alpha}(\boldsymbol{x})$ may not be consistent outside $(supp(F_{H_{\mathcal{S}}}))^\circ$, where the value of $\alpha(\boldsymbol{x})$ is set to be zero in the definition for convenience; see (4.3). However, this has no impact on the accuracy of the exploitation estimator as $\mathbf{1}_{\{H_{\mathcal{S}}\le\boldsymbol{x}\}}$ is constant in this region.*

Our second intermediate result shows that the exploitation estimator for the CDF of Y is consistent asymptotically in both the exploration sample count m and the exploitation sample count $N_{\mathcal{S}}$.

Lemma 5.10 (Uniform asymptotic consistency of the exploitation CDF estimator). *Under Assumptions 5.3, 5.4, and 5.6, then with probability one,* $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{F}_{\mathcal{S}}(\boldsymbol{x}) - F_Y(\boldsymbol{x})| \to 0$ *as* $m, N_{\mathcal{S}} \to \infty$.

See Appendix A.3 for the proof. The proof of our main result, Theorem 5.7, is in Appendix A.4, which leverages the results in Lemma 5.8 and Lemma 5.10. One additional high-level step needed to prove Theorem 5.7 is to show that cvMDL in Algorithm 5.2 for asymptotically large budget B results in both m and $N_{\mathcal{S}}$ going to infinity. This is the first part of the proof presented in Appendix A.4.

6. Numerical simulations.

In this section, we compare cvMDL and its variants with other algorithms on three forward uncertainty quantification scenarios. In Section 6.1, we examine a scalar-valued parametric linear elasticity PDE problem, followed by a vector-valued stochastic differential equation problem concerning the extrema of a geometric Brownian motion over a finite interval in Section 6.2. Lastly, in Section 6.3 we evaluate the cvMDL method on a a scalar-valued practical engineering problem of brittle fracture in a fiber-reinforced matrix. We label algorithms under consideration as follows:

- (ECDF) The empirical CDF estimator for F_Y using the high-fidelity samples only;
- (AETC-d) The AETC-d algorithm from [36];
- (cvMDL) The cvMDL algorithm in Algorithm 5.2;
- (cvMDL-sorted) cvMDL with the exploitation CDF monotonicity fix using Algorithm 5.1;
- (cvMDL*) cvMDL that estimates $\widehat{\alpha}(\boldsymbol{x})$ in the tail using (5.11) with $\tau = 0.05$ when $d = 1$;
- (cvMDL*-sorted) cvMDL* with the CDF monotonicity fix using Algorithm 5.1.

For the weight function in the cvMDL algorithm and its variants, we choose $\omega(x) \equiv 1$ for all $x \in \mathbb{R}$ when Y is scalar-valued, but in a case-dependent manner when Y is vector-valued. Since the estimators produced by the cvMDL-type and AETC-d algorithms are random (depending on the exploration data), for every budget value B, we repeat the experiment 100 times and report both the average of the mean ω-weighted L^2 error and the corresponding 5%-95% quantiles to measure the uncertainty.

6.1. Linear elasticity.

We consider a suite of models with varying fidelities associated with a parametric elliptic equation, where lower-fidelity models are identified through mesh coarsening. The setup is taken from [35, Section 7.1]. The elliptic PDE governs displacement in linear elasticity over a square spatial domain $\mathcal{D} = [0,1]^2$, with $\boldsymbol{x} = (x_1, x_2)^\top$; see Figure 2. The parametric version of this problem equation seeks the displacement field $\boldsymbol{u} = (u, v)^\top$ that satisfies the PDE

$$-\nabla \cdot (\kappa(\boldsymbol{p}, \boldsymbol{x})\, \boldsymbol{\sigma}(\boldsymbol{x})) = \boldsymbol{F}(\boldsymbol{x}), \quad \forall (\boldsymbol{p}, \boldsymbol{x}) \in \mathcal{P} \times \mathcal{D}$$

where $\boldsymbol{p}$ is a random input vector that parameterizes the scalar κ. Moreover, $\boldsymbol{\sigma}$ is the Cauchy stress tensor, given as

$$\boldsymbol{\sigma}(\boldsymbol{x}) = \begin{bmatrix} \sigma_1(\boldsymbol{x}) & \sigma_{12}(\boldsymbol{x}) \\ \sigma_{12}(\boldsymbol{x}) & \sigma_2(\boldsymbol{x}) \end{bmatrix}, \qquad \begin{bmatrix} \sigma_1(\boldsymbol{x}) \\ \sigma_2(\boldsymbol{x}) \\ \sigma_{12}(\boldsymbol{x}) \end{bmatrix} = \frac{1}{1-\nu^2} \begin{bmatrix} \frac{\partial u(\boldsymbol{x})}{\partial x_1} + \frac{\partial v(\boldsymbol{x})}{\partial x_2} \\ \frac{\partial v(\boldsymbol{x})}{\partial x_2} + \nu \frac{\partial u(\boldsymbol{x})}{\partial x_1} \\ \frac{1-\nu}{2}\left(\frac{\partial u(\boldsymbol{x})}{\partial x_1} + \frac{\partial v(\boldsymbol{x})}{\partial x_2}\right) \end{bmatrix}$$

where we set the Poisson ratio to $\nu = 0.3$. Here, $\kappa(\boldsymbol{p}, \boldsymbol{x})$ is a scalar modeled with a Karhunen-Loève expansion with four modes, given by

$$\kappa(\boldsymbol{p}, \boldsymbol{x}) = 1 + 0.5 \sum_{i=1}^{4} \sqrt{\lambda_i} \phi_i(\boldsymbol{x}) p_i,$$

where (λ_i, ϕ_i) are ordered eigenpairs of the exponential covariance kernel K on $\mathcal{D}$, i.e.,

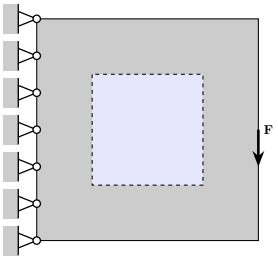
$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|_1 / \eta),$$

where $\|\cdot\|_1$ is the ℓ^1-norm on vectors, and we choose $\eta = 0.7$. Hence, $\boldsymbol{p} \in \mathbb{R}^4$ is a random input vector with independent components uniformly distributed on $[-1, 1]$.

The displacement $\boldsymbol{u}$ is used to compute a scalar QoI, the structural *compliance* or energy norm of the solution, which is the measure of elastic energy absorbed in the structure as a result of loading:

$$E(\boldsymbol{u}; \boldsymbol{p}) \coloneqq \int_D (\boldsymbol{u}(\boldsymbol{x}) \cdot \boldsymbol{F}(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}.$$

We solve the above system for each fixed $\boldsymbol{p}$ via the finite element method with standard bilinear square isotropic finite elements on a rectangular mesh [2].



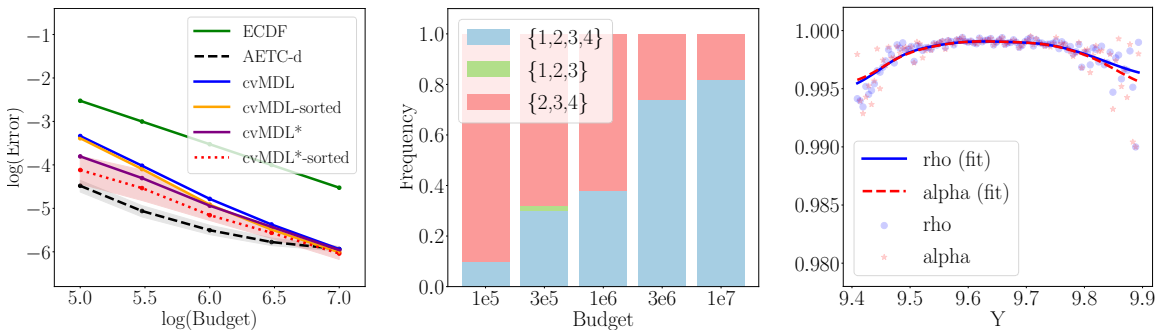
Model $\mathcal{S}$	$\{1\}$	$\{2\}$	$\{3\}$	$\{4\}$	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$
$\gamma_{\mathcal{S}}$	123.8	149.3	203.9	304.8	25.2	48.6	93.7	62.2
$m^*_{\mathcal{S}}$ when $B = 10^7$	1998	2231	2337	2390	1253	1657	1909	2054

Model $\mathcal{S}$	$\{2,4\}$	$\{3,4\}$	$\{1,2,3\}$	$\{1,2,4\}$	$\{1,3,4\}$	$\{2,3,4\}$	**{1,2,3,4}**
$\gamma_{\mathcal{S}}$	107.6	129.7	11.8	11.7	14.3	11.9	**11.5**
$m^*_{\mathcal{S}}$ when $B = 10^7$	2175	2292	669	734	976	1540	**638**

Figure 2: Linear elasticity. Left: Geometry and boundary conditions for the linear elastic structure. Right: Oracle scaled loss $\gamma_{\mathcal{S}}$ (3.11) and optimal exploration sample count $m^*_{\mathcal{S}}$ (3.11) for different choices of subsets of low-fidelity model indices $\mathcal{S}$. The optimal model subset $\mathcal{S}$ is typed in boldface. Oracle statistics are computed with 50,000 samples.

In this example, we form a multifidelity hierarchy through mesh coarsening via the mesh parameter h. The compliance E is our scalar-valued QoI for every model, i.e., $d = d_i = 1$ for $i = 1, \ldots, 4$. A mesh parameter of $h = 2^{-7}$ corresponds to the high-fidelity model Y, and $h = 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}$ correspond to lower-fidelity models $X_1, \ldots, X_4$, respectively.

The cost for each model is the computational time, which we take to be inversely proportional to the mesh size squared, i.e., h^2. This corresponds to using a linear solver of optimal linear complexity. We normalize cost so that the model with the lowest fidelity has unit cost, i.e., $(c_0, c_1, c_2, c_3, c_4) = (4096, 64, 16, 4, 1)$. The correlations between the QoIs of Y and X_1, X_2, X_3, X_4 are 0.976, 0.940, 0.841, -0.146, respectively. The total budget B is taken on the interval $[10^5, 10^7]$.



(a) (b) (c)

Figure 3: Linear elasticity. (a). Mean $\omega(x)$-weighted L^2 error between F_Y and the estimated CDFs given by ECDF, AETC-d, cvMDL, and its variants, with the 5%-95% quantiles (for ease of visualization, we only plot the quantiles for AETC-d and cvMDL*-sorted) to measure the uncertainty. (b). Frequency of different models selected by cvMDL. (c). Scatter plot of the estimated $\alpha(x)$ and $\rho(x)$ when $\mathcal{S} = \{1, 2, 3, 4\}$ using 50,000 i.i.d. samples in the 1%-99%-quantile regime of Y. Gaussian kernel smoothing is applied to both data Gaussian kernel with bandwidth bd = 0.0358 chosen using 5-fold cross-validation.

6.1.1. Results for estimating the CDF. Figure 3 shows the simulation results given by different multifidelity estimators as well as more refined statistics of the cvMDL-related algorithms. In Figure 3(a), we see that AETC-d has the smallest error for smaller budgets but its asymptotic convergence is constrained by the model misspecification effects (associated with theoretical assumptions on the applicability of AETC-d), i.e., the error curve starts to plateau when B exceeds 10^6. Although this can be mitigated by including additional nonlinear (e.g. polynomial) terms as additional covariates, trustworthy practical guidance is still lacking for this approach. On the other hand, both cvMDL and its variants demonstrate superior performance over ECDF, with cvMDL*-sorted achieving a result competitive to AETC-d without the plateau effect.

In Figure 3(b), we note that as the budget increases, the model $\widehat{\mathcal{S}}^*$ selected by cvMDL converges to $\{1, 2, 3, 4\}$, which is the same as the optimal model computed under oracle statistics in Figure 2 (right). We note that the suboptimal model $\mathcal{S} = \{2, 3, 4\}$ is selected often by cvMDL, but not other models whose $\gamma_{\mathcal{S}}$ is close to that of $\{1, 2, 3, 4\}$. We believe this occurrence is due to the aggressive exploration steps taken by cvMDL, in particular when we double exploration samples causing suboptimal models $\mathcal{S}$ with large values of $m^*_{\mathcal{S}}$ (e.g., $\mathcal{S} = \{2, 3, 4\}$) become the preferred model.

The significant error reduction achieved by cvMDL is indicated by near-unity values of $\rho_{\mathcal{S}}(x) = \mathrm{Corr}[\mathbf{1}_{\{Y \le x\}}, h(X_{\mathcal{S}}; x)]$ where $\mathcal{S} = \{1, 2, 3, 4\}$, shown in Figure 3(c). For cvMDL variants, either estimating $\alpha(x)$ in the tail regime through (5.8) (cvMDL*) or sorting CDF values to ensure monotonicity (cvMDL/cvMDL*-sorted) can help further reduce the errors. The former is particularly helpful in the small-budget regime where exploration data are not

sufficient to estimate the full support of the QoI.

The weight function $\omega(x)$ in this scenario is constant on $\mathbb{R}$ thus the estimates produced by cvMDL-type estimators are expected to capture the global structure of F_Y (e.g. bulk and tails). To inspect this, we fix $B = 10^7$ and plot the estimated CDFs in the tail and bulk regimes separately. The CDFs of the pointwise errors (at 1000 discretization points) in the three regimes are shown in Figure 4. It can be seen that cvMDL*-sorted has the smallest errors in all three regimes.

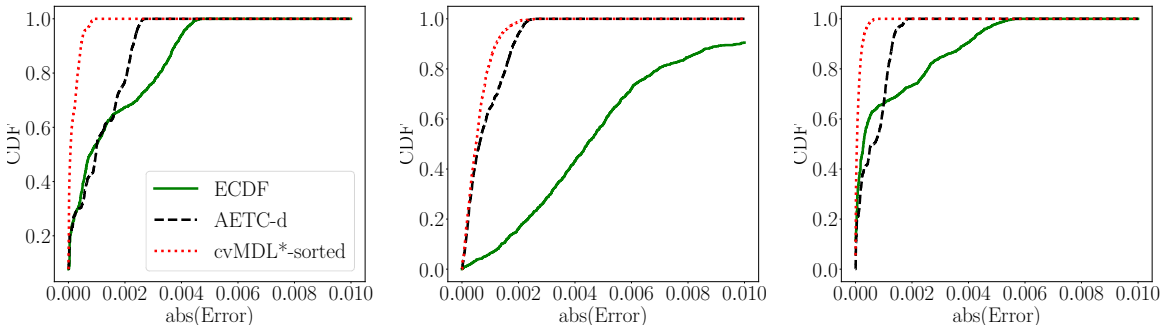


Figure 4: Linear elasticity. One realization of the distribution of pointwise CDF errors computed by cvMDL*-sorted, AETC-d, and ECDF for budget $B = 10^7$. We plot CDFs of errors in three different regimes: (a) the lower tail of Y defined by the $0 - 0.05$ quantile region, (b) the bulk defined by the $0.05 - 0.95$ quantile region, (c) the upper tail defined by the $0.95 - 1.00$ quantile region.

6.1.2. Risk metrics. We now compare some risk metrics of the estimated CDFs. For example, one frequently used metric is the conditional value-at-risk (CVaR), also called the expected shortfall, which is defined as the conditional expectation of Y in a tail regime (here, Y being large):

$$\mathsf{CVaR}_a(Y) := \mathbb{E}[Y|F_Y(Y) \geq a] = \frac{1}{1-a}\int_a^1 F_Y^{-1}(x)\mathrm{d}x \qquad 0 < a < 1,$$

where a is the quantile level. Assuming F_Y is known, CVaR can be numerically computed using linear interpolation of F_Y^{-1}. Fixing $B = 10^7$ as before, we use the estimated CDFs by ECDF, AETC-d, and cvMDL*-sorted to compute the CVaR of Y for $a = 0.95$ and 0.99, respectively. The experiment is repeated 50 times, and the corresponding statistics are summarized using boxplots in Figure 5 (a)-(b). For both choices of a, cvMDL*-sorted outperforms the other methods by a noticeable margin. It is worth noting that although AETC-d and cvMDL*-sorted have similar errors under the tested budget globally (Figure 3 (a)), the model misspecification effects result in the former estimates being biased upward. The cvMDL*-sorted estimates, on the other hand, remain unbiased.

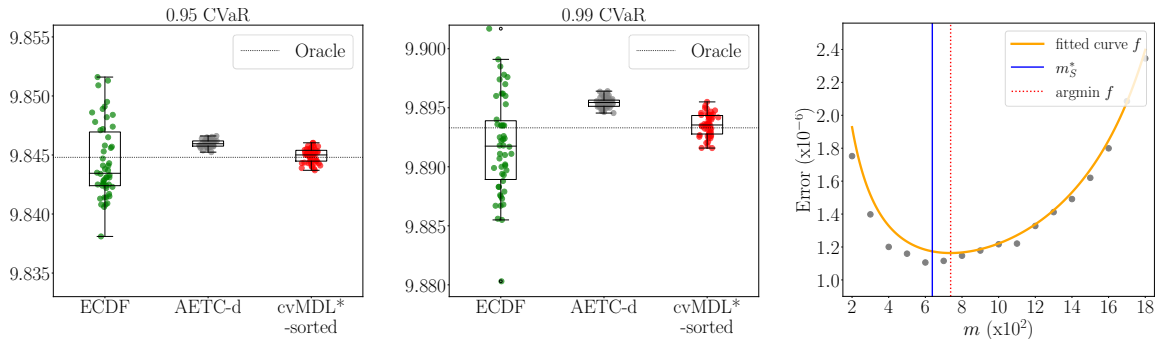


Figure 5: Linear elasticity. (a): Boxplots of the $\mathsf{CVaR}_{0.95}(Y)$ computed using the estimated CDFs given by ECDF, AETC-d, and cvMDL*-sorted when $B = 10^7$ with 50 experiments. (b): Same but for $\mathsf{CVaR}_{0.99}(Y)$. (c): Inspection of how well the estimated loss $\widehat{L}_{\mathcal{S}}$ mimics the oracle loss curve as a function of m. The discrete data are fitted using a function $f(m; a, b)$ of the form $\frac{a}{m} + \frac{b}{B - c_{\text{epr}} m}$, with fitted values for a and b given by 2.64×10^{-4} and 5.57, respectively.

6.1.3. Oracle versus estimated loss. We investigate the model selection criteria used in cvMDL. Note for each $\mathcal{S} \subseteq \{1 : n\}$, there is a discrepancy between the exact loss function versus the estimator $\widehat{L}_{\mathcal{S}}$ constructed with empirical data. We inspect if this approximation is reasonable. To numerically determine if the exploration-exploitation trade-off is optimal, we fix $B = 10^7$ and $\mathcal{S} = \{1, 2, 3, 4\}$. For a given value of m, we first take m exploration samples to estimate the control variates parameters and then use them to build an estimator $\widehat{F}_{\mathcal{S}}$ for F_Y as in (5.10). We then compute the (exact) mean weighted L^2 loss associated with this value of m. We repeat the experiment 10 times and compute the average loss value. We compile results of the above for m in the range from 200 to 1800. The results are reported in Figure 5 (c). It can be seen that the optimal exploration rate under oracle loss $L_{\mathcal{S}}$, 638 (see Figure 2, right), almost matches the empirically identified optimal exploration rate, which is 736. The small gap can be attributed to the underestimation of exploration error committed due to the finite-sample estimation of parameters.

6.2. Extrema of Geometric Brownian Motion. Geometric Brownian motion is a continuous-time stochastic process that is widely used in financial modeling. In a simple setting, a geometric Brownian motion S_t is a random process with a constant initial state $s_0 > 0$ whose evolution is described using the stochastic differential equation

$$dS_t = \mu S_t dt + \sigma S_t dW_t \qquad t \geq 0, \qquad S_0 = s_0,$$

where both μ and $\sigma > 0$ are constants, and W_t is a standard Brownian motion process. A unique explicit solution for S_t exists and can be written as

$$S_t = s_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t\right).$$

Set $\mu = 0.05, \sigma = 0.2, s_0 = 1$. We are interested in estimating the joint distribution of the extreme values of S_t over the time interval $[0, 1]$:

$$(S_{\min}, S_{\max})^\top \in \mathbb{R}^2 \qquad S_{\min} := \min_{0\le t\le 1} S_t, \quad S_{\max} := \max_{0\le t\le 1} S_t.$$

We thus choose as the QoI the random vector $(S_{\min}, S_{\max})^\top$. We evaluate these quantities by discretizing the stochastic differential equation in time using the Euler–Maryuama scheme with time step Δt and computing the discrete extrema. The computational complexity (cost) of the corresponding procedure is proportional to the number of grid points used for discretization.

We construct a multifidelity model for this problem based on time discretization. In particular, we consider four different time scales $\Delta t \in \{2^{-4}, 2^{-6}, 2^{-8}, 2^{-14}\}$, with the high-fidelity model Y corresponding to $\Delta t = 2^{-14}$ and X_1, X_2, X_3 corresponding to $\Delta t = 2^{-8}, 2^{-6}, 2^{-4}$, respectively. This results in (normalized) model costs $(c_0, c_1, c_2, c_3) = (1024, 16, 4, 1)$. The total budget B takes values in $[10^4, 10^6]$. To generate joint samples, the randomness of W_t is simulated from the same realization used in the high-fidelity model. The oracle CDF of the high-fidelity model is computed using MC with 10^5 samples. The oracle correlations between the QoIs of the high- and low-fidelity models in Table 1.

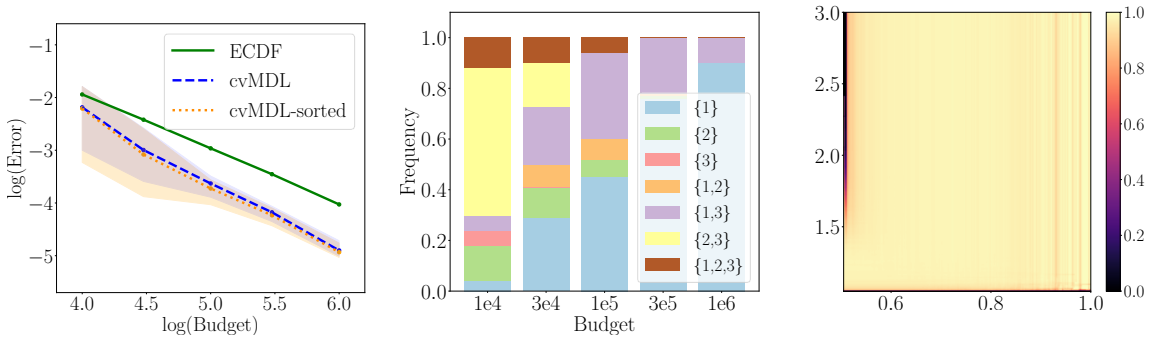
Table 1: Geometric Brownian motion. Oracle correlations between the high-fidelity and low-fidelity model QoIs computed using 50,000 samples.

Model QoIs	$S_{\min}(2^{-8})$	$S_{\max}(2^{-8})$	$S_{\min}(2^{-6})$	$S_{\max}(2^{-6})$	$S_{\min}(2^{-4})$	$S_{\max}(2^{-4})$
$S_{\min}(2^{-14})$	0.999	0.682	0.997	0.682	0.984	0.680
$S_{\max}(2^{-14})$	0.681	0.999	0.681	0.998	0.674	0.988

In this example, all model QoIs are two-dimensional random vectors so AETC-d cannot be directly applied. For cvMDL and its variants, setting $\omega(\boldsymbol{x}) \equiv 1$ violates Assumption 5.5. Instead, since $S_{\min} \le s_0 = 1 \le S_{\max}$, we choose $\omega(\boldsymbol{x}) = \mathbf{1}_{\mathcal{T}}(\boldsymbol{x})$ as an indicator function on a two-dimensional bounded region $\mathcal{T} \subset \mathbb{R}^2$ where the most likely outcomes reside. For instance, here we take $\mathcal{T} = [0.5, 1] \times [1, 3]$. The statistics of the estimation errors and the selected models by cvMDL are reported in Figure 6(a),(b). Panel (c) shows that the correlation coefficient $\rho_{\mathcal{S}}(\boldsymbol{x})$, is close to unity over the entire domain, suggesting that our chosen control variate (3.6) is a good choice.

Figure 6 shows that cvMDL is consistent on the region $\mathcal{T}$, and the corresponding estimation error is on average much lower than that of ECDF. As the budget goes to infinity, the model selected by cvMDL converges to the single low-fidelity model $\{1\}$, which coincides with the optimal model computed using oracle statistics. With additional sorting to stabilize the algorithm, cvMDL-sorted slightly further reduces the errors of cvMDL, which is consistent with the observations in the 1d case. In the pre-asymptotic regime when the budget is small, the models selected by cvMDL have relatively large fluctuations, but these stabilize for larger budgets. More results from this experiment are presented in Appendix B.1.

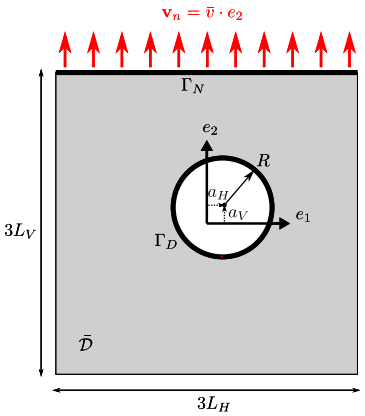
6.3. Brittle fracture behavior of a fiber-reinforced matrix. We investigate a two-dimensional fiber-reinforced matrix, a subject commonly explored in the field of fracture mechanics. Our QoI is the maximum load that induces brittle fracture within the matrix region adjacent to the



(a) (b) (c)

Figure 6: Geometric Brownian motion. (a): Mean $\omega(\boldsymbol{x})$-weighted L^2 error between F_Y and the estimated CDFs given by ECDF, cvMDL, and cvMDL-sorted, with the 5%-50%-95% quantiles to measure the uncertainty. (b): Frequency of different models selected by cvMDL; cf. optimal model losses in Figure 11. (c) Estimated $\rho_{\mathcal{S}}(\boldsymbol{x})$ from (3.6) when $\mathcal{S} = \{1\}$ using 50,000 i.i.d. samples for $\boldsymbol{x} \in \mathcal{T}$.

fiber. To obtain the QoI, we solve a quasi-static, two-dimensional finite element problem. Fig-

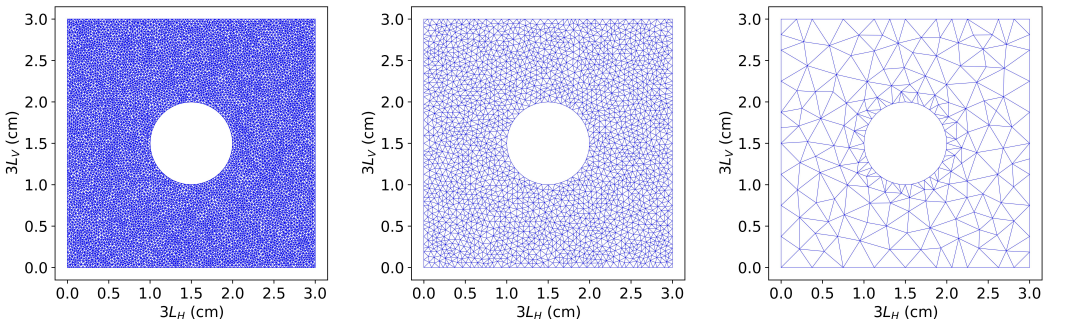


Random variable	Property	Mean	COV$_i$ (%)[a]	Lower boundary	Upper boundary	Probability distribution
p_1	κ (GPa)	55	5	45	60	Truncated normal
p_2	ν	0.25	5	0.20	0.30	Truncated normal
p_3	G_c (GPa)	1	10	0	∞	Lognormal
p_4	L_H(cm)	1	–	0.9	1.1	Uniform
p_5	L_V (cm)	1	–	0.9	1.1	Uniform
p_6	R (cm)	0.5	–	0.4	0.6	Uniform
p_7	a_H (cm)	0	–	−0.05	0.05	Uniform
p_8	a_V (cm)	0	–	−0.05	0.05	Uniform

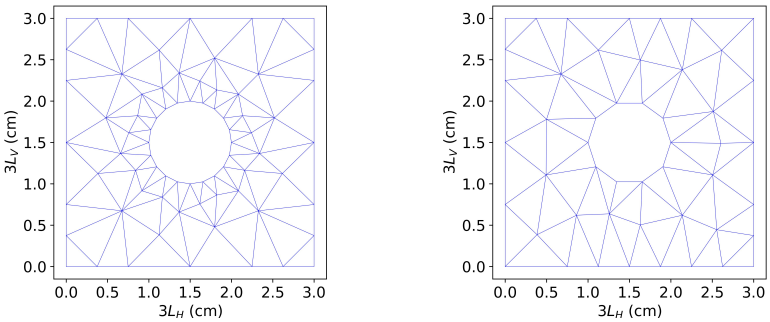
a. Coefficient of variation COV$_i = 100 \times \sigma_i / \mathbb{E}[p_i]$ for $i = 1, \ldots, 8$.

Figure 7: Fiber-reinforced matrix. Left: Geometry, loading, and boundary condition. We consider the domain $\mathcal{D} = [-1.5L_H, 1.5L_H] \times [-1.5L_V, 1.5L_V] \in \mathbb{R}^2$ including a circular hole of radius R at (a_H, a_V) in the e_1 and e_2 directions of the center lines. Right: Properties of the eight random inputs in the fiber-reinforced matrix. Here, κ is the Young's modulus and ν is the Poisson ratio, see Appendix B.2 for details.

ure 7 (left) shows a square plate of length $3L_H$ in the e_1 direction and $3L_V$ in the direction with a circular inclusion of radius R. In the domain $\mathcal{D} = [-1.5L_H, 1.5L_H] \times [-1.5L_V, 1.5L_V] \in \mathbb{R}^2$, the loading is given by an applied normal displacement $\mathbf{v}_n = \bar{v} \cdot e_2$ on the boundary Γ_N. The other boundaries, denoted Γ_D, are free, corresponding to zero displacement on Γ_D. The



(a) Fine mesh (53,766 DOFs) (b) Coarse mesh (13,416 DOFs) (c) Coarse mesh (975 DOFs)



(d) Coarse mesh (324 DOFs) (e) Coarse mesh (195 DOFs)

Figure 8: Fiber-reinforced matrix. The fine finite-element mesh in (a) is used to generate the high-fidelity QoI Y, while the coarse meshes in (b)–(e) are used to generate the low-fidelity QoIs $X_1, X_2, X_3, X4$, respectively.

closure of the domain is $\bar{\mathcal{D}} \equiv \mathcal{D} \cup \Gamma$, with $\Gamma = \Gamma_D \cup \Gamma_N$. The unknowns are the displacement field $\mathbf{u} = (u, v)^\intercal \in \mathbb{R}^2$ and the scalar damage variable $\phi_d \in \mathbb{R}$ in the domain $\bar{\mathcal{D}}$ of the elastic body. This setup models the traction experiment of a fiber-reinforced matrix [1, 6], with the corresponding boundary value problem as described in [25]. The PDE formulation is: find $\mathbf{u}(\mathbf{x})$ and $\phi_d(\mathbf{x})$ for $\mathbf{x} = (x_1, x_2)^\intercal \in \bar{\mathcal{D}}$, such that

$$[(1 - \phi_d(\mathbf{x}))^2 + q]\nabla \cdot \boldsymbol{\sigma}(\mathbf{x}) = \mathbf{0}, \tag{6.1}$$

$$-G_c \ell_0 \nabla^2 \phi_d(\mathbf{x}) + \left[\frac{G_c}{\ell_0} + 2H(\mathbf{x})\right] \phi_d(\mathbf{x}) = 2H(\mathbf{x}), \tag{6.2}$$

with corresponding boundary conditions on Γ_N and Γ_D. The full model details, including definitions for ℓ_0, $\boldsymbol{\sigma}$, q, and H are shown in Appendix B.2. We consider eight input random variables, $p_1, \ldots p_8$, which are stemming from material properties and geometries, see Figure 7 (right).

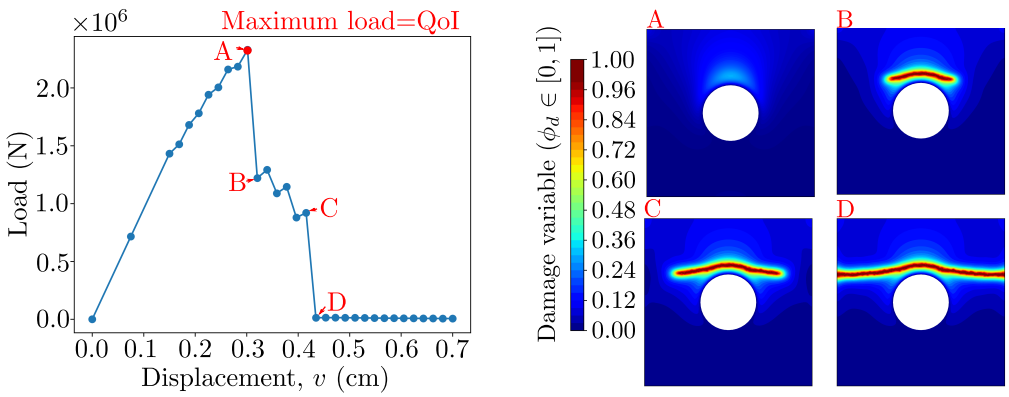
6.3.1. High-fidelity and low-fidelity models. The model (6.1) and (6.2) for brittle fracture analysis is solved using an iterative solver, wherein we solve for the scalar damage variable (ϕ_d) using the displacement fields ($\mathbf{u}$). Subsequently, the updated damage variable is used to solve for the displacement field, and the process is repeated until the difference between

the current and previous iterates becomes less than the user-defined tolerance $\delta \ll 1$. We set $\delta = 5 \times 10^{-3}$ for the high-fidelity model, and $\delta = 5 \times 10^{-2},\ 0.2,\ 0.4$ for the lower-fidelity models. Figure 8 shows a fine mesh and several coarse meshes used for the high and low-fidelity models, respectively. The details of the high and low-fidelity models, including CPU times to implement finite element analysis, are reported in Table 2, which also reports (normalized) model costs. The oracle correlations between the QoI (Y) of the high-fidelity model and its low-fidelity QoIs X_1, X_2, X_3, X_4 are 0.96, 0.93, 0.87, and 0.74, respectively.

Table 2: Fiber-reinforced matrix. Comparison of the high-fidelity and four different low-fidelity finite element models to compute the QoI.

Model type	Tolerance (δ)	DOFs	CPU time (s)[a]	Normalized cost[b]
High-fidelity, Y	5×10^{-3}	53,766	250.51	108.9
Low-fidelity 1, X_1	5×10^{-2}	13,416	20.95	9.1
Low-fidelity 2, X_2	0.2	975	2.97	1.3
Low-fidelity 3, X_3	0.2	324	2.50	1.1
Low-fidelity 4, X_4	0.4	195	2.30	1

a. The CPU time is averaged over 5 trials.
b. The cost is normalized so that sampling X_4 has unit cost.



(a) Load-displacement curve (b) Damage variable contour

Figure 9: Fiber-reinforced matrix. Finite element analysis results: (a) The ultimate tensile load in the load-displacement curve is recorded as the QoI. (b) the damage variable contour shows the degree of damage ($0 < \phi_d \leq 1$) that occurred in regimes 'A'–'D' of the load-displacement curve, indicating that brittle fracture occurred at the top of circular hole advances in the regime 'A' to 'B' before a complete fracture occurs in regime 'D'.

We measure the maximum tensile load as a QoI from the load-displacement curve. Figure 9a presents the relationship between the resulting load and the imposed displacement on the top of the fiber-reinforced matrix. As the applied displacement $\bar{v}$ at the top increases from

884 0 cm to 7.5×10^{-2} cm, the resulting load exhibits an almost linear increase until the structure
885 begins to sustain damage. Upon reaching a peak load, the rate of change of the resulting load
886 over displacement significantly decreases. This behavior is observed from regimes 'A' to 'B'
887 in Figure 9a. The regimes occur due to the partial fracturing of the matrix, as indicated by a
888 damage variable of value $\phi_d = 1$ in Figure 9b. There is another substantial drop in load from
889 regimes 'C' to 'D', presenting complete fracture throughout the entire domain of the matrix.
890 The maximum tensile load at 'A' is represented as a scalar; thus, for the high-fidelity QoI we
891 have $d = 1$, while for the low-fidelity QoIs we have $d_1 = d_2 = d_3 = d_4 = 1$.

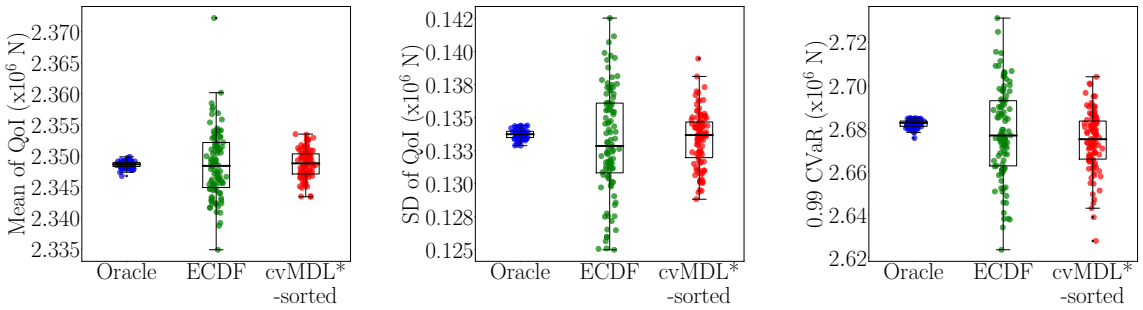
Table 3: Fiber-reinforced matrix. Comparison of the accuracy of cvMDL*-sorted and ECDF in estimating CDF, mean, standard deviation, and CVaR at $\beta = 0.99$ for the QoI (the ultimate tensile load). We present the mean error of these estimates relative to oracle estimates obtained from 5500 i.i.d. high-fidelity samples ($B = 577{,}170$) over 100 trials.

Method	CDF (error)[a]	Mean (-)[b]	Standard deviation (-)[b]	$\text{CVaR}_{0.99}$ (-)[b]
		Budget $B = 20{,}000$		
cvMDL*-sorted	2.192×10^{-4}	1.559×10^{-3}	2.291×10^{-2}	6.698×10^{-3}
ECDF	3.916×10^{-4}	2.991×10^{-3}	3.837×10^{-2}	9.435×10^{-3}
		Budget $B = 35{,}000$		
cvMDL*-sorted	1.194×10^{-4}	1.168×10^{-3}	1.575×10^{-2}	5.531×10^{-3}
ECDF	2.269×10^{-4}	2.311×10^{-3}	2.955×10^{-2}	7.577×10^{-3}
		Budget $B = 50{,}000$		
cvMDL*-sorted	8.441×10^{-5}	8.957×10^{-4}	1.340×10^{-2}	4.846×10^{-3}
ECDF	1.627×10^{-4}	1.983×10^{-3}	2.670×10^{-2}	7.102×10^{-3}

a. We determine the mean $\omega(x)$-weighted L^2 error between F_Y and the estimated CDFs given by cvMDL*-sorted and ECDF. The mean $\omega(x)$-weighted L^2 errors are averaged over independent 100 trials.
b. We use the mean relative error of the estimates in the comparison of the oracle estimates over independent 100 trials.

892 **6.3.2. Results for CDF, mean, standard deviation, and CVaR estimation.** The high-
893 fidelity simulations are costly enough here that we must approximate the oracle solution with
894 limited samples: 6000 high-fidelity simulations are generated, and we randomly select 5500
895 to estimate a quantity. We generate an ensemble of 100 such instances and use the average
896 as the oracle. For the multi-fidelity procedure, 6000 joint high- and low-fidelity samples are
897 used as the pool from which model samples are drawn. We investigate three budget values
898 as reported in Table 3. Over the corresponding 100 trials, cvMDL*-sorted predominantly
899 selects the model subset $\mathcal{S} = \{2, 4\}$ (selected 95, 96, and 98 times for the 3 budget values,
900 respectively) and less frequently selects the model subset $\mathcal{S} = \{4\}$ (selected 5, 4, and 2 times,
901 respectively) from the model set $\{1, 2, 3, 4\}$. This process yields averaged optimal exploration
902 sample numbers $m^*_{\mathcal{S}} = 140$, 245, and 350 for each respective budget B.
903 In Table 3, cvMDL*-sorted surpasses ECDF in terms of mean errors for CDF, mean,
904 standard deviation, and CVaR at $a = 0.99$ for the QoI. The second column of Table 3 reports
905 the mean-weighted L^2 error for F_Y over $Y \in [1.5 \times 10^6, 3 \times 10^6]$ (N). The last four columns of

that table show that the proposed cvMDL approach yields nearly twice the accuracy compared to the ECDF method, and this increased accuracy extends to the estimated statistics and risk metric.



(a) Mean (b) Standard deviation (c) $\text{CVaR}_{0.99}$

Figure 10: Fiber-reinforced matrix. (a): Boxplots of the mean of QoI, computed by ECDF and cvMDL*-sorted when $B = 50{,}000$ with 100 experiments. (b): Same boxplots for the standard deviation of QoI. (c): Same boxplots for CVaR (a=0.99).

Figures 10a–10c present the results of the statistical mean, standard deviation, and CVaR at $a = 0.99$ via boxplots. The cvMDL*-sorted method achieves higher accuracy than the ECDF by a significant margin when compared to the oracle results. The box plots demonstrate that the statistical estimates by the cvMDL*-sorted method exhibit a lower spread than those of the ECDF, showing that cvMDL*-sorted estimates have smaller variance for this example.

7. Conclusions. We developed a versatile framework for efficiently estimating the CDFs of QoI subject to a budget constraint. To implement this framework, we constructed a set of binary control variables based on linear surrogates and used them in an adaptive meta algorithm (cvMDL) that estimates the CDFs. We established both uniform consistency and trade-off optimality for the corresponding algorithm as the budget tends to infinity.

Although the proposed framework is built upon an existing bandit-learning paradigm, our treatment of exploration and exploitation distinguishes itself from the previous works. In particular, the new approach employed in our framework leads to innovative estimators that dramatically relaxes the reliance on underlying model assumptions. Furthermore, the approach allows for the treatment of different types of QoIs, both vector- and scalar-valued. To the best of our knowledge, our framework provides the first robust multifidelity CDF estimator under a budget constraint that can deal with both heterogeneous model sets and multi-valued QoIs at the same time, meanwhile requiring no a priori cross-model statistics.

Appendix A. Proofs of the main results.

A.1. Proof of Theorem 5.2. Sort the entries of $\boldsymbol{T}$ in increasing order: $z^{(1)} < \cdots < z^{(M)}$, where M is the product of each dimension M_i of $\boldsymbol{T}$: $M = M_1 \cdots M_d$. At the beginning of the algorithm, the index of $z^{(1)}$ is strictly decreasing in each direction. As a result, $z^{(1)}$

arrives at the entry of $\boldsymbol{T}$ with index $(1,\dots,1)$ after a finite number of iterations, and after that, it remains unchanged in the subsequent iteration. In fact, for every $s < M$, assuming $z^{(1)},\dots,z^{(s)}$ have reached their final positions after which no change occurs, the index of $z^{(s+1)}$ is decreasing in each direction if the algorithm has not converged yet. The result follows by noting that M is finite, and a stationary point must possess the desired monotonicity.

A.2. Proof of Lemma 5.8. This section contains the proofs of statements (i) through (vi) in Lemma 5.8. The proof of statement (ii), the asymptotic consistency of $\widehat{\boldsymbol{B}}_{\mathcal{S}^+}$, is a direct result of the strong law of large numbers, so we omit this proof.

A.2.1. Proof of Statement (i). We only prove (5.15b) as the proof for (5.15a) is similar. Denote by $\boldsymbol{e}_i$ the ith unit vector in $\mathbb{R}^d$, i.e., $\boldsymbol{e}_d^{(j)} = \delta_{ij}, j \in \{1:d\}$, where δ is the Kronecker notation, and $\boldsymbol{e} := \sum_{i\in\{1:d\}} \boldsymbol{e}_i$ is the all-ones vector in $\mathbb{R}^d$. For fixed $\boldsymbol{x} \in \mathbb{R}^d$, without loss of generality, we assume $F_{V(\boldsymbol{A})\vee Y}(\boldsymbol{x}) \le F_{V(\boldsymbol{B}_{\mathcal{S}^+})\vee Y}(\boldsymbol{x})$, as the other case is similar by reversing $F_{V(\boldsymbol{A})\vee Y}(\boldsymbol{x})$ and $F_{V(\boldsymbol{B}_{\mathcal{S}^+})\vee Y}(\boldsymbol{x})$. Meanwhile,

$$V(\boldsymbol{B}_{\mathcal{S}^+}) \vee Y \le \boldsymbol{x} \Rightarrow V(\boldsymbol{A}) \vee Y \le \boldsymbol{x} + \Delta\boldsymbol{x} \Rightarrow V(\boldsymbol{A}) \vee Y - \|\Delta\boldsymbol{x}\|_\infty \boldsymbol{e} \le \boldsymbol{x},$$

where

$$\Delta\boldsymbol{x} = \sum_{i\in\{1:d\}} |X_{\mathcal{S}^+}^\top (\boldsymbol{A}^{(i)} - \boldsymbol{B}_{\mathcal{S}^+}^{(i)})| \boldsymbol{e}_i.$$

Hence, under Assumptions 5.3 and 5.6, for $t > 0$,

$$\begin{aligned}
|F_{V(\boldsymbol{A})\vee Y}(\boldsymbol{x}) - F_{V(\boldsymbol{B}_{\mathcal{S}^+})\vee Y}(\boldsymbol{x})| &= F_{V(\boldsymbol{B}_{\mathcal{S}^+})\vee Y}(\boldsymbol{x}) - F_{V(\boldsymbol{A})\vee Y}(\boldsymbol{x}) \\
&\le F_{V(\boldsymbol{A})\vee Y - \|\Delta\boldsymbol{x}\|_\infty \boldsymbol{e}}(\boldsymbol{x}) - F_{V(\boldsymbol{A})\vee Y}(\boldsymbol{x}) \\
&\le F_{V(\boldsymbol{A})\vee Y - t\boldsymbol{e}}(\boldsymbol{x}) - F_{V(\boldsymbol{A})\vee Y}(\boldsymbol{x}) + \mathbb{P}\left(\|\Delta\boldsymbol{x}\|_\infty \ge t\right) \\
&= F_{V(\boldsymbol{A})\vee Y}(\boldsymbol{x} + t\boldsymbol{e}) - F_{V(\boldsymbol{A})\vee Y}(\boldsymbol{x}) + \mathbb{P}\left(\|\Delta\boldsymbol{x}\|_\infty \ge t\right) \\
&\lesssim C\sqrt{d}t + \sum_{i\in\{1:d\}} \mathbb{P}\left(\|X_{\mathcal{S}^+}^\top (\boldsymbol{A}^{(i)} - \boldsymbol{B}_{\mathcal{S}^+}^{(i)})\|_2 \ge t\right) \\
&\overset{(5.13)}{\lesssim} C\sqrt{d}t + \frac{1}{t^2} \sum_{i\in\{1:d\}} \|\boldsymbol{A}^{(i)} - \boldsymbol{B}_{\mathcal{S}^+}^{(i)}\|_2^2,
\end{aligned}$$

where the penultimate inequality follows from the Lipschitz condition on $F_{V(\boldsymbol{A})\vee Y}$ and a union bound, and the last inequality follows from Markov's inequality. Taking $t = \|\boldsymbol{A} - \boldsymbol{B}_{\mathcal{S}^+}\|_F^{2/3}$ yields the desired result.

A.2.2. Proof of Statement (iii). We only prove the first statement; the second can be proved similarly. Recall that

$$\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) = G(\widehat{\boldsymbol{B}}_{\mathcal{S}^+}; \boldsymbol{x}) \qquad G(\boldsymbol{A}; \boldsymbol{x}) := \frac{1}{m} \sum_{\ell\in\{1:m\}} \mathbf{1}_{\{(X_{\mathrm{epr},\ell,\mathcal{S}^+}^\top \boldsymbol{A})^\top \le \boldsymbol{x}\}} \quad \boldsymbol{A} \in \mathbb{R}^{(d_{\mathcal{S}}+1)\times d},$$

where $X_{\mathrm{epr},\ell,\mathcal{S}^+}$ denotes the ℓth exploration sample of $X_{\mathcal{S}}$ with intercept. It follows from the direct computation that

$$
\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) - F_{H_{\mathcal{S}}}(\boldsymbol{x})| \\
&\le \sup_{\boldsymbol{x}\in\mathbb{R}^d} \left| G(\widehat{\boldsymbol{B}}_{\mathcal{S}^+};\boldsymbol{x}) - \mathbb{P}((X_{\mathcal{S}^+}^\top \widehat{\boldsymbol{B}}_{\mathcal{S}^+})^\top \le \boldsymbol{x}) \right| + \sup_{\boldsymbol{x}\in\mathbb{R}^d} \left| \mathbb{P}((X_{\mathcal{S}^+}^\top \widehat{\boldsymbol{B}}_{\mathcal{S}^+})^\top \le \boldsymbol{x}) - \mathbb{P}((X_{\mathcal{S}^+}^\top \boldsymbol{B}_{\mathcal{S}^+})^\top \le \boldsymbol{x}) \right| \\
&\le \sup_{\boldsymbol{A}\in\mathbb{R}^{(d_{\mathcal{S}}+1)\times d}} \sup_{\boldsymbol{x}\in\mathbb{R}^d} \left| G(\boldsymbol{A};\boldsymbol{x}) - \mathbb{P}((X_{\mathcal{S}^+}^\top \boldsymbol{A})^\top \le \boldsymbol{x}) \right| + \sup_{\boldsymbol{x}\in\mathbb{R}^d} \left| \mathbb{P}((X_{\mathcal{S}^+}^\top \widehat{\boldsymbol{B}}_{\mathcal{S}^+})^\top \le \boldsymbol{x}) - \mathbb{P}((X_{\mathcal{S}^+}^\top \boldsymbol{B}_{\mathcal{S}^+})^\top \le \boldsymbol{x}) \right| \\
&= \underbrace{\sup_{\boldsymbol{A}\in\mathbb{R}^{(d_{\mathcal{S}}+1)\times d}} \left| G(\boldsymbol{A};\boldsymbol{0}) - \mathbb{E}[G(\boldsymbol{A};\boldsymbol{0})] \right|}_{\Lambda_{m,1}} + \underbrace{\sup_{\boldsymbol{x}\in\mathbb{R}^d} \left| \mathbb{P}((X_{\mathcal{S}^+}^\top \widehat{\boldsymbol{B}}_{\mathcal{S}^+})^\top \le \boldsymbol{x}) - \mathbb{P}((X_{\mathcal{S}^+}^\top \boldsymbol{B}_{\mathcal{S}^+})^\top \le \boldsymbol{x}) \right|}_{\Lambda_{m,2}}
\end{aligned}
$$

where $X_{\mathcal{S}^+}$ a general notation that is independent of $\widehat{\boldsymbol{B}}_{\mathcal{S}^+}$ and $\boldsymbol{0}$ is the all-zeros vector. Note that $\Lambda_{m,1}$ has no supremum over $\boldsymbol{x}$ since one is able to alter the intercept coefficients in $\boldsymbol{A}$ to yield different values of $\boldsymbol{x}\in\mathbb{R}^d$ without changing the coefficients of $X_{\mathcal{S}}$. In what follows, we show that both $\Lambda_{m,1}$ and $\Lambda_{m,2}$ converge to 0 a.s.

To bound $\Lambda_{m,1}$, we appeal to the empirical process theory. For any $\boldsymbol{A}\in\mathbb{R}^{(d_{\mathcal{S}}+1)\times d}$, the indicator function $\mathbf{1}_{\{(X_{\ell,\mathcal{S}^+}^\top \boldsymbol{A})^\top \le \boldsymbol{0}\}} \le 1$. According to Massart concentration inequality [7, Theorem 14.2], we have for any $t>0$ such that

$$
\mathbb{P}(\Lambda_{m,1} > \mathbb{E}[\Lambda_{m,1}] + t) \le \exp\bigl(-mt^2/8\bigr).
$$

Taking $t = 4\sqrt{\log m/m}$,

$$
\mathbb{P}\left(\Lambda_{m,1} > \mathbb{E}[\Lambda_{m,1}] + 4\sqrt{\frac{\log m}{m}}\right) \le m^{-2}. \tag{A.1}
$$

Since $\sum_{m=1}^{\infty} m^{-2} < \infty$, by the Borel-Cantelli lemma, we conclude that

$$
\Lambda_{m,1} \le \mathbb{E}[\Lambda_{m,1}] + 4\sqrt{\frac{\log m}{m}} \tag{A.2}
$$

for all sufficiently large m a.s. To bound $\mathbb{E}[\Lambda_{m,1}]$, note that the supremum in $\mathbb{E}[\Lambda_{m,1}]$ is taken over all indicator functions defined on d intersections of hyperplanes in $\mathbb{R}^{d_{\mathcal{S}}}$ (the constant dimension is only a shift), which has a finite Vapnik–Chervonenkis (VC) dimension of order $d_{\mathcal{S}} d\log d$ [4]. According to [34, Theorem 8.3.23], there exists a universal constant C' such that $\mathbb{E}[\Lambda_{m,1}] \le C'\sqrt{d_{\mathcal{S}} d\log d/m}$. This combined with (A.2) shows that $\Lambda_{m,1}\to 0$ a.s.

To bound $\Lambda_{m,2}$, note that by Statement (ii) in Lemma 5.8, a.s., for all sufficiently large m, $\|\widehat{\boldsymbol{B}}_{\mathcal{S}^+} - \boldsymbol{B}_{\mathcal{S}^+}\|_F < \varepsilon$ where ε is the same as in Assumption 5.6. Since $X_{\mathcal{S}^+}$ is independent of $\widehat{\boldsymbol{B}}_{\mathcal{S}^+}$, conditioning on $\|\widehat{\boldsymbol{B}}_{\mathcal{S}^+} - \boldsymbol{B}_{\mathcal{S}^+}\|_F < \varepsilon$ and applying Statement (i) of Lemma 5.8, $\Lambda_{m,2} \lesssim \|\widehat{\boldsymbol{B}}_{\mathcal{S}^+} - \boldsymbol{B}_{\mathcal{S}^+}\|_F^{2/3}$. Now taking $m\to\infty$ shows $\Lambda_{m,2}\to 0$ a.s.

A.2.3. Proof of Statement (iv). Note $\mathcal{K}_1(\boldsymbol{x}) + \mathcal{K}_2(\boldsymbol{x}) = \widehat{F}_Y(\boldsymbol{x})(1 - \widehat{F}_Y(\boldsymbol{x}))$, which is a consistent estimator for $F_Y(\boldsymbol{x})(1 - F_Y(\boldsymbol{x}))$ for all $\boldsymbol{x} \in \mathbb{R}^d$ a.s. as a result of the strong law of large numbers. Therefore, it suffices to prove the consistency for $\mathcal{K}_2(\boldsymbol{x})$ only.

Note $\mathcal{K}_2(\boldsymbol{x})$ in (5.3) can be rewritten as

$$\mathcal{K}_2(\boldsymbol{x}) = \widehat{\rho}^2_{\mathcal{S}}(\boldsymbol{x})\widehat{F}_Y(\boldsymbol{x})(1 - \widehat{F}_Y(\boldsymbol{x})) = \begin{cases} \frac{(\widehat{F}_{Y\vee\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) - \widehat{F}_Y(\boldsymbol{x})\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}))^2}{\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x})(1-\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}))} & \boldsymbol{x} \in (\mathrm{supp}(\widehat{F}_{\widehat{H}_{\mathcal{S}}}))^\circ \\ 0 & \text{otherwise} \end{cases} \tag{A.3}$$

where

$$\widehat{\rho}^2_{\mathcal{S}}(\boldsymbol{x}) = \begin{cases} \frac{(\widehat{F}_{Y\vee\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) - \widehat{F}_Y(\boldsymbol{x})\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}))^2}{\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x})(1-\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}))\widehat{F}_Y(\boldsymbol{x})(1-\widehat{F}_Y(\boldsymbol{x}))} & \boldsymbol{x} \in (\mathrm{supp}(\widehat{F}_Y))^\circ \cap (\mathrm{supp}(\widehat{F}_{\widehat{H}_{\mathcal{S}}}))^\circ \\ 0 & \text{otherwise} \end{cases}$$

is the empirical estimator for $\rho^2_{\mathcal{S}}(\boldsymbol{x})$. On the other hand,

$$\rho^2_{\mathcal{S}}(\boldsymbol{x})F_Y(\boldsymbol{x})(1 - F_Y(\boldsymbol{x})) = \begin{cases} \frac{(F_{Y\vee H_{\mathcal{S}}}(\boldsymbol{x}) - F_Y(\boldsymbol{x})F_{H_{\mathcal{S}}}(\boldsymbol{x}))^2}{F_{H_{\mathcal{S}}}(\boldsymbol{x})(1-F_{H_{\mathcal{S}}}(\boldsymbol{x}))} & \boldsymbol{x} \in (\mathrm{supp}(F_{H_{\mathcal{S}}}))^\circ \\ 0 & \text{otherwise} \end{cases} \tag{A.4}$$

Comparing (A.3) and (A.4), the desired result follows from statement (iii) in Lemma 5.8.

A.2.4. Proof of statement (v). We prove the consistency of $\widehat{k}_2(\mathcal{S})$; the consistency of $\widehat{k}_1(\mathcal{S})$ can be proved similarly. By statement (iv) in Lemma 5.8, $\mathcal{K}_2(\boldsymbol{x})$ converges to $\rho^2_{\mathcal{S}}(\boldsymbol{x})F_Y(\boldsymbol{x})(1 - F_Y(\boldsymbol{x}))$ for all $\boldsymbol{x} \in \mathbb{R}^d$ as $m \to \infty$ a.s.

We first prove the first case where $d = 1$ and $\|\omega\|_{L^\infty(\mathbb{R})} = C < \infty$, and we change the notation $\boldsymbol{x}$ to the lowercase x. Under the moment condition in Assumption 5.3, according to [5, Theorem 2.13],

$$W_1\left(F_Y, \widehat{F}_Y\right) = \int_{\mathbb{R}} |\widehat{F}_Y(x) - F_Y(x)|\mathrm{d}x \to 0 \qquad\qquad m \to \infty,$$

where W_1 is the Wasserstein-1 metric. Fix an arbitrary trajectory in the sample space such that $\mathcal{K}_2(x) \to \rho^2_{\mathcal{S}}(x)F_Y(x)(1 - F_Y(x))$ and $\int_{\mathbb{R}} |\widehat{F}_Y(x) - F_Y(x)|\mathrm{d}x \to 0$. In the following, we treat $\mathcal{K}_2(x)$ as a deterministic sequence.

To show the consistency of $\widehat{k}_2(\mathcal{S})$, it remains to justify the change of order of taking limit and integration:

$$\begin{aligned} \lim_{m\to\infty} \widehat{k}_2(\mathcal{S}) = \lim_{m\to\infty} c_{\mathcal{S}} \int_{\mathbb{R}} \omega(x)\mathcal{K}_2(x)\mathrm{d}x &= c_{\mathcal{S}} \int_{\mathbb{R}} \lim_{m\to\infty} \omega(x)\mathcal{K}_2(x)\mathrm{d}x \\ &= c_{\mathcal{S}} \int_{\mathbb{R}} \omega(x)\rho^2_{\mathcal{S}}(x)F_Y(x)(1 - F_Y(x))\mathrm{d}x = k_2(\mathcal{S}), \end{aligned}$$

for which we appeal to the Vitali convergence theorem. To apply the Vitali convergence theorem, we need to verify that the sequence $\omega(x)\mathcal{K}_2(x)$ is uniformly integrable and has

absolutely continuous integrals. To this end, recall the representation $\mathcal{K}_2(x)$ in (A.3). Since the square of the empirical correlation estimator is bounded by 1, a.s.,

$$\omega(x)\mathcal{K}_2(x) \le \omega(x)\widehat{F}_Y(x)(1-\widehat{F}_Y(x)) \le \frac{C}{4} < C.$$

The absolutely continuous integrals part follows immediately from the uniform boundedness. For uniform integrability, we first observe

$$\begin{aligned}\int_{\mathbb{R}} |\omega(x)F_Y(x)(1-F_Y(x)) - \omega(x)\widehat{F}_Y(x)(1-\widehat{F}_Y(x))|\mathrm{d}x &\le \int_{\mathbb{R}} \omega(x)|\widehat{F}_Y(x) - F_Y(x)|\mathrm{d}x \\ &\le C\int_{\mathbb{R}} |\widehat{F}_Y(x) - F_Y(x)|\mathrm{d}x \to 0.\end{aligned}$$

Thus,

$$\begin{aligned}\int_{|x|>M} \omega(x)\mathcal{K}_2(x)\mathrm{d}x &\le \int_{|x|>M} \omega(x)\widehat{F}_Y(x)(1-\widehat{F}_Y(x))\mathrm{d}x \\ &\le \int_{|x|>M} \omega(x)F_Y(x)(1-F_Y(x))\mathrm{d}x + C\int_{|x|>M} |\widehat{F}_Y(x) - F_Y(x)|\mathrm{d}x \\ &\lesssim \int_{|x|>M} \frac{C}{x^2}\mathrm{d}x + C\int_{\mathbb{R}} |\widehat{F}_Y(x) - F_Y(x)|\mathrm{d}x,\end{aligned}$$

where the last step follows from Assumption 5.3 and Chebyshev's inequality. For every $\varepsilon > 0$, we can choose m and M sufficiently large so the right-hand side is less than ε. The uniform integrability follows by enlarging M to accommodate the first m terms.

The proof for (b) is similar. It suffices to verify the change of order for the sequence $\omega(\boldsymbol{x})\mathcal{K}_2(\boldsymbol{x})$. Since $\omega(\boldsymbol{x})\mathcal{K}_2(\boldsymbol{x}) \le \omega(\boldsymbol{x})$ and the latter is integrable and independent of m, the dominated convergence does the rest.

A.2.5. Proof of statement (vi). For $\boldsymbol{x} \in (\mathrm{supp}(F_{H_{\mathcal{S}}}))^\circ$, it is easy to show via a contradiction argument that $\boldsymbol{x} \in \mathrm{supp}(\widehat{F}_{\widehat{H}_{\mathcal{S}}})$ for all sufficiently large m a.s. By statement (iii) in Lemma 5.8, $\widehat{F}_{Y\vee\widehat{H}_{\mathcal{S}}}$ and $\widehat{F}_{\widehat{H}_{\mathcal{S}}}$ are consistent estimators. Meanwhile, $\widehat{F}_Y(\boldsymbol{x})$ is a consistent estimator for $F_Y(\boldsymbol{x})$ due to the strong law of large numbers. Therefore, we obtain

$$\widehat{\alpha}(\boldsymbol{x}) = \frac{\widehat{F}_{Y\vee\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) - \widehat{F}_Y(\boldsymbol{x})\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x})}{\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x})(1-\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}))} \to \alpha(\boldsymbol{x}) = \frac{F_{Y\vee H_{\mathcal{S}}}(\boldsymbol{x}) - F_Y(\boldsymbol{x})F_{H_{\mathcal{S}}}(\boldsymbol{x})}{F_{H_{\mathcal{S}}}(\boldsymbol{x})(1-F_{\boldsymbol{H}_{\mathcal{S}}}(\boldsymbol{x}))}$$

as $m \to \infty$ almost surely.

A.3. Proof of Lemma 5.10. Recall in (5.10) that

$$\begin{aligned}\widetilde{F}_{\mathcal{S}}(\boldsymbol{x}) &= \widehat{F}_Y(\boldsymbol{x}) - \frac{1}{m}\sum_{\ell\in\{1:m\}}\left(\widehat{\alpha}(\boldsymbol{x})\widehat{h}_{\mathcal{S}}(X_{\mathrm{epr},\ell,\mathcal{S}};\boldsymbol{x}) - \frac{1}{N_{\mathcal{S}}}\sum_{j\in\{1:N_{\mathcal{S}}\}}\widehat{\alpha}(\boldsymbol{x})\widehat{h}_{\mathcal{S}}(X_{\mathrm{ept},\ell,\mathcal{S}};\boldsymbol{x})\right) \\ &= \widehat{F}_Y(\boldsymbol{x}) - \widehat{\alpha}(\boldsymbol{x})\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x}) + \widehat{\alpha}(\boldsymbol{x})\left(\frac{1}{N_{\mathcal{S}}}\sum_{j\in\{1:N_{\mathcal{S}}\}}\mathbf{1}_{\{(X_{\mathrm{ept},j,\mathcal{S}^+}^\top\widehat{\boldsymbol{B}}_{\mathcal{S}^+})^\top\le\boldsymbol{x}\}}\right).\end{aligned}$$

Thus,

$$\begin{aligned}\sup_{\boldsymbol{x}\in\mathbb{R}^d}|\widetilde{F}_{\mathcal{S}}(\boldsymbol{x})-F_Y(\boldsymbol{x})| &\le \sup_{\boldsymbol{x}\in\mathbb{R}^d}|\widehat{F}_Y(\boldsymbol{x})-F_Y(\boldsymbol{x})|+\sup_{\boldsymbol{x}\in\mathbb{R}^d}|\widehat{\alpha}(\boldsymbol{x})(\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x})-F_{H_{\mathcal{S}}}(\boldsymbol{x}))| \\ &\quad+\sup_{\boldsymbol{x}\in\mathbb{R}^d}\left|\widehat{\alpha}(\boldsymbol{x})\left(\frac{1}{N_{\mathcal{S}}}\sum_{j\in\{1:N_{\mathcal{S}}\}}\mathbf{1}_{\{(X_{\mathrm{ept},j,\mathcal{S}^+}^\top\widehat{\boldsymbol{B}}_{\mathcal{S}^+})^\top\le\boldsymbol{x}\}}-F_{H_{\mathcal{S}}}(\boldsymbol{x})\right)\right| \\ &\overset{(5.9)}{\le} \underbrace{\sup_{\boldsymbol{x}\in\mathbb{R}^d}|\widehat{F}_Y(\boldsymbol{x})-F_Y(\boldsymbol{x})|}_{(i)}+\underbrace{\sup_{\boldsymbol{x}\in\mathbb{R}^d}|\widehat{F}_{\widehat{H}_{\mathcal{S}}}(\boldsymbol{x})-F_{H_{\mathcal{S}}}(\boldsymbol{x})|}_{(ii)} \\ &\quad+\underbrace{\sup_{\boldsymbol{x}\in\mathbb{R}^d}\left|\left(\frac{1}{N_{\mathcal{S}}}\sum_{j\in\{1:N_{\mathcal{S}}\}}\mathbf{1}_{\{(X_{\mathrm{ept},j,\mathcal{S}^+}^\top\widehat{\boldsymbol{B}}_{\mathcal{S}^+})^\top\le\boldsymbol{x}\}}-F_{H_{\mathcal{S}}}(\boldsymbol{x})\right)\right|}_{(iii)}.\end{aligned}$$

Note (i) converges to 0 as $m\to\infty$ due to the multivariate Glivenko-Cantelli theorem. (ii) converges to 0 as $m\to\infty$ due to statement (iii) in Lemma 5.8. A similar argument as in the proof of statement (iii) of Lemma 5.8 can be used to prove that (iii) converges to 0 as $N_{\mathcal{S}}\to\infty$, which is not repeated here.

A.4. Proof of Theorem 5.7. To reduce notational confusion with m, we use t to denote the number of exploration samples. The exploration rate m grows nonlinearly with respect to an index that counts the iterations of the exploration loop in Algorithm 5.2. We let q denote the loop iteration index, and t_q the corresponding exploration rate, i.e., $t_1=n+2$. Let $q(B)$ be the total number of exploration iteration steps in Algorithm 5.2, which is random. It follows from the definition that $t_{q(B)}=m(B)$ and

$$n+1\le t_q\le t_{q+1}\le 2t_q \qquad 1\le q<q(B). \tag{A.5}$$

We first show that $m(B)$ diverges as $B\to\infty$ a.s. According to statement (v) in Lemma 5.8, $\widehat{k}_1(\mathcal{S})\to k_1(\mathcal{S}),\widehat{k}_2(\mathcal{S})\to k_2(\mathcal{S})$ for $\mathcal{S}\subseteq\{1:n\}$ a.s. As a result, for almost every realization $\omega\in\Omega$, where Ω denotes the product space of exploration samples, there exists an $0<L(\omega),L'(\omega)<\infty$,

$$\sup_{t>n+1}\max_{\mathcal{S}\subseteq\{1:n\}}\widehat{k}_1(\mathcal{S};\omega)<L(\omega)<\infty \qquad \inf_{t>n+1}\min_{\mathcal{S}\subseteq\{1:n\}}\widehat{k}_2(\mathcal{S};\omega)>L'(\omega)>0$$

The exploration stopping criterion of cvMDL in Algorithm 5.2 requires that

$$m(B;\omega)\ge\widehat{m}^*_{\mathcal{S}(B;\omega)}\ge\frac{B}{c_{\mathrm{epr}}+\sqrt{\frac{c_{\mathrm{epr}}L(\omega)}{L'(\omega)}}}\to\infty \qquad B\to\infty.$$

Thus,

$$\lim_{B\to\infty}m(B;\omega)=\infty. \tag{A.6}$$

We now work with a fixed realization ω along which $m(B;\omega) \to \infty$ as $B \to \infty$, and $\widehat{k}_1(\mathcal{S}), \widehat{k}_2(\mathcal{S})$ converge to the true parameters as $t \to \infty$. We prove that both (5.14a) and (5.14b) hold for such an ω. Fix $\delta < 1/2$ sufficiently small. Since $\mathcal{S}^*$ is assumed unique, a continuity argument implies that there exists a sufficiently large $T(\delta;\omega)$, such that for all $t \geq T(\delta;\omega)$,

$$\max_{(1-\delta)m^*_{\mathcal{S}^*} \leq m \leq (1+\delta)m^*_{\mathcal{S}^*}} \widehat{L}_{\mathcal{S}^*}(m;t) < \min_{\mathcal{S}\subseteq\{1:n\}, \mathcal{S}\neq\mathcal{S}^*} \widehat{L}^*_{\mathcal{S}}(t). \tag{A.7}$$

$$1-\delta \leq \frac{\widehat{m}^*_{\mathcal{S}}(t;\omega)}{m^*_{\mathcal{S}}} \leq 1+\delta \qquad \forall \mathcal{S} \subseteq \{1:n\}, \tag{A.8}$$

where $\widehat{L}_{\mathcal{S}^*}(\cdot;t)$ is the estimated loss function for $\mathcal{S}^*$ using t exploration samples, and $\widehat{L}^*_{\mathcal{S}}(t)$ is the estimated $L^*_{\mathcal{S}}$ in (3.11) using t exploration samples.

Since $m^*_{\mathcal{S}}$ scales linearly in B and $m(B;\omega)$ diverges as $B \to \infty$, there exists a sufficiently large $B(\delta;\omega)$ such that for $B > B(\delta;\omega)$,

$$\min_{\mathcal{S}\subseteq\{1:n\}} m^*_{\mathcal{S}} > 4T(\delta;\omega) \tag{A.9}$$

$$t_{q(B)} = m(B;\omega) > 4T(\delta;\omega). \tag{A.10}$$

Consider $q' < q(B)$ that satisfies $t_{q'-1} < T(\delta;\omega) \leq t_{q'}$. Such a q' always exists due to (A.10), and satisfies

$$t_{q'} \overset{\text{(A.5)}}{\leq} 2t_{q'-1} < 2T(\delta;\omega) \overset{\text{(A.9)}}{\leq} \frac{1}{2} \min_{\mathcal{S}\subseteq\{1:n\}} m^*_{\mathcal{S}} \overset{\text{(A.8)},\delta<1/2}{\leq} \widehat{m}^*_{\mathcal{S}}(t_{q'};\omega).$$

This inequality tells us that in the q'-th loop iteration, for all $\mathcal{S} \subseteq \{1:n\}$, the corresponding estimated optimal exploration rate is larger than the current exploration rate. In this case,

$$\widehat{L}_{\mathcal{S}}(t_{q'} \vee \widehat{m}^*_{\mathcal{S}}(t_{q'};\omega); t_{q'}) = \widehat{L}_{\mathcal{S}}(\widehat{m}^*_{\mathcal{S}}(t_{q'};\omega); t_{q'}) = \widehat{L}^*_{\mathcal{S}}(t_{q'}) \qquad \forall \mathcal{S} \subseteq \{1:n\}.$$

This, along with (A.7) and (A.8), tells us that $\mathcal{S}^*$ is the estimated optimal model in the current step, and more exploration is needed.

To see what $t_{q'+1}$ should be, we consider two separate cases. If $2t_{q'} \leq \widehat{m}^*_{\mathcal{S}^*}(t_{q'};\omega)$, then

$$T(\delta;\omega) < t_{q'+1} = 2t_{q'} \leq \widehat{m}^*_{\mathcal{S}^*}(t_{q'};\omega) \leq (1+\delta)m^*_{\mathcal{S}^*},$$

which implies

$$(1-\delta)m^*_{\mathcal{S}^*} \overset{\text{(A.8)}}{\leq} t_{q'+1} \vee \widehat{m}^*_{\mathcal{S}^*}(t_{q'+1};\omega) \leq (1+\delta)m^*_{\mathcal{S}^*}. \tag{A.11}$$

If $t_{q'} \leq \widehat{m}^*_{\mathcal{S}^*}(t_{q'};\omega) < 2t_{q'}$, then

$$t_{q'+1} = \left\lceil \frac{t_{q'} + \widehat{m}^*_{\mathcal{S}^*}(t_{q'};\omega)}{2} \right\rceil \leq \widehat{m}^*_{\mathcal{S}^*}(t_{q'};\omega) \overset{\text{(A.8)}}{\leq} (1+\delta)m^*_{\mathcal{S}^*},$$

which also implies (A.11). But (A.11) combined with (A.7) and (A.8) implies that $\mathcal{S}^*$ is again the estimated optimal model in the $(q'+1)$-th loop iteration. Applying the above argument

inductively proves $\mathcal{S}(B) = \mathcal{S}^*$, i.e. (5.14b). Note (A.11) holds true until the algorithm terminates, which combined with the termination criteria $t_{q(B)} \geq \widehat{m}^*_{\mathcal{S}^*}(t_{q(B)}; \omega) \geq (1-\delta) m^*_{\mathcal{S}^*}$ implies

$$1 - \delta \leq \frac{m(B;\omega)}{m^*_{\mathcal{S}^*}} = \frac{t_{q(B)}}{m^*_{\mathcal{S}^*}} \leq 1 + \delta.$$

(5.14a) now follows by noting that δ can be set arbitrarily small.

Finally, let $\widetilde{F}'(\boldsymbol{x}; B)$ be chosen as in (5.10) with $\mathcal{S} = \mathcal{S}^*$, $m = m^*_{\mathcal{S}^*}$ and $N_{\mathcal{S}} = (B - c_{\text{epr}} m^*_{\mathcal{S}^*})/c_{\mathcal{S}^*}$. Note both $m, N_{\mathcal{S}}$ are deterministic and diverge as $B \to \infty$. By the triangle inequality,

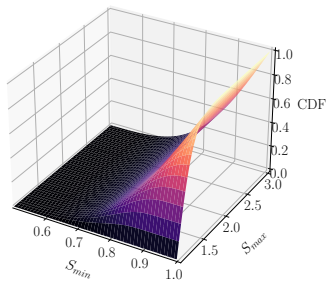
$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{F}(\boldsymbol{x}; B) - F_Y(\boldsymbol{x})| \leq \sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{F}(\boldsymbol{x}; B) - \widetilde{F}'(\boldsymbol{x}; B)| + \sup_{\boldsymbol{x}\in\mathbb{R}^d} |\widetilde{F}'(\boldsymbol{x}; B) - F_Y(\boldsymbol{x})|.$$

As $B \to \infty$, the first term on the right-hand side converges to 0 due to (5.14a) and (5.14b) in Theorem 5.7, and the second term on the right-hand side converges to 0 due to Theorem 5.10. This proves (5.14c).

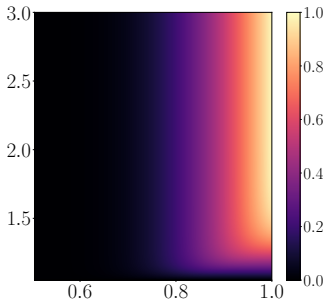
Appendix B. Additional numerical results.

B.1. Additional results for geometric Brownian motion in Subsection 6.2.

We present two figures that provide experimental results to supplement those presented in Subsection 6.2. A plot of the oracle CDF is visualized in Figure 11 (left, middle). The oracle model loss and exploration sample count are in Figure 11 (right). Figure 12 shows an instance of a heatmap of the absolute estimation errors of ECDF, cvMDL, and cvMDL-sorted when $B = 10^6$, providing supporting evidence that cvMDL is more accurate than ECDF on $\mathcal{T}$.



(a)



(b)

Model $\mathcal{S}$	$\gamma_{\mathcal{S}}$	$m^*_{\mathcal{S}}$
{1}	**11.3**	**613**
{2}	13.7	23.2
{3}	23.2	902
{1, 2}	12.2	596
{1, 3}	11.6	606
{2, 3}	14.2	790
{1, 2, 3}	12.4	581

(c)

Figure 11: Geometric Brownian motion. (a)-(b): Oracle CDF of $(S_{\min}, S_{\max})$ in the high-fidelity model computed using 10^5 Monte Carlo samples. (c): Oracle scaled loss $\gamma_{\mathcal{S}}$ (3.11) and the optimal exploration sample count $m^*_{\mathcal{S}}$ (3.11) for budget $B = 10^6$, computed using 50,000 samples.

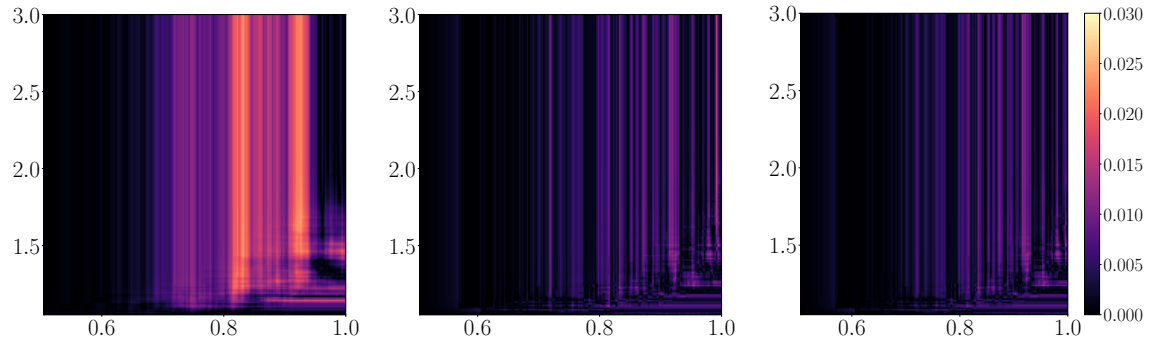


Figure 12: Geometric Brownian motion. An instance of absolute pointwise estimation errors of (a) ECDF, (b) cvMDL, and (c) cvMDL-sorted for budget $B = 10^6$.

B.2. Additional results for brittle fracture in Subsection 6.3. We present additional experimental details that supplement those presented in Subsection 6.3.

Recall the boundary value problem in (6.1) and (6.2). The boundary conditions on Γ_N and Γ_D are

$$\begin{aligned} [(1-\phi_d(\mathbf{x}))^2 + q]\nabla \cdot \boldsymbol{\sigma}(\mathbf{x}) &= \mathbf{v}_n, \quad \mathbf{x} \text{ on } \Gamma_N, \\ \mathbf{u}(\mathbf{x}) &= \mathbf{0}, \quad \mathbf{x} \text{ on } \Gamma_D, \\ \nabla \phi_d(\mathbf{x}) \cdot \mathbf{n} &= 0, \quad \mathbf{x} \text{ on } \Gamma_N, \end{aligned}$$

where $q \ll 1$, $\boldsymbol{\sigma}(\mathbf{x}) = \dfrac{\partial \Psi(\boldsymbol{\epsilon}(\mathbf{x}))}{\partial \boldsymbol{\epsilon}(\mathbf{x})}$ is the Cauchy stress tensor, and $\Psi(\boldsymbol{\epsilon}(\mathbf{x})) = \dfrac{1}{2}\lambda(\mathrm{tr}(\boldsymbol{\epsilon}(\mathbf{x})))^2 + \mu \mathrm{tr}(\boldsymbol{\epsilon}(\mathbf{x})^2)$ is the elastic energy density with μ and λ the Lamé constants, i.e.,

$$\lambda = \frac{\nu\kappa}{(1+\nu)(1-2\nu)}, \qquad \mu = \frac{\kappa}{2(1+\nu)}$$

with Young's modulus κ and Poisson's ratio ν, and $\boldsymbol{\epsilon}(\mathbf{x}) = \dfrac{1}{2}\left[\nabla \mathbf{u}(\mathbf{x}) + \nabla \mathbf{u}(\mathbf{x})^\intercal\right]$ is the small strain tensor. In (6.2) the history variable $H(\mathbf{x})$ is defined as:

$$H(\mathbf{x}) = \begin{cases} \Psi(\boldsymbol{\epsilon}(\mathbf{x})), & \Psi(\boldsymbol{\epsilon}(\mathbf{x})) < H_i(\mathbf{x}) \\ H_i(\mathbf{x}), & \text{otherwise} \end{cases}, \quad i = 1, 2, \ldots, n,$$

where $H_i(\mathbf{x})$ is the strain energy computed at ith step of the discretized load, which corresponds to the iterative solver stage $\bar{v}_i \cdot e_2$, with $\bar{v}_i \in [0, \bar{v}]$.

REFERENCES

[1] H. Amor, J.-J. Marigo, and C. Maurini, *Regularized formulation of the variational brittle fracture with unilateral contact: Numerical experiments*, Journal of the Mechanics and Physics of Solids, 57 (2009), pp. 1209–1229.

[2] E. Andreassen, A. Clausen, M. Schevenels, B. S. Lazarov, and O. Sigmund, *Efficient topology optimization in matlab using 88 lines of code*, Structural and Multidisciplinary Optimization, 43 (2011), pp. 1–16.

[3] Q. Ayoul-Guilmard, S. Ganesh, S. Krumscheid, and F. Nobile, *Quantifying uncertain system outputs via the multi-level Monte Carlo method–distribution and robustness measures*, International Journal for Uncertainty Quantification, 13 (2023).

[4] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, *Learnability and the vapnik-chervonenkis dimension*, Journal of the ACM (JACM), 36 (1989), pp. 929–965.

[5] S. Bobkov and M. Ledoux, *One-dimensional empirical measures, order statistics, and kantorovich transport distances*, Memoirs of the American Mathematical Society, 261 (2019).

[6] B. Bourdin, G. A. Francfort, and J.-J. Marigo, *Numerical experiments in revisited brittle fracture*, Journal of the Mechanics and Physics of Solids, 48 (2000), pp. 797–826.

[7] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media, 2011.

[8] M. Croci, K. Willcox, and S. Wright, *Multi-output multilevel best linear unbiased estimators via semidefinite programming*, Computer Methods in Applied Mechanics and Engineering, 413 (2023), p. 116130.

[9] I.-G. Farcas, *Context-aware model hierarchies for higher-dimensional uncertainty quantification*, PhD thesis, Technische Universität München, 2020.

[10] I.-G. Farcas, B. Peherstorfer, T. Neckel, F. Jenko, and H.-J. Bungartz, *Context-aware learning of hierarchies of low-fidelity models for multi-fidelity uncertainty quantification*, Computer Methods in Applied Mechanics and Engineering, 406 (2023), p. 115908.

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

[12] M. B. Giles, T. Nagapetyan, and K. Ritter, *Multilevel Monte Carlo approximation of distribution functions and densities*, SIAM/ASA Journal on Uncertainty Quantification, 3 (2015), pp. 267–295.

[13] M. B. Giles, T. Nagapetyan, and K. Ritter, *Adaptive multilevel Monte Carlo approximation of distribution functions*, arXiv preprint arXiv:1706.06869, (2017).

[14] P. Glasserman, *Monte Carlo methods in financial engineering*, vol. 53, Springer, 2004.

[15] P. Glasserman, P. Heidelberger, and P. Shahabuddin, *Efficient Monte Carlo methods for value-at-risk*, Mastering Risk, 2 (2000), pp. 5–18.

[16] A. A. Gorodetsky, G. Geraci, M. S. Eldred, and J. D. Jakeman, *A generalized approximate control variate framework for multifidelity uncertainty quantification*, Journal of Computational Physics, 408 (2020), p. 109257.

[17] A. Gruber, M. Gunzburger, L. Ju, and Z. Wang, *A multifidelity monte carlo method for realistic computational budgets*, Journal of Scientific Computing, 94 (2023), p. 2.

[18] M. Heinkenschloss, B. Kramer, and T. Takhtaganov, *Adaptive reduced-order model construction for conditional value-at-risk estimation*, SIAM/ASA Journal on Uncertainty Quantification, 8 (2020), pp. 668–692.

[19] M. Heinkenschloss, B. Kramer, T. Takhtaganov, and K. Willcox, *Conditional-value-at-risk estimation via reduced-order models*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 1395–1423.

[20] T. C. Hesterberg and B. L. Nelson, *Control variates for probability and quantile estimation*, Management Science, 44 (1998), pp. 1295–1312.

[21] S. Krumscheid and F. Nobile, *Multilevel Monte Carlo approximation of functions*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 1256–1293.

[22] T. Lattimore and C. Szepesvári, *Bandit algorithms*, Cambridge University Press, 2020.

[23] D. Lu, G. Zhang, C. Webster, and C. Barbier, *An improved multilevel monte carlo method for estimating probability distribution functions in stochastic oil reservoir simulations*, Water resources research, 52 (2016), pp. 9642–9660.

[24] A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative risk management: concepts, techniques and tools-revised edition*, Princeton University press, 2015.
[25] S. Natarajan and R. K. Annabattula, *A FEniCS implementation of the phase field method for quasi-static brittle fracture*, Frontiers of Structural and Civil Engineering, 13 (2019), pp. 380–396.
[26] A. B. Owen, *Monte Carlo theory, methods and examples*, Stanford, 2013.
[27] B. Peherstorfer, *Multifidelity monte carlo estimation with adaptive low-fidelity models*, SIAM/ASA Journal on Uncertainty Quantification, 7 (2019), pp. 579–603.
[28] B. Peherstorfer, B. Kramer, and K. Willcox, *Multifidelity preconditioning of the cross-entropy method for rare event simulation and failure probability estimation*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 737–761.
[29] 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.
[30] E. Qian, B. Peherstorfer, D. O'Malley, V. V. Vesselinov, and K. Willcox, *Multifidelity Monte Carlo estimation of variance and sensitivity indices*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 683–706.
[31] C. R. Rao, C. R. Rao, M. Statistiker, C. R. Rao, and C. R. Rao, *Linear statistical inference and its applications*, vol. 2, Wiley New York, 1973.
[32] D. Schaden and E. Ullmann, *On multilevel best linear unbiased estimators*, SIAM/ASA Journal on Uncertainty Quantification, 8 (2020), pp. 601–635.
[33] D. Schaden and E. Ullmann, *Asymptotic analysis of multilevel best linear unbiased estimators*, SIAM/ASA Journal on Uncertainty Quantification, 9 (2021), pp. 953–978.
[34] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge University Press, 2018.
[35] Y. Xu, V. Keshavarzzadeh, R. M. Kirby, and A. Narayan, *A bandit-learning approach to multifidelity approximation*, SIAM Journal on Scientific Computing, 44 (2022), pp. A150–A175.
[36] Y. Xu and A. Narayan, *Budget-limited distribution learning in multifidelity problems*, Numerische Mathematik, 153 (2023), pp. 171–212.