Budget-limited distribution learning in multifidelity problems*

Yiming Xu† and Akil Narayan†

Abstract. Multifidelity methods are widely used for statistical estimation of quantities of interest (QoIs) in uncertainty quantification using simulation codes of differing costs and accuracies. Many methods approximate numerical-valued statistics that represent only limited information of the QoIs. In this paper, we introduce a semi-parametric approach that aims to effectively describe the distribution of a scalar-valued QoI in the multifidelity setup. Under a linear model hypothesis, we propose an exploration-exploitation strategy to reconstruct the full distribution of a scalar-valued QoI using samples from a subset of low-fidelity regressors. We derive an informative asymptotic bound for the mean 1-Wasserstein distance between the estimator and the true distribution, and use it to adaptively allocate computational budget for parametric estimation and non-parametric reconstruction. Assuming the linear model is correct, we prove that such a procedure is consistent, and converges to the optimal policy (and hence optimal computational budget allocation) under an upper bound criterion as the budget goes to infinity. A major advantage of our approach compared to several other multifidelity methods is that it is automatic, and its implementation does not require a hierarchical model setup, cross-model information, or *a priori* known model statistics. Numerical experiments are provided in the end to support our theoretical analysis.

Key words. multifidelity, Wasserstein distance, sequential decision-making, empirical measure, linear regression

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

1. Introduction.

Estimation of output quantities of interest (QoIs) from complex and large-scale simulations is an important task in many areas of computational science. A concrete example is a forward uncertainty quantification setup, where the QoI is an output of a physical system subject to modeled uncertainty, and the goal is to identify the typical behavior of the QoI by computing its expectation. A universal solution for this specific task is through Monte Carlo (MC) simulation [12], which in practice requires many repeated evaluations of an accurate forward model, and can be computationally infeasible for expensive models.

A modern collection of approaches for addressing this computational challenge is multifidelity methods [22]. Instead of operating on a single (high-fidelity) model alone, multifidelity methods combine several models of different accuracies and costs to accelerate computation. The lower fidelity models used in multifidelity methods typically arise from simplification or reduction of the high-fidelity model, and thus are cheaper but less accurate. However, they typically contain information which, if utilized properly, can contribute to characterizing QoIs.

A prototypical example of a multifidelity method is the multilevel approach [7, 21, 10], which approximates the expectation of a scalar-valued QoI given by the high-fidelity model. Leveraging a telescoping sum using hierarchical models, multilevel estimators make use of cross-model correlations and typically attain smaller variance compared to a single-model MC

*Submitted to the editors.

Funding: Y. Xu and A. Narayan are partially supported by NSF DMS-1848508. A. Narayan is partially supported by AFOSR award FA9550-20-1-0338.

†Department of Mathematics, and Scientific Computing and Imaging (SCI) Institute, University of Utah (yxu@math.utah.edu, akil@sci.utah.edu).

estimator. Recent work has introduced a uniform perspective for many existing methods within the multilevel framework, and provided a way to realize the optimal variance reduction among all linear unbiased estimators [28, 27]. Multilevel estimators are considered universal in the sense that they rely only on correlation, which is used as an input for the estimator construction. A similar approach has recently been developed in [33] which assumes a linear model assumption but does *not* require *a priori* knowledge of correlation statistics.

Since multifidelity methods have been so successfully applied to parametric estimation of QoIs, it is natural to ask if it is possible to extend the same technique to also characterize their distributions, or equivalent statistics such as characteristic functions. This question has been studied in recent works [8, 9, 18, 16], where the major application under consideration is solving stochastic differential equations, and strong coupling assumptions and hierarchical structure are assumed for models of different fidelities. For more general (i.e., non-hierarchical) mutlifidelity setups, paradigms for efficiently learning distributions of QoIs are nascent.

Learning a distribution is possible when independent and identically distributed (i.i.d.) samples are available: Given enough samples, universal approaches using empirical measures estimate the true distribution [29]. However, the general nonparametric nature of this approach is balanced by rather slow convergence rates, which considerably limit direct usage of this approach in applications where sampling is costly. Alternative parametric methods limit the space of expressible distributions, but enable use of classical tools such as maximum likelihood estimation to accelerate estimation. In practice, models are often so complex that it is difficult to identify an appropriately expressive parametric family.

1.1. Contributions of this paper. In this paper, we adopt a semi-parametric approach based on ideas from multifidelity methods. In particular, we assume that the high-fidelity model and the lower fidelity surrogates interact in a way that can be described using a parametric framework, but we do not make specific assumptions on the distribution of the model outputs. (This is different from the commonly used semi-parametric framework in regression analysis [26].) In an exploration phase, we learn the interaction between the high and lower fidelity models via parametric strategies, and we construct a linear regression emulator for the high-fidelity output using a selection of the inexpensive low-fidelity models. Once this parametric emulator is computed, our exploitation phase utilizes the nonparametric approach of empirical measures to approximate the high-fidelity distribution by collecting a sufficient number of low-fidelity samples that are passed to the learned emulator. Such an approach leverages models of different fidelities and costs as well as various statistical procedures to produce an efficient estimator for the unknown distribution, which is difficult and costly to estimate directly. Our procedure does not require an established hierarchy of models (only a high-fidelity model must be identified), and does not require *a priori* knowledge of any model or cross-model statistics.

In summary, our contributions in this article are threefold:

- We introduce a semi-parametric formulation of the multifidelity problem for the budget-limited learning of distributions of scalar-valued QoIs, and propose a general exploration-exploitation strategy.
- We define a loss function for the estimator produced by an exploration-exploitation strategy using Wasserstein metrics, and derive an asymptotically informative and es-

timable upper bound for the estimator error.
- We utilize the upper bound to design a consistent and adaptive algorithm, AETC-d, which optimally balances exploration and exploitation along every trajectory under a sufficiently large budget.

We numerically investigate the proposed algorithm on several datasets, and demonstrate its consistency when the model assumptions hold. We also discuss procedures that can be used to mitigate the the impact of model misspecification, when our assumed linear regression assumptions are violated. Our methodology enjoys great generality as it does not require any particular knowledge of the models, i.e., nested structure or specific coupling assumptions to guarantee convergence. The implementation requires only the identification of a trusted high-fidelity model, the ability to query the models themselves, and a cost estimate of each model relative to the high-fidelity model.

The rest of the paper is organized as follows. In Section 2, we set up the budget-limited distribution learning problem. In Section 3, we briefly review results of convergence of empirical measures under Wasserstein metrics, and prove a required technical result regarding the difference between the cumulative distribution functions of linear sketches of a jointly sub-exponential random vector. In Section 4, we propose an exploration-exploitation strategy for distribution learning, and derive an asymptotically informative upper bound for the mean 1-Wasserstein error of the estimator. We then utilize this upper bound in Section 5 to devise an adaptive algorithm, AETC-d, and establish a trajectory-wise optimality result for it. In Section 6, we provide a detailed numerical study of the AETC-d algorithm, investigating consistency, model misspecification, and optimality of exploration rates. In Section 7, we conclude by summarizing the main results in the paper. Several technical details and proofs are collected in Appendix 8.

X_i	output of the i-th low-fidelity model
Y	output of the high-fidelity model
n	the number of low-fidelity models
m	exploration rate
B	the total budget
S	the model index set
$(c_i)_{i=0}^n / c_{\text{epr}} / c_{\text{ept}}(S)$	cost parameters/unit exploration cost/unit exploitation cost for model S
N_S	exploitation rate for model S
Z_S	the design matrix for model S
β_S	the coefficient vector for model S
σ_S^2	the model variance for model S
F_Y	the CDF of Y

Table 1
Notation used throughout this article.

2. Problem setup.

2.1. Notation. Let $Y, X_1, \ldots, X_n \in \mathbb{R}$ be scalar-valued random outputs (QoIs) associated with the high-fidelity model and n low-fidelity surrogates, respectively. Let c_0 and c_i ($i \in [n]$) be the respective cost of sampling Y and X_i. The costs are assumed known and deterministic.

No additional assumptions about the accuracy or costs of X_i relative to those of X_{i+1} are assumed in the following discussion. In particular, the index i does not represent an ordering based on cost, accuracy, or hierarchy.

Many recent advances on the multifidelity methods centered around the efficient estimation of the expectation of scalar-valued QoIs associated with the high-fidelity model [7, 21, 10, 28, 27, 33]. Under appropriate correlation/cost conditions and for a fixed budget, the estimators from these methods are significantly more accurate than the classical MC estimator using i.i.d. samples of Y. However, for some applications, obtaining precise estimates for only statistics is not sufficient. A complete description of the randomness (the distribution) of a QoI is needed to quantify the uncertainty of interest. Letting $F_Y(y) = \mathbb{P}(Y \leq y)$ be the cumulative distribution function (CDF) of Y, we wish to find an efficient estimate for F_Y instead of only certain functionals (statistics) of it. To make use of the low-fidelity models to this end, assumptions on cross-model correlations are not enough for this purpose. As opposed to imposing strong hierarchical assumptions on the models as in [8], we introduce an alternative parametric assumption on the relationship $X_i, i \in [n]$ and Y such that efficient estimation of $F_Y(y)$ is possible through X_i's.

2.2. Linear regression. A simple yet useful assumption relating X_i and Y is through linear regression. For any $S \subset [n]$, we assume that

$$Y = X_S^T \beta_S + \varepsilon_S \qquad\qquad X_S = \left(1, (X_i)_{i\in[S]}\right)^T, \tag{2.1}$$

where ε_S is a centered random variable (noise) with variance σ_S^2 and is independent of X_S. Under this assumption, (2.1) provides an alternative way to simulate Y using only samples of X_S and a noise generator for ε_S. Using samples generated under a fixed budget, this could potentially lead to an estimator for F_Y with better accuracy compared to directly sampling Y alone. More details about the construction of the estimator will be discussed in Section 4.1.

Even though (2.1) considers only linear interactions, it is not overly restrictive since additional regressors that are nonlinear functions of the models can be added. For many applications of interest, the high-fidelity model approximately satisfies (2.1) with a sufficient number of appropriately chosen regressors. Nevertheless, identifying an optimal set of such regressors is often problem-dependent, which is beyond the scope of this paper.

We close this section by noting that we do require some additional technical assumptions (see Assumptions 4.1, 4.2, 4.3) about the models X_i, Y and the noise ε_S for some of our theoretical results. Our numerical results investigate the performance of our algorithm when some of these assumptions are violated.

3. Distribution metrics and related results. Before directly addressing the multifidelity problem, this section provides some necessary technical discussion regarding empirical distributions, metrics, and error bounds.

3.1. Distance between distributions. The goal of distribution learning is to approximate a probability measure on $\mathbb{R}$. In this section, we discuss metrics to measure the discrepancy between F_Y and an estimated distribution. For two Borel probability measures μ, ν on $\mathbb{R}$, numerous metrics are available to measure discrepancy [6], such as the Kolmogorov distance,

Wasserstein distances, Kullback-Leibler divergence, etc. In this article, we will focus on the Wasserstein metrics, which are defined below.

Definition 3.1 (p-th Wasserstein distance). *Let $1 \le p \le \infty$. The p-th Wasserstein distance between two Borel probability measures μ, ν on $\mathbb{R}$ is defined as*

$$W_p(\mu,\nu) = \inf_{\pi} \mathbb{E}_\pi \left[|x-y|^p\right]^{1/p},$$

where infimum is taken over all Borel probability measures π on $\mathbb{R}^2$ with marginals satisfying $\pi_x \equiv \mu, \pi_y \equiv \nu$.

Intuitively, $W_p^p(\mu,\nu)$ corresponds to the minimal amount of work needed to transform μ to ν, with the cost function given by the p-th power of the moving distance. Hence it is frequently labeled as the optimal transport distance. A comprehensive study of the subject can be found in [31]. In the following discussion we are mostly concerned with $p = 1$.

Wasserstein distances are hard to compute in general. But when the metric space is the real line equipped with the Borel algebra, explicit formulas exist using inverse CDFs [4]:

$$W_p^p(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(t) - F_\nu^{-1}(t)\right|^p dt,$$

where $F_\mu^{-1}(t) := \inf\{x \in \mathbb{R} : F_\mu(x) \ge t\}$ is the inverse CDF of μ. When $p = 1$, $W_1(\mu,\nu)$ is the $L^1(\mathbb{R})$-norm of $F_\mu(t) - F_\nu(t)$:

$$W_1(\mu,\nu) = \int_0^1 \left|F_\mu^{-1}(t) - F_\nu^{-1}(t)\right| dt = \int_{\mathbb{R}} |F_\mu(t) - F_\nu(t)|\, dt. \tag{3.1}$$

A classical result in optimal transport is the Kantorovich-Rubinstein duality [31], which provides an alternative characterization for $W_1(\mu,\nu)$ using Lipschitz test functions:

$$W_1(\mu,\nu) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left| \int f d\mu - \int f d\nu \right|, \tag{3.2}$$

where $\|\cdot\|_{\mathrm{Lip}}$ is the Lipschitz constant. Assuming F_μ has a uniformly bounded density f_μ, one can bound the Kolmogorov distance between μ and ν by $W_1(\mu,\nu)$ using (3.2) [5]:

$$d_K(\mu,\nu) := \sup_{x\in\mathbb{R}} |F_\mu(x) - F_\nu(x)| \le 2\sqrt{\|f_\mu\|_\infty W_1(\mu,\nu)}.$$

In the situations of this article, at least one of the measures under comparison has a uniformly bounded density. Thus, in what follows we will work with the Wasserstein metrics, i.e., the W_1-metric.

3.2. Convergence of empirical measures. Our analysis in Section 4 relies on sharp pre-asymptotic convergence rates of one-dimensional empirical measures under the mean W_p metric. Given i.i.d. samples of $\mu : Z_1, \ldots, Z_N$, the associated empirical measure is

$$\mu_N = \frac{1}{N} \sum_{i\in[N]} \delta_{Z_i}. \tag{3.3}$$

The question is to quantify the average convergence rate for $\mathbb{E}[W_p^p(\mu_N,\mu)]^{1/p}$ at fixed N. A comprehensive analysis quantifying $\mathbb{E}[W_p^p(\mu_N,\mu)]^{1/p}$ for fixed N can be found in [2]. Here we only collect relevant results to be used later.

Theorem 3.2 ([2]). *Suppose μ is a Borel probability measure on $\mathbb{R}$ with finite $(2+\delta)$-th moment for some $\delta > 0$, i.e., $\int |x|^{2+\delta} d\mu(x) < \infty$. Let μ_N be the empirical measure defined in (3.3). Then for every $N \geq 1$,*

$$\frac{J_0(\mu)}{\sqrt{2}} \frac{1}{\sqrt{N}} \leq \mathbb{E}[W_1(\mu_N,\mu)] \leq \frac{J_1(\mu)}{\sqrt{N}}, \tag{3.4}$$

where

$$J_0(\mu) := \int_{\mathbb{R}} F_\mu(x)(1-F_\mu(x))dx \qquad J_1(\mu) := \int_{\mathbb{R}} \sqrt{F_\mu(x)(1-F_\mu(x))}dx. \tag{3.5}$$

The moment assumption on μ ensures that $J_1(\mu) < \infty$ so that the upper bound in (3.4) is nonvacuous. For completeness, we give the proof here:

Proof of Theorem 3.2. The upper bound follows from (3.1) and Jensen's inequality:

$$\mathbb{E}[W_1(\mu,\mu_N)] = \int_{\mathbb{R}} \mathbb{E}[|F_\mu(x) - F_{\mu_N}(x)|]dx \leq \int_{\mathbb{R}} \mathbb{E}[|F_\mu(x) - F_{\mu_N}(x)|^2]^{1/2} dx = \frac{J_1(\mu)}{\sqrt{N}},$$

where the last equality follows by noting that for fixed x, $NF_{\mu_N}(x) \sim \text{Binomial}(N, F_\mu(x))$, i.e., $F_\mu(x) - F_{\mu_N}(x) = N^{-1}\sum_{i\in[N]} \xi_i$, where $\xi_i = F_\mu(x) - \mathbf{1}_{X_i \leq x}$. For the lower bound, let ε_i be i.i.d. symmetric Bernoulli random variables which are independent of ξ_i, and $\boldsymbol{\xi}_N = (|\xi_i|)_{i=1}^N$. Applying the symmetrization technique [30] combined with Khinchine's inequality [11], we have

$$\mathbb{E}[|F_\mu(x) - F_{\mu_N}(x)|] = \frac{1}{N}\mathbb{E}\left[\left|\sum_{i\in[N]} \xi_i\right|\right] \geq \frac{1}{2N}\mathbb{E}\left[\left|\sum_{i\in[N]} \varepsilon_i\xi_i\right|\right] \geq \frac{\mathbb{E}\left[\|\boldsymbol{\xi}_N\|_2\right]}{2\sqrt{2}N} \geq \frac{\|\mathbb{E}[\boldsymbol{\xi}_N]\|_2}{2\sqrt{2}N}.$$

Integrating against x yields the desired lower bound. ∎

A refinement of the result in Theorem 3.2 based on a two-sided bound is given below:

Theorem 3.3 ([2]). *Under the same assumptions as in Theorem 3.2, for every $N \geq 1$,*

$$\frac{T_N}{1250\sqrt{N}} \leq \mathbb{E}[W_1(\mu_N,\mu)] \leq \frac{T_N}{\sqrt{N}}, \tag{3.6}$$

where

$$T_N = \int_{I_N} \sqrt{F_\mu(x)(1-F_\mu(x))}dx + 2\int_{I_N^c} F_\mu(x)(1-F_\mu(x))dx,$$
$$I_N = \left[F_\mu^{-1}\left(\frac{1}{2} - \frac{1}{2}\sqrt{\frac{N-1}{N}}\right), F_\mu^{-1}\left(\frac{1}{2} + \frac{1}{2}\sqrt{\frac{N-1}{N}}\right)\right].$$

Note that for large N, the interval I_N covers most of the mass of μ, implying that the upper bound in (3.4) is asymptotically tight.

Similar nonasymptotic results for $\mathbb{E}[W_p^p(\mu_N,\mu)]^{1/p}$ when $p>1$ can be found in [2, Corollary 5.10]. In those cases, the sufficient and necessary condition to achieve the optimal convergence rate has a more subtle dependence on the moments of a distribution [2, Corollary 6.14].

3.3. A CDF estimate for marginals of a random vector. We have seen in (3.1) that if two one-dimensional random variables have 'similar' CDFs, then they are close in the W_1-metric. In this section, we will prove a type of converse of this statement. In particular, we show that two close marginals of a jointly sub-exponential vector have similar CDFs.

Let $X\in\mathbb{R}^k$ be a *jointly sub-exponential* random vector, i.e., $\sup_{\|a\|_2=1}\|a^TX\|_{\psi_1}<\infty$, where $\|\cdot\|_{\psi_1}$ is the 1-Orlicz norm [30]. Equivalently, there exists an $C_1>0$ such that

$$\sup_{\|a\|_2=1}\mathbb{P}\left(|a^TX|>x\right)\le 2\exp(-x/C_1)\qquad\forall x\ge 0. \tag{3.7}$$

Consider two marginals Z_1,Z_2 of X:

$$Z_1=a_1^TX\qquad Z_2=a_2^TX, \tag{3.8}$$

where $a_1,a_2\in\mathbb{R}^k$. Denote by $F_1(x)$ and $F_2(x)$ the CDFs of Z_1 and Z_2, respectively. If $\|a_1-a_2\|_2$ is small, then $F_1(x)\approx F_2(x)$ for every fixed $x\in\mathbb{R}$. Since (3.7) ensures that the CDFs of both Z_1 and Z_2 quickly decay to zero as $|x|\to\infty$, the distance between $F_1(x)$ and $F_2(x)$ can be bounded in a uniform sense.

Theorem 3.4. *Let $X\in\mathbb{R}^k$ be a sub-exponential random vector satisfying (3.7), and Z_1,Z_2 be the marginals defined in (3.8). Denote by*

$$\delta:=\|a_1-a_2\|_2<1\qquad C_2:=\sqrt{\|a_1\|_2^2+\|a_2\|_2^2}<\infty.$$

Suppose the CDFs of Z_1 and Z_2, $F_1(x)$ and $F_2(x)$, are globally Lipschitz continuous; i.e., $\max\{\|F_1\|_{Lip},\|F_2\|_{Lip}\}\le C_3$ for some $C_3<\infty$. Then, for any $p\in(0,\infty)$, there exists a constant $K>0$ depending only on $C_i, i\in[3]$, such that

$$\|F_1(x)-F_2(x)\|_{L^p(\mathbb{R})}\le K\delta\log^{\frac{p+1}{p}}(1/\delta)\xrightarrow{\delta\downarrow 0}0. \tag{3.9}$$

Proof. See Appendix A. ∎

4. Empirical CDF estimator under linear regression. In this section, we develop a rigorous framework for efficient estimation of $F_Y(y)$ under the linear regression assumption (2.1). We begin with the basic idea assuming oracles statistics are available.

4.1. Basic idea. Suppose that (2.1) holds and is known *a priori*. Then for fixed $S\subset[n]$, one may expend a given and fixed total budget B to sample X_S and build an empirical CDF estimator for F_Y:

$$\frac{1}{N}\sum_{i\in[N]}\mathbf{1}_{Z_i\le y}\qquad Z_i\overset{\text{iid}}{\sim}X_S^T\beta_S+\varepsilon_S, \tag{4.1}$$

where

$$N = \left\lfloor \frac{B}{c_{\text{ept}}(S)} \right\rfloor \qquad c_{\text{ept}}(S) = \sum_{i \in S} c_i. \tag{4.2}$$

As opposed to the direct construction of the empirical CDF from i.i.d. samples of Y, the emulator (4.1) will admit a much larger sampling rate under a fixed budget if $c_{\text{ept}}(S) \ll c_0$, which can substantially accelerate convergence. However, practical issues exist:

- Oracle information of the parameters in (2.1) is often not available.
- The most effective model choice $S \subset [n]$ is not known without exploration.
- Simulating ε_S is difficult without making further assumptions.

We resolve the first two problems by taking a similar bandit-learning approach as in [33], with the loss function articulated in Section 4.2 and 4.3. For the third issue, we will assume that ε_S is Gaussian to make the estimation procedure more convenient and efficient. This assumption may be hard to verify/satisfy in practice. However, analyzing this simplified model provides valuable insight into the general case, and the resulting procedures are observed to enjoy certain robustness regarding the Gaussian assumption. (See Section 6.2 for numerical evidence.) In general one can always resort to nonparametric procedures when the normality assumption is severely violated or the exploration rate is large.

4.2. Exploration and exploitation. In this section, we introduce an exploration-exploitation strategy following the ideas in [33]. Since neither the best parametric model nor the corresponding coefficients are known, it is necessary to expend some effort (budget) to decide on S before committing to (4.1). We split the total budget into two parts, one for *exploration* and the other for *exploitation*. In the exploration stage, we collect m independent joint samples of all fidelities to estimate β_S and σ_S^2 for every $S \subset [n]$, based on which we decide the best model for exploitation. In the exploitation stage, we follow the decision made in the exploration phase and use (4.1) to construct an estimator for F_Y by plugging in the estimated coefficients.

Denote

$$\begin{aligned}
&\text{(exploration samples)} && X_{\text{epr},\ell} := (1, X_{1,\ell}, \cdots, X_{n,\ell}, Y_\ell)^T && \ell \in [m] \\
&\text{(design matrix for } S\text{)} && Z_S := \begin{pmatrix} 1, (X_{i,1})_{i\in S} \\ \vdots \\ 1, (X_{i,m})_{i\in S} \end{pmatrix} \in \mathbb{R}^{m\times(s+1)} \\
&\text{(exploration responses)} && Y_{\text{epr}} := (Y_1, \cdots, Y_m)^T,
\end{aligned}$$

where ℓ is the sampling index, and $N_S = \lfloor (B - c_{\text{epr}} m)/c_{\text{ept}}(S) \rfloor$ is the number of affordable samples for exploiting S, where $c_{\text{epr}} = \sum_{i=0}^n c_i$ is the cost for an exploration sample. For $m > |S| + 1$, β_S and σ_S^2 can be estimated using standard least-squares and the average of residuals squared, respectively. Assuming that Z_S has full column rank, then the estimators for β_S and σ_S^2 are given by

$$\widehat{\beta}_S = Z_S^\dagger Y_{\text{epr}} = \left(Z_S^T Z_S\right)^{-1} Z_S^T Y_{\text{epr}} \qquad \widehat{\sigma}_S^2 = \frac{1}{m - |S| - 1} \left\| Y_{\text{epr}} - Z_S \widehat{\beta}_S \right\|_2^2,$$

where the scaling constant in $\widehat{\sigma}_S^2$ ensures that this estimator is unbiased. The resulting empirical estimator from model S is

$$\widehat{F}_{Y,S}(y) := \frac{1}{N_S} \sum_{j\in[N_S]} \mathbf{1}_{A_j \le y} \qquad A_j \overset{\text{iid}}{\sim} Y' := X_S^T \widehat{\beta}_S + \mathcal{N}(0, \widehat{\sigma}_S^2). \tag{4.3}$$

To measure the average quality of (4.3) as an estimator for F_Y, we use the following mean 1-Wasserstein distance as the *loss function*:

$$L_S(m) := \mathbb{E}_{A_j,\varepsilon_S}\left[W_1(\widehat{F}_{Y,S}, F_Y)\right], \tag{4.4}$$

where the expectation averages out the *randomness in exploitation* as well as the *noise in exploration*. Note that the $L_S(m)$ defined in (4.4) is random due to the remaining randomness in Z_S, and one could alternatively define it by averaging all the randomness. But this would lead to the term $\mathbb{E}[(Z_S^T Z_S)^{-1}]$, which is difficult to analyze. Thus we will pursue (4.4) in this article and consider the case when $B \to \infty$.

Obtaining exact asymptotically equivalent expressions for (4.4) is difficult. To find computably informative substitutes, we compute sharp estimates for an upper bound of (4.4) in the next section.

4.3. Upper bounds. We start by writing (4.4) in two parts using the triangle inequality:

$$L_S(m) \le \mathbb{E}_{A_j,\varepsilon_S}\left[W_1(F_Y, F_{Y'})\right] + \mathbb{E}_{A_j,\varepsilon_S}\left[W_1(\widehat{F}_{Y,S}, F_{Y'})\right], \tag{4.5}$$

where $F_{Y'}$ is the CDF of Y' (see (4.3) for the definition of Y'). According to Sanov's theorem in large deviation theory [25, 3], the probability that empirical measures are 'closer' to a different distribution than the sampling distribution is exponentially small with respect to the sampling rate, which heuristically suggests that (4.5) is tight.

We provide some intuition for the bound (4.5). The first term measures the mean W_1 distance between $X_S^T\beta_S + \mathcal{N}(0,\sigma_S^2)$ and $X_S^T\widehat{\beta}_S + \mathcal{N}(0,\widehat{\sigma}_S^2)$, which depends on the accuracy of $\widehat{\beta}_S$ and $\widehat{\sigma}_S^2$, hence on the exploration rate m. The second term measures the mean convergence rate of empirical measures, which depends on the exploitation rate N_S. A good exploration-exploitation strategy (i.e., determination of m) will balance these quantities. We will now produce a computable asymptotic upper bound for (4.5) that can be used to find such an m. We formalize our needed assumptions:

Assumption 4.1. *For every $S \subset [S]$, $\Lambda_S := \mathbb{E}[X_S X_S^T]$ is invertible.*

Assumption 4.2. *$X = (X_1, \cdots, X_n)$ is jointly sub-exponential in the sense of (3.7). Moreover, denoting by $F_a(x)$ the CDF of $a^T X$ for $a \in \mathbb{R}^n$, we assume that $\sup_{\|a\|_2=1} \|F_a\|_{Lip} < \infty$.*

Assumption 4.3. *For $S \subset [n]$, the model noise ε_S is Gaussian, i.e., $\varepsilon_S \sim \mathcal{N}(0,\sigma_S^2)$.*

Lemma 4.4. *Under Assumption 4.1, 4.2 and 4.3, and given any $\delta > 0$, the following is true with probability 1: For large enough m (realization-dependent) and every $B > c_{epr} m$,*

$$\mathbb{E}_{A_j,\varepsilon_S}\left[W_1(F_Y, F_{Y'})\right] \le (1+\delta)\sigma_S \sqrt{\frac{s+2}{m}}, \tag{4.6a}$$

$$\mathbb{E}_{A_j,\varepsilon_S}\left[W_1(\widehat{F}_{Y,S}, F_{Y'})\right] \le (1+\delta)\frac{J_1(F_Y)}{\sqrt{N_S}}. \tag{4.6b}$$

The proof of Lemma 4.4 is given in Appendix B. We now state an immediate consequence, that L_S can be estimated by a more computable quantity that serves as an asymptotic upper bound. This fact will be used for algorithm design in Section 5.

Theorem 4.5. *Under Assumptions 4.1, 4.2 and 4.3, with probability* 1,

$$\limsup_{B,m\uparrow\infty} \frac{L_S(m)}{G_S(m)} \leq 1 \qquad\qquad G_S(m) := \sqrt{\frac{k_1(S)}{m}} + \sqrt{\frac{k_2(S)}{B - c_{epr}m}},$$

where

$$k_1(S) = \sigma_S^2 \cdot (s+2) \qquad\qquad k_2(S) = c_{ept}(S) J_1^2(F_Y).$$

Proof. Combining Lemma 4.4 and (4.5) yields that for any $\delta > 0$,

$$\limsup_{B,m\uparrow\infty} \frac{L_S(m)}{G_S(m)} \leq 1 + \delta \qquad\qquad a.s.$$

The proof is finished by sending $\delta \downarrow 0$. ∎

5. Algorithm. In this section, we first use the asymptotic upper bound G_S to analyze the optimal exploration rate for each exploitation choice $S \subset [n]$ in an exploration-exploitation policy, which allows us to find a deterministic strategy that explores and exploits optimally. Then we propose an adaptive procedure which, along each trajectory, resembles the best exploration-exploitation policy. Integer rounding effects defining N_S are subsequently ignored to simplify analysis.

5.1. Optimal exploration based on G_S. The quantity $G_S(m)$ defined in Theorem 4.5 provides a computable criterion to evaluate the model S as a simulator for Y. For fixed S, $G_S(m)$ is a strictly convex function of m in the domain, attaining its minimum at

$$m^*(S) = \frac{B}{c_{\text{epr}} + \left(\frac{c_{\text{epr}}^2 k_2(S)}{k_1(S)}\right)^{1/3}}, \tag{5.1}$$

with optimum value

$$G_S^* := G_S(m^*(S)) = \frac{\left[(c_{\text{epr}}k_1(S))^{1/3} + k_2(S)^{1/3}\right]^{3/2}}{\sqrt{B}} \propto \left[(c_{\text{epr}}k_1(S))^{1/3} + k_2(S)^{1/3}\right]^{3/2}. \tag{5.2}$$

(5.2) is the (asymptotic) minimum loss of exploiting model S with optimal exploration rate (5.1). The model with the smallest value G_S^* is considered the optimal model under the "upper bound criterion", which is our terminology for using G_S as a criterion. We assume the optimal model is unique and denoted by S_{opt}, i.e.,

$$S_{\text{opt}} = \arg\min_{S\subset[n]} \left[(c_{\text{epr}}k_1(S))^{1/3} + k_2(S)^{1/3}\right]^{3/2}. \tag{5.3}$$

We call the policy that spends $m^*(S_{\text{opt}})$ rounds on exploration and then selects model S_{opt} for exploitation a *perfect exploration-exploitation policy*. This is similar to the perfect uniform exploration policies in [33] where exact asymptotics of the loss function are used.

To investigate the performance of a perfect exploration-exploitation policy, we compare it to the empirical CDF estimator for Y using samples of Y only. Let $\widehat{F}_{Y,\text{dir}}(B)$ denote the empirical CDF of Y with the whole budget devoted to sampling Y. The number of admissible samples for $\widehat{F}_{Y,\text{dir}}(B)$ is B/c_0, which, combined with (3.4), implies the following lower bound for the mean W_1 distance between $\widehat{F}_{Y,\text{dir}}(B)$ and F_Y: For every $B > c_0$,

$$\mathbb{E}\left[W_1\left(\widehat{F}_{Y,\text{dir}}(B), F_Y\right)\right] \geq \sqrt{\frac{c_0 J_0^2(F_Y)}{2B}}. \tag{5.4}$$

Under the same budget, the average W_1 distance between F_Y and the estimator given by the perfect exploration-exploitation policy is asymptotically bounded by $G^*_{S_{\text{opt}}}$, as guaranteed by Theorem 4.5. This motivates us to introduce the ratio r between the lower bound in (5.4) and $G^*_{S_{\text{opt}}}$ as a measure for the efficiency of perfect exploration-exploitation policies relative to $\widehat{F}_{Y,\text{dir}}(B)$:

$$\begin{aligned} r := \frac{\sqrt{\frac{c_0 J_0^2(F_Y)}{2B}}}{G^*_{S_{\text{opt}}}} &= \sqrt{\frac{c_0 J_0^2(F_Y)}{2\left[(c_{\text{epr}} k_1(S_{\text{opt}}))^{1/3} + k_2(S_{\text{opt}})^{1/3}\right]^3}} \overset{\text{Jensen}}{\geq} \sqrt{\frac{c_0 J_0^2(F_Y)}{8(c_{\text{epr}} k_1(S_{\text{opt}}) + k_2(S_{\text{opt}}))}} \\ &\geq \frac{1}{\sqrt{8\left[\left(\frac{\sigma_{S_{\text{opt}}}(n+1)}{J_0(F_Y)}\right)^2 + \left(\sqrt{\frac{c_{\text{ept}}(S_{\text{opt}})}{c_0}}\frac{J_1(F_Y)}{J_0(F_Y)}\right)^2\right]}} = \frac{1}{4\max\{\kappa_0, \kappa_1\}}, \end{aligned}$$

where

$$\kappa_0 = \frac{\sigma_{S_{\text{opt}}}(n+1)}{J_0(F_Y)} \qquad \kappa_1 = \sqrt{\frac{c_{\text{ept}}(S_{\text{opt}})}{c_0}}\frac{J_1(F_Y)}{J_0(F_Y)}.$$

If $\max\{\kappa_0, \kappa_1\} \ll 1$, then $r \gg 1$. (This condition implies that the exploration is efficient and exploitation sampling rate is large.) In this case, the perfect exploration-exploitation policy is expected to demonstrate a superior performance over the empirical CDF estimator based only on the samples of Y.

5.2. An adaptive algorithm. Finding perfect exploration-exploitation policies requires evaluation of (5.1) and (5.2), which uses oracle information of model statistics such as σ_S^2 and $J_1^2(F_Y)$. These statistics are not available in practice but can be computed in an online fashion using exploration data. At each step $t \geq n+2$ in exploration, we define

$$\widehat{\sigma}_S^2(t) = \frac{1}{t - |S| - 1}\left\|Y_{\text{epr}}(t) - Z_S(t)\widehat{\beta}_S(t)\right\|_2^2 \qquad \widehat{J}_1(t) = J_1(F_{Y_{\text{epr}}(t)}(y)), \tag{5.5}$$

where the parameter t indicates that estimates/data are based on the first t rounds of exploration, and $F_{Y_{\text{epr}}(t)}(y)$ is the empirical CDF of Y based on the exploration samples $Y_{\text{epr}}(t)$.

Plugging (5.5) into G_S and (5.1) allows us to estimate the optimal loss for each model S at exploration step t:

$$\rho_S = \widehat{G}_S(t \vee \widehat{m}^*(S)), \tag{5.6}$$

where

$$\widehat{k}_1(S) = \widehat{\sigma}_S^2(t) \cdot (s+2), \qquad \widehat{k}_2(S) = c_{\text{ept}}(S)\widehat{J}_1(t) \tag{5.7}$$

$$\widehat{G}_S(t) = \sqrt{\frac{\widehat{k}_1(S)}{t}} + \sqrt{\frac{\widehat{k}_2(S)}{B - c_{\text{epr}}t}} \qquad \widehat{m}^*(S) = \frac{B}{c_{\text{epr}} + \left(\frac{c_{\text{epr}}^2 \widehat{k}_2(S)}{\widehat{k}_1(S)}\right)^{1/3}}. \tag{5.8}$$

(5.6) can be used to decide which model is optimal to exploit at time t, and the corresponding estimated optimal stopping time $\widehat{m}^*$ will indicate if more exploration is needed. The details are given in Algorithm 5.1, the adaptive Explore-Then-Commit algorithm for multifidelity distribution learning (AETC-d), where regularization parameters α_t are added to $\widehat{k}_2(S)$ to encourage exploration at the beginning (as is common in bandit learning algorithms). In our case, the regularization parameters are mostly used for theoretical analysis. In practice, we observe that setting them as a rapidly decreasing sequence (i.e., exponential decay) is often sufficient. An asymptotic analysis of Algorithm 5.1 will be given in the next section.

Algorithm 5.1 AETC algorithm for multifidelity distribution learning (AETC-d)

Input: B: total budget, c_i: cost parameters, $\alpha_t \downarrow 0$: regularization parameters
Output: An estimate for F_Y

compute the maximum exploration round $M = \lfloor B/c_{\text{epr}} \rfloor$
collect $(n+2)$ samples for exploration
while $n + 2 < t \leq M$ **do**
  for $S \subset [n]$ **do**
    compute $\widehat{k}_1(S)$ and $\widehat{k}_2(S)$ using (5.7), and $\widehat{k}_2(S) \leftarrow \widehat{k}_2(S) + \alpha_t$
    compute $\widehat{m}^*(S)$ and ρ_S as in (5.8) and (5.6)
  end for
  find the optimal model $S^* = \arg\min_{S \subset [n]} \rho_S$
  if $\widehat{m}^*(S^*) > t$ **then**
    take a new exploration sample, and set $t \leftarrow t + 1$
  else
    compute the estimate for F_Y using (4.3) with $S = S^*$, and set $t \leftarrow M + 1$
  end if
end while

5.3. Asymptotic performance of the AETC-d. Regularization parameters used in Algorithm 5.1 encourages $m \to \infty$ as the total budget $B \to \infty$. This, combined with the fact that all estimators used in the algorithm are strongly consistent and $\alpha_t \downarrow 0$, implies the following asymptotic performance guarantee for Algorithm 5.1:

Theorem 5.1. *Let $\alpha_t \downarrow 0$. Let $m(B)$ and $S(B)$ denote the exploration rate and selected model in Algorithm 5.1, respectively. Under Assumptions 4.1, 4.2 and 4.3, with probability* 1,

$$\lim_{B\uparrow\infty} \frac{m(B)}{m^*(S_{opt})} = 1 \qquad \lim_{B\uparrow\infty} S(B) = S_{opt}, \tag{5.9}$$

where S_{opt} is defined in (5.3).

Proof. See Appendix C. ∎

Remark 5.2. *Theorem 5.1 tells us that as the budget goes to infinity, almost every policy realization of Algorithm 5.1 resembles a perfect exploration-exploitation policy. This establishes a trajectory-wise optimality result for the AETC-d under the upper bound criterion. However, this does not imply that the loss associated with the AETC-d is also asymptotically bounded by $G^*_{S_{opt}}$, due to the adaptive selection of $m(B)$. In fact, m is assumed deterministic in the analysis in Section 4.3.*

As a consequence, we have the following consistency result for the CDF estimator produced by the AETC-d algorithm:

Corollary 5.3. *Denote the CDF estimator produced by Algorithm 5.1 as $\widehat{F}_{Y,aetcd}(B)$. Under Assumptions 4.1, 4.2 and 4.3, then with probability* 1,

$$\lim_{B\uparrow\infty} W_1\left(\widehat{F}_{Y,aetcd}(B), F_Y\right) = 0.$$

Proof. See Appendix D. ∎

6. Numerical experiments. In this section, we demonstrate the performance of the AETC-d (Algorithms 5.1) for multifidelity estimation of univariate distributions. We will focus on consistency, model misspecification, and optimality of exploration rates. The regularization parameter α_t is set as $\alpha_t = 4^{-t}$ in all simulations. Five methods will be considered for estimating $F_Y(y)$:

- (ECDF-Y): Empirical CDF estimator for F_Y based on the samples of Y only.
- (AETC-d): Algorithm 5.1.
- (AETC-d-em): A modification of Algorithm 5.1, where the noise in (4.3) is generated using the empirical measure of the residuals in exploration (i.e., bootstrapping).
- (AETC-d-no): A modification of Algorithm 5.1, where the noise in (4.3) is omitted.
- (AETC-d-q): A modification of Algorithm 5.1 using quantile regression. See Appendix E for a detailed description of the algorithm.

To evaluate results, we compute and report an empirical mean W_1 distance (error) between F_Y and the estimated CDF given by the algorithms over 200 samples. Since AETC-d produces random estimators due to the exploration phase, the experiment is repeated 100 times with the 5-50-95-quantiles recorded to measure this extra uncertainty.

6.1. Ishigami function. In this example, we investigate the performance of the proposed AETC-d algorithm (Algorithm 5.1) on a multifidelity algebraic system consisting of Ishigami functions [13]. We adopt a modified version of the setup in [23]. The high-fidelity model output corresponds to the following random variable:

$$Y = \sin Z_1 + a \sin^2 Z_2 + bZ_3^4 \sin Z_1 + cZ_4 + dZ_5, \tag{6.1}$$

where a, b, c, d are deterministic constants, and $Z_i, i \in [5]$ are independent random variables,

$$Z_{1,2,3} \overset{\text{iid}}{\sim} \text{Unif}(-\pi, \pi) \qquad\qquad Z_{4,5} \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

In the following experiments, we set $a = 5$ and $b = 0.1$. For the low-fidelity models, we will consider two different scenarios with different model assumptions.

6.1.1. Perfect model assumptions. Let $c = 1, d = 0.1$. We first consider a synthetic dataset consisting of two low-fidelity models:

$$\begin{aligned} X_1 &= \sin Z_1 + a \sin^2 Z_2 + bZ_3^4 \sin Z_1 + cZ_4 + 5 \\ X_2 &= \sin Z_1 + a \sin^2 Z_2 + bZ_3^4 \sin Z_1 + 5. \end{aligned} \tag{6.2}$$

In this case, both the linear model assumption (2.1) and the noise normality assumption (Assumption 4.3) are satisfied, i.e., $\mathbb{E}[Y|X_2] = X_2 - 5$, $\mathbb{E}[Y|X_1] = \mathbb{E}[Y|X_1, X_2] = X_1 - 5$. The correlation between Y and X_1, X_2 are approximately 0.999 and 0.956, respectively. The cost of sampling Y, X_1 and X_2 are assumed hierarchical, assigned as $(c_0, c_1, c_2) = (1, 0.05, 0.001)$.

We first compare the performance of AETC-d with ECDF-Y. The total budget B in the experiment ranges from 10 to 10^5. The ground truth is taken as an empirical CDF of Y computed using 10^7 independent samples. Accuracy results for ECDF-Y and AETC-d are reported in Figure 1. In this example, the AETC-d consistently outperforms ECDF-Y by a factor of around 3.8. The average estimation error decays to 0 as the budget goes to infinity, verifying the consistency result in Corollary 5.3.

To inspect the optimality of exploration rates selected by AETC-d, we fix the total budget at $B = 10^3$, and consider a deterministic version of Algorithm 5.1 with fixed exploration rate m. The new algorithm collects m exploration samples to estimate the parametric coefficients, and then selects S_{opt} for exploitation. In this example, $S_{\text{opt}} = \{1\}$, and the maximum exploration round is $\lfloor B/1.051 \rfloor = 951$. We apply the deterministic algorithm and compute the average W_1 distance between the resulting estimator and F_Y over 100 independent experiments for $m = (10, 30, 50, 100, 200, 300, 400, 500, 600)$. According to Theorem 4.5, for fixed m, the average error of the estimator given by the deterministic algorithm is asymptotically bounded by a function of the form

$$f(m; \alpha_1, \alpha_2) = \frac{\alpha_1}{\sqrt{m}} + \frac{\alpha_2}{\sqrt{\frac{B}{c_{\text{epr}}} - m}} \qquad\qquad 0 < m < \frac{B}{c_{\text{epr}}}, \tag{6.3}$$

for some constants $\alpha_1, \alpha_2 > 0$. Assuming the analysis in Theorem 4.5 is tight, we would expect the mean error of the deterministic algorithm, as a function of m, to fit a curve of the form (6.3). We use a nonlinear least-squares procedure to obtain such as a curve by optimizing over α_1, α_2. We also run the AETC-d 100 times and record the 5-50-95 quantiles of the exploration rates. The results are illustrated in Figure 1.

By comparing the mean errors committed by the deterministic algorithm at various fixed exploration rates, we find that the 5-50-95 quantiles of the exploration rate chosen by the AETC-d algorithm, 188-202-218, is relatively close to the optimal trade-off point, which is around 130. This empirically suggests that the upper bound criterion is informative in terms of balancing the exploration and exploitation errors.

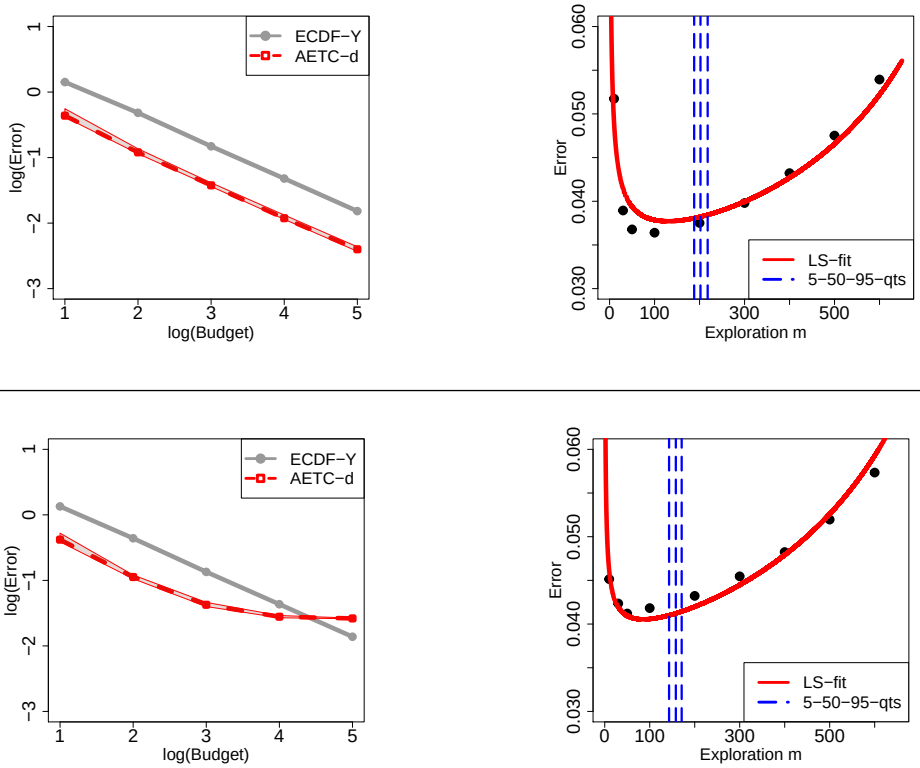

Figure 1. *Numerical results for Model (6.2) (**Top panel**) and Model (6.4) (**Bottom panel**). Comparison of the ($\log_{10}$) mean W_1 distance between F_Y and the estimated CDFs given by the ECDF-Y and the AETC-d as the total budget increases from 10 to 10^5. The 5-50-95 quantiles are plotted for the AETC-d to measure its uncertainty in the exploration phase (**Left**). Fixing the total budget $B = 10^3$, we compute the mean W_1 error of the deterministic algorithm at different exploration rates m, and the quantiles of the exploration rate from AETC-d (**Right**).*

6.1.2. Approximate linear assumptions. Let $c = d = 0$. We now consider the same low-fidelity models which were used in [23] for sensitivity analysis:

$$X_1 = \sin Z_1 + 0.95a \sin^2 Z_2 + bZ_3^4 \sin Z_1 \tag{6.4}$$
$$X_2 = \sin Z_1 + 0.6a \sin^2 Z_2 + 9bZ_3^2 \sin Z_1.$$

The correlation between Y and X_1, X_2 are approximately 0.999 and 0.950, respectively, which is similar to the previous example. However, under the current setup, neither the linear model relation (2.1) nor the Gaussian noise assumption is satisfied. Nevertheless, the correlations between Y and X_1, X_2 suggest that the relationship between the high-fidelity and low-fidelity models is approximately linear. The cost of sampling Y, X_1 and X_2 as well as the budget range are the same as in the previous section. We repeat the same experiments (both the accuracy and the optimality of exploration rates) and report the corresponding results in Figure 1.

Despite model misspecification, the AETC-d still demonstrates reasonable performance until the total budget exceeds 10^4. When the budget is sufficiently large, both exploration

and exploitation errors are so small that model misspecification errors start to dominate. A further investigation of the model misspecification effects will be carried out in Section 6.2. The optimal model selected by the AETC-d is $S_{\text{opt}} = \{1, 2\}$. By comparing the two plots in the right panel in Figure 6.2, we can see that the AETC-d exploration rate is only slightly above the optimal trade-off point.

Finally, to visualize the estimated CDFs, we conduct an instance study under three different budgets, $B = 10^k$, $k \in \{3, 4, 5\}$. We apply AETC-d and ECDF-Y to the same training dataset to estimate F_Y. The estimated CDFs are empirical CDFs of i.i.d. data points. We use a kernel density estimator [1] to obtain a mollified version of their histograms. For consistency, we apply the `density` function in `R` [24] with the default bandwidth (data-dependent), and the Gaussian kernel for smoothing. The results are given in Figure 2. We see that, despite model specification, AETC-d can more quickly capture certain features of the density such as the symmetric hump structures near $x = 0$ and $x = 5$ compared to ECDF-Y.

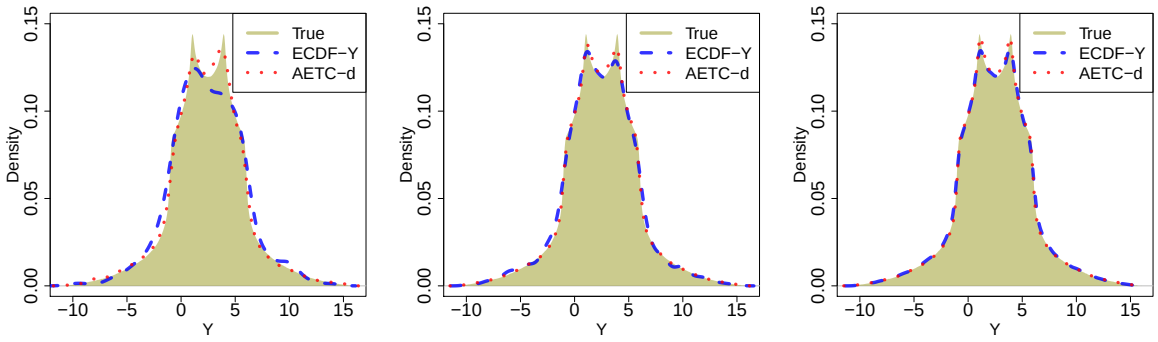

Figure 2. *Three instances of density estimates based on the ECDF-Y and the AETC-d at budget* $B = 10^3$ ***(Left)***, $B = 10^4$ ***(Middle)***, *and* $B = 10^5$ ***(Right)***.

6.2. Short column. To further discuss the model misspecification effects, we consider an analytic multifidelity model of a short column with cross-sectional area subject to bending and axial force, which were originally used in [14, 21]. The high-fidelity model is,

$$Y = 1 - \frac{4Z_4}{Z_1 Z_2^2 Z_3} - \left(\frac{Z_5}{Z_1 Z_2 Z_3}\right)^2,$$

where $Z_i, i \in [5]$ are independent random variables, and their respective distributions are

$$\begin{aligned} Z_1 &\sim \text{Unif}(5, 15) & Z_2 &\sim \text{Unif}(15, 25) \\ Z_3 &\sim \text{Lognormal}(5,\ 0.5) & Z_4 &\sim \mathcal{N}(2000, 400^2) \\ Z_5 &\sim \mathcal{N}(500, 100^2). \end{aligned}$$

[1] Kernel density estimators commit additional approximation errors, but they are more illustrative than CDFs, which are difficult to inspect visually.

The three low-fidelity model outputs are

$$X_1 = 1 - \frac{Z_4}{Z_1 Z_2^2 Z_3} - \left(\frac{Z_5}{Z_1 Z_2 Z_3}\right)^2 \qquad X_2 = 1 - \frac{Z_4}{Z_1 Z_2^2 Z_3} - \left(\frac{Z_5 + Z_4 Z_5}{Z_1 Z_2 Z_3}\right)^2$$
$$X_3 = 1 - \frac{Z_4}{Z_1 Z_2^2 Z_3} - \left(\frac{Z_5 + Z_4 Z_5}{Z_2 Z_3}\right)^2.$$

The correlation between Y and $X_i, i \in [3]$ are approximately 0.991, 0.829, and 0.712, respectively. The cost of sampling Y and $X_i, i \in [3]$, are assumed as 1 and 0.01, respectively, i.e., $c_0 = 1, c_1 = c_2 = c_3 = 0.01$.

We compare the performance of AETC-d and ECDF-Y at four different budgets, $B = 10^k, k \in [4]$, and report accuracy in the first plot of Figure 3. In this example, the estimation error of AETC-d plateaus when the budget is only moderately large. A possible explanation, as discussed in Section 6.1, is due to model misspecification. Such effects are often ubiquitous in practice, suggesting that AETC-d has dubious practical value for a sufficiently large budget.

To mitigate the impact of model misspecification, we identify the misspecified assumptions that prevent the model from converging. For every $S \subset [n]$, the following decomposition holds:

$$Y = \underbrace{\mathbb{E}[Y|X_S]}_{f_S(X_S)} + \underbrace{(Y - \mathbb{E}[Y|X_S])}_{\varepsilon_S},$$

where $f_S : \mathbb{R} \to \mathbb{R}$ is a measurable function, and ε_S is a centered random variable which is uncorrelated with $f_S(X_S)$. Assumption (2.1) assumes that f_S a linear function of X_S and ε_S is independent of $f_S(X_S)$. It was further imposed later that ε_S is Gaussian. Violation of any of these hypotheses may cause the inconsistent behavior of the AETC-d in this example.

To inspect the linearity assumption, we expand the feature space by incorporating nonlinear terms of the existing regressors to fit a larger linear model. To test the noise independence condition, we can either drop the noise in (4.3) to reduce the erroneous noisy effect[2], or use a different method such as quantile regression to model the potential heteroscedasticity of the noise. A brief description of using quantile regression combined with AETC-d in our setup is given in Appendix E. We investigate the Gaussian assumption by replacing the Gaussian noise emulator in (4.3) by an empirical measure sampler generated by residuals collected in the exploration phase.

In the following experiment, we expand the original model by including higher order (polynomial) terms of the existing features. We include additional regressors X_1^j, $j \in \{2, 3, 4, 5\}$, and call this enlarged model L. (Note that $Y \sim X_1 + \text{intercept}$ is the limiting model selected by AETC-d in the original case.) Note that in L, an exploitation model is no longer defined by the subset of regressors, but instead by the σ-field generated by the regressors. In fact, since both X_1 and X_1^2 generate the same σ-field, $\mathbb{E}[Y|X_1] = \mathbb{E}[Y|X_1^2]$. As a result, the linear model assumption (2.1) cannot hold simultaneously for both regressors as two distinct models unless X_1 is a constant. Thus, for each σ-field, we define its associated model as the union of the

[2]When the variance ratio between ε_S and Y is small, $Y \approx X_S^T \beta_S$, so that adding the noise emulator has little impact on the accuracy of the resulting estimator. When the ratio is moderate, adding ε_S as an independent component will degrade the quality of the estimator if the independence assumption is violated.

subsets of regressors that generate the σ-field, with exploitation cost given by the total cost of the low-fidelity models used to compute the regressors[3]. In this case, any model generating the same σ-field can be viewed as a sub-model under our definition. Also, in the exploitation stage of AETC-d, we consider three alternative methods to estimate F_Y: AETC-em (empirical noise), AETC-no (no noise), and AETC-q (quantile regression noise). The performance of these modified estimators in both the original and the expanded models (L) are compared in the second plot in Figure 3. For illustration, we plot the mean errors instead of the 5-50-95 quantiles region in the 100 experiments.

Figure 3 shows that the overall performance of all the methods under comparison improves in the enlarged model L when the budget exceeds 10^2. This implies that adding nonlinear terms of the existing regressors can help reduce the model misspecification effect that limit the asymptotic accuracy of AETC-d. The improvement for AETC-d in model L diminishes when the budget reaches a new threshold, after which the noise misspecification effect starts to emerge. By comparing the exploitation methods, we find that the independence assumption, as opposed to the Gaussian assumption, plays a more relevant role in causing the inconsistent behavior of the model. In this example, either dropping the noise or using the quantile regression for reconstruction is beneficial for mitigating the noise misspecification effects.

We end this section by providing an instance of the pointwise absolute error between F_Y and the estimated CDFs given by the ECDF-Y, the AETC-d, the AETC-d (L), the AETC-d-no (L), and the AETC-d-q (L) when $B = 10^4$, as in the last plot in Figure 3.

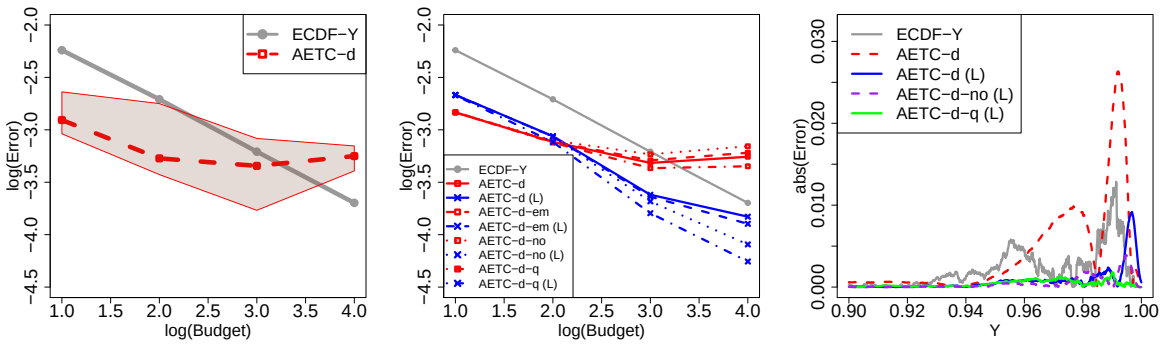

Figure 3. *Comparison of the* ($\log_{10}$) *mean* W_1 *distance between* F_Y *and the estimated CDFs given by the ECDF-Y and the AETC-d as the total budget increases from* 10 *to* 10^4. *The* 5-50-95 *quantiles are plotted for the AETC-d to measure its uncertainty in the exploration phase (**Left**). Average estimation errors of the estimated CDFs given by the ECDF-Y and the AETC-d with different exploitation strategies in both the original and the expanded model L (**Middle**). An instance of pointwise absolute errors of the CDFs given by the ECDF-Y, the AETC-d, the AETC-d (L), the AETC-d-no (L), and the AETC-d-q (L) when* $B = 10^4$ (***Right***).

[3]For example, if $S = \{1\}$, then the corresponding regressors are $1, X_1^j, j \in [5]$, and the exploitation cost per sample is simply c_1, not $5c_1$, since sampling X_1 with cost c_1 allows one to generate the corresponding sample of $X_1^j, j = 2, 3, 4, 5$ at no additional cost.

6.3. Parametrized PDEs. In the last experiment, we consider a multifidelity setup given by a parametric elliptic equation. The particular setup is taken from [33, Section 7.1]: We consider an elliptic PDE over a square spatial domain $D = [0,1]^2$ that governs displacement in linear elasticity.

The parametric version of this problem equation seeks the solution u to the PDE,

$$\begin{cases} \nabla \cdot (\kappa(\boldsymbol{p}, \boldsymbol{x}) \nabla u(\boldsymbol{p}, \boldsymbol{x})) = f(\boldsymbol{x}) & \forall (\boldsymbol{p}, \boldsymbol{x}) \in \mathcal{P} \times D \\ u(\boldsymbol{p}, \boldsymbol{x}) = 0 & \forall (\boldsymbol{p}, \boldsymbol{x}) \in \mathcal{P} \times \partial D \end{cases} \tag{6.5}$$

where $\boldsymbol{p} \in \mathbb{R}^4$ is a random vector with independent components uniformly distributed on $[-1,1]$, and κ is modeled as a truncated Karhunen-Loéve expansion, 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 an exponential covariance kernel on D. (See [33, Section 7.1] for more details.) The displacement u is used to compute a scalar QoI, the *compliance* or energy norm of the solution, which is the measure of elastic energy absorbed in the structure as a result of loading,

$$\mathsf{cpl} := \int_D \kappa(\boldsymbol{x}, \boldsymbol{p}) \nabla u(\boldsymbol{p}, \boldsymbol{x})^T \nabla u(\boldsymbol{p}, \boldsymbol{x}) d\boldsymbol{x}. \tag{6.6}$$

We solve (6.5) for each fixed $\boldsymbol{p}$ via the finite element method with standard bilinear square isotropic finite elements on a rectangular mesh.

In this example, we form a multifidelity hierarchy through mesh coarsening through mesh parameter h. The model solved with mesh size $h = 2^{-7}$ is the high-fidelity model. We create three low-fidelity models based on more economical discretizations: $h = 2^{-3}, 2^{-2}, 2^{-1}$. The outputs of these models are the compliance QoI computed from the respective approximate solutions.

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., $c_0 = 4096, c_1 = 16, c_2 = 4, c_3 = 1$. The correlations between the outputs of Y and X_1, X_2, X_3 are $0.940, 0.841, -0.146$, respectively. The total budget B ranges from 10^5 to 10^7.

To mitigate the potential model misspecification effects, we include the second-order interactions between $X_i, i \in [3]$ as additional regressors, i.e., $X_i X_j, i, j \in [3]$. In this case, we have ten different σ-fields generated by the regressors, namely,

$$\sigma(X_{i_1} X_{j_1}, X_{i_2} X_{j_2}, X_{i_3} X_{j_3}) \qquad\qquad i_1, i_2, i_3, j_1, j_2, j_3 \in [3].$$

Their corresponding models are the maximal subsets of regressors generating the σ-fields, with exploitation cost defined as the sum of the cost of the low-fidelity models used to build the regressors. The oracle F_Y is taken as an empirical CDF constructed from 1.8×10^6 independent samples of the high-fidelity model. Accuracy results for ECDF-Y, AETC-d, AETC-d-no and

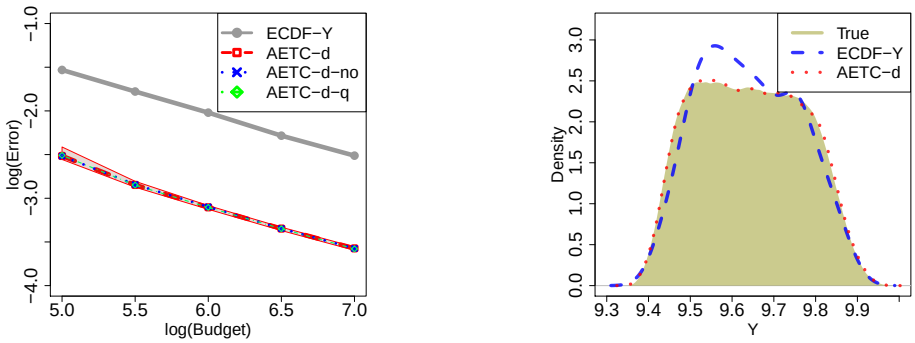

Figure 4. *Comparison of the* ($\log_{10}$) *mean* W_1 *distance between* F_Y *and the estimated CDFs given by the ECDF-Y, the AETC-d, the AETC-d-no and the AETC-d-q as the total budget increases from* 10^5 *to* 10^7*. The* 5-50-95 *quantiles are plotted for the AETC-d algorithm to measure its uncertainty in the exploration phase (**Left**). An instance of the densities given by the ECDF-Y and the AETC-d when* $B = 10^6$ *(**Right**).*

AETC-d-q are reported in the first plot in Figure 4. For visualization, we provide an instance of the estimated densities (default `density` function in `R` applied to the estimated CDFs) given by ECDF-Y and AETC-d using the same training dataset when $B = 10^6$ in Figure 4.

Figure 4 shows that within the budget range of this experiment, AETC-d is asymptotically consistent and outperforms ECDF-Y by a substantial margin. The estimated density given by AETC-d at $B = 10^6$ almost matches the density obtained from 1.8×10^6 independent samples of Y, which approximately costs 10^{10} budget units. Little difference between AETC-d, AETC-d-no and AETC-d-q is visible. A possible explanation for this is that the magnitude of the variance of the model selected for exploitation (which is the full model in this example) is much smaller than the variance of Y, making the potential noise misspecification effect negligible under the budget range of this example. We substantiate this hypothesis by fitting a linear model using the complete training data, and we approximately compute the variance ratio between the model noise and Y, which is $3 \times 10^{-5} \ll 1$.

7. Conclusions.

In this paper, we introduce an efficient strategy for learning the distribution of a scalar-valued QoI in the multifidelity setup. Under a linear model assumption, we propose a semi-parametric approach for approximating the distribution by leveraging samples of models of different resolutions/costs. The main novelty in our analysis is to provide an asymptotically informative and computationally estimable upper bound for the average 1-Wasserstein distance between the estimator and the true distribution, and using it to devise an adaptive algorithm, AETC-d, for efficient budget allocation. We show that, for a large budget, the AETC-d is consistent, and explores and exploits optimally under a proposed upper bound criterion. Our setup and algorithm require neither a model hierarchy nor an *a priori* estimate of cross-model correlations. We also discuss various approaches to mitigating the model misspecification impact when the linear regression, independence, and/or normality assumptions are violated.

Distribution learning is considered a much harder problem compared to parameter es-

timation. Our method takes an initial step towards addressing this problem, and provides a potentially effective way to fully quantify the uncertainty of the QoI associated with a high-fidelity model. The implementation of our algorithm is automatic, and enjoys certain robustness when the model noise is relatively small.

Acknowledgement. Y. Xu thanks Xiaoou Pan for clarifying a uniform consistency result in quantile regression under a random design setup. We also thank Ruijian Han for a careful reading of an early draft, and for providing several comments that improved the presentation of the manuscript.

8. Appendices.

A. Proof of Theorem 3.4. Without loss of generality, we assume $F_1(x) \geq F_2(x)$ for all $x \in \mathbb{R}$, otherwise one can divide into two cases and discuss separately. We first give a bound for $|F_1(x) - F_2(x)|$ which is tight in the asymptotic regime $|x| \to \infty$:

$$|F_1(x) - F_2(x)| = F_1(x) - F_2(x) \leq \min\{F_1(x), 1 - F_2(x)\} \leq 2\exp\left(-\frac{|x|}{C_1C_2}\right). \tag{A.1}$$

To utilize the coupling information, note that for $t > 0$,

$$\{Z_1 \leq x\} \subseteq \{Z_2 \leq x + t\} \cup \{Z_2 - Z_1 > t\},$$

which implies

$$F_1(x) - F_2(x) \leq F_2(x+t) - F_2(x) + \mathbb{P}\left(Z_2 - Z_1 > t\right) \leq C_3 t + 2\exp\left(-\frac{t}{C_1\delta}\right), \tag{A.2}$$

where the first inequality follows from applying (3.7) to the normalized sub-exponential random variable $(Z_1 - Z_2)/\delta$. Combining estimates (A.1) and (A.2), we can bound $\|F_1(x) - F_2(x)\|^p_{L^p(\mathbb{R})}$ as follows: Fixing $T > 0$ (which will be determined later),

$$\begin{aligned}
\|F_1(x) - F_2(x)\|^p_{L^p(\mathbb{R})} &= \int_{|x|>T} |F_1(x) - F_2(x)|^p dx + \int_{|x|\leq T} |F_1(x) - F_2(x)|^p dx \\
&\overset{(A.1),(A.2)}{\leq} \int_{|x|>T} 2^p \exp\left(-\frac{p|x|}{C_1C_2}\right) dx + \int_{|x|\leq T} \left[C_3 t + 2\exp\left(-\frac{t}{C_1\delta}\right)\right]^p dx \\
&\leq \frac{2^{p+1}C_1C_2}{p} \exp\left(-\frac{pT}{C_1C_2}\right) + 2T\left[C_3 t + 2\exp\left(-\frac{t}{C_1\delta}\right)\right]^p. \qquad \text{(A.3)}
\end{aligned}$$

Setting $t = C_1\delta\log(1/\delta), T = \frac{C_1C_2}{p}\log\left(\frac{2^{p+1}C_1C_2}{p\delta^p}\right)$ and substituting back into (A.3) yields

$$\begin{aligned}
\|F_1(x) - F_2(x)\|^p_{L^p(\mathbb{R})} &\leq \delta^p + \frac{2C_1C_2}{p}\log\left(\frac{2^{p+1}C_1C_2}{p\delta^p}\right)\left[C_1C_3\delta\log\left(\frac{1}{\delta}\right) + 2\delta\right]^p \\
&\lesssim \log\left(\frac{1}{\delta}\right)\left[\delta\log\left(\frac{1}{\delta}\right)\right]^p. \qquad \text{(A.4)}
\end{aligned}$$

Taking the p-th root on both sides of (A.4) completes the proof.

B. Proof of Lemma 4.4. We first prove (4.6a). Since it is easier to work with the W_2 distance when independence is assumed, we start by deriving an estimate for the W_2 distance between F_Y and $F_{Y'}$. Conditional on $\widehat{\beta}_S$ and $\widehat{\sigma}_S^2$, the natural coupling (V_1, V_2) for F_Y and $F_{Y'}$

$$V_1 = X_S^T \beta_S + \sigma_S \xi \qquad V_2 = X_S^T \widehat{\beta}_S + \widehat{\sigma}_S \xi \qquad\qquad \xi \sim \mathcal{N}(0,1) \perp\!\!\!\perp X_S$$

yields an upper bound for the average W_2 distance between F_Y and $F_{Y'}$:

$$\begin{aligned} W_2^2(F_Y, F_{Y'}) &\le \mathbb{E}_{X_S,\xi}\left[|V_1 - V_2|^2 \middle| \widehat{\beta}_S, \widehat{\sigma}_S^2\right] \\ &= \mathbb{E}\left[(X_S^T(\widehat{\beta}_S - \beta_S))^2 \middle| \widehat{\beta}_S\right] + |\widehat{\sigma}_S - \sigma_S|^2 \cdot \mathbb{E}[\xi^2] \\ &= (\widehat{\beta}_S - \beta_S)^T \Lambda_S (\widehat{\beta}_S - \beta_S) + |\widehat{\sigma}_S - \sigma_S|^2. \end{aligned} \tag{B.1}$$

Note that (B.1) does not involve exploitation samples. We now average out the randomness of noise in exploration for both terms in (B.1). For the first term, the fact

$$\widehat{\beta}_S - \beta_S \sim \mathcal{N}(0, \sigma_S^2 (Z_S^T Z_S)^{-1}) \tag{B.2}$$

and the strong law of large numbers (SLLN) together yields that

$$\mathbb{E}_{\varepsilon_S}\left[(\widehat{\beta}_S - \beta_S)^T \Lambda_S (\widehat{\beta}_S - \beta_S)\right] = \frac{\sigma_S^2}{m} \operatorname{tr}\left(\Lambda_S (m^{-1} Z_S^T Z_S)^{-1}\right) \overset{\text{SLLN}}{\simeq} \frac{\sigma_S^2 (s+1)}{m}. \tag{B.3}$$

For the second term, note that when ε_S is Gaussian, Cochran's theorem tells us that the distribution of $\widehat{\sigma}_S$ is independent of Z_S (or exploration samples), i.e.,

$$\widehat{\sigma}_S \overset{\mathcal{D}}{=} \frac{\sigma_S}{\sqrt{m - |S| - 1}} \chi_{m-|S|-1}, \tag{B.4}$$

where $\chi_{m-|S|-1}$ is the Chi-distribution with $(m - |S| - 1)$ degrees of freedom. Consequently, using the exact moment formulas for $\chi_{m-|S|-1}$, we have

$$\begin{aligned} &\mathbb{E}_{\varepsilon_S}[|\widehat{\sigma}_S - \sigma_S|^2] \\ =& \frac{\sigma_S^2}{m - |S| - 1} \mathbb{E}\left[\left(\chi_{m-|S|-1} - \sqrt{m - |S| - 1}\right)^2\right] \\ =& \frac{\sigma_S^2}{m - |S| - 1}\left[\mathbb{V}[\chi_{m-|S|-1}] + \left(\mathbb{E}[\chi_{m-|S|-1}] - \sqrt{m - |S| - 1}\right)^2\right] \\ =& \frac{\sigma_S^2}{m - |S| - 1}\left[m - |S| - 1 - \left(\frac{\sqrt{2}\Gamma(\frac{m-|S|}{2})}{\Gamma(\frac{m-|S|-1}{2})}\right)^2 + \left(\frac{\sqrt{2}\Gamma(\frac{m-|S|}{2})}{\Gamma(\frac{m-|S|-1}{2})} - \sqrt{m - |S| - 1}\right)^2\right] \\ =& \frac{2\sigma_S^2}{m - |S| - 1}\left(m - |S| - 1 - \sqrt{2(m - |S| - 1)} \frac{\Gamma(\frac{m-|S|}{2})}{\Gamma(\frac{m-|S|-1}{2})}\right) \\ \le& \frac{2\sigma_S^2}{m - |S| - 1}\left(m - |S| - 1 - \sqrt{(m - |S| - 1)(m - |S| - 2)}\right) \\ \le& \frac{\sigma_S^2}{m - |S| - 1} \cdot \sqrt{\frac{m - |S| - 1}{m - |S| - 2}} \simeq \frac{\sigma_S^2}{m}. \end{aligned} \tag{B.5}$$

(B.3), (B.5) combined with Jensen's inequality finishes the proof of (4.6a).

We next prove (4.6b). Conditional on $\widehat{\beta}_S$ and $\widehat{\sigma}_S^2$, the sub-exponential assumption ensures that Y' is a random variable with bounded r-th moments for all $r > 0$, i.e., for $r > 2$. Appealing to the non-asymptotic estimates on the convergence rate of empirical measures in Theorem 3.2 and averaging over exploration noise ε_S, we have

$$\mathbb{E}_{A_j,\varepsilon_S}\left[W_1\left(\widehat{F}_{Y,S}, F_{Y'}\right)\right] \leq \frac{\mathbb{E}_{\varepsilon_S}[J_1(F_{Y'})]}{\sqrt{N_S}}, \tag{B.6}$$

where J_1 is defined in (3.5). The desired result would follow if we can show that $\mathbb{E}_{\varepsilon_S}[J_1(F_{Y'})] \to J_1(F_Y)$ almost surely as $m \to \infty$.

Note from (B.4) that $\widehat{\sigma}_S^2$'s distribution does not depend on Z_S. The strong consistency of $\widehat{\sigma}_S^2$ follows immediately from tail probability estimates of Chi-distributions and the Borel-Cantelli lemma. On the other hand, under the sub-exponential assumption on X_S, for sufficiently large m, with overwhelming probability, say $1 - m^{-2}$, that the largest and smallest eigenvalues of $Z_S^T Z_S$, $\lambda_{\max}$ and $\lambda_{\min}$, are both of order m [19], i.e., $\lambda_{\min} \to \infty$ and $\log(\lambda_{\max}) = o(\lambda_{\min})$ as $m \to \infty$. An application of the Borel-Cantelli lemma yields that $\log(\lambda_{\max}) = o(\lambda_{\min})$ a.s. According to a result in [17, Theorem 1], $\widehat{\beta}_S$ is a strongly consistent estimator for β_S. Hence, with probability 1,

$$Y' := X_S^T\widehat{\beta}_S + \widehat{\sigma}_S\xi \xrightarrow[a.s.]{m\to\infty} X_S^T\beta_S + \sigma_S\xi := Y \qquad \xi \sim \mathcal{N}(0,1). \tag{B.7}$$

Now define $V_S = (X_S, \xi)$. It can be verified using independence between X_S and ξ that under Assumption 4.2, V_S is a jointly sub-exponential random vector and satisfies the global Lipschitz condition for all unit marginals. Write Y and Y' in a coupled form using V_S: $Y = a_1^T V_S, Y' = a_2^T V_S$, where $a_1 = (\beta_S, \sigma_S)^T, a_2 = (\widehat{\beta}_S, \widehat{\sigma}_S)^T$. Conditional on Z_S, the strong consistency of $\widehat{\beta}_S$ and $\widehat{\sigma}_S$ implies that for large m (realization-dependent), $\|a_2 - a_1\|_2 < 1$. We are now in a position to apply Theorem 3.4 to bound $\mathbb{E}_{\varepsilon_S}[J_1(F_{Y'})] - J_1(F_Y)$: For large m,

$$\begin{aligned}
|\mathbb{E}_{\varepsilon_S}[J_1(F_{Y'})] - J_1(F_Y)| &\leq \mathbb{E}_{\varepsilon_S}\left[\int_{\mathbb{R}}\left|\sqrt{F_{Y'}(y)(1-F_{Y'}(y))} - \sqrt{F_Y(y)(1-F_Y(y))}\right| dy\right] \\
&\overset{(i)}{\leq} \mathbb{E}_{\varepsilon_S}\left[\int_{\mathbb{R}}\sqrt{|F_{Y'}(y)(1-F_{Y'}(y)) - F_Y(y)(1-F_Y(y))|}dy\right] \\
&\overset{(ii)}{\leq} \mathbb{E}_{\varepsilon_S}\left[\int_{\mathbb{R}}\sqrt{|F_{Y'}(y) - F_Y(y)|}dy\right] \\
&\overset{(3.9)}{\lesssim} \mathbb{E}_{\varepsilon_S}\left[(\|\widehat{\beta}_S - \beta_S\|_2^2 + \|\widehat{\sigma}_S - \sigma_S\|_2^2)^{1/6}\right] \\
&\overset{\text{Jensen}}{\leq} \mathbb{E}_{\varepsilon_S}\left[\|\widehat{\beta}_S - \beta_S\|_2^2 + \|\widehat{\sigma}_S - \sigma_S\|_2^2\right]^{1/6} \\
&\overset{(B.5)}{\lesssim} \left[\frac{\sigma_S^2}{m}\left(\mathrm{tr}\big(m^{-1}Z_S^T Z_S\big)^{-1} + 1\right)\right]^{1/6} \xrightarrow[a.s.]{m\to\infty} 0,
\end{aligned} \tag{B.8}$$

where (i) follows from the elementary inequality $\sqrt{a} - \sqrt{b} \leq \sqrt{|a-b|}$, and (ii) follows from

the calculation

$$|F_{Y'}(y)(1-F_{Y'}(y)) - F_Y(y)(1-F_Y(y))| = |F_{Y'}(y) - F_Y(y)||1-(F_{Y'}(y)+F_Y(y))| \\ \leq |F_{Y'}(y) - F_Y(y)|.$$

The proof is finished by plugging (B.8) into (B.6).

C. Proof of Theorem 5.1. We only sketch the proof as the details are similar to [33, Theorem 5.2]. First note that the strong consistency of the estimators used in the procedure and the regularization step force $m(B) \uparrow \infty$ as $B \uparrow \infty$ with probability 1. Whenever B is so large that $m(B)/2$ is greater than some threshold, the estimated $\widehat{G}_S$ function will be sufficiently close to G_S for every $S \subset [n]$, so that the S_{opt} will be selected as the candidate for exploitation afterwards. Since the threshold arrives before the exploration stops, the model selected in the end is optimal. This proves the second part of (5.9). The first part of (5.9) follows from a continuity perturbation argument.

D. Proof of Corollary 5.3. For fixed B, let $N(B)$ be the exploitation sampling rate in Algorithm 5.1, and denote the parametric model used for exploitation as $Y'(B)$, i.e.,

$$Y'(B) = X_{S(B)}^T \widehat{\beta}_{S(B)}(m(B)) + \mathcal{N}\left(0, \widehat{\sigma}^2_{S(B)}(m(B))\right), \tag{D.1}$$

where the two terms on the right-hand side of (D.1) are independent. By Theorem 5.1, with probability 1, $S(B) = S_{\text{opt}}$ for all sufficiently large B, and $m(B)/m^*(S_{\text{opt}}) \to 1$, i.e., $m(B) \to \infty$ and $N(B) \to \infty$ as $B \to \infty$. Under Assumption 4.2, we apply (3.9) with $p = 1$ together with the strong consistency of $\widehat{\beta}_{S_{\text{opt}}}$ and $\widehat{\sigma}^2_{S_{\text{opt}}}$ to conclude that $W_1(Y'(B), Y) \to 0$ as $B \to \infty$ almost surely. (The randomness here only depends on exploration.)

Now fix a realization along which $N(B) \to \infty$ and $W_1(Y'(B), Y) \to 0$ as $B \to \infty$. Let $\{B_k\}$ be an arbitrary sequence such that $B_k \uparrow \infty$ as $k \uparrow \infty$. Since convergence in W_1 implies convergence in distribution, $\{F_{Y'(B_k)}\}$ is δ-tight[4] [32, Section 17.5]. This observation, combined with the fact that $\widehat{F}_{Y,\text{aetcd}}(B_k)$ is an empirical measure of $Y'(B_k)$ consisting of $N(B_k)$ samples, implies that $\widehat{F}_{Y,\text{aetcd}}(B_k)$ converges to F_Y in distribution almost surely [1, Theorem 1]. To lift the convergence to W_1, it only remains to show $\int |x| d\widehat{F}_{Y,\text{aetcd}}(B_k) \to \int |x| dF_Y$ as $k \to \infty$, which can be verified using (D.1) and the strong law of large numbers. The proof is finished by noting that $\{B_k\}$ is arbitrary.

E. A quantile regression framework. Quantile regression offers an alternative approach to simulating Y through a random coefficient interpretation [15]. For any $S \subset [n]$ and $\tau \in (0,1)$, we assume the conditional τ-th quantile of Y on X_S satisfies

$$F^{-1}_{Y|X_S}(\tau) = X_S^T \beta_S(\tau), \tag{E.1}$$

where $\beta_S(\tau)$ the τ-th coefficient vector. (E.1) is a standard quantile regression formulation, and can be used to model heteroscedastic noise effects. $\beta_S(\tau)$ can be computed by minimizing

[4] A sequence of probability measures $\{P_k\}$ defined on a metric space is called δ-tight if for every $\varepsilon > 0$, there exist a compact measurable set K and a sequence $\delta_k \downarrow 0$ such that $P_k(K^{\delta_k}) > 1 - \varepsilon$ for every k, where $K^{\delta_k} := \{x : \text{dist}(x, K) < \delta_k\}$.

the following empirical ρ_τ-risk:

$$\widehat{\beta}_S(\tau) = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^{s+1}} \frac{1}{m}\sum_{\ell\in[m]} \rho_\tau(Y_\ell - X_{\mathrm{epr},\ell}^T\beta) \qquad \rho_\tau(x) = x(\tau - \mathbf{1}_{x<0}).$$

Thus, (E.1) approximately equals

$$\widehat{F}^{-1}_{Y|X_S}(\tau) = X_S^T\widehat{\beta}_S(\tau). \tag{E.2}$$

As opposed to (4.3), (E.2) provides a way to simulate Y based on X_S via inverse transform sampling:

$$Y \approx X_S^T\widehat{\beta}_S(U). \qquad U \sim \mathrm{Unif}(0,1) \perp\!\!\!\perp X_S. \tag{E.3}$$

In our case, $X_{\mathrm{epr},\ell}, \ell \in [m]$ are i.i.d. samples so (E.1) fits into a random design quantile regression framework as analyzed in [20], where the authors established a strong consistency result for $\widehat{\beta}_S(\tau)$ under mild conditions. The consistency result can be further proven to hold uniformly for all $\tau \in [\delta, 1-\delta]$ for any fixed $\delta > 0$, which justifies the asymptotic behavior of the procedure in (E.2) as $m, N_S \to \infty$.

In the quantile regression framework, obtaining the optimal choices for m and S is much harder than in the linear regression setup. The AETC-d-q algorithm in Section 6 implements (E.3) with m set as the adaptive exploration rate given by the AETC-d, S as the corresponding model output for exploitation, and U approximated via $\frac{1}{K}\sum_{j\in[K]}\delta_{\frac{j}{K+1}}$ with $K = 100$.

REFERENCES

[1] R. Beran, L. Le Cam, and P. Millar, *Convergence of stochastic empirical measures*, Journal of multivariate analysis, 23 (1987), pp. 159–168.

[2] S. Bobkov and M. Ledoux, *One-dimensional empirical measures, order statistics, and kantorovich transport distances*, Memoirs of the American Mathematical Society, 261 (2019), pp. 0–0, https://doi.org/10.1090/memo/1259, https://doi.org/10.1090%2Fmemo%2F1259.

[3] F. Bolley, A. Guillin, and C. Villani, *Quantitative concentration inequalities for empirical measures on non-compact spaces*, Probability Theory and Related Fields, 137 (2006), pp. 541–593, https://doi.org/10.1007/s00440-006-0004-7, https://doi.org/10.1007%2Fs00440-006-0004-7.

[4] S. Cambanis, G. Simons, and W. Stout, *Inequalities for e k (x, y) when the marginals are fixed*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 36 (1976), pp. 285–294.

[5] S. Chatterjee, *Lecture notes on stein's method*, Stanford lecture notes, (2007).

[6] A. L. Gibbs and F. E. Su, *On choosing and bounding probability metrics*, International Statistical Review, 70 (2002), pp. 419–435, https://doi.org/10.1111/j.1751-5823.2002.tb00178.x, https://doi.org/10.1111%2Fj.1751-5823.2002.tb00178.x.

[7] M. B. Giles, *Multilevel monte carlo path simulation*, Operations research, 56 (2008), pp. 607–617.

[8] 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, https://doi.org/10.1137/140960086, https://doi.org/10.1137%2F140960086.

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

[10] 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, https://doi.org/10.1016/j.jcp.2020.109257, https://doi.org/10.1016%2Fj.jcp.2020.109257.

[11] U. Haagerup, *The best constants in the khintchine inequality*, Studia Mathematica, 70 (1981), pp. 231–283, https://doi.org/10.4064/sm-70-3-231-283, https://doi.org/10.4064%2Fsm-70-3-231-283.
[12] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Methods*, Methuen, London, 1964.
[13] T. Ishigami and T. Homma, *An importance quantification technique in uncertainty analysis for computer models*, in [1990] Proceedings. First International Symposium on Uncertainty Modeling and Analysis, IEEE Comput. Soc. Press, 1990, https://doi.org/10.1109/isuma.1990.151285, https://doi.org/10.1109%2Fisuma.1990.151285.
[14] C. Kirjner-Neto, E. Polak, and A. Kiureghian, *Algorithms for reliability-based optimal design*, in Reliability and Optimization of Structural Systems, Springer US, 1995, pp. 144–152, https://doi.org/10.1007/978-0-387-34866-7_13, https://doi.org/10.1007%2F978-0-387-34866-7_13.
[15] R. Koenker, *Fundamentals of quantile regression*, in Quantile Regression, Cambridge University Press, pp. 26–67, https://doi.org/10.1017/ccol0521845734.002, https://doi.org/10.1017%2Fccol0521845734.002.
[16] S. Krumscheid and F. Nobile, *Multilevel monte carlo approximation of functions*, SIAM/ASA Journal on Uncertainty Quantification, 6 (2018), pp. 1256–1293, https://doi.org/10.1137/17m1135566, https://doi.org/10.1137%2F17m1135566.
[17] T. L. Lai and C. Z. Wei, *Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems*, The Annals of Statistics, 10 (1982), pp. 154–166, https://doi.org/10.1214/aos/1176345697, https://doi.org/10.1214%2Faos%2F1176345697.
[18] 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, https://doi.org/10.1002/2016wr019475, https://doi.org/10.1002%2F2016wr019475.
[19] S. Mendelson and A. Pajor, *On singular values of matrices with independent rows*, Bernoulli, 12 (2006), pp. 761–773, https://doi.org/10.3150/bj/1161614945, https://doi.org/10.3150%2Fbj%2F1161614945.
[20] X. Pan and W.-X. Zhou, *Multiplier bootstrap for quantile regression: non-asymptotic theory under random design*, Information and Inference: A Journal of the IMA, (2020), https://doi.org/10.1093/imaiai/iaaa006, https://doi.org/10.1093%2Fimaiai%2Fiaaa006.
[21] B. Peherstorfer, K. Willcox, and M. Gunzburger, *Optimal model management for multifidelity monte carlo estimation*, SIAM Journal on Scientific Computing, 38 (2016), pp. A3163–A3194, https://doi.org/10.1137/15m1046472, https://doi.org/10.1137%2F15m1046472.
[22] B. Peherstorfer, K. Willcox, and M. Gunzburger, *Survey of multifidelity methods in uncertainty propagation, inference, and optimization*, SIAM Review, 60 (2018), pp. A550–A591.
[23] 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, https://doi.org/10.1137/17m1151006, https://doi.org/10.1137%2F17m1151006.
[24] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020, https://www.R-project.org/.
[25] F. Rassoul-Agha and T. Seppäläinen, *A course on large deviations with an introduction to Gibbs measures*, vol. 162, American Mathematical Soc., 2015.
[26] P. M. Robinson, *Root-n-consistent semiparametric regression*, Econometrica, 56 (1988), p. 931, https://doi.org/10.2307/1912705, https://doi.org/10.2307%2F1912705.
[27] D. Schaden and E. Ullmann, *Asymptotic analysis of multilevel best linear unbiased estimators*, arXiv preprint arXiv:2012.03658, (2020).
[28] D. Schaden and E. Ullmann, *On multilevel best linear unbiased estimators*, SIAM/ASA Journal on Uncertainty Quantification, 8 (2020), pp. 601–635, https://doi.org/10.1137/19m1263534, https://doi.org/10.1137%2F19m1263534.
[29] A. W. Van der Vaart, *Asymptotic statistics*, vol. 3, Cambridge university press, 2000.
[30] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018.
[31] C. Villani, *The metric side of optimal transportation*, in Graduate Studies in Mathematics, American Mathematical Society, mar 2003, pp. 205–235, https://doi.org/10.1090/gsm/058/08, https://doi.org/10.1090%2Fgsm%2F058%2F08.

[32] D. Williams, *Probability with martingales*, Cambridge university press, 1991.
[33] Y. Xu, V. Keshavarzzadeh, R. M. Kirby, and A. Narayan, *A bandit-learning approach to multifidelity approximation*, arXiv preprint arXiv:2103.15342, (2021).