Multifidelity preconditioning of the cross-entropy method for rare event simulation and failure probability estimation*

Benjamin Peherstorfer†, Boris Kramer‡, and Karen Willcox‡

Abstract. Accurately estimating rare event probabilities with Monte Carlo can become costly if for each sample a computationally expensive high-fidelity model evaluation is necessary to approximate the system response. Variance reduction with importance sampling significantly reduces the number of required samples if a suitable biasing density is used. This work introduces a multifidelity approach that leverages a hierarchy of low-cost surrogate models to efficiently construct biasing densities for importance sampling. Our multifidelity approach is based on the cross-entropy method that derives a biasing density via an optimization problem. We approximate the solution of the optimization problem at each level of the surrogate-model hierarchy, reusing the densities found on the previous levels to precondition the optimization problem on the subsequent levels. With the preconditioning, an accurate approximation of the solution of the optimization problem at each level can be obtained from a few model evaluations only. In particular, at the highest level, only few evaluations of the computationally expensive high-fidelity model are necessary. Our numerical results demonstrate that our multifidelity approach achieves speedups of several orders of magnitude in a thermal and a reacting-flow example compared to the single-fidelity cross-entropy method that uses a single model alone.

Key words. multifidelity, Monte Carlo, rare event simulation, failure probability estimation, surrogate models, reduced models, importance sampling, multilevel, cross-entropy method, variance reduction

AMS subject classifications. 35Q62, 62F25, 65C60, 65C05, 60H35, 35R60

1. Introduction.

Rare event simulation with standard Monte Carlo typically requires a large number of samples to derive accurate estimates of rare event probabilities, which can become computationally infeasible if for each sample a computationally expensive high-fidelity model evaluation is necessary to simulate the system response. Importance sampling is a variance reduction strategy for Monte Carlo estimation that samples from a problem-dependent biasing distribution. The biasing distribution is chosen such that fewer samples are necessary to obtain an acceptable estimate of the rare event probability than with standard Monte Carlo. The bias introduced by the sampling from the biasing distribution is corrected by reweighing the samples in the importance sampling estimator [16, 31].

Traditionally, importance sampling consists of two steps. In the first step, the biasing distribution is constructed, and in the second step, samples are drawn from the biasing distribution and the estimate is derived [6, 38]. The challenge of rare event probability estimation

*Submitted to the editors May 9, 2018.

Funding: This work was supported by the DARPA EQUiPS Program, Award UTA15-001067, Program Manager F. Fahroo, by the AFOSR MURI on multi-information sources of multi-physics systems, Award Number FA9550-15-1-0038, Program Manager J.-L. Cambier, and by the Air Force Center of Excellence on Multi-Fidelity Modeling of Rocket Combustor Dynamics, Award Number FA9550-17-1-0195.

†Department of Mechanical Engineering and Wisconsin Institute for Discovery, University of Wisconsin-Madison (peherstorfer@wisc.edu).

‡Department of Aeronautics and Astronautics, Massachusetts Institute of Technology (bokramer@mit.edu, kwillcox@mit.edu).

with importance sampling is the construction of a suitable biasing distribution that leads to variance reduction. In principle, the optimal biasing distribution that leads to an estimator with zero variance is known, but evaluating the density of this zero-variance distribution requires the probability of the rare event, i.e., the quantity that is to be estimated. The cross-entropy (CE) method [34, 33, 35, 10] provides a practical way to approximate the zero-variance density. The CE method optimizes for a density that minimizes the Kullback-Leibler divergence from the zero-variance density in a set of feasible densities. Even though solving for a biasing density with the CE method typically requires fewer high-fidelity model evaluations than estimating the rare event probability with a standard Monte Carlo approach, the costs of the optimization problem in the CE method can still be significant if the high-fidelity model is expensive to evaluate.

In this paper, we introduce a multifidelity method that leverages a hierarchy of low-cost surrogate models to reduce the costs of constructing a CE-optimal biasing density. Examples of surrogate models include projection-based reduced models [32, 2], data-fit interpolation and regression models [17], machine-learning-based models such as support vector machines [9], and other simplified models [23]. At each level of the hierarchy, a CE-optimal density is derived with respect to the surrogate model corresponding to that current level. The optimization is initialized with the CE-optimal density of the previous level, which leads to preconditioned optimization problems that can be solved accurately with few model evaluations only. Thus, at higher levels, where the models are expensive to evaluate, only few model evaluations are necessary to obtain an accurate approximation of the solution of the optimization problem, which can lead to significant runtime savings while obtaining biasing densities that lead to a similar variance reduction as the biasing densities derived with the single-fidelity CE method that uses the high-fidelity model alone.

Multifidelity methods have been extensively used to speedup rare event probability estimation [30]. We distinguish here between three categories of multifidelity methods for rare event simulation. First, there are two-fidelity methods that use a single surrogate model and combine it with the high-fidelity model. The work [21, 20, 22] introduces a two-fidelity approach that switches between a single surrogate model and the high-fidelity model depending on the error of the surrogate model, which can lead to unbiased estimators if the error of the surrogate model is known. In [8], an error estimator of a reduced-basis model is used to decide whether to evaluate the reduced or the high-fidelity model. In [12], the zero-variance biasing density is approximated with a Kriging model. Unbiasedness of the estimator is guaranteed by using the Kriging model as a proxy in the biasing density only. Similarly, in [26], unbiased estimators of rare event probabilities are obtained by using a surrogate model to construct biasing densities and the high-fidelity model to derive the actual estimates.

Second, there are methods that use a multilevel hierarchy of models of a single type to speedup the estimation. Typically, these methods are developed for high-fidelity models that stem from partial differential equations (PDEs). The model hierarchy then often corresponds to different discretizations of the underlying PDE. There are extensions [13, 14, 15] of the multilevel Monte Carlo method [19, 18] for rare event probability estimation, which are based on variance reduction with control variates, instead of importance sampling. The subset method [1, 42] is another approach that has been extended to exploit a hierarchy of coarse-grid approximations in [40]. The subset method has also been combined with classification

methods of machine learning such as support vector machines and neural networks in, e.g., [5, 25].

Third, multifidelity methods have been proposed that use multiple surrogate models of any type and combine them with the high-fidelity model. The method introduced in [24, 29] uses a control variate framework based on multiple surrogate models to accelerate the Monte Carlo estimation of statistics of the outputs of the high-fidelity model; however, the approach does not target rare event probability estimation. The multifidelity approach presented in [27] uses multiple surrogate models for speeding up the construction of biasing densities in importance sampling and guarantees unbiased estimators of the rare event probabilities by using the high-fidelity model to derive the estimate; however, the approach proposed in [27] has not been demonstrated on small probabilities below 10^{-6}. The new multifidelity approach proposed in this paper also falls in this third category of multifidelity methods because we aim to exploit a hierarchy of surrogate models of any type. In contrast to [24, 29, 27], our approach explicitly targets rare event probabilities and we show that we successfully estimate probabilities as low as $\approx 10^{-9}$.

Section 2 of this paper provides preliminaries and the problem setup. Section 3 introduces our multifidelity preconditioner for the CE method, provides an error analysis, and summarizes our multifidelity approach in Algorithm 1. Section 4 demonstrates that our multifidelity approach achieves up to two orders of magnitude speedup compared to the single-fidelity CE method in a thermal and a reacting-flow example. Section 5 gives concluding remarks.

2. Importance sampling with the cross-entropy method.

We first introduce the problem setup and then discuss importance sampling with the classical CE method that uses a single model alone.

2.1. Notation and problem setup.

Let the value of the function $f : \mathcal{D} \to \mathbb{R}$ denote the system response to an input $\boldsymbol{z} \in \mathcal{D}$ with the input domain $\mathcal{D} \subset \mathbb{R}^d$ in $d \in \mathbb{N}$ dimensions. For example, if the system of interest is a cantilever beam, then the input could define material properties and the system response could be the displacement of the tip of the beam. Let $Z : \Omega \to \mathcal{D}$ be a random variable with sample space Ω and with probability density function p. We denote a realization of Z as $\boldsymbol{z} \in \mathcal{D}$.

Let $t \in \mathbb{R}$ with $t > 0$ be a rare event threshold and define the rare event probability as

$$P_t = \mathbb{P}_p[f \leq t] = \int_{\mathcal{D}} I_t(\boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \,,$$

with the indicator function $I_t : \mathcal{D} \to \{0, 1\}$ defined as

$$I_t(\boldsymbol{z}) = \begin{cases} 1 \,, & f(\boldsymbol{z}) \leq t \,, \\ 0 \,, & f(\boldsymbol{z}) > t \end{cases} .$$

Note that $P_t = \mathbb{E}_p[I_t]$, where $\mathbb{E}_p$ denotes the expected value with respect to p. Let $\mathrm{Var}_p[I_t]$ be the variance of I_t with respect to p and assume $\mathrm{Var}_p[I_t] \in \mathbb{R}$ such that $\mathrm{Var}_p[I_t] = P_t(1 - P_t)$. Let $\rho \in (0, 1)$ and define the ρ-quantile of Z as $\gamma \in \mathcal{D}$, i.e.,

$$\mathbb{P}_p[Z \leq \gamma] = \rho \,.$$

Note that t is the P_t-quantile of $f(Z)$.

Consider now models $f^{(\ell)} : \mathcal{D} \to \mathbb{R}$ of the system of interest, where $\ell \in \mathbb{N}$ is a level parameter. For example, the models $f^{(\ell)}$ can be derived via finite-element discretization from the governing equations of the system of interest; in this case, the level parameter ℓ determines the mesh width. Note, however, that we will also consider models where the level parameter refers to more general concepts than mesh widths, e.g., the number of reduced basis vectors in reduced models and the number of data points in support vector regression machines. The costs of evaluating a model $f^{(\ell)}$ are denoted as $0 < w_\ell \in \mathbb{R}$. For each model $f^{(\ell)}$, we define the indicator function $I_t^{(\ell)} : \mathcal{D} \to \{0, 1\}$ as

$$I_t^{(\ell)}(\boldsymbol{z}) = \begin{cases} 1\,, & f^{(\ell)}(\boldsymbol{z}) \leq t\,, \\ 0\,, & f^{(\ell)}(\boldsymbol{z}) > t\,, \end{cases}$$

with the rare event threshold t. The rare event probability with respect to a model $f^{(\ell)}$ is $P_t^{(\ell)} = \mathbb{P}_p[f^{(\ell)} \leq t]$. In the following, we choose a maximal level $L \in \mathbb{N}$ such that the indicator function $I_t^{(L)}$ leads to a rare event probability $P_t^{(L)}$ that is a sufficiently accurate approximation of the rare event probability P_t of the system of interest for the current application at hand.

Let $\hat{P}_t^{(L)}$ be an unbiased estimator of the rare event probability $P_t^{(L)}$. We assess the quality of an estimator with respect to its error and costs. We measure the error of an estimator $\hat{P}_t^{(L)}$ with the squared coefficient of variation

$$e(\hat{P}_t^{(L)}) = \frac{\mathrm{Var}_p[\hat{P}_t^{(L)}]}{\left(\mathbb{E}_p[\hat{P}_t^{(L)}]\right)^2}\,. \tag{1}$$

The costs $c(\hat{P}_t^{(L)})$ are quantified with the costs of the model evaluations required by the estimator.

2.2. Standard Monte Carlo estimators. Let $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \in \mathcal{D}$ be $m \in \mathbb{N}$ realizations of the random variable Z and let

$$\hat{P}_t^{\mathrm{MC}} = \frac{1}{m} \sum_{i=1}^{m} I_t^{(L)}(\boldsymbol{z}_i) \tag{2}$$

be the standard Monte Carlo estimator of $P_t^{(L)}$. The squared coefficient of variation $e(\hat{P}_t^{\mathrm{MC}})$ of $\hat{P}_t^{\mathrm{MC}}$ is

$$e(\hat{P}_t^{\mathrm{MC}}) = \frac{1 - P_t^{(L)}}{m P_t^{(L)}}\,.$$

To achieve $e(\hat{P}_t^{(L)}) < \epsilon$ for a given tolerance $0 < \epsilon \in \mathbb{R}$, the standard Monte Carlo estimator requires

$$m > \frac{1 - P_t^{(L)}}{\epsilon P_t^{(L)}}$$

evaluations of $I_t^{(L)}$, and thus m evaluations of $f^{(L)}$. Since m depends inverse proportionally on the rare event probability $P_t^{(L)}$, the number of evaluations m can become large for small rare event probabilities, which means that standard Monte Carlo estimators become computationally infeasible if the costs w_L of evaluating $f^{(L)}$ are high.

2.3. Importance sampling with the cross-entropy method. Importance sampling estimators draw samples from a problem-dependent biasing distribution with the aim of reducing the variance compared to standard Monte Carlo estimators. This section discusses the CE method that iteratively constructs biasing distributions.

2.3.1. Importance sampling. Let $\operatorname{supp}(p) = \{\boldsymbol{z} \in \mathcal{D} \,|\, p(\boldsymbol{z}) > 0\}$ be the support of the density p. For a biasing density q with $\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$, the importance sampling estimator $\hat{P}_t^{\mathrm{IS}}$ of $P_t^{(L)}$ is

$$\hat{P}_t^{\mathrm{IS}} = \frac{1}{m} \sum_{i=1}^{m} I_t^{(L)}(\boldsymbol{z}_i') \frac{p(\boldsymbol{z}_i')}{q(\boldsymbol{z}_i')},$$

with m realizations $\boldsymbol{z}_1', \dots, \boldsymbol{z}_m'$ of the random variable Z_q with the biasing density q. Because $\operatorname{supp}(p) \subseteq \operatorname{supp}(q)$, and if the variance of the importance sampling estimator

$$\operatorname{Var}_q[\hat{P}_t^{\mathrm{IS}}] = \frac{1}{m} \operatorname{Var}_q\left[I_t^{(L)} \frac{p}{q} \right]$$

is finite, then the importance sampling estimator $\hat{P}_t^{\mathrm{IS}}$ is an unbiased estimator of $P_t^{(L)}$. The biasing density q^* that minimizes the variance $\operatorname{Var}_q[\hat{P}_t^{\mathrm{IS}}]$ is

$$q^*(\boldsymbol{z}) = \frac{I_t^{(L)}(\boldsymbol{z}) p(\boldsymbol{z})}{P_t^{(L)}},$$

which leads to an importance sampling estimator with variance 0. The density q^*, however, depends on $P_t^{(L)}$, which is the quantity we want to estimate.

2.3.2. CE-optimal biasing density. The CE method [34, 33, 35, 10] provides a practical way of approximating the zero-variance density q^*. Consider a set of parametrized densities $\mathcal{Q} = \{q_{\boldsymbol{v}} \,|\, \boldsymbol{v} \in \mathcal{P}\}$, where $\boldsymbol{v} \in \mathcal{P}$ is a parameter in the set $\mathcal{P}$. For example, $\mathcal{Q}$ could be the set of normal distributions with the parameter $\boldsymbol{v}$ corresponding to the mean and covariance matrix. To ease the presentation, we assume in the following without loss of generality that the nominal density p of the random variable Z is in the set $\mathcal{Q}$. The CE method optimizes for a parameter $\boldsymbol{v}_* \in \mathcal{P}$ such that the corresponding density $q_{\boldsymbol{v}_*} \in \mathcal{Q}$ minimizes the Kullback-Leibler divergence (also called the cross entropy) from the zero-variance density q^*. Transformations show that a solution of the problem

$$\boldsymbol{v}_* = \arg\max_{\boldsymbol{v} \in \mathcal{P}} \mathbb{E}_p[I_t^{(L)} \log(q_{\boldsymbol{v}})] \tag{3}$$

is a parameter $\boldsymbol{v}_*$ that corresponds to a CE-optimal density $q_{\boldsymbol{v}_*}$, see, e.g., [10]. Solving the stochastic counterpart of (3)

$$\max_{\hat{\boldsymbol{v}} \in \mathcal{Q}} \frac{1}{m} \sum_{i=1}^{m} I_t^{(L)}(\boldsymbol{z}_i) \log(q_{\hat{\boldsymbol{v}}}(\boldsymbol{z}_i)), \tag{4}$$

with realizations $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ of Z, typically fails, because the stochastic counterpart (4) is affected by the rareness of $I_t^{(L)}(Z)$, just as the standard Monte Carlo estimator (2).

In [34, 33, 35], the CE method is proposed. The CE method iteratively derives an estimate $\hat{\boldsymbol{v}}_*$ of the solution of the optimization problem (3). Our description of the CE method follows [10]. Consider the first iteration $k = 1$. In the first iteration, the CE method is initialized with the nominal random variable Z and the nominal density $p \in \mathcal{Q}$. Define the rare event threshold for the first iteration $t_1 \in \mathbb{R}$ to be the ρ-quantile of the distribution of $f^{(L)}(Z)$, where $\rho \in (0, 1)$ is a parameter that is typically in the range $[10^{-2}, 10^{-1}]$. Note that typically $t_1 > t$. Then, a solution $\hat{\boldsymbol{v}}_1 \in \mathcal{Q}$ of the optimization problem

$$\max_{\hat{\boldsymbol{v}} \in \mathcal{Q}} \frac{1}{m} \sum_{i=1}^{m} I_{t_1}^{(L)}(\boldsymbol{z}_i) \log(q_{\hat{\boldsymbol{v}}}(\boldsymbol{z}_i)) \,, \qquad \boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \sim Z \,, \tag{5}$$

is obtained, where $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \sim Z$ denotes that $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ are realizations of Z. The optimization problem (5) uses the indicator function $I_{t_1}^{(L)}$ with threshold t_1 instead of t, and therefore (5) avoids the rare event induced by the original threshold t. Note that the gradient of $\log(q_{\hat{\boldsymbol{v}}})$ with respect to the parameter $\hat{\boldsymbol{v}}$ is known analytically for certain sets of distributions $\mathcal{Q}$, see Section 3.5. In the second iteration $k = 2$, the threshold t_2 is selected with respect to the distribution of $f^{(L)}(Z_1)$, where Z_1 is the random variable with density $q_{\hat{\boldsymbol{v}}_1}$ derived in the first iteration. To guarantee termination of the CE method, the threshold t_2 is set to the minimum of the ρ-quantile of $f^{(L)}(Z_1)$ and $t_1 - \delta$, where $0 < \delta \in \mathbb{R}$ is a small constant [11, 10]. Then, the parameter $\hat{\boldsymbol{v}}_2$ is derived from the optimization problem

$$\max_{\hat{\boldsymbol{v}} \in \mathcal{Q}} \frac{1}{m} \sum_{i=1}^{m} I_{t_2}^{(L)}(\boldsymbol{z}_i) \frac{p(\boldsymbol{z}_i)}{q_{\hat{\boldsymbol{v}}_1}(\boldsymbol{z}_i)} \log(q_{\hat{\boldsymbol{v}}}(\boldsymbol{z}_i)) \,, \qquad \boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \sim Z_1 \,,$$

which is formulated with respect to the indicator function $I_{t_2}^{(L)}$ that depends on the threshold t_2. This process is continued until step $K \in \mathbb{N}$ where $t_K \leq t$, and where an estimate $\hat{\boldsymbol{v}}_*$ of the CE-optimal parameter $\boldsymbol{v}_*$ is obtained.

The CE method depends on two parameters: the quantile parameter ρ that determines the ρ-quantile for selecting the thresholds $t_1, t_2, t_3, \ldots, t_K$, and the minimal-step-size parameter δ that defines the minimal reduction of the threshold in each iteration. With the parameter δ, the CE method terminates after at most

$$K = \frac{t_1 - t}{\delta} \tag{6}$$

iterations with an estimate $\hat{\boldsymbol{v}}_*$ of $\boldsymbol{v}_*$. Thus, K is an upper bound on the number of CE iterations. Note that we sometimes use K but implicitly mean $\lceil K \rceil$ to get an integer number. Details on the CE method, including a convergence analysis, are given in [10, 11]. A cross-entropy method that optimizes for non-parametric densities, i.e., where it is unnecessary to specify a family $\mathcal{Q}$ of parametrized distributions, is introduced in [36] and further extended in [3, 4].

In each iteration $k = 1, \ldots, K$ of the CE method, the model $f^{(L)}$ is evaluated at m realizations. Therefore, a bound on the costs of deriving an importance-sampling estimate

$\hat{P}_t^{\mathrm{IS}}$ of $P_t^{(L)}$ with the CE method from m samples is

$$c(\hat{P}_t^{\mathrm{IS}}) \leq Kmw_L\,.$$

The squared coefficient of variation of the importance sampling estimator $\hat{P}_t^{\mathrm{IS}}$ depends on the variance reduction achieved by the biasing density

$$e(\hat{P}_t^{\mathrm{IS}}) = \frac{\mathrm{Var}_{\hat{\boldsymbol{v}}_*}\left[I_t^{(L)}\frac{p}{q_{\hat{\boldsymbol{v}}_*}}\right]}{\left(\mathbb{E}_{\hat{\boldsymbol{v}}_*}\left[\hat{P}_t^{\mathrm{IS}}\right]\right)^2 m}\,. \tag{7}$$

Note that we abbreviate $\mathbb{E}_{q_{\hat{\boldsymbol{v}}_*}}$ and $\mathrm{Var}_{q_{\hat{\boldsymbol{v}}_*}}$ with $\mathbb{E}_{\hat{\boldsymbol{v}}_*}$ and $\mathrm{Var}_{\hat{\boldsymbol{v}}_*}$, respectively, in (7) and in the following.

3. A multifidelity preconditioner for the cross-entropy method. We propose a multifidelity-preconditioned CE (MFCE) method that exploits a hierarchy of models $f^{(1)},\ldots,f^{(L)}$ to reduce the runtime of constructing a biasing density compared to the classical, single-fidelity CE method that uses $f^{(L)}$ only. Section 3.1 introduces our MFCE approach. Section 3.2 and Section 3.3 formalize our MFCE method and present an analysis of the savings obtained with our MFCE method compared to the classical, single-fidelity CE method in terms of the bounds on the number of CE iterations. Section 3.4 summarizes the MFCE method in Algorithm 1, and Section 3.5 provides practical considerations.

3.1. The MFCE method. Let p be the nominal density and let $q \in \mathcal{Q}$ be a density in $\mathcal{Q}$. Consider the classical, single-fidelity CE method that uses model $f^{(L)}$ alone, as discussed in Section 2.3.2. Let the CE method be initialized with density p and let t_p be the ρ-quantile of $f^{(L)}(Z)$. Then, the bound $K_p = (t_p - t)/\delta$ on the number of CE iterations is obtained from (6). Similarly, the bound on the number of CE iterations is $K_q = (t_q - t)/\delta$ if the CE method is initialized with q, where t_q is the ρ-quantile of $f^{(L)}(Z_q)$ and where Z_q is a random variable with density q. This shows that the bound on the number of iterations of the CE method depends on the density with which the CE method is initialized. If $t_q \leq t_p$, then the bound on the number of iterations of the CE method initialized with q is lower or equal than the bound on the number of iterations of the CE method initialized with p.

We propose to exploit that the bound on the number of CE iterations can be reduced by a suitable choice of the density with which the CE method is initialized. Our MFCE method iterates through the levels $\ell = 1,\ldots,L$. At level $\ell = 1$, our MFCE method constructs a biasing density $q_{\hat{\boldsymbol{v}}_*^{(1)}}$ with parameter $\hat{\boldsymbol{v}}_*^{(1)} \in \mathcal{P}$ with the classical CE method, initialized with the nominal density p and using model $f^{(1)}$. At level $\ell = 2$, our MFCE method uses the CE method to derive a density $q_{\hat{\boldsymbol{v}}_*^{(2)}}$ with model $f^{(2)}$; however, the CE method on level $\ell = 2$ is initialized with the density $q_{\hat{\boldsymbol{v}}_*^{(1)}}$ of the previous level, instead of the nominal density p as in the classical CE method. This hierarchical process is continued until level $\ell = L$, where density $q_{\hat{\boldsymbol{v}}_*^{(L-1)}}$ and model $f^{(L)}$ are used to derive density $q_{\hat{\boldsymbol{v}}_*^{(L)}}$.

3.2. Effect of the MFCE preconditioning. Consider now our MFCE method on level ℓ, where we have obtained an estimate $\hat{\boldsymbol{v}}_*^{(\ell)}$ and the corresponding biasing density $q_{\hat{\boldsymbol{v}}_*^{(\ell)}}$. Using

$q_{\hat{\boldsymbol{v}}_*^{(\ell)}}$ on level $\ell+1$ to obtain an estimate of the solution of the stochastic counterpart of

$$\max_{\boldsymbol{v}\in\mathcal{Q}} \mathbb{E}_{\hat{\boldsymbol{v}}_*^{(\ell)}}\left[I_t^{(\ell+1)} \frac{p}{q_{\hat{\boldsymbol{v}}_*^{(\ell)}}} \log(q_{\boldsymbol{v}}) \right]$$

means that in the first CE iteration on level $\ell+1$ the threshold parameter $t_1^{(\ell+1)}$ is set to the ρ-quantile of $f^{(\ell+1)}(Z_\ell)$, where Z_ℓ is a random variable with density $q_{\hat{\boldsymbol{v}}_*^{(\ell)}}$. In contrast, the classical CE method uses the ρ-quantile of $f^{(\ell+1)}(Z)$ instead, where Z corresponds to the nominal density. If the ρ-quantile $t_1^{(\ell+1)}$ of $f^{(\ell+1)}(Z_\ell)$ is smaller than the ρ-quantile of $f^{(\ell+1)}(Z)$, then the bound on the iterations of the CE method on level $\ell+1$ is smaller if the CE method is initialized with $q_{\hat{\boldsymbol{v}}_*^{(\ell)}}$ than if the CE method is initialized with the nominal density p. The following proposition formalizes this notion.

Proposition 1. *Let $\ell \in \mathbb{N}$ and let $q_{\hat{\boldsymbol{v}}_*^{(\ell)}} \in \mathcal{Q}$ be the biasing density obtained with the CE method on level ℓ of our MFCE approach. Let further $t_1^{(\ell+1)}$ be the ρ-quantile of $f^{(\ell+1)}(Z_\ell)$ with respect to the random variable Z_ℓ with density $q_{\hat{\boldsymbol{v}}_*^{(\ell)}}$. If*

$$\mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}\left[f^{(\ell+1)} \leq t_1^{(\ell+1)} \right] \geq \mathbb{P}_p\left[f^{(\ell+1)} \leq t_1^{(\ell+1)} \right], \tag{8}$$

then the bound on the number of iterations of the CE method initialized with $q_{\hat{\boldsymbol{v}}_*^{(\ell)}}$ on level $\ell+1$ of our MFCE approach is less or equal to the bound of the classical CE method initialized with the nominal density p.

Proof. Let t_p be the ρ-quantile of $f^{(\ell+1)}(Z)$, then we obtain with the monotonicity of the cumulative distribution function of $f^{(\ell+1)}(Z)$ and (8) that $t_p \geq t_1^{(\ell+1)}$. The proposition follows with (6). ∎

3.3. Error analysis of multifidelity approach. Proposition 1 states under which condition the bound on the number of iterations of our MFCE approach is lower than the bound on the number of iterations of the classical CE method. In this section, we analyze which properties of the models $f^{(1)}, \ldots, f^{(L)}$ are required such that the condition (8) of Proposition 1 is met. The following analysis is based on the framework introduced in [13, 14]. We first make similar assumptions on the models as in [13, 14].

Assumption 1. *Let $0 < \alpha < 1$ and let t be a threshold parameter. The models $f^{(\ell)}$ satisfy*

$$|f^{(\ell)}(\boldsymbol{z}) - f^{(\ell+1)}(\boldsymbol{z})| \leq \alpha^\ell \text{ or } |f^{(\ell)}(\boldsymbol{z}) - f^{(\ell+1)}(\boldsymbol{z})| \leq |f^{(\ell)}(\boldsymbol{z}) - t|, \qquad \boldsymbol{z} \in \mathcal{D},$$

for $\ell = 1, \ldots, L-1$.

Assumption 2. *Consider a density $q \in \mathcal{Q}$ and the corresponding random variable Z_q. Let further $F_q^{(\ell)}$ be the cumulative distribution function of $f^{(\ell)}(Z_q)$ for $\ell = 1, \ldots, L$. The cumulative distribution function $F_q^{(\ell)}$ is Lipschitz continuous with Lipschitz constant $C_q^{(\ell)}$. Furthermore, there exists a constant $C \in \mathbb{R}$ that bounds $C_q^{(\ell)} \leq C$ for all $\ell = 1, \ldots, L$ and $q \in \mathcal{Q}$. We therefore have*

$$|F_q^{(\ell)}(t) - F_q^{(\ell)}(t')| \leq C|t - t'|,$$

where $t, t' \in \mathbb{R}$.

Assumption 1 is an assumption on the accuracy of the models $f^{(1)}, \dots, f^{(L)}$. In many settings in numerical analysis, accuracy assumptions of the form $|f^{(\ell)}(\boldsymbol{z}) - f^{(\ell+1)}(\boldsymbol{z})| \leq \alpha^\ell$ need to hold for all $\boldsymbol{z} \in \mathcal{D}$. Such a uniform assumption over $\mathcal{D}$ is too restrictive for our problem setup. We can tolerate large model errors at inputs $\boldsymbol{z} \in \mathcal{D}$ that lead to model outputs far away from our threshold t and need a high accuracy only at inputs that lead to model outputs that are close to t. Assumption 1 allows a large deviation of $f^{(\ell+1)}$ from $f^{(\ell)}$ in regions of $\mathcal{D}$ for which $f^{(\ell)}$ is far from t, see the right inequality in Assumption 1. The error $|f^{(\ell)}(\boldsymbol{z}) - f^{(\ell+1)}(\boldsymbol{z})|$ has to be low only in regions of $\mathcal{D}$ that lead to model outputs near t. We refer to [13, 14], where Assumption 1 is discussed in detail and further motivated. In particular, the work [13, 14] presents a selective refinement algorithm to establish Assumption 1, which can be applied in our setting as well.

Consider now Assumption 2 and note that $F_q^\ell(t') - F_q^{(\ell)}(t) = \mathbb{P}[t \leq f^{(\ell)}(Z_q) \leq t']$ for $t \leq t'$. With the monotonicity of the cumulative distribution function, we obtain that the probability $\mathbb{P}[t \leq f^{(\ell)}(Z_q) \leq t']$ decreases with $|t - t'|$. Assumption 2 relates the decrease of $\mathbb{P}[t \leq f^{(\ell)}(Z_q) \leq t']$ to the decrease in $|t - t'|$. To gain intuition for the constant C in Assumption 2, assume the cumulative distribution functions in Assumption 2 are absolutely continuous and let $g_q^{(\ell)}$ be the density of the random variable $f^{(\ell)}(Z_q)$. If there exists a bound $\bar{g}^{(\ell)} \in \mathbb{R}$ with $g_q^{(\ell)}(\xi) \leq \bar{g}^{(\ell)}$ for all $\xi \in \mathbb{R}$ and all $q \in \mathcal{Q}$, then we obtain

$$|F_q^\ell(t) - F_q^{(\ell)}(t')| = \mathbb{P}[t \leq f^{(\ell)}(Z_q) \leq t'] = \int_t^{t'} g_q^{(\ell)}(\xi) \mathrm{d}\xi \leq \bar{g}^\ell |t - t'|\,,$$

with $t \leq t'$. Thus, in this case, Assumption 2 is satisfied by setting the constant C to the maximum of the bounds $\bar{g}^{(1)}, \dots, \bar{g}^{(L)}$. The mean value theorem shows that there exists $\bar{t} \in [t, t']$ such that

$$\frac{F_q^{(\ell)}(t') - F_q^{(\ell)}(t)}{t' - t} = g_q^{(\ell)}(\bar{t})\,,$$

which gives further intuition for the constant $\bar{g}^{(\ell)}$. Assumption 2 is used in [14] in a similar context.

Under Assumption 1 and Assumption 2 we obtain the following proposition.

Proposition 2. *Let $\ell \in \{1, \dots, L-1\}$ and let $\hat{\boldsymbol{v}}_*^{(\ell)}$ be the parameter of the biasing density estimated on level ℓ. Let further $t_1^{(\ell+1)}$ be the ρ-quantile with $\mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell+1)} \leq t_1^{(\ell+1)}] = \rho$ and let $\mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell)} \leq t_1^{(\ell+1)}] \geq \mathbb{P}_p[f^{(\ell)} \leq t_1^{(\ell+1)}]$, then*

$$\mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell+1)} \leq t_1^{(\ell+1)}] \geq \mathbb{P}_p[f^{(\ell+1)} \leq t_1^{(\ell+1)}] - 8C\alpha^\ell\,, \tag{9}$$

where α and C are the constants of Assumptions 1–2.

Before we prove Proposition 2, we first show Lemma 3 and Lemma 4.

Lemma 3. *Define $\gamma = t_1^{(\ell+1)}$ and the set $\mathcal{B} = \{\boldsymbol{z} \in \mathcal{D} : |f^{(\ell)}(\boldsymbol{z}) - \gamma| \leq \alpha^\ell\}$. With Assumption 1 follows that $I_\gamma^{(\ell)}(\boldsymbol{z}) = I_\gamma^{(\ell+1)}(\boldsymbol{z})$ for $\boldsymbol{z} \in \mathcal{D} \setminus \mathcal{B}$.*

Proof. This proof follows similar arguments as the proof of Lemma 3.3 in [14]. We show that $f^{(\ell)}(\boldsymbol{z}) \leq \gamma \iff f^{(\ell+1)}(\boldsymbol{z}) \leq \gamma$ holds for $\boldsymbol{z} \in \mathcal{D} \setminus \mathcal{B}$, which is equivalent to $I_\gamma^{(\ell)}(\boldsymbol{z}) = I_\gamma^{(\ell+1)}(\boldsymbol{z})$ for $\boldsymbol{z} \in \mathcal{D} \setminus \mathcal{B}$. Consider first $f^{(\ell)}(\boldsymbol{z}) \leq \gamma \Rightarrow f^{(\ell+1)}(\boldsymbol{z}) \leq \gamma$. We obtain

$$0 \leq \gamma - f^{(\ell)}(\boldsymbol{z}) \leq \gamma - f^{(\ell+1)}(\boldsymbol{z}) + |f^{(\ell+1)}(\boldsymbol{z}) - f^{(\ell)}(\boldsymbol{z})| \leq \gamma - f^{(\ell+1)}(\boldsymbol{z}) + |f^{(\ell)}(\boldsymbol{z}) - \gamma| \,, \tag{10}$$

because for all $\boldsymbol{z} \in \mathcal{D}\setminus\mathcal{B}$ we have $|f^{(\ell)}(\boldsymbol{z})-\gamma| > \alpha^\ell$ by definition of $\mathcal{B}$ and therefore $|f^{(\ell+1)}(\boldsymbol{z}) - f^{(\ell)}(\boldsymbol{z})| \leq |f^{(\ell)}(\boldsymbol{z}) - \gamma|$ because of Assumption 1. Since $0 \leq \gamma - f^{(\ell)}(\boldsymbol{z})$, we have $\gamma - f^{(\ell)}(\boldsymbol{z}) = |\gamma - f^{(\ell)}(\boldsymbol{z})|$, and therefore follows from (10)

$$|\gamma - f^{(\ell)}(\boldsymbol{z})| \leq \gamma - f^{(\ell+1)}(\boldsymbol{z}) + |f^{(\ell)}(\boldsymbol{z}) - \gamma| \,. \tag{11}$$

Subtracting $|\gamma - f^{(\ell)}(\boldsymbol{z})|$ on both sides of (11) leads to

$$0 \leq \gamma - f^{(\ell+1)}(\boldsymbol{z}) \,,$$

which shows $f^{(\ell)}(\boldsymbol{z}) \leq \gamma \Rightarrow f^{(\ell+1)}(\boldsymbol{z}) \leq \gamma$. For $f^{(\ell+1)}(\boldsymbol{z}) \leq \gamma \Rightarrow f^{(\ell)}(\boldsymbol{z}) \leq \gamma$, we show $f^{(\ell)}(\boldsymbol{z}) \geq \gamma \Rightarrow f^{(\ell+1)}(\boldsymbol{z}) \geq \gamma$ with similar arguments. We obtain

$$0 \leq f^{(\ell)}(\boldsymbol{z}) - \gamma \leq |f^{(\ell)}(\boldsymbol{z}) - f^{(\ell+1)}(\boldsymbol{z})| + f^{(\ell+1)}(\boldsymbol{z}) - \gamma \leq |f^{(\ell)}(\boldsymbol{z}) - \gamma| + f^{(\ell+1)}(\boldsymbol{z}) - \gamma \,. \tag{12}$$

Since $f^{(\ell)}(\boldsymbol{z}) - \gamma \geq 0$, we have $f^{(\ell)}(\boldsymbol{z}) - \gamma = |f^{(\ell)}(\boldsymbol{z}) - \gamma|$. Subtracting $|f^{(\ell)}(z) - \gamma|$ from the inequality in (12), shows $0 \leq f^{(\ell+1)}(\boldsymbol{z}) - \gamma$. ∎

Lemma 4. *With Assumptions 1–2, we obtain*

$$P_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell)} \leq \gamma] \leq P_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell+1)} \leq \gamma] + 4C\alpha^\ell \,, \tag{13}$$

where $\gamma = t_1^{(\ell+1)}$ *as in Lemma 3.*

Proof. Let $\mathcal{B}$ be the set defined in Lemma 3. For $\boldsymbol{z} \in \mathcal{B}$, we obtain with Assumption 1 that $|f^{(\ell+1)}(\boldsymbol{z}) - f^{(\ell)}(\boldsymbol{z})| \leq \alpha^\ell$ and $|f^{(\ell+1)}(\boldsymbol{z}) - \gamma| \leq 2\alpha^\ell$. Consider now $\mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell)} \leq \gamma]$, which we write as

$$\begin{aligned}
\mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell)} \leq \gamma] &= \int_{\mathcal{B}} I_\gamma^{(\ell)}(\boldsymbol{z}) q_{\hat{\boldsymbol{v}}_*^{(\ell)}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \int_{\mathcal{D}\setminus\mathcal{B}} I_\gamma^{(\ell)}(\boldsymbol{z}) q_{\hat{\boldsymbol{v}}_*^{(\ell)}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \\
&= \int_{\mathcal{B}} I_\gamma^{(\ell)}(\boldsymbol{z}) q_{\hat{\boldsymbol{v}}_*^{(\ell)}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \int_{\mathcal{D}\setminus\mathcal{B}} I_\gamma^{(\ell+1)}(\boldsymbol{z}) q_{\hat{\boldsymbol{v}}_*^{(\ell)}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \qquad (14) \\
&\leq \int_{\mathcal{B}} I_\gamma^{(\ell)}(\boldsymbol{z}) q_{\hat{\boldsymbol{v}}_*^{(\ell)}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell+1)} \leq \gamma] \,, \qquad (15)
\end{aligned}$$

where we obtain equality in (14) because $I_\gamma^{(\ell)}(\boldsymbol{z}) = I_\gamma^{(\ell+1)}(\boldsymbol{z})$ for $\boldsymbol{z} \in \mathcal{D} \setminus \mathcal{B}$, see Lemma 3, and $\leq$ in (15) because $I_\gamma^{(\ell+1)}$ is non-negative. Consider now the first term in (15), for which we obtain

$$\begin{aligned}
\int_{\mathcal{B}} I_\gamma^{(\ell)}(\boldsymbol{z}) q_{\hat{\boldsymbol{v}}_*^{(\ell)}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} &\leq \int_{\mathcal{B}} q_{\hat{\boldsymbol{v}}_*^{(\ell)}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \\
&\leq \mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}\left[|f^{(\ell+1)} - \gamma| \leq 2\alpha^\ell\right] \\
&= F_{\hat{\boldsymbol{v}}_*^{(\ell)}}^{(\ell+1)}(\gamma - 2\alpha^\ell) - F_{\hat{\boldsymbol{v}}_*^{(\ell)}}^{(\ell+1)}(\gamma + 2\alpha^\ell) \\
&\leq 4C\alpha^\ell \,, \qquad (16)
\end{aligned}$$

where we used Assumption 2 in (16). Combining the bound in (16) with (15) leads to (13). ∎

We now state the proof of Proposition 2.

Proof of Proposition 2. Let γ and $\mathcal{B}$ be defined as in Lemma 3. Consider $\mathbb{P}_p[f^{(\ell+1)} \leq \gamma]$, which we write as

$$\begin{aligned}
\mathbb{P}_p[f^{(\ell+1)} \leq \gamma] &= \int_{\mathcal{B}} I_\gamma^{(\ell+1)}(\boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \int_{\mathcal{D}\setminus\mathcal{B}} I_\gamma^{(\ell+1)}(\boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \\
&= \int_{\mathcal{B}} I_\gamma^{(\ell+1)}(\boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \int_{\mathcal{D}\setminus\mathcal{B}} I_\gamma^{(\ell)}(\boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \qquad (17) \\
&\leq \int_{\mathcal{B}} I_\gamma^{(\ell+1)}(\boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \mathbb{P}_p[f^{(\ell)} \leq \gamma] \\
&\leq \int_{\mathcal{B}} I_\gamma^{(\ell+1)}(\boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} + \mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell)} \leq \gamma] \qquad (18) \\
&\leq \int_{\mathcal{B}} I_\gamma^{(\ell+1)}(\boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} + 4C\alpha^\ell + \mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell+1)} \leq \gamma]\,, \qquad (19)
\end{aligned}$$

where we obtain equality in (17) because of Lemma 3, and $\leq$ in (18) because $\mathbb{P}_p[f^{(\ell)} \leq \gamma] \leq \mathbb{P}_{\hat{\boldsymbol{v}}_*^{(\ell)}}[f^{(\ell)} \leq \gamma]$ as assumed in the statement of Proposition 2. The inequality $\leq$ in (19) is obtained because of Lemma 4. Consider now the first term in (19), for which we obtain

$$\begin{aligned}
\int_{\mathcal{B}} I_\gamma^{(\ell+1)}(\boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} &\leq \int_{\mathcal{B}} p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \\
&\leq \mathbb{P}_p\left[|f^{(\ell+1)} - \gamma| \leq 2\alpha^\ell\right] \qquad (20) \\
&= F_p^{(\ell+1)}(\gamma - 2\alpha^\ell) - F_p^{(\ell+1)}(\gamma + 2\alpha^\ell) \\
&\leq 4C\alpha^\ell\,, \qquad (21)
\end{aligned}$$

where we used $|f^{(\ell+1)} - \gamma| \leq 2\alpha^\ell$ for $\boldsymbol{z} \in \mathcal{B}$ in (20) as in the proof of Lemma 4 and Assumption 2 in (21). Combining the bound in (21) with (19) leads to (9). ∎

Proposition 2 shows that with Assumptions 1–2 we obtain condition (8) of Proposition 1 up to the factor $8C\alpha^\ell$. Note that the factor $8C\alpha^\ell$ decays with the level ℓ, because we have $|\alpha| < 1$. With the parameter $\hat{\boldsymbol{v}}_*^{(L)}$ derived at level L, we define our MFCE estimator as

$$\hat{P}_t^{\mathrm{MFCE}} = \frac{1}{m} \sum_{i=1}^{m} I_t^{(L)}(\boldsymbol{z}_i) \frac{p(\boldsymbol{z}_i)}{q_{\hat{\boldsymbol{v}}_*^{(L)}}(\boldsymbol{z}_i)}\,,$$

where $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ are realizations of the random variable with density $q_{\hat{\boldsymbol{v}}_*^{(L)}}$. The MFCE estimator is unbiased with respect to the rare event probability $P_t^{(L)}$ if the biasing density defined by the parameter $\hat{\boldsymbol{v}}_*^{(L)}$ has a support that is a superset of the support of the nominal density p and the variance of the MFCE estimator is finite. The squared coefficient of variation of the MFCE estimator is

$$e(\hat{P}_t^{\mathrm{MFCE}}) = \frac{\mathrm{Var}_{\hat{\boldsymbol{v}}_*^{(L)}}\left[I_t^{(L)} \frac{p}{q_{\hat{\boldsymbol{v}}_*^{(L)}}}\right]}{\left(\mathbb{E}[\hat{P}_t^{\mathrm{MFCE}}]\right)^2 m}$$

Algorithm 1 Cross-entropy method with multifidelity preconditioning

procedure MFCE($f^{(1)}, \ldots, f^{(L)}$, t, p, m, ρ, δ)
  Initialize $\hat{\boldsymbol{v}}_*^{(0)} \in \mathcal{P}$ such that $q_{\hat{\boldsymbol{v}}_*^{(0)}} = p$; redefine $\mathcal{Q} = \mathcal{Q} \cup \{p\}$ if necessary
  for $\ell = 1, \ldots, L$ **do**
    Initialize $\hat{\boldsymbol{v}}_0^{(\ell)} = \hat{\boldsymbol{v}}_*^{(\ell-1)}$ and set $k = 0$
    while 1 **do**
      Draw realizations $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ from random variable with density $q_{\hat{\boldsymbol{v}}_k^{(\ell)}}$
      Compute model outputs $\mathcal{G}_k^{(\ell)} = \{f^{(\ell)}(\boldsymbol{z}_1), \ldots, f^{(\ell)}(\boldsymbol{z}_m)\}$
      Estimate ρ-quantile $\hat{\gamma}_k^{(\ell)}$ from $\mathcal{G}_k^{(\ell)}$
      if $k == 0$ **then** $t_0^{(\ell)} = \hat{\gamma}_k^{(\ell)} + \delta$
      end if
      Set $t_{k+1}^{(\ell)} = \max\{t, \min\{t_k^{(\ell)} - \delta, \hat{\gamma}_k^{(\ell)}\}\}$
      Estimate parameter $\hat{\boldsymbol{v}}_{k+1}^{(\ell)} \in \mathcal{P}$ by solving

$$\max_{\hat{\boldsymbol{v}} \in \mathcal{P}} \frac{1}{m} \sum_{i=1}^{m} I_{t_{k+1}^{(\ell)}}^{(\ell)}(\boldsymbol{z}_i) \frac{p(\boldsymbol{z}_i)}{q_{\hat{\boldsymbol{v}}_k^{(\ell)}}(\boldsymbol{z}_i)} \log\left(q_{\hat{\boldsymbol{v}}}(\boldsymbol{z}_i)\right) \tag{22}$$

      if $t_{k+1}^{(\ell)} == t$ **then break**
      end if
      Set $k = k + 1$
    end while
    Set $\hat{\boldsymbol{v}}_*^{(\ell)} = \hat{\boldsymbol{v}}_k^{(\ell)}$ and $\mathcal{G}_*^{(\ell)} = \mathcal{G}_k^{(\ell)}$
  end for
  Return estimate $\hat{P}_t^{\mathrm{MFCE}}$ with $\mathcal{G}_*^{(L)} = \{f^{(L)}(\boldsymbol{z}_1), \ldots, f^{(L)}(\boldsymbol{z}_m)\}$ and $\hat{\boldsymbol{v}}_*^{(L)}$ as

$$\hat{P}_t^{\mathrm{MFCE}} = \frac{1}{m} \sum_{i=1}^{m} I_t^{(L)}(\boldsymbol{z}_i) \frac{p(\boldsymbol{z}_i)}{q_{\hat{\boldsymbol{v}}_*^{(L)}}(\boldsymbol{z}_i)}$$

end procedure

and depends on the parameter $\hat{\boldsymbol{v}}_*^{(L)}$.

3.4. Computational procedure. Algorithm 1 summarizes our MFCE method. Inputs are the models $f^{(1)}, \ldots, f^{(L)}$, the threshold t, the nominal density p, the number of samples $m \in \mathbb{N}$, the quantile parameter ρ, and the minimal step size δ. Note that the parameters ρ and δ are the same parameters as in the classical, single-fidelity CE method, see Section 2.3.2. The `for` loop in line 3 iterates through the levels $\ell = 1, \ldots, L$. At each level ℓ, the density with parameter $\hat{\boldsymbol{v}}_*^{(\ell)}$ with respect to model $f^{(\ell)}$ is derived. At iteration $k = 0, 1, 2, \ldots$ of the `while` loop in line 5, realizations $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ are drawn from the random variable with the density $q_{\hat{\boldsymbol{v}}_k^{(\ell)}}$ of the current iteration, the model $f^{(\ell)}$ is evaluated at the realizations, and the

ρ-quantile is estimated from the model outputs. In line 9, the threshold $t_0^{(\ell)}$ is set to the ρ-quantile estimate $\hat{\gamma}_k^{(\ell)} + \delta$ in the first iteration of the loop. The threshold $t_{k+1}^{(\ell)}$ is selected in line 11, where the max{} operation guarantees that the threshold $t_{k+1}^{(\ell)}$ is greater or equal to t and the min{} operation guarantees that the threshold parameter is reduced by at least δ in each iteration of the `while` loop, except in the first and the last iteration. In line 12, the parameter $\hat{\boldsymbol{v}}_{k+1}^{(\ell)}$ is estimated. Line 13 exits the `while` loop if $t_{k+1}^{(\ell)}$ is equal to the threshold t. In line 17, the estimate $\hat{\boldsymbol{v}}_k^{(\ell)}$ and the model outputs $\mathcal{G}_k^{(\ell)}$ of the last iteration of the `while` loop are stored, and the `for` loop starts a new iteration for the next level $\ell + 1$. After the `for` loop iterated through all levels $\ell = 1, \ldots, L$, the MFCE estimate $\hat{P}_t^{\mathrm{MFCE}}$ is returned using the density $p_{\hat{\boldsymbol{v}}_*^{(L)}}$ and the model outputs in $\mathcal{G}_*^{(L)}$.

Typically, the computationally most expensive step of Algorithm 1 is the computation of the model outputs on line 7, which scales linearly with the number of realizations m. Solving the optimization problem (22) on line 12 typically incurs small costs if the gradients of the objective can be computed analytically. If the gradients of the objective have to be approximated numerically, then solving the optimization on line 12 can become expensive.

We use in Algorithm 1 the same number of realizations m for all models in the model hierarchy, i.e., in each iteration of the `for` loop in line 3 of Algorithm 1 the same number of realizations m are used. An adaptive selection of m might help to further reduce the costs of our MFCE approach. There are several options for selecting the number of realizations m adaptively. For example, the number of realizations within the ρ-quantile in line 8 of Algorithm 1 can be used to guide the selection of m. If only few realizations are within the ρ-quantile, then the estimation of the parameters in (22) can become inaccurate. Increasing the number of realizations m in such a situation can help to improve the estimation of the parameters and so lead to a more suitable biasing density. If many samples are within the ρ-quantile, then reducing m is unlikely to have a significant effect on the accuracy of the estimated parameter in (22) while the reduction of m saves costs. We also refer to [26] where similar strategies are discussed in the context of importance sampling.

3.5. Practical considerations. The gradient of the objective of the stochastic counterpart (22) can be derived analytically in many situations. In particular, if $\mathcal{Q}$ corresponds to distributions that belong to the natural exponential family, the gradient can be derived analytically [10, 37]. In the following, we will use Gaussian, log-normal, and Gamma distributions, for which we derive the gradients here.

Let the parameter $\boldsymbol{v} \in \mathcal{P}$ describe the mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ of a d-dimensional Gaussian distribution. The corresponding density is

$$q_{\boldsymbol{v}}(\boldsymbol{z}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{z}-\boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{z}-\boldsymbol{\mu}\right)\right),$$

where $|2\pi\boldsymbol{\Sigma}|$ denotes the determinant of the matrix $2\pi\boldsymbol{\Sigma}$. We obtain the gradient of $\log(q_{\boldsymbol{v}}(\boldsymbol{z}))$ with respect to $\boldsymbol{\mu}$ as

$$\nabla_{\boldsymbol{\mu}} \log(q_{\boldsymbol{v}}(\boldsymbol{z})) = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{z}-\boldsymbol{\mu}\right),$$

and the gradient with respect to $\boldsymbol{\Sigma}$ as

$$\nabla_{\boldsymbol{\Sigma}} \log(q_{\boldsymbol{v}}(\boldsymbol{z})) = -\frac{1}{2}\left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{z}-\boldsymbol{\mu}\right)\left(\boldsymbol{z}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\right) .$$

We use the gradients of $\log(q_{\boldsymbol{v}}(\boldsymbol{z}))$ and plug them into the gradient of the objective of the stochastic counterpart (22). Then, setting the gradient of the objective of (22) to zero leads to a system of equations that is linear in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Therefore, in the case where $\mathcal{Q}$ is the set of Gaussian distributions of dimension d, solving the optimization problem (22) means solving a system of linear equations. To find the parameters of a log-normal distribution, we fit a Gaussian to $\log(\boldsymbol{z})$. Consider now a one-dimensional Gamma distribution with density

$$q_{\boldsymbol{v}}(z) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} z^{\alpha-1}\mathrm{e}^{-z/\beta} ,$$

where Γ is the Gamma function, $\boldsymbol{v} = [\alpha, \beta]^T$, and α and β are the shape and scaling parameter, respectively. When we fit a Gamma distribution to data in the following, we keep the shape parameter α fixed and modify the scaling parameter β only. The gradient of $\log(q_{\boldsymbol{v}}(z))$ with respect to β can be obtained analytically and is

$$\nabla_{\beta} \log(q_{\boldsymbol{v}}(z)) = \frac{-\alpha\beta + z}{\beta^2} .$$

As in the Gaussian and the log-normal case, setting the gradient $\nabla_{\beta} \log(q_{\boldsymbol{v}}(z)) = 0$ and solving for the scaling parameter β leads to a system of linear equations.

4. Numerical results. We demonstrate the efficiency of our MFCE method on a heat transfer and a reacting flow example. In all of the following experiments, the quantile parameter is set to $\rho = 0.1$ and the minimal step size is $\delta = 10^{-2}$, which are similar to the parameters chosen in, e.g., [10]. Furthermore, we set $\mathcal{Q}$ to the set of Gaussian distributions of the respective dimensions, except if noted otherwise. We constrain the optimization problem (22) in Algorithm 1 to covariance matrices $\boldsymbol{\Sigma}$ with a minimal absolute value of 10^{-3}, which avoids convergence of the biasing distributions to outliers and single points, see [39] for a similar technique. All runtime measurements were performed on compute nodes with Intel Xeon E5-1620 and 32GB RAM on a single core using a MATLAB implementation.

4.1. Heat transfer. We consider rare event probability estimation with a one-dimensional heat problem with two inputs.

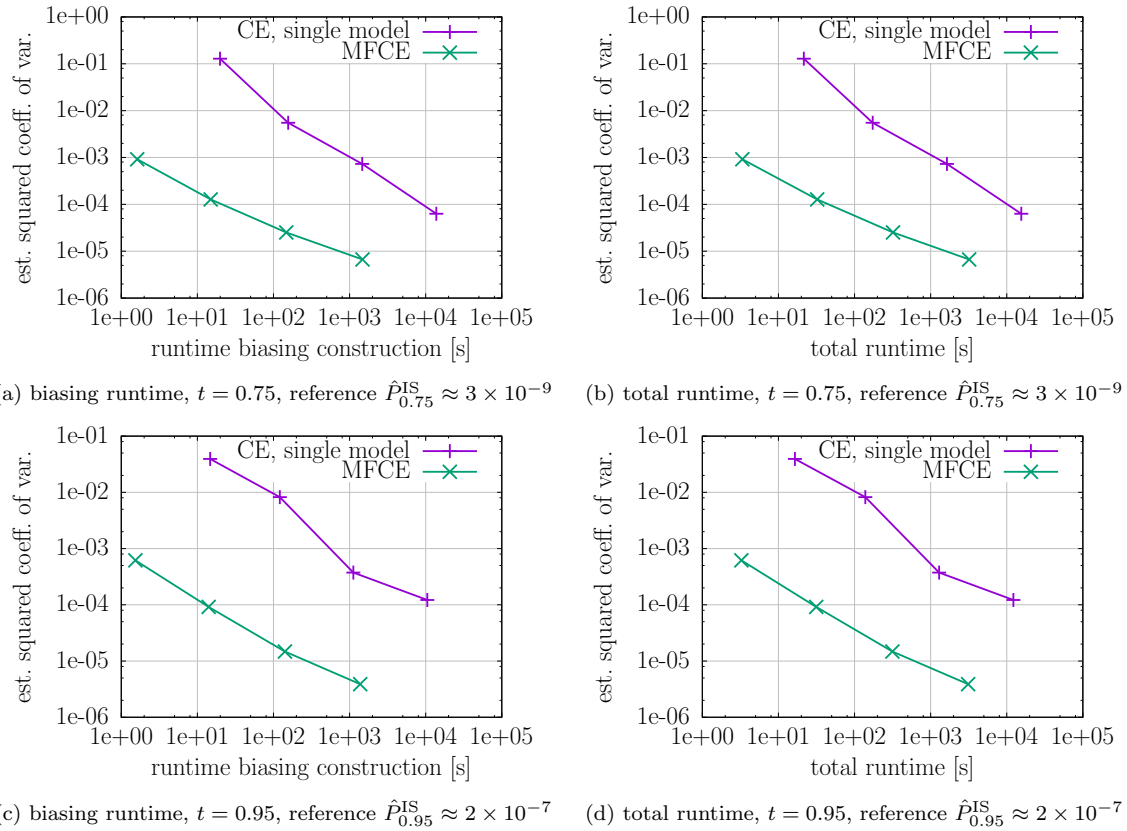
4.1.1. Problem setup. Let $\mathcal{X} = (0,1) \in \mathbb{R}$ be a domain with boundary $\partial\mathcal{X} = \{0,1\}$. Consider the linear elliptic PDE with random coefficients

$$-\nabla \cdot (a(\omega,x)\nabla u(\omega,x)) = 1 , \qquad x \in \mathcal{X} , \tag{23}$$

$$u(\omega,0) = 0 , \tag{24}$$

$$\partial_n u(\omega,1) = 0 , \tag{25}$$

where $u : \Omega \times \bar{\mathcal{X}} \to \mathbb{R}$ is the solution function defined on the set of outcomes Ω and where $\bar{\mathcal{X}}$ is the closure of $\mathcal{X}$. We impose homogeneous Dirichlet boundary condition on the left



(a) biasing runtime, $t = 0.75$, reference $\hat{P}^{\mathrm{IS}}_{0.75} \approx 3 \times 10^{-9}$

(b) total runtime, $t = 0.75$, reference $\hat{P}^{\mathrm{IS}}_{0.75} \approx 3 \times 10^{-9}$

(c) biasing runtime, $t = 0.95$, reference $\hat{P}^{\mathrm{IS}}_{0.95} \approx 2 \times 10^{-7}$

(d) total runtime, $t = 0.95$, reference $\hat{P}^{\mathrm{IS}}_{0.95} \approx 2 \times 10^{-7}$

Figure 1: Heat transfer: Our MFCE approach achieves up to two orders of magnitude speedup compared to using the single-fidelity CE method with the model $f^{(L)}$ for $t \in \{0.75, 0.95\}$.

boundary $x = 0$ of the domain $\mathcal{X}$ and homogeneous Neumann boundary conditions on the right boundary $x = 1$. The coefficient a is given as

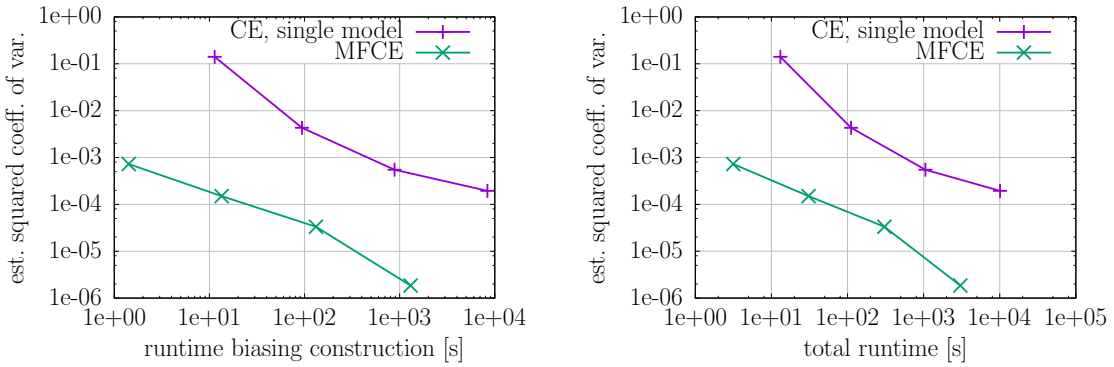
$$a(\omega, x) = \sum_{i=1}^{n} \exp\left(z_i(\omega)\right) \exp\left(-0.5 \frac{(x - v_i)^2}{0.0225}\right),$$

where $n = 2$ and where $Z = [z_1, z_2]^T$ is a random vector with components that are normally distributed with mean $\boldsymbol{\mu} = [1, 1]^T$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

The vector $\boldsymbol{v} = [v_1, v_2]^T \in \mathbb{R}^2$ is $\boldsymbol{v} = [0.5, 0.8]^T$. The quantity of interest is the value of the solution function at the right boundary and given by the output of the function $f : \mathcal{D} \to \mathbb{R}$ defined as

$$f(Z(\omega)) = u(\omega, 1).$$



(a) biasing runtime, $t = 1.14$, reference $\hat{P}^{\mathrm{IS}}_{1.14} \approx 4 \times 10^{-6}$ (b) total runtime, $t = 1.14$, reference $\hat{P}^{\mathrm{IS}}_{1.14} \approx 4 \times 10^{-6}$

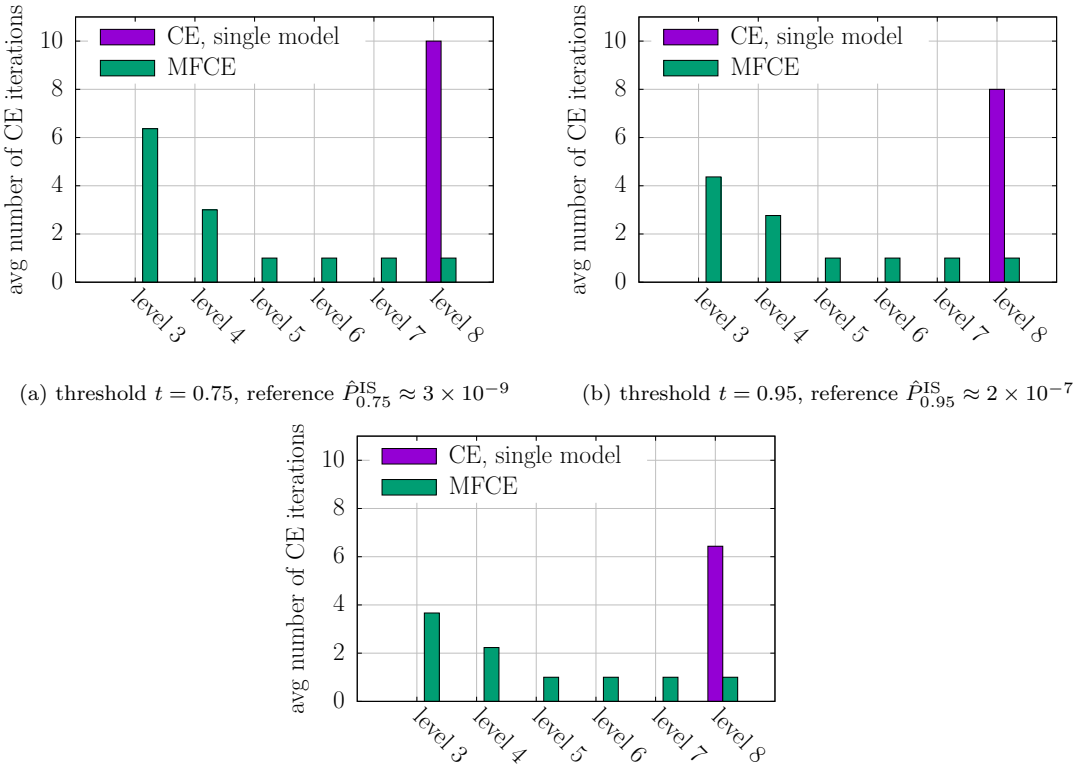
Figure 2: Heat transfer: Our MFCE approach achieves a speedup of about one order of magnitude for rare event probabilities of $\approx 10^{-6}$ with $t = 1.14$ in this example.

We discretize (23)–(25) with linear finite elements on an equidistant grid with mesh width $h^{(\ell)} = 2^{-\ell}$ on level $\ell \in \mathbb{N}$. The solution of the discretized problem on level ℓ leads to models $f^{(\ell)}$. We set the maximal level to $L = 8$.

4.1.2. Rare event probability estimation. Our goal is to estimate the rare event probabilities $P_t^{(L)}$ for $t \in \{0.75, 0.95, 1.14\}$. We derive reference rare event probabilities with the classical, single-fidelity CE method with 10^7 realizations. To obtain these reference probabilities, we run Algorithm 1 with the model $f^{(L)}$ only. We average over 30 runs and obtain the reference probability $\hat{P}^{\mathrm{IS}}_{0.75} \approx 3 \times 10^{-9}$ for $t = 0.75$, the reference probability $\hat{P}^{\mathrm{IS}}_{0.95} \approx 2 \times 10^{-7}$ for $t = 0.95$, and the reference probability $\hat{P}^{\mathrm{IS}}_{1.14} \approx 4 \times 10^{-6}$ for $t = 1.14$. For our MFCE method, we consider the levels $\ell = 3, \ldots, 8$ and run Algorithm 1 with $m \in \{10^3, 10^4, 10^5, 10^6\}$ realizations. We repeat the estimation with MFCE 30 times and estimate the squared coefficient of variation (1) with respect to the reference probabilities.

Figure 1a and Figure 1c compare the runtime of constructing the biasing density with our MFCE method to the runtime of the single-fidelity CE method that uses model $f^{(L)}$ alone. Our MFCE approach achieves a speedup of up to two orders of magnitude. Figure 1b and Figure 1d show similar speedups for the total runtime, which includes the runtime of constructing the biasing density and the final estimation step. In this example, the total runtime is dominated by the runtime of constructing the biasing densities.

Figure 2 shows the speedup of our MFCE method for the threshold $t = 1.14$, which is smaller than the speedup obtained with the thresholds $t = 0.75$ and $t = 0.95$ shown in Figure 1. The threshold $t = 1.14$ corresponds to a reference probability of $\approx 10^{-6}$, which is significantly higher than the reference probabilities $\approx 10^{-7}$ and $\approx 10^{-9}$ corresponding to $t = 0.95$ and $t = 0.75$, respectively. Typically fewer CE iterations are sufficient to construct a biasing density to estimate a rare event probability of $\approx 10^{-6}$ than of $\approx 10^{-9}$. Thus, the results in Figure 2 confirm that our MFCE approach is particularly beneficial in cases where the CE method requires many iterations to obtain a biasing density, which typically is the case for small rare event probabilities. Overall, the results reported in Figure 1 and Figure 2 show



(a) threshold $t = 0.75$, reference $\hat{P}^{\mathrm{IS}}_{0.75} \approx 3 \times 10^{-9}$

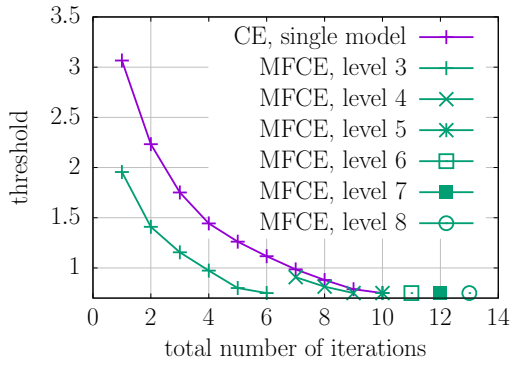
(b) threshold $t = 0.95$, reference $\hat{P}^{\mathrm{IS}}_{0.95} \approx 2 \times 10^{-7}$

(c) threshold $t = 1.14$, reference $\hat{P}^{\mathrm{IS}}_{1.14} \approx 4 \times 10^{-6}$

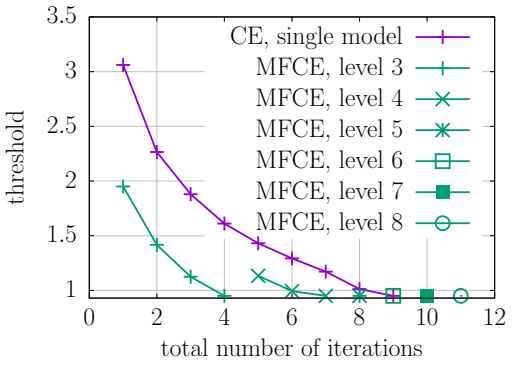
Figure 3: Heat transfer: Our MFCE approach achieves runtime speedups by shifting most of the iterations onto the models with coarse grids.

that our MFCE method successfully leverages the hierarchy of models to estimate rare event probabilities that vary by about three orders of magnitude (i.e., from $\approx 10^{-6}$ to $\approx 10^{-9}$).

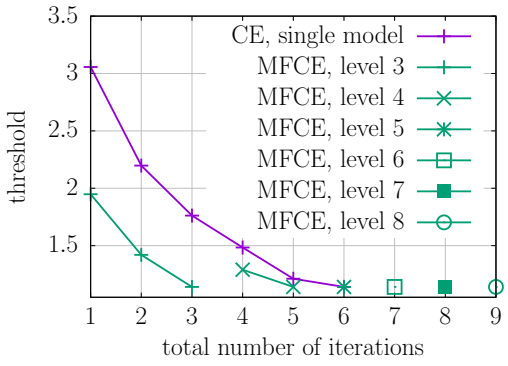
4.1.3. Number of model evaluations. Figure 3 compares the number of iterations spent at each level $\ell = 3, \ldots, 8$ of our MFCE method to the number of iterations of the single-fidelity CE method on level $L = 8$. The reported numbers of iterations are averaged over 30 runs. First, note that our MFCE approach and the single-fidelity CE method require most iterations for $t = 0.75$, which corresponds to the smallest rare event probability $\approx 10^{-9}$ of the three cases $t \in \{0.75, 0.95, 1.14\}$. Second, the results confirm that our multifidelity approach spends most of the iterations with models on the coarse grids, where model evaluations are cheap compared to the model $f^{(L)}$ on the finest grid. Consider now Figure 4, which reports the intermediate thresholds that are selected by Algorithm 1. Figure 4a shows that the single-fidelity CE method requires about 10 iterations on level $L = 8$ to obtain an intermediate threshold that is equal to or below $t = 0.75$. Our MFCE approach requires six iterations on the lowest level $\ell = 3$; however, the parameter $\hat{\boldsymbol{v}}_*^{(3)}$ estimated on level $\ell = 3$ is then further corrected on level $\ell = 4$ in three iterations. On levels $\ell = 5, 6, 7, 8$ only a single iteration is necessary to slightly



(a) threshold $t = 0.75$, reference $\hat{P}^{\text{IS}}_{0.75} \approx 3 \times 10^{-9}$



(b) threshold $t = 0.95$, reference $\hat{P}^{\text{IS}}_{0.95} \approx 2 \times 10^{-7}$



(c) threshold $t = 1.14$, reference $\hat{P}^{\text{IS}}_{1.14} \approx 4 \times 10^{-6}$

Figure 4: Heat transfer: The plot shows the intermediate thresholds selected by Algorithm 1. Our multifidelity approach achieves speedups because only a single iteration is required with the computationally expensive models on the fine grids (high level) to correct estimates obtained with the cheap models on the coarse grids (low levels).

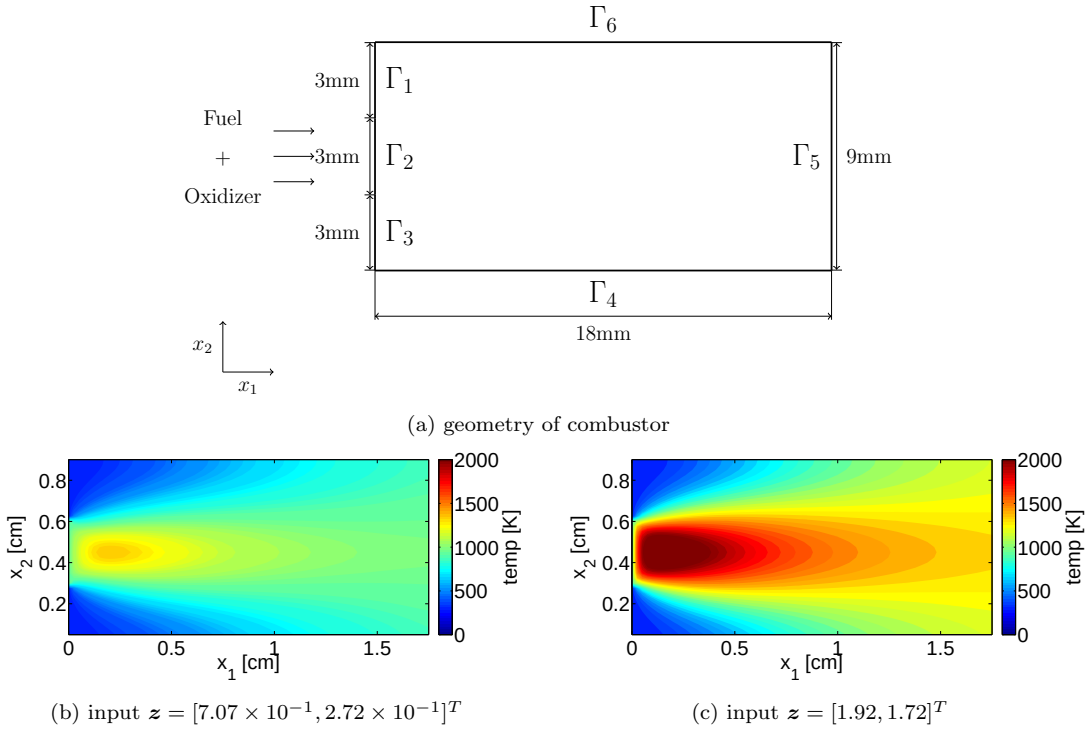
correct the estimated parameter. Similar results are obtained for $t \in \{0.95, 1.14\}$ shown in Figure 4b and Figure 4c.

4.2. Reacting flow problem. This section demonstrates the MFCE approach on a reacting-flow problem.

4.2.1. Problem setup. We consider the simplified combustor model described in [7], which is based on the one-step reaction

$$2\mathrm{H}_2 + \mathrm{O}_2 \rightarrow 2\mathrm{H}_2\mathrm{O}\,,$$

with the fuel H_2, the oxidizer O_2, and the product $\mathrm{H}_2\mathrm{O}$. The governing equations are nonlinear advection-diffusion-reaction equations with an Arrhenius-type reaction term [7]. The geometry of the combustor is shown in Figure 5. Dirichlet boundary conditions are imposed on Γ_1, Γ_2, and Γ_3. On Γ_4, Γ_5, and Γ_6, homogeneous Neumann boundary conditions are imposed. The governing equations are discretized with finite differences on a mesh with equidistant grid



(a) geometry of combustor

(b) input $\boldsymbol{z} = [7.07 \times 10^{-1}, 2.72 \times 10^{-1}]^T$

(c) input $\boldsymbol{z} = [1.92, 1.72]^T$

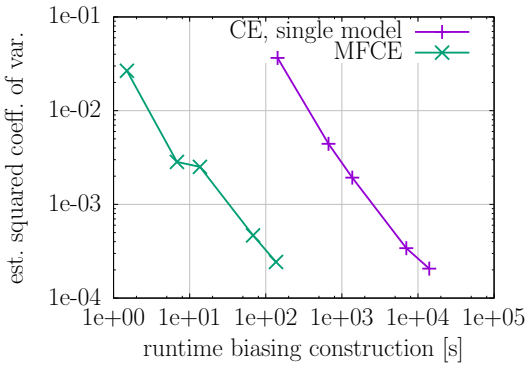
Figure 5: Reacting flow: The geometry of the combustor is shown in (a). The plots in (b) and (c) show the temperature of the reaction for two different inputs.

points. The problem has two inputs $\boldsymbol{z} = [z_1, z_2]^T$ that define properties of the reaction. The input z_1 is the normalized pre-exponential factor of the Arrhenius-type reaction term and z_2 is the normalized activation energy. We refer to [7, 28] for details on the problem. The inputs are realizations of a random variable with normal distribution with mean $\boldsymbol{\mu} = [1, 1]^T$ and covariance

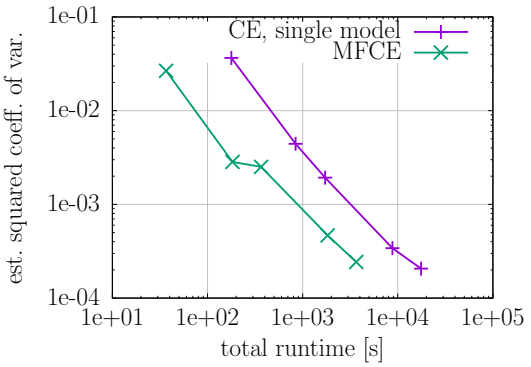
$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.0060 & 0 \\ 0 & 0.0037 \end{bmatrix}.$$

The output of the model is the maximum temperature in the combustion chamber, see Figure 5.

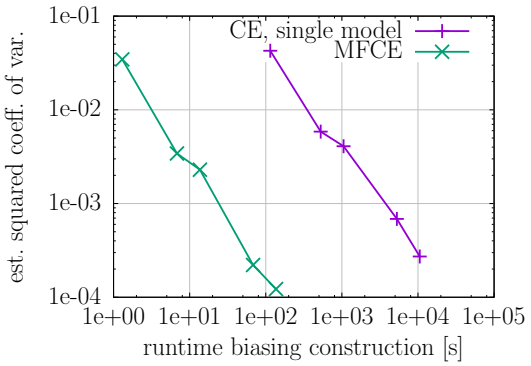
The high-fidelity model in this experiment is given by the finite-difference model on a mesh with 54×27 equidistant grid points. Furthermore, we derive a reduced model with proper orthogonal decomposition and the discrete empirical interpolation method as described in [41]. To construct the reduced model, we derive 100 snapshots with the high-fidelity model that correspond to inputs on an equidistant 10×10 grid in the domain $[0.7, 1.92] \times [0.27, 1.72]$ and derive proper orthogonal decomposition and empirical interpolation bases with 4 and 8 basis vectors, respectively. Additionally, we derive a piecewise-linear interpolant of the input-output map given by the high-fidelity model from four data points drawn from the distribution of the inputs. Thus, we have an interpolant $f^{(1)}$, a reduced model $f^{(2)}$, and a high-fidelity finite-difference model $f^{(3)} = f^{(L)}$.



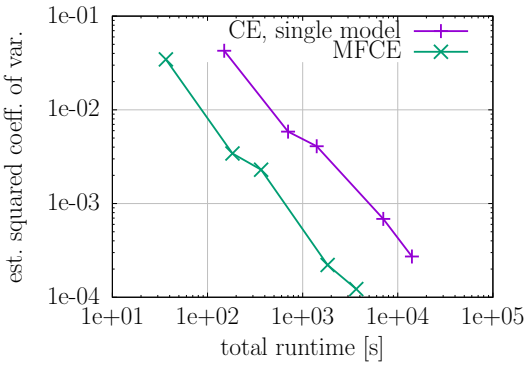
(a) biasing runtime, $t = 2021.3$, $\hat{P}^{\text{IS}}_{2021.3} \approx 2 \times 10^{-6}$



(b) total runtime, $t = 2021.3$, reference $\hat{P}^{\text{IS}}_{2021.3} \approx 2 \times 10^{-6}$



(c) biasing runtime, $t = 2043$, reference $\hat{P}^{\text{IS}}_{2043} \approx 2 \times 10^{-5}$

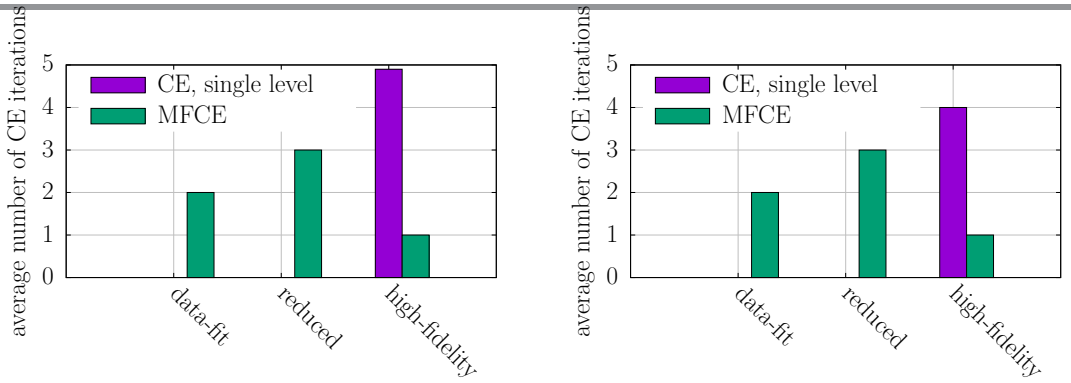


(d) total runtime, $t = 2043$, reference $\hat{P}^{\text{IS}}_{2043} \approx 2 \times 10^{-5}$

Figure 6: Reacting flow: The total runtime is dominated by the estimation step with the high-fidelity model because a speedup of multiple orders of magnitude is achieved for constructing the biasing density and a speedup of at most one order of magnitude is obtained in the total runtime.

Our goal in this experiment is to estimate the probability that the temperature is below a threshold value, which can indicate a poor mixing in the reaction. We estimate the rare event probabilities for the thresholds $t \in \{2021.3, 2043\}$. We first run Algorithm 1 with the high-fidelity model $f^{(L)}$ alone to obtain the reference probabilities $\hat{P}^{\text{IS}}_{2021.3} \approx 2 \times 10^{-6}$ and $\hat{P}^{\text{IS}}_{2043} \approx 2 \times 10^{-5}$, respectively. The reference probabilities are estimated from 10^4 realizations and are averaged over 30 runs. We then run Algorithm 1 with the models $f^{(1)}, f^{(2)}, f^{(3)}$ 30 times with $m \in \{10^2, 5 \times 10^2, 10^3, 5 \times 10^3, 10^4\}$ and estimate the squared coefficient of variation with respect to the reference probabilities.

4.2.2. Comparison of multi- and single-fidelity approaches. Figure 6 compares the runtime of our multifidelity approach with the runtime of the single-fidelity CE method that uses $f^{(L)}$ alone. Our multifidelity approach achieves speedups of more than two orders of magnitude in the construction of the biasing densities. The large speedups are obtained because the data-fit $f^{(1)}$ and the reduced model $f^{(2)}$ are five and two orders of magnitude cheaper to evaluate than the high-fidelity model $f^{(3)}$, respectively. Consider now the total runtime,



(a) threshold $t = 2021.3$, reference $\hat{P}^{\mathrm{IS}}_{2021.3} \approx 2 \times 10^{-6}$ (b) threshold $t = 2043$, reference $\hat{P}^{\mathrm{IS}}_{2043} \approx 2 \times 10^{-5}$

Figure 7: Reacting flow: Our MFCE approach shifts most of the iteration onto the reduced and data-fit interpolation models.

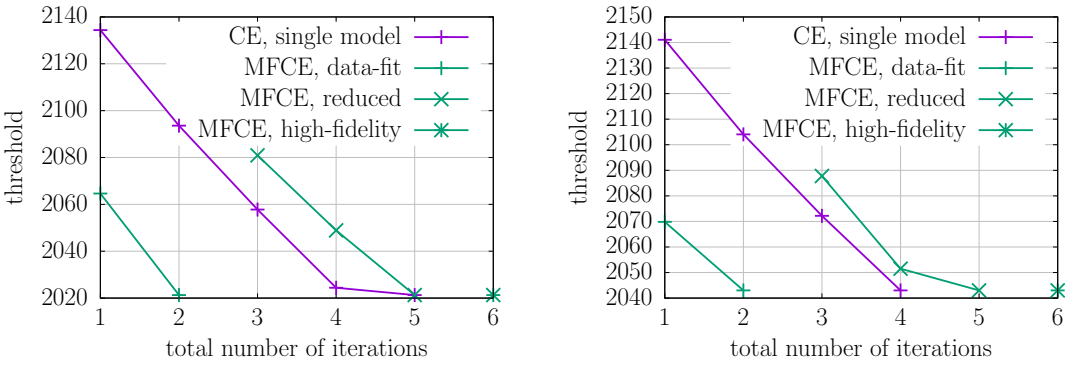
which includes the construction of the biasing densities and the final estimate step. The total runtime in this example is dominated by the costs of the final estimation step, which means that the speedup of our multifidelity approach obtained in constructing the biasing densities is smaller when considered with respect to the total runtime. Our method achieves up to an order of magnitude speedup in the total runtime.

Figure 7 reports the number of iterations per level. In our MFCE method, a single iteration with the high-fidelity model is sufficient, whereas the single-fidelity CE method requires up to almost 5 iterations. The intermediate thresholds selected by Algorithm 1 are shown in Figure 8. The results confirm that model $f^{(1)}$ is a poor approximation of the high-fidelity model because the intermediate threshold selected on level $\ell = 1$ needs to be corrected with three iterations on level $\ell = 2$.

4.3. Reacting flow problem with five inputs. We now extend the reacting flow problem of Section 4.2 with three additional inputs to demonstrate our MFCE approach on a five-dimensional problem with non-Gaussian input random variables.

4.3.1. Problem setup. Consider the combustor model described in Section 4.2. The two inputs z_1 and z_2 describe the properties of the Arrhenius-type reaction term. We now additionally have the input z_3, which is the temperature that is imposed on the boundaries Γ_1 and Γ_3, see Figure 5. The input z_4 is the temperature imposed on Γ_2, and the input z_5 controls the fuel content of the inflow on Γ_2. The inputs are described in detail in [7].

The inputs z_1 and z_2 are realizations of a random variable that is distributed as described in Section 4.2. The input z_3 is the realization of a random variable with a Gamma distribution with shape parameter $\alpha = 18000$ and scaling parameter $\beta = 1/60$. The random variable of input z_4 is also a Gamma distribution with parameters $\alpha = 180500$ and $\beta = 1/190$. The mean of the random variables is 300 and 950, respectively, which are the values used in Section 4.2. The input z_5 is the realization of a random variable with a log-normal distribution with mean -5×10^{-7} and standard deviation 0.001. In Section 4.2, the input z_5 is set to 1.



(a) threshold $t = 2021.3$, reference $\hat{P}^{\text{IS}}_{2021.3} \approx 2 \times 10^{-6}$ (b) threshold $t = 2043$, reference $\hat{P}^{\text{IS}}_{2043} \approx 2 \times 10^{-5}$

Figure 8: Reacting flow: The data-fit model $f^{(1)}$ is a poor approximation of the high-fidelity model and therefore three iterations with the more accurate reduced model $f^{(2)}$ are necessary to correct the intermediate thresholds. Overall, our multifidelity method leverages the data-fit and the reduced model to reduce the number of iterations required on the highest level with the high-fidelity model.

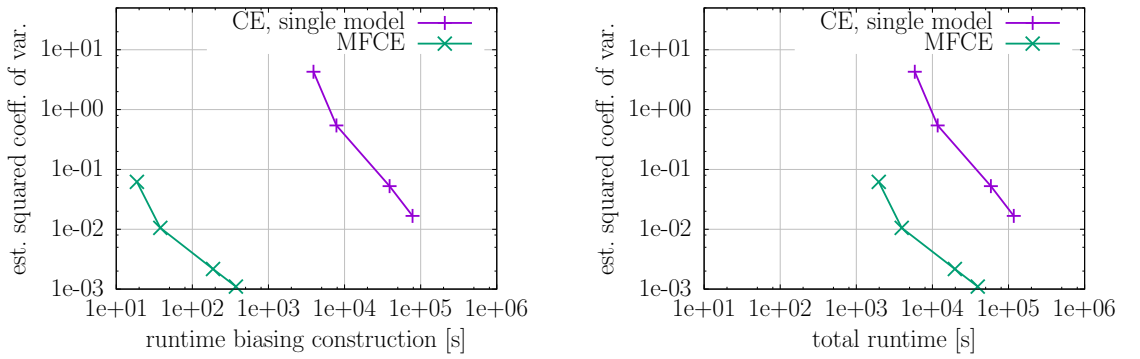
A finite-difference high-fidelity model $f^{(3)}$ is used as in Section 4.2. A reduced model $f^{(2)}$ is derived similar to Section 4.2, except that the snapshots correspond to inputs $\boldsymbol{z} = [z_1, z_2, \ldots, z_5]$ on a $5 \times 5 \times 5 \times 5 \times 5$ grid in the domain

$$[0.7, 1.92] \times [0.27, 1.72] \times [280, 320] \times [920, 980] \times [0.95, 1.05]\,. \tag{26}$$

The grid points are logarithmically spaced in the first two dimensions corresponding to z_1 and z_2 and equidistant on a linear scale in the dimensions corresponding to z_3, z_4, and z_5. Additionally, we construct a data-fit model $f^{(1)}$ that interpolates the input-output map given by the high-fidelity model. The data-fit model is a spline interpolant on a $4 \times 4 \times 4 \times 4 \times 4$ grid in the domain (26). The grid points are logarithmically spaced in the first two dimensions and linearly spaced in dimensions 3, 4, and 5. The spline interpolant is obtained with MATLAB's `griddedInterpolant` method.

4.3.2. Estimation of rare event probability. We estimate the probability that the temperature is below the threshold $t = 2021.3$. Running Algorithm 1 with the high-fidelity model $f^{(3)}$ alone gives a reference probability $\hat{P}^{\text{IS}}_{2021.3} \approx 2 \times 10^{-6}$, which is similar to what we obtained in Section 4.2. For the parameter estimation step in (22) in Algorithm 1, we use the formulas obtained in Section 3.5. The reference probability is estimated from 10^5 realizations and averaged over 30 runs. We run Algorithm 1 with $m \in \{5 \times 10^3, 10^4, 5 \times 10^4, 10^5\}$ and models $f^{(1)}, f^{(2)}$, and $f^{(3)}$. Algorithm 1 is run 30 times and the averages of the coefficients of variations of the single-fidelity CE estimator and our MFCE estimator are reported in Figure 9. Our MFCE approach achieves a speedup of about two orders of magnitude compared to the single-fidelity approach that uses the high-fidelity model alone.

5. Conclusions. We presented a multifidelity preconditioner for the CE method to accelerate the estimation of rare event probabilities. Our multifidelity approach leverages a



(a) biasing runtime, $t = 2021.3$, $\hat{P}^{\mathrm{IS}}_{2021.3} \approx 2 \times 10^{-6}$ (b) total runtime, $t = 2021.3$, reference $\hat{P}^{\mathrm{IS}}_{2021.3} \approx 2 \times 10^{-6}$

Figure 9: Reacting flow with five inputs: Our MFCE approach achieves up to two orders of magnitude speedup in this example with five inputs. The first two inputs are normally distributed, the third and fourth input follow a Gamma distribution, and the fifth input is log-normally distributed.

hierarchy of surrogate models to reduce the costs of constructing biasing densities compared to the single-fidelity CE method that uses the high-fidelity model alone. Our approach can exploit multiple surrogate models that include general surrogate models such as projection-based reduced models and data-fit models, which goes beyond the classical setting of multilevel techniques that are often restricted to hierarchies of coarse-grid approximations. In our numerical examples, our approach achieved speedups of up to two orders of magnitude compared to the single-fidelity CE method in estimating probabilities on the order of 10^{-5} to 10^{-9}.

Acknowledgements. The authors would like to thank Elizabeth Qian for providing the reacting-flow problem with five inputs. The authors are very grateful to Elisabeth Ullmann for reporting an inconsistency in an earlier draft of this work. Several examples were computed on the computer clusters of the Munich Centre of Advanced Computing.

REFERENCES

[1] S.-K. Au and J. L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263 – 277, 2001.

[2] P. Benner, S. Gugercin, and K. Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review*, 57(4):483–531, 2015.

[3] Z. Botev, D. Kroese, and T. Taimre. Generalized cross-entropy methods with applications to rare-event simulation and optimization. *SIMULATION*, 83(11):785–806, 2007.

[4] Z. I. Botev and D. P. Kroese. The generalized cross entropy method, with applications to probability density estimation. *Methodology and Computing in Applied Probability*, 13(1):1–27, Mar 2011.

[5] J.-M. Bourinet, F. Deheeger, and M. Lemaire. Assessing small failure probabilities by combined subset simulation and support vector machines. *Structural Safety*, 33(6):343 – 353, 2011.

[6] J. Bucklew. *Introduction to Rare Event Simulation*. Springer, 2003.

[7] M. Buffoni and K. Willcox. Projection-based model reduction for reacting flows. In *40th Fluid Dynamics Conference and Exhibit*, Fluid Dynamics and Co-located Conferences, pages 1–14. American Institute of Aeronautics and Astronautics, 2010.

[8] P. Chen and A. Quarteroni. Accurate and efficient evaluation of failure probability for partial differ-

ent equations with random input data. *Computer Methods in Applied Mechanics and Engineering*, 267:233–260, 2013.

[9] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[10] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

[11] T. H. de Mello and R. Y. Rubinstein. Estimation of rare event probabilities using cross-entropy. In *Winter Simulation Conference*, volume 01, pages 310–319, 2002.

[12] V. Dubourg, B. Sudret, and F. Deheeger. Metamodel-based importance sampling for structural reliability analysis. *Probabilistic Engineering Mechanics*, 33:47 – 57, 2013.

[13] D. Elfverson, D. J. Estep, F. Hellman, and A. Målqvist. Uncertainty quantification for approximate p-quantiles for physical models with stochastic inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):826–850, 2014.

[14] D. Elfverson, F. Hellman, and A. Målqvist. A multilevel Monte Carlo method for computing failure probabilities. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):312–330, 2016.

[15] F. Fagerlund, F. Hellman, A. Målqvist, and A. Niemi. Multilevel Monte Carlo methods for computing failure probability of porous media flow systems. *Advances in Water Resources*, 94:498 – 509, 2016.

[16] G. Fishman. *Monte Carlo*. Springer, 1996.

[17] A. Forrester, A. Sóbester, and A. Keane. *Engineering design via surrogate modelling: a practical guide*. Wiley, 2008.

[18] M. B. Giles. Multilevel monte carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

[19] S. Heinrich. Multilevel Monte Carlo Methods. In S. Margenov, J. Waśniewski, and P. Yalamov, editors, *Large-Scale Scientific Computing*, number 2179 in Lecture Notes in Computer Science, pages 58–67. Springer, 2001.

[20] J. Li, J. Li, and D. Xiu. An efficient surrogate-based method for computing rare failure probability. *Journal of Computational Physics*, 230(24):8683–8697, 2011.

[21] J. Li and D. Xiu. Evaluation of failure probability via surrogate models. *Journal of Computational Physics*, 229(23):8966–8980, 2010.

[22] J. Li and D. Xiu. Surrogate based method for evaluation of failure probability under multiple constraints. *SIAM Journal on Scientific Computing*, 36(2):A828–A845, 2014.

[23] A. J. Majda and B. Gershgorin. Quantifying uncertainty in climate change science through empirical information theory. *Proceedings of the National Academy of Sciences of the United States of America*, 107(34):14958–14963, Aug. 2010.

[24] L. Ng and K. Willcox. Multifidelity approaches for optimization under uncertainty. *International Journal for Numerical Methods in Engineering*, 100(10):746–772, 2014.

[25] V. Papadopoulos, D. G. Giovanis, N. D. Lagaros, and M. Papadrakakis. Accelerated subset simulation with neural networks for reliability analysis. *Computer Methods in Applied Mechanics and Engineering*, 223–224:70 – 80, 2012.

[26] B. Peherstorfer, T. Cui, Y. Marzouk, and K. Willcox. Multifidelity importance sampling. *Computer Methods in Applied Mechanics and Engineering*, 300:490 – 509, 2016.

[27] B. Peherstorfer, B. Kramer, and K. Willcox. Combining multiple surrogate models to accelerate failure probability estimation with expensive high-fidelity models. *Journal of Computational Physics*, 341:61–75, 2017.

[28] B. Peherstorfer and K. Willcox. Online adaptive model reduction for nonlinear systems via low-rank updates. *SIAM Journal on Scientific Computing*, 37(4):A2123–A2150, 2015.

[29] B. Peherstorfer, K. Willcox, and M. Gunzburger. Optimal model management for multifidelity Monte Carlo estimation. *SIAM Journal on Scientific Computing*, 38(5):A3163–A3194, 2016.

[30] B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 2017. (to appear).

[31] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[32] G. Rozza, D. Huynh, and A. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Archives of Computational Methods in Engineering*, 15(3):1–47, 2007.

[33] R. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology And Computing In Applied Probability*, 1(2):127–190, 1999.

[34] R. Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89 – 112, 1997.

[35] R. Y. Rubinstein. Combinatorial optimization, cross-entropy, ants and rare events. In S. Uryasev and P. M. Pardalos, editors, *Stochastic Optimization: Algorithms and Applications*, pages 303–363, Boston, MA, 2001. Springer US.

[36] R. Y. Rubinstein. A stochastic minimum cross-entropy method for combinatorial optimization and rare-event estimation*. *Methodology and Computing in Applied Probability*, 7(1):5–50, Mar 2005.

[37] R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer, 2004.

[38] R. Srinivasan. *Importance Sampling*. Springer, 2002.

[39] I. Szita and A. Lőrincz. Learning tetris using the noisy cross-entropy method. *Neural Computation*, 18:2936–2941, 2006.

[40] E. Ullmann and I. Papaioannou. Multilevel estimation of rare events. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):922–953, 2015.

[41] Y. B. Zhou. *Model reduction for nonlinear dynamical systems with parametric uncertainties*. Thesis (S.M.), Massachusetts Institute of Technology, Dept. of Aeronautics and Astronautics, 2012.

[42] K. M. Zuev, J. L. Beck, S.-K. Au, and L. S. Katafygiotis. Bayesian post-processor and other enhancements of subset simulation for estimating failure probabilities in high dimensions. *Computers & Structures*, 92–93:283 – 296, 2012.