SAND2020-3026J

sMF-BO-2CoGP: A sequential multi-fidelity constrained Bayesian optimization framework for design applications

Anh Tran[*] **, Tim Wildey**
Optimization and Uncertainty Quantification
Sandia National Laboratories, Albuquerque, NM 87123
Email: {anhtran, tmwilde}@sandia.gov

Scott McCann
Xilinx Inc., San Jose, CA 95124
Email: smccann@xilinx.com

Bayesian optimization (BO) is an effective surrogate-based method that has been widely used to optimize simulation-based applications. While the traditional Bayesian optimization approach only applies to single-fidelity models, many realistic applications provide multiple levels of fidelity with various computational complexity and predictive capability. In this work, we propose a multi-fidelity Bayesian optimization method for design applications with both known and unknown constraints. The proposed framework, called sMF-BO-2CoGP, is built on a multi-level CoKriging method to predict the objective function. An external binary classifier, which we approximate using a separate CoKriging model, is used to distinguish between feasible and infeasible regions. The sMF-BO-2CoGP method is demonstrated using a series of analytical examples and a flip-chip application for design optimization to minimize the deformation due to warping under thermal loading conditions.

1 Introduction

High-fidelity engineering models are frequently utilized to predict quantities of interests, such as properties or performances, with respect to a specific design. These predictions are then fed back into the design process to find a better design that outperform the previous ones by changing the design parameters. This process, often called design optimization, is ubiquitous in industrial settings. Simulation-based optimization is challenging due to the tremendous computational cost associated with high-fidelity models. However, for many practical applications, multiple models of various fidelity can be developed and a multi-fidelity optimization framework, such as Bayesian optimization (BO), can be then applied to optimize the objective function at the highest level of fidelity, but at a reduced computational cost by leveraging lower-fidelity data.

Multi-fidelity methods provide an effective framework to reduce the computational cost in optimization and uncertainty quantification by leveraging the correlations with lower-fidelity models to reduce the computational burden on high-fidelity models. Multi-fidelity frameworks are particularly practical for engineering simulation-based applications, such as computational fluid dynamics and solid mechanics problems, because most of these involve discretizations (spatial and/or temporal), where a finer discretization typically corresponds to a higher level of fidelity and the coarser discretization corresponds to a lower level of fidelity.

Incorporating physical and practical constraints into the optimization formulation is also a critical task. Digabel and Wild [1] proposed the QRAK taxonomy to classify constrained optimization problems. In engineering settings, constraints arise from multiple sources and many problems require both known and unknown constraints to be incorporated into the formulation. Constraints are *known* if the feasibility of the input can be determined directly from the input sampling location, i.e., without actually evaluating the model. Such known constraints are often formulated as a set of inequalities. On the other hand, constraints are *unknown* if the feasibility of the input must be evaluated indirectly by evaluating the model. Common examples of unknown constraints are ill-conditioning induced by the parameters, singularity in the design, and mesh generation problems. These constraints are implicitly imposed and feasibility cannot be determined without evaluating the computational model.

Gaussian process (GP) methods provide an efficient framework to model a response surface that approximates the objective function for single-fidelity formulations. In the traditional BO approach, an acquisition function $a(\boldsymbol{x})$ is con-

[*] Corresponding author: anhtran@sandia.gov

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

structed based on a utility function, which rewards the BO method if the new sampling location outperforms the rest. The acquisition function is typically constructed based on the posterior mean and posterior variance of the GP. Because of its flexibility, many extensions based on the traditional BO framework have been proposed to solve other optimization problems, including both constrained and multi-fidelity problems. Incorporating constraints is a well-studied subject in the context of BO methods, and typically involves adopting a penalty scheme to penalize the infeasible sampling locations that do not satisfy all of the constraints. Multi-fidelity BO problems are more complicated to deal with. To generalize to multiple levels of fidelity, one needs to consider the correlation between levels of fidelity from the objective function and fuse the predictions across all levels of fidelity. For example, Kennedy and O'Hagan [2] proposed an autoregressive approach to form a link between lower-fidelity to the next higher-fidelity by a linear regression between two levels of fidelity. The terms CoGP and CoKriging are used interchangeably in this work to describe the recursive autoregressive GP model. Because the constrained problems have been relatively well studied, we will focus the literature review in Section 2 on multi-fidelity GP.

In this work, we develop a sequential constrained multi-fidelity method sMF-BO-2CoGP, as an extension of sBF-BO-2CoGP [3], using a CoKriging approach to approximate the objective function at the highest level of fidelity. The known constraints are implemented by penalizing the acquisition function directly for infeasible input sampling locations. The unknown constraints are adaptively learned using another CoKriging model, which acts as a probabilistic binary classifier. The unknown constrained acquisition function is conditioned on this predicted probability mass function, in addition to the penalty scheme for known constraints. The optimal location for the next sample is determined by maximizing the constrained acquisition function. Next, an uncertainty reduction scheme, where uncertainty is measured by the integrated mean-square error, is proposed to determine the appropriate level of fidelity to evaluate. Compared to the maximum mean square error criteria, the integrated mean square error is demonstrated to be more robust and efficient.

The content of this paper is invited following the conference paper presented at the ASME IDTEC CIE 2019 (August 18–21, 2019) at Anaheim, CA [3]. The main difference is that this paper generalizes our previous work [3] from bi-fidelity to multi-fidelity problems. The remainder of this paper is organized as follows. Section 2 provides a brief introduction to the BO method. Section 3 describes the multi-fidelity sMF-BO-2CoGP method proposed in this paper, including the constrained acquisition function, the fidelity selection criteria. Section 4 demonstrates the application of the proposed sMF-BO-2CoGP methodology using several analytical examples and an engineering application in designing flip-chip package based on finite element model. Section 5 discusses and Section 6 concludes the paper.

2 Related works

Let f denote a function of $\boldsymbol{x}$, where $\boldsymbol{x} \in \mathcal{X}$ is a d-dimensional input, and y is the observation. The optimization considered in this paper is formulated as

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\operatorname{argmax}} f(\boldsymbol{x}), \quad (1)$$

subjected to a set of inequality constraints

$$g_j(\boldsymbol{x}) \leq 0, \quad j = 1, \ldots, J, \quad (2)$$

where J is the number of inequality constraints.

We briefly review the classical BO method, CoKriging method, the most common acquisition functions in BO, in Sections 2.1, 2.2, and 2.3, respectively. Readers are referred to other comprehensive reviews and tutorials [4, 5, 6, 7] for rigorous literature reviews on GP and BO methods and its variants.

2.1 Gaussian process

In this section, we followed the notation of Shahriari et al. [5] in the GP formulation. Let $\mathcal{D} = (\boldsymbol{x}_i, y_i)_{i=1}^n$ denote the dataset, where n is the number of observations and $\boldsymbol{x} \in \mathcal{X}$ is the d-dimensional input. A GP regression approach assumes that $\boldsymbol{f} = f_{1:n}$ is jointly Gaussian, and the observation y is normally distributed given f,

$$\boldsymbol{f}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}), \quad (3)$$

$$\boldsymbol{y}|\boldsymbol{f}, \sigma^2 \sim \mathcal{N}(\boldsymbol{f}, \sigma^2 \boldsymbol{I}), \quad (4)$$

where $m_i := \mu(\boldsymbol{x}_i)$ and $K_{i,j} := k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

The choice of the kernel $\boldsymbol{K}$ depends on the covariance between inputs. At an unknown sampling location $\boldsymbol{x}$, the predicted response is described by a posterior Gaussian distribution, where the posterior mean is

$$\mu_n(\boldsymbol{x}) = \mu_0(\boldsymbol{x}) + \boldsymbol{k}(\boldsymbol{x})^T (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{m}), \quad (5)$$

and the posterior variance is

$$\sigma_n^2 = k(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{k}(\boldsymbol{x}), \quad (6)$$

where $\mu_0(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X} \mapsto \mathbb{R}$ is the prior mean function and $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is the covariance function between the query point $\boldsymbol{x}$ and $\boldsymbol{x}_{1:n}$. The classical GP formulation assumes a stationary covariance matrix, which only depends on the distance $r = \|\boldsymbol{x} - \boldsymbol{x}'\|$ where $\|\cdot\|$ is usually the classical L^2-norm, but other choices, e.g., the L^1-norm and weighted variants, have also been explored. Common kernels for GP include [5]

$$\begin{aligned}
&k_{\text{Matérn1}}(\boldsymbol{x},\boldsymbol{x}') = \theta_0^2 \exp\left(-r\right),\\
&k_{\text{Matérn3}}(\boldsymbol{x},\boldsymbol{x}') = \theta_0^2 \exp\left(-\sqrt{3}r\right)(1+\sqrt{3}r),\\
&k_{\text{Matérn5}}(\boldsymbol{x},\boldsymbol{x}') = \theta_0^2 \exp\left(-\sqrt{5}r\right)\left(1+\sqrt{5}r+\frac{5}{3}r^2\right),\\
&k_{\text{sq-exp}}(\boldsymbol{x},\boldsymbol{x}') = \theta_0^2 \exp\left(-\frac{1}{2}r^2\right).
\end{aligned}$$

The log-likelihood function can be written as

$$\begin{aligned}
\log p(\boldsymbol{y}|\boldsymbol{x}_{1:n},\theta) = &-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{K}^{\theta}+\sigma^2\boldsymbol{I}| \\
&-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{m}_{\theta})^T(\boldsymbol{K}^{\theta}+\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\boldsymbol{m}_{\theta}).
\end{aligned} \tag{7}$$

Optimizing the log-likelihood function yields the optimal hyper-parameter θ at the computational cost of $O(n^3)$ due to the inversion of the covariance matrix.

2.2 Multi-fidelity CoKriging

One of the advantages of CoKriging is that it can exploit the correlation between low- and high-fidelity and improve the prediction at high-fidelity level by adding more low-fidelity training data points. If the computational costs of evaluation between the high- and low-fidelity differ significantly, this advantage offers a reduction in the number of training data points, thus increase the efficiency of the optimization problem. Kennedy and O'Hagan [2] proposed an auto-regressive model that couples all levels of fidelity together. Le Gratiet and Garnier [8] proposed a nested scheme $\mathcal{D}_1 \subset \mathcal{D}_2 \subseteq \cdots \subseteq \mathcal{D}_s$ to decouple s levels of fidelity into s standard levels of GP regression, where the GP is used to model the discrepancy between two consecutive levels of fidelity. Karniadakis et al. [9, 10, 11] employed the same method to approximate the highest level of fidelity and extended for noisy evaluations using the same method. Perdikaris et al. [12] proposed a generalized framework that can model nonlinear and space-dependent cross-correlations between models of variable fidelity. The multi-fidelity Bayesian optimization approach is sometimes known as multi-information source optimization [13] or multi-task Bayesian optimization [14]; they are all closely related with each other. For example, Ghoreishi and Allaire have proposed several approaches to solve the multi-information source optimization problem in the context of constraints [15], knowledge-gradient acquisition function [16], Monte Carlo-based approach [17], and applications to computational micromechanics [18]. In this paper, we follow the formulation of Xiao et al. [19] in developing a multi-fidelity CoKriging framework due to its simplicity and the relaxation of the nested requirement, compared to Le Gratiet and Garnier [8] and Perdikaris et al. [12] .

Assuming that the prediction at highest level of fidelity, i.e., level s, can be written as an auto-regressive model [19],

$$f_s(\boldsymbol{x}) = \sum_{t=1}^{s-1} \rho_t f_t(\boldsymbol{x}) + \delta(\boldsymbol{x}), \tag{8}$$

where s is the high-fidelity level, the remaining $(s-1)$ levels correspond to the low-fidelity levels, and ρ_t's are the scaling factors. Two important assumptions are typically made. First, $\delta(\boldsymbol{x})$ is assumed to be independent of $f_t(\boldsymbol{x})$, i.e.,

$$\text{Cov}[f_t(\boldsymbol{x}),\ \delta(\boldsymbol{x})] = 0, \quad t = 1,\ldots,s-1. \tag{9}$$

Second, we assume that $(s-1)$ low-fidelity levels are uncorrelated, i.e.,

$$\text{Cov}[f_i(\boldsymbol{x}),\ f_j(\boldsymbol{x})] = 0, \quad 1 \le i \ne j \le s-1. \tag{10}$$

Then, the covariance matrix for s levels of fidelity is given by

$$\boldsymbol{K} = \begin{pmatrix}
\sigma_1^2\boldsymbol{K}_1(\boldsymbol{X}_1,\boldsymbol{X}_1) & 0 & \cdots & \rho_1\sigma_1^2\boldsymbol{K}_1(\boldsymbol{X}_1,\boldsymbol{X}_e) \\
0 & \sigma_2^2\boldsymbol{K}_2(\boldsymbol{X}_2,\boldsymbol{X}_2) & \cdots & \rho_2\sigma_2^2\boldsymbol{K}_2(\boldsymbol{X}_2,\boldsymbol{X}_e) \\
\vdots & \vdots & \ddots & \vdots \\
\rho_1\sigma_1^2\boldsymbol{K}_1(\boldsymbol{X}_e,\boldsymbol{X}_1) & \rho_2\sigma_2^2\boldsymbol{K}_2(\boldsymbol{X}_e,\boldsymbol{X}_2) & & \sum_{t=1}^{s}\rho_t^2\sigma_t^2\boldsymbol{K}_t(\boldsymbol{X}_e,\boldsymbol{X}_e)+\sigma_d^2\boldsymbol{K}_e(\boldsymbol{X}_e,\boldsymbol{X}_e)
\end{pmatrix}, \tag{11}$$

where σ_t is the intrinsic variance of noisy observations at the t-th level of fidelity. The hyper-parameters $\{\theta_t\}_{t=1}^{s}$ are obtained by optimizing the maximum likelihood function,

$$\log p(\boldsymbol{y}_t|\boldsymbol{x}_{1:n_t},\theta_t) = -\frac{n_t}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{K}_t^{\theta_t}+\sigma^2\boldsymbol{I}|, \tag{12}$$

where as the hyper-parameters θ_δ that corresponds to the discrepancy are obtained by maximizing the likelihood function

$$\begin{aligned}
\log p(\boldsymbol{y}_\delta|\boldsymbol{x}_{1:n_s},\theta_\delta) = &-\frac{n_s}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{K}_\delta^{\theta_\delta}(\boldsymbol{X}_s,\boldsymbol{X}_s)+\sigma^2\boldsymbol{I}| \\
&-\frac{1}{2}(\boldsymbol{\delta}-\boldsymbol{m}_{\theta_\delta})^T(\boldsymbol{K}^{\theta_\delta}+\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{\delta}-\boldsymbol{m}_{\theta_\delta}).
\end{aligned} \tag{13}$$

The coefficients, $\{\rho_t\}_{t=1}^{s-1}$, are obtained by maximizing

$$-\frac{n_s}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{K}_\delta^{\theta_\delta}(\boldsymbol{X}_s,\boldsymbol{X}_s)+\sigma^2\boldsymbol{I}| \tag{14}$$

If the number of fidelity levels are two ($s = 2$), then the conventional bi-fidelity CoKriging framework is conveniently recovered from the multi-fidelity CoKriging. The dataset $\mathcal{D}$ is divided into $\mathcal{D}_c$ and $\mathcal{D}_e$, corresponding to cheap and expensive datasets, respectively. The bi-fidelity formulation is closely related with Couckuyt et al [20, 21, 22] and Forrester et al [23]. Following the autoregressive scheme described above, the first GP models the low-fidelity response $\{\boldsymbol{x}_c, y_c\}$, whereas the second GP models the discrepancy between the high- and low-fidelity model $\delta(\boldsymbol{x})$.

The correlation vector $k(\boldsymbol{x})$ and the covariance matrix $K(\boldsymbol{x})$ are then updated [21, 19] as

$$k(\boldsymbol{x}) = (\rho\cdot\sigma_c^2\cdot k_c(\boldsymbol{x}) \quad \rho\cdot\sigma^2\cdot k_c(\boldsymbol{x},\boldsymbol{X})), \tag{15}$$

$$\boldsymbol{K} = \begin{pmatrix} \sigma_c^2 \cdot \boldsymbol{K}_c & \rho \cdot \sigma_c^2 \cdot \boldsymbol{K}_c(\boldsymbol{X}_c, \boldsymbol{X}_e) \\ \rho \cdot \sigma_c^2 \cdot \boldsymbol{K}_c(\boldsymbol{X}_e, \boldsymbol{X}_c) & \rho^2 \cdot \sigma_c^2 \cdot \boldsymbol{K}_c(\boldsymbol{X}_e, \boldsymbol{X}_e) + \sigma_d^2 \cdot \boldsymbol{K}_e(\boldsymbol{X}_e, \boldsymbol{X}_e) \end{pmatrix}, \tag{16}$$

The hyper-parameters for the low-fidelity level, θ_c, are obtained by maximizing the likelihood function at the lower fidelity level,

$$\log p(\boldsymbol{y}_c | \boldsymbol{x}_{n_c}, \theta_c) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{K}_{\boldsymbol{c}}^{\theta_c} + \sigma_c^2 \boldsymbol{I}|. \tag{17}$$

The hyper-parameters for the high-fidelity level, θ_e, are obtained along with ρ, again by maximizing the likelihood function,

$$\log p(\boldsymbol{y}_e | \boldsymbol{x}_{n_e}, \theta_e) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{K}_{\boldsymbol{e}}^{\theta} + \sigma_e^2 \boldsymbol{I}| \\ - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{m}_{\theta_e})^T (\boldsymbol{K}_{\boldsymbol{e}}^{\theta_e} + \sigma^2 \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{m}_{\theta_e}). \tag{18}$$

The predicted distribution of CoKriging is also characterized by a Gaussian distribution, where the posterior mean and posterior variance are given by (5) and (6), respectively.

2.3 Acquisition function

In the traditional BO method, the next sampling location is determined by maximizing an acquisition function, i.e.,

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathcal{X}}{\operatorname{argmax}}\ a(\boldsymbol{x}), \tag{19}$$

where $a(\boldsymbol{x})$ denotes the acquisition function and $\boldsymbol{x}^*$ is the next sampling location. The acquisition function is deeply connected to the utility function, which corresponds to the rewarding scheme for BO methods, if the next sampling point outperforms the other sampling locations.

There are three acquisition functions that are widely used: the probability of improvement (PI), the expected improvement (EI), and the upper-confident bounds (UCB), but other forms also exist, for example, entropy-based approaches, GP-PES [24, 25, 26], GP-ES [27], GP-EST [28], GP-EPS [29].

The PI acquisition function [30] is defined as

$$a_{\text{PI}}(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta) = \Phi(\gamma(\boldsymbol{x})), \tag{20}$$

where

$$\gamma(\boldsymbol{x}) = \frac{\mu(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta) - f(\boldsymbol{x}_{\text{best}})}{\sigma(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta)}, \tag{21}$$

indicates the deviation away from the best sample. The PI acquisition function is constructed based on the idea of binary utility function, where a unit reward is received if a new best-so-far sample is found and zero otherwise.

The EI acquisition function [31, 32, 33, 34] is defined as

$$a_{\text{EI}}(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta) = \\ \sigma(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta) \cdot (\gamma(\boldsymbol{x}) \Phi(\gamma(\boldsymbol{x})) + \phi(\gamma(\boldsymbol{x})). \tag{22}$$

The EI acquisition is constructed based on an improvement utility function, where the reward is the relative difference if a new best-so-far sample is found and zero otherwise.

The UCB acquisition function [35, 36, 37] is defined as

$$a_{\text{UCB}}(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta) = \\ \mu(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta) + \kappa \sigma(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^n, \theta), \tag{23}$$

where κ is a hyper-parameter describing the acquisition exploitation-exploration balance. We adopt the κ computation from Daniel et al. [37], which is based on Srinivas et al. [35, 36], instead of fixing κ as a constant.

3 Methodology

In this section, we describe the sMF-BO-2CoGP method solving the multi-fidelity optimization problem in Section 2.

3.1 Constraints

We adopt the method from our previous work [38,39,40] to handle the known and unknown constraints. For known constraints, where the sampling location is known to be infeasible without running any functional evaluation, the acquisition function is penalized by setting it to zero. The penalization scheme is equivalent with multiplying the acquisition function $a(\boldsymbol{x})$ with another indicator function $I(\boldsymbol{x})$, where

$$I(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \forall j (1 \leq j \leq J) : g_j(\boldsymbol{x}) \leq 0, \\ 0, & \text{if } \exists j (1 \leq j \leq J) : g_j(\boldsymbol{x}) > 0. \end{cases} \tag{24}$$

The indicator function can be easily implemented by iterating through the known constraints.

To handle the unknown-constrained problem, an external binary probabilistic classifier is employed to predict the probability of feasibility. Practically speaking, the approach employed to approximate the binary classifier for feasibility is up to users. Some examples are k-NN [41], AdaBoost [42], RandomForest [43], support vector machine [44] (SVM), least squares support vector machine (LSSVM) [45], GP [46], and convolutional neural network [47]. One notable choice for the binary classifier is the GP classifier, which performs relatively well compared to other binary classifiers. In sMF-BO-2CoGP, another CoGP is adopted as a binary classifier to predict the probability of feasibility of the sampling location considered.

At an unknown sampling location $\boldsymbol{x}$, the coupled binary classifier predicts a probability of feasibility based on the trained dataset, where the probability of being feasible is $Pr(\text{clf}(\boldsymbol{x}) = 1)$, whereas the probability of being infeasible is $Pr(\text{clf}(\boldsymbol{x}) = 0) = 1 - Pr(\text{clf}(\boldsymbol{x}) = 1)$. Again, we condition the sampling point on this predicted probability mass function by assigning zero value to the probability of being infeasible. Taking the expectation of the acquisition function conditioned on this probability mass function results in a new acquisition function, which can be rewritten in a product form as

$$a^*(\boldsymbol{x}) = a(\boldsymbol{x}) \cdot I(\boldsymbol{x}) \cdot Pr(\text{clf}(\boldsymbol{x}) = 1). \tag{25}$$

Maximizing the new acquisition function $a^*(\boldsymbol{x})$ yields the next sampling location of sMF-BO-2CoGP:

$$\begin{aligned} \boldsymbol{x}^* &= \underset{\boldsymbol{x} \in \mathcal{X}}{\text{argmax}}\ a^*(\boldsymbol{x}) \\ &= \underset{\boldsymbol{x} \in \mathcal{X}}{\text{argmax}}\ a(\boldsymbol{x}) \cdot I(\boldsymbol{x}) \cdot Pr(\text{clf}(\boldsymbol{x}) = 1). \end{aligned} \tag{26}$$

In practice, we adopt the covariance matrix adaptation evolution strategy (CMA-ES) from Hansen et al. [48, 49] to maximize the new acquisition function $a^*(\boldsymbol{x})$.

3.2 Fidelity selection criteria

In this section, we propose the fidelity selection criteria for the multi-fidelity frameworks. The computational cost, as well as the reduction of uncertainty are used as the two factors to determine the fidelity level at which the evaluation will be performed.

To determine the level of fidelity in evaluating the new sampling location, a fidelity selection criteria balancing the computational cost and integrated mean squared error (IMSE) reduction is utilized based on a one-step hallucination. The process of hallucination is adopted from our previous work [38]. For the sake of completeness, the process is summarized here.

The CoKriging is hallucinated at a point $\boldsymbol{x}^*$ if the observation, i.e., training data, is assumed to be exactly the same with the GP posterior mean prediction *temporarily*. The CoKriging posterior distribution is then updated accordingly based on the assumption. The posterior variance $\sigma^2(\boldsymbol{x}^*)$ at $\boldsymbol{x}^*$ is σ^2 for the posterior prediction (in particular, $\sigma^2(\boldsymbol{x}^*)$ at $\boldsymbol{x}^*$ is 0 for deterministic functional evaluation). Then, if the sampling point $\boldsymbol{x}^*$ is feasible, the GP is updated with the true observation, instead of the posterior GP prediction. If the sampling point $\boldsymbol{x}^*$ is infeasible with respect to unknown constraints, then the hallucination process will take place for every iteration at $\boldsymbol{x}^*$.

If $t = 1, \dots, s$ are the s levels of fidelity, then the optimal fidelity level t^* to perform the functional evaluation is determined by

$$\begin{aligned} t^* &= \underset{t}{\text{argmin}} \left(\text{IMSE}_{\text{t, hallucinated}} \cdot C_t\right) \\ &= \underset{t}{\text{argmin}} \left(\int_{\mathcal{X}} \sigma^2(\boldsymbol{x}) d\boldsymbol{x} \cdot C_t \right), \end{aligned} \tag{27}$$

where IMSE is the integrated mean-square error, and the computational cost at level t is denoted as C_t. The term $\left(\text{IMSE}_{\text{t, hallucinated}} \cdot C_t\right)$ quantifies the performance of querying at level t of fidelity, which is measured as a product between the estimated IMSE and the computational cost. The integrated mean-square error, IMSE, is calculated as

$$\text{IMSE} = \int_{\mathcal{X}} \sigma^2(\boldsymbol{x}) d\boldsymbol{x}, \tag{28}$$

where the posterior variance field $\sigma^2(\boldsymbol{x})$ is hallucinated at the sampling location $\boldsymbol{x}^*$, i.e. assuming that $y(\boldsymbol{x}^*) = \mu(\boldsymbol{x}^*)$. The optimal level t^*, corresponding with the optimal product $\left(\text{IMSE}_{\text{t, hallucinated}} \cdot C_t\right)$ as a measure of cost and effectiveness, is selected to query the model.

Additionally, to promote the functional evaluation at highest fidelity level, i.e., level $t = s$, we choose the highest fidelity data (instead of $t = t^*$) if the ratio of number of training data points is larger than the computational cost ratio since the goal is to optimize at the highest level of fidelity $t = s$. In particular, let t^* be the optimal level of fidelity to query for the next sampling point, and $|\mathcal{D}^{(t)}|$ be the cardinality of the training dataset at level t of fidelity (i.e., the number of training data points at level t) and C_t be the computational cost at level t. We compare two quantities, $\left(C_{t^*} \cdot |\mathcal{D}^{(t^*)}|\right)$ and $\left(C_s \cdot |\mathcal{D}^{(s)}|\right)$. If $C_{t^*} \cdot |\mathcal{D}^{(t^*)}| \geq C_s \cdot |\mathcal{D}^{(s)}|$, which means some of the computational cost building $\mathcal{D}^{(t^*)}$ could be traded for building $\mathcal{D}^{(s)}$ (which is consistent with the policy of promoting evaluation at highest fidelity level s), then level s, i.e., the highest level of fidelity is chosen instead of t^*.

For the case of bi-fidelity, the criteria selection is obtained by restricting the multi-fidelity in (27) to the bi-fidelity settings. We compare the measure of the high-fidelity level $\left(\text{IMSE}_{\text{h, hallucinated}} \cdot C_h\right)$, and that of the low-fidelity level $\left(\text{IMSE}_{\text{l, hallucinated}} \cdot C_l\right)$. For the sake of convenience, we define a_{fidelity} ratio of measure at the high-fidelity level to that of low-fidelity as

$$a_{\text{fidelity}} := \frac{\text{IMSE}_{\text{h, hallucinated}}}{\text{IMSE}_{\text{l, hallucinated}}} \cdot \frac{C_h}{C_l}, \tag{29}$$

where C_h and C_l are the computational costs at the high- and low-fidelity levels, respectively. If $a_{\text{fidelity}} \leq 1$, then the function evaluator is called at the high-fidelity level, whereas if $a_{\text{fidelity}} > 1$, then the function is evaluated at the low-fidelity level. The proposed fidelity selection criteria defined in (29) determines the trade-off between running at low-fidelity and high-fidelity levels. If the high-fidelity return is higher than the low-fidelity, then the high-fidelity level is chosen, and vice versa.

In practice, to promote the high-fidelity evaluations, a hard condition is proposed to prevent the imbalance between low- and high-fidelity data sets based on the comparison between the number of data points available and the relative computational cost between high- and low-fidelity data. If the ratio of low-to-high fidelity data points is higher the relative computational cost, then the high-fidelity level will be chosen to evaluate the sampling locations. The IMSE is computed by Monte Carlo sampling in high-dimensional space. It is noted that if the relative computational cost between the high- and low-fidelity is 1, then fidelity criteria selection always promotes evaluating the sampling data point at the high-fidelity level.

4 Applications

In this section, we demonstrate the proposed sMF-BO-2CoGP through several analytical functions in Section 4.1, including 1d Forrester function [23] and a subset of benchmark functions from Kandasamy et al. [50] and Xiong et al. [51], including 2d Currin exponential function (two levels of fidelity), 8d borehole function (two and three levels of fidelity), welded beam design problem (four levels of fidelity), and an 11d real-world engineering application (two levels of fidelity) in designing flip-chip package (Section 4.6). Some implementations are adopted from Surjanovic and Bingham [52]. In all the benchmark of different acquisition functions, the computational cost between the high- and low-fidelity model is fixed at 2.50.

4.1 Forrester function (1d) with two fidelity levels

In this section, we consider the Forrester function [23], where $x \in [0,1]$, the low-fidelity function is

$$f_L(x) = 0.5(6x-2)^2 \sin(12x-4) + 10(x-0.5) - 5, \quad (30)$$

and the high-fidelity function is

$$f_H(x) = (6x-2)^2 \sin(12x-4). \quad (31)$$

First, consider a baseline set of 4 low-fidelity and 2 high-fidelity data points. We compare the effects of adding low- and high-fidelity observations on the prediction of CoKriging. Figure 2 shows the comparison between the posterior mean $\mu(\boldsymbol{x})$ and posterior variance $\sigma^2(\boldsymbol{x})$ between adding 4 more low-fidelity and 2 more high-fidelity data points. The common low-fidelity data points are denoted as blue squares, the common high-fidelity data points are denoted as black diamonds, and the added data points are denoted as red circles (low-fidelity for Figures 1a and 2a; high-fidelity for Figures 1b and 2b).

For the low-fidelity level, Figures 1a and 2a show the updated posterior mean $\mu(\boldsymbol{x})$ and posterior variance $\sigma^2(\boldsymbol{x})$ after 4 more low-fidelity data points are added, respectively. The posterior mean $\mu(\boldsymbol{x})$ prediction slightly improves near the end of the domain $x = 1$, but does not improve significantly near the other end of the domain $x = 0$ (Figure 1a). The posterior variance $\sigma^2(\boldsymbol{x})$ slightly reduces at the location where the low-fidelity data points are added.

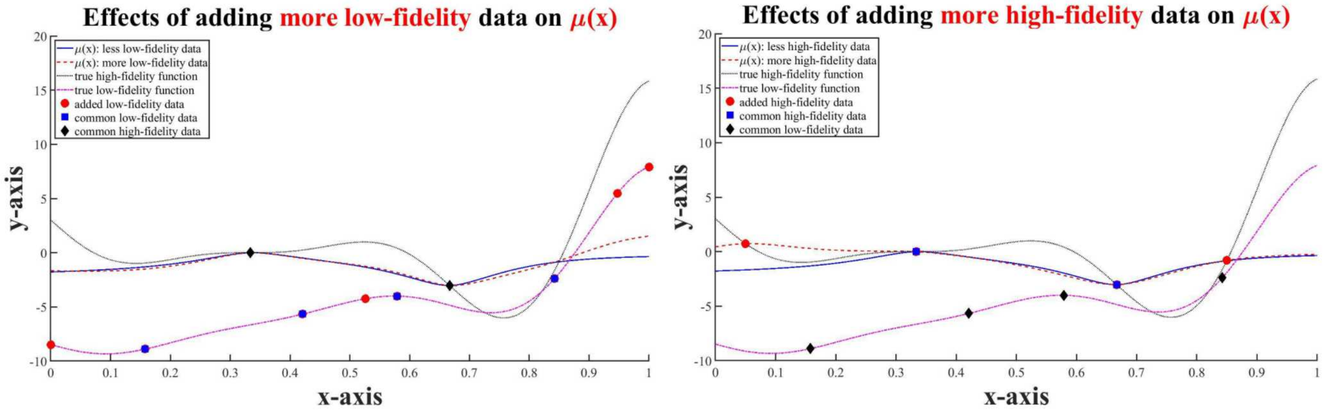
For the high-fidelity level, Figure 1b and Figure 2b show the updated posterior mean $\mu(\boldsymbol{x})$ and posterior variance $\sigma^2(\boldsymbol{x})$ after 2 more high-fidelity data points are added, respectively. The posterior mean $\mu(\boldsymbol{x})$ improves as expected, as shown in Figure 1b. The posterior variance $\sigma^2(\boldsymbol{x})$ reduces to zero for noiseless evaluations at the two added sampling locations.

Next, we test the numerical implementation of the sMF-BO-2CoGP method by considering the minimization problem $\operatorname{argmin} f_H(\boldsymbol{x})$ with no constraint and various computational relative cost ratio between the high- and low-fidelity levels. Figure 3a and Figure 3b show the convergence plot with respect to iterations and total computation cost, respectively. The case where the relative cost ratio is 1.0 serves as a benchmark for traditional sequential BO using only high-fidelity. We verified that when the relative cost ratio is 1.0, all the evaluations are evaluated only at high-fidelity level. When the relative cost ratio is higher than 1.0, the sMF-BO-2CoGP selects the fidelity criteria on-the-fly, using the fidelity criteria selection described above. It is worth noting that Figure 3a only shows the convergence plot at high-fidelity level. That means, the convergence plot only updates when a better high-fidelity result is available. The numerical performance at high-fidelity level of the multi-fidelity sMF-BO-2CoGP framework degrades when the computational cost ratio increases, because more low-fidelity points are selected at high computational cost ratio, according to Equation 29.

As shown in Figure 3a, when the computational cost ratio is 1.0, the sMF-BO-2CoGP converges to a sequential BO with high-fidelity, and is the fastest with respect to the number of iterations. Figure 3b shows comparable performances between the cases of ratio 1.0 and 2.5, where the performance degrades when the computational cost ratio increases. However, they all converge after approximately 7 iterations.

In this example, we consider an initial sampling data set comprised of 4 low-fidelity and 2 high-fidelity data points. The numerical performances are expected to change with different initial samples, as well as the behavior of high- and low-fidelity models.

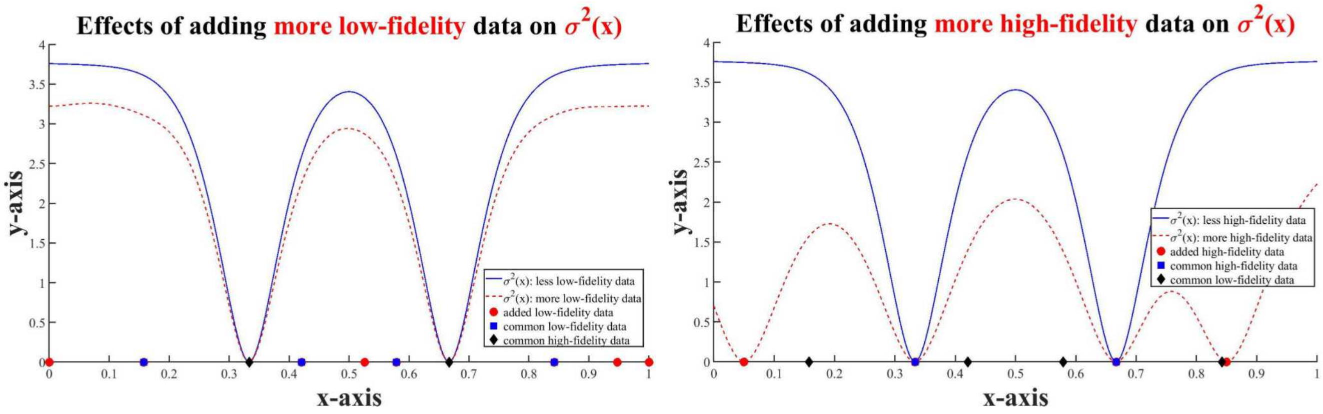
Figure 4 shows the convergence plot of 1d Forrester function. Five trial runs are performed with different initial samples. Bands are covered by the lower and upper bounds at a fixed iteration with respect to different trial runs. Solid lines denotes the mean objective function at a fixed iteration. The EI acquisition function is denoted with blue circles and blue band. The UCB acquisition function is denoted with red crosses and red band. The PI acquisition function is denoted with green squares and green band. Readers are referred to color version online. The UCB acquisition function converges slowly at the beginning, but outperforms the EI acquisition function later on. The PI acquisition function does not perform very well. In a case of UCB, 5 high-fidelity and 3 low-fidelity evaluations are performed to achieve convergence.



(a) $\mu(\boldsymbol{x})$: more low-fidelity data.

(b) $\mu(\boldsymbol{x})$: more high-fidelity data.

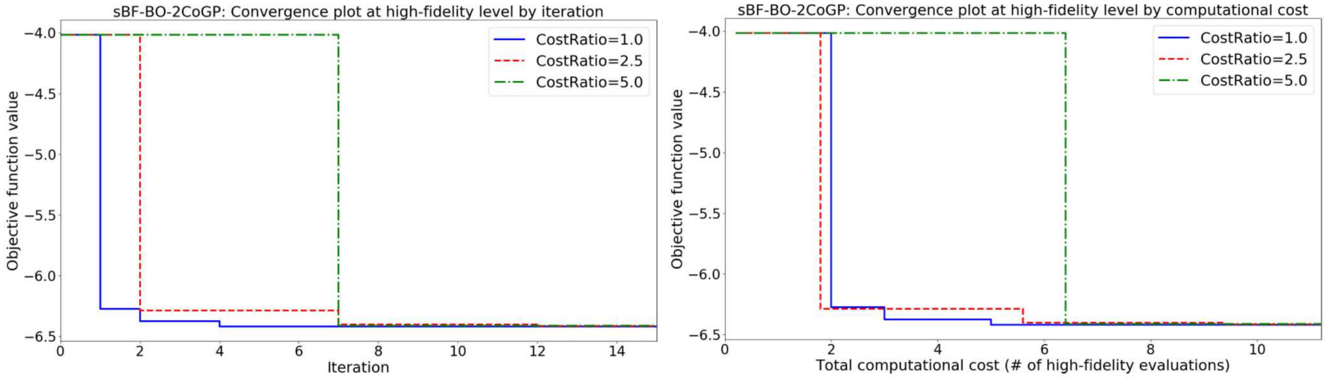
Fig. 1: Effects of adding more low-fidelity and high-fidelity on $\mu(\boldsymbol{x})$. Figure 1a shows the update posterior mean with 4 low-fidelity data points (denoted as red dots) added, whereas Figure 1b shows the update posterior mean with 2 high-fidelity data points (denoted as red dots) added. Readers are referred to the color online version.



(a) $\sigma^2(\boldsymbol{x})$: more low-fidelity data.

(b) $\sigma^2(\boldsymbol{x})$: more high-fidelity data.

Fig. 2: Effects of adding more low-fidelity and high-fidelity $\sigma^2(\boldsymbol{x})$. Figure 2a shows the update posterior variance with 4 low-fidelity data points (denoted as red dots) added, whereas Figure 2b shows the update posterior variance with 2 high-fidelity data points (denoted as red dots) added.



(a) Convergence plots by iteration with different relative computational cost ratios.

(b) Convergence plots by total computational cost with different relative computational cost ratios.

Fig. 3: Convergence plot of sMF-BO-2CoGP by iteration (Figure 3a) and by total computational cost (Figure 3b).

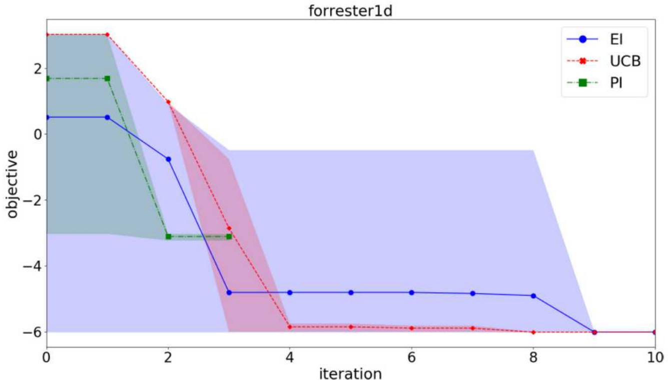


Fig. 4: Convergence plot of 1d Forrester function, two levels of fidelity, with different acquisition functions. Bands are encapsulated by the lower and upper bounds of objectives at a particular iteration.

4.2 Currin exponential function (2d) with two fidelity levels

We adopted the multi-fidelity Currin functions, where the high- and low-fidelity functions are written as, respectively,

$$f_H(\boldsymbol{x}) = \left[1 - \exp\left(-\frac{1}{2x_2}\right)\right] \frac{2300x_1^3 + 1900x_1^2 + 2092x_1 + 60}{100x_1^3 + 500x_1^2 + 4x_1 + 20}, \tag{32}$$

$$\begin{aligned} f_L(\boldsymbol{x}) = & \frac{1}{4}\Big[f_H(x_1 + 0.05, x_2 + 0.05) \\ & + f_H(x_1 + 0.05, \max(0, x_2 - 0.05))\Big] \\ & + \frac{1}{4}\Big[f_H(x_1 - 0.05, x_2 + 0.05) \\ & + f_H(x_1 - 0.05, \max(0, x_2 - 0.05))\Big] \end{aligned}, \tag{33}$$

where the domain is $\mathcal{X} = [0,1] \times [0,1]$.

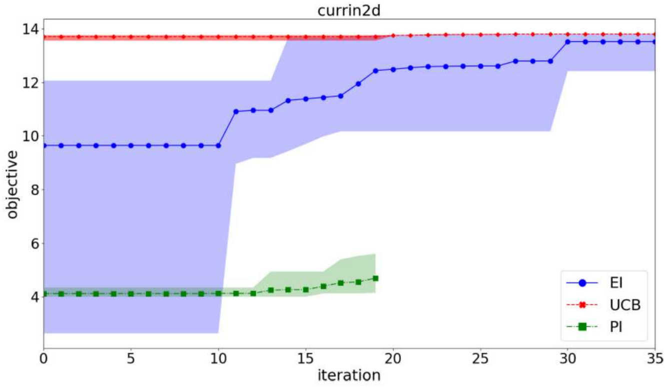


Fig. 5: Convergence plot of 2d Currin function, two levels of fidelity, with different acquisition functions. Bands are encapsulated by the lower and upper bounds of objectives at a particular iteration.

Figure 5 shows the convergence plot of 2d Currin function. The explanation of the plots is similar to the case of 1d Forrester function. In this example, the UCB acquisition function outperforms both the EI and PI acquisition functions. The PI acquisition function performs poorly. In a case of EI, 16 high-fidelity and 14 low-fidelity evaluations are performed to achieve convergence.

4.3 Welded beam design problem (4d) with four fidelity levels

In this example, we adopted the welded beam design problem from one of our previous work [39], illustrated in Figure 6. The objective function f is calculated as

$$f(w,m,h,l,t,b) = (1 + C_1)(wt + l)h^2 + C_2 tb(L + l) \tag{34}$$

subject to five known inequality constraints,

$$\text{shear stress}(\tau): g_1 = 0.577\sigma_d - \tau(\boldsymbol{x}) \geq 0, \tag{35}$$

$$\text{bending stress in the beam}(\sigma): g_2 = \sigma_d - \sigma(\boldsymbol{x}) \geq 0, \tag{36}$$

$$\text{buckling load on the bar}(P_c): g_3 = b - h \geq 0, \tag{37}$$

$$\text{deflection of the beam}: g_4 = P_c(\boldsymbol{x}) - F \geq 0, \tag{38}$$

$$\text{side constraints}: g_5 = \delta_{\max} - \delta(\boldsymbol{x}) \geq 0, \tag{39}$$

where

$$\sigma(\boldsymbol{x}) = \frac{6FL}{t^2 b}, \tag{40}$$

$$\delta(\boldsymbol{x}) = \frac{4FL^3}{Et^3 b}, \tag{41}$$

$$P_c(\boldsymbol{x}) = \frac{4.013tb^3\sqrt{EG}}{6L^2}\left(1 - \frac{t}{4L}\sqrt{\frac{E}{G}}\right), \tag{42}$$

$$\tau = \sqrt{(\tau')^2 + (\tau'')^2 + 2\tau'\tau''\cos\theta}, \tag{43}$$

$$\tau' = \frac{F}{A}, \tag{44}$$

$$\tau'' = \frac{F(L + 0.5l)R}{J} \tag{45}$$

$$w = 0: \begin{cases} A = \sqrt{2}hl \\ J = \sqrt{2}hl\left[\frac{(h+t)^2}{4} + \frac{l^2}{12}\right] \\ R = \frac{1}{2}\sqrt{l^2 + (h+t)^2} \\ \cos\theta = \frac{l}{2R} \end{cases}, \tag{46}$$

$$w = 1: \begin{cases} A = \sqrt{2}h(t+l) \\ J = \sqrt{2}hl\left[\frac{(h+t)^2}{4} + \frac{l^2}{12}\right] + \sqrt{2}ht\left[\frac{(h+l)^2}{4} + \frac{t^2}{12}\right] \\ R = \max\left\{\frac{1}{2}\sqrt{l^2 + (h+t)^2}, \frac{1}{2}\sqrt{t^2 + (h+l)^2}\right\} \\ \cos\theta = \frac{l}{2R} \end{cases}. \tag{47}$$

Table 1: Material-dependent parameters and constants in the welded beam design problem.

Constants	Description	steel	cast iron	aluminum	brass
C_1	cost per volume of the welded material ($/in^3)	0.1047	0.0489	0.5235	0.5584
C_2	cost per volume of the bar stock ($/in^3)	0.0481	0.0224	0.2405	0.2566
σ_d	design normal stress of the bar material (psi)	$30 \cdot 10^3$	$8 \cdot 10^3$	$5 \cdot 10^3$	$8 \cdot 10^3$
E	Young's modulus of bar stock (psi)	$30 \cdot 10^6$	$14 \cdot 10^6$	$10 \cdot 10^6$	$16 \cdot 10^6$
G	shear modulus of bar stock (psi)	$12 \cdot 10^6$	$6 \cdot 10^6$	$4 \cdot 10^6$	$6 \cdot 10^6$

$C_1(m)$, $C_2(m)$, $\sigma_d(m)$, $E(m)$, $G(m)$ are parameters that depend on the bulk materials m, as listed in Table 1. The lower and upper bounds of the problem are $0.0625 \leq h \leq 2$, $0.1 \leq l \leq 10$, $2.0 \leq t \leq 20.0$, and $0.0625 \leq b \leq 2.0$. $m \in \{1,2,3,4\}$ encodes the bulk materials as steel, cast iron, aluminum, and brass, respectively.

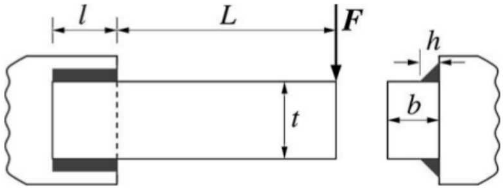


Fig. 6: Welded beam design problem [53].

The goal is to minimize the objective function $f(w,m,h,l,t,b)$, which is the estimated cost. The physical meaning of the input parameters is as follows. h is the thickness of the weld; l is the length of the welded join, t is the width of the beam; b is the thickness of the beam. m is a discrete variable enumerating the bulk material of the beam, which can be steel, cast iron, aluminum, or brass. w is a binary variable to model the type of weld: $w = 0$ for two-sided welding, and $w = 1$ for four-sided welding. Different bulk materials are associated with different materials property, as described in Table 1. The fixed design parameters of the beam are $L = 14$ inch, $\delta_{\max} = 0.25$ inch, and $F = 6{,}000$ lb.

Compared to the example [39], here, we fix $w = 0$, and consider four levels of fidelity. The high-fidelity function and three low-fidelity functions are

$$f_H(\boldsymbol{x}) = f(w=0, m=1, h, l, t, b), \tag{48}$$

$$f_{L_1}(\boldsymbol{x}) = f(w=0, m=2, h, l, t, b), \tag{49}$$

$$f_{L_2}(\boldsymbol{x}) = f(w=0, m=3, h, l, t, b), \tag{50}$$

$$f_{L_3}(\boldsymbol{x}) = f(w=0, m=4, h, l, t, b). \tag{51}$$

We wish to minimize the cost of using steel as the bulk material at the high-fidelity level, where the low-fidelity functions $f_{L_1}(\boldsymbol{x})$, $f_{L_2}(\boldsymbol{x})$, and $f_{L_3}(\boldsymbol{x})$ are used to estimate the cost of using cast iron, aluminum, and brass, respectively. The computational cost of the high-fidelity level are 2.5 times higher than that of low-fidelity levels.

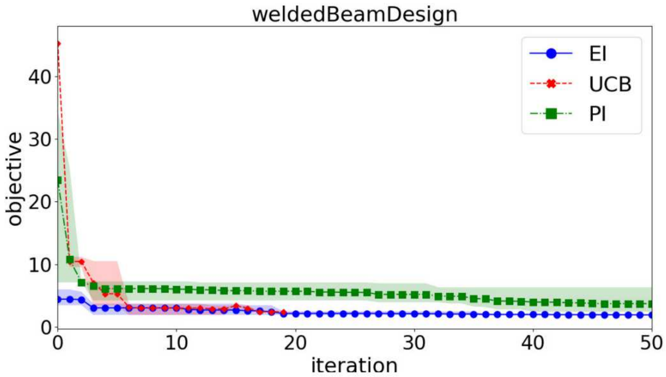


Fig. 7: Convergence plot of 4d welded beam design function, four levels of fidelity, with different acquisition functions. Bands are encapsulated by the lower and upper bounds of objectives at a particular iteration.

Figure 7 shows the convergence plot of the welded beam design problem with four levels of fidelity. In this example, the UCB and EI acquisition functions perform on par with each other and outperform the PI acquisition functions. In a case of EI, 6 high-fidelity and 14 low-fidelity evaluations are performed to achieve convergence.

4.4 Borehole function (8d) with two fidelity levels

In this example, we adopted the multi-fidelity borehole function from Xiong et al. [51], where two fidelity levels are considered. The high- and low-fidelity functions are, respectively,

$$f_H(\boldsymbol{x}) = \frac{2\pi x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1 + \frac{2x_7x_3}{\log(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)}, \tag{52}$$

$$f_L(\boldsymbol{x}) = \frac{5x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1.5 + \frac{2x_7x_3}{\log(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)}. \tag{53}$$

The domain of the 8d borehole function is $x_1 \in [0.05, 0.15]$, $x_2 \in [100, 50000]$, $x_3 \in [63070, 115600]$, $x_4 \in [990, 1110]$, $x_5 \in [63.1, 116]$, $x_6 \in [700, 820]$, $x_7 \in [1120, 1680]$, $x_8 \in [9855, 12045]$.

Figure 8 shows the convergence plot of 8d borehole function. The explanation of the plots is similar to the case of 1d Forrester function. In this example, the EI acquisition

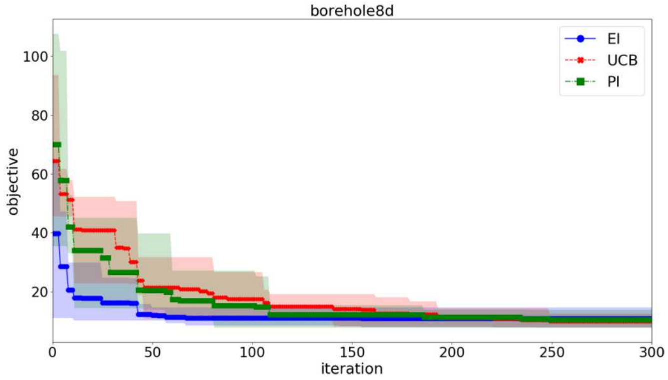


Fig. 8: Convergence plot of 8d borehole function, two levels of fidelity, with different acquisition functions. Bands are encapsulated by the lower and upper bounds of objectives at a particular iteration.

function outperforms both the UCB and PI acquisition functions, where the PI and UCB acquisition functions perform on-par with each other. In the case of EI, 22 high-fidelity and 53 low-fidelity evaluations are performed to achieve convergence.

4.5 Borehole function (8d) with three fidelity levels

We further modify and extend the previous 8d borehole function in the previous example with three levels of fidelity. The high-fidelity and low-fidelity functions are described as,

$$f_H(\boldsymbol{x}) = \frac{2\pi x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1 + \frac{2x_7x_3}{\log(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)}, \tag{54}$$

$$f_{L_1}(\boldsymbol{x}) = \frac{5x_3(x_4 - x_6)}{\log(x_2/x_1)\left(1.5 + \frac{2x_7x_3}{\log(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)}, \tag{55}$$

$$f_{L_2}(\boldsymbol{x}) = \frac{7x_3(x_4 - x_6)}{\log(x_2/x_1)\left(0.5 + \frac{2x_7x_3}{\log(x_2/x_1)x_1^2x_8} + \frac{x_3}{x_5}\right)}. \tag{56}$$

It is obvious to see that $f_{L_1}(\boldsymbol{x}) \leq f_H(\boldsymbol{x}) \leq f_{L_2}(\boldsymbol{x})$; thus, the high-fidelity function is bounded between two low-fidelity functions. Figure 9 shows the convergence plot of the 8d borehole function with three levels of fidelity. In this example, all the acquisition functions, including EI, UCB, and PI, perform on par with each other, almost throughout the optimization process. In the case of PI, 28 high-fidelity and 82 low-fidelity evaluations are performed to achieve convergence.

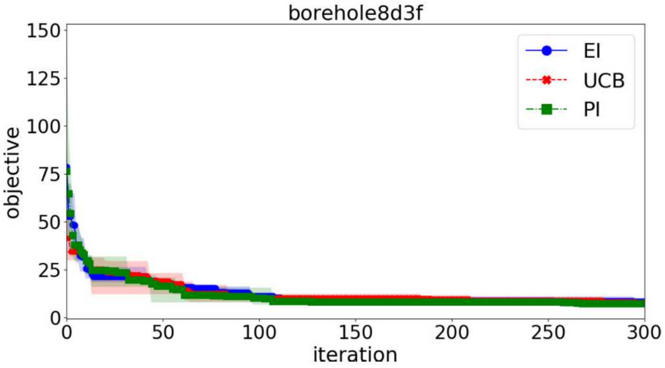


Fig. 9: Convergence plot of 8d borehole function, three levels of fidelity, with different acquisition functions. Bands are encapsulated by the lower and upper bounds of objectives at a particular iteration.

4.6 Flip-chip package design (11d) with two fidelity levels

In this section, we demonstrate the design application of a flip-chip package using the proposed sMF-BO-2CoGP, where the details of development and implementation are fully described in our previous work [54]. A lidless flip-chip package with a monolithic silicon die (FCBGA) mounted on a printed circuit board (PCB) with a stiffener ring is considered in this example. The computational model is constructed based on a 2.5D, half symmetry to reduce the computational time.

Figure 10 shows the geometric model of the thermomechanical finite element model (FEM), where the mesh density varies for different levels of fidelity. Two design variables are associated with the die, three are associated with the substrate, three more are associated with the stiffener ring, two are with the underfill, and the last one is with the PCB board. Only two levels of fidelity are considered in this example. Table 2 show the design variables, the physical meaning of the design variables, as well as their lower and upper bounds in this case study.

Table 2: Design variables for the FCBGA design optimization.

Variable	Design part	Lower bound	Upper bound	Optimal value
x_1	die	20000	30000	20702
x_2	die	300	750	320
x_3	substrate	30000	40000	35539
x_4	substrate	100	1800	1614
x_5	substrate	$10 \cdot 10^{-6}$	$17 \cdot 10^{-6}$	$17 \cdot 10^{-6}$
x_6	stiffener ring	2000	6000	4126
x_7	stiffener ring	100	2500	1646
x_8	stiffener ring	$8 \cdot 10^{-6}$	$25 \cdot 10^{-6}$	$8.94 \cdot 10^{-6}$
x_9	underfill	1.0	3.0	1.52
x_{10}	underfill	0.5	1.0	0.804
x_{11}	PCB board	$12.0 \cdot 10^{-6}$	$16.7 \cdot 10^{-6}$	$16.7 \cdot 10^{-6}$

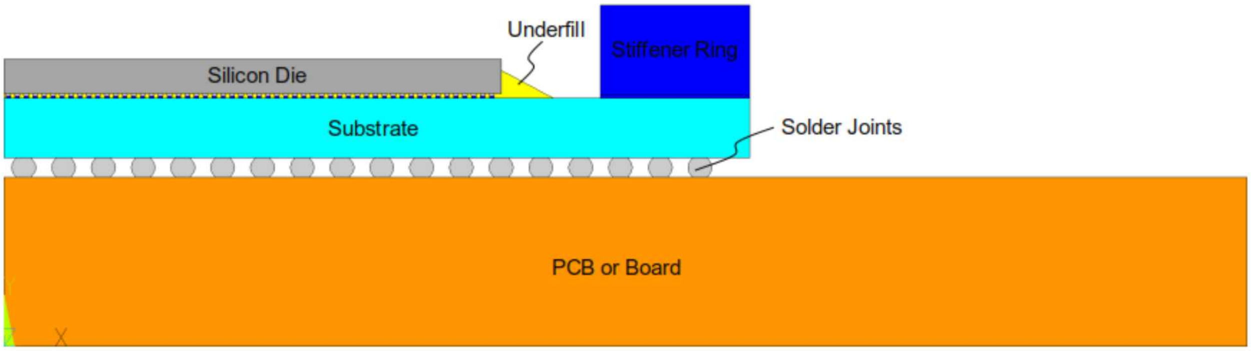


Fig. 10: Finite element model geometry.

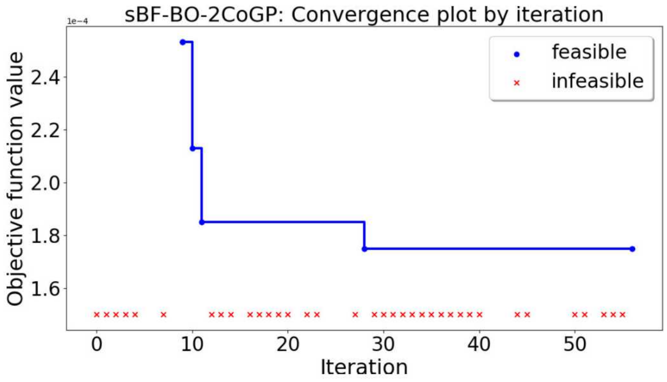


Fig. 11: Convergence plot of flip-chip package design evaluation to minimize the flip-chip warpage.

After the numerical solution is obtained, the component warpage at 20°C, 200°C, and the strain energy density of the furthest solder joint are calculated. The strain energy density is one of accurate predictors to estimate the fatigue life of the solder joints during thermal cycling [55].

A vectorized 11-dimensional input is used to parameterize the design. 9 low-fidelity and 3 high-fidelity data points are used as initial samples. It is noted that not all of the initial samples are feasible. There are some unknown constraints, but no known constraint is imposed in this example. We consider that the sampling locations where the FEM solutions diverge are infeasible. This condition can be regarded as an unknown constraint, because no prior knowledge regarding divergence is known beforehand but only after the simulation is finished. ANSYS Parametric Design Language (APDL) software is used to evaluate the model in the batch mode with no graphical user interface. The sMF-BO-2CoGP is implemented in MATLAB, where an interface using Python is devised to communicate with the APDL FEM model. The average computational time for one iteration is approximately 0.4 CPU hour.

Figure 11 presents the convergence plot of the FCBGA design optimization, where the feasible sampling points are plotted as blue circles, whereas the infeasible sampling points are plotted as red squares. The EI acquisition function is used in this example to locate the next sampling point. It is observed that the predicted warpage is converging steadily. The numerical solver fails to converge on many cases. It has also demonstrated that the proposed sMF-BO-2CoGP is robust against diverging simulations, by its convergent objective despite numerous failed cases.

The optimization results are relatively close to designs commonly used in the microelectronics packing industry. It is observed that thin and small die, as well as thick substrate, are suggested in order to minimize the component warpage.

5 Discussion

The main contribution of this work is the proposal of the fidelity selection criteria. The criteria is inspired by the work of Huang et al. [56], where the original criteria is proposed based on the EI acquisition function as

$$EI(\boldsymbol{x}, l) = \quad EI_m(\boldsymbol{x}) \tag{57}$$

$$\times \quad \mathrm{Corr}(f_l^p(\boldsymbol{x}), f_m^p(\boldsymbol{x})) \tag{58}$$

$$\times \quad \left(1 - \frac{\sigma_{\varepsilon,l}}{\sqrt{s_l^2(\boldsymbol{x}) + \sigma_{\varepsilon,l}^2}}\right) \tag{59}$$

$$\times \quad \frac{C_m}{C_l}, \tag{60}$$

where m is an arbitrary level of fidelity, and l is the highest level of fidelity. In this scheme, after each point is nominated at a level of fidelity, a unique sampling point is chosen by looping over all the levels. The uncertainty reduction is measured in the second term of the above equation, $\left(1 - \frac{\sigma_{\varepsilon,l}}{\sqrt{s_l^2(\boldsymbol{x}) + \sigma_{\varepsilon,l}^2}}\right)$. In our scheme, the uncertainty is measured by $\frac{\mathrm{IMSE}_{\text{h, hallucinated}}}{\mathrm{IMSE}_{\text{l, hallucinated}}}$ in Equation 29. One advantage of the proposed criteria is that it truly estimates the reduction of uncertainty at a particular level. While the uncertainty could be measured by the maximum $\sigma^2(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{X}$ for the uncertainty reduction, the maximal location is often found on the border of the bounded domain. Another advantage of the proposed criteria is that it removes the restriction of using EI acquisition, and generalizes to any arbitrary acquisition function. The choice of the acquisition function is left to users as a choice. Previous work by Gauthier et al. [57, 58] and Silvestrini et al. [59] have shown that the performance of IMSE supersedes the performance of maximal MSE. The scheme proposed by Huang et al. [56] in Equation 57 can be further generalized to some other commonly used acquisition functions, such as PI and UCB. Furthermore, multiple acquisition functions can be considered simultaneously based on their performance, as in GP-Hedge scheme [60].

In the implementation, the CMA-ES framework is adopted to maximize the acquisition function $a^*(\boldsymbol{x})$. For computationally expensive high-fidelity simulations, the CMA-ES parameters must be tuned to search carefully with multiple restarts to avoid local minima. In practice, optimizing the acquisition function takes some amount of time, thus it also reduces the efficiency of the method. However, it has been rarely discussed in the literature, and there has not been much work dedicated to benchmarking and quantifying the computational cost of this process. For batch-sequential parallel BO approaches, the computational cost is much more severe, particularly with simulations that are associated with large infeasible space.

The use of the probabilistic binary classifier to learn and distinguish feasible and infeasible region also depends many factors of the problems. Essentially, the classifier needs to accurately predict the feasibility before the optimal point is obtained. This depends largely on the dimensionality of the problem considered. However, once the feasibility is accurately predicted through Equation 25, the convergence to the global optimal point is guaranteed through the classical BO framework. The analytical convergence rate can be found in the seminal work of Rasmussen [46].

BO is indeed a flexible framework that allows for numerous possible extensions in engineering domains [61]. One of those are multi-fidelity, which is studied in this paper. Other extensions include batch-sequential parallel sampling [38], asynchronously parallel sampling [40], mixed-integer optimization with a small number of discrete/categorical variables [39], latent variable model [62]. More sophisticated GP models, including local GP, [63, 64], sparse GP [65], heteroscedastic [66], and even deep learning [67], have been developed to widen the range of multi-disciplinary applications. This area remains active for further research.

While the proposed sequential multi-fidelity sMF-BO-2CoGP aims at improving the efficiency compared to the sequential high-fidelity BO, the efficiency can be further improved by performing parallel optimization. That is to sample multiple locations concurrently (i.e. at the same time) and asynchronously (i.e. sampling points do not have to wait for others to complete). The proposed multi-fidelity framework serves as a foundation work to tackle the constrained multi-fidelity problem in an asynchronously parallel manner. The research question remains open and poses as a potential future work.

6 Conclusion

In this paper, a sequential multi-fidelity BO optimization, called sMF-BO-2CoGP, is proposed to solve the constrained simulation-based optimization problem. A fidelity selection criteria is proposed to determine the level of fidelity for evaluating the objective function value. Another CoKriging model is coupled into the method to classify the next sampling point and distinguish between feasible and infeasible regions.

The proposed sMF-BO-2CoGP method is demonstrated using a simple analytic 1D example, as well as an engineering thermomechanical FEM for flip-chip package design optimization. The preliminary results provided in this study demonstrates the applicability of the proposed sMF-BO-2CoGP method.

Acknowledgements

This research was supported by the U.S. Department of Energy, Office of Science, Early Career Research Program, under award 17020246 and the Sandia LDRD program. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

References

[1] Digabel, S. L., and Wild, S. M., 2015. "A taxonomy of constraints in simulation-based optimization". *arXiv preprint arXiv:1505.07881*.

[2] Kennedy, M. C., and O'Hagan, A., 2000. "Predicting the output from a complex computer code when fast approximations are available". *Biometrika,* **87**(1), pp. 1–13.

[3] Tran, A., Wildey, T., and McCann, S., 2019. "sBF-BO-2CoGP: A sequential bi-fidelity constrained Bayesian optimization for design applications". In Proceedings of the ASME 2019 IDETC/CIE, Vol. Volume 1: 39th Computers and Information in Engineering Conference of *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, American Society of Mechanical Engineers. V001T02A073.

[4] Brochu, E., Cora, V. M., and De Freitas, N., 2010. "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning". *arXiv preprint arXiv:1012.2599*.

[5] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N., 2016. "Taking the human out of the loop: A review of Bayesian optimization". *Proceedings of the IEEE,* **104**(1), pp. 148–175.

[6] Frazier, P. I., 2018. "A tutorial on Bayesian optimization". *arXiv preprint arXiv:1807.02811*.

[7] Jones, D. R., Schonlau, M., and Welch, W. J., 1998. "Efficient global optimization of expensive black-box functions". *Journal of Global Optimization,* **13**(4), pp. 455–492.

[8] Le Gratiet, L., and Garnier, J., 2014. "Recursive co-kriging model for design of computer experiments with multiple levels of fidelity". *International Journal for Uncertainty Quantification,* **4**(5).

[9] Raissi, M., and Karniadakis, G., 2016. "Deep multi-fidelity Gaussian processes". *arXiv preprint arXiv:1604.07484*.

[10] Raissi, M., Perdikaris, P., and Karniadakis, G. E., 2017. "Machine learning of linear differential equations using Gaussian processes". *Journal of Computational Physics,* **348**, pp. 683–693.

[11] Perdikaris, P., Venturi, D., Royset, J., and Karniadakis, G., 2015. "Multi-fidelity modelling via recursive co-kriging and Gaussian–Markov random fields". In Proc. R. Soc. A, Vol. 471, The Royal Society, p. 20150018.

[12] Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N., and Karniadakis, G. E., 2017. "Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling". *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences,* **473**(2198), p. 20160751.

[13] Poloczek, M., Wang, J., and Frazier, P., 2017. "Multi-information source optimization". In Advances in Neural Information Processing Systems, pp. 4288–4298.

[14] Swersky, K., Snoek, J., and Adams, R. P., 2013. "Multi-task Bayesian optimization". In Advances in neural information processing systems, pp. 2004–2012.

[15] Ghoreishi, S. F., and Allaire, D., 2019. "Multi-information source constrained Bayesian optimization". *Structural and Multidisciplinary Optimization,* **59**(3), pp. 977–991.

[16] Ghoreishi, S. F., and Allaire, D. L., 2018. "A fusion-based multi-information source optimization approach using knowledge gradient policies". In 2018 AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, p. 1159.

[17] Ghoreishi, S. F., and Allaire, D. L., 2018. "Gaussian process regression for Bayesian fusion of multi-fidelity information sources". In 2018 Multidisciplinary Analysis and Optimization Conference, p. 4176.

[18] Ghoreishi, S. F., Molkeri, A., Srivastava, A., Arroyave, R., and Allaire, D., 2018. "Multi-information source fusion and optimization to realize ICME: Application to dual-phase materials". *Journal of Mechanical Design,* **140**(11), p. 111409.

[19] Xiao, M., Zhang, G., Breitkopf, P., Villon, P., and Zhang, W., 2018. "Extended co-kriging interpolation method based on multi-fidelity data". *Applied Mathematics and Computation,* **323**, pp. 120–131.

[20] Couckuyt, I., Forrester, A., Gorissen, D., De Turck, F., and Dhaene, T., 2012. "Blind kriging: Implementation and performance analysis". *Advances in Engineering Software,* **49**, pp. 1–13.

[21] Couckuyt, I., Dhaene, T., and Demeester, P., 2013. "ooDACE toolbox, A Matlab Kriging toolbox: Getting started". *Universiteit Gent*, pp. 3–15.

[22] Couckuyt, I., Dhaene, T., and Demeester, P., 2014. "ooDACE toolbox: a flexible object-oriented Kriging implementation". *The Journal of Machine Learning Research,* **15**(1), pp. 3183–3186.

[23] Forrester, A. I., Sóbester, A., and Keane, A. J., 2007. "Multi-fidelity optimization via surrogate modelling". *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences,* **463**(2088), pp. 3251–3269.

[24] Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z., 2014. "Predictive entropy search for efficient global optimization of black-box functions". In Advances in neural information processing systems, pp. 918–926.

[25] Hernández-Lobato, J. M., Gelbart, M., Hoffman, M., Adams, R., and Ghahramani, Z., 2015. "Predictive entropy search for Bayesian optimization with unknown constraints". In International Conference on Machine Learning, pp. 1699–1707.

[26] Hernández-Lobato, D., Hernández-Lobato, J., Shah, A., and Adams, R., 2016. "Predictive entropy search for multi-objective Bayesian optimization". In International Conference on Machine Learning, pp. 1492–1501.

[27] Hennig, P., and Schuler, C. J., 2012. "Entropy search for information-efficient global optimization". *Journal of Machine Learning Research,* **13**(Jun), pp. 1809–1837.

[28] Wang, Z., Zhou, B., and Jegelka, S., 2016. "Optimization as estimation with Gaussian processes in bandit settings". In Artificial Intelligence and Statistics, pp. 1022–1031.

[29] Shahriari, B., Wang, Z., Hoffman, M. W., Bouchard-Côté, A., and de Freitas, N., 2014. "An entropy search portfolio for Bayesian optimization". *arXiv preprint arXiv:1406.4625*.

[30] Kushner, H. J., 1964. "A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise". *Journal of Basic Engineering,* **86**(1), pp. 97–106.

[31] Mockus, J., 1975. "On Bayesian methods for seeking the extremum". In Optimization Techniques IFIP Technical Conference, Springer, pp. 400–404.

[32] Mockus, J., 1982. "The Bayesian approach to global optimization". *System Modeling and Optimization*, pp. 473–481.

[33] Bull, A. D., 2011. "Convergence rates of efficient global optimization algorithms". *Journal of Machine Learning Research,* **12**(Oct), pp. 2879–2904.

[34] Snoek, J., Larochelle, H., and Adams, R. P., 2012. "Practical Bayesian optimization of machine learning algorithms". In Advances in neural information processing systems, pp. 2951–2959.

[35] Srinivas, N., Krause, A., Kakade, S. M., and Seeger,

M., 2009. “Gaussian process optimization in the bandit setting: No regret and experimental design”. *arXiv preprint arXiv:0912.3995*.

[36] Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W., 2012. “Information-theoretic regret bounds for Gaussian process optimization in the bandit setting”. *IEEE Transactions on Information Theory,* **58**(5), pp. 3250–3265.

[37] Daniel, C., Viering, M., Metz, J., Kroemer, O., and Peters, J., 2014. “Active reward learning.”. In Robotics: Science and Systems.

[38] Tran, A., Sun, J., Furlan, J. M., Pagalthivarthi, K. V., Visintainer, R. J., and Wang, Y., 2019. “pBO-2GP-3B: A batch parallel known/unknown constrained Bayesian optimization with feasibility classification and its applications in computational fluid dynamics”. *Computer Methods in Applied Mechanics and Engineering,* **347**, pp. 827–852.

[39] Tran, A., Tran, M., and Wang, Y., 2019. “Constrained mixed-integer Gaussian mixture Bayesian optimization and its applications in designing fractal and auxetic metamaterials”. *Structural and Multidisciplinary Optimization*, pp. 1–24.

[40] Tran, A., Scott, M., Furlan, J. M., Pagalthivarthi, K. V., Visintainer, R. J., and Wildey, T., 2019. “aphBO-2GP-3B: A budgeted asynchronously-parallel multi-acquisition for known/unknown constrained Bayesian optimization on high-performing computing architecture”. *Reliability Engineering and System Safety*.

[41] Bentley, J. L., 1975. “Multidimensional binary search trees used for associative searching”. *Communications of the ACM,* **18**(9), pp. 509–517.

[42] Hastie, T., Rosset, S., Zhu, J., and Zou, H., 2009. “Multi-class AdaBoost”. *Statistics and its Interface,* **2**(3), pp. 349–360.

[43] Breiman, L., 2001. “Random forests”. *Machine learning,* **45**(1), pp. 5–32.

[44] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B., 1998. “Support vector machines”. *IEEE Intelligent Systems and their applications,* **13**(4), pp. 18–28.

[45] Suykens, J. A., and Vandewalle, J., 1999. “Least squares support vector machine classifiers”. *Neural processing letters,* **9**(3), pp. 293–300.

[46] Rasmussen, C. E., 2004. “Gaussian processes in machine learning”. In *Advanced lectures on machine learning*. Springer, pp. 63–71.

[47] LeCun, Y., Bengio, Y., and Hinton, G., 2015. “Deep learning”. *nature,* **521**(7553), p. 436.

[48] Hansen, N., and Ostermeier, A., 2001. “Completely derandomized self-adaptation in evolution strategies”. *Evolutionary computation,* **9**(2), pp. 159–195.

[49] Hansen, N., Müller, S. D., and Koumoutsakos, P., 2003. “Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)”. *Evolutionary computation,* **11**(1), pp. 1–18.

[50] Kandasamy, K., Dasarathy, G., Schneider, J., and Poczos, B., 2017. “Multi-fidelity Bayesian optimisation with continuous approximations”. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, pp. 1799–1808.

[51] Xiong, S., Qian, P. Z., and Wu, C. J., 2013. “Sequential design and analysis of high-accuracy and low-accuracy computer codes”. *Technometrics,* **55**(1), pp. 37–46.

[52] Surjanovic, S., and Bingham, D. Virtual library of simulation experiments: test functions and datasets. https://www.sfu.ca/ssurjano/optimization.html.

[53] Datta, D., and Figueira, J. R., 2011. “A real-integer-discrete-coded particle swarm optimization for design problems”. *Applied Soft Computing,* **11**(4), pp. 3625–3633.

[54] McCann, S., Ostrowicki, G. T., Tran, A., Huang, T., Bernhard, T., Tummala, R. R., and Sitaraman, S. K., 2017. “Determination of energy release rate through sequential crack extension”. *Journal of Electronic Packaging,* **139**(4), p. 041003.

[55] Darveaux, R., 2000. “Effect of simulation methodology on solder joint crack growth correlation”. In Electronic Components & Technology Conference, 2000. 2000 Proceedings. 50th, IEEE, pp. 1048–1058.

[56] Huang, D., Allen, T. T., Notz, W. I., and Miller, R. A., 2006. “Sequential kriging optimization using multiple-fidelity evaluations”. *Structural and Multidisciplinary Optimization,* **32**(5), pp. 369–382.

[57] Gauthier, B., and Pronzato, L., 2014. “Spectral approximation of the IMSE criterion for optimal designs in kernel-based interpolation models”. *SIAM/ASA Journal on Uncertainty Quantification,* **2**(1), pp. 805–825.

[58] Gauthier, B., and Pronzato, L., 2017. “Convex relaxation for IMSE optimal design in random-field models”. *Computational Statistics & Data Analysis,* **113**, pp. 375–394.

[59] Silvestrini, R. T., Montgomery, D. C., and Jones, B., 2013. “Comparing computer experiments for the Gaussian process model using integrated prediction variance”. *Quality Engineering,* **25**(2), pp. 164–174.

[60] Hoffman, M. D., Brochu, E., and de Freitas, N., 2011. “Portfolio allocation for Bayesian optimization.”. In UAI, Citeseer, pp. 327–336.

[61] Travaglino, S., Murdock, K., Tran, A., Martin, C., Liang, L., Wang, Y., and Sun, W., 2020. “Computational optimization study of transcatheter aortic valve leaflet design using porcine and bovine leaflets”. *Journal of Biomechanical Engineering,* **142**.

[62] Lawrence, N. D., 2004. “Gaussian process latent variable models for visualisation of high dimensional data”. In Advances in neural information processing systems, pp. 329–336.

[63] Tran, A., He, L., and Wang, Y., 2018. “An efficient first-principles saddle point searching method based on distributed kriging metamodels”. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering,* **4**(1), p. 011006.

[64] Tran, A., Furlan, J. M., Pagalthivarthi, K. V., Visin-

tainer, R. J., Wildey, T., and Wang, Y., 2019. “WearGP: A computationally efficient machine learning framework for local erosive wear predictions via nodal Gaussian processes”. *Wear,* **422**, pp. 9–26.

[65] Snelson, E., and Ghahramani, Z., 2006. “Sparse Gaussian processes using pseudo-inputs”. In Advances in neural information processing systems, pp. 1257–1264.

[66] Kersting, K., Plagemann, C., Pfaff, P., and Burgard, W., 2007. “Most likely heteroscedastic Gaussian process regression”. In Proceedings of the 24th international conference on Machine learning, ACM, pp. 393–400.

[67] Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M. M. A., Prabhat, M., and Adams, R. P., 2015. “Scalable Bayesian optimization using deep neural networks.”. In ICML, pp. 2171–2180.