SAND2019-9537C

sBF-BO-2CoGP: A sequential bi-fidelity constrained Bayesian optimization for design applications

Anh Tran*, Tim Wildey
Optimization and Uncertainty Quantification
Sandia National Laboratories, Albuquerque, NM 87123
Email: anhtran, twilde@sandia.gov

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

Bayesian optimization is an effective surrogate-based optimization method that has been widely used for simulation-based applications. However, the traditional Bayesian optimization (BO) method is only applicable to single-fidelity applications, whereas multiple levels of fidelity exist in reality. In this work, we propose a bi-fidelity known/unknown constrained Bayesian optimization method for design applications. The proposed framework, called sBF-BO-2CoGP, is built on a two-level CoKriging method to predict the objective function. An external binary classifier, which is also another CoKriging model, is used to distinguish between feasible and infeasible regions. The sBF-BO-2CoGP method is demonstrated using a numerical example and a flip-chip application for design optimization to minimize the warpage deformation under thermal loading conditions.

1 Introduction

Numerous high-fidelity engineering models are developed nowadays. These models are usually used to predict some properties and performances of interests, with respect to a specific design. The properties and performance predictions are then feedback into the design process to find a better design that outperform the previous ones by changing a few design parameters. Such process is ubiquitous in industrial settings, so called design optimization process. Simulation-based optimization is a challenging and practical problem due to its high computational cost. However, in practice, the simulation can be further divided into multiple fidelity levels. A multi-fidelity framework can be then applied to optimize the objective function at the highest level of fidelity, but at a reduced computational cost by fusing with other low-fidelity data. Thus, the multi-fidelity approach aims at reducing the optimization cost by fusing information at different levels of fidelity. The fused information can be incorporated into a traditional optimization framework, such as Bayesian optimization (BO), to improve the efficiency.

Multi-fidelity approach is an effective framework to reduce the computational cost to approximate the objective function, by improving the accuracy of prediction through low- and high-fidelity. In particular, most of the multi-fidelity methods seek to exploit the correlation between low- and high-fidelity models in order to approximate the high-fidelity model more accurately with more low-fidelity data points. The framework is practical for engineering simulation-based applications, because most of them are mesh-based approaches. Some examples include computational fluid dynamics and solid mechanics problems, which are widely used by engineers on a daily basis. Regarding the mesh-based approach, a finer mesh corresponds to a higher level of fidelity, because of smaller discretization error, whereas the coarser mesh corresponds to a lower level of fidelity.

Constrained optimization problem is also an important topic. Digabel and Wild [1] proposed the QRAK taxonomy to classifies constrained optimization problems. In engineering settings, constraints arise from multiple sources, thus both known and unknown constraints are usually observed in the optimization problem. On one hand, constraints are *known* if the feasibility of the input can be determined directly from the input sampling location, without actually running the simulation or the functional evaluator. Such known constraints are often formulated as a set of inequalities, which can be evaluated before sampling. On the other hand, constraints are *unknown* if the feasibility of the input must be evaluated indirectly through running the functional evaluator or the simulation. Some common examples for un-

*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.

known constraints are ill-conditioned problems, singularity in design, mesh problem, that leads to divergent solutions, to name a few. These constraints are implicitly imposed, and cannot be evaluated without evaluating the function or running the simulation.

Gaussian process (GP) is an efficient methodology to model a response surface that approximates the objective function for a single-fidelity. In the traditional BO approach, an acquisition function $a(\boldsymbol{x})$ is constructed based on another utility function, which rewards the BO method if the new sampling location outperforms the rest. The acquisition function is constructed based on the posterior mean and posterior variance of the GP. Because of its flexibility, many extensions have been proposed to solve other optimization problems, based on the traditional BO framework, including constrained and multi-fidelity problems. For constrained problems, constrained BO is a well-studied subject in the context of traditional BO methods. In order to include constraints in the BO framework, typically a penalty scheme is adopted to penalize the infeasible sampling locations that do not satisfy some constraints. For multi-fidelity problems, to generalize to multiple levels of fidelity, one needs to consider the correlation at different levels of fidelity. 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 on multi-fidelity GP. The literature on GP, CoGP, and BO is briefly reviewed in Section 2.

In this work, we proposed a sequential constrained bi-fidelity sBF-BO-2CoGP method, using CoKriging approach to approximate the objective function at the high-level of fidelity. The known constraints are implemented by penalizing the acquisition function directly for infeasible input sampling locations. The unknown constraints are learned adaptively via another CoKriging model, which acts as a probabilistic binary classifier. The unknown constrained acquisition function is also conditioned on this predicted probability mass function, in addition to the penalty scheme for known constraints. The next sampling location is determined by maximizing the constrained acquisition function. Next, the uncertainty reduction scheme, where uncertainty is measured by the integrated mean-square error, is proposed to determine the fidelity level, which the function evaluation is performed. Compared to the maximum mean square error criteria, the integrated mean square error has been shown to be more robust and efficient.

In the rest of this paper, Section 2 provides a brief introduction to the BO method. Section 3 describes the bi-fidelity sBF-BO-2CoGP method proposed in this paper, including the constrained acquisition function, the fidelity selection criteria. Section 4 demonstrates the application of the proposed sBF-BO-2CoGP methodology on 1D and an engineering application in designing flip-chip package. Section 5 discusses and Section 6 concludes the paper.

2 Related works

The bi-fidelity optimization considered in this paper is formulated as

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

subjected to a set of inequality constraints

$$g_j(\boldsymbol{x}) \leq 0, \tag{2}$$

where $j = 1,\ldots,J$ is the number of inequality constraints.

In this section, we briefly review the relevant GP model that is used as a surrogate model for BO, CoKriging method, the acquisition functions, in Section 2.1, 2.2, and 2.3, respectively. Readers are referred to our previous work [3,4,5] and others [6, 7, 8, 9] for rigorous literature reviews on GP and BO methods and its variants.

2.1 Gaussian process

Assume that f is a function of $\boldsymbol{x}$, where $\boldsymbol{x} \in \mathcal{X}$ is a d-dimensional input, and y is the observation. Let the dataset $\mathcal{D} = (\boldsymbol{x}_i, y_i)_{i=1}^N$, where N is the number of observations. A GP regression assumes that $\boldsymbol{f} = f_{1:N}$ is jointly Gaussian, and the observation y is normally distributed given f,

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

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

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

The covariance kernel $\boldsymbol{K}$ is a choice of modeling 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

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

and the posterior variance is

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

where $k(\boldsymbol{x})$ is the covariance vector between the query point $\boldsymbol{x}$ and $\boldsymbol{x}_{1:N}$.

2.2 CoKriging

One of the advantages of CoKriging is that it exploits the correlation between low- and high-fidelity to improve the prediction. We follow the formulation of Karniadakis et al. [10, 11, 12] in formulating the multi-fidelity GP regression.

Let s be the number of fidelity levels, f_s be the highest fidelity model, f_1 be the lowest fidelity model, and assume that

$$f_t(\boldsymbol{x}) = \rho_{t-1} f_{t-1}(\boldsymbol{x}) + \delta_t(\boldsymbol{x}), t = 2, \ldots, s, \tag{7}$$

where $\delta_t(\boldsymbol{x})$ is a Gaussian random field and

$$\mathrm{Cov}[f_t(\boldsymbol{x}),\ f_{t-1}(\boldsymbol{x}')] = 0, \forall \boldsymbol{x} \neq \boldsymbol{x}'. \tag{8}$$

Kennedy and O'Hagan [2] and Le Gratiet and Garnier [13] 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. Karniadakis et al. [10, 11, 12] employed the same method to approximate the highest level of fidelity and extended for noisy evaluations using the same method. Perikaris et al. [14] proposed a generalized framework that can model nonlinear and space-dependent cross-correlations between models of variable fidelity.

In this paper, we only consider two levels of fidelity, and divide the dataset $\mathcal{D}$ into $\mathcal{D}_c$ and $\mathcal{D}_e$, corresponding to cheap and expensive datasets, respectively. The bi-fidelity formulation is adopted from Couckuyt et al [15, 16, 17]. 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 residual between the high- and low-fidelity model $\delta(\boldsymbol{x})$. Since there are only two levels, the notion of $\delta_t(\boldsymbol{x})$ is dropped and becomes $\delta(\boldsymbol{x})$.

The correlation vector $k(\boldsymbol{x})$ and the covariance matrix $K(\boldsymbol{x})$ are then updated [16, 18] as

$$k(\boldsymbol{x}) = (\rho \cdot \sigma_c^2 \cdot k_c(\boldsymbol{x}) \quad \rho^2 \cdot \sigma^2 \cdot k_c(\boldsymbol{x}, \boldsymbol{X})), \tag{9}$$

$$\boldsymbol{K} = \begin{pmatrix} \sigma_c^2 \cdot \boldsymbol{K}_c & \rho \cdot \sigma_c^2 \cdot \boldsymbol{K}_c(\boldsymbol{X}_c, \boldsymbol{X}_e) \\ \rho \cdot \sigma_c^2 \cdot \boldsymbol{K}_c(\boldsymbol{X}_e, \boldsymbol{X}_c) & \rho^2 \cdot \sigma_c^2 \cdot \boldsymbol{K}_c(\boldsymbol{X}_e, \boldsymbol{X}_e) + \sigma_d^2 \cdot \boldsymbol{K}_e(\boldsymbol{X}_e, \boldsymbol{X}_e) \end{pmatrix}, \tag{10}$$

respectively. The predicted distribution of CoKriging is also characterized by a Gaussian distribution, where the posterior mean and posterior variance are still described by Equation 5 and Equation 6.

2.3 Acquisition function

In the traditional BO method, the acquisition function $a(\boldsymbol{x})$ is used to locate the next sampling location by maximizing its acquisition function. 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 location in the dataset.

There are mainly 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, such as GP-PES [19, 20, 21], GP-ES [22], GP-EST [23], GP-EPS [24].

The PI acquisition function [25] is defined as

$$a_{\mathrm{PI}}(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^N, \theta) = \Phi(\gamma(\boldsymbol{x})), \tag{11}$$

where

$$\gamma(\boldsymbol{x}) = \frac{\mu(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^N, \theta) - f(\boldsymbol{x}_{\mathrm{best}})}{\sigma(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^N, \theta)}, \tag{12}$$

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 [26, 27, 28, 29] is defined as

$$a_{\mathrm{EI}}(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^N, \theta) = \sigma(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^N, \theta) \cdot (\gamma(\boldsymbol{x})\Phi(\gamma(\boldsymbol{x})) + \phi(\gamma(\boldsymbol{x})). \tag{13}$$

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 [30, 31, 32] is defined as

$$a_{\mathrm{UCB}}(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^N, \theta) = \mu(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^N, \theta) + \kappa\sigma(\boldsymbol{x}; \{\boldsymbol{x}_i, y_i\}_{i=1}^N, \theta), \tag{14}$$

where κ is a hyper-parameter describing the acquisition exploitation-exploration balance.

3 Methodology

In this section, the sBF-BO-2CoGP method solving the bi-fidelity optimization problem in Section 2 is described.

3.1 Constraints

We adopted the method from our previous work [3, 4, 33] 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 as zero. The penalization scheme is equivalent with multiplying the acquisition function $a(\boldsymbol{x})$ with another indicator function $I(\boldsymbol{x})$, where

$$I(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \forall j (1 \leq j \leq J) : g_j(\boldsymbol{x}) \leq 0, \\ 0, & \text{if } \exists j (1 \leq j \leq J) : g_j(\boldsymbol{x}) > 0. \end{cases} \tag{15}$$

The indicator function can be easily implemented by looping over all the known constraints.

To handle the unknown-constrained problem, an external binary probabilistic classifier is employed to predict the probability of feasibility. Theoretically speaking, the binary classifier for feasibility is free and up to users. Some examples are k-NN [34], AdaBoost [35], RandomForest [36], support vector machine [37] (SVM), least squares support vector machine (LSSVM) [38], GP [39], and convolutional neural network [40]. However, some classifiers tend to outperform others. One notable choice for the binary classifier is the GP classifier, which performs relatively well on dataset. In sBF-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

$$a^*(\boldsymbol{x}) = a(\boldsymbol{x}) \cdot I(\boldsymbol{x}) \cdot Pr(\text{clf}(\boldsymbol{x}) = 1). \tag{16}$$

Maximizing the new acquisition function $a^*(\boldsymbol{x})$ yields the next sampling location of sBF-BO-2CoGP. In practice, we adopt the covariance matrix adaptation evolution strategy (CMA-ES) from Hansen et al. [41, 42] to maximize the new acquisition function $a^*(\boldsymbol{x})$.

3.2 Fidelity selection criteria

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 proposed based on one-step hallucination. The CoKriging surrogate model will briefly consider two scenarios whether the low-fidelity or the high-fidelity function should be evaluated, and calculate the IMSE reduction in two cases. The IMSE reduction ratio is then compared with the computational cost ratio.

The hallucination process is performed by temporarily assuming that the observation at the next sampling location is exactly the same with the CoKriging prediction, and fitting that sampling location into the CoKriging. The new CoKriging model is then said to be hallucinated at the next sampling location point.

Define a_{fidelity} to quantify the benefit to cost ratio of running at the high-fidelity level as

$$a_{\text{fidelity}} := \frac{\text{IMSE}_{\text{h, hallucinated}}}{\text{IMSE}_{\text{l, hallucinated}}} \cdot \frac{C_h}{C_l}, \tag{17}$$

where a_{fidelity} quantifies the value of adding high-fidelity data compared to that of adding low-fidelity data, C_h and C_l are the computational costs at the high- and low-fidelity levels, respectively, and $\text{IMSE}_{(\cdot),\text{hallucinated}}$ denotes the integrated mean-squared error if the sampling point is hallucinated.

In the proposed fidelity selection criteria, the IMSE is calculated as

$$\text{IMSE} = \int_{\mathcal{X}} \sigma^2(\boldsymbol{x}) d\boldsymbol{x}, \tag{18}$$

where the $\sigma^2(\boldsymbol{x})$ field is updated by assuming that $y(\boldsymbol{x}) = \mu(\boldsymbol{x})$, where $\boldsymbol{x}$ is the new sampling point.

If a_{fidelity} ratio is less than 1, then the function evaluator is called at the high-fidelity level, whereas if this ratio is more than 1, then the function is evaluated at the low-fidelity level. The proposed fidelity selection criteria defined in Equation 17 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.

Also, to promote the high-fidelity evaluations, a hard condition is proposed to prevent the imbalance between low- and high-fidelity datasets, 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. In practice, 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 sBF-BO-2CoGP using a simple analytical example in 1D (Section 4.1), and a real-world engineering application in designing flip-chip package (Section 4.2).

4.1 Numerical example

In this section, we consider a simple analytic 1D example, where the low-fidelity function is

$$f_L(x) = 0.5(6x-2)^2 \sin(12x-4) + 10(x-0.5) - 5, \tag{19}$$

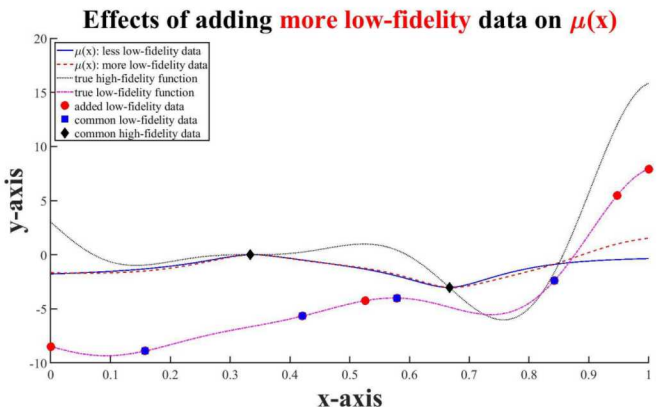
and the high-fidelity function is

$$f_H(x) = (6x-2)^2 \sin(12x-4), \tag{20}$$

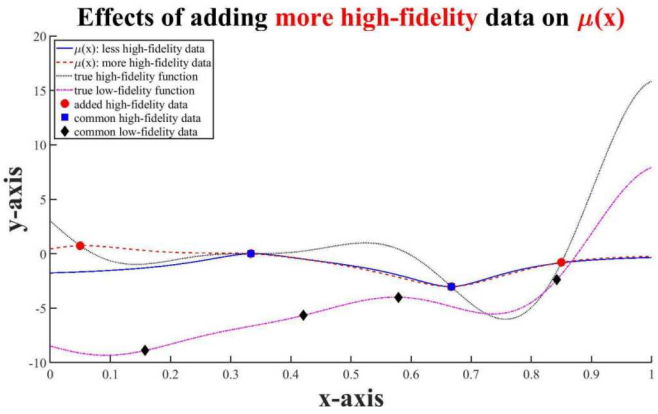
where on $x \in [0,1]$.

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, where the common data points are denoted as blue squares, and the added data points are denotes as red circles and black diamonds.

For the low-fidelity level, Figure 1a and Figure 2a shows 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).

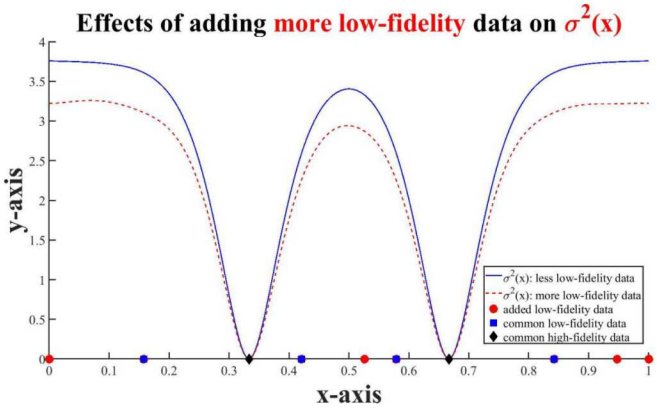


(a) $\mu(\boldsymbol{x})$: more low-fidelity data.

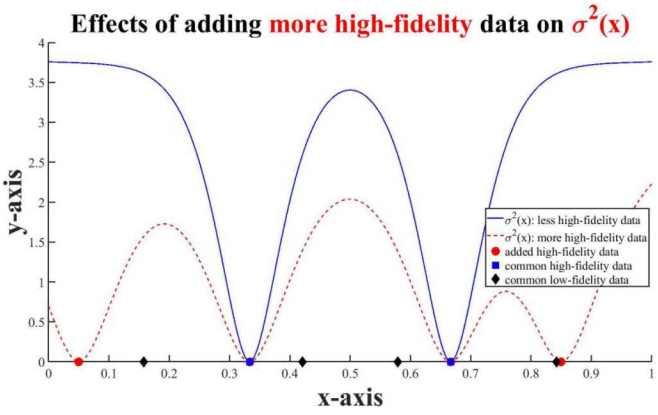


(b) $\mu(\boldsymbol{x})$: more high-fidelity data.

Fig. 1: Effects of adding more low-fidelity and high-fidelity on $\mu(\boldsymbol{x})$.

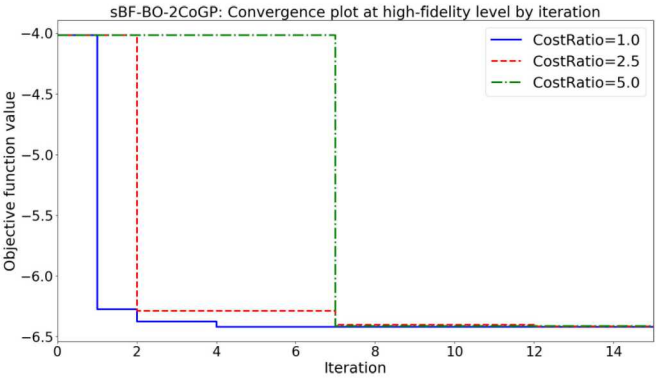


(a) $\sigma^2(\boldsymbol{x})$: more low-fidelity data.

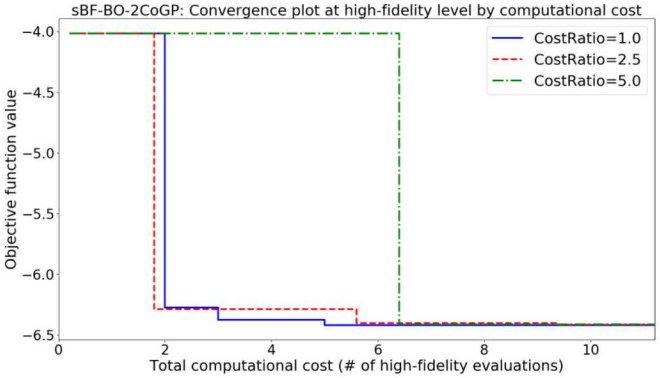


(b) $\sigma^2(\boldsymbol{x})$: more high-fidelity data.

Fig. 2: Effects of adding more low-fidelity and high-fidelity $\sigma^2(\boldsymbol{x})$.



(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 sBF-BO-2CoGP by iteration (Figure 3a) and by total computational cost (Figure 3b).

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 shows 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.

Here, we verify the numerical implementation of the sBF-BO-2CoGP method by considering the minimization problem argmin $f_H(\boldsymbol{x})$ with no constraint and various compu-

tational relative cost ratio between the high- and low-fidelity levels. Figure 3a and Figure 3b shows 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 worthy to note 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 bi-fidelity sBF-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 17.

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 iteration. Figure 3b shows on-par 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 dataset comprised of 4 low-fidelity and 2 high-fidelity. The numerical performances are expected to change with different initial samples, as well as the behavior of high- and low-fidelity models.

4.2 Flip-chip package design

In this section, we demonstrate the design application of a flip-chip package using the proposed sBF-BO-2CoGP. 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 4 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.

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 [43].

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 sBF-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 5 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. 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 sBF-BO-2CoGP is robust against diverging simulations, by its convergent objective despite numerous failed cases.

The optimization results are relatively close with to the design 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. [44], where the original criteria is proposed based on the EI acquisition function as

$$
\begin{aligned}
EI(\boldsymbol{x}, l) = \quad & EI_m(\boldsymbol{x}) && (21)\\
\times \quad & \mathrm{Corr}(f_l^p(\boldsymbol{x}), f_m^p(\boldsymbol{x})) && (22)\\
\times \quad & \left(1 - \frac{\sigma_{\varepsilon,l}}{\sqrt{s_l^2(\boldsymbol{x}) + \sigma_{\varepsilon,l}^2}}\right) && (23)\\
\times \quad & \frac{C_m}{C_l}, && (24)
\end{aligned}
$$

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 17. 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

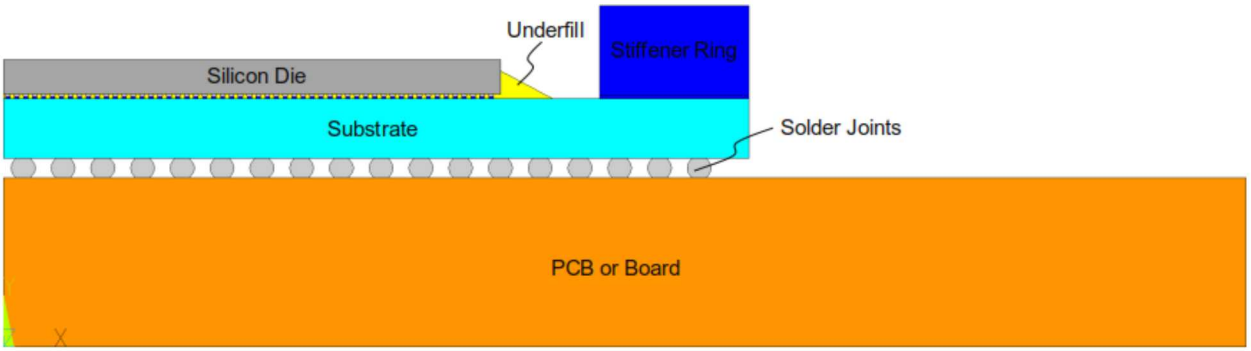


Fig. 4: Finite element model geometry.

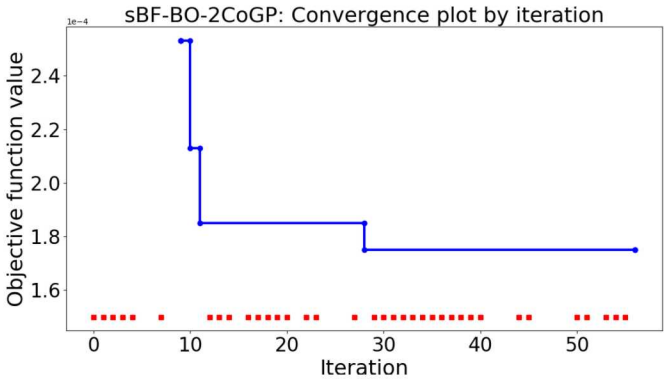


Fig. 5: Convergence plot of flip-chip package design evaluation to minimize the flip-chip warpage.

is left to users as a choice. Previous work by Gauthier et al. [45, 46] and Silvestrini et al. [47] have shown that the performance of IMSE supersedes the performance of maximal MSE. The scheme proposed by Huang et al. [44] in Equation 21 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 [48].

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 is not so many work dedicated to benchmark and quantify 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 16, 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 [39].

While the proposed sequential bi-fidelity sBF-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 bi-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 bi-fidelity BO optimization, called sBF-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 sBF-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 sBF-BO-2CoGP method. However, more benchmark studies are needed to draw a conclusion.

Acknowledgements

A portion of this research was supported by the U.S. Department of Energy, Office of Science, Early Career Research 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., 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.

[4] 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.

[5] Tran, A., Furlan, J. M., Pagalthivarthi, K. V., Visintainer, R. J., Wildey, T., and Wang, Y., 2018. "WearGP: A computationally efficient machine learning framework for local erosive wear predictions via nodal Gaussian processes". *Wear*.

[6] 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*.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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).

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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*.

[25] 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.

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

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

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

[29] 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.

[30] 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*.

[31] 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.

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

[33] Tran, A., Scott, M., Furlan, J. M., Pagalthivarthi, K. V., Visintainer, R. J., and Wildey, T., 2019. "aphBO-2GP-3B: 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*.

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

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

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

[37] 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.

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

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

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

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

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