srMO-BO-3GP: A sequential regularized multi-objective constrained Bayesian optimization for design applications

Anh Tran[*]**, Mike Eldred**
Optimization and Uncertainty Quantification
Sandia National Laboratories
Albuquerque, NM 87123
Email: anhtran, mseldre@sandia.gov

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

Yan Wang
Woodruff School of Mechanical Engineering
Georgia Institute of Technology
Atlanta, GA 30332
Email: yan.wang@me.gatech.edu

Bayesian optimization (BO) is an efficient and flexible global optimization framework that is applicable to a very wide range of engineering applications. To leverage the capability of the classical BO, many extensions, including multi-objective, multi-fidelity, parallelization, latent-variable model, have been proposed to improve the limitation of the classical BO framework. In this work, we propose a novel multi-objective (MO) extension, called srMO-BO-3GP, to solve the MO optimization problems in a sequential setting. Three different Gaussian processes (GPs) are stacked together, where each of the GP is assigned with a different task: the first GP is used to approximate the single-objective function, the second GP is used to learn the unknown constraints, and the third GP is used to learn the uncertain Pareto frontier. At each iteration, a MO augmented Tchebycheff function converting MO to single-objective is adopted and extended with a regularized ridge term, where the regularization is introduced to smoothen the single-objective function. Finally, we couple the third GP along with the classical BO framework to promote the richness and diversity of the Pareto frontier by the exploitation and exploration acquisition function. The proposed framework is demonstrated using several numerical benchmark functions, as well as a thermomechanical finite element model for flip-chip package design optimization.

1 Introduction

Optimization, in general, is a common problem that appears in many contexts, including engineering, machine learning, physics, finance, and mathematics. Numerous methods, such as evolutionary algorithms, and global optimization method, have been proposed to achieve the optimality with improved efficiency and effectiveness. Bayesian optimization (BO) method, also known as efficient global optimization method, is a derivative-free optimization technique that has been used extensively in engineering domains, particularly design optimization. Traditionally, the classical BO method only consider single-objective function, whereas in practice, multiple objectives are interested. Objectives are often found to conflict each other, where a trade-off between objectives is desired to achieve the optimality. Thus, it is important to derive the set of optimal solutions in the Pareto sense, where the limits of objectives are fully exposed for decision makers.

Multi-objective Bayesian optimization (MOBO) is an extension of the classical BO method for multi-objective optimization. Here, we limit ourselves in reviewing MO extension to most recent developments. Jeong and Obayashi [1] proposed an EI-based criteria and employed genetic algorithm (GA) to search for potential sampling points on the Pareto frontier. Knowles [2] introduced the ParEGO framework with the augmented Tchebycheff function to scalarize the multi-objective functions, which was later extended by Davins-Valldaura et al. [3] by including the probability of Pareto frontier with another GP. Zhang et al. [4] proposed the MOEA/D framework with the classical

[*]Corresponding author: anhtran@sandia.gov

Tchebycheff scalarization function and decomposed a MO into a number of single-objective optimization subproblems. Li et al. [5,6] improve the MO GA methods with the K-MOGA framework, where the posterior mean and posterior variance is considered to assist GA methods. Feliot et al. [7] proposed a multi-objective EI acquisition function considering constraints based on the posterior mean and posterior variance. Qian and Yu [8] extended the REMBO framework from Wang et al. [9] for solving MO problems by using random embeddings to significantly reduce the effective dimensionality of the problem considered. Gupta et al. [10] proposed a batch MOBO framework with a weighted linear average objective and demonstrated by optimizing heat treatment process of an Al-Sc alloy. Shu et al. [11] developed a composite acquisition function for MOBO to efficiently sample with simultaneous improvement of convergence and diversity in constructing the Pareto frontier. Hernández-Lobato et al. [12,13] proposed a predictive entropy search acquisition function for MO problems considering constraints. Abdolshah et al. [14] proposed a MO BO framework where various computational efforts are taken into account when exploring the input domain. Gaudrie et al. [15] proposed a sequential and batch extension for MO BO method by maximizing the expected hyper-volume improvement. Palar et al. [16] compared the impact of different covariance functions in MO BO and concluded that Matérn-3/2 is the most robust kernel for design optimization applications. Further interested readers are referred to other survey and literature review [17,18] in MO BO methodologies and applications.

In this paper, we propose to stack three GPs to solve the MO constrained optimization problems, where s is the number of objectives considered. Furthermore, we explore the possibility of introducing regularization in the Tchebycheff scalarization method. The approach is unique in three different perspectives. First, we employ GP as an uncertain machine learning technique to probabilistically search for the Pareto frontier, simply by stacking this classifier over the objective GP. The acquisition function in the classical BO method is employed to balance the exploitation and the exploration of this Pareto GP, which enhances its richness and diversity in the Pareto frontier. Second, we combine the multi-objective functions to a single-objective function with a randomized weight vector and regularize this single-objective function by a ridge regularization to smoothen the single-objective function. The philosophy of our approach is intrinsically similar to that of classical BO methods. Compared to other approach in literature, the heuristic proposed in this research is greatly simplified, yet its efficiency is comparable to other methods. Our approach is more general in the sense that it does not restrict to a particular form of the acquisition function, and leaves the choice of acquisition functions up to the users. The idea is extended based on the work of Davins-Valldaura et al. [3]; however, our work differs from their work in treating the Pareto GP, as well as the acquisition function. In our approach, we opt to include both exploitation and exploration of this Pareto GP by considering uncertainty, whereas in their approach, only the probability of Pareto frontier (i.e. exploitation) is considered.

Another novelty in our approach is the ridge regularized term in the acquisition to smooth the acquisition function.

Two significant advantages of the proposed srMO-BO-3GP algorithms are highlighted as follows. First, the regularization terms in the acquisition can significantly mitigate the effects of noise on observations. Second, our approach is not limited by the number of objective functions, as in other approaches. For example, if hypervolume-based approach is considered, the number of objectives would have a scalability effect on the computation of the hypervolume. Our proposed approach avoids limitation by converting the identification of Pareto frontier into an uncertain classification problems in machine learning context, with a flavor of uncertainty quantification. Thus, as the Pareto GP classifier gains more accuracy, it converges to the true Pareto frontier through the mean of classical Bayesian optimization approach.

For the remaining of this paper, Section 2 describes the proposed srMO-BO-3GP framework. Section 3 provides the numerical results for several numerical analytical functions, as well as an engineering thermomechanical finite element model (FEM) using the proposed approach. Section 4 discusses and Section 5 concludes the paper.

2 Methodology

We follow the formulation of Lin et al. [19] in defining the MO BO problem. For the sake of clarity, we denote $\boldsymbol{x} = \{x_i\}_{i=1}^d \in \mathcal{X} \subseteq \mathbb{R}^d$ as the input of d continuous variables in d-dimensional space, whereas $\boldsymbol{y} = \{y_j\}_{j=1}^s$ as s outputs. Traditionally, BO solves the single-objective optimization problem

$$y = \underset{\boldsymbol{x} \in \mathcal{X}}{\arg\max}\, c(\boldsymbol{x}), \tag{1}$$

subject to the constraints $c(\boldsymbol{x}) \leq 0$. In this paper, we consider the scenario of MO optimization problem,

$$Y = \underset{\boldsymbol{x} \in \mathcal{X}}{\arg\max}\,(f_1(\boldsymbol{x}), \ldots, f_s(\boldsymbol{x})), \tag{2}$$

subject to the constraints $c(\boldsymbol{x}) \leq 0$.

2.1 Gaussian process

Gaussian process regression is an efficient and flexible framework to approximate a response surface for a single-fidelity, single-objective function. We briefly summarize the GP theoretical formulation for the sake of completeness [20]. Let $\mathcal{D} = (\boldsymbol{x}_i, y_i)_{i=1:n}$ denote the dataset of n observations with output y and d-dimensional input $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$. A GP is a nonparametric model characterized by its prior mean function $\mu_0(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ and a covariance function $k(\boldsymbol{x}, \boldsymbol{x}') : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Assuming that the observations $\boldsymbol{f} = f_{1:n}$ are jointly Gaussian, and the observation y is normally distributed given $\boldsymbol{f}$, i.e.

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

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

where $m_i := \mu(x_i)$ and $K_{i,j} := k(x_i, x_j)$.

The classical GP regression formulation assumes a stationary covariance matrix and only considers the weighted distance $r^2(x, x') = (x - x')^T \Lambda (x - x')$, where Λ is a diagonal matrix of d squared length scales θ_i^2. Matérn kernels offer a broad class for stationary kernels, controlled by a smoothness parameter $\nu > 0$ (cf. Section 4.2, [21]), including the square-exponential ($\nu \to \infty$) and exponential ($\nu = 1/2$) kernels widely used in the literature. The $\nu = 3/2$ Matérn kernel $k(x, x') = \theta_0^2 \exp(-\sqrt{3}r)(1 + \sqrt{3}r)$ is used in this work. At a known sampling point $x \in \mathcal{X}$, the posterior mean $\mu(x)$ is calculated by

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

and the posterior variance $\sigma^2(x)$ is given by

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

where $k(x)$ is a vector of covariance $k(x)_i = k(x, x_i)$, $\sigma^2 = \frac{1}{n}(y - \mu_0(x))^T K^{-1}(y - \mu_0(x))$ is the intrinsic variance. To obtain the hyper-parameter $\theta = (\theta_i)_{i=1:d}$, we maximize the log marginal likelihood, which is computed as

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

Here, K^θ is emphasized to be strongly dependent on θ.

2.2 Multi-objective function

We adopt the definition of Pareto-dominant from Rojas-Gonzalez et al. [18]

Definition-1: x_1 is said to dominate x_2, denoted as $x_1 \preceq x_2$, if and only if $\forall 1 \leq j \leq s$, such that $y_j(x_1) \leq y_j(x_2)$, and $\exists 1 \leq j \leq s$, such that $y_j(x_1) < y_j(x_2)$.

Definition-2: x_1 is said to strictly dominate x_2, denoted as $x_1 \prec x_2$, if and only if $\forall 1 \leq j \leq s$, such that $y_j(x_1) < y_j(x_2)$.

MO optimization problems are typically solved by converting a MO problem to a single-objective problem, where the single-objective function is scalarized as a weighted sum of s multiple objectives [18], such as,

1. weighted Tchebycheff scalarization function $y = \max_{1 \leq i \leq s} w_i(y_i(x) - z_i^*)$,

2. the weighted sum scalarization function: $y = \sum_{i=1}^{s} w_i y_i(x)$,

3. and the augmented Tchebycheff scalarization function $y = \max_{1 \leq i \leq s} w_i(y_i(x) - z_i^*) + \rho \sum_{i=1}^{s} w_i y_i(x)$,

where z_i^* denotes the ideal value for the i-th objective, the weights $0 \leq w_i \leq 1$, $\sum_{i=1}^{m} w_i = 1$, ρ is a small positive constant ($\rho = 0.05$ in [2,3]).

We closely follow the idea of Davins-Valldaura et al. [3] in converting multiple objectives to a single-objective function. We propose another objective function based on the augmented Tchebycheff scalarization function, with an addition of the regularized term for differentiability, which could be interpreted in terms of ridge regression,

$$y = \max_{1 \leq i \leq s} w_j y_j(x) + \rho \sum_{i=1}^{s} w_j y_j(x) + \lambda \|x\|_2, \tag{8}$$

where λ is an appropriate constant for regularization.

2.3 Constraints

We consider the known constraints and hidden constraints, where the known constraints are known before the functional evaluation, whereas the hidden constraints must be learn indirectly through the functional evaluation.

2.3.1 Known constraints

Known constraints can be represented as a set of inequalities $c(x) \leq 0$, where c is a relatively cheap function to evaluate, compared to the real objective function f. The known constraints can be easily implemented by directly penalizing the acquisition by setting it to zero if known constraints are violated, while leaving unviolated sampling points as is.

In practice, the penalty scheme is implemented by multiplying the acquisition with another indicator function $I(x)$,

$$I(x) = \begin{cases} 1, & \text{if } \forall k : c_k(x) \leq 0, \\ 0, & \text{if } \exists k : c_k(x) > 0, \end{cases} \tag{9}$$

where c_k denotes the k-th constraint in the set of known constraints.

2.3.2 Unknown/hidden constraints

We adopt our previous strategy by employing a probabilistic binary classifier to learn the hidden constraints. As a result, feasible and infeasible regions are separated in the input domain $\mathcal{X}$. These two regions are mutually exclusive, i.e. disjoint, because a sampling point cannot be both feasible and infeasible at the same time. The labels of feasible/infeasible for sampling points

are fixed, in the sense that as the optimization process advances, the labels do not change.

Denote the feasibility dataset as $\{\boldsymbol{x}_i, c_i\}_{i=1}^n$, where n is the number of data points. We assign $c_i = 1$ if $\boldsymbol{x}_i$ is feasible, and $c_i = 0$ if $\boldsymbol{x}_i$ is infeasible. At an unknown $\boldsymbol{x}$, the feasibility classifier provides a probability mass function, with $\Pr(\boldsymbol{x}|c(\boldsymbol{x}) = 1)$ as the predicted probability of passing the hidden constraints, and $\Pr(\boldsymbol{x}|c(\boldsymbol{x}) = 0)$ as the predicted probability of failing the hidden constraints. It is noteworthy that their sum adds up $\Pr(\boldsymbol{x}|c(\boldsymbol{x}) = 0) + \Pr(\boldsymbol{x}|c(\boldsymbol{x}) = 1)$, as they are mutually exclusive and there are only two possibilities. The probability of passing the unknown constraints will be used to condition on the acquisition, resulting in the multiplication of $\Pr(\boldsymbol{x}|c(\boldsymbol{x}) = 1)$ in the conditioned acquisition.

Even though there are many available probabilistic binary classifier in the context of machine learning, for example, k-NN [22], AdaBoost [23], RandomForest [24], support vector machine [25] (SVM), least squares support vector machine (LSSVM) [26], and convolutional neural network [27], in this work, we restrict our methodology to GP as a binary probabilistic classifier. Labeling feasibility as described above, the posterior mean of feasibility GP can be used to predict the probability of passing unknown constraints, i.e. $\mu_{\text{feasible}}(\boldsymbol{x}) = \Pr(\boldsymbol{x}|c(\boldsymbol{x}) = 1)$.

2.4 Pareto frontier with an uncertain GP classifier

At each iteration, we construct the current Pareto frontier, which is subjected to change as the optimization process advances. If a sampling point is currently Pareto-dominant, the point is labeled as 1, and if the sampling point is not Pareto-dominant (i.e. it is dominated by another point in the dataset), the sampling point is labeled as 0. The classification process is thus uncertain, in the sense that the labels change from one iteration to another. This is to contrast with the constraint classifier $c(\boldsymbol{x})$, where the labels are fixed and do not change, as the optimization process advances. The uncertainty in the Pareto frontier classification gradually decreases, as the number of sampling data points increases.

We explore the possibility of using GP as an uncertain Pareto frontier classifier. Surprisingly, GP is one of a few well-established machine learning techniques that considers uncertainty in prediction through its posterior variance function. Because of this particular reason, GP is employed as an uncertain classifier to construct the Pareto frontier, where the uncertainty is quantified by the GP posterior variance function (Eq 6).

The main idea is to force the Pareto GP classifier to balance its learning by exploitation and exploration, especially when the Pareto frontier prediction is not accurate by focusing on the unknown region at the beginning of the optimization process. As the optimization advances, if the Pareto frontier can be classified with high accuracy, the BO framework should be exploited to promote richness and explored to promote the diversity in the MO optimization settings. This is consistent with the philosophy of the classical BO approach, which strikes for the balance of exploration and exploitation. Furthermore, the richness and diversity, which are the two keys measure of MO optimization problems [18], are promoted by employing common acquisition functions on the Pareto frontier GP. As a result, the acquisition of Pareto GP classifier is included as another in the main acquisition function for the classical BO method (Equation 10).

It is noteworthy to point out that the Pareto classification problem is also mutually exclusive, in the sense that a sampling point cannot be both Pareto-dominant and Pareto-non-dominant at the same time. For an unknown location $\boldsymbol{x}$, the Pareto GP classifier provides both the probability of being Pareto-dominant $\mu_{\text{Pareto}}(\boldsymbol{x})$, which is bounded between 0 and 1, as well as the uncertainty associated with the probability as $\sigma^2_{\text{Pareto}}(\boldsymbol{x})$.

2.5 Acquisition function

We propose an extended criteria in learning the Pareto frontier considering uncertainty, as opposed to directly couple the Pareto-dominant probability into the EI acquisition function in Davins-Valldaura et al [3]. The composite acquisition function is defined as

$$a(\boldsymbol{x}) = \underbrace{a_{\text{obj}}(\boldsymbol{x})}_{\text{objective GP}} \cdot \underbrace{a_{\text{Pareto}}(\boldsymbol{x})}_{\text{uncertain Pareto}} \cdot \underbrace{\Pr(\boldsymbol{x}|c(\boldsymbol{x}) = 1)}_{\text{unknown constraints}} \cdot \underbrace{I(\boldsymbol{x})}_{\text{known constraints}}.$$
(10)

Equation 10 is explained as follows. The first term, $a_{\text{obj}}(\boldsymbol{x}; \mu_{\text{obj}}(\boldsymbol{x}), \sigma^2_{\text{obj}}(\boldsymbol{x}))$, is the regularized augmented Tchebycheff single-objective function in Equation 8. At each time step, a random weight vector $\boldsymbol{w} = (w_1, \ldots, w_s)$ is sampled to combine multiple objectives $\{y_j\}_{j=1}^s$ to a single-objective function y. A GP model is fitted using the dataset of n data points $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$. An acquisition function $a_{\text{obj}}(\boldsymbol{x})$ is formed, as a function of posterior mean $\mu_{\text{obj}}(\boldsymbol{x})$ and posterior variance $\sigma^2_{\text{obj}}(\boldsymbol{x})$. The second term, $a_{\text{Pareto}}(\boldsymbol{x}; \mu_{\text{Pareto}}(\boldsymbol{x}), \sigma^2_{\text{Pareto}}(\boldsymbol{x}))$, is the acquisition function based on the Pareto frontier GP classifier. The third term, $\Pr(\boldsymbol{x}|c(\boldsymbol{x}) = 1) = \mu_{\text{feasible}}(\boldsymbol{x})$, is the probability of passing unknown constraints, provided by the second GP.

The next sampling point $\boldsymbol{x}^*$ is obtained by maximizing the acquisition function described in Equation 10, i.e.

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in X}{\arg\max} \, a(\boldsymbol{x}) \qquad (11)$$

2.6 Summary

The srMO-BO-3GP framework is summarized in Algorithm 1. CMA-ES [28, 29] is used as the sub-optimizer to obtain the next sampling point in Equation 11, where the default settings are retained. An interface between the srMO-BO-3GP optimizer and the engineering applications is constructed by MATLAB, Python, and Shell scripts. Also, since the maximization is set as a default setting, the Tchebycheff function is algebraically manipulated to conform with the default setting.

3 Numerical results

In this section, we investigate the effectiveness and the efficiency of the proposed srMO-BO-3GP algorithm by ZDT and

Algorithm 1 srMO-BO-3GP algorithm.

Input: dataset $\mathcal{D}_n$ consisting of input, observation, feasibility $(\boldsymbol{x},\boldsymbol{y},c)_{i=1}^n$
Input: multi-objective $(\boldsymbol{x}_i,\boldsymbol{y}_i)_{i=1}^n$, constraint GP $(\boldsymbol{x},c_i)_{i=1}^n$,

  1: **for** $n = 1, 2, \ldots,$ **do**
  2:　　randomize a weight vector $\boldsymbol{w}$
  3:　　combine $\{y_j\}_{j=1}^s$ to y (Eq 8)　　　　　　　　　　　　　　　　　　▷ multi- to single-objective
  4:　　construct single-objective GP　　　　　　　　　　　　　　　　　　　▷ GP #1: $\mu_{\text{obj}}(\boldsymbol{x}), \sigma_{\text{obj}}^2(\boldsymbol{x})$
  5:　　construct Pareto front　　　　　　　　　　　　　　　　　　　　　　▷ GP #2: $\mu_{\text{Pareto}}(\boldsymbol{x}), \sigma_{\text{Pareto}}^2(\boldsymbol{x})$
  6:　　　find current Pareto front
  7:　　　construct Pareto classifier GP
  8:　　construct constraints classifier GP　　　　　　　　　　　　　　　　▷ GP #3: $\mu_{\text{feasible}}(\boldsymbol{x}), \sigma_{\text{feasible}}^2(\boldsymbol{x})$
  9:　　locate the next sampling point $\boldsymbol{x}_{n+1}$ (Eq. 10)
 10:　　query for $\boldsymbol{y}_{n+1} = \{y_j\}_{j=1}^s$, feasibility c_{n+1}
 11:　　augment dataset $\mathcal{D}_{n+1} = \{\mathcal{D}_n, (\boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1}, c_{n+1})\}$
 12: **end for**

LDTZ benchmarking functions, where the geometry of Pareto frontiers are well documented by Jin et al [30]. The hypervolume metric, which is a strictly monotonic measure, is computed using Walking Fish Group algorithms used [31, 32]. The regularization λ parameter is set to 0.05, even though its effect requires a further benchmark study. We compare the numerical performance between variants of srMO-BO-3GP, particularly for

- variations of acquisition function for the objective GP: PI, EI, and UCB;
- variations of acquisition function for the Pareto GP: PI, EI, and UCB;
- regularized vs. non-regularized Tchebycheff objective function.

This results in 18 different variants of the srMO-BO-3GP. The name convention for these variants is Reg/NoReg-{PI,EI,UCB}-{PI,EI,UCB}, which corresponds the option of regularization, the objective GP, and the Pareto GP.

3.1 ZDT1

The ZDT1 benchmark function for a 12-dimensional input $\boldsymbol{x} \in [0,1]^{12}$ is described as

$$f_1 = x_1, \tag{12}$$
$$f_2 = gh, \tag{13}$$

where

$$g = 1 + \frac{9}{11} \sum_{p=2}^{12} x_p, \tag{14}$$

$$h = 1 - \sqrt{\frac{f_1}{g}}. \tag{15}$$

It can be proved analytically that since $1 \le g \le 10$, the Pareto frontier is obtained when $g = 1$. The Pareto frontier is a collection of $\{f_1, f_2\}$, where $f_2 = 1 - \sqrt{f_1}$.

(a) ZDT1 Pareto frontier comparison.

(b) ZDT1 Pareto frontier comparison: Pareto frontier details at the [0, 1] × [0, 1] window.
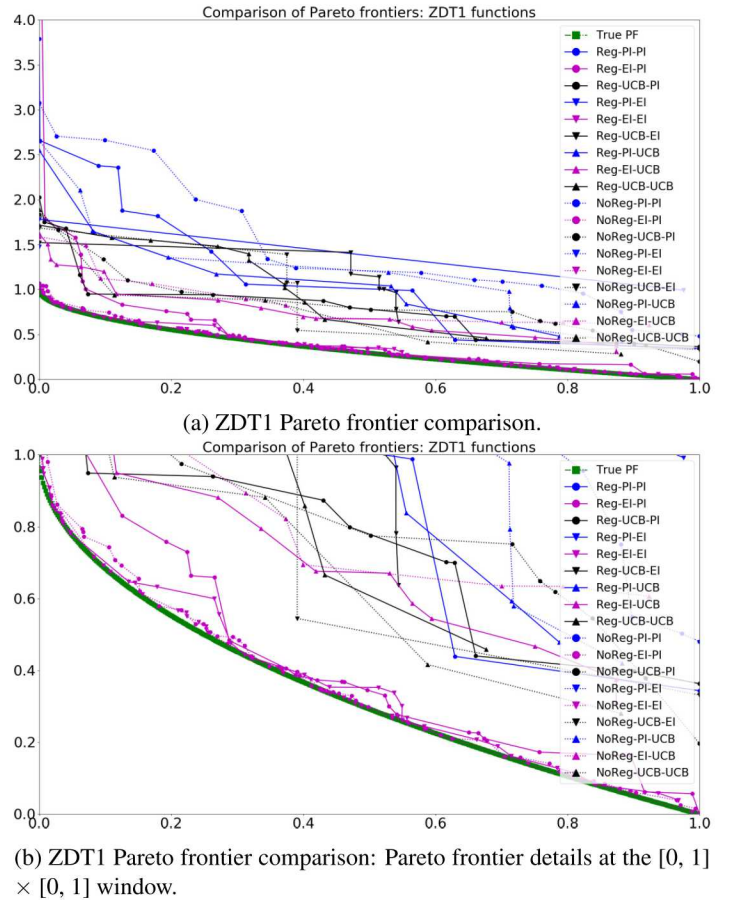
Fig. 1: Comparison of srMO-BO-3GP variants on ZDT1 benchmark function.

Figure 1a shows a comparison between 18 variants of the srMO-BO-3GP algorithm, where Figure 1b shows the magnified view in the window of $[0,1] \times [0,1]$. Readers are referred to the color version online. The true Pareto front is plotted with the dash-dotted green square. The regularized acquisition functions are plotted as solid lines, whereas the non-regularized acquisi-

tion functions are plotted as dash-dotted lines. The objective GPs with PI acquisition function are plotted as blue, whereas those with EI acquisition function are plotted as magenta, and those with PI acquisition function are plotted as black. The Pareto GPs with PI acquisition function are plotted as circles, whereas those with EI acquisition function are plotted as downward triangles, and those with PI acquisition functions are plotted as upward triangles.

Overall, the regularized objective function ($\lambda = 0.05$) slightly outperforms the non-regularized objective function ($\lambda = 0$). The objective GPs with the EI acquisition function seems to outperform the objective GPs with UCB and PI acquisition function, with PI acquisition function being the worst. However, the Pareto GPs with the EI and PI acquisition functions are more accurate in classifying the true Pareto frontier, compared to those with the UCB acquisition functions. The best performers among these variants are NoReg-EI-PI and NoReg-EI-EI.

3.2 DTLZ1

The DTLZ1 function for a 12-dimensional input $\boldsymbol{x} \in [0, 1]^{12}$ is described as

$$f_1 = 0.5(1 + g)x_1, \tag{16}$$
$$f_2 = 0.5(1 + g)(1 - x_1), \tag{17}$$

where $g = 10011 + \sum_{p=2}^{12} \langle (x_p - 0.5)^2 - \cos[20\pi(x_p - 0.5)] \rangle$

Figure 2a shows a comparison between 18 variants of the srMO-BO-3GP algorithm, where Figure 2b shows the magnified view in the window of $[2500, 2525] \times [2475, 2500]$. Readers are referred to the color version online. The true Pareto front is plotted with the dash-dotted green square. The regularized acquisition functions are plotted as solid lines, whereas the non-regularized acquisition functions are plotted as dash-dotted lines. The objective GPs with PI acquisition function are plotted as blue, whereas those with EI acquisition function are plotted as magenta, and those with PI acquisition function are plotted as black. The Pareto GPs with PI acquisition function are plotted as circles, whereas those with PI acquisition functions are plotted as upward triangles. All variants of the sMO-BO-3GP converges on the true Pareto frontier with little variations.

3.3 Engineering applications

We demonstrate the applicability of our proposed framework to a thermomechanical FEM model for flip-chip package design, where five objectives are considered. Figure 3 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 1 show the design variables, the physical

(a) DTLZ1 Pareto frontier comparison.

(b) DTLZ1 Pareto frontier comparison: Pareto frontier details at the $[2500, 2525] \times [2475, 2500]$ window.
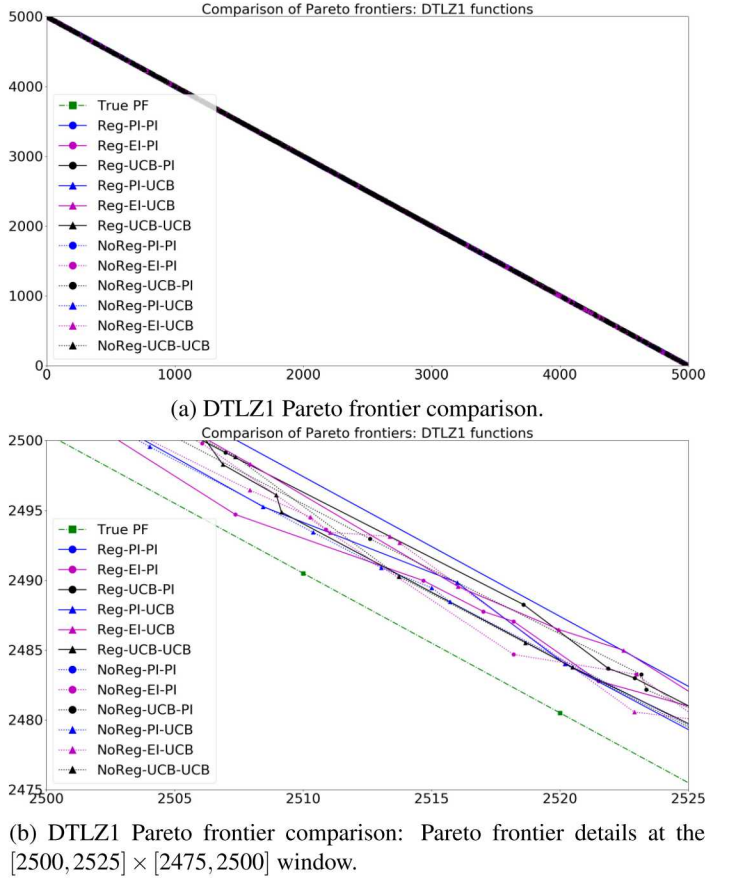
Fig. 2: Comparison of srMO-BO-3GP variants on DTLZ1 benchmark function.

meaning of the design variables, as well as their lower and upper bounds in this case study.

Table 1: 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	$12.0 \cdot 10^{-6}$	$16.7 \cdot 10^{-6}$	$16.7 \cdot 10^{-6}$

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 five objectives are listed as follows,
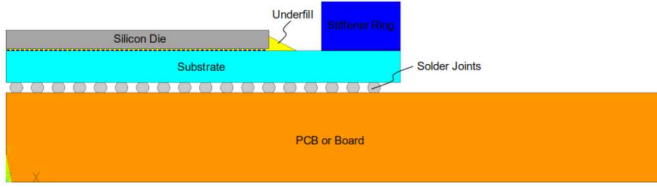
Fig. 3: Finite element model geometry.

Objective-1: warpage at 20°C,
Objective-2: warpage at 250°C,
Objective-3: damage metric for BGA lifetime prediction,
Objective-4: damage metric for C4 interconnection lifetime prediction,
Objective-5: first principal stress at C4 corner, where cracking and delamination frequently occurs,

where the goal is to minimize all these objectives.

Figure 4a shows the scatter plot matrix between the five objectives and their joint densities, whereas Figure 4b presents the Pearson correlation, which is bounded between (-1) and (+1). 851 simulations are performed, in which only 256 simulations are feasible. Out of 256 feasible simulations, 84 of those represent the current Pareto frontier of this numerical study. The objective 1 is very positively correlated with the objective 2, as both of them corresponds to the warpages at different temperatures. This implies that minimizing the object 1 would also minimize the objective 2, thus no trade-off is found. The same argument applies for objectives 3 and 5, where positive correlation is found. The objective 4 is poorly correlated with other objectives. Trade-offs are found between members of the group of objectives 1 and 2 and those of the group of objective 3 and 5. Overall, it is challenging to visualize the Pareto front on high-dimensional space for a practical. However, we have demonstrated that our methodology is applicable to problems with many objective functions. In practice, the number of objectives are often reduced to minimum, before the optimization study is conducted.

4   Discussion

BO is a powerful and flexible framework which allows for many useful extension. For example, the local GP approach [33, 34] could be used to improve the scalability of GP. An extended version of local GP has also been developed [35] to solve the mixed-integer optimization problems, where the number of discrete/categorical variables is relatively small. Accelerated BO methods by exploiting computational resource on high-performance computing platform have been proposed [36, 37] to reduce the amount of physical waiting time. Multi-fidelity BO approaches [38, 39] have been developed to couple information between different levels of fidelity by exploiting the correlation between low- and high-fidelity models to reduce the computational efforts. Its success has been demonstrated, at least in the field of bioengineering [40] and computational fluid dynam-
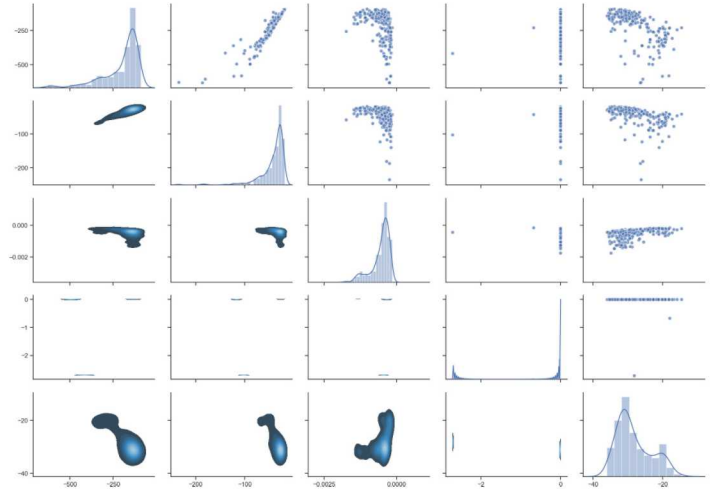
(a) Correlation between objectives and their joint densities.
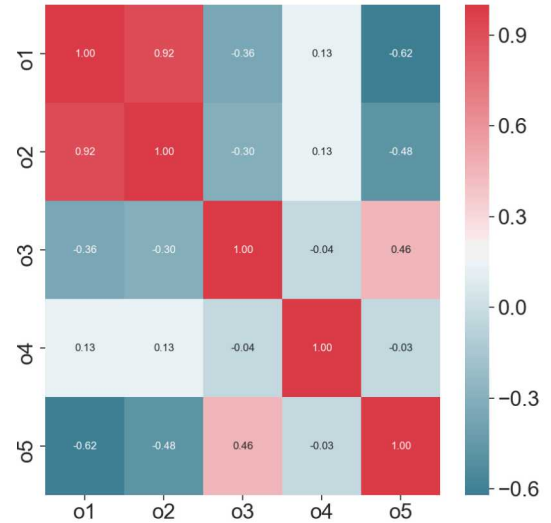
(b) Correlation heatmap between objectives.

Fig. 4: Correlation between five objectives in the FEM example and their joint densities.

ics [41, 42].

The uncertain GP Pareto frontier was tested on both multi-objective and single-objective function. Surprisingly, the results for uncertain multi-objective Pareto GP classifier are not as good as those for uncertain single-objective Pareto GP classifier. Particularly, the uncertain single-objective Pareto GP classifier tends to diversify results more, using the exploration feature that is already readily available in the acquisition function. Additionally, the exploitation feature will push the uncertain Pareto frontier to the true Pareto frontier. As the optimization advances, the uncertain Pareto frontier provided by the Pareto GP will converge to the true Pareto GP.

Two significant advantages of the proposed srMO-BO-3GP algorithms are highlighted as follows. First, the regularization terms in the acquisition can significantly mitigate the effects of noise on observations. Second, our approach is not limited by the number of objective functions, as in other approaches.

For example, if hypervolume-based approach is considered, the number of objectives would have a scalability effect on the computation of the hypervolume. Our proposed approach avoids limitation by converting the identification of Pareto frontier into an uncertain classification problems in machine learning context, with a flavor of uncertainty quantification. Thus, as the Pareto GP classifier gains more accuracy, it converges to the true Pareto frontier through the mean of classical Bayesian optimization approach. The second advantage is particularly a huge improvement, compared to other conventional approaches, such as first/second-order reliability method (FORM/SORM) [43], because these approaches parameterize and fit the Pareto frontier with a polynomial. This leads to a significant drawback when the Pareto frontier is discontinuous. Our proposed approach srMO-BO-3GP does not suffer from this limitation, and does not limit the number of objective functions. By converging the multi-objective optimization problem into an uncertain classification and solving through GP classifier, the discontinuous problem is completely resolved by classification techniques in the context of machine learning.

5   Conclusion

In this paper, we propose a MO BO framework, called srMO-BO-3GP in a sequential setting with applications to engineering-based simulations. In this framework, three distinct GPs are coupled together. The first GP is used to approximate the single-objective function, which is converted from MO function using a regularized augmented Tchebycheff with a ridge regularization term. The second GP is used to learn the unknown constraints, which are evaluated simultaneously with the objective function, e.g. the output does not exist because of various numerical reasons. The third GP is used as an uncertain binary classifier to learn the Pareto frontier, where its own acquisition function is embedded in the main acquisition function. The srMO-BO-3GP framework is demonstrated using two numerical benchmarking functions, as well as a thermomechanical FEM.

References

[1] Jeong, S., and Obayashi, S., 2005. "Efficient global optimization (EGO) for multi-objective problem and data mining". In 2005 IEEE congress on evolutionary computation, Vol. 3, IEEE, pp. 2138–2145.

[2] Knowles, J., 2006. "ParEGO: a hybrid algorithm with online landscape approximation for expensive multiobjective optimization problems". *IEEE Transactions on Evolutionary Computation,* **10**(1), pp. 50–66.

[3] Davins-Valldaura, J., Moussaoui, S., Pita-Gil, G., and Plestan, F., 2017. "ParEGO extensions for multi-objective optimization of expensive evaluation functions". *Journal of Global Optimization,* **67**(1-2), pp. 79–96.

[4] Zhang, Q., Liu, W., Tsang, E., and Virginas, B., 2009. "Expensive multiobjective optimization by MOEA/D with Gaussian process model". *IEEE Transactions on Evolutionary Computation,* **14**(3), pp. 456–474.

[5] Li, M., Li, G., and Azarm, S., 2008. "A kriging metamodel assisted multi-objective genetic algorithm for design optimization". *Journal of Mechanical Design,* **130**(3), p. 031401.

[6] Li, G., Li, M., Azarm, S., Al Hashimi, S., Al Ameri, T., and Al Qasas, N., 2009. "Improving multi-objective genetic algorithms with adaptive design of experiments and online metamodeling". *Structural and Multidisciplinary Optimization,* **37**(5), pp. 447–461.

[7] Feliot, P., Bect, J., and Vazquez, E., 2017. "A Bayesian approach to constrained single-and multi-objective optimization". *Journal of Global Optimization,* **67**(1-2), pp. 97–133.

[8] Qian, H., and Yu, Y., 2017. "Solving high-dimensional multi-objective optimization problems with low effective dimensions". In Thirty-First AAAI Conference on Artificial Intelligence.

[9] Wang, Z., Zoghi, M., Hutter, F., Matheson, D., Freitas, N., et al., 2013. "Bayesian optimization in high dimensions via random embeddings". AAAI Press/International Joint Conferences on Artificial Intelligence.

[10] Gupta, S., Shilton, A., Rana, S., and Venkatesh, S., 2018. "Exploiting strategy-space diversity for batch Bayesian optimization". In International Conference on Artificial Intelligence and Statistics, pp. 538–547.

[11] Shu, L., Jiang, P., Shao, X., and Wang, Y., 2020. "A new multi-objective bayesian optimization formulation with the acquisition function for convergence and diversity". *Journal of Mechanical Design,* **127**(6), pp. 1–38.

[12] 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.

[13] Garrido-Merchán, E. C., and Hernández-Lobato, D., 2019. "Predictive entropy search for multi-objective bayesian optimization with constraints". *Neurocomputing,* **361**, pp. 50–68.

[14] Abdolshah, M., Shilton, A., Rana, S., Gupta, S., and

Venkatesh, S., 2019. "Cost-aware multi-objective Bayesian optimisation". *arXiv preprint arXiv:1909.03600.*

[15] Gaudrie, D., Le Riche, R., Picheny, V., Enaux, B., and Herbert, V., 2019. "Targeting solutions in Bayesian multi-objective optimization: sequential and batch versions". *Annals of Mathematics and Artificial Intelligence*, pp. 1–26.

[16] Palar, P. S., Zuhal, L. R., Chugh, T., and Rahat, A., 2020. "On the impact of covariance functions in multi-objective Bayesian optimization for engineering design". In AIAA Scitech 2020 Forum, p. 1867.

[17] Allmendinger, R., Emmerich, M. T., Hakanen, J., Jin, Y., and Rigoni, E., 2017. "Surrogate-assisted multicriteria optimization: Complexities, prospective solutions, and business case". *Journal of Multi-Criteria Decision Analysis,* **24**(1-2), pp. 5–24.

[18] Rojas-Gonzalez, S., and Van Nieuwenhuyse, I., 2019. "A survey on kriging-based infill algorithms for multi-objective simulation optimization". *FEB Research Report KBI_1907.*

[19] Lin, X., Zhen, H.-L., Li, Z., Zhang, Q., and Kwong, S., 2018. "A batched scalable multi-objective Bayesian optimization algorithm". *arXiv preprint arXiv:1811.01323.*

[20] 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.

[21] Rasmussen, C. E., 2006. *Gaussian processes in machine learning.* MIT Press.

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

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

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

[25] 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.

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

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

[28] 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.

[29] Hansen, N., and Kern, S., 2004. "Evaluating the CMA evolution strategy on multimodal test functions". In International Conference on Parallel Problem Solving from Nature, Springer, pp. 282–291.

[30] Jin, Y.-F., Yin, Z.-Y., Zhou, W.-H., and Huang, H.-W., 2019. "Multi-objective optimization-based updating of predictions during excavation". *Engineering Applications of Artificial Intelligence,* **78**, pp. 102–123.

[31] Cox, W., and While, L., 2016. "Improving the IWFG algorithm for calculating incremental hypervolume". In 2016 IEEE Congress on Evolutionary Computation (CEC), IEEE, pp. 3969–3976.

[32] While, L., 2005. "A new analysis of the LebMeasure algorithm for calculating hypervolume". In International Conference on Evolutionary Multi-Criterion Optimization, Springer, pp. 326–340.

[33] 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.

[34] Bostanabad, R., Chan, Y.-C., Wang, L., Zhu, P., and Chen, W., 2019. "Globally approximate Gaussian processes for Big Data with application to data-driven metamaterials design". *Journal of Mechanical Design,* **141**(11).

[35] 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.

[36] 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.

[37] Tran, A., McCann, S., Furlan, J. M., Pagalthivarthi, K. V., Visintainer, R. J., and Wildey, T., 2020. "aphBO-2GP-3B: A budgeted asynchronously-parallel multi-acquisition for known/unknown constrained Bayesian optimization on high-performing computing architecture". *arXiv preprint arXiv:2003.09436.*

[38] 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.

[39] Tran, A., Wildey, T., and McCann, S., 2020. "sMF-BO-2CoGP: A sequential multi-fidelity constrained Bayesian optimization for design applications". *Journal of Computing and Information Science in Engineering,* **20**(3), pp. 1–15.

[40] 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**.

[41] Tran, A., Furlan, J. M., Pagalthivarthi, K. V., Visintainer, 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.

[42] Tran, A., Wang, Y., Furlan, J., Pagalthivarthi, K. V., Garman, M., Cutright, A., and Visintainer, R. J., 2020. "WearGP: A UQ/ML wear prediction framework for slurry pump impellers and casings". In ASME 2020 Fluids Engineering Division Summer Meeting, American Society of Mechanical Engineers.

[43] Zhao, Y.-G., and Ono, T., 1999. "A general procedure for first/second-order reliability method (FORM/SORM)". *Structural safety,* **21**(2), pp. 95–112.

[44] PACE, 2017. *Partnership for an Advanced Computing Environment (PACE).*