Mixed-Input Bayesian Optimization Method for Structural Damage Diagnosis

Congfang Huang, Jaesung Lee, Yang Zhang, Shiyu Zhou, and Jiong Tang

***Abstract*— Structural health monitoring (SHM) is of significant importance in the operation of engineering systems to ensure the durability and reliability. In this article, we introduce a Bayesian optimization method using a multioutput Gaussian process to solve the structural fault diagnosis problem. This method utilizes a high fidelity finite element model (FE) of the structure and the impedance/admittance measurements from the structure to identify the location and severity of the damage. The method improves the accuracy of the damage diagnosis by adopting a multioutput Gaussian process as the surrogate model for the full FE model and Thompson sampling approach is used to guide the search for the structural damage in the Bayesian optimization. The detailed algorithms are presented, and the convergence analysis of the method is conducted. We apply our proposed method on simulated synthetic functions and it achieves better performance and higher convergence speed than the traditional mixed input optimization methods. We then apply our method on a real world structural damage identification problem using measured piezoelectric admittance data and illustrate the effectiveness of the proposed method.**

***Index Terms*—Structural health monitoring, system reliability, Bayesian optimization, multiple output Gaussian process, multiarmed bandit, black-box optimization.**

I. Introduction

Structural health monitoring (SHM) is of vital importance in ensuring the durability and reliability of engineering systems. Structural damage in the systems can cause performance degradation and even lead to catastrophic consequences [1]–[4]. Early detection and identification of the location and severity of structural damage is the key for mitigation such risk. To enhance the reliability of the systems, a number of structural damage diagnosis techniques have been developed in recent years. Structural health monitoring is mostly facilitated through measuring and comparing the dynamic responses of structures [5]–[7]. Methods of SHM utilize vibration measurements from which the natural frequencies and mode shapes are extracted to infer damage occurrence. These methods are easy to implement, as they employ off-the-shelf sensors and the vibration responses can be modeled in a straightforward manner. However, typically only the lower order natural frequencies and mode shapes can be extracted experimentally [8], [9], because it is generally hard and even impossible to excite and measure high-frequency vibration responses using simple experimental setup. As such, the wavelengths involved are large, leading to relatively low sensitivity in damage detection and identification.

Alternatively, a different class of methods utilize the wave propagation information. As waves pass through damage sites, their propagation pattern may exhibit changes [10]. Using transducers such as piezoelectric actuators/sensors with high bandwidth, high-frequency waves can be excited and senses, leading to high detection sensitivity. Nevertheless, the interaction between transient wave and local damage which may have arbitrary profile could be extremely complicated. The identification of damage severity using transient wave propagation becomes difficult [11].

In recent years, a new class of structural damage detection methods, called piezoelectric impedance or admittance based methods [12], [13], has been suggested. The approach works in such a way that a piezoelectric transducer is bonded locally to the host structure to be monitored. The transducer has a two-way electromechanical coupling effect and can be used as both the actuator and sensor simultaneously. When a harmonic voltage excitation is applied, the piezoelectric transducer can excite the structure and measure the electromechanical signatures. When there is damage in the structure, it will result in impedance change around the structural resonances. These impedance changes can be used to monitor structural conditions. These methods preserve the high-bandwidth nature of piezoelectric transducers and thus lead to high detection sensitivity. When a high-fidelity finite element (FE) model in healthy state is available, the damage identification can be carried out to locate and quantify the damage inversely with the electromechanical impedance changes as input. For example, Shuai *et al.*[12] applied the FE modeling, and formulated a linearization of the impedance response through sensitivity matrix computation to represent the relationship of the impedance changes and possible damage. In practice, however, the number of unknowns to be identified, i.e., the location and severity of damage, is large while the data points of impedance response measurements are limited. In another words, the sensitivity matrix is rank deficient [9]. If the conventional least squares method is adopted

Manuscript received April 5, 2022; accepted May 29, 2022. This work was supported by the National Science Foundation under Grant 1824761 and Grant 1825324. Associate Editor: Steven Li. *(Corresponding author: Shiyu Zhou.)*

Congfang Huang, Jaesung Lee, and Shiyu Zhou are with the Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI 53706 USA (e-mail: chuang286@wisc.edu; jaesung.lee@wisc.edu; szhou@engr.wisc.edu).

Yang Zhang and Jiong Tang are with the Department of Mechanical Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: yang.3.zhang@uconn.edu; jiong.tang@uconn.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TR.2022.3179602.

Digital Object Identifier 10.1109/TR.2022.3179602

to solve the underdetermined problem, it may yield untrue solutions.

To address this limitation, the underdetermined identification problem can be formulated as a global optimization problem [14], [15], aiming at minimizing the difference of admittance change between experimental measurements and model prediction. Since the sensitivity matrix is rank-deficient, the underdetermined problem has many or infinite solutions that may not be the true damage location and severity. To address this issue, some sparsity constraints are added to the model to remove the undesired solutions [16], [17]. In [8], a multiobjective DIRECT [18] algorithm to find the damage location and severity that is closest to the observed admittance under the assumption of sparsity is proposed.

Nevertheless, there still exists a research gap. The model remains the linear formulation, which is accurate only when the damage severity is small enough. In addition, even with sparsity constraint, these methods still provide multiple solutions and requires manually select the most plausible solution from them. A better algorithm is expected to be developed to find a more accurate solution of the problem and provide deterministic solution.

In this work, we propose a mixed-input global optimization approach for structure damage diagnosis. In this approach, we treat the structure FE model as a black-box function and then we search the input to the function (i.e., damage location and severity) that fits the obtained observations the best. Using the full FE model, we can avoid the accuracy loss in the linearization step. Black-box optimization has been studied extensively. Bayesian optimization (BO) is a popular black-box function optimization approach [19]–[22], particularly for functions that are expensive to evaluate. It has been applied to a number of scientific domains in recent years [23], [24]. BO uses statistical surrogate model fitted to the data and utilizes the surrogate model to find the next point to evaluate, so that the optimization process can efficiently converge to the optimal solution. Benefiting from the properties inherited from the normal distribution, Gaussian process (GP) has been widely applied in data modeling, simulation optimization, and prediction problems [25]–[27]. The majority of applications of BO also utilizes GP as a surrogate model to be fitted to the data. It has been shown that BO that uses GP as the surrogate model will converge to the global optimal under some nonrestrictive constraints [28].

Most of the existing BO methods and applications focus on the cases where there are only continuous variables in the input domain of the function [20]–[22], [29]. However, for the structure damage diagnosis problem at hand, the input variables are of mixed types: the location of the damage is a discrete variable while the severity of the damage is a continuous variable. Optimization with both discrete and continuous inputs are interesting topics in the optimization field. Unlike the continuous input variables, discrete variables contain the information that is not easy to be explored. Recently, Nguyen *et al.* used multiarmed bandit model to frame the BO problems (MAB-BO) with discrete and continuous input variables [30].

Although MAB-BO is able to handle both discrete and continuous variables, it does not consider the correlation of the responses corresponding to different discrete inputs in each optimization iteration. In other words, the responses corresponding to different discrete inputs are modeled independently by separate GP models in MAB-BO method [30]. However, in many real world cases, the responses corresponding to different discrete inputs often show similar patterns. Take the structure damage diagnosis problem as an example. Intuitively, damages occurring on close locations may show very similar patterns in the structure dynamic response and the information from adjacent locations may contribute to the prediction for each other. Thus, if we have a model to simultaneously describe multiple responses corresponding to multiple potential damage locations, then we will have more accurate model to describe the underlying function, which will lead to better performance of the BO algorithm.

In this work, we adopt a recently developed multioutput Gaussian process (MGP) model with nonseparable covariance functions as the surrogate model in the BO algorithm. Similar to MAB-BO in [30], a Thompson sampling approach is used to guide the sequential sample of the underlying function, which can provide a good tradeoff between exploitation and exploration of the search. A rigorous convergence analysis is conducted to show the converging property of the proposed method. Numerical studies based on both simulated data and real world structure damage diagnostics problem compare the proposed method and several other methods and demonstrate the effectiveness of our method.

The main contributions of this work can be summarized into three folds:

1) apply BO, a black-box function optimization method, to the structure damage diagnosis problem, which can directly work on the full FE model without linearization;
2) extend the existing BO method by adopting a flexible nonseparable MGP model as the statistical surrogate model in the BO framework;
3) rigorous convergence results are obtained for the proposed BO method.

The rest of the article is organized as follows. In Section II, we introduce the related background of Bayesian optimization (BO), including a brief introduction of BO and the Gaussian process used in BO. In Section III, the multiarmed bandit multioutput Gaussian process Bayesian Optimization (MAB-MGP-BO) is presented and the convergence of the method is analyzed. In Section IV, numerical simulation on two synthetic functions is conducted. The results show that our method can achieve a higher optimal value and a faster convergence than several other methods including MAB-BO method in [30]. In Section V, case study of structural damage identification based on both simulated data and real world data are conducted. It shows that our method can successfully detect the location and severity of the damage. Finally, we conclude our work and have a discussion on future work in Section VI.

II. Review of Bayesian Optimization

In this section, we give a brief introduction of related background knowledge on the Bayesian optimization (BO) we will utilize in the proposed method. Generally, BO methods consist

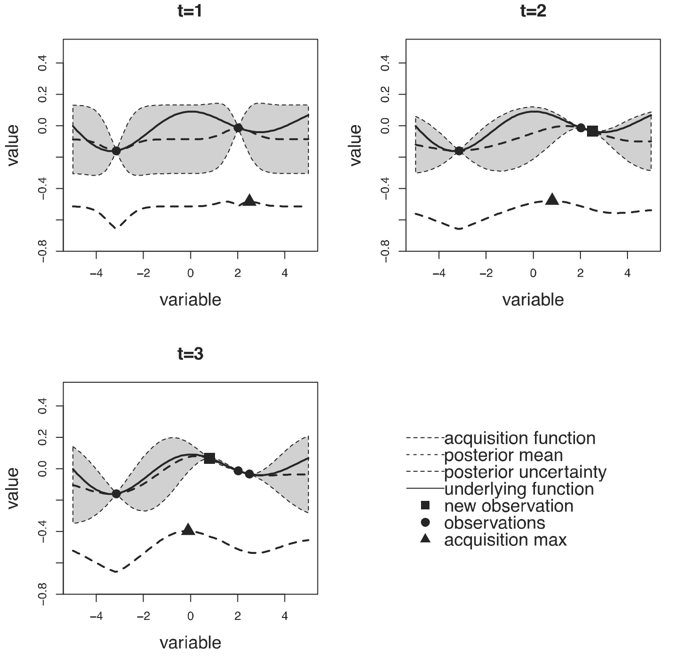


Fig. 1. Example of Bayesian optimization method.

of two parts. First, a machine learning method is utilized to build a surrogate model of the input and the objective function. Then, an acquisition function is designed to decide where to take the next observation (also called "evaluation") from the objective function. Repeating the two steps sequentially, we can build a surrogate model close to the underlying black-box function and find a solution near the global optimal point. Fig. 1 is an illustration of BO method using GP as the surrogate model and upper confidence bound [31] of prediction as the acquisition function under noiseless condition, where t is the number of iterations, i.e., the number of evaluations we made on the underlying black-box function f. The x-axis is the input of the function and the y-axis is the value of function f. The values of the acquisition function, the upper confidence bounds with one standard deviation, are shown in dash-dotted line on the bottom of the box at each t. The true underlying functions are presented in solid lines above them. The observations we already have are presented in by round points and the square point is the new observation point we made in the current iteration, which is chosen based on the acquisition function of the former iteration. The dotted and short dotted lines are the posterior predictions and the confidence intervals. As we can see from the figure, the prediction is closer to the underlying function and the uncertainty is lower through the iterations.

Formally, let $f(\mathbf{x})$ be the black-box function with global optimizer $\mathbf{x}^* = \text{argmax}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$, where $\mathbf{x}$ is the vector of variables and $\mathcal{X}$ is the domain of $\mathbf{x}$. Assume that till the tth iteration, the set of observation points $D_t = \{(\mathbf{x}_i, y_i)|i = 1, \ldots, t\}$, where $y_i = f(\mathbf{x}_i) + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. If we fit the surrogate model to the data in each iteration and we obtain the prediction of $f(\mathbf{x})$ with mean $\mu_t(\mathbf{x})$ and variance $\sigma_t^2(\mathbf{x})$. Taking the upper confidence bound $U_t(\mathbf{x}) = \mu_t(\mathbf{x}) + \beta\sigma_t(\mathbf{x})$, (in Fig. 1, $\beta = 1$) as the acquisition function, we take $\mathbf{x}_{t+1} = \text{argmax}_{\mathbf{x}\in\mathcal{X}} U_t(\mathbf{x})$ as the next evaluation point. Then, we evaluate $\mathbf{x}_{t+1}$ with the underlying model and get the response y_{t+1}. Finally, the tth iteration is finished and the dataset of the $(t+1)$th iteration is $D^{t+1} = D^t \cup \{(\mathbf{x}_{t+1}, y_{t+1})\}$. After a large enough number of iterations, we can achieve the global extreme point x^*.

As we mentioned above, the most frequently used surrogate model in BO is the GP model. A GP is a generalization of the Gaussian probability distribution and can be considered as a Gaussian distribution prior over functional data. It has been well established that GP model is a very flexible and expressive model that can be used to describe a wide range of functions [32], [33]. It is a stochastic process, which is a collection of random variables and any finite subcollection of which follows a multivariate normal distribution. Thus, the distribution of a GP is the joint distribution of all those (infinitely many) random variables.

A GP $f(\mathbf{x})$ is specified by its mean function $\mu(\mathbf{x})$, $\mu : \mathcal{X} \to \mathbb{R}$ and its covariance function $c(\mathbf{x}, \mathbf{x}')$, $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where $c(\mathbf{x}_i, \mathbf{x}_j)$ is the covariance between $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$. Please note that without causing confusion, we use the same symbol f to represent the underlying function and a GP. We denote the GP as $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), c(\mathbf{x}, \mathbf{x}'))$. Specifically, for a set of inputs $\mathbf{X}_n = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$, the vector of output $f(\mathbf{X}_n)$ is Gaussian distributed with mean $\mu(\mathbf{X}_n) = [\mu(\mathbf{x}_1), \mu(\mathbf{x}_2), \ldots, \mu(\mathbf{x}_n)]^T$ and $n \times n$ covariance matrix $C(\mathbf{X}_n, \mathbf{X}_n)$:

$$C(\mathbf{X}_n, \mathbf{X}_n) = \begin{bmatrix} c(\mathbf{x}_1, \mathbf{x}_1) & c(\mathbf{x}_1, \mathbf{x}_2) & \cdots & c(\mathbf{x}_1, \mathbf{x}_n) \\ c(\mathbf{x}_2, \mathbf{x}_1) & c(\mathbf{x}_2, \mathbf{x}_2) & \cdots & c(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ c(\mathbf{x}_n, \mathbf{x}_1) & c(\mathbf{x}_n, \mathbf{x}_2) & \cdots & c(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}. \tag{1}$$

A common covariance function is the squared exponential kernel, defined as $c(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2\text{exp}(-||\mathbf{x}_i - \mathbf{x}_j||_2^2)/(2l)^2$, where l is a length scale parameter and σ^2 is the parameter dictating the uncertainty in $f(\mathbf{x})$.

When we utilize a GP in BO as used in Fig. 1, the posterior GP with a newly observed $\mathbf{y}$ at iteration t can be expressed as

$$\begin{aligned} &f(\mathbf{x})|D_t \sim \mathcal{N}(\mu_t(\mathbf{x}), \sigma_t^2(\mathbf{x})) \\ &\quad \mu_t(\mathbf{x}) = C(\mathbf{x}, \mathbf{X}_t)[C(\mathbf{X}_t, \mathbf{X}_t) + \sigma_\epsilon^2\mathbf{I}]^{-1}\mathbf{y} \\ &\quad \sigma_t^2(\mathbf{x}) = C(\mathbf{x}, \mathbf{x}) - C(\mathbf{x}, \mathbf{X}_t)[C(\mathbf{X}_t, \mathbf{X}_t) + \sigma_\epsilon^2\mathbf{I}]^{-1}C(\mathbf{X}_t, \mathbf{x}). \end{aligned} \tag{2}$$

Another essential part of BO is the acquisition function, which determines how the search space are explored during the iterations. Acquisition functions are based on the posterior distributions of the GP fitted in each iteration. There is a tradeoff between two directions of search: exploitation and exploration. Exploitation improves the region near the current best result, while exploration develops more unfamiliar regions that have higher uncertainties. Both directions are necessary to get to the optimal point in the long run of the algorithm. Otherwise the algorithm may stuck in a local optimum or jump around with no sense. Most common acquisition functions includes

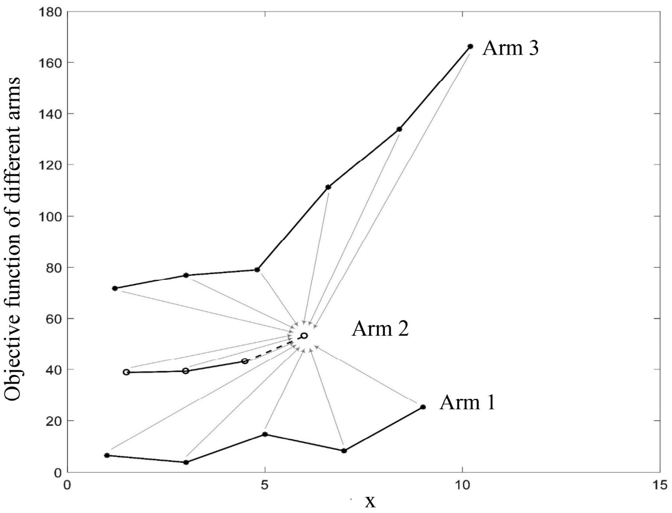


Fig. 2. Illustration of MGP. Three functions are shown, which correspond to three different categorical input values.

upper confidence bound, expected improvement, and Thompson sampling [34], [35]. We will provide more detailed discussion on these acquisition functions in the next section.

III. Multiarmed Bandit Bayesian Optimization With Multioutput Gaussian Process

In this section, we first introduce the detailed way to construct covariance function of multioutput Gaussian process (MGP) in Section III-A. Then, we introduce the proposed method multiarmed bandit multioutput Gaussian process Bayesian optimization (MAB-MGP-BO) in Section III-B. Finally, the convergence analysis of the proposed method is conducted in Section III-C.

A. Convolved Process and Multioutput Gaussian Process

As we mentioned before, in many cases, not only the observations under the same categorical input value but also the observations from different categorical input values contribute information to the fitting of the model. A model that can consider both the information within the same category and the information across different categories is preferred. In Fig. 2, we show three functions that correspond to the objective function under three different categorical input values. In the multiarmed Bandit formulation, the objective function under a categorical input value is also regarded as an arm.

For an MGP model, we should not only specify the covariance function for a single output (as that in the conventional univariate output GP model), but also the cross covariance among different outputs. Thus, the covariance function of MGP is in the form of $c(i, j, \mathbf{x}, \mathbf{x}')$, where $c(i, j, \mathbf{x}, \mathbf{x}') = \text{cov}(f_i(\mathbf{x}), f_j(\mathbf{x}'))$ and i, j are output indices. The parameterization of the covariance function $c(i, j, \mathbf{x}, \mathbf{x}')$ plays a critical role in MGP model because it characterizes the relationship between any two outputs. A popular approach is to use a separable covariance structure, i.e., letting $c(i, j, \mathbf{x}, \mathbf{x}') = \tau(i, j) \times \text{cov}(\mathbf{x}, \mathbf{x}')$ [36]–[38]. Other formats combining separable covariance functions are also designed in recent research [39], [40], but they still calculate the covariance of continuous and categorical variables separately. In the current state-of-the-art open source implementation of GP model such as GPyTorch [41], the separable covariance function is used. The separable structure is appealing due to the simplified covariance structure, however it restricts all outputs to share the same set of covariance parameters, i.e., the part cov $(\mathbf{x}, \mathbf{x}')$ is the same for all the outputs. However, different outputs may have different characteristics and thus it is too restrictive to use the same covariance function for the continuous variables across different outputs.

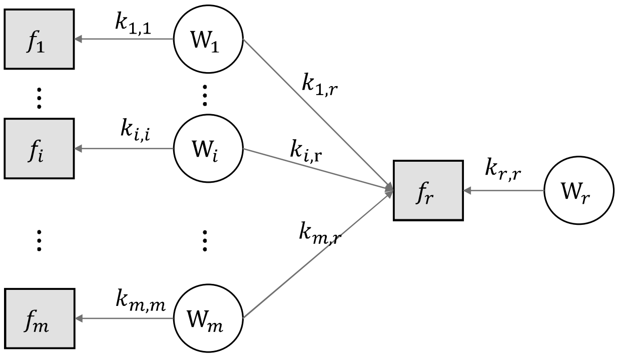


Fig. 3. Illustration of the construction of MGP with shared white noise Gaussian Processes.

To overcome the limitation of separable covariance function, we adopt a nonseparable covariance structure that is based on convolution processes (CP) [42]–[44]. The CP-based nonseparable covariance function is based on that a GP can be constructed by convolving a latent Gaussian white noise process with a smoothing kernel [45]. In more details, we can construct a GP $f(\mathbf{x})$ by convolving a Gaussian white noise process $W(\mathbf{x})$ with a smoothing kernel $k(\mathbf{x}) = \exp(-\mathbf{x}^2/(2l^2))$, where $\text{cov}(W(\mathbf{x}), W(\mathbf{x}')) = \delta(\mathbf{x} - \mathbf{x}')$, l is the length scale parameter, δ is the Dirac delta function, defined in [46] as

$$\int_{-\infty}^{\infty} \delta(x)dx = 1 \quad , \delta(x) = \begin{cases} 0, & x \neq 0 \\ \infty, & x = 0 \end{cases} \tag{3}$$

such that

$$f(\mathbf{x}) = \int_{-\infty}^{\infty} k(\mathbf{x} - \mathbf{u})W(\mathbf{u})d\mathbf{u}. \tag{4}$$

Then, the corresponding covariance function can be expressed as

$$\text{cov}(f(\mathbf{x}), f(\mathbf{x}')) = \int_{-\infty}^{\infty} k(\mathbf{x} - \mathbf{u})k(\mathbf{x}' - \mathbf{u})d\mathbf{u}. \tag{5}$$

Thus, the GP is parameterized by the parameters in the smoothing kernel k, which is required to be square or absolutely integrable, i.e., $\int_{-\infty}^{\infty} |k(\mathbf{x})|^2 d\mathbf{x} < \infty$.

For a MGP setting, if each output is constructed in this way, and if we share these latent Gaussian white noise process across multiple outputs, then multiple outputs can be expressed as a jointly distributed GP [47], [48]. For example, a nonseparable MGP construction is shown in Fig. 3. This construction is first proposed by [49] and is adopted in this work. In Fig. 3, we try to construct a surrogate MGP model for the function corresponding to the rth arm, denoted as f_r, while $f_i,\ i \in \{1, ..., m\} \backslash r$ correspond to the functions from other arms. Then, according to this construction, $f_r(\mathbf{x})$ can be written as

$$f_r(\mathbf{x}) = k_{r,r}(\mathbf{x}) \star W_r(\mathbf{x}) + \sum_{i=1}^{m} k_{i,r}(\mathbf{x}) \star W_i(\mathbf{x}) \qquad (6)$$

where $k_{i,j}(\mathbf{x}) \star W_i(\mathbf{x}) = \int_{-\infty}^{\infty} k_{i,j}(\mathbf{x}-\mathbf{u}) W_i(\mathbf{u}) d\mathbf{u}$ and $k_{i,j}$ is the kernel used in the convolution of the ith latent function and the jth output, $i, j \in \{1, 2, \ldots, m\}$. For the functions from other arms, it is constructed as $f_i(\mathbf{x}) = k_{i,i}(\mathbf{x}) \star W_i(\mathbf{x})$. Clearly, W_i is shared across f_i and f_r in the construction and thus, the cross correlation between f_i and f_r can be modeled.

With this construction, we can obtain the covariance matrix of the MGP model for f_r as [49]

$$C_{N \times N} = \begin{bmatrix} C^{1,1}_{n_1 \times n_1} & \mathbf{0}_{n_1 \times n_2} & \cdots & \mathbf{0}_{n_1 \times n_m} & C^{1,r}_{n_1 \times n_r} \\ \mathbf{0}_{n_2 \times n_1} & C^{2,2}_{n_2 \times n_2} & \cdots & \mathbf{0}_{n_2 \times n_m} & C^{2,r}_{n_2 \times n_r} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_{n_m \times n_1} & \mathbf{0}_{n_m \times n_2} & \cdots & C^{m,m}_{n_m \times n_m} & C^{m,r}_{n_m \times n_r} \\ C^{r,1}_{n_r \times n_1} & C^{r,2}_{n_r \times n_2} & \cdots & C^{r,m}_{n_r \times n_m} & C^{r,r}_{n_r \times n_r} \end{bmatrix} \qquad (7)$$

where n_i is the number of observed data points from function f_i, $N = \sum_{i=1}^{m} n_i$, $C^{i,i}_{n_i \times n_i}$ is the covariance within function f_i, $C^{r,i}_{n_r \times n_i}$ is the cross covariance between f_r and f_i, and $\mathbf{0}_{n_i \times n_j}$ is the zero matrix with dimension $n_i \times n_j$. Please note the covariance matrix in (7) is a function of all the parameters of kernel functions used in the construction, including $k_{i,i}$, $k_{i,r}$, and $k_{r,r}, i = 1, ..., m$. Once the covariance matrix is constructed, we can actually view the MGP as a conventional GP model and use the maximum likelihood estimation method to estimate the kernel function parameters based on the observed data from the functions of all the arms. With the fitted MGP model, we can also make probabilistic predictions of f_r at other input locations using the formula in (2). More technical details of the model construction and estimation can be found in [49].

One point we want to emphasize here is that the MGP model in Fig. 3 is to model the function f_r only. If we want to model the function from a different arm, then we need to put that function in the place of f_r in Fig. 3 and build a different MGP model. In other words, in our proposed MAB-MGP-BO method (details are shown in Section III-B), for each arm, we have a different MGP surrogate model. It is possible to establish a complex CP based MGP model to model all the function simultaneously. However, such a model may involve a very large number of parameters and difficult to estimate and use. The arrangement in Fig. 3 allows us to use multiple simpler models to describe the functions and each MGP is relatively easy to estimate.

B. MAB-MGP-BO Algorithm

Suppose we have a collection of black-box function $\{f_a(\mathbf{x}_a)\}_{a=1}^{m}$ with both categorical variable $a \in \{1, 2, \ldots, m\}$ and continuous variable $\mathbf{x}_a \in \mathcal{X} \subset \mathbb{R}^d$. The goal of the algorithm is to find the optimum point $[a^*, \mathbf{x}^*_{a^*}] = \text{argmax}_{[a,\mathbf{x}] \in \{1,2,\ldots,m\} \times \mathcal{X}_a} f_a(\mathbf{x})$, which can also be written as $a^* = \text{argmax}_{a \in \{1,2,\ldots,m\}} f(\mathbf{x}^*_a)$ and $\mathbf{x}^*_a = \text{argmax}_{\mathbf{x} \in \mathcal{X}_a} f(\mathbf{x}_a)$.

Algorithm 1: MAB-MGP-BO.

```
Input: $m$: number of arms/outputs, $t$: number of iterations
Output: $\mathbf{x}^*$: the optimal solution of $f$ and $f^* = f(\mathbf{x}^*)$
1:  for $t = 1, 2, \ldots$ do
2:      for $a = 1, 2, \ldots, m$ do
3:          Fit the Multi-output Gaussian process model to
            $D_t = \{(a_i, \mathbf{x}_i, y_i) | i = 1, \ldots, t\}$
4:          Draw a sample $\tilde{f}_a(\mathbf{x})$ from the fitted model $\hat{f}_a(\mathbf{x})$:
            $\tilde{f}_a(\mathbf{x}) \sim p(\hat{f}_a(\mathbf{x}) | D_t)$
5:          Find $\tilde{\mathbf{x}}^*_a = \text{argmax}_{\mathbf{x} \in \mathcal{X}_a} \tilde{f}_a(\mathbf{x})$
6:          Let $\tilde{f}^*_a = \tilde{f}_a(\tilde{\mathbf{x}}^*_a)$
7:      end for
8:      $a_{t+1} = \text{argmax}_{1 \le a \le m} \tilde{f}^*_a$
9:      $\mathbf{x}_{t+1} = \tilde{\mathbf{x}}^*_{a_{t+1}}$
10:     Evaluate $y_{t+1} = f(\mathbf{x}_{t+1}) + \epsilon_{t+1}$
11:     if $y_{t+1} > f^*$ then
12:         $f^* = y_{t+1}, \mathbf{x}^* = \mathbf{x}_{t+1}$
13:     end if
14:     $D_{t+1} = \{(a_{t+1}, \mathbf{x}_{t+1}, y_{t+1})\} \bigcup D_t$
15: end for
```

To solve the problem, we can formulate it into a multiarmed bandit (MAB) problem [50]. For each arm a, we use BO to find the optimal point $\mathbf{x}^*_a = \text{argmax}_{\mathbf{x} \in \mathcal{X}_a} f(\mathbf{x}_a)$. Then, to find the optimal arm $a^* = \text{argmax}_{a \in \{1,2,\ldots,m\}} f(\mathbf{x}^*_a)$. In the proposed MAB-MGP-BO, we use MGP model in each arm as the surrogate model for the black-box function corresponding to the arm. The MGP model can borrow information from the black-box functions of other arms, which could lead to more accurate surrogate model and in turn lead to better optimization performance. The detailed algorithm is presented in Algorithm 1.

Take m as the number of arms (also called number of outputs) and t as the number of iterations or evaluations. In each iteration t, we fit the MGP to D_t, the total dataset at iteration t. The fitted model is denoted by $\hat{f}_a(\mathbf{x})$. Then, we draw a sample function $\tilde{f}_a(\mathbf{x})$ from the posterior Gaussian distribution of the MGP: $\tilde{f}_a(\mathbf{x}) \sim p(\hat{f}_a(\mathbf{x}) | D_t)$. We locate $\tilde{\mathbf{x}}^*_a$, the optimal point at arm a, by maximizing the sampled function $\tilde{f}$. The value of $\tilde{f}$ at $\tilde{\mathbf{x}}^*_a$ is denoted by $\tilde{f}^*_a$. After finding the optimal solution for the every arm, find the best arm and data point a_{t+1} and $\mathbf{x}_{t+1}$ for evaluation at the next iteration. Then, evaluate the underlying true function f and get the new observation value $y_{t+1} = f(\mathbf{x}_{t+1}) + \epsilon_{t+1}$. Finally, the dataset is updated as D_{t+1} by adding the new observed data point $(a_{t+1}, \mathbf{x}_{t+1}, y_{t+1})$.

We would like to mention a couple of points regarding the above algorithm:

1) Line 5 to Line 13 of Algorithm 1 is the Thompson sampling (TS) approach to determine the next input point to evaluate. As mentioned in the introduction section, there are many different method to sequentially select the next point to evaluate, including the popular methods such as the expected improvement (EI), the upper confidence bound (UCB), and TS. EI is a greedy improvement-based

strategy that suggests the point that could improve the expectation of the function the most over the current evaluated point. Although EI is effective and provides reasonable performance in the practice, it is often too greedily focusing on the existing optimum point and collecting little information about the unfamiliar regions in the domain [51]. UCB, as we mentioned earlier, chooses the point that maximum the upper confidence bound $U_t(\mathbf{x}) = \mu_t(\mathbf{x}) + \beta\sigma_t(\mathbf{x})$ of the posterior prediction of the underlying function for the next iteration. This algorithm can balance the exploration and exploitation of the optimization. However, the performance of the optimization may depend on the confidence parameter β and it holds a lower regret bounds than the TS method as an acquisition function of BO [52]. TS proposed by [53] is a sequential sampling heuristic that can handle the exploration–exploitation dilemma. Instead of setting a closed form expression, this method samples a function from the fitted Gaussian process and suggests the next iteration point by the maximum value of the sampled function. TS has desirable theoretical properties [52] with a pretty tight regret bound and provides basis to the convergence analysis for MAB-MGP-BO as discussed in Section III-C.

2) On Line 8 of Algorithm 1, we need to find the maximum value from a sample (i.e., a realization) of GP from the fitted MGP model. A straightforward grid search is often used to solve the optimization problem on Line 8: we just treat the MGP as a multivariate Gaussian distribution and sample it at a large number of regularly distributed input points (also called grid points) and then find the largest value among the sampled values. This method is effective and viable only when the dimension of input is low. If the dimension of input is high, the number of grids points to fill the input space will be overwhelming. For example, for a problem with 10 dimensional input and 1000 grid points for each dimension, we will have 10^{30} grid points. It is infeasible to sample such a large number of points simultaneously. Instead, we need to adopt a sequential sampling strategy when we solve the problem in Line 8: We just sample one or a small batch of points from the MGP at an iteration and let a searching algorithm (such as a gradient descent searching method) to determine what next points to sample for the next iteration. One critical point for the sequential sampling method is that the samples from different iterations should not be *independent* samples from the MGP. Rather, *the later samples depend on the previous samples so that the samples obtained from the sequential procedure should be the same as if they are sampled simultaneously.* We have established a sequential sampling method as shown in Algorithm 2.

The sequential sampling algorithm from MGP is shown in Algorithm 2.

Algorithm 2: Sequential Sampling from MGP.

```
Initialize: Global $D_t^s = \emptyset$: the set of sampled points from
  the fitted MGP, denoted by $\hat{f}_a(\mathbf{x})$.
1: function SAMP x
2:     Calculate the posterior Gaussian distribution of given
         both the evaluated and sampled data $\hat{f}_t^a|D_t, D_t^s$.
3:     Sample a new point from the posterior distribution
     $\tilde{f} \sim p(\hat{f}_a(\mathbf{x})|D_t, D_t^s)$
4:     Save the new sampled point: $D_t^s = D_t^s \bigcup\{(\mathbf{x}, \tilde{f})\}$
5:     return $\tilde{f}$
6: end function
7: Optimize the sampling function: $\tilde{\mathbf{x}}_a^* = \text{argmax}_{\mathbf{x}\in\mathcal{X}_a}$
     SAMP x
```

At each iteration t and for each arm a of the Bayesian optimization model, we wrap the sampling process as a function of the new sample point $\mathbf{x}$ given the set of previously sampled points D_t^s. The sampled points D_t^s are regarded as evaluated points and the new sample value $\tilde{f}$ is based on the posterior distribution of not only the evaluated points D_t, but also the previously sampled points. Given the fitted MGP $\hat{f}_a(\mathbf{x})|D_t$ and a sampled dataset D_t^s, we can use the following lemma to obtain the posterior MGP sample $\tilde{f} \sim \hat{f}_a(\mathbf{x})|D_t, D_t^s$. This result is used in Line 4 of Algorithm 2.

Lemma 1: $D_t = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_t}$ is a set of the observed data points, where $y_i \sim f(\mathbf{x}) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. Suppose $D_t^s = \{(\mathbf{x}_j^s, \tilde{f}_j)\}_{j=1}^{n_s}$ is the set of sampled functions from the posterior distribution of the fitted MGP $\hat{f}(\mathbf{x})$ in the tth iteration. Let $\mathbf{X}_t = [\mathbf{x}_1, \ldots, \mathbf{x}_{n_t}]^T$, $\mathbf{X}_s = [\mathbf{x}_1^s, \ldots, \mathbf{x}_{n_s}^s]^T$, $\mathbf{y} = [y_1, \ldots, y_{n_t}, \tilde{f}_1, \ldots, \tilde{f}_{n_s}]^T$. Then, we have $f_a(\mathbf{x})|D_t, D_t^s \sim \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, where

$$\boldsymbol{\mu}_s = \begin{bmatrix} C(\mathbf{x}, \mathbf{X}_t) \\ C(\mathbf{x}, \mathbf{X}_s) \end{bmatrix} \begin{bmatrix} C(\mathbf{X}_t, \mathbf{X}_t) + \sigma_\epsilon^2\mathbf{I} & C(\mathbf{X}_t, \mathbf{X}_s) \\ C(\mathbf{X}_s, \mathbf{X}_t) & C(\mathbf{X}_s, \mathbf{X}_s) \end{bmatrix}^{-1} \mathbf{y}$$

$$\boldsymbol{\Sigma}_s = c(\mathbf{x}, \mathbf{x}) - \begin{bmatrix} C(\mathbf{x}, \mathbf{X}_t) \\ C(\mathbf{x}, \mathbf{X}_s) \end{bmatrix}$$

$$\begin{bmatrix} C(\mathbf{X}_t, \mathbf{X}_t) + \sigma_\epsilon^2\mathbf{I} & C(\mathbf{X}_t, \mathbf{X}_s) \\ C(\mathbf{X}_s, \mathbf{X}_t) & C(\mathbf{X}_s, \mathbf{X}_s) \end{bmatrix}^{-1} [C(\mathbf{x}, \mathbf{X}_t), C(\mathbf{x}, \mathbf{X}_s)]. \tag{8}$$

Following Algorithm 2, we can sequentially sample a MGP while guarantee the sequentially sampled data points have the same property as if they are sampled simultaneously from the MGP. The proof of this property utilizes the basic properties of conditional multivariate normal distribution and is omitted here. With the sequential sample approach, gradient-based optimization algorithms can be used in Line 8 of Algorithm 1 to find the optimal point of the sampled GP from the fitted posterior MGP directly.

C. Convergence Analysis

In this section, we present the convergence analysis of the proposed MAB-MGP-BO method. Bayesian regret has been used for a performance measure for the Bayesian optimization with Thompson sampling due to its sampling scheme [52]. The Bayesian regret is defined as follows. Assume the underlying function is f and $\mathbf{x}^*$ is the optimal solution of f,

$\mathbf{x}^* = \text{argmax} f(\mathbf{x})$. In iteration t, the algorithm suggests $\mathbf{x}_t$ as the optimal solution, then $r_t = f(\mathbf{x}^*) - f(\mathbf{x}_t)$ is the instantaneous regret of the algorithm. The Bayesian regret is the expectation of the summation of the instantaneous regret by the Tth iteration: $BayesRegret(T) = E[\sum_{t=1}^{T} r_t]$. With the Bayesian regret, we can assess the convergence rate of the algorithm by finding the bound of the marginal increment of the Bayesian regret $BayesRegret(T)/T$.

For the observed data in the tth iteration of the algorithm, suppose we choose the a_tth arm and the optimal point is $\mathbf{x}_t$, then we have

$$y_t = f_{a_t}(\mathbf{x}_t) + \epsilon_t, \quad t = 1, \ldots, T. \tag{9}$$

We define $f_a(\mathbf{x})$ is the MGP for the ath arm with mean zero and covariance function $\sigma^2 c(\mathbf{x}, \mathbf{x}')$, and $\mathbf{x}, \mathbf{x}' \in \mathcal{X}_a \subset \mathbb{R}^d$. We denote the observed data of the ath output as $\mathcal{D}_t^a = \{(a_t, x_t, y_t) | a_t = a\}_{t=1}^T$ and the whole dataset by $\mathcal{D}_t = \cup_{a=1}^m \mathcal{D}_t^a$, where m is the number of arms.

According to our formulation, Bayesian regret of the MAB-MGP-BO model by the T iteration is defined as the expected difference between the underlying optimal value and the optimal value we find by the Bayesian optimization:

$$BayesRegret(T) = \mathbb{E}\left[\sum_{t=1}^{T} \{f_{a^*}(x^*) - f_{a_t}(x_t)\}\right]. \tag{10}$$

To find the Bayesian regret of the MAB-MGP-BO method, we need two following assumptions. These assumptions are not restrictive and have been used in other convergence analysis works [31] and [30].

Assumption 1: For all a and for any sample f from $\mathcal{GP}$, there exist constants $r, s > 0$ such that its partial derivatives satisfy the following condition:

$$\begin{aligned} &P(|\partial f / \partial x_i| < L) \geq 1 - dr \exp(-L^2/s^2) \\ &\forall L > 0 \quad \forall i \in \{1, \ldots, d\} \end{aligned} \tag{11}$$

where d is the dimension of the model input $\mathbf{x}$.

With this assumption, the partial derivatives of f is bounded in probability. Another assumption is on the maximum information gain of the model. The information gain is the mutual information shared by f and the observations $y_o = f(\mathbf{x}_o) + \epsilon_o$ on $\mathbf{x}_o \in O \subset \mathcal{X}$, where $\epsilon_o \sim \mathcal{N}(0, \sigma^2)$. Then, the information gain is defined as

$$I(y_o; f) = H(y_o) - H(y_o | f). \tag{12}$$

$H(y_o)$ is the marginal entropy of the observation y_o and $H(y_o|f)$ is the conditional entropy of the observation y_o given the corresponding function value $f(\mathbf{x}_o)$. For a GP $f \sim \mathcal{N}(\mu, \Sigma)$, $H(\mathcal{N}(\mu, \Sigma)) = \log|2\pi e \Sigma|/2$. The information gain is a measure of the reduction in uncertainty of function f after knowing the observations y_o. The maximum information gain after T iterations is then defined as

$$\gamma_T = \max_{O \subset D : |O| = T} I(y_o; f). \tag{13}$$

With the following Assumption 2, we can guarantee the maximum information gain of the model within a sublinear bound.

Assumption 2: The maximum information gain γ_{T_a} about f_a in Bayesian Optimization with Gaussian process is sublinear in T_a which is the number of observations in the ath arm; there exists an α such that $\gamma_{T_a} \sim \mathcal{O}(T_a^\alpha)$ where $0 \leq \alpha < 1$.

Both assumptions 1 and 2 hold for our multioutput Gaussian process. The kernel used for the GP in this work reduced to the form of Gaussian kernel with paired scale parameters corresponding to each category. Assumption 1 holds for any four-times differentiable covariance functions as stated in [31]; thus, it holds for the Gaussian kernel function. Assumption 2 holds for the Gaussian kernel function [30]. In other words, our multioutput Gaussian process satisfies both the Assumptions 1 and 2.

Under the assumptions, the Bayesian regret of the MAB-MGP-BO method in Algorithm 1 can be bounded as follows. This result stated in Theorem 1 shows that the Bayesian regret bound of the algorithm is growing sublinearly in T with factor $\sqrt{m}$. A sublinear Bayesian regret implies the solution of MAB-MGP-BO will converge to the global optimal point.

Theorem 1:

$$\begin{aligned} BayesRegret(T) &= \mathbb{E}\left[\sum_{t=1}^{T} \{f_{a^*}(x^*) - f_{a_t}(x_t)\}\right] \\ &\leq \mathcal{O}\left(\sqrt{mT^{\alpha+1} \log T}\right) \end{aligned} \tag{14}$$

where $0 \leq \alpha < 1$.

We provide an outline of the proofs of the theorem. We use $f_a(x)|\mathcal{D}_t$ for the posterior distribution of the MGP instead of the posterior distribution of $f_a(x)|\mathcal{D}_t^a$, which is a conventional univariate output GP as they used. In other words, instead of using the dataset for specific a, we used whole dataset $\mathcal{D}_t$ leveraging all the information across a by using MGP. Theorem 1 can be proved after the following decomposition of the Bayesian regret

$$\begin{aligned} BayesRegret(T) &= \mathbb{E}\left[\sum_{t=1}^{T} \{f_{a^*}(x^*) - f_{a_t}(x_t)\}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \left\{f_{a^*}(x^*) - f_{a_t}(x_{a_t}^*)\right\}\right] \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{T} \left\{f_{a_t}(x_{a_t}^*) - f_{a_t}(x_t)\right\}\right] \qquad (15) \\ &= R_T^{MAB} + R_T^{BO}. \qquad (16) \end{aligned}$$

The boundary of Bayesian regret will be obtained by having the boundaries of R_T^{MAB} and R_T^{BO} in Lemmas 2 and 3. For the proofs of Lemmas 2 and 3, we state them in Appendix A. In the MAB-MGP-BO model, the Bayesian regret consists of two parts. One is the Bayesian regret of the multiarmed bandit model $R_T^{MAB} = \mathbb{E}[\sum_{t=1}^{T} \{f_{a^*}(x^*) - f_{a_t}(x_{a_t}^*)\}]$, which is the expected cumulative difference between the underlying optimal value of all the arms and the optimal value of the selected optimal arm a_t from the MAB model. We define the Gaussian process model used for the function at the arm a as $\mathcal{GP}_a$. The boundary of R_T^{MAB} is presented in Lemma 2. Given the observations by the Tth iteration, the Bayesian regret generated from multiarmed bandit model can be bounded by an sublinear upper bound.

Lemma 2: The Bayesian regret from multiarmed bandit R_T^{MAB} in the MAB-MGP-BO model has an upper bound of $\mathcal{O}(\sqrt{mT^{\alpha+1}\log T})$

$$\begin{aligned} R_T^{MAB} &= \mathbb{E}\left[\sum_{t=1}^{T}\{f_{a^*}(x^*) - f_{a_t}(x^*_{a_t})\}\middle|\mathcal{D}_T, \mathcal{GP}_1, \ldots, \mathcal{GP}_m\right] \\ &\le \mathcal{O}(\sqrt{mT^{\alpha+1}\log T}) \qquad (17)\end{aligned}$$

where $0 \le \alpha < 1$.

Another part of the Bayesian regret (16) is the regret of the Bayesian optimization $R_T^{BO} = \mathbb{E}[\sum_{t=1}^{T}\{f_{a_t}(x^*_{a_t}) - f_{a_t}(x_t)\}]$, which is the expected cumulative difference between the underlying optimal value within arm a_t and the optimal value obtained by the Bayesian optimization algorithm with arm a_t. Similarly, the boundary of R_T^{BO} is presented in Lemma 3. Given the observations by the Tth iteration, the Bayesian regret generated from Bayesian optimization can be bounded by an sublinear upper bound.

Lemma 3: The regret from Bayesian optimization R_T^{BO} in the MAB-MGP-BO model has an upper bound of $b\sqrt{mT^{\alpha+1}\log T}$

$$\begin{aligned} R_T^{BO} &= \mathbb{E}\left[\sum_{t=1}^{T}\{f_{a_t}(x^*_{a_t}) - f_{a_t}(x_t)\}\middle|\mathcal{D}_t, \mathcal{GP}_1, \ldots, \mathcal{GP}_m\right] \\ &= \sum_{a=1}^{m}\sum_{t_a=1}^{T_a}\mathbb{E}\left[f_a(x^*_a) - f_a(x_{t_a})|\mathcal{D}_t, \mathcal{GP}_a\right] \qquad (18) \\ &\le b\sqrt{mT^{\alpha+1}\log T} \qquad (19)\end{aligned}$$

where b is an arbitrary constant, $0 \le \alpha < 1$.

Combining Lemmas 2 and 3, we have the Bayesian regret bound in Theorem 1. Then we have $\lim_{T\to\infty} BayesRegret(T)/T = 0$. The expected cumulative regret is sublinear and the average expected regret of the proposed model is convergent.

IV. Illustration Through Benchmark Example

In this section, we present two benchmark examples to compare the performance of the proposed MAB-MGP-BO method with other three Bayesian Optimization methods.

A. Methods Considered

To show the advantage of our methods, we compared the following four methods on the simulated functions:

1) *Onehot-BO [54]:* BO method using one-hot encoding method to transform the categorical variable into additional continuous variables.
2) *MAB-BO [30]:* Multiarmed bandit Bayesian Optimization method using traditional univariate output Gaussian process in each arm.
3) *MAB-SMGP-BO:* This method is the same as our proposed MAB-MGP-BO method except that a separable covariance functions for the MGP is used. We adopt the separable covariance functions developed in [38]. In this approach, one positive definite matrix with unit diagonal elements (PDUDE) is constructed to measure the correlation among the functions from different arms. A spherical coordinate parameterization method is used to simplify the fitting algorithm for the hyperparameters of the PDUDE matrix.
4) *MAB-MGP-BO:* (Our proposed model in Section III-B).

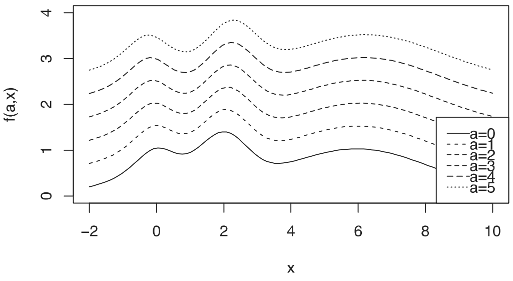


Fig. 4. 2-D synthetic function used in simulation.

B. Simulated Function

In this simulation, we take the same synthetic functions used in [30]. One of the function we try to maximize is a two-dimensional (2-D) function containing one continuous variable x and a categorical variable a. The expression of the function is shown in (20) and Fig. 4

$$\begin{aligned} f([a,x]) &= \exp(-(z_1-2)^2) + \exp\left(\frac{(-z_1-6)^2}{10}\right) \\ &\quad + \frac{1}{z_2^2+1} + \frac{a}{2} \qquad (20)\end{aligned}$$

where $z_i = x + 0.05a(-1)^i, x \in [-2, 10], a = 0, 1, \ldots, 5$.

We can see the set of functions follow a similar pattern and the maximum value are obtained with similar value of x. There are positive correlations between different functions and the proposed algorithms can take advantage of information from other functions to predict each function value. There are also fluctuations in the functions which are common properties of real world functions.

The second simulated function contains four continuous variables $\mathbf{x}$ and one categorical variable a. The expression of the function is as follows:

$$f([a, \mathbf{x}]) = \prod_{i=1}^{4}\sqrt{z_i}\sin(z_i) + 2a \qquad (21)$$

where $z_i = x_i + 2a, \mathbf{x} \in [1, 10]^4, a = 0, 1, \ldots, 5$.

Though there are five dimensions of continuous variables and we cannot present a figure to show the similarities of the functions, we can see from the expression they share similar patterns and have positive correlations. The simulations are both validated ten times and the mean of the best function values are shown in Fig. 5. We can see from both function optimizations that our proposed method can not only achieve a higher maximum value but also converge faster than the other three methods. Among the other three methods, MAB-BO model produces a higher optimal value, while the MAB-SMGP-BO

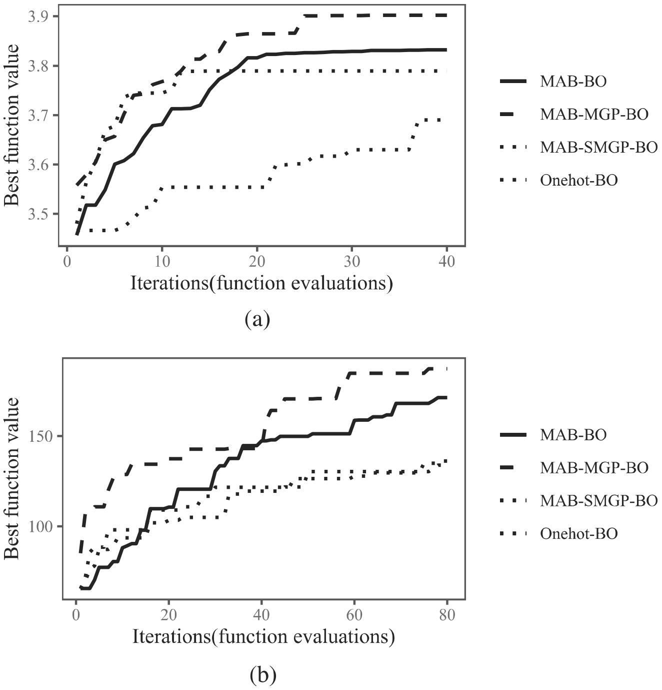


Fig. 5. Comparison of our proposed method and other methods.

model has a faster convergence speed in the earlier iterations. This result may be because the MAB-SMGP-BO model has more data information than the MAB-BO model in the earlier iterations and the fitting and prediction of the Gaussian process is faster at the beginning. However, as the model accumulates enough information in the later iterations, the separable structure of the MAB-SMGP-BO method restricts the fitting and cannot achieve a higher optimal value than the MAB-BO method does. The proposed MAB-MGP-BO method, on the other hand, is not restricted by the separable covariance structure and can fit a better Gaussian process of the underlying function and obtain the best optimum in a faster rate.

V. Application to Structural Damage Diagnosis

In this section, we apply the proposed MAB-MGP-BO method for a real-world structural damage identification using piezoelectric admittance measurement [8]. We will first give the problem description in Section V-A. Then, a simulation study is conducted in Section V-B. Finally, the structural damage identification using real experimental data is shown in Section V-C.

A. Problem Description

As shown in Fig. 6, a piezoelectric transducer is attached to the host structure and the dynamic response of the structure in terms of structural admittance at different frequencies can be measured and observed. To build a finite element analysis (FE) model, the host structure is divided into 11 250 elements. It is then divided into 25 segments, which are regarded as 25 locations for damage identification. With the FE model, we can link the structural damage, which is often modeled as a property change such as the stiffness loss at a specific element, with the observed dynamic response. Now the structural damage identification problem can also be considered as a black-box function optimization problem with categorical and continuous inputs, where the FE model is the black-box function, the location (i.e., segment) of the damage is the categorical input, and the severity of the damage is the continuous input. In structural damage identification, we try to find the location and severity of the damage by minimizing the difference between the FE model prediction and the experimental measurements. As presented in Fig. 6, the PZT segments are located consecutively in a line and it is intuitive to assume there are correlations between adjacent segments. If the damage locations are near to each other, the admittance change should follow similar patterns.

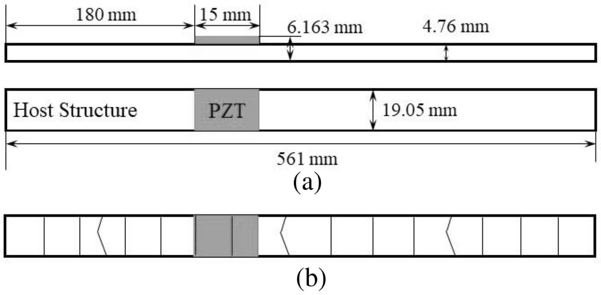


Fig. 6. (a) Dimensions of the structure and PZT patch. (b) Diagram of division of segments.

We can consider the difference of the observed and the computed admittance change $||\Delta I_{\text{exp}} - \Delta I_{\text{FE}}||$ as the objective function we need to optimize. The location and severity of the damage $[l, s]$ is the input of the model, where $l \in \{1, \ldots, m\}$ is the categorical variable and $s \in \mathcal{S} \subset \mathbb{R}$ is the continuous variable. To identify the damage, we need to find

$$[L^*, s^*] = \text{argmin}_{[l,s]}||\Delta I_{\text{exp}} - \Delta I_{\text{FE}}(l, s)||. \tag{22}$$

In this case study, the FE model of the host structure contains $m = 25$ segments. To make comparison, we consider the same cases of damages as that considered in [8]. In admittance-based damage detection, the damage occurrence causes admittance changes around resonant peaks. Without loss of generality, we pick two frequencies, 14th (1893.58 Hz) and 21st (3704.05 Hz) natural frequencies, to conduct frequency sweeping. Two frequency ranges around the two natural frequencies (from 1891.69 to 1895.47 Hz and from 3700.35 to 3707.75 Hz) are used in the inverse analysis. The experimental setup is shown in Fig. 7. The data points for the experimental measurements contained in each frequency range are detailed in Table I. The voltage drop is measured across a small resistor $R = 100\Omega$ which is connected in serial to the transducer. And the current in the circuit can be obtained which then yields the admittance information. A Dynamic signal analyzer (Agilent 35670 A) with a source channel and the sweep sine capability is utilized. The source channel is used to generate the sinusoidal voltage V_{in} sent to the piezoelectric transducer, and the output voltage V_{out} across the resistor is recorded.

A small mass block is introduced to emulate the damage, as shown in the Fig. 7. The damage is introduced under assumption

TABLE I
EXPERIMENTAL CASES CONSIDERED

Experiment Case	Segment	Severity	Frequency Range (Hz)	# of Frequency Points
Case I	12	0.0016	[1891.69,1895.47]	100
Case II	14	0.0028	[3700.35,3707.75]	85

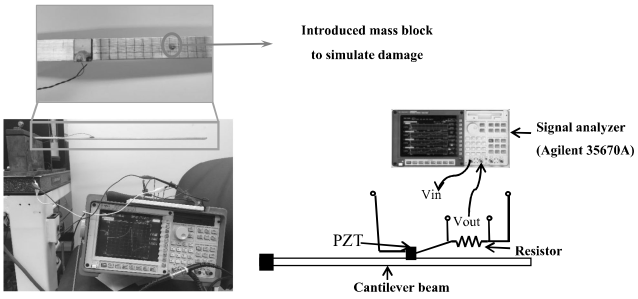
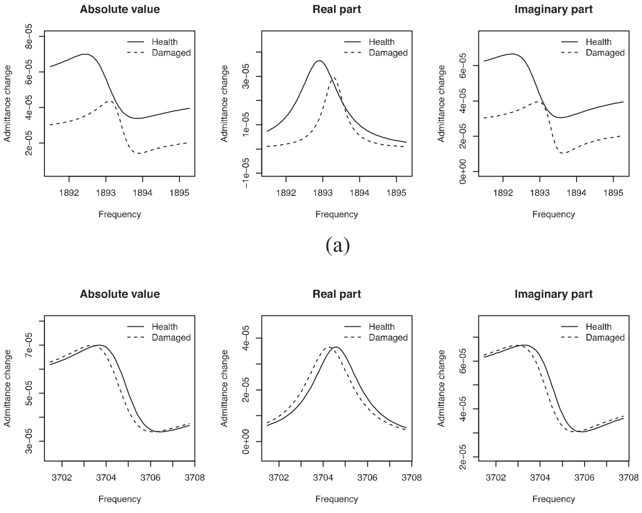


Fig. 7. Experimental setup.



(a)

(b)

Fig. 8. Admittance change of the health and damaged structure.

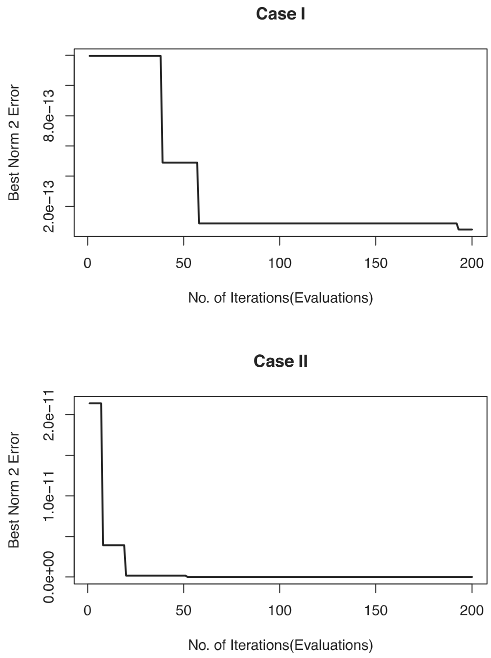


Fig. 9. Performance on the simulated structural damage identification.

that the mass of the whole system now is unchanged, thus resulting in equivalent stiffness reduction. In Case I, the damage locates on the 12th segment and the severity is equivalent to a stiffness loss of 0.16%. In Case II, the real damage locates on the 14th segment and the severity is equivalent to a stiffness loss of 0.28%.

B. Structural Damage Identification Using Simulated Observations

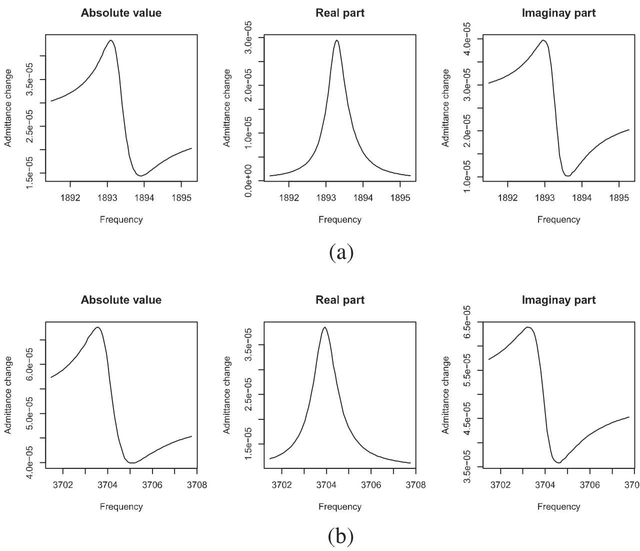
First, we take the simulated admittance as shown in Fig. 8 from the FE model as if they were the true observed admittance. The figures of each case are the absolute value, the real part and the imaginary part of the complex admittance, respectively. We can see there is an overlay of admittance with and without damage in the figures. We attempt to locate the damage location and severity of the structure based on the admittance change.

Since there is no noise and other uncertainties in the simulated observations, we expect the proposed MAB-MGP-BO method can identify the damage quickly and accurately. Indeed, as shown in Fig. 9, MAB-MGP-BO method identifies the correct damage location and has a very close damage severity estimation to the underlying damage severity. The comparison with the multi-DIRECT method proposed in [8] is shown in Table II. Our method can achieve more accurate estimation of the damage severity, which confirms the significance of nonlinearity in the model.

Other three models are also applied on this simulation study, but there are some limitations. Because all the three other methods presented in the numerical study showed too bad performance, we did not include the results. For the MAB-SMGP-BO model, when the number of arms becomes large (25 in this case), the number of hyperparameters in the correlation matrix becomes $25^2 = 625$ in each iteration of Gaussian process fitting. This model takes too long time to complete, and it is not feasible on the simulation case. For the MAB-BO model, the convergence is relatively slow because the Bayesian optimization only considers the information of the current arm. For the Onehot-BO method, similar problem happens to the model convergence. The transformed data are too sparse for the Gaussian process to fit. Neither of the two methods can locate the correct damaged location within 200 iterations.

TABLE II
RESULTS COMPARISON OF TWO MODELS

True damage [location, severity]	Model	Prediction	Estimation error
[12, 0.0016]	MAB-MGP-BO	[12, 0.00159]	**0.625%**
	Multi-DIRECT	[12, 0.0017]	6.25%
[14, 0.0028]	MAB-MGP-BO	[14, 0.00277]	**1.074%**
	Multi-DIRECT	[14, 0.003]	7.14%



(a)

(b)

Fig. 10. Experimental observations of Admittance change for Case I and Case II.

C. Structural Damage Identification Based on Real Observations

In this section, we conduct the structural damage identification based on real observations as shown in Fig. 10. Structural damage identification based on real observed responses is more challenging. First, there are always noises in the real observations. The measurement noise will not only make the damage severity estimation less accurate, but also possibly cause alias in the damage location. In other words, it is possible that other damage locations have closer simulated admittance change to the real observations. These noises may lead to the incorrect damage identification results. Second, the FE model itself is not perfect. There will always be modeling errors which will lead to bias in the damage identification. As a result, we generally cannot guarantee that the model based structural damage identification using Bayesian Optimization can always find the unique optimal solution. To avoid these difficulties in structural damage identification using real observations, we provide an extended algorithm that can give a set of possible solutions instead of a single solution. The algorithm we used to generate the group of optimal results are shown in Algorithm 3. The basic idea is simple: once we identify the current best solution, we remove it from the solution space and then find the best solution in the rest solution space. The result shows that the correct damage locations are included in the first several solutions.

Algorithm 3: Optimal results generation.

```
$t$: number of iterations, $tol$: largest number of iterations of one optimal result.
$\alpha_t^*$: optimal arm in iteration $t$. $s_t^*$: optimal severity in iteration $t$.
$\mathcal{G}$: Group of optimal results.
for $t = 1, 2, \ldots$ do
    if $r == tol$ then
        Save the optimal result. $\mathcal{G} = \mathcal{G} \bigcup \{[\alpha_t^*, s_t^*]\}$
        Remove all the points from the temporal optimal arm $\alpha_{t-1}^*$.
        $r = 1$. Restart the count of number of iterations with the same optimal result.
    end if
    Run the MAB-MGP-BO model and find the optimal arm $\alpha_t^*$.
    if $\alpha_t^* == \alpha_{t-1}^*$ then
        $r = r + 1$.
    end if
end for
```

TABLE III
OPTIMAL RESULTS: EXPERIMENTAL CASE I

Optimal results	Residuals	Damage [location,severity]
1	5.99337×10^{-8}	[23, 0.00109]
2	6.30363×10^{-8}	[12, 0.00181]
3	6.17052×10^{-8}	[9, 0.00140]

TABLE IV
OPTIMAL RESULTS: EXPERIMENTAL CASE II

Optimal results	Residuals	Damage [location,severity]
1	6.57408×10^{-8}	[23, 0.00209]
2	6.593299×10^{-8}	[24, 0.00209]
3	6.96688×10^{-8}	[25, 0.00217]
4	6.593457×10^{-8}	[9, 0.00299]
5	6.595832×10^{-8}	[14, 0.00298]

The first several best results we found using MAB-MGP-BO for the two cases are shown in Tables III and IV, respectively. For both cases, the underlying true damage location are successfully detected in the first several solutions. Considering the noise of measurement and the bias of the FE model, this result is satisfactory and we can expect to utilize the proposed model in practical applications.

VI. CONCLUSION

In this article, we propose a multiarmed bandit method using Bayesian Optimization with multioutput Gaussian process to solve Structural Health Monitoring problems, especially in complex systems. This method can be also applied to other fields of engineering applications, such as design optimization with complicated structures and hyperparameter tuning of surrogate models.

The proposed method utilizes the information of the data collected under all the categorical input values to make prediction of the black-box function under a specific categorical input value. We also presented the convergence analysis of the proposed method and provide a Bayesian regret bound of the algorithm. Numerical benchmark examples are conducted to show the proposed model has a better performance than several other existing Bayesian Optimization methods. We also apply our model to a real world case study on structural damage identification. The results show the proposed method can identify the location and the severity of the damage with a better performance than the existing fault diagnostics multi-DIRECT algorithm.

In addition to the diagnosis problem, the mixed input Bayesian optimization strategy proposed in this article can also be applied to many design problems, where both categorical inputs (e.g., material selection, geometry configuration selection) and continuous inputs influence the design performance. The proposed strategy can utilize the information under different categorical inputs and achieve the optimal design within the design space spanned by both categorical and continuous variables.

A limitation of MAB-MGP-BO method is that the computation load is heavy when the number of categories is very large. A MGP is fitted in each category of the data and it is hard to find a way to simplify the computation because of the complexity of the mathematical formulation. Though the data we fit the MGP to for different categories are the same, the assumption and the category of interest and the calculated covariance matrix of fitted categories cannot be reused. Other kinds of constructions of MGP can be explored to make the algorithm more scalable. We will extend the algorithm in this direction and report the findings in the near future.

Data Availability Statement: Data supporting the findings of this study are available from the corresponding author on request.

Appendix
Proof of Convergence

A. Lemmas Used

Following Lemmas 4 and 5 will be used to prove Lemmas 2 and 3.

Lemma 4: The Bayesian regret of the T_ath iterations of the ath arm has an upper bound of $\mathcal{O}(\sqrt{T_a^{\alpha+1}\log T_a})$.

$$\begin{aligned} BayesRegret(T_a) &= \mathbb{E}\left[\sum_{t_a=1}^{T_a}\left\{f_{a^*}(x^*) - f_{a_{t_a}}(x_t)\right\}\middle|\mathcal{D}_{T_a}\right] \\ &\le \mathcal{O}\left(\sqrt{\gamma_{T_a} T_a \log T_a}\right) && (23) \\ &\le \mathcal{O}\left(\sqrt{T_a^{\alpha+1}\log T_a}\right) && (24) \end{aligned}$$

where γ_{T_a} is the maximum information gain about $f_a(x)$ after T_a iterations.

The boundary in (23) is proved by [31], [52]. Using the assumption 2 (i.e., $\gamma_{T_a} \sim \mathcal{O}(T_a^{\alpha})$), the boundary of (24) can be obtained.

Bayesian simple regret is defined for the Bayesian regret of individual function rather than summation over time. Then, the boundary of the Bayesian simple regret can be obtained as follows.

Lemma 5: The Bayesian regret of individual function by the T_ath iteration has an upper bound of $\mathcal{O}(\sqrt{\frac{\log T_a}{T_a^{1-\alpha}}})$

$$\begin{aligned} BayesSimpleRegret(T_a) &= \mathbb{E}\left[f_{a^*}(x^*) - \max_{t\le T_a} f_{a_t}(x_t)\middle|\mathcal{D}_{T_a}\right] \\ &\le \mathcal{O}\left(\sqrt{\frac{\gamma_{T_a}\log T_a}{T_a}}\right) \\ &\le \mathcal{O}\left(\sqrt{\frac{\log T_a}{T_a^{1-\alpha}}}\right). && (25) \end{aligned}$$

B. Proof of Lemma 2

Proof:

$$\begin{aligned} R_T^{MAB} =& \mathbb{E}\left[\sum_{t=1}^{T}\left\{f_{a^*}(x^*) - f_{a_t}(x^*_{a_t})\right\}\middle|\mathcal{D}_T, \mathcal{GP}_1, \ldots, \mathcal{GP}_m\right] \\ =& \mathbb{E}\left[\sum_{t=1}^{T}\left\{f_{a^*}(x^*) - U_t(a^*)\right\}\middle|\mathcal{D}_T, \mathcal{GP}_1, \ldots, \mathcal{GP}_m\right] \\ &+\mathbb{E}\left[\sum_{t=1}^{T}\left\{U_t(a_t) - f_{a_t}(x^*_{a_t})\right\}\middle|\mathcal{D}_T, \mathcal{GP}_1, \ldots, \mathcal{GP}_m\right] \end{aligned} \quad (26)$$

where $U_t(a)$ is an upper confidence bound that is a determined function of a. a_t is a random variable selected from the posterior sampling. Here, our function $f_a(x)$ is a random variable, and we do not know the true function $f_a(x)$. Therefore, we select the a_t based on the posterior samples of functions from each arm. In particular, we select a_t by $\max_{a,x}\tilde{f}_a(x)$, where $\tilde{f}_a(x)$ is posterior samples given $\mathcal{D}$. Because posterior sampling $\tilde{f}_a(x)|\mathcal{D}, \mathcal{GP}_a$ is precise, we can claim that the distributions $p(a^*)$ and $p(a_t|\mathcal{D}, \mathcal{GP}_1, \ldots, \mathcal{GP}_m)$ are identically distributed. This argument is fundamental in the Bayesian regret proof of Thompson sampling used in [52]; since the Thompson sampling at each arm precisely uses the posterior distribution to propose a_t at iteration t, both a_t and a^* are identically distributed conditioned on $\mathcal{D}_T$ (26).

We select $U_t(a)$ as follows.

$$U_t(a) = \mathbb{E}\left[\max_{t'\le t_a} f_a(x^a_{t'}) + a\sqrt{\frac{\log t_a}{t_a^{1-\alpha}}}\middle|\mathcal{D}_t, \mathcal{GP}_a\right] \quad (27)$$

where t_a is the number of times that a is selected during t times of total iterations, and a is an arbitrary constant. Then, the first term in (26) is nonpositive by the definition of (27), that is,

$$\mathbb{E}\left[\sum_{t=1}^{T}\{f_{a^*}(x^*)-U_t(a^*)\}\middle|\mathcal{D}_T,\mathcal{GP}_1,\ldots,\mathcal{GP}_m\right]\le 0.$$

The second term in (26)

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T}\{U_t(a_t)-f_{a_t}(x^*_{a_t})\}\middle|\mathcal{D}_T,\mathcal{GP}_1,...,\mathcal{GP}_m\right]\\
&\le\mathbb{E}\left[\sum_{t=1}^{T}\{U_t(a_t)-L_t(a_t)\}\middle|\mathcal{D}_T,\mathcal{GP}_1,...,\mathcal{GP}_m\right]\\
&=\sum_{a=1}^{m}\mathbb{E}\left[\sum_{t_a=1}^{T_a}\{U_{t_a}(a)-L_{t_a}(a)\}\middle|\mathcal{D}_T,\mathcal{GP}_a\right]\\
&=a\sum_{a=1}^{m}\sum_{t_a=1}^{T_a}\sqrt{\frac{\log t_a}{t_a^{1-\alpha}}}\\
&\le a\sqrt{\log T}\sum_{a=1}^{m}\sum_{t_a=1}^{T_a}\frac{1}{\sqrt{t_a^{1-\alpha}}}\\
&\le 2a\sqrt{mT^{\alpha+1}\log T}\end{aligned}\tag{28}$$

where $L_t(a_t)$ is defined as $L_t(a_t)=\mathbb{E}[\max_{t'\le t_a}f_a(x^a_{t'})|\mathcal{D}_t]$ because $L_t(a_t)\le\mathbb{E}[f_{a^*}(x^*)|\mathcal{D}_t]$. Here, again, we sum the Bayesian regret bound obtained for each arm. Therefore, the Bayesian regret bound of R_T^{MAB} is (17). ■

C. *Proof of Lemma 3*

Proof:

$$\begin{aligned}R_T^{BO}&=\mathbb{E}\left[\sum_{t=1}^{T}\{f_{a_t}(x^*_{a_t})-f_{a_t}(x_t)\}\middle|\mathcal{D}_t\right]\\
&=\mathbb{E}\left[\sum_{a=1}^{m}\sum_{t_a=1}^{T_a}\{f_a(x^*_a)-f_a(x_{t_a})\}\middle|\mathcal{D}_t\right]\\
&\le\mathbb{E}\left[\sum_{a=1}^{m}b\sqrt{T_a^{\alpha+1}\log T_a}\middle|\mathcal{D}_t\right]\\
&\le b\sqrt{mT^{\alpha+1}\log T}\end{aligned}\tag{29}$$

where b is an arbitrary constant. ■

REFERENCES

[1] A. He and X. Jin, "Deep variational autoencoder classifier for intelligent fault diagnosis adaptive to unseen fault categories," *IEEE Trans. Rel.*, vol. 70, no. 4, pp. 1581–1595, Dec. 2021.

[2] L. Guo, N. Li, F. Jia, Y. Lei, and J. Lin, "A recurrent neural network based health indicator for remaining useful life prediction of bearings," *Neurocomputing*, vol. 240, pp. 98–109, 2017.

[3] Y. Liu, Q. Shuai, S. Zhou, and J. Tang, "Prognosis of structural damage growth via integration of physical model prediction and Bayesian estimation," *IEEE Trans. Rel.*, vol. 66, no. 3, pp. 700–711, Sep. 2017.

[4] H. Hanachi, C. Mechefske, J. Liu, A. Banerjee, and Y. Chen, "Performance-based gas turbine health monitoring, diagnostics, and prognostics: A survey," *IEEE Trans. Rel.*, vol. 67, no. 3, pp. 1340–1363, Sep. 2018.

[5] J.-T. Kim and N. Stubbs, "Crack detection in beam-type structures using frequency data," *J. Sound Vib.*, vol. 259, no. 1, pp. 145–160, 2003.

[6] P. Cao, D. Yoo, Q. Shuai, and J. Tang, "Structural damage identification with multi-objective direct algorithm using natural frequencies and single mode shape," in *Proc. Health Monit. Struct. Biol. Syst.*, 2017, vol. 10170, Art. no. 101702H.

[7] P. Lall, R. Lowe, and K. Goebel, "Extended Kalman filter models and resistance spectroscopy for prognostication and health monitoring of leadfree electronics under vibration," *IEEE Trans. Rel.*, vol. 61, no. 4, pp. 858–871, Dec. 2012.

[8] P. Cao, S. Qi, and J. Tang, "Structural damage identification using piezoelectric impedance measurement with sparse inverse analysis," *Smart Mater. Struct.*, vol. 27, no. 3, 2018, Art. no. 035020.

[9] J. Kim and K. Wang, "An enhanced impedance-based damage identification method using adaptive piezoelectric circuitry," *Smart Mater. Struct.*, vol. 23, no. 9, 2014, Art. no. 095041.

[10] J. E. Michaels and T. E. Michaels, "Guided wave signal processing and image fusion for in situ damage localization in plates," *Wave Motion*, vol. 44, no. 6, pp. 482–492, 2007.

[11] D. A. T. Burgos, R. C. G. Vargas, C. Pedraza, D. Agis, and F. Pozo, "Damage identification in structural health monitoring: A brief review from its implementation to the use of data-driven applications," *Sensors*, vol. 20, no. 3, p. 733, 2020.

[12] Q. Shuai, K. Zhou, S. Zhou, and J. Tang, "Fault identification using piezoelectric impedance measurement and model-based intelligent inference with pre-screening," *Smart Mater. Struct.*, vol. 26, no. 4, 2017, Art. no. 045007.

[13] J. Min, S. Park, C.-B. Yun, C.-G. Lee, and C. Lee, "Impedance-based structural health monitoring incorporating neural network technique for identification of damage type and severity," *Eng. Struct.*, vol. 39, pp. 210–220, 2012.

[14] R. Perera, S.-E. Fang, and A. Ruiz, "Application of particle swarm optimization and genetic algorithms to multiobjective damage identification inverse problems with modelling errors," *Meccanica*, vol. 45, no. 5, pp. 723–734, 2010.

[15] S. Seyedpoor, S. Shahbandeh, and O. Yazdanpanah, "An efficient method for structural damage detection using a differential evolution algorithm-based optimisation approach," *Civil Eng. Environ. Syst.*, vol. 32, no. 3, pp. 230–250, 2015.

[16] Y. Wang and H. Hao, "Damage identification scheme based on compressive sensing," *J. Comput. Civil Eng.*, vol. 29, no. 2, 2015, Art. no. 04014037.

[17] Y. Huang, J. L. Beck, and H. Li, "Bayesian system identification based on hierarchical sparse Bayesian learning and Gibbs sampling with application to structural damage assessment," *Comput. Methods Appl. Mech. Eng.*, vol. 318, pp. 382–411, 2017.

[18] D. R. Jones, C. D. Perttunen, and B. E. Stuckman, "Lipschitzian optimization without the lipschitz constant," *J. Optim. Theory Appl.*, vol. 79, no. 1, pp. 157–181, 1993.

[19] J. Močkus, "On Bayesian methods for seeking the extremum," in *Proc. Optim. Techn. IFIP Tech. Conf.*, 1975, pp. 400–404.

[20] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 2951–2959, 2012.

[21] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, "Taking the human out of the loop: A review of Bayesian optimization," *Proc. IEEE*, vol. 104, no. 1, pp. 148–175, Jan. 2016.

[22] J. T. Springenberg, A. Klein, S. Falkner, and F. Hutter, "Bayesian optimization with robust Bayesian neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29, pp. 4134–4142.

[23] R. Lam, M. Poloczek, P. Frazier, and K. E. Willcox, "Advances in Bayesian optimization with applications in aerospace engineering," in *Proc. AIAA Non-Deterministic Approaches Conf.*, 2018, Art. no. 1656.

[24] S. Cakmak, D. Wu, and E. Zhou, "Solving Bayesian risk optimization via nested stochastic gradient estimation," *IISE Trans.*, to be published, doi: 10.1080/24725854.2020.1869352.

[25] H. Wang, J. Yuan, and S. H. Ng, "Gaussian process based optimization algorithms with input uncertainty," *IISE Trans.*, vol. 52, no. 4, pp. 377–393, 2020.

[26] M. Kim and K. Liu, "A Bayesian deep learning framework for interval estimation of remaining useful life in complex systems by incorporating general degradation characteristics," *IISE Trans.*, vol. 53, no. 3, pp. 326–340, 2020.

[27] S. Jahani, S. Zhou, D. Veeramani, and J. Schmidt, "Multioutput Gaussian process modulated poisson processes for event prediction," *IEEE Trans. Rel.*, vol. 70, no. 4, pp. 1569–1580, Dec. 2021.

[28] A. D. Bull, "Convergence rates of efficient global optimization algorithms," *J. Mach. Learn. Res.*, vol. 12, no. 10, pp. 2879–2904, 2011.

[29] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyperparameter optimization," in *Proc. 25th Annu. Conf. Neural Inf. Process. Syst.*, pp. 2546–2554, 2011, vol. 24.

[30] D. Nguyen, S. Gupta, S. Rana, A. Shilton, and S. Venkatesh, "Bayesian optimization for categorical and category-specific continuous inputs," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 5256–5263. [Online]. Available: https://doi.org/10.1609/aaai.v34i04.5971

[31] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3250–3265, May 2012. [Online]. Available: https://doi.org/10.1109/tit.2011.2182033

[32] J. Wiebe, I. Cecílio, J. Dunlop, and R. Misener, "A robust approach to warped Gaussian process-constrained optimization," *Math. Program.*, pp. 1–35, 2022.

[33] Z. Xu, Y. Guo, and J. H. Saleh, "Accurate remaining useful life prediction with uncertainty quantification: A deep learning and nonstationary Gaussian process approach," *IEEE Trans. Rel.*, vol. 71, no. 1, 2022, pp. 443–456.

[34] H. Wang, B. van Stein, M. Emmerich, and T. Back, "A new acquisition function for Bayesian optimization based on the moment-generating function," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2017, pp. 507–512.

[35] I. Frazier, "Bayesian optimization," in *Recent advances optim. model. contemporary problems informs*, 2018, pp. 225–.

[36] S. Conti and A. O'Hagan, "Bayesian emulation of complex multi-output and dynamic computer models," *J. Statist. Plan. Inference*, vol. 140, no. 3, pp. 640–651, 2010.

[37] P. Z. G. Qian, H. Wu, and C. J. Wu, "Gaussian process models for computer experiments with qualitative and quantitative factors," *Technometrics*, vol. 50, no. 3, pp. 383–396, 2008.

[38] Q. Zhou, P. Z. Qian, and S. Zhou, "A simple approach to emulation for computer models with qualitative and quantitative factors," *Technometrics*, vol. 53, no. 3, pp. 266–273, 2011.

[39] B. Ru, A. Alvi, V. Nguyen, M. A. Osborne, and S. Roberts, "Bayesian optimisation over multiple continuous and categorical inputs," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, vol. 119, pp. 8276–8285.

[40] C. Oh, E. Gavves, and M. Welling, "Mixed variable Bayesian optimization with frequency modulated kernels," in *Proc. 37th Conf. Uncertainty Artif. Intell.*, 2021, vol. 161, pp. 950–960.

[41] J. R. Gardner, G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson, "GPyTorch: Blackbox matrix-matrix gaussian process inference with GPU acceleration," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7587–7597.

[42] A. Majumdar and A. E. Gelfand, "Multivariate spatial modeling for geostatistical data using convolved covariance functions," *Math. Geol.*, vol. 39, no. 2, pp. 225–245, 2007.

[43] T. E. Fricker, J. E. Oakley, and N. M. Urban, "Multivariate Gaussian process emulators with nonseparable covariance structures," *Technometrics*, vol. 55, no. 1, pp. 47–56, 2013.

[44] A. Melkumyan and F. Ramos, "Multi-kernel Gaussian processes," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, vol. 2, pp. 1408–1413.

[45] J. M. Ver Hoef and R. P. Barry, "Constructing and fitting models for cokriging and multivariable spatial prediction," *J. Statist. Plan. Inference*, vol. 69, no. 2, pp. 275–294, 1998.

[46] G. B. Arfken and H.-J. Weber, *Mathematical Methods for Physicists*. Orlando, FL, USA: Academic Press, 1967.

[47] P. Boyle and M. Frean, "Dependent Gaussian processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 217–224.

[48] M. Alvarez and N. D. Lawrence, "Sparse convolved Gaussian processes for multi-output regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 57–64.

[49] R. Kontar, S. Zhou, C. Sankavaram, X. Du, and Y. Zhang, "Nonparametric modeling and prognosis of condition monitoring signals using multivariate Gaussian convolution processes," *Technometrics*, vol. 60, no. 4, pp. 484–496, 2018.

[50] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, 2012. [Online]. Available: https://doi.org/10.1561/2200000024

[51] I. O. Ryzhov, "On the convergence rates of expected improvement methods," *Oper. Res.*, vol. 64, no. 6, pp. 1515–1528, 2016.

[52] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Math. Oper. Res.*, vol. 39, no. 4, pp. 1221–1243, 2014.

[53] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[54] D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. Karro, and D. Sculley, "Google vizier: A service for black-box optimization," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 1487–1495.