IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 23, 2020, accepted June 11, 2020, date of publication June 16, 2020, date of current version June 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3002819

An Active Learning Methodology for Efficient Estimation of Expensive Noisy Black-Box Functions Using Gaussian Process Regression

RAJITHA MEKA[1], ADEL ALAEDDINI[1], SAKIKO OYAMA[2], AND KRISTINA LANGER[3]

[1]Department of Mechanical Engineering, The University of Texas at San Antonio, San Antonio, TX 78249, USA
[2]Department of Kinesiology, Health, and Nutrition, The University of Texas at San Antonio, San Antonio, TX 78249, USA
[3]Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH 454331, USA

Corresponding author: Adel Alaeddini (adel.alaeddini@utsa.edu)

This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-16-1-0171.

ABSTRACT Estimation of black-box functions often requires evaluating an extensive number of expensive noisy points. Learning algorithms can actively compare the similarity between the evaluated and unevaluated points to determine the most informative subsequent points for efficient estimation of expensive functions in a sequential procedure. In this paper, we propose an active learning methodology based on the integration of Laplacian regularization and active learning - Cohn (ALC) measure for identification of the most informative points for efficient estimation of noisy black-box functions using Gaussian processes. We propose two simple greedy search algorithms for sequential optimization of the tuning parameters and determination of subsequent points based on the information from the previously evaluated points. We also enhance the graph Laplacian with the information of both the predictor and response variables to capture the similarity between the points more effectively. The proposed methodology is particularly suited for problems involving estimation of expensive black-box functions with a high level of noise and plenty of unevaluated points. Using a case study for analysis of the kinematics of pitching in baseball as well as simulation experiments, we demonstrate the performance of the proposed methodology against existing methods in the literature in terms of estimation error.

INDEX TERMS Active learning, Gaussian process regression, kernel ridge regression, Laplacian regularization.

I. INTRODUCTION

In many real-world problems, we encounter situations involving estimation of expensive noisy black-box functions. These problems typically require a large number of evaluations that could take hours or days for evaluating one single point [1], [2]. Traditional response surface methods [3], which are based on simple parametric models, may not properly approximate these expensive black-box functions. Surrogate modeling based on Gaussian process (GP), which can be viewed as an extension of standard regression models, is one of the most popular non-parametric probabilistic models for estimating black-box functions [4]. GP has key advantages over most estimation methods, which includes: (1) ability to fit highly nonlinear functions with minimal risk of overfitting, (2) built-in capability for uncertainty estimation and quantification, and (3) small estimation bias [5], [6]. GP is widely used in a variety of fields. It was first used for time series analysis in the 1880s by astronomer Thiele [7] and then in 1940s in Wiener-Kolmogorov prediction theory [8], [9]. The use of GP in geo-statistics where it is referred to as Kriging dates back to 1960s, where it is used to approximate the function to determine the optimum location for mining exploration [10]. Since then, it has been applied to address statistical problems and widely being used in spatial statistics. In statistics, it is one of the well-known approaches for modeling and optimization of expensive functions like complex computer models or codes [11]. In the machine

The associate editor coordinating the review of this manuscript and approving it for publication was Yuan Zhuang.

learning community, the usage of GP is first explained by Williams and Rasmussen [4]. Neal [12], shows that the neural networks with infinite hidden units converge to GP. Various aspects of GP, including model selection and adaptation of hyperparameters, applications in regression and classification problems, and relationship with other estimation models are extensively discussed in the literature [13], [14].

A common application of GP is to emulate a physical system when experiments are too costly to perform or infeasible, i.e. wind tunnel testing [11], [15]. For such applications with high experimentation cost generally the labeled/evaluated data is very scarce. GP is typically built in a sequential process, starting with a small number of initial points based on a space-filling design, such as Latin hypercube design (LHD) [16], sphere packing [17], uniform designs [18], and then using a strategy to select the most promising points for the next iteration, until some convergence criterion is met. The selection of points should be in a way that additional points shall improve the information content of the data that describe the design space. Semi supervised learning approach is one area of research that seeks to exploit the information from both labeled and unlabeled data to improve the estimation of the underlying function [19]. This approach assumes the labeled data is given and fixed [20]. Then, active learning, a subsection of semi supervised learning came into play to use in conjunction with semi supervised learning to pick the next evaluation point that augments the evaluated data. Active learning, which uses both labeled data (evaluated points) and unlabeled data (unevaluated points) for efficient estimation of statistical models have been successfully applied to many problems and received a lot of attention in machine learning community during last few years [21]–[23]. Using the active learning scenario with GP, the selection process of evaluation point is such that it incorporates as much new information into the model as possible, after seeing the training data [24]. Assuming that the given model is correct, active learning - Mckay (ALM) and active learning - Cohn (ALC) are two popular algorithms for selecting the next evaluation points. ALM measure is based on entropy (or cross entropy) for selecting a point that maximizes the expected information gain [25]. It is similar to maximum entropy measure that is based on selecting the points with highest uncertainty. The entropy criterion tends to pick the points near to the boundary of the area of interest as the high uncertainty points are the ones that are far away from each other [26]. ALC on the other hand, tries to minimize the generalization error. It selects a point that reduces the average predictive variance at unevaluated points [24]. Seo *et al.* [27] show that Cohn's criterion of minimizing the average variance performs well with GP. Pasolli and Melgani [28] propose two active learning strategies for GP regression, one based on the distances in kernel space from samples in the training set and the other one is based on the variance. There are several other methods based on maximum entropy [29], integrated mean square error, maximum mean square error [11] that guide in selecting the promising subsequent evaluation point. Wu *et al.* discuss two active learning based greedy sampling approaches, greedy sampling on the output (GSy) that selects samples to increase the diversity of output space and improved greedy sampling approach (iGS) that selects sample to increase the diversity in both input and output spaces [30]. To our knowledge, GSy and iGS are the only approaches that consider the output space information in selecting the next evaluation point. However as the name suggests, these two approaches do not consider the model uncertainty and only select the samples greedily, which may result in high predictive variance (discussed in Section IV). Recently, Zhang *et al.* propose a graph based active learning (GBAL) approach to select the next evaluation point based on the uncertainty information using L_1 measure, which enables it to use any surrogate model [31].

Apart from estimation of expensive black-box functions, the sequential approach using GP is similar for optimization of the black-box functions as well. For optimization, although the objective is to find the global minimum or maximum of the function, after each functional evaluation the GP model is updated to improve the estimate of the underlying black-box function. One of the popular global optimization algorithms is efficient global optimization algorithm (EGO) that uses expected improvement (EI) acquisition function to efficiently explore and exploit the design space and find global optimization point of expensive black-box functions [32]. Similar to EI, there are several other acquisition functions to find the global optimum point by estimating the black-box function using GP [33]–[37]. GP is generally used as it represents the prediction and uncertainty of true function which is utilized in building the strategy for selection of subsequent evaluation point. Chen *et al.* [38] propose an over-complete basis surrogate method (OBSM) which uses linear combinations of over-complete bases to globally approximate the surface. Chen *et al.* [39] also develop a stochastic search variable selection (SSVS) method to derive the prediction uncertainty by specifying some priors on the coefficients of OBSM and generating the posterior samples followed by an algorithm similar to EGO for selecting the subsequent points. Vu *et al.* [1] discuss iterative construction of the surrogate models to reach the global solution of expensive black-box functions. There also exist a group of algorithms in sequential decision making for multi-armed bandit problem. Thompson sampling (TS) [40] and upper confidence bound (UCB) [41] algorithms are popular algorithms in solving exploration versus exploitation problems. Given a multi-armed bandit problem, TS selects an arm randomly according to its probability of being optimal [42], [43]. Kullback - Leibler upper confidence bound is one of the UCB methods in which informational upper confidence bounds are computed using Kullback-Leibler (KL) divergence [44]. Other popular bandit strategies and their empirical evaluation can be found in [45]. While TS and UCB methods provide considerable performance for problems with simple information structures, information direct sampling (IDS) is another approach to address complex information structure problems using mutual information measure [46]. Krause *et al.* [47] propose

various algorithms based on mutual information criterion to actively select best possible locations over the design space that is modeled as GP. When the hyperparameters of the covariance of the model are approximately known, the algorithm selects the set of locations, which maximizes the mutual information between evaluated points and unevaluated points. Ben-Gal and Caramanis propose a sequential design of experiments via dynamic programming that also uses the mutual information measure to optimize the location and the number of points to evaluate [48].

Later, semi supervised learning algorithms based on manifold regularization have been widely utilized to effectively exploit the information from unevaluated points [49]–[53]. Laplacian regularization is one of the popular manifold regularization techniques that uses graph Laplacian to determine the information of underlying manifold [54]–[57]. Laplacian regularization has been successfully applied to many classification and regression problems [58]–[61]. Specifically, Laplacian regularized optimal design of experiments has been successfully used for image retrieval and interactive video indexing [21], [22], [62]. The graph Laplacian uses similarity matrix for constructing the Laplacian matrix. Recently, Liu *et al.* [63] has proposed a structured optimal graph based sparse feature extraction method in which they replace the similarity matrix used in constructing the Laplacian matrix with structured optimal graph, to capture the local manifold information by adaptively modifying the graph. When the labeled data is scarce, Zhu *et al.* [20] proposes combining the semi supervised and active learning methodologies using Gaussian fields and harmonic functions. Later, they show how the Gaussian random fields and harmonic energy minimizing framework can be viewed as GP with covariance matrix derived from graph Laplacian [64]. Alaeddini *et al.* [2] propose an active learning methodology based on sequential Laplacian regularized V-optimal design of experiments for efficient estimation of the black box functions.

All the Laplacian regularized methods discussed above use classical graph Laplacian that only considers the information from input space. In our study, we propose to extend classical graph Laplacian to incorporate the information of both input and output space. We expect the proposed graph Laplacian to be suitable for most of semi-supervised graph-based learning algorithms that use classical graph Laplacian. We provide a sample result of comparing the classical graph Laplacian with the proposed graph Laplacian measure in the Appendix. The main goal of this study is to develop an active learning methodology based on sequential Laplacian regularized Gaussian process (SLRGP) for efficient estimation of expensive noisy black-box functions, which uses the information from not only evaluated points but also (abundant) cheap unevaluated points to determine the most informative settings to evaluate subsequently. The proposed methodology has two major contributions. First, it considers the intrinsic manifold structure of evaluated and unevaluated points based on a novel similarity measure which considers both predictor and response variables. Second, it provides a unified active learning framework for identification of the most informative points for the construction of GP in a principled manner. In many applications, the resources are usually limited, or the cost of evaluating the points is very high. Thus, the selection of informative points is very crucial for training a good statistical model. This framework increases the efficiency of the learning process which consequently reduces the number of required points and improves the estimation accuracy. The proposed methodology is most suited for applications involving efficient estimation of expensive black-box functions with a high level of noise and plenty of unevaluated points.

The organization of this paper is as follows. Section II presents the related works and preliminaries to the proposed methodology. Section III explains the proposed SLRGP for efficient estimation of expensive noisy black-box function and its core components. Section IV discusses a case study for analysis of the kinematics of pitching in baseball as well as simulated experiments evaluating the performance of SLRGP in comparison to some of the existing methods in the literature. Finally, Section V provides a summary and concluding remarks.

II. RELATED WORKS

In this paper, we develop an active learning methodology for identification of the most informative points for GP regression [4]. The proposed algorithm is based on the integration of several components: (1) space-filling design of experiments [65] for identifying the initial set of points, (2) an extension of active learning - Cohn (ALC) [24] criterion for identification of the subsequent points to evaluate, (3) active learning [24] regularization for leveraging the information of unevaluated points (4) GP regression for fitting the evaluated data, and (5) an extension of bilateral kernel for formulating the similarity between evaluated and unevaluated points. In this section, we provide a brief description of the major components of the proposed algorithm. Throughout the paper, we use $\boldsymbol{z}$ to denote the design vector of evaluated points, and x to denote the design vector of any (either evaluated or unevaluated) point. We also use m to denote the number of evaluated points, q to denote the number of unevaluated points, n to denote the number of all (evaluated and unevaluated) points, and d to denote the dimensionality of predictor variables (X).

A. SPACE-FILLING DESIGN OF EXPERIMENTS

Space-filling designs are often used in computer experiments because there is no cost for changing the factor levels and the focus can be on good coverage of the region instead of the number of levels it might produce [65]. Space-filling designs may also help to avoid the localized effects as they sample throughout the design space [66]. Latin hypercube design (LHD) is one of the most popular space filling designs first introduced by [67]. For creating a m point LHD, each of the d dimension in the design space D is divided into m equal intervals such that the design space consists of m^d identical cells. Then, the m points are assigned to the centers

of m^d cells [1]. Some of the other popular space filling designs include maximin distance design [17], and uniform design [18].

B. ACTIVE LEARNING PROBLEM

The generic problem of active learning is the following: Given a set of points $X = (x_1, x_2, \ldots, x_n)$ in $\mathbb{R}^d$, we would like to find a subset of points $Z = (z_1, z_2, \ldots, z_m) \subset X$ which contains the most information about the response variable. In other words, the points $z_i (i = 1, \ldots, m)$ can improve the estimation the most, if they are evaluated and used as training points [22]. In the remainder of the paper we consider Z to represent the set of evaluated points, U to represent the set of unevaluated points, and X to represent the set of all points including evaluated and unevaluated points ($X = Z + U$).

C. GAUSSIAN PROCESS REGRESSION

Having some observed input-output pairs (z_i, y_i) where y_i might be corrupted by some noise ϵ_i, GP defines a prior over an unknown link function f, and gives the posterior after seeing some data [14]. More specifically, the GP regression is defined as $y_i = f(z_i) + \epsilon_i$ for $i = 1, .., m$, where ϵ is the additive independent identically distributed Gaussian noise with variance σ_m^2. The functional evaluation at the test point $x \subset U$ is denoted as f_*. $Y = (y_1, y_2, .., y_m)^T$ is the observed outputs at training points $Z = (z_1, z_2, \ldots, z_m)$. According to the joint distribution of observed outputs and test output we have:

$$\begin{bmatrix} Y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(Z,Z) + \sigma_m^2 I & K(Z,x) \\ K(x,Z) & K(x,x) \end{bmatrix}\right) \quad (1)$$

where $K(Z,Z), K(Z,x), K(x,Z), K(x,x)$ are the covariance between the training and training points, training and test points, test and training points, test and test points respectively, and $K(.,.)$ is an appropriate kernel function to evaluate the covariance. Here, we consider the squared exponential kernel $K(z_i, z_j) = \sigma_f^2 \, exp(-\frac{\| z_i - z_j \|^2}{2l^2})$ where σ_f^2 the signal variance, and l the characteristic length scale are two hyper-parameters of the kernel. Let, $K(Z,Z) = K_{ZZ}$, $K(Z,x) = K_{Zx}$, $K(x,Z) = K_{xZ}$, $K(x,x) = K_{xx}$, and by conditional distribution, we get:

$$E(f_*) = K_{xZ}\, \boldsymbol{a}, \quad \text{where } \boldsymbol{a} = (K_{ZZ} + \sigma_m^2 I)^{-1} y \quad (2)$$

$$cov(f_*) = K_{xx} - K_{xZ}[K_{ZZ} + \sigma_m^2 I]^{-1} K_{Zx} \quad (3)$$

The predicted variance $\sigma^2(f(U))$ of all unevaluated points U is the diagonal of $cov(f(U))$ calculated from (3) using the measured data Z.

D. ACTIVE LEARNING - COHN (ALC)

Active learning - Cohn (ALC) algorithm selects the next evaluation point that maximizes the expected reduction in the squared error averaged over input space for each $x_i \in U$ added to the training set (Z) [27], [68]:

$$argmax_{x_i \in U} \frac{\sum_{j=1}^{q-1} (\sigma_Z^2 f(x_j) - \sigma_{Z+x_i}^2 f(x_j))}{q-1} \quad (4)$$

where $\sigma_Z^2 f(x_j) = K_{x_j x_j} - K_{x_j Z}[K_{ZZ} + \sigma_m^2 I]^{-1} K_{Z x_j}$, $\sigma_{Z+x_i}^2 f(x_j) = K_{x_j x_j} - K_{x_j, Z+x_i}[K_{Z+x_i, Z+x_i} + \sigma_m^2 I]^{-1} K_{Z+x_i, x_j}$, and q is the number of unevaluated points ($q = n - m$).

III. PROPOSED ACTIVE LEARNING CRITERIA FOR SELECTING THE MOST INFORMATIVE POINTS

We begin with extending the ALC measure by adding a penalty term to integrate the information of both evaluated and unevaluated points. This penalty term is regularized to achieve the right balance for selecting the most informative point without increasing the uncertainty of the model. We then propose a novel formulation for calculating the similarity between evaluated and unevaluated points to further improve the proposed method. Finally, we discuss the relationship between the proposed measure and the Laplacian regularized Kernel ridge regression.

A. LAPLACIAN REGULARIZED ACTIVE LEARNING (LR-AL)

Different from the classical criteria for selecting the next most informative points which makes use of only evaluated points, i.e. classical ALC, the Laplacian regularized active learning (LR-AL) makes use of both evaluated and unevaluated points to learn the underlying geometrical structure in the data. It is assumed that if two points (x_i, x_j), are sufficiently close to each other, then their responses $(f(x_i), f(x_j))$ are close as well.

Assuming there is a set of pre-specified unevaluated points (U) from which the next evaluation point should be selected from, we introduce a graph Laplacian penalty term to the ALC measure to incorporate the information of unevaluated points as well as evaluated points to identify the points with most information content. Specifically, the proposed LR-AL measure selects the next point that minimizes the regularized predicted variance that is averaged over all unevaluated points $U - x_i$, when $x_i \in U$ is added to the training set (Z). The selection of points follows (4) where

$$\sigma_Z^2 f(x_j) = K_{x_j x_j} - K_{x_j Z}[\sigma_m^2 I + K_{ZZ} + \lambda K_{(ZX)} L K_{(XZ)}]^{-1} K_{Z x_j} \quad (5)$$

$$\sigma_{Z+x_i}^2 f(x_j) = K_{x_j x_j} - K_{x_j, Z+x_i}[\sigma_m^2 I + K_{Z+x_i, Z+x_i} + \lambda K_{(Z+x_i, X)} L K_{(X, Z+x_i)}]^{-1} K_{Z+x_i, x_j} \quad (6)$$

where $K_{(Z,X)} L K_{(X,Z)}$ is the graph Laplacian penalty for (5), $K_{(Z+x_i,X)} L K_{(X,Z+x_i)}$ is the graph Laplacian penalty for (6), and $\lambda \geq 0$ is the tuning parameter which should be set to a small number. The matrix L is called graph Laplacian in spectral graph theory [69] and is calculated as $L = D - S$, where D is a diagonal matrix with $D_{ii} = \sum_j S_{ij}$, and S is a similarity to quantify the similarity between points (X). An appropriate choice of similarity matrix should contain symmetric weights $S_{ij} (S_{ij} = S_{ji})$ which imposes a heavy penalty if neighboring points x_i and x_j are mapped far apart, i.e. nearest neighbour

(Alaeddini, Craft *et al.* 2019). In Section III-B we propose a novel formulation for constructing graph Laplacian which considers similarity of both predictor and response variable spaces. For the tuning parameter, we propose to set $\lambda = \lambda_* \sigma_m^2 K_{(Z,Z)}^{-1}$ for (5), and $\lambda = \lambda_* \sigma_m^2 K_{(Z+x_\mathbf{i}, Z+x_\mathbf{i})}^{-1}$ for (6) to automatically adjust the significance of the graph Laplacian penalty with respect to the other components of the inverse term, namely K and $\sigma_m^2 I$, as well as the hyperparameter λ_*. Therefore, whenever there is an update in the variance of the Gaussian noise (σ_m^2), or the covariance between the set of points (K), the tuning parameter will be automatically updated. Such formulation has an intuitive relation with the tuning parameter of the Laplacian regularized ridge regression which is briefly discussed in Section III-C. Setting $\lambda_* = 0$, simplifies the equations (5) and (6) to the variance of regular GP using training set Z and $Z + x_\mathbf{i}$ respectively. We propose to select the λ_* value such that it maximizes the expected reduction in squared error averaged over the input space. In Section III-D we provide a simple greedy algorithm for optimizing $\lambda_* \geq 0$ parameter.

B. PROPOSED GRAPH LAPLACIAN

Let S_X and S_Y denote the similarity matrices of the data points in the predictor variables (input) space (X) and the response variable (output) space (Y), where $S_{X_{ij}}$ and $S_{Y_{ij}}$ are measured using squared Euclidean distance, namely $S_{X_{ij}} = \left\|(x_\mathbf{i} - x_\mathbf{j})\right\|_2$, and $S_{Y_{ij}} = \left\|(y_i - y_j)\right\|_2$. We propose to define the graph Laplacian as

$$L = \frac{L_Y}{L_X} = \frac{D_Y - S_Y}{D_X - S_X} \tag{7}$$

where $D_{X_{ii}} = \sum_j S_{X_{ij}}$, $D_{Y_{ii}} = \sum_j S_{Y_{ij}}$. The proposed graph Laplacian utilizes the information of both the predictor and response variables over the hypothetical line between each pair of points. Using extensive simulation studies, we found the proposed graph Laplacian in (7) outperforms the classical graph Laplacian, where S is defined as

$$S_{ij} = \begin{cases} 1 & \text{if } [i,j] \text{ are among } p \text{ nearest} \\ & \text{neighbors of each other} \\ 0 & \text{otherwise,} \end{cases}$$

where p can be set using cross-validation, and D is the degree matix with diagonal elements as $D_{ii} = \sum_j S_{ij}$ [70], [71]. We provide a sample result of comparing the classical graph Laplacian with the proposed graph Laplacian in the Appendix. We also compared the performance of the proposed Laplacian with other intuitive forms including L_X, L_Y, $\sqrt{\frac{L_Y}{L_X}}$, $\sqrt{\frac{L_X}{L_Y}}$, $\frac{L_X}{L_Y}$ and $L_X L_Y$ and found the proposed measure has the most competitive performance.

Fig. 1 provides a graphical representation of the proposed graph Laplacian measure over the contour plot of a nonlinear function with three evaluated and three unevaluated points. In the space of predictor variables, we can see that $x_\mathbf{4}$ and $x_\mathbf{6}$ are the two points that are almost equidistant from the evaluated point $x_\mathbf{3}$. By using the kernels without the information

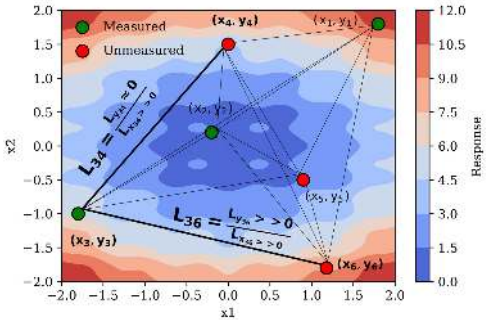

FIGURE 1. Graphical representation of the proposed graph Laplacian measure.

from responses, we tend to pick either $x_\mathbf{4}$ or $x_\mathbf{6}$ as our next evaluation point. However, by incorporating the information from the responses, we can see that the unevaluated point $x_\mathbf{4}$ is more similar to $x_\mathbf{3}$ (compared to $x_\mathbf{6}$). Thus, it helps picking $x_\mathbf{6}$ which provides more information.

C. RELATIONSHIP WITH LAPLACIAN REGULARIZED KERNEL RIDGE REGRESSION

Suppose there are a total of n possible points out of which m points are already evaluated. Also, let S be a similarity matrix. Then, the Laplacian regularized ridge regression solves the following optimization problem:

$$J[f] = \frac{\lambda_1}{2} \|f\|_{\mathcal{H}}^2 + \frac{\lambda_2}{2} \sum_{i,j=1}^{n} (f(x_\mathbf{i}) - f(x_\mathbf{j}))^2 S_{ij} + \frac{1}{\sigma_m^2} \sum_{i=1}^{m} (y_i - f(\boldsymbol{z_i}))^2 \tag{8}$$

where the first term is the ridge penalty in the form of squared Euclidean norm of the vector of regression coefficients in Hilbert space with λ_1 denoting the weight of ridge regularization term, the second term is the Laplacian penalty with λ_2 denoting the weight of the Laplacian regularization term, and the third term is the standard least square loss function. Substituting $f(x) = \sum_{i=1}^{m} a_i k(x, \boldsymbol{z_i})$ and using $\langle k(\cdot, x_\mathbf{i}), k(\cdot, x_\mathbf{j}) \rangle_{\mathcal{H}} = k(x_\mathbf{i}, x_\mathbf{j})$, and setting $\lambda_1 = 1$, result in the following kernel Laplacian regularized least squared problem.

$$J[\boldsymbol{a}] = \frac{1}{2} \boldsymbol{a}^T K_{ZZ} \boldsymbol{a} + \frac{\lambda_2}{2} \boldsymbol{a}^T K_{ZX} L_{XX} K_{XZ} \boldsymbol{a} + \frac{1}{2\sigma_m^2} |\boldsymbol{y} - K_{ZZ} \boldsymbol{a}|^2 \tag{9}$$

The Laplacian regularized least squared model in (9) can be minimized by solving the system of equations resulting from differentiating J with respect to the vector of coefficients $\boldsymbol{a}$

$$\frac{\partial J}{\partial \boldsymbol{a}} = \frac{\partial}{\partial \boldsymbol{a}} \left(\frac{1}{2} \boldsymbol{a}^T \left(K_{ZZ} + \frac{1}{\sigma_m^2} K_{ZZ}^2 + \lambda_2 K_{ZX} L_{XX} K_{XZ}\right) \boldsymbol{a} - \frac{1}{\sigma_m^2} \boldsymbol{y}^T K_{ZZ} \boldsymbol{a} + \frac{1}{2\sigma_m^2} \boldsymbol{y}^T \boldsymbol{y}\right) = 0 \tag{10}$$

$$\boldsymbol{a} = (\sigma_m^2 I + K_{ZZ} + \sigma_m^2 \lambda_2 K_{ZZ}^{-1} K_{ZX} L_{XX} K_{XZ})^{-1} \boldsymbol{y} \tag{11}$$

The inverse term in (11) is very similar to the one in (5). In fact, setting $\lambda = \lambda_2 \sigma_m^2 K_{ZZ}^{-1}$ in (5), makes the inverse terms identical in both equations.

D. PROPOSED ALGORITHMS

In this section, we provide the pseudo codes of the proposed sequential Laplacian regularized Gaussian process (SLRGP) algorithm and the sequential algorithm for optimizing the tuning parameter of graph Laplacian used in SLRGP. The implementation of the proposed algorithms can be found in https://github.com/rajithameka/SLRGP.

Algorithm 1 Sequential Laplacian Regularized Gaussian Process (SLRGP)

Input:	Set of n pre-specified points (X)
Output:	Design vector for the m evaluated points (Z) Estimated GP ($f(x)$)
Step 1.	Determine Z by selecting m points from X using a space filling design
Step 2.	Until satisfying some desired stopping criteria
Step 2.1	Optimize the tuning parameter using Algorithm 2
Step 2.2	For each $x_\mathbf{i} \in U$ select x^* such that $argmax_{x_\mathbf{i} \in U} \frac{\sum_{j=1}^{q-1} (\sigma_Z^2 f(x_\mathbf{j}) - \sigma_{Z+x_i}^2 f(x_\mathbf{j}))}{q-1}$ using (5) and (6)
Step 2.3	$Z \leftarrow Z + x^*$
Step 2.4	$f(x) = K_{xZ}(\sigma_n^2 I + K_{ZZ})^{-1}\mathbf{y}$ Calculate MSE Go to Step 2.1

1) SEQUENTIAL LAPLACIAN REGULARIZED GAUSSIAN PROCESS (SLRGP)

Algorithm 1 illustrates the proposed sequential Laplacian regularized Gaussian process (SLRGP). The algorithm essential input includes a set of pre-specified settings (X) from which the evaluation points should be selected. The outputs of the algorithm include the matrix of design vector for the selected points (Z), and the estimated GP ($f(x)$). The algorithm begins with determining a set of m points (Z) from all feasible settings (X) using a space filling design such as LHD and obtaining their response values (y) (Step 1). Next, it optimizes the tuning parameter of the proposed sequential Laplacian regularized Gaussian process (SLRGP) using Algorithm 2 (Step 2.1), and uses (4) to sequentially identify the most informative unevaluated points to be evaluated until a desired stopping criterion is met (Step 2.2). Each selected setting (x^*) is then evaluated and moved to the evaluated points (Z) before checking the stopping criterion for initiating another iteration (Step 2.3). The stopping criterion can be based on a pre-specified number of design points, reduction in MSE, etc. After each evaluation, the set of evaluated points (Z), which are expected to have the most information content of all settings (X), are used to update the GP model fit and calculate the associated error (Step 2.4).

2) OPTIMIZATION OF THE TUNING PARAMETER

Algorithm 2 demonstrates the proposed sequential algorithm for optimizing the tuning parameter of the graph Laplacian regularization. The algorithm input includes the set of m existing evaluated (Z) and q unevaluated points ($U = X - Z$), and a set of candidate values (λ_c) for the tuning parameter λ. We consider a prespecified finite set of candidate values of the tuning parameter λ to reduce the computational complexity of the optimization algorithm. The output of the algorithm is the optimal value of the tuning parameter, λ^*. The algorithm begins with initializing the vector P to store the LR-AL values for different choices of the tuning parameter, i.e. $\lambda_c = 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2$ (Step 1). Next, for each choice of the tuning parameter in λ_c, it uses (4) to select the most informative point from the unevaluated set (Step 2.1). Adding the selected point for each λ_c to the set of evaluated points, it then calculates the maximum variance of the remaining unevaluated points using (3) and store it in vector P (Step 2.2). Finally, it selects the optimal tuning parameter λ_* which corresponds to the minimum of the maximum variances stored in P (Step 3).

Algorithm 2 Optimization of the Tuning Parameter of the Laplacian Regularization Penalty (λ_*)

Input:	Set of m evaluated points (Z), Set of q unevaluated points ($X - Z$), Set of candidate values for tuning parameters (λ_c)
Output:	Optimized value of the tuning parameter (λ_*)
Step 1.	Initialize, $P[i] = 0$, $i = 1, .., size(\lambda_c)$
Step 2.	For each $\lambda \in \lambda_c$, select x^* such that
Step 2.1	$x^* = argmax_{x_\mathbf{i} \in U}[max[K_{(U-x_\mathbf{i}, U-x_\mathbf{i})} - K_{(U-x_\mathbf{i}, Z+x_\mathbf{i})}[\sigma_m^2 I + K_{(Z+x_\mathbf{i}, Z+x_\mathbf{i})} + \lambda K_{(Z+x_\mathbf{i}, X)} L K_{(X, Z+x_\mathbf{i})}]^{-1} \times K_{(Z+x_\mathbf{i}, U-x_\mathbf{i})}]]$
Step 2.2	$P[i] = \frac{\sum_{j=1}^{q-1} (\sigma_Z^2 f(x_\mathbf{j}) - \sigma_{Z+x^*}^2 f(x_\mathbf{j}))}{q-1}$ where, $\sigma_Z^2 f(x_\mathbf{j}) = K_{x_\mathbf{j} x_\mathbf{j}} - K_{x_\mathbf{j} Z}[K_{ZZ} + \sigma_m^2 I]^{-1} K_{Z x_\mathbf{j}}$ $\sigma_{Z+x^*}^2 f(x_\mathbf{j}) = K_{x_\mathbf{j} x_\mathbf{j}} - K_{x_\mathbf{j}, Z+x^*}$ $[K_{Z+x^*, Z+x^*} + \sigma_m^2 I]^{-1} K_{Z+x^*, x_\mathbf{j}}$ $i \leftarrow i + 1$
Step 3	$\lambda_* = argmax_{\lambda \in \lambda_c}(P)$

IV. CASE STUDY AND SIMULATED EXPERIMENTS

In this section, we validate the performance of the proposed methodology along with a number of existing methods in the literature including expected improvement (EI), maximum entropy (ME), integrated mean square error (IMSE),

maximum mean square error (MMSE), Gaussian random field (GRF), active learning - Cohn (ALC), sequential Laplacian regularized V-optimal (SLRV), improved greedy sampling (iGS) and graph based active learning (GBAL) methods using both case study and simulated experiments. In our paper, we use MATLAB for coding and GPML library [72] for optimizing the hyperparameters of the GP model. While there are several possible choices of kernel functions to build the GP model, i.e. squared exponential, Matern, etc. we consider the squared exponential kernel for all of our experiments, as it is widely used and has the capacity to learn any well behaved function with infinite data [4]. We begin with a brief discussion of each of the comparing methods and the performance metric chosen for the analysis of the results. Next, we illustrate the result of a case study on kinematics of pitching in baseball. Finally, we describe the result of a simulation study based on eight nonlinear response models of 2 to 10 dimensions with different levels of noise.

A. COMPARING METHODS

Here we provide a brief discussion of the comparing methods, except ALC which has been presented earlier in Section II.

1) EXPECTED IMPROVEMENT (EI)

EI selects the next evaluation point $x_\mathbf{i} \in U$ that maximizes the expected improvement given the training set (Z) [32]:

$$argmax_{x_\mathbf{i}\in U} E(I(x_\mathbf{i})) = (f_{min} - \hat{y})\Phi(\frac{f_{min} - \hat{y}}{s}) + s\phi(\frac{f_{min} - \hat{y}}{s}) \quad (12)$$

where f_{min} is the current best value of the estimated function, $\phi(.)$ is the standard normal density function, and $\Phi(.)$ is the standard normal distribution functions. Here, $\hat{y} = f_*$ is the predicted value at $x_\mathbf{i} \in U$ given Z.

2) MAXIMUM ENTROPY (ME)

ME selects the next evaluation point $x_\mathbf{i} \in U$ that maximizes the entropy given the training set (Z) [47]:

$$argmax_{x_\mathbf{i}\in U}[\frac{1}{2}log\sigma^2_{x_\mathbf{i}\,|\,Z} + \frac{1}{2}(log(2\pi) + 1)] \quad (13)$$

3) INTEGRATED MEAN SQUARE ERROR (IMSE)

IMSE selects the next evaluation point that minimizes the trace of the matrix of the predicted variance of the (remaining) unevaluated points $U - x_\mathbf{i}$, when $x_\mathbf{i} \in U$ is added to the training set (Z) [11]:

$$argmin_{x_\mathbf{i}\in U}[\sum_{j=1}^{q}[K_{(x_\mathbf{j}x_\mathbf{j})} - K_{(x_\mathbf{j},Z+x_\mathbf{i})} \times [\sigma_m^2 I + K_{(Z+x_\mathbf{i},Z+x_\mathbf{i})}]^{-1}K_{(Z+x_\mathbf{i},x_\mathbf{j})}]] \quad (14)$$

4) MAXIMUM MEAN SQUARE ERROR (MMSE)

MMSE selects the next evaluation point that minimizes the maximum predicted variance of the (remaining) unevaluated points $U - x_\mathbf{i}$, when $x_\mathbf{i} \in U$ is added to the training set (Z) [11]:

$$argmin_{x_\mathbf{i}\in U}[max_{x_\mathbf{j}\in U - x_\mathbf{i}}[K_{(x_\mathbf{j}x_\mathbf{j})} - K_{(x_\mathbf{j},Z+x_\mathbf{i})} \times [\sigma_m^2 I + K_{(Z+x_\mathbf{i},Z+x_\mathbf{i})}]^{-1}K_{(Z+x_\mathbf{i},x_\mathbf{j})}]] \quad (15)$$

5) GAUSSIAN RANDOM FIELDS (GRF)

GRF is a semi-supervised learning method which represents evaluated (labeled) and unevaluated (unlabeled) data points using a weighted graph, where the graph weights are calculated based on a similarity function like radial basis function (RBF) [73]. As the Gaussian field conditioned on the evaluated data points is a multivariate normal distribution $y \sim \mathcal{N}(0, \Delta^{-1})$, it can be seen as GP, where Δ is the Laplacian matrix calculated as $\Delta = D - W$. W is an edge matrix calculated using any kernel function K, and D is a diagonal matrix with entries $D_{ii} = \sum_j W_{ij}$. Here, we consider $y_u \sim \mathcal{N}(0, (\beta(\Delta + \frac{1}{\sigma^2})^{-1})$ for construction of the GRF, where β controls the sharpness of the distribution, and σ^2 controls the amount of regularization as described in [64].

6) SEQUENTIAL LAPLACIAN REGULARIZED V-OPTIMAL (SLRV)

SLRV selects the next evaluation point that minimizes the Laplacian regularized V-optimality criterion based on the locally weighted regression (LOESS) using both evaluated and unevaluated points [2]:

$$argmin_{x_\mathbf{i}\in U}[\sum_{j=1}^{q} \text{avg}(X_*^T(Z_*(x_\mathbf{i})^T W(x_\mathbf{i}) Z_*(x_\mathbf{i}) + \lambda_1 X^T L X + \lambda_2 I)^{-1} X_*)] \quad (16)$$

where

$$Z_*(x_\mathbf{i}) = \begin{pmatrix} 1 & (x_\mathbf{1} - x_\mathbf{i})^T \\ \vdots & \vdots \\ 1 & (x_\mathbf{m} - x_\mathbf{i})^T \end{pmatrix}$$

is the transform matrix of evaluated points,

$$X_* = \begin{pmatrix} 1 & (x_\mathbf{1})^T \\ \vdots & \vdots \\ 1 & (x_\mathbf{n})^T \end{pmatrix}$$

is the transform matrix of all points, $W(x_\mathbf{i})$ is a weight matrix based on the scaled distances between the target point $x_\mathbf{i}$ and the evaluated points $x_\mathbf{j}$, $j = 1, \ldots, m$, namely $W(x_\mathbf{i}) = diag(K_h(x_\mathbf{i}, x_\mathbf{1}), \ldots, K_h(x_\mathbf{i}, x_\mathbf{m}))$. While there are several choices for calculating the scaled distances, the tricube weight function is usually used in practice, with

$$K_h(x_\mathbf{i}, x_\mathbf{j}) = \begin{cases} (1 - \left|\frac{x_\mathbf{i} - x_\mathbf{j}}{h}\right|^3)^3 & \text{if } \left|\frac{x_\mathbf{i} - x_\mathbf{j}}{h}\right| < 1 \\ 0 & \text{if } \left|\frac{x_\mathbf{i} - x_\mathbf{j}}{h}\right| \geq 1. \end{cases}$$

The Laplacian matrix is calculated as $L = D_x - S_x$, where

$$S_{X_{ij}} = \begin{cases} S_{X_{ij}} = 1 & \text{if i, j are among } p \text{ nearest neighbors} \\ S_{X_{ij}} = 0 & \text{otherwise,} \end{cases}$$

and $D_{X_{ii}} = \sum_j S_{X_{ij}}$. As opposed to the proposed Laplacian matrix in (7), the Laplacian matrix of SLRV is developed using only the information from the input variables.

7) IMPROVED GREEDY SAMPLING (IGS)

IGS initially calculates $d^x_{UZ} = \|x_\mathbf{i} - x_\mathbf{j}\|_2$ and $d^y_{UZ} = \|f(x_\mathbf{i}) - y_j\|_2$ for each $x_\mathbf{i} \in U$ and $x_\mathbf{j} \in Z$, then selects the next evaluation point that maximizes d^{xy}_U which is calculated as [30]:

$$d^{xy}_U = min(d^x_{UZ}\, d^y_{UZ}) \tag{17}$$

8) GRAPH BASED ACTIVE LEARNING (GBAL)

GBAL calculates the measure of uncertainty for each unevaluated point $x_\mathbf{i} \in U$ as $\theta(x_\mathbf{i}) = min_{x_\mathbf{z} \in Z} L_1(x_\mathbf{i}, x_\mathbf{z})$, then selects the next evaluation point that maximizes $Q(x_\mathbf{i})$ when $x_\mathbf{i} \in U$ is added to the training set (Z) [31]:

$$Q(x_\mathbf{i}) = \sum_{i \in U} \theta(x_\mathbf{i}) - \sum_{j \in U - x_\mathbf{i}} \theta^{x_\mathbf{i}}(x_\mathbf{j}) \tag{18}$$

B. PERFORMANCE METRIC

Since the main objective of the proposed SLRGP algorithm is efficient estimation of expensive noisy black-box functions, we choose to use the root mean squared error (RMSE) of the estimated and true models for performance evaluation. We also study the average predicted variance (APV) at test points, as another performance metric, which shows similar general trend as RMSE and hence not reported in the manuscript for the economy of space, except for one case which is discussed in Section IV-D. To calculate the RMSE, we use a randomly selected out-of-sample of size $t = 1000$ from the true response models and compare their associated response values $(y_i^{Tru}, i = 1, \ldots, t)$ against those provided by the estimated model (y_i^{Est}) using the RMSE metric, $RMSE = \sqrt{\frac{\sum_{i=1}^{t}(y_i^{Est} - y_i^{Tru})^2}{t}}$. In order to achieve a high level of confidence over the results, all simulated experiments are repeated hundred times and the average result is reported.

C. CASE STUDY

In this section, we illustrate the results of a case study for analysis of the kinematics of pitching in baseball comparing the proposed methodology with expected improvement (EI), maximum entropy (ME), integrated mean square error (IMSE), maximum mean square error (MMSE), Gaussian random field (GRF), active learning - Cohn (ALC), sequential Laplacian regularized V-optimal (SLRV), improved greedy sampling (iGS) and graph based active learning (GBAL) methods. The study is based on a secondary analysis of the effect of 5 kinematic (explanatory) variables, including (1) maximum axial shoulder external rotation angles, (2) trunk forward flexion angle at ball release, (3) stride length, (4) release height, and (5) release distance, on the velocity (response variable) of the baseball for 73 pitchers. Fig. 2 gives the graphical representation of the kinematic variables. Ball velocity is one of the key performance measures for baseball pitchers. The maximum axial shoulder external rotation angle and trunk forward flexion angle at ball release have been linked to ball velocity in studies using regression analyses. Stride length and the position of the hand at ball release are also considered to affect ball velocity, and thus often evaluated by coaches.

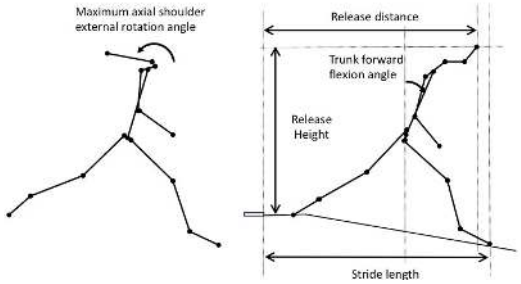

FIGURE 2. Kinematic variables associated with ball velocity.

To construct the design matrix of the comparing methods from the available 73 data points (pitchers), we begin with randomly selecting 15 points for testing. Next, from the remaining 58 points, we randomly select 20 points for training the initial surrogate model. Finally, from the remaining 38 points, we select 10 augmenting points one-at-a-time using each of the comparing methods. To increase the confidence we repeat the procedure 100 times and report the the average result.

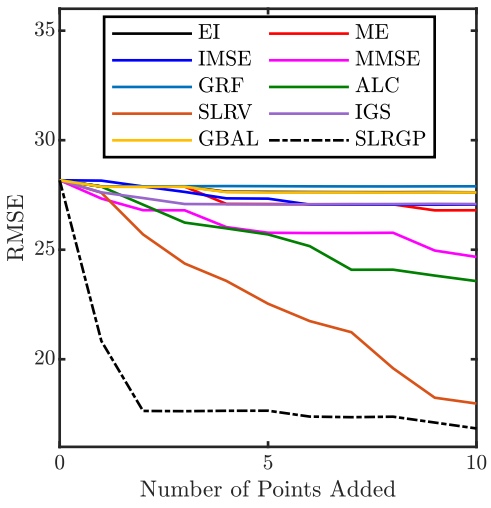

FIGURE 3. RMSE performance of the EI, ME, IMSE, MMSE, ALC, SLRV, IGS, GBAL and SLRGP - pitching case study.

Fig. 3 illustrates the performance of the comparing methods based on the RMSE of estimated and true response model parameters using 100 replications of the procedure. As shown in Fig. 3, the proposed SLRGP method outperforms all other methods by a significant margin across all

TABLE 1. P-values of the Wilcoxon rank test - Case study.

	EI	ME	IMSE	MMSE	GRF	ALC	SLRV	IGS	GBAL
SLRGP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

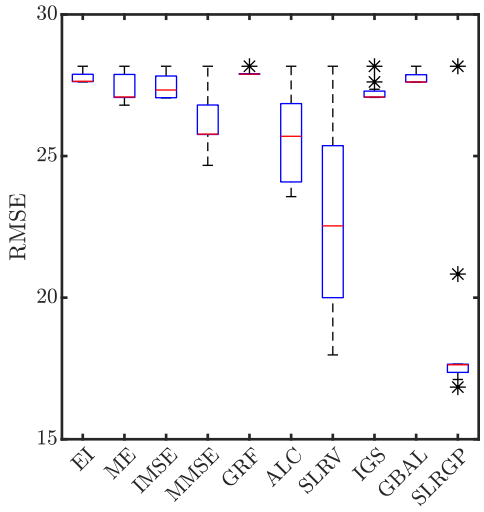

FIGURE 4. Box plot of RMSE performance of the EI, ME, IMSE, MMSE, ALC, SLRV, IGS, GBAL and SLRGP - pitching case study.

data points. The performance patterns also show the proposed method improves the RMSE much quicker than the other methods for the first few additional points before stabilizing. This can be attributed to the proposed graph Laplacian measure and the sequential algorithm for selection of new experiments. After the proposed method, SLRV and ALC provide the best performance, followed by the MMSE and iGS methods. Table 1 verifies the observation by providing the result of the Wilcoxon rank test for the significance of the difference between the RMSE of the comparing methods (See also Fig. 4).

D. SIMULATED EXPERIMENTS: NONLINEAR RESPONSE MODELS

In this section, we evaluate the performance of the proposed SLRGP method along with EI, ME, IMSE, MMSE, GRF, ALC, SLRV, iGS and GBAL methods over nonlinear response models of two, three, six and ten variables at different levels of noise including 1%, 3% and 5% of the mean value of the response models. These response models are presented in Table 2. All simulated experiments are repeated hundred times and the average results are reported.

For each of the comparing methods, we begin with creating a LHD of 200 points for all the functions and then randomly select $4d$ number of points as the initial set of evaluated points. This gives 8, 12, 24 and 40 initial points for each of the 2, 3, 6 and 10 variable functions respectively. Next, each of the comparing methods is used to select 40 additional points to improve the initial prediction.

Fig. 5 illustrates the root mean squared error (RMSE) of the estimated responses from each of the comparing methods after adding each point at 1%, 3% and 5% noise levels. As shown in the Fig. 5, the SLRGP method outperforms almost all other methods in terms of RMSE performance. Apart from few exceptions, the improvement made by the proposed method is generally more evident in cases with larger standard deviation of errors. Also, the performance of the SLRGP for high dimensional response models improves compared to lower dimensional models, indicating its robustness towards increasing the number of variables. In addition, as the number of points increases, the proposed method generally maintains or increases its advantage over other methods, which demonstrates the effectiveness of both the proposed LR-AL measure and SLRGP algorithm. Among the other comparing methods, the result is mixed but IMSE and SLRV provide relatively better performance for lower and higher dimension functions respectively.

For low dimensional functions, most of the comparing methods provide competitive performance, especially for the initial set and the first few additional points. Meanwhile, for some of the response models, i.e. 2.2, the proposed method starts with a higher RMSE compared to other leading methods, i.e. IMSE and ALC, though it catches up after few additional points. This may be attributed to requiring additional points for better prediction of unevaluated points in the Laplacian matrix, and optimizing the tuning parameter. For the higher dimensional functions, SLRGP method provides a significant improvement over the comparing methods. Our conjecture is that as the complexity of the response model increases, the information of the unevaluated points provide more contribution compared to lower dimension functions. After SLRGP, SLRV is the next best performing methods especially for higher dimensional functions, namely 6D and 10D, and low noise, namely 1%, 3%. We think this because SLRV also utilized the information of the unevaluated points.

Among comparing methods, the iGS is the only method that does not use the uncertainty information for selecting the next evaluation point. iGS shows competitive performance for almost all low dimensional functions, namely 2.1, 2.2, 3.1 and 3.2. Specifically, for the response model 3.2, the iGS method provides even better RMSE performance compared to the proposed method. However, while iGS method provides competitive performance in terms of RMSE, there is a good chance that the selected point might increase the uncertainty of the updated/augmented model, which can be shown using average predicted variance at test points.

Fig. 6 compares the RMSE and APV of the response model 3.2. As shown in the Fig. 6, the average predicted variance of iGS is considerably larger than the proposed method, even though their RMSE's are comparable. The result shows that SLRGP is overall a better performing method compared to iGS, as the point selection process should not only improve the performance of the model in terms of RMSE, but also it

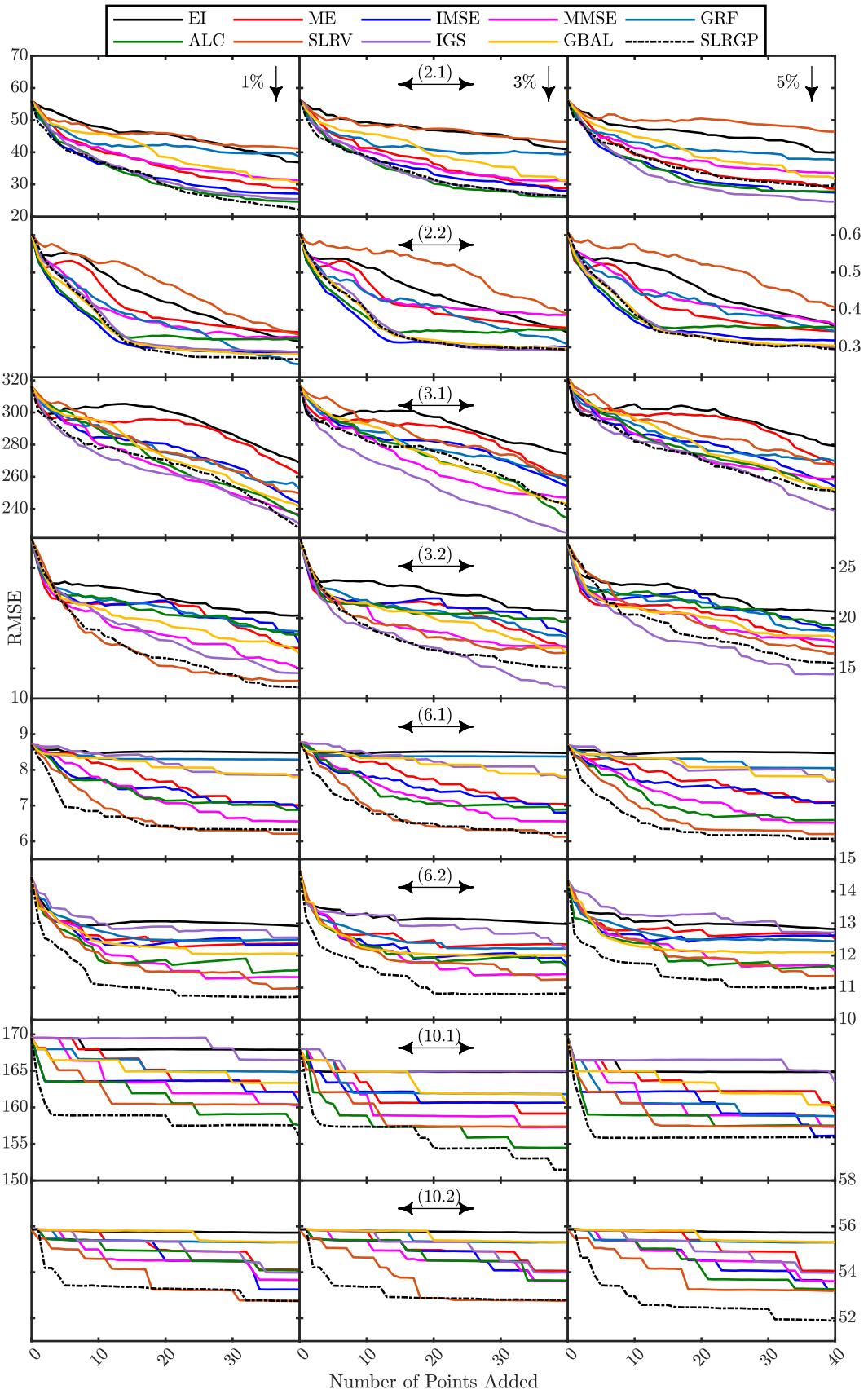

FIGURE 5. RMSE performance of the EI, ME, IMSE, MMSE, GRF, ALC, SLRV, IGS, GBAL and SLRGP.

TABLE 2. Non-linear response models considered for the comparisons.

	Bounds	Response model
2.1	$x_1 = [-5, 10], x_2 = [0, 15]$	$y = (x_2 - \frac{5.1}{(4\pi^2)}x_1^2 + \frac{5}{\pi}x_1 - 6)^2 + 10(1 - \frac{1}{8\pi}cos(x_1)) + 10 + \epsilon$
2.2	$x_i = [-1, 1], i = 1, 2$	$y = \sum_{i=1}^{2} ix_i^4 + \epsilon$
3.1	$x_i = [-5, 5], i = 1, 2, 3$	$y = \sum_{i=1}^{3}(x_i^2 - 1)^2 + \epsilon$
3.2	$x_i = [0, 1], i = 1, 2, 3$	$y = 4(x_1 - 2 + 8x_2 - 8x_2^2)^2 + (3 - 4x_2)^2 + 16\sqrt{(x_3 + 1)}(2x_3 - 1)^2 + \epsilon$
6.1	$x_i = [-5, 5], i = 1, .., 6$	$y = \sum_{i=1}^{6} \lvert x_i \sin x_i + 0.1x_i \rvert + \epsilon$
6.2	$x_i = [-2, 2], i = 1, .., 6$	$y = \sum_{i=1}^{6} x_i^2 + 2x_{i+1}^2 - 0.3cos(3\pi x_i) - 0.4cos(4\pi_{i+1}) + 0.7 + \epsilon$
10.1	$x_i = [-5, 5], i = 1, .., 10$	$y = \frac{1}{2}\sum_{i=1}^{10} x_i^2 - 10cos(2\pi x_i) + \epsilon$
10.2	$x_i = [-5, 5], i = 1, .., 10$	$y = \sin^2\pi(1 + \frac{x_1 - 1}{4}) + \sum_{i=1}^{9}((1 + \frac{x_i - 1}{4}) - 1)^2[1 + 10\sin^2(\pi(1 + \frac{x_i - 1}{4}) + 1)] + ((1 + \frac{x_{10} - 1}{4}) - 1)^2[1 + \sin^2(2\pi(1 + \frac{x_{10} - 1}{4}))] + \epsilon$

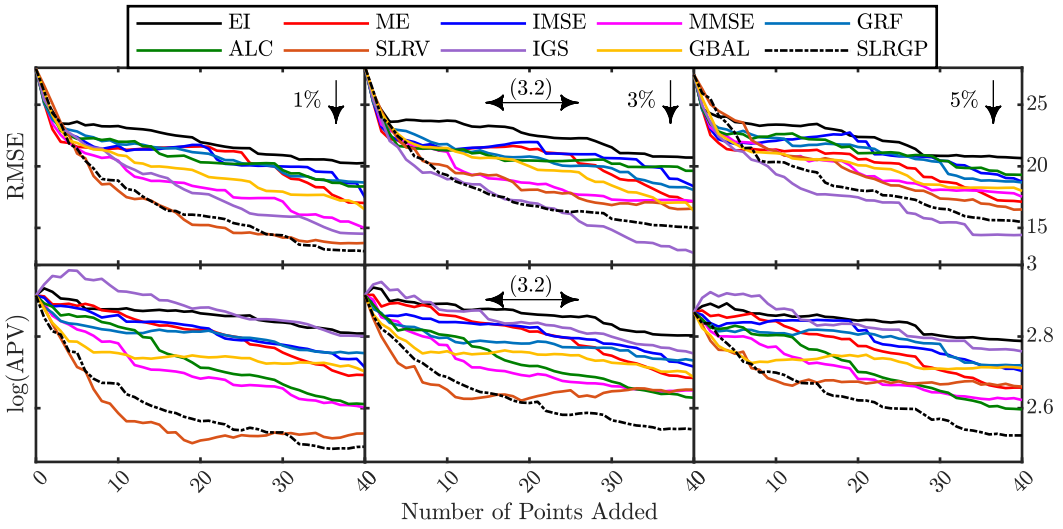

FIGURE 6. RMSE and APV of the EI, ME, IMSE, MMSE, GRF, ALC, SLRV, IGS, GBAL and SLRGP of response model 3.2.

should decrease or maintain the uncertainty of the model as new information is added.

Fig. 7 complements the results of the earlier analysis by providing the box plot of the RMSE performance of each of the comparing methods for each of the response models over all of their selected points (initial + 40 additional points). As shown in the Fig. 7 the proposed SLRGP method generally provides the lowest quantiles, i.e. 25*th*, 50*th*, and 75*th* compared to the others which demonstrate its superior performance. Also, for most cases, SLRGP shows a larger variance in the boxplot, which can be attributed to the greater RMSE reduction over the selected points in comparison to the other methods. This is due to the selection of more informative points by the proposed SLRGP algorithm using the LR-AL criteria. For high dimension models, with better quality of prediction information, the proposed method converged very fast compared to other models and due to the more evaluations needed before next good approximation of the function, the variance of the proposed method is small. Meanwhile, other than few exceptions, the box plots show EI and GRF provide the smallest changes in the RMSE performance from the initial set of points.

Finally, Table 3 provides the result of the Wilcoxon rank test for the significance of the difference between the RMSE performance of the proposed method against other methods, where lower values show an increased probability of

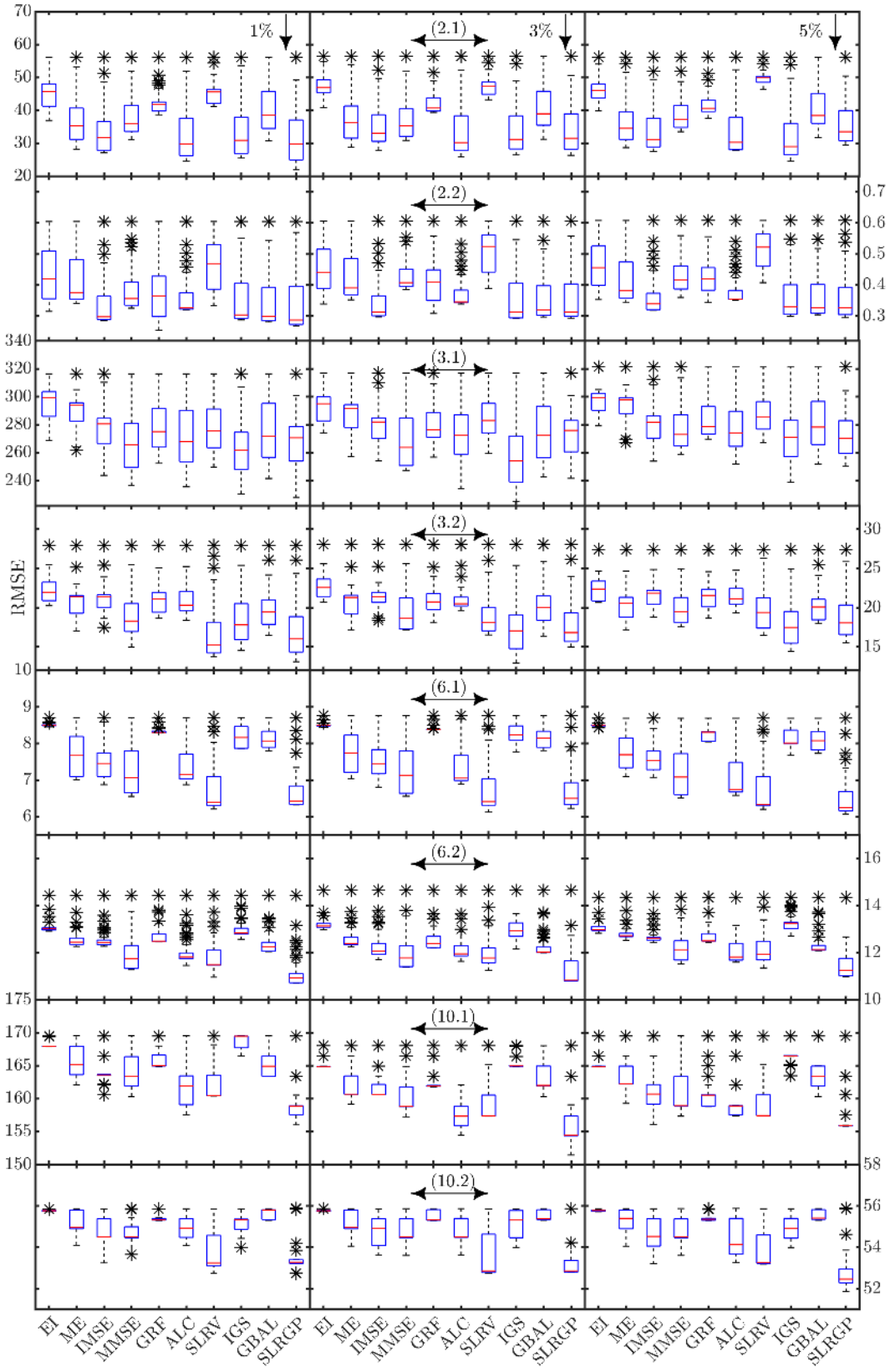

FIGURE 7. Box plot of RMSE of the EI, ME, IMSE, MMSE, GRF, ALC, SLRV, IGS, GBAL and SLRGP.

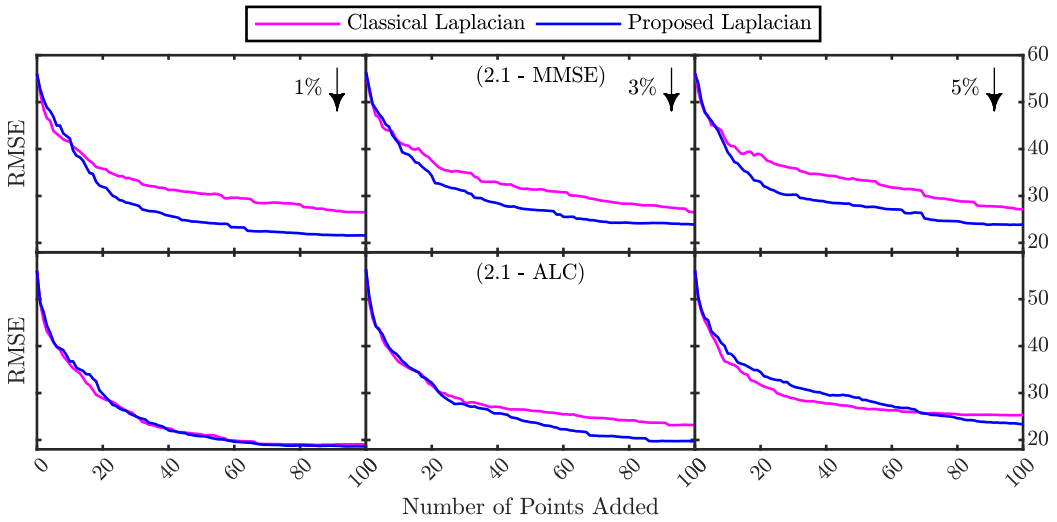

FIGURE 8. Comparison between the classical and the proposed graph Laplacian based on response model 2.1 using MMSE and ALC.

TABLE 3. P-values of the Wilcoxon rank test - simulated experiments.

		2.1	2.2	3.1	3.2	6.1	6.2	10.1	10.2
		SLRGP							
EI	1%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ME		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IMSE		0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MMSE		0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0
GRF		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ALC		0.5	0.0	0.5	0.0	0.0	0.0	0.0	0.0
SLRV		0.0	0.0	0.0	0.1	0.4	0.0	0.0	0.8
IGS		0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0
GBAL		0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
EI	3%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ME		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IMSE		0.1	0.7	0.0	0.0	0.0	0.0	0.0	0.0
MMSE		0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0
GRF		0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
ALC		0.5	0.0	0.9	0.0	0.0	0.0	0.0	0.0
SLRV		0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.6
IGS		1.0	0.4	0.0	0.3	0.0	0.0	0.0	0.0
GBAL		0.0	0.4	0.9	0.0	0.0	0.0	0.0	0.0
EI	5%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ME		0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IMSE		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MMSE		0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
GRF		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ALC		0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
SLRV		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
IGS		0.0	0.7	0.8	0.1	0.0	0.0	0.0	0.0
GBAL		0.0	0.4	0.1	0.0	0.0	0.0	0.0	0.0

difference in the RMSE performance. As shown in the Table 3, the Wilcoxon rank test also signifies the improvement by the proposed method which further validates the earlier results.

V. CONCLUSION

We proposed an integrated methodology for active learning in Gaussian process regression to estimate black-box functions with a fewer number of points. For this purpose, we introduced a Laplacian regularization term to the popular active learning - Cohn (ALC) criteria to explore the transfer of information among evaluated and unevaluated points as well as predictor variables in a dynamic setting. We also developed two simple greedy algorithms for optimizing the tuning parameter, and sequential selection of the most informative subsequent points to evaluate. In addition, we extended the classical graph Laplacian matrix to consider the similarity between points in both predictor variables (input), and response variable (output) spaces to better capture the relationship between the points. For the development of the proposed SLRGP method, we considered a common scenario, in which evaluation points are to be selected from a set of pre-specified points. We used a case study for analysis of the kinematics of pitching in baseball and conducted a simulation study to evaluate the performance of the proposed methodology against popular methods in the literature in terms of root mean squared error (RMSE). The simulation results suggest the SLRGP algorithm provides better performance when there are plenty of unevaluated points available and the standard deviation of error is large. The models developed through this study can be used to reduce the number of points for estimating expensive noisy black-box functions.

APPENDIX
COMPARISON BETWEEN CLASSICAL AND PROPOSED GRAPH LAPLACIAN

In our study, we propose a sequential Laplacian regularized Gaussian process (SLRGP) algorithm based on Laplacian regularized active learning (LR-AL) by extending the active learning - Cohn (ALC) measure. Meanwhile, the proposed Laplacian penalty can be used to extend any measure that

selects the next evaluation point based on the uncertainty information, i.e. MMSE, maximum entropy, etc. We conduct several simulation studies to test different forms of graph Laplacian against classical graph Laplacian. Here, we provide a sample result based on response model 2.1 to compare the proposed graph Laplacian with the classical Laplacian based on ALC and also MMSE measures. Figure 8 illustrates the RMSE of the classical and proposed Laplacian using a set of 8 initial points augmented with 100 additional points. As shown in the Figure 8, the proposed graph Laplacian generally provides a better performance compared to the classical Laplacian for both MMSE and ALC measures and across different levels of noise. For the MMSE, even though the result is not as good as ALC, the proposed graph Laplacian is outperforming the classical Laplacian after as little as 20 additional points. For the ALC, when the number of evaluated points and the noise level are both small, there is not much difference between the classical and proposed Laplacian. However, as the number of evaluated points and the level of noise increase the proposed graph Laplacian performs better. Meanwhile, it should be noted that the remarkable performance of the proposed SLRGP is based on the combination of the proposed LR-AL measure and the proposed graph Laplacian.

REFERENCES

[1] K. K. Vu, C. D'Ambrosio, Y. Hamadi, and L. Liberti, "Surrogate-based methods for black-box optimization," *Int. Trans. Oper. Res.*, vol. 24, no. 3, pp. 393–424, May 2017.

[2] A. Alaeddini, E. Craft, R. Meka, and S. Martinez, "Sequential Laplacian regularized V-optimal design of experiments for response surface modeling of expensive tests: An application in wind tunnel testing," *IISE Trans.*, vol. 51, no. 5, pp. 559–576, May 2019.

[3] R. H. Myers, D. C. Montgomery, and C. M. Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Hoboken, NJ, USA: Wiley, 2016.

[4] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*, vol. 2. Cambridge, MA, USA: MIT Press, 2006, no. 3.

[5] J. Kocijan, R. Murray-Smith, C. E. Rasmussen, and A. Girard, "Gaussian process model based predictive control," in *Proc. Amer. Control Conf.*, 2004, pp. 2214–2219.

[6] H. (Heidi) Xia, Y. Ding, and J. Wang, "Gaussian process method for form error assessment using coordinate measurements," *IIE Trans.*, vol. 40, no. 10, pp. 931–946, Aug. 2008.

[7] S. L. Lauritzen, "Time series analysis in 1880: A discussion of contributions made by tn Thiele," in *Proc. Int. Stat. Review/Revue Internationale Statistique*, 1981, pp. 319–331.

[8] N. Wiener, "Cybernetýcs," *Sci. Sayı Sci. Amer.*, vol. 179, no. 5, pp. 14–19, 1948.

[9] D. J. MacKay, "Gaussian processes-a replacement for supervised neural networks?" Cavendish Lab., Cambridge Univ., Cambridge, U.K., Lecture Notes, 1997.

[10] G. Matheron, "Principles of geostatistics," *Econ. Geol.*, vol. 58, no. 8, pp. 1246–1266, Dec. 1963.

[11] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and analysis of computer experiments," *Stat. Sci.*, vol. 4, pp. 409–423, Oct. 1989.

[12] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. Berlin, Germany: Springer, 2012.

[13] C. K. Williams and C. E. Rasmussen, "Gaussian processes for regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 1996, pp. 514–520.

[14] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[15] T. J. Santner, B. J. Williams, W. Notz, and B. J. Williams, *Th Design and Analysis of Computer Experiments*. vol. 1. New York, NY, USA: Springer, 2003.

[16] L. Pronzato and W. G. Müller, "Design of computer experiments: Space filling and beyond," *Statist. Comput.*, vol. 22, no. 3, pp. 681–701, May 2012.

[17] M. E. Johnson, L. M. Moore, and D. Ylvisaker, "Minimax and maximin distance designs," *J. Stat. Planning Inference*, vol. 26, no. 2, pp. 131–148, Oct. 1990.

[18] K.-T. Fang, D. K. J. Lin, P. Winker, and Y. Zhang, "Uniform design: Theory and application," *Technometrics*, vol. 42, no. 3, pp. 237–248, Aug. 2000.

[19] M. Seeger, "Learning with labeled and unlabeled data," Edinburgh Univ., Edinburgh, U.K., Lecture Notes 161327, 2000.

[20] X. Chen and T. Wang, "Combining active learning and semi-supervised learning by using selective label spreading," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017.

[21] X. He, W. Min, D. Cai, and K. Zhou, "Laplacian optimal design for image retrieval," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2007, pp. 119–126.

[22] X. He, "Laplacian regularized D-Optimal design for active learning and its application to image retrieval," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 254–263, Jan. 2010.

[23] L. Zhang, C. Chen, J. Bu, D. Cai, X. He, and T. S. Huang, "Active learning based on locally linear reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2026–2038, Oct. 2011.

[24] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, pp. 129–145, Mar. 1996.

[25] D. J. C. MacKay, "Information-based objective functions for active data selection," *Neural Comput.*, vol. 4, no. 4, pp. 590–604, Jul. 1992.

[26] N. Ramakrishnan, C. Bailey-Kellogg, S. Tadepalli, and V. N. Pandey, "Gaussian processes for active data mining of spatial aggregates," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2005, pp. 427–438.

[27] S. Seo, M. Wallat, T. Graepel, and K. Obermayer, "Gaussian process regression: Active data selection and test point rejection," in *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Netw.*, 2000, pp. 27–34.

[28] E. Pasolli and F. Melgani, "Gaussian process regression within an active learning scheme," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 3574–3577.

[29] M. C. Shewry and H. P. Wynn, "Maximum entropy sampling," *J. Appl. Statist.*, vol. 14, no. 2, pp. 165–170, 1987.

[30] D. Wu, C.-T. Lin, and J. Huang, "Active learning for regression using greedy sampling," *Inf. Sci.*, vol. 474, pp. 90–105, Feb. 2019.

[31] H. Zhang, S. Ravi, and I. Davidson, "A graph-based approach for active learning in regression," in *Proc. SIAM Int. Conf. Data Mining*, 2020, pp. 280–288.

[32] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *J. Global Optim.*, vol. 13, no. 4, pp. 455–492, 1998.

[33] E. Vazquez, J. Villemonteix, M. Sidorkiewicz, and E. Walter, "Global optimization based on noisy evaluations: An empirical study of two statistical approaches," *J. Phys., Conf. Ser.*, vol. 135, no. 1, 2008, Art. no. 012100.

[34] J. Villemonteix, E. Vazquez, and E. Walter, "An informational approach to the global optimization of expensive-to-evaluate functions," *J. Global Optim.*, vol. 44, no. 4, p. 509, 2009.

[35] D. Huang, T. T. Allen, W. I. Notz, and N. Zeng, "Global optimization of stochastic black-box systems via sequential kriging meta-models," *J. Global Optim.*, vol. 34, no. 3, pp. 441–466, Mar. 2006.

[36] R. Benassi, J. Bect, and E. Vazquez, "Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion," in *Proc. Int. Conf. Learn. Intell. Optim.* Berlin, Germany: Springer, 2011, pp. 176–190.

[37] R. Preuss and U. von Toussaint, "Global optimization employing Gaussian process-based Bayesian surrogates," *Entropy*, vol. 20, no. 3, p. 201, Mar. 2018.

[38] R.-B. Chen, W. Wang, and C. F. J. Wu, "Building surrogates with overcomplete bases in computer experiments with applications to bistable laser diodes," *IIE Trans.*, vol. 43, no. 1, pp. 39–53, Oct. 2010.

[39] R.-B. Chen, W. Wang, and C. F. J. Wu, "Sequential designs based on Bayesian uncertainty quantification in sparse representation surrogate modeling," *Technometrics*, vol. 59, no. 2, pp. 139–152, Apr. 2017.

[40] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, nos. 3–4, pp. 285–294, Dec. 1933.

[41] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2, pp. 235–256, 2002.

[42] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 2249–2257.
[43] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. Conf. Learn. Theory*, 2012, pp. 1–39.
[44] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz, "Kullback-Leibler upper confidence bounds for optimal sequential allocation," *Ann. Statist.*, vol. 41, no. 3, pp. 1516–1541, Jun. 2013.
[45] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," in *Proc. Eur. Conf. Mach. Learn.* Porto, Portugal: Springer, 2005, pp. 437–448.
[46] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1583–1591.
[47] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," *J. Mach. Learn. Res.*, vol. 9, no. 2, pp. 235–284, 2008.
[48] I. BEN-GAL and M. Caramanis, "Sequential DOE via dynamic programming," *IIE Trans.*, vol. 34, no. 12, pp. 1087–1100, Dec. 2002.
[49] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.
[50] W. Deng, Q. Zheng, and L. Chen, "Regularized extreme learning machine," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, Mar. 2009, pp. 389–395.
[51] K. I. Kim, F. Steinke, and M. Hein, "Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 979–987.
[52] Y. Lei, L. Cen, X. Chen, and Y. Xie, "A hybrid regularization semi-supervised extreme learning machine method and its application," *IEEE Access*, vol. 7, pp. 30102–30111, 2019.
[53] S. C. H. Hoi and M. R. Lyu, "A semi-supervised active learning framework for image retrieval," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Dec. 2005, pp. 302–309.
[54] P. Shen, X. Du, and C. Li, "Distributed semi-supervised metric learning," *IEEE Access*, vol. 4, pp. 8558–8571, 2016.
[55] Y. Zhou, B. Liu, S. Xia, and B. Liu, "Semi-supervised extreme learning machine with manifold and pairwise constraints regularization," *Neurocomputing*, vol. 149, pp. 180–186, Feb. 2015.
[56] Y. Gu, Y. Chen, J. Liu, and X. Jiang, "Semi-supervised deep extreme learning machine for Wi-Fi based localization," *Neurocomputing*, vol. 166, pp. 282–293, Oct. 2015.
[57] J. Liu, Y. Chen, M. Liu, and Z. Zhao, "SELM: Semi-supervised ELM with application in sparse calibrated location estimation," *Neurocomputing*, vol. 74, no. 16, pp. 2566–2572, Sep. 2011.
[58] G. Huang, S. Song, J. N. D. Gupta, and C. Wu, "Semi-supervised and unsupervised extreme learning machines," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2405–2417, Dec. 2014.
[59] Q. She, B. Hu, H. Gan, Y. Fan, T. Nguyen, T. Potter, and Y. Zhang, "Safe semi-supervised extreme learning machine for eeg signal classification," *IEEE Access*, vol. 6, pp. 49399–49407, 2018.
[60] H. Gan, N. Sang, and X. Chen, "Semi-supervised kernel minimum squared error based on manifold structure," in *Proc. Int. Symp. Neural Netw.* Dalian, China: Springer, 2013, pp. 265–272.
[61] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.
[62] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua, "Interactive video indexing with statistical active learning," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 17–27, Feb. 2012.
[63] Z. Liu, Z. Lai, W. Ou, K. Zhang, and R. Zheng, "Structured optimal graph based sparse feature extraction for semi-supervised learning," *Signal Process.*, vol. 170, May 2020, Art. no. 107456.
[64] X. Zhu, J. D. Lafferty, and Z. Ghahramani, "Semi-supervised learning: From Gaussian fields to Gaussian processes," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-03-175, 2003.
[65] V. R. Joseph, "Space-filling designs for computer experiments: A review," *Qual. Eng.*, vol. 28, no. 1, pp. 28–35, Jan. 2016.
[66] A. D. MacCalman, "Flexible space-filling designs for complex system simulations," Ph.D. dissertation, Naval Postgraduate School, Monterey CA, USA, 2013.
[67] M. D. McKay, R. J. Beckman, and W. J. Conover, "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, May 1979.
[68] R. B. Gramacy and H. K. H. Lee, "Adaptive design and analysis of supercomputer experiments," *Technometrics*, vol. 51, no. 2, pp. 130–145, May 2009.
[69] F. R. Chung and F. C. Graham, *Spectral Graph Theory*. Providence, RI, USA: American Mathematical Society, 1997, no. 92.
[70] S. Boyd, "Convex optimization of graph Laplacian eigenvalues," in *Proc. Int. Congr. Math.*, 2006, vol. 3, nos. 1–3, pp. 1311–1319.
[71] A. Marsden, "Eigenvalues of the Laplacian and their relationship to the connectedness of a graph," Univ. Chicago, Chicago, IL, USA, Res. Exper. Undergraduates Paper, 2013.
[72] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (GPML) toolbox," *J. Mach. Learn. Res.*, vol. 11, pp. 3011–3015, Nov. 2010.
[73] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, 2003, pp. 912–919.

RAJITHA MEKA received the bachelor's degree in mechanical engineering from Acharya Nagarjuna University, India, in 2010, and the master's degree in industrial engineering from the University of Houston, USA, in 2016. She is currently pursuing the Ph.D. degree in mechanical engineering with The University of Texas at San Antonio, USA. She worked for Hyundai Motors, India, from 2010 to 2014. Her research interests include design of experiments, nonparametric regression, and data analytics.

ADEL ALAEDDINI received the Ph.D. degree in industrial and systems engineering from Wayne State University. He also did a Postdoctoral Research with the University of Michigan, Ann Arbor. He is currently an Associate Professor of mechanical engineering with The University of Texas at San Antonio. He has contributed to over 35 peer-reviewed publications in journals, such as IEEE Access, *IISE Transactions*, *Production and Operations Management* (POMS), and *Information Sciences*. His main research interests include statistical learning in systems modeling and control and data analytics in health care and manufacturing.

SAKIKO OYAMA received the Ph.D. degree in human movement science from the University of North Carolina at Chapel Hill. She is currently an Associate Professor with the Department of Kinesiology, Health, and Nutrition, The University of Texas at San Antonio. She has contributed to over 30 peer-reviewed publications in journals, such as the *American Journal of Sports Medicine*, *Journal of Athletic Training*, and the *Clinical Journal of Biomechanics*. Her main research interest includes prevention and rehabilitation of shoulder and elbow injuries in youth/adolescent baseball pitchers through identification of biomechanical risk factors.

KRISTINA LANGER received the Ph.D. degree from Stanford University, where she developed multiscale models of dynamic crack propagation to explain experimentally observed nonplanar fracture in brittle materials. She is currently an Aerospace Engineer with the Air Force Research Laboratory, Wright-Patterson Air Force Base, where she serves as the Technology Advisor for basic and applied research in aerospace structural integrity and advanced structural concepts. She also works in the area of engineered residual stress (ERS) for airframe fatigue life enhancement, with a specific focus on developing modeling tools to enable cost-effective and timely deployment for residual stress solutions to fatigue challenges in the fielded air force fleet.