ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095008

BAYESIAN OPTIMIZATION WITH ENSEMBLE LEARNING MODELS AND ADAPTIVE EXPECTED IMPROVEMENT

Konstantinos D. Polyzos, Qin Lu, Georgios B. Giannakis

Dept. of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA

ABSTRACT

Optimizing a *black-box* function that is *expensive* to evaluate emerges in a gamut of machine learning and artificial intelligence applications including drug discovery, policy optimization in robotics, and hyperparameter tuning of learning models to list a few. Bayesian optimization (BO) provides a principled framework to find the global optimum of such functions using a limited number of function evaluations. BO relies on a statistical surrogate model to actively select new query points, that is typically captured by a Gaussian process (GP). Unlike most existing approaches that hinge on a *single* GP surrogate model with a *pre-selected* kernel function that may confine the expressiveness of the sought function especially under the limited evaluation budget, the present work puts forth a *weighted ensemble* of GPs as a surrogate model. Building on the advocated Gaussian mixture (GM) posterior, the EGP framework adapts to the most fitted surrogate model as data arrive on-the-fly, offering a richer function space. For the acquisition of next evaluation points, the EGP-based posterior is coupled with an adaptive expected improvement (EI) criterion to balance exploration and exploitation of the search space. Numerical tests on a set of benchmark synthetic functions and two robotic tasks, demonstrate the impressive benefits of the proposed approach.

Index Terms— Bayesian optimization, Gaussian processes, ensemble learning, expected improvement, adaptive learning

1. INTRODUCTION

In machine learning and artificial intelligence, several major tasks boil down to optimizing a function. When the analytic expression of the function is *known*, plain-vanilla optimization techniques can be applied depending on the nature of the optimization function; e.g convexity and nonlinearities. Nonetheless, these methods may not be applicable in practical settings where the function is *unknown* and/or each function evaluation is *costly*, such as hyperparameter tuning [1], robotics [2, 3], sensor networks [4], and drug discovery [5] to name a few. Bayesian optimization (BO) provides a principled framework to efficiently and effectively optimize a black-box function capitalizing on a statistical surrogate model for the black-box function that enables the sequential acquisition of new query points [6, 7]. Gaussian processes (GPs) are widely adopted as a surrogate model in various BO settings due to their ability to learn a non-parametric function with data efficiency and additional uncertainty quantification [8].

Building on the GP surrogate model, there exist several acquisition criteria or *acquisition functions* (AFs) to select query points on-the-fly, including expected improvement (EI) [9], Thomson sampling (TS) [10], upper confidence bound (UCB), [11], and entropy search (ES) [12]. The present work will focus on the EI criterion because of its well-documented merits in balancing exploration and exploitation of the search space [7, 9]. In [1], the 'EI per second' criterion is adopted that aims to acquire new query points that are not only closer to the global optimum solution(s) but are also quick to evaluate. To effectively handle highly noisy observations and constrained BO problems, the work in [13] introduces a constrained EI criterion that is efficiently optimized via quasi-Monte Carlo approximation. To further reduce convergence time, the EI criterion can be coupled with a parallel operation with function evaluations distributed at different computing resources, where extra hyperparameters or selection rules are needed to guarantee the acquisition of diverse query points at different locations [14]. Albeit interesting, these approaches use a single GP surrogate model whose performance hinges on a *pre-selected* kernel function that may confine function space expressiveness.

Kernel selection is a critical component of GP surrogate models in BO. Several existing approaches to discover the form of the kernel function typically operate in a batch mode and require a large number of data which may become prohibitive in the BO context where data points are scarce due to costly evaluations, and are acquired in an online fashion; see e.g., [15–17]. Without any prior information about the BO problem at hand, selecting the form of the kernel function is a nontrivial task. Alternatively, one can resort to ensemble methods by combining the benefits of different approaches, that have markedly improved the empirical performance in

This work was supported in part by NSF grants 1901134, 2128593, 2126052, 2212318, and 2220292. The work of Konstantinos D. Polyzos was also suppoorted by the Onassis Foundation Scholarship. Emails: {polyz003,qlu,georgios}@umn.edu

hyperparameter tuning tasks [18] and different contexts such as high-dimensional inputs [19]. Ensembling rules have been used for combining different acquisition criteria for BO given a single GP surrogate model [20, 21], but the complementary setup of adopting an ensemble of surrogate models given a specific AF for BO has not been explored in existing literature. Although a recently developed GP-based online kernel selection framework is used for conventional prediction-oriented and graph-guided learning tasks [22–24], the notion of GP ensembles for the BO context, where extra design of the acquisition step is necessitated, has *not* been touched upon yet.

Contributions. To ensure a richer function space than that of a single GP with a pre-selected kernel function, the present work puts forth a weighted ensemble (E) of GPs as the surrogate model for BO that sequentially adjusts to the proper model fit by judiciously updating the per-GP weight as new data arrive on-the-fly. The novel EGP-based surrogate model is coupled with an adaptive and properly adjusted EI-based acquisition criterion to effectively balance exploration and exploitation of the search space. Numerical tests on synthetic benchmark functions and two robotic tasks, corroborate the benefits of the advocated EGP-EI approach compared to the single GP-based EI counterparts.

2. PRELIMINARIES

Let us consider the following optimization problem

$$\mathbf{x}_* = \arg\max_{\mathbf{x}\in\mathcal{X}} \; f(\mathbf{x}) \tag{1}$$

where $\mathcal{X}$ denotes the set comprising all feasible values of the $d \times 1$ optimization variable $\mathbf{x}$, and $f(\mathbf{x})$ is a *black-box* objective function whose analytic expression is unknown and/or is *expensive* to evaluate. For instance, in the hyperparameter tuning task of machine learning models where $\mathbf{x}$ consists of all the hyperparameters to be tuned and $f(\mathbf{x})$ represents the mapping from $\mathbf{x}$ to the validation accuracy, the latter cannot be expressed analytically and each evaluation is computationally costly, especially for large data sizes and deep learning architectures [1]. Since f is not expressed analytically, plain vanilla gradient-based methods are not suitable for obtaining $\mathbf{x}_*$ and due to the expensive function evaluations, an exhaustive enumeration is impractical. To overcome these challenges, BO judiciously selects query pairs for a given evaluation budget in a data-efficient manner [6, 7].

Capitalizing on input-output evaluation pairs collected at $\mathcal{D}_t := \{(\mathbf{x}_\tau, y_\tau)\}_{\tau=1}^t$, BO typically leverages a statistical surrogate model for f, to obtain the next query input $\mathbf{x}_{t+1}$. Building on this surrogate model, the so-termed acquisition function $\alpha(\cdot)$, often given in closed form, is employed to balance *exploration* and *exploitation* of the search space. Specifically, each iteration of the BO process alternates between the following steps

s1. Obtain $p(f(\mathbf{x})|\mathcal{D}_t)$ based on the surrogate model;

s2. Obtain $\mathbf{x}_{t+1} = \arg\max_{\mathbf{x}\in\mathcal{X}} \alpha_{t+1}(\mathbf{x}|\mathcal{D}_t)$ given $p(f(\mathbf{x})|\mathcal{D}_t)$.

Although there exist several choices for both the surrogate model and the acquisition function, we will focus on the widely adopted GP-based surrogate model and the EI acquisition function with well-documented benefits; see e.g [6, 7].

2.1. GP-based surrogate model and EI acquisition

GPs offer a nonparametric Bayesian approach to learn an unknown function along with its corresponding probability density function (pdf) in a sample-efficient manner, which is of great interest in the BO setting where each function evaluation is expensive. To learn the function $f(\cdot)$ that maps the input vector $\mathbf{x}_\tau$ to the scalar output y_τ as $\mathbf{x}_\tau \to f(\mathbf{x}_\tau) \to y_\tau$, a GP prior is postulated on f as $f \sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}'))$ where $\kappa(\cdot,\cdot)$ is a positive-definite kernel function that captures the pairwise similarity between $\mathbf{x}$ and $\mathbf{x}'$. This implies that the random vector $\mathbf{f}_t := [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_t)]^\top$ comprising all function evaluations at $\mathbf{X}_t := [\mathbf{x}_1, \ldots, \mathbf{x}_t]^\top$ $(\forall t)$ is Gaussian distributed as $\mathbf{f}_t \sim \mathcal{N}(\mathbf{f}_t; \mathbf{0}_t, \mathbf{K}_t)$ with $\mathbf{K}_t$ denoting the $t \times t$ kernel (covariance) matrix whose (m, m') entry is $[\mathbf{K}_t]_{m,m'} = \mathrm{cov}(f(\mathbf{x}_m), f(\mathbf{x}_{m'})) := \kappa(\mathbf{x}_m, \mathbf{x}_{m'})$ [8].

Focusing on the regression task, y_τ can be expressed as $y_\tau = f(\mathbf{x}_\tau) + n_\tau$ $(\forall \tau)$ where the noise sequence is independently and identically distributed (iid) as: $n_\tau \sim \mathcal{N}(0, \sigma_n^2)$. Equivalently, the collection of outputs $\mathbf{y}_t := [y_1, \ldots, y_t]^\top$ are related to $\mathbf{f}_t$ via the batch conditional likelihood as $p(\mathbf{y}_t|\mathbf{f}_t; \mathbf{X}_t) = \prod_{\tau=1}^t p(y_\tau|f(\mathbf{x}_\tau)) = \prod_{\tau=1}^t \mathcal{N}(y_\tau; f(\mathbf{x}_\tau), \sigma_n^2)$. Then for any input $\mathbf{x}$ the joint pdf of $f(\mathbf{x})$ and $\mathbf{y}_t$ is

$$\begin{bmatrix} \mathbf{y}_t \\ f(\mathbf{x}) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}_{t+1}, \begin{bmatrix} \mathbf{K}_t + \sigma_n^2 \mathbf{I}_t & \mathbf{k}_t(\mathbf{x}) \\ \mathbf{k}_t^\top(\mathbf{x}) & \kappa(\mathbf{x}, \mathbf{x}) + \sigma_n^2 \end{bmatrix}\right)$$

where $\mathbf{k}_t(\mathbf{x}) := [\kappa(\mathbf{x}_1, \mathbf{x}), \ldots, \kappa(\mathbf{x}_t, \mathbf{x})]^\top$. With the joint pdf at hand, the function posterior pdf of $f(\mathbf{x})$ is [8]

$$p(f(\mathbf{x})|\mathcal{D}_t) = \mathcal{N}(f(\mathbf{x}); \mu_t(\mathbf{x}), \sigma_t^2(\mathbf{x})) \tag{2}$$

with mean and variance expressed in closed form as follows

$$\mu_t(\mathbf{x}) = \mathbf{k}_t^\top(\mathbf{x})(\mathbf{K}_t + \sigma_n^2 \mathbf{I}_t)^{-1} \mathbf{y}_t \tag{3a}$$

$$\sigma_t^2(\mathbf{x}) = \kappa(\mathbf{x}, \mathbf{x}) - \mathbf{k}_t^\top(\mathbf{x})(\mathbf{K}_t + \sigma_n^2 \mathbf{I}_t)^{-1} \mathbf{k}_t(\mathbf{x}). \tag{3b}$$

Capitalizing on the function posterior pdf, the next evaluation point can be obtained by utilizing the so-termed EI acquisition function, which is given by [9]

$$\alpha_{t+1}^{\mathrm{EI}}(\mathbf{x}|\mathcal{D}_t) := \mathbb{E}_{p(f(\mathbf{x})|\mathcal{D}_t)}[\max(0, f(\mathbf{x}) - \hat{f}_t^{\max})] \tag{4}$$

where $\hat{f}_t^{\max}$ is an estimate of the maximum function value at slot t, which is typically given by $\hat{f}_t^{\max} = \max(y_1, \ldots, y_t)$; see e.g., [7, 9]. For the single GP-based surrogate model

where the function posterior pdf in (2) is Gaussian, the EI acquisition function in (4) can be re-written as

$$\alpha_{t+1}^{\text{EI}}(\mathbf{x}|\mathcal{D}_t) = \sigma_t(\mathbf{x})\phi\left(\frac{\Delta_t(\mathbf{x})}{\sigma_t(\mathbf{x})}\right) + \Delta_t(\mathbf{x})\Phi\left(\frac{\Delta_t(\mathbf{x})}{\sigma_t(\mathbf{x})}\right) \quad (5)$$

with $\Delta_t(\mathbf{x}) := \mu_t(\mathbf{x}) - \hat{f}_t^{\max}$ and $\phi(\cdot), \Phi(\cdot)$ denoting the Gaussian pdf and Gaussian cumulative density function (cdf) respectively.

The EI criterion is employed in several practical BO settings, since it can readily balance exploration and exploitation [7, 9]. Nonetheless, the EI criterion is coupled here with the single GP based surrogate model that relies on a *pre-selected* kernel function $\kappa(\cdot)$, which may exhibit limited expressiveness of the sought function f, thus motivating the ensemble (E)GP surrogate model as outlined next.

3. ENSEMBLE GPS SURROGATE MODEL AND EI

Targeting at a richer function space, the present work advocates an ensemble of M GP models as a surrogate model for the black-box function f. Each GP model hinges on a distinct kernel function from a kernel dictionary $\mathcal{K} := \{\kappa_1, \ldots, \kappa_M\}$ that consists of kernels with different hyperparameters and of different types as in [22]. Specifically per GP model $m \in \mathcal{M} := \{1, \ldots, M\}$, a unique GP prior is postulated as $f|m \sim \mathcal{GP}(0, \kappa^m(\mathbf{x}, \mathbf{x}'))$. Further combining these individual GP priors with the weights $\{w_0^m\}_{m=1}^M$ yields the EGP prior expressed as

$$f(\mathbf{x}) \sim \sum_{m=1}^{M} w_0^m \mathcal{GP}(0, \kappa^m(\mathbf{x}, \mathbf{x}')), \quad \sum_{m=1}^{M} w_0^m = 1 \quad (6)$$

which is a Gaussian mixture (GM) with each weight $w_0^m := \Pr(i = m)$ denoting the probability that measures the contribution of the corresponding GP model in the EGP surrogate model. Although the notion of using a GM as an EGP prior has been employed in typical prediction-oriented tasks [22], the novelty of this work lies on its adaptation in the BO setting, where additional design step is required to select the next input vector $\mathbf{x}_{t+1}$ at the end of slot t. Relying on the EGP prior in (6) and the set of evaluated data $\mathcal{D}_t$, the EGP posterior pdf can be obtained via the sum-product rule as follows

$$p(f(\mathbf{x})|\mathcal{D}_t) = \sum_{m=1}^{M} \Pr(i=m|\mathcal{D}_t) p(f(\mathbf{x})|i=m, \mathcal{D}_t) \quad (7)$$

which is a GM posterior pdf with per-GP model weight $w_t^m := \Pr(i=m|\mathcal{D}_t)$ being computed via Bayes' rule as

$$w_t^m \propto \Pr(i=m) p(\mathcal{D}_t|i=m) = w_0^m p(\mathcal{D}_t|i=m) \quad (8)$$

where for GP model m, $p(\mathcal{D}_t|i=m)$ denotes the marginal likelihood of the acquired data $\mathcal{D}_t$ up to slot t, which is expressed as

$$p(\mathcal{D}_t|i=m) = \int p(\mathbf{y}_t|\mathbf{f}_t, i=m; \mathbf{X}_t) p(\mathbf{f}_t|i=m; \mathbf{X}_t) d\mathbf{f}_t$$

$$= \mathcal{N}(\mathbf{y}_t; \mathbf{0}_t, \mathbf{K}_t^m + (\sigma_n^m)^2 \mathbf{I}_t) \quad (9)$$

where $\mathbf{K}_t^m$ and $(\sigma_n^m)^2$ denote the kernel (covariance) matrix and noise variance of the mth GP model respectively. Note that the per-GP model kernel hyperparameters along with the noise variance are estimated at every iteration by optimizing the marginal likelihood [6].

Leveraging the posterior pdf in (7) along with the corresponding weights in (8), the acquisition of the next instance $\mathbf{x}_{t+1}$ to be queried is carried out by first selecting a specific GP model from the ensemble as follows

$$m_t \sim \mathcal{CAT}(\mathcal{M}, \mathbf{w}_t) \quad (10)$$

with $\mathcal{CAT}(\mathcal{M}, \mathbf{w}_t)$ denoting a categorical distribution that selects one of the values from $\mathcal{M}$ with probabilities $\mathbf{w}_t := [w_t^1, \ldots, w_t^M]^\top$. Then, $\mathbf{x}_{t+1}$ is obtained through the novel EGP-based EI acquisition criterion as

$$\mathbf{x}_{t+1} = \arg\max_{\mathbf{x} \in \mathcal{X}} \ \alpha_{t+1}^{\text{EGP-EI}}(\mathbf{x}|\mathcal{D}_t), \quad (11)$$

$$\alpha_{t+1}^{\text{EGP-EI}}(\mathbf{x}|\mathcal{D}_t) = \sigma_t^{m_t}(\mathbf{x})\phi\left(\frac{\Delta_t^{m_t}(\mathbf{x})}{\sigma_t^{m_t}(\mathbf{x})}\right) + \Delta_t^{m_t}(\mathbf{x})\Phi\left(\frac{\Delta_t^{m_t}(\mathbf{x})}{\sigma_t^{m_t}(\mathbf{x})}\right)$$

where $\Delta_t^{m_t}(\mathbf{x}) := \mu_t^{m_t}(\mathbf{x}) - \hat{f}_t^{\max}$, and $\mu_t^{m_t}(\mathbf{x})$ and $\sigma_t^{m_t}(\mathbf{x})$ represent the Gaussian posterior mean and variance of the m_tth GP model respectively. Intuitively, the larger the weight w_t^m of GP model m is, the more probable the latter is to be utilized in the EI criterion in (11) at slot t. In that sense, the EGP-based EI criterion properly adjusts to the m_tth GP model at each slot t as new data arrive on-the-fly.

Remark. Instead of sampling a single GP model from the ensemble as in (10), one can directly apply the EI criterion in (4) using the EGP posterior pdf in (7); though since the latter is a GM, the EI criterion cannot be written in the form of (5). The adoption of the EGP posterior pdf in (4) belongs to our future research agenda.

4. NUMERICAL TESTS

In this section, the performance of the novel EGP-EI approach is assessed on a set of benchmark synthetic functions and two robotic tasks as detailed next.

Synthetic functions. Three standard synthetic functions for BO are employed to corroborate the effectiveness of the advocated EGP-EI; that is, Ackley5d, Zakharov, and Drop-wave [25] with the latter being a particularly challenging function to optimize due to its multiple local optima.

Robot pushing tasks. The EGP-EI method is also evaluated on a practical robotic task, where a robot needs to select the proper action so as to push an object towards a pre-specified target location. Following [26], we have tested two different scenarios, namely, 'robot pushing 3D' and 'robot pushing 4D', where the former optimizes the 2D position coordinates of the robot and the push duration and the latter additionally optimizes the push angle. Here, the objective is to minimize the distance between the target and the terminal location.

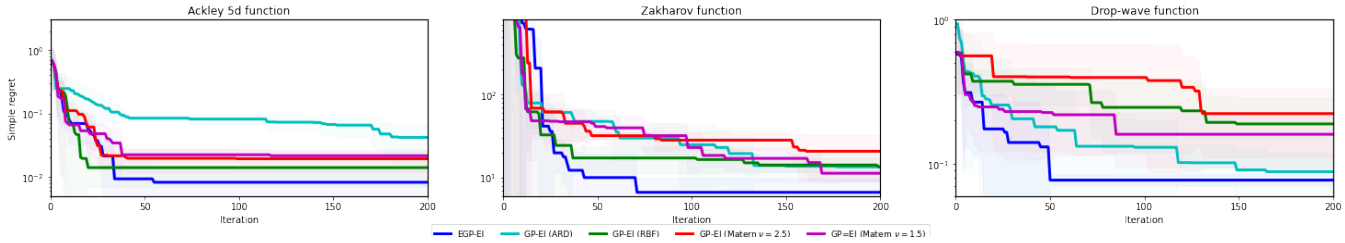

Fig. 1: Simple regret performance of EGP-EI and single GP-EI baselines on `Ackley-5d`, `Zakharov`, and `DropWave` function (from left to right). The kernel dictionary comprises four different kernels: RBF with(out) ARD and Matérn with $\nu = 1.5, 2.5$.

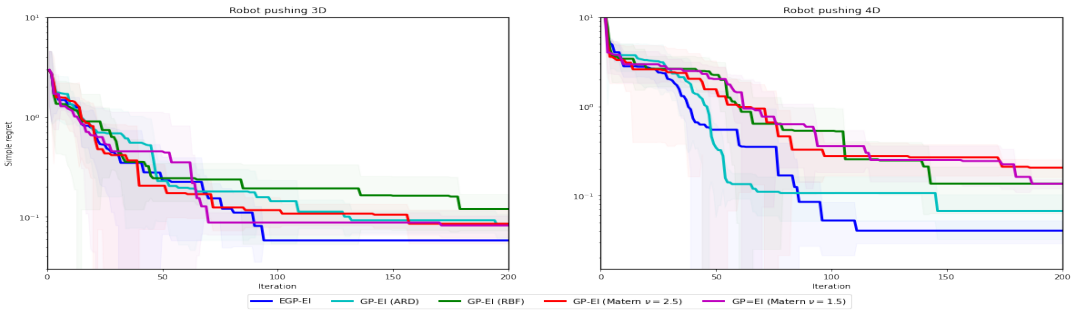

Fig. 2: Simple regret performance of EGP-EI and single GP-EI baselines on `Robot pushing 3D`, `Robot pushing 4D` tasks (from left to right). The kernel dictionary comprises four different kernels: RBF with(out) ARD and Matérn with $\nu = 1.5, 2.5$.

The trajectory of the pushed object is generated utilizing the github code[1] in [26].

We compare the performance of the EGP-EI method with the single GP-EI counterpart with kernel being pre-selected as an RBF with and without auto-relevance determination (ARD), and Matern kernel with $\nu = 1.5$ and $\nu = 2.5$ respectively. The kernel dictionary of EGP-EI consists of these four kernels; i.e., $M = 4$ and the weights of the M GP models are initialized as $w_0^m = 1/M$, $\forall m \in \mathcal{M}$. For all competing approaches, 10 initial evaluation pairs are used to obtain the kernel hyperparameters for each GP model maximizing the marginal log-likelihood, and the hyperparameters are then refitted every iteration. As a figure of merit, the simple regret (SR) metric is utilized which per slot t is expressed as $\mathcal{SR}(t) := f(\mathbf{x}_*) - \max_{\tau \in \{1,\ldots,t\}} f(\mathbf{x}_\tau)$.

The average performance of all competing approaches along with the corresponding standard deviation are reported for 10 independent runs. As shown in Fig. 1, the advocated EGP-EI approach not only enjoys the lowest SR at the end of the BO process but also converges faster to the corresponding minimum SR value compared to the single GP-based counterparts. In the robotic tasks, it is evident in Fig. 2 that although the EGP-EI method requires more iterations to converge compared to the synthetic benchmark functions, it consistently outperforms all competing alternatives in terms of SR upon 80 iterations in both tasks. The superior performance of EGP-EI in all cases corroborates the merits of adopting an ensemble of GPs as a surrogate model, which not only offers a more expressive function space but can readily guide the EI-based acquisition step in the BO process.

[1] https://github.com/zi-w/Max-value-Entropy-Search

5. CONCLUSIONS AND FUTURE DIRECTIONS

This work puts forth a novel weighted ensemble of GPs as a surrogate model for BO. Building on the resultant GM posterior pdf with adaptive weights being updated on-the-fly, the advocated approach sequentially selects the proper surrogate model fit, bypassing the need for selecting a priori a specific kernel that may exhibit limited function space expressiveness. Based on the EGP surrogate model, the acquisition of new query points is carried out by utilizing a properly adjusted EI acquisition criterion. The empirical performance on both synthetic functions and practical robotic tasks showcases the merits of the proposed EGP-EI approach.

Future directions include the combination of the advocated EGP surrogate model with other acquisition criteria such as Thomson sampling (TS) and upper confidence bound (UCB), and theoretical analysis for the convergence of the proposed method to the global optimum through the notion of Bayesian regret.

6. REFERENCES

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

[2] A. Cully, J. Clune, D. Tarapore, and J.-B. Mouret, "Robots that can adapt like animals," *Nature*, vol. 521, no. 7553, pp. 503–507, 2015.

[3] R. Marchant and F. Ramos, "Bayesian optimisation for informative continuous path planning," in *Proc. Int. Conf. Robotics Autom.*, 2014, pp. 6136–6143.

[4] ——, "Bayesian optimisation for intelligent environmental monitoring," in *Int. Conf. Intel. Robots Sys.*, 2012, pp. 2242–2249.

[5] K. Korovina, S. Xu, K. Kandasamy, W. Neiswanger, B. Poczos, J. Schneider, and E. Xing, "Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations," *Proc. Int. Conf. Artif. Intel. and Stats.*, pp. 3393–3403, 2020.

[6] 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, 2015.

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

[8] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

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

[10] 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.

[11] N. Srinivas, A. Krause, S. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," in *Proc. Int. Conf. Mach. Learn.*, 2010.

[12] P. Hennig and C. J. Schuler, "Entropy search for information-efficient global optimization." *J. Mach. Learn. Res.*, vol. 13, no. 6, 2012.

[13] B. Letham, B. Karrer, G. Ottoni, and E. Bakshy, "Constrained bayesian optimization with noisy experiments," *Bayesian Analysis*, vol. 14, no. 2, pp. 495–519, 2019.

[14] J. Wang, S. C. Clark, E. Liu, and P. I. Frazier, "Parallel Bayesian global optimization of expensive functions," *arXiv preprint arXiv:1602.05149*, 2016.

[15] T. Teng, J. Chen, Y. Zhang, and B. K. H. Low, "Scalable variational Bayesian kernel selection for sparse Gaussian process regression," *Proc. AAAI Conf. Artif. Intel.*, vol. 34, no. 4, pp. 5997–6004, 2020.

[16] H. Kim and Y. W. Teh, "Scaling up the automatic statistician: Scalable structure discovery using Gaussian processes," *Proc. Int. Conf. Artif. Intel. and Stats.*, pp. 575–584, 2018.

[17] G. Malkomes, C. Schaff, and R. Garnett, "Bayesian optimization for automated model selection," *Proc. Adv. Neural Inf. Process. Syst.*, 2016.

[18] R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon, "Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020," *arXiv preprint arXiv:2104.10201*, 2021.

[19] Z. Wang, C. Gehring, P. Kohli, and S. Jegelka, "Batched large-scale Bayesian optimization in high-dimensional spaces," *Proc. Int. Conf. Artif. Intel. and Stats.*, pp. 745–754, 2018.

[20] M. Hoffman, E. Brochu, N. de Freitas *et al.*, "Portfolio allocation for bayesian optimization." *Proc. Conf. Uncerntainty in Artif. Intel.*, pp. 327–336, 2011.

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

[22] Q. Lu, G. Karanikolas, Y. Shen, and G. B. Giannakis, "Ensemble Gaussian processes with spectral features for online interactive learning with scalability," *Proc. Int. Conf. Artif. Intel. and Stats.*, pp. 1910–1920, 2020.

[23] K. D. Polyzos, Q. Lu, and G. B. Giannakis, "Ensemble Gaussian processes for online learning over graphs with adaptivity and scalability," *IEEE Trans. Sig. Process.*, vol. 70, pp. 17–30, 2022.

[24] ——, "Active sampling over graphs for Bayesian reconstruction with Gaussian ensembles," in *Proc. Asilomar Conf. Sig., Syst., Comput.*, 2022.

[25] "Synthetic functions," https://www.sfu.ca/~ssurjano.

[26] Z. Wang and S. Jegelka, "Max-value entropy search for efficient Bayesian optimization," *Proc. Int. Conf. Mach. Learn.*, pp. 3627–3635, 2017.