Weighted Ensembles for Adaptive Active Learning

Konstantinos D. Polyzos, *Student Member, IEEE,* Qin Lu, *Member, IEEE,*
and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—Labeled data can be expensive to acquire in several application domains, including medical imaging, robotics, computer vision and wireless networks to list a few. To efficiently train machine learning models under such high labeling costs, active learning (AL) judiciously selects the most informative data instances to label on-the-fly. This active sampling process can benefit from a statistical function model, that is typically captured by a Gaussian process (GP) with well-documented merits especially in the regression task. While most GP-based AL approaches rely on a single kernel function, the present contribution advocates an *ensemble* of GP (EGP) models with weights adapted to the labeled data collected incrementally. Building on this novel EGP model, a suite of acquisition functions emerges based on the uncertainty and disagreement rules. An adaptively weighted *ensemble* of EGP-based *acquisition functions* is advocated to further robustify performance. Extensive tests on synthetic and real datasets in the regression task showcase the merits of the proposed EGP-based approaches with respect to the single GP-based AL alternatives.**

***Index Terms*—Active learning, Gaussian processes, ensemble learning**

1. Introduction

Identifying machine learning models usually relies on a sufficient number of labeled input-output data pairs, which may not be feasible due to labeling costs or privacy concerns in a number of application domains, including healthcare [11], computer vision [15], robotics [43], graph signal processing [1], and wireless networks [42]. To cope with sampling cost constraints, active learning (AL) selects a set of most informative data to label incrementally [39], [45]. For instance, consider the channel modeling task in wireless communications, where the goal is to find the mapping from the environment features to the channel impulse response. How to select the few most representative input-output pairs to effectively and efficiently learn such a mapping is of great importance for subsequent tasks including beamforming, transmission, as well as sensing. This selection process can benefit from a statistical model for the learning function $f(\mathbf{x})$ that maps each input feature vector $\mathbf{x}_\tau$ to the output/label y_τ [6]. Capable of learning nonlinear functions with *uncertainty quantification* in a *sample-efficient* fashion, Gaussian processes (GPs) are widely adopted to model the aforementioned f in AL; see e.g., [15], [19]. Given labeled samples $\{\mathbf{x}_\tau, y_\tau\}_{\tau=1}^t$ collected in the set $\mathcal{L}_t$, GP modeling yields a function posterior probability density function (pdf) $p(f(\mathbf{x})|\mathcal{L}_t)$. For regression, the latter is Gaussian with mean and variance available in closed form, while the model uncertainty captured by the variance is essential to guide the selection of subsequent instances to be queried.

The expressiveness of GP models depends on how well the chosen covariance (kernel) is adapted to the data at hand. Typically, GP-based AL approaches rely on a single kernel with preselected form, which may exhibit limited expressiveness for AL, where labeled data are collected incrementally. On par with this online interactive operation, a key desideratum is a more expressive function space to improve adaptation to the labeled instances on-the-fly. Besides expressiveness of the function model, the AL performance is critically affected by the data acquisition rule. For regression, GP-AL typically leverages the function variance to select the next instance to be queried. Devising alternative acquisition rules for more expressive GP models is an open problem. Clearly, with multiple acquisition rules available, devising a strategy to combine and adapt them will be conducive to robust performance [1] across tasks.

To address the aforementioned challenges, our contributions here can be summarized in the following aspects.

c1) Relative to GP-based AL that relies on a single GP with a *preselected* kernel, we introduce *weighted ensembles* (E) of GPs for enhanced expressiveness with weights capturing the contributions of individual GP models, and adapting to labeled data collected online.

c2) Utilizing the advocated EGP model, we devise a suite of acquisition functions (AFs), including novel *weighted ensembles* of AFs that further robustify performance.

c3) Our thorough tests on synthetic and real data corroborate the impressive merits of GP and AF ensembles.

2. Related works

This section outlines the context of the present work.

Statistical models for AL. The performance of model-based AL depends critically on the chosen statistical model. GPs are widely used because they come with uncertainty quantification and sample efficiency [15], [19], [35]. However, most of the existing approaches operate in a batch mode and rely on a preselected GP kernel with limited expressiveness, which may fall short in characterizing the incrementally collected data in AL. To that end, existing approaches aim to accommodate online GP model updates and inference as new coming data are processed in an online fashion [2], [26], [28], [41], [49]. Yet, they rely on a single GP model which may confine the expressiveness of the sought function. Broadening the scope

The authors are with Dept. of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455. This work was supported by NSF grants 1901134, 2128593, 2126052, 2212318, and 2220292. K. D. Polyzos was also supported by the Onassis Foundation Scholarship. E-mails: polyz003@umn.edu;qlu@umn.edu; georgios@umn.edu

[1] In the present work, the term 'robustness' refers to the consistency or stability in the superior empirical performance of a certain method across different scenarios.

of a single GP, a mixture of GPs is viable by training each GP component on a subset of labeled data [50]. This GP mixture model can account for multi-modalities in the function space but needs to be refitted per iteration using nontrivial variational methods. On the other hand, the proposed EGP model requires minimal refitting efforts and trains each GP on the whole labeled set. Apart from GP mixtures, the work in [35] considers a fully Bayesian approach and can be viewed as an ensemble approach with infinite number of members since a pdf is maintained for the hyperparameters of GPs. Although effective, it relies on Hamiltonian Monte Carlo (HMC) sampling to approximate the posterior pdf of the hyperparameters which can prove to be computationally challenging. To alleviate this challenge, the work in [36] constructs a mixture of GPs relying on different sets of parameters drawn from a prior pdf with each set representing a GP. Besides GP-based statistical models, Bayesian convolutional networks have been leveraged for image data [9].

Acquisition rules for AL. The acquisition function (AF) also plays a performance-critical role in AL. Different AFs have been devised based on distinct selection criteria, learning tasks, as well as whether or not a statistical model is capitalized on; see e.g., [39]. In GP-AL settings, the quantifiable uncertainty offered by the function posterior variance is leveraged to build the variance [15], entropy [18], [30], and mutual information [18] - based AFs. Inspired by the intuition that uncertainty increases at instances far away from the labeled set, [22], [31], [47] adopt AFs that rely on the distance of an instance from the labeled set. Beyond statistical learning models for the regression task, linear and nonlinear regression learning models are coupled with the 'Expected Model Change' (EMC) AF to select instances that cause the largest change on the corresponding learning model [4]. In [3], the so-termed 'Query by Committee' (QBC) AF is employed, which selects the instance in which the committee members disagree the most. Similar disagreement based criteria are presented in [13]. Albeit interesting, these approaches fall short in uncertainty quantification which can effectively and efficiently guide the AL process.

Selection of AFs. How to appropriately select the AF from the available choices usually calls for domain expertise. It has been shown empirically that no acquisition rule excels in all tasks [12]. This motivates a strategy that can combine and adapt candidate selection rules. [23] and [17] learn or/and select the acquisition rule in a data-driven fashion, but necessitate additional training data to learn the acquisition rule offline with possibly prohibitive complexity. On the contrary, the advocated weighted ensemble of AFs method combines all candidate AFs on-the-fly without need for a training phase. Ensemble acquisition rule has also been leveraged in [12] by adapting the multi-armed bandit framework with each arm representing an acquisition rule. In spite of this similarity, the advocated differs from [12] in the following three aspects i) it relies on a small validation set to evaluate the performance of each individual EGP-based AF, whereas the per-AF loss in [12] is defined as the estimated test error; ii) rather than selecting a subset of the AFs based on a certain rule in [12], it selects the next query instance by optimizing a weighted combination of all AFs; and iii) it updates the per-AF weight by knowing the losses of all AFs, which is different from the bandit setting in [12], where only losses of the selected AFs are revealed. Part of this work is presented in [32] which accounts only for graph-guided learning.

Kernel selection for GPs. Adaptively choosing the form of the kernel from training data has been reported for conventional GP learning; see, e.g., [7], [16], [27], [44]. These approaches usually operate in a batch offline mode, and rely on a large number of samples – what discourages their use for AL, where data are not only acquired online, but are also scarce due to the potentially high labeling cost. Recently, an online scalable kernel selection approach has been introduced that combines a set of GP experts in a Bayesian model averaging fashion [24]. Still, the focus is not on AL, which entails additional design of the AF.

3. Preliminaries

This section outlines the motivation and preliminaries for the AL approach of interest.

Typical learning approaches boil down to estimating the mapping $f(\cdot)$ that relates the $d \times 1$ input feature vector $\mathbf{x}_\tau$ to the output y_τ (that is either a real number in regression or it belongs to a finite alphabet in classification) as $\mathbf{x}_\tau \rightarrow f(\mathbf{x}_\tau) \rightarrow y_\tau$. This estimation task relies on a sufficient number of labeled training samples $\{\mathbf{x}_\tau, y_\tau\}_{\tau=1}^{T}$. In several applications however, the input can be readily obtained whereas the corresponding label can be expensive to acquire due to sampling costs or privacy concerns. In healthcare for instance, many labels describing the medical condition of patients may not be revealed to preserve confidentiality. Faced with this challenge, one can resort to the AL paradigm, which judiciously and proactively selects the most *informative* instances to label so that the sought mapping can be inferred in a *sample-efficient* manner.

AL starts with a small-size set $\mathcal{L}_0 := \{(\mathbf{x}_\tau, y_\tau)\}_{\tau=-L_0+1}^{0}$ of labeled samples,[2] and a larger collection of unlabeled features $\mathcal{U}_0 := \{\mathbf{x}_\tau^u\}_{\tau=1}^{U_0}$ ($L_0 \ll U_0$). Given corresponding sets $\mathcal{L}_t$ and $\mathcal{U}_t$ up to index $t > 0$, model-based AL entails a statistical function model, namely the pdf $p(f(\mathbf{x})|\mathcal{L}_t)$. The latter is utilized by the so-termed *acquisition function* (AF) $\alpha(\cdot)$ to select the instance $\mathbf{x}_{t+1}$ from the corresponding unlabeled set $\mathcal{U}_t$, as [6]

$$\mathbf{x}_{t+1} = \arg\max_{\mathbf{x}\in\mathcal{U}_t} \ \alpha(\mathbf{x}; \mathcal{L}_t) \ . \tag{1}$$

Intuitively, α is chosen to guide exploration of the space $f(\cdot)$ belongs to, which hinges on quantifying the uncertainty of the belief model $p(f(\mathbf{x})|\mathcal{L}_t)$. Upon querying an *oracle* for the associated label y_{t+1}, the labeled set is then augmented with the new pair and the feature vector is removed from the unlabeled set, that is $\mathcal{L}_{t+1} := \mathcal{L}_t \cup \{(\mathbf{x}_{t+1}, y_{t+1})\}$ and $\mathcal{U}_{t+1} := \mathcal{U}_t \setminus \{\mathbf{x}_{t+1}\}$. Apparently, the two critical choices are the model for f, and the AF design for α. Focusing on the regression task, we will outline the GP-based model for f, and the associated acquisition rules next.

[2]The negative instance index here is used for notational brevity as more labeled data are included next.

Algorithm 1 RFF-based EGP-AL.

1: **Initialization**: $\mathcal{U}_0$, $\mathcal{L}_0$, $\mathcal{K}$;
2: **for** $t = 0, \ldots, T$ **do**
3: Obtain $p(f(\mathbf{x})|\mathcal{L}_t)$ via $\boldsymbol{\Xi}_t$;
4: Select $\mathbf{x}_{t+1}$ based on one from (18)- (20), (22)-(23);
5: Query the oracle to obtain y_{t+1};
6: $\mathcal{L}_{t+1}:=\mathcal{L}_t \cup \{(\mathbf{x}_{t+1}, y_{t+1})\}$, $\mathcal{U}_{t+1}:=\mathcal{U}_t \setminus \{\mathbf{x}_{t+1}\}$;
7: **end for**

A. GP-based active learning

GPs estimate a nonparametric function model in a sample-efficient manner, while also offering quantification of the associated model uncertainty [33]. Sample efficiency justifies their wide adoption in AL. The GP model postulates f as being randomly drawn from a GP prior; that is $f \sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}'))$ with $\kappa(\mathbf{x}, \mathbf{x}')$ being a positive-definite kernel function that measures the pairwise similarity between two distinct inputs $\mathbf{x}$ and $\mathbf{x}'$. With $^\top$ denoting transposition, this means that the random vector $\mathbf{f}_t := [f(\mathbf{x}_1) \ldots f(\mathbf{x}_t)]^\top$ consisting of the function evaluations at instances $\mathbf{X}_t := [\mathbf{x}_1 \ldots \mathbf{x}_t]^\top$ is Gaussian distributed as $p(\mathbf{f}_t|\mathbf{X}_t) = \mathcal{N}(\mathbf{f}_t; \mathbf{0}_t, \mathbf{K}_t)(\forall t)$, where $\mathbf{K}_t$ denotes the $t \times t$ covariance matrix whose (i, j) entry is $[\mathbf{K}_t]_{i,j} = \text{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) := \kappa(\mathbf{x}_i, \mathbf{x}_j)$ [33].

With $\mathbf{y}_t := [y_1 \cdots y_t]^\top$ denoting the (possibly noisy) output data, it can be shown that in the regression task where the per-datum likelihood can be expressed as $p(y_\tau|f(\mathbf{x}_\tau)) = \mathcal{N}(y_\tau; f(\mathbf{x}_t), \sigma_n^2)$, the posterior pdf of $f(\mathbf{x})$ is [33]

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

where

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

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

and $\mathbf{k}_t(\mathbf{x}) := [\kappa(\mathbf{x}_1, \mathbf{x}), \ldots, \kappa(\mathbf{x}_t, \mathbf{x})]^\top$.

Note that the mean in (3a) is a point prediction of $f(\mathbf{x})$, while the variance in (3b) quantifies the associated uncertainty. In the AL context, this uncertainty is used by the acquisition function that selects the next query instance as

$$\mathbf{x}_{t+1} = \arg\max_{\mathbf{x} \in \mathcal{U}_t} \ \sigma_t^2(\mathbf{x}) \ . \tag{4}$$

It is worth mentioning that for a Gaussian pdf, (4) is tantamount to maximizing the entropy [25].

The posterior mean and variance in (3) rely on all t instances in $\mathbf{X}_t$ to form $\mathbf{K}_t$, and the associated complexity for its inversion is thus $\mathcal{O}(t^3)$. Although this complexity can be affordable in AL where t is small, it can be further reduced. In addition, $\mu_t(\mathbf{x})$ and $\sigma_t^2(\mathbf{x})$ require direct access to $\{\mathbf{x}_\tau\}_{\tau=1}^t$, which may be discouraged due to privacy concerns as in e.g medical records and financial statements. Further, GP-based AL relies on a preselected kernel function, which may exhibit limited expressiveness. These limitations can be ameliorated through our novel ensemble approach that leverages also random spectral features, as delineated next.

4. Ensemble GPs for AL

The chosen function model affects critically the performance of AL approaches. Unlike most existing works that rely on a single GP with a *preselected* kernel, we advocate an ensemble (E) of M GPs to enhance expressiveness. Each GP has a distinct kernel function selected from a given dictionary $\mathcal{K} := \{\kappa^1, \ldots, \kappa^M\}$, that is formed using kernels of different types and with different hyperparameters. Specifically, each GP $m \in \mathcal{M} := \{1, \ldots, M\}$ places a unique prior on f as $f|m \sim \mathcal{GP}(0, \kappa^m(\mathbf{x}, \mathbf{x}'))$. The EGP prior of $f(\mathbf{x})$ is then a weighted ensemble of the individual GP priors as

$$f(\mathbf{x}) \sim \sum_{m=1}^M w_0^m \mathcal{GP}(0, \kappa^m(\mathbf{x}, \mathbf{x}')), \quad \sum_{m=1}^M w_0^m = 1 \tag{5}$$

where $w_0^m := \Pr(i = m)$ is the prior probability that measures the contribution of GP model m. With labeled data collected on-the-fly, the sum-product rule allows one to express the EGP-based function posterior pdf as

$$p(f(\mathbf{x})|\mathcal{L}_t) = \sum_{m=1}^M \Pr(i=m|\mathcal{L}_t) p(f(\mathbf{x})|i=m, \mathcal{L}_t) \tag{6}$$

which is a mixture of posterior GPs with weights $w_t^m := \Pr(i = m|\mathcal{L}_t)$ that signify the significance of the GP experts. These weights thus enable online model adaptation.

To efficiently update this EGP function model across t, we will leverage a parametric function approximant, formed by the so-termed random features (RFFs), as outlined next.

A. RFF-based approximation per GP

The RFF approximation will be applied to each GP with kernel κ^m in EGP (5). Here, we drop the superscript m in κ^m for notational brevity. Consider a standardized ans shift-invariant kernel $\bar{\kappa}(\mathbf{x}, \mathbf{x}') = \bar{\kappa}(\mathbf{x} - \mathbf{x}')$ satisfying $\kappa = \sigma_\theta^2\bar{\kappa}$, which can be expressed as the inverse Fourier transform of a spectral density $\pi_{\bar{\kappa}}(\boldsymbol{\zeta})$ as [37]

$$\bar{\kappa}(\mathbf{x} - \mathbf{x}') = \int \pi_{\bar{\kappa}}(\boldsymbol{\zeta}) e^{j\boldsymbol{\zeta}^\top(\mathbf{x}-\mathbf{x}')} d\boldsymbol{\zeta} = \mathbb{E}_{\pi_{\bar{\kappa}}}\left[e^{j\boldsymbol{\zeta}^\top(\mathbf{x}-\mathbf{x}')}\right]$$

where $\int \pi_{\bar{\kappa}}(\boldsymbol{\zeta}) d\boldsymbol{\zeta} = 1$, allowing $\pi_{\bar{\kappa}}$ to be deemed as a pdf. Since $\bar{\kappa}$ is real, the last expectation is equal to $\mathbb{E}_{\pi_{\bar{\kappa}}}\left[\cos(\boldsymbol{\zeta}^\top(\mathbf{x} - \mathbf{x}'))\right]$; and after drawing a sufficient number D of independent and identically distributed (i.i.d.) samples $\{\boldsymbol{\zeta}_j\}_{j=1}^D$ from $\pi_{\bar{\kappa}}(\boldsymbol{\zeta})$, kernel $\bar{\kappa}$ is approximated by[3]

$$\breve{\bar{\kappa}}(\mathbf{x}, \mathbf{x}') := \frac{1}{D}\sum_{j=1}^D \cos\left(\boldsymbol{\zeta}_j^\top(\mathbf{x} - \mathbf{x}')\right) \ . \tag{7}$$

Defining the $2D \times 1$ RFF vector as [21]

$$\boldsymbol{\phi}_{\boldsymbol{\zeta}}(\mathbf{x}) := \frac{1}{\sqrt{D}}\left[\sin(\boldsymbol{\zeta}_1^\top\mathbf{x}), \cos(\boldsymbol{\zeta}_1^\top\mathbf{x}), \ldots, \sin(\boldsymbol{\zeta}_D^\top\mathbf{x}), \cos(\boldsymbol{\zeta}_D^\top\mathbf{x})\right]^\top \tag{8}$$

[3]Here we don't have a fully Bayesian treatment of the frequencies $\{\boldsymbol{\zeta}_i\}$, which would otherwise raise the issue of scalability. The notion of Bayesian active learning refers to modeling the unknown function via a Bayesian statistical model.

the sample average (7) can be re-written as $\check{\kappa}(\mathbf{x},\mathbf{x}') = \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{\top}(\mathbf{x})\boldsymbol{\phi}_{\boldsymbol{\zeta}}(\mathbf{x}')$. This allows the *parametric linear* function

$$\check{f}(\mathbf{x}) = \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{\top}(\mathbf{x})\boldsymbol{\theta}, \quad \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}_{2D}, \sigma_{\theta}^2 \mathbf{I}_{2D}) \tag{9}$$

to have an approximate GP prior. To interpret this point, we can see that the joint prior pdf for any number t of function values $\check{\mathbf{f}}_t := [\check{f}(\mathbf{x}_1), \ldots, \check{f}(\mathbf{x}_t)]$ is given by the $p(\check{\mathbf{f}}_t|\mathbf{X}_t) = \mathcal{N}(\check{\mathbf{f}}_t; \mathbf{0}_t, \tilde{\mathbf{K}}_t)$, where $\tilde{\mathbf{K}}_t = \sigma_{\theta}^2 \boldsymbol{\Phi}_t \boldsymbol{\Phi}_t^{\top}$ (with $\boldsymbol{\Phi}_t := [\boldsymbol{\phi}_{\boldsymbol{\zeta}}(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_{\boldsymbol{\zeta}}(\mathbf{x}_t)]^{\top}$) is a low rank ($2D$) approximant of the original kernel matrix $\mathbf{K}_t$. Such an RFF-based parametric function readily yields an efficient model update by propagating the posterior pdf $p(\boldsymbol{\theta}|\mathbf{y}_t; \mathbf{X}_t) = \mathcal{N}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_t, \boldsymbol{\Sigma}_t)$ per slot t in a recursive Bayes fashion. It is also worth stressing that the model learning step does not require direct access to $\mathbf{x}_t$, but relies only on the RFF vector $\boldsymbol{\phi}_{\boldsymbol{\zeta}}(\mathbf{x}_t)$, which can be viewed as an encrypted version of $\mathbf{x}_t$. This may be appealing if $\mathbf{x}_t$, which may e.g., comprise private medical data, cannot be revealed during model learning.

Having outlined the RFF-based approximation per GP, we can proceed with updating our RFF-based EGP function model, as labeled instances become available incrementally.

B. EGP parametric model updates

When kernels in the dictionary are shift-invariant, the RFF vector $\boldsymbol{\phi}_{\boldsymbol{\zeta}}^m(\mathbf{x})$ per GP m can be formed via (8) by first drawing i.i.d. random vectors $\{\boldsymbol{\zeta}_j^m\}_{j=1}^D$ from $\pi_{\bar{\kappa}}^m(\boldsymbol{\zeta})$, which is the spectral density of the standardized kernel $\bar{\kappa}^m$. Let $\sigma_{\theta^m}^2$ be the kernel magnitude, so that $\kappa^m = \sigma_{\theta^m}^2 \bar{\kappa}^m$. The generative model for the sought function and the noisy output y per GP m are characterized by the $2D \times 1$ vector $\boldsymbol{\theta}^m$ as

$$\begin{aligned} p(\boldsymbol{\theta}^m) &= \mathcal{N}(\boldsymbol{\theta}^m; \mathbf{0}_{2D}, \sigma_{\theta^m}^2 \mathbf{I}_{2D}) \\ p(f(\mathbf{x}_\tau)|i=m, \boldsymbol{\theta}^m) &= \delta(f(\mathbf{x}_\tau) - \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{x}_\tau)\boldsymbol{\theta}^m) \\ p(y_\tau|\boldsymbol{\theta}^m, \mathbf{x}_\tau) &= \mathcal{N}(y_\tau; \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{x}_\tau)\boldsymbol{\theta}^m, \sigma_n^2) \,. \end{aligned} \tag{10}$$

This parametric form allows one to capture the function posterior pdf per GP m via $p(\boldsymbol{\theta}^m|\mathcal{L}_t) = \mathcal{N}(\boldsymbol{\theta}^m; \hat{\boldsymbol{\theta}}_t^m, \boldsymbol{\Sigma}_t^m)$, which together with the weight w_t^m, approximates the EGP function posterior (6) via

$$p(\check{f}(\mathbf{x})|\mathcal{L}_t) = \sum_{m=1}^{M} w_t^m \mathcal{N}(\check{f}(\mathbf{x}); \check{\mu}_t^m(\mathbf{x}), (\check{\sigma}_t^m(\mathbf{x}))^2) \tag{11}$$

with

$$\check{\mu}_t^m(\mathbf{x}) = \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{x})\hat{\boldsymbol{\theta}}_t^m \tag{12a}$$

$$(\check{\sigma}_t^m(\mathbf{x}))^2 = \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{x})\boldsymbol{\Sigma}_t^m \boldsymbol{\phi}_{\boldsymbol{\zeta}}^m(\mathbf{x}) \,. \tag{12b}$$

Next, we will see how RFF-based EGP propagates the function model by updating across t the parameter set

$$\boldsymbol{\Xi}_t := \{w_t^m, \hat{\boldsymbol{\theta}}_t^m, \boldsymbol{\Sigma}_t^m, m \in \mathcal{M}\} \,. \tag{13}$$

Upon acquiring the newly labeled pair $\{\mathbf{x}_{t+1}, y_{t+1}\}$, the updated weight $w_{t+1}^m := \Pr(i=m|\mathcal{L}_{t+1})$ can be obtained per GP m via Bayes' rule as

$$w_{t+1}^m = \frac{\Pr(i=m|\mathcal{L}_t) p(y_{t+1}|\mathbf{x}_{t+1}, i=m, \mathcal{L}_t)}{p(y_{t+1}|\mathbf{x}_{t+1}, \mathcal{L}_t)} .$$

Since the per-model predictive likelihood is given by

$$\begin{aligned} p(y_{t+1}|i=m, \mathcal{L}_t, \mathbf{x}_{t+1}) &= \int p(y_{t+1}|\boldsymbol{\theta}^m, \mathbf{x}_{t+1}) p(\boldsymbol{\theta}^m|\mathcal{L}_t) d\boldsymbol{\theta}^m \\ &= \mathcal{N}(y_{t+1}; \hat{y}_{t+1|t}^m, (\sigma_{t+1|t}^m)^2) \end{aligned}$$

with

$$\begin{aligned} \hat{y}_{t+1|t}^m &= \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{x}_{t+1})\hat{\boldsymbol{\theta}}_t^m \\ (\sigma_{t+1|t}^m)^2 &= \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{x}_{t+1})\boldsymbol{\Sigma}_t^m \boldsymbol{\phi}_{\boldsymbol{\zeta}}^m(\mathbf{x}_{t+1}) + \sigma_n^2 \end{aligned}$$

the updated weight can thus be expressed as

$$w_{t+1}^m = \frac{w_t^m \mathcal{N}\left(y_{t+1}; \hat{y}_{t+1|t}^m, (\sigma_{t+1|t}^m)^2\right)}{\sum_{m'=1}^{M} w_t^{m'} \mathcal{N}\left(y_{t+1}; \hat{y}_{t+1|t}^{m'}, (\sigma_{t+1|t}^{m'})^2\right)} \,. \tag{15}$$

Bayes' rule further allows updating the posterior of $\boldsymbol{\theta}^m$ as

$$\begin{aligned} p(\boldsymbol{\theta}^m|\mathcal{L}_{t+1}) &= \frac{p(\boldsymbol{\theta}^m|\mathcal{L}_t) p(y_{t+1}|\boldsymbol{\theta}^m, \mathbf{x}_{t+1})}{p(y_{t+1}|\mathbf{x}_{t+1}, i=m, \mathcal{L}_t)} \\ &= \mathcal{N}(\boldsymbol{\theta}^m; \hat{\boldsymbol{\theta}}_{t+1}^m, \boldsymbol{\Sigma}_{t+1}^m) \end{aligned} \tag{16}$$

where the mean $\hat{\boldsymbol{\theta}}_{t+1}^m$ and covariance matrix $\boldsymbol{\Sigma}_{t+1}^m$ are

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{t+1}^m &= \hat{\boldsymbol{\theta}}_t^m + (\sigma_{t+1|t}^m)^{-2} \boldsymbol{\Sigma}_t^m \boldsymbol{\phi}_{\boldsymbol{\zeta}}^m(\mathbf{x}_{t+1})(y_{t+1} - \hat{y}_{t+1|t}^m) \\ \boldsymbol{\Sigma}_{t+1}^m &= \boldsymbol{\Sigma}_t^m - (\sigma_{t+1|t}^m)^{-2} \boldsymbol{\Sigma}_t^m \boldsymbol{\phi}_{\boldsymbol{\zeta}}^m(\mathbf{x}_{t+1}) \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{x}_{t+1}) \boldsymbol{\Sigma}_t^m . \end{aligned}$$

Remark. While most of kernel functions in GPs induce stationary functions, there are nonstatinary kernel functions, that could be accommodated by our ensemble GP framework by using a generalized random feature approximation; see [34]. Further, the ensembling rule (c.f. (15)) that adaptively weights the kernels also readily accommodates nonstationary functions.

C. Acquisition rules for EGP-based AL

Using the EGP posterior in (11), we are ready to devise AFs that select the next query point based on different rules.

1) Weighted variance: Motivated by (4), the first AF relies on the uncertainty expressed by the variance. With GP expert m forming the function posterior with variance $(\check{\sigma}_t^m(\mathbf{x}))^2$, a weighted combination over all the M experts yields the AF

$$\alpha^{\mathrm{wVar}}(\mathbf{x}; \mathcal{L}_t) := \sum_{m=1}^{M} w_t^m (\check{\sigma}_t^m(\mathbf{x}))^2 \,. \tag{18}$$

2) Weighted entropy: Relying alternatively on the entropy as the uncertainty measure, one can take a weighted sum of the entropy values given by the M GP experts, yielding

$$\alpha^{\mathrm{wEnt}}(\mathbf{x}; \mathcal{L}_t) := \frac{1}{2} \sum_{m=1}^{M} w_t^m \ln(2\pi(\check{\sigma}_t^m(\mathbf{x}))^2) \,. \tag{19}$$

3) Query-by-Committee (QBC): Besides capturing uncertainty by variance or entropy, an alternative disagreement-based AF – QBC, has been reported for classification [40], and regression using neural networks [20]. With the M GP experts forming a committee, the novel EGP-based QBC rule

is (cf. (12a))

$$\alpha^{\mathrm{QBC}}(\mathbf{x};\mathcal{L}_t) := \sum_{m=1}^{M} w_t^m(\breve{\mu}_t^m(\mathbf{x}) - \bar{\mu}_t(\mathbf{x}))^2 \qquad (20)$$

where $\bar{\mu}_t(\mathbf{x})$ is the consensus of the committee given by

$$\bar{\mu}_t(\mathbf{x}) = \sum_{m=1}^{M} w_t^m \breve{\mu}_t^m(\mathbf{x}) \; . \qquad (21)$$

Unlike previous QBC approaches that have equal weights per committee member, the weights w_t^m in (20) and (21) are generally different across m.

4) Variance of GP mixtures: Rather than directly summing per-GP weighted variances in (18), one can alternatively obtain the variance based on the GP mixture of the function posterior (cf. (11)) as

$$\alpha^{\mathrm{GPM-Var}}(\mathbf{x};\mathcal{L}_t) := \sum_{m=1}^{M} w_t^m((\breve{\sigma}_t^m(\mathbf{x}))^2 + (\breve{\mu}_t^m(\mathbf{x}) - \bar{\mu}_t(\mathbf{x}))^2 \qquad (22)$$

which, interestingly, is the sum of (18) and (20).

5) Entropy of GP mixtures: The last AF is given by the entropy of the GP mixture in (11), which unfortunately, has no analytic expression. Aiming at a tractable form, we will resort to its analytic lower bound [14], which is expressed as

$$\begin{aligned} &-\sum_{m=1}^{M} w_t^m \int \mathcal{N}(\breve{f}(\mathbf{x}); \breve{\mu}_t^m(\mathbf{x}), (\breve{\sigma}_t^m(\mathbf{x}))^2) \log p(\breve{f}(\mathbf{x})|\mathcal{L}_t) d\breve{f}(\mathbf{x}) \\ &\overset{(a)}{\geq} -\sum_{m=1}^{M} w_t^m \log\Big(\int \mathcal{N}(\breve{f}(\mathbf{x}); \breve{\mu}_t^m(\mathbf{x}), (\breve{\sigma}_t^m(\mathbf{x}))^2) \\ &\qquad \times p(\breve{f}(\mathbf{x})|\mathcal{L}_t) d\breve{f}(\mathbf{x}) \Big) \end{aligned}$$

where (a) holds due to Jensen's inequality. Upon obtaining the analytic expression for the term inside the logarithm, the last AF is then given by

$$\alpha^{\mathrm{GPM-Ent}}(\mathbf{x};\mathcal{L}_t) := -\sum_{m=1}^{M} w_t^m \log\left(\sum_{m'=1}^{M} w_t^{m'} z_t^{m,m'} \right) \qquad (23)$$

with $z^{m,m'}$ accounting for the interaction of any two GP models as

$$\begin{aligned} z_t^{m,m'} &:= \int \mathcal{N}(\breve{f}(\mathbf{x}); \breve{\mu}_t^m(\mathbf{x}), (\breve{\sigma}_t^m(\mathbf{x}))^2) \\ &\quad \times \mathcal{N}(\breve{f}(\mathbf{x}); \breve{\mu}_t^{m'}(\mathbf{x}), (\breve{\sigma}_t^{m'}(\mathbf{x}))^2) d\breve{f}(\mathbf{x}) \\ &= \mathcal{N}(\breve{\mu}_t^m(\mathbf{x}); \breve{\mu}_t^{m'}(\mathbf{x}), (\breve{\sigma}_t^m(\mathbf{x}))^2 + (\breve{\sigma}_t^{m'}(\mathbf{x}))^2) \; . \end{aligned}$$

Based on our novel EGP-based AFs, implementation of the proposed EGP-AL approach is summarized in Alg. 1. In the diagram of Fig. 1 the AL process of the proposed EGP-AL approach is illustrated. A discussion about the pros and cons of these AFs is deferred to Sec. 1 of the supplementary file.

5. Ensemble of EGP-based AFs

So far, we have introduced a novel EGP-based function model along with several choices for the AF. In the context of Bayesian optimization though, it is known that no single AF excels at all tasks [10]. Hence, combining candidate AFs can intuitively offer robustness and improved performance. To this end, we will rely on a validation set $\mathcal{V} := \{(\mathbf{x}_\tau^v, y_\tau^v)\}_{\tau=1}^V$ to evaluate the performance of different AFs. Similar to EGP, each of the K candidate AFs will come with a weight (probability) $\omega_t^k \in [0,1]$ to capture its contribution per slot t, such that $\sum_{k=1}^K \omega_t^k = 1$.

Algorithm 2 EGP-MultiAFs for AL.

Initialization: $\mathcal{L}_0$, $\mathcal{U}_0$, $\mathcal{V}$, $\mathcal{K}$, D, σ_θ^2;
$\boldsymbol{\omega}_0 = \frac{1}{K}[1,\ldots,1]^\top$;
for $t = 0, 1, \ldots, T$ **do**
  Obtain EGP $\boldsymbol{\Xi}_t$ based on $\mathcal{L}_t$ using (15)-(16);
  for $k = 1, \ldots, K$ **do**
    Obtain instance $\tilde{\mathbf{x}}_{t+1}^k \in \mathcal{U}_t$ by (24);
    Obtain pseudo-label $\tilde{y}_{t+1}^k$ using $\boldsymbol{\Xi}_t$ via (25);
    Using pseudo pair $\{\tilde{\mathbf{x}}_{t+1}^k, \tilde{y}_{t+1}^k\}$ obtain $\tilde{\boldsymbol{\Xi}}_{t+1}^k$;
    Obtain error $\epsilon_{t+1}^{v,k}$ on $\mathcal{V}$ via (27);
  end for
  Update per AF weight using (29);
  Obtain $\mathbf{x}_{t+1} \in \mathcal{U}_t$ by (30);
  Query the oracle to obtain true label y_{t+1};
  $\mathcal{L}_{t+1} = \mathcal{L}_t \cup (\mathbf{x}_{t+1}, y_{t+1})$;
  $\mathcal{U}_{t+1} = \mathcal{U}_t \setminus \{\mathbf{x}_{t+1}\}$;
end for
Output: $\mathcal{L}_T$, $\boldsymbol{\Xi}_T$

Upon identifying the RFF-based EGP set $\boldsymbol{\Xi}_t$ in (13) using the labeled set $\mathcal{L}_t$ at slot t, each AF k selects its query point $\tilde{\mathbf{x}}_{t+1}^k$ at slot $t+1$ by optimizing the associated criterion as

$$\tilde{\mathbf{x}}_{t+1}^k = \arg\max_{\mathbf{x}\in\mathcal{U}_t} \; \alpha^k(\mathbf{x};\mathcal{L}_t) \; . \qquad (24)$$

Upon obtaining $\tilde{\mathbf{x}}_{t+1}^k$, AF k constructs a *'pseudo label'* $\tilde{y}_{t+1}^k$ using the EGP parameters in $\boldsymbol{\Xi}_t$, as

$$\tilde{y}_{t+1}^k = \sum_{m=1}^{M} w_t^m \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\tilde{\mathbf{x}}_{t+1}^k)\hat{\boldsymbol{\theta}}_t^m \; . \qquad (25)$$

This pseudo pair $\{\tilde{\mathbf{x}}_{t+1}^k, \tilde{y}_{t+1}^k\}$ allows one to leverage (15)–(16) to find the updated EGP parameter vector as

$$\tilde{\boldsymbol{\Xi}}_{t+1}^k = \{\tilde{w}_{t+1}^{m,k}, \tilde{\boldsymbol{\theta}}_{t+1}^{m,k}, \tilde{\boldsymbol{\Sigma}}_{t+1}^{m,k}, m \in \mathcal{M}\} \qquad (26)$$

based on which the loss per AF can be evaluated.

To find this loss, AF k capitalizes on $\tilde{\boldsymbol{\Xi}}_{t+1}^k$ in order to obtain the prediction error at the validation set

$$\epsilon_{t+1}^{v,k} = V^{-1} \sum_{\tau=1}^{V} (y_\tau^v - \hat{y}_{\tau|t+1}^{v,k})^2 \qquad (27)$$

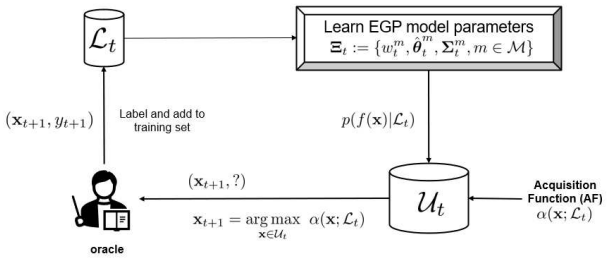

Fig. 1: Diagram of the AL process of the advocated EGP-based AL methods.

TABLE I: Additional experimental details

Dataset	$\mathcal{L}_0$ size	$\mathcal{V}$ size	$\mathcal{U}_0$ size	$\mathcal{T}$ size	η
Ackley-5D	10	50	500	100	1
Branin	10	50	500	100	100
Currin exponential	10	50	500	100	100
Gramacy	10	50	500	100	100
Higdon	10	50	500	100	100
Diabetes	15	55	261	111	100
Robot pushing 3D	20	50	500	200	20
Robot pushing 4D	20	50	500	200	20
California housing	50	70	1000	1032	0.05
DeepMIMO_2feat	100	50	1000	2000	1

TABLE II: Analytical expression of all synthetic functions

Function	Analytical expression
Ackley-5D	$-20e^{-0.2\sqrt{(x_1^2+x_2^2+x_3^2+x_4^2+x_5^2)/5}} - e^{(\cos(2\pi x_1)+\cos(2\pi x_2)+\cos(2\pi x_3)+\cos(2\pi x_4)+\cos(2\pi x_5))/5} + 20 + e^1$
Branin	$(x_2 - 5.1/(4\pi^2)x_1^2 + 5x_1/\pi - 6)^2 + 10(1 - 1/(8\pi))\cos(x_1) + 10$
Currin exponential	$(1 - e^{-1/(2x_2)})(2300x_1^3 + 1900x_1^2 + 2092x_1 + 60)/(100x_1^3 + 500x_1^2 + 4x_1 + 20)$
Gramacy	$\sin(10\pi x)/(2x) + (x-1)^4$
Higdon	$\sin(2\pi x/10) + 0.2\sin(2\pi x/2.5)$

where the predicted label per validation sample τ is

$$\hat{y}_{\tau|t+1}^{v,k} = \sum_{m=1}^{M} \tilde{w}_{t+1}^{m,k} \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{x}_\tau^v) \tilde{\boldsymbol{\theta}}_{t+1}^{m,k} \ . \tag{28}$$

Having available the prediction error over the validation set per AF k, the associated weight can then be updated as

$$\omega_{t+1}^k = \frac{\omega_t^k \exp(-\eta \epsilon_{t+1}^{v,k})}{\sum_{k'=1}^{K} \omega_t^{k'} \exp(-\eta \epsilon_{t+1}^{v,k'})} \tag{29}$$

where η denotes the learning rate. Here, the weight update rule is similar to that in EGP (cf. (15)), and belongs to the exponential weight update in online learning with expert advice; see e.g., [5].

Given the updated weights, the next query point is identified by maximizing the weighted ensemble of AFs as

$$\mathbf{x}_{t+1} = \arg\max_{\mathbf{x}\in\mathcal{U}_t} \sum_{k=1}^{K} \omega_{t+1}^k \alpha^k(\mathbf{x}; \mathcal{L}_t) \ . \tag{30}$$

Upon querying the oracle for the label y_{t+1} of instance $\mathbf{x}_{t+1}$, the labeled and unlabeled sets are updated, thus completing one iteration of the novel "EGP-MultiAFs" approach, that is implemented as listed in Alg. 2.

Computational complexity of EGP-MultiAFs. Per AL iteration, the computational complexity of EGP-MultiAF emanates from updating the EGP model and optimizing the AF. Leveraging the random feature (RF) approximation per GP, the former incurs complexity $\mathcal{O}((2D)^2 M)$, where M is the number of GPs in the EGP, and D is the number of spectral features in the RFF vector (cf. Eq.(8)). For the latter in the pool-based AL, the major computation originates from the steps in (24)-(27), and (30), which respectively, incur complexity $\mathcal{O}((2DM)^2|\mathcal{U}_t|)$, $\mathcal{O}(5(2D)M)$, $\mathcal{O}((2D)^2 M)$, $\mathcal{O}(2DM|\mathcal{V}|)$

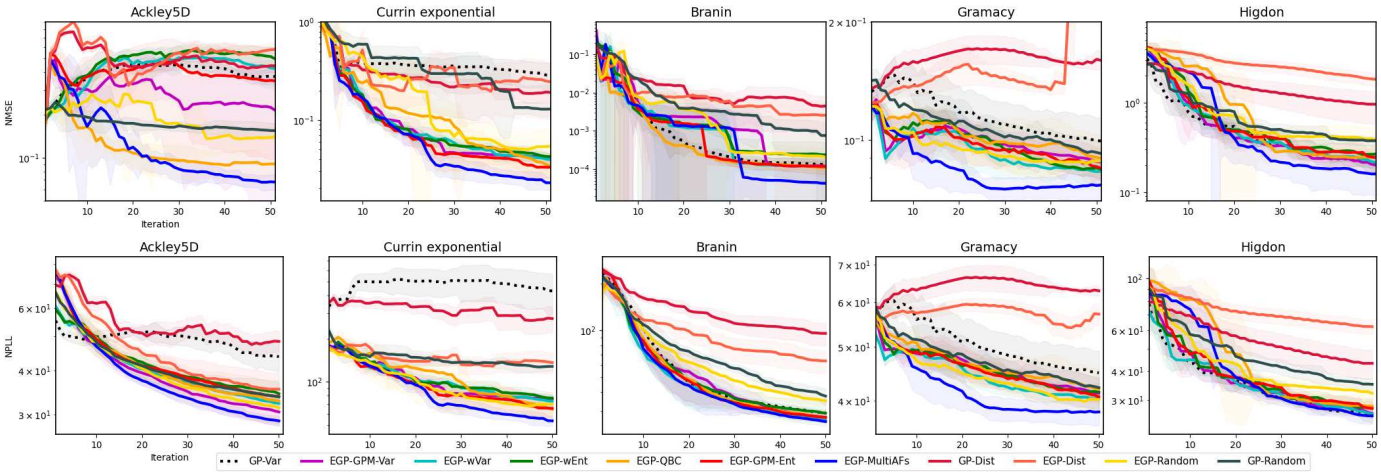

Fig. 2: NMSE and NPLL versus iterations for Ackley5D, Currin exponential, Branin, Gramacy and Higdon functions

and $\mathcal{O}((2DM)^2|\mathcal{U}_t|)$, where $|\mathcal{V}|$ and $|\mathcal{U}_t|$ are the cardinalities of the fixed validation set and the unlabeled set per-iteration. Complexity of AF optimization step in (24) and (30) is dominated by the EGP-GPM-Ent approach. Considering only the dominating factors, the overall complexity of EGP-MultiAFs is $\mathcal{O}((2DM)^2|\mathcal{U}_t|)$.

6. Numerical tests

In this section, the performance of the proposed EGP-based AL models will be compared against several benchmark synthetic functions, and it will be tested with real datasets ranging from biomedical to robotic based ones. Based on the novel EGP model, the innovative acquisition criteria to be tested are the ones described in (18) - (23) and (30), which henceforth will be abbreviated as "EGP-wVar," "EGP-wEnt," "EGP-QBC," "EGP-GPM-Var," "EGP-GPM-Ent," and "EGP-multiAFs," respectively. The competing baselines are (i) "GP-var" – a single GP model coupled with the maximum variance (entropy) AF in (4) that has been extensively used in AL; see e.g., [15], [31], [38], (ii) "GP-dist" – a single GP model together with the maximum distance-based AF as in [31], (iii) "EGP-dist" – the EGP model with the same AF, and (iv) "EGP-random" – the EGP model with random sampling. For all approaches, a few initially labeled data collected in $\mathcal{L}_0$ are utilized to obtain the kernel hyperparameters per GP expert by maximizing the marginal likelihood using the *sklearn* package. The RFF-based GP approximate models rely on $D = 50$ RFFs. For EGP-based approaches, the kernel dictionary $\mathcal{K}$ consists of radial basis functions (RBFs) with lengthscales $\{10^c\}_{c=-4}^{6}$. For the EGP-multiAFs approach, each $\alpha^k(\mathbf{x}; \mathcal{L}_t)$ in (30) is divided by its maximum value so that to range between 0 and 1.

The performance of the competing methods is evaluated on a held-out test set $\mathcal{T}^e := (\mathbf{x}^e_\tau, y^e_\tau)_{\tau=1}^{T^e}$ (superscript ^{e} stands for evaluation) using two metrics. The first performance metric is the normalized mean-square error (NMSE) that for iteration t, is given by

$$\text{NMSE}_t := \frac{1}{T^e} \sum_{\tau=1}^{T^e} (\hat{y}^e_{\tau|t} - y^e_\tau)^2 / \sigma_y^2$$

where $\hat{y}^e_{\tau|t}$ denotes the point prediction of test instance τ, and $\sigma_y^2 := \mathbb{E}\|\mathbf{y}^e_{T^e} - \mathbb{E}\{\mathbf{y}^e_{T^e}\}\|^2$, where $\mathbf{y}^e_{T^e} := [y^e_1 \dots y^e_{T^e}]^\top$. A second metric used to assess the associated uncertainty is the negative predictive log-likelihood (NPLL)

$$\text{NPLL}_t := -\log\, p(\mathbf{y}^e_{T^e} | \mathcal{L}_t, \mathbf{X}_{T^e})$$

where the matrix $\mathbf{X}_{T^e} := [\mathbf{x}^e_1 \dots \mathbf{x}^e_{T^e}]^\top$ collects the feature vectors of all T^e test instances. All methods are tested over 10 realizations, and their average performance is reported along with the corresponding standard deviation. More details about the experimental set up can be found in Table 1.

A. Synthetic functions

The tests here are run for known synthetic functions, including Ackley5D, Currin exponential, Branin, Gramacy and Higdon; see Table 2 for their analytic expression. Fig. 2 demonstrates that all EGP-based approaches with a single AF outperform the single GP-based baselines in the Currin exponential and Gramacy functions, in terms of NMSE and most have superior performance in the remaining three datasets. Further, all single AF EGP-based approaches achieve lower NMSE than the EGP-Dist baseline in all synthetic datasets and most of them outperform the EGP-Random baseline in four out of five datasets. In addition, all EGP-based methods enjoy the lowest NPLL in four out of five datasets compared to the single GP-based AL approaches, which corroborates the merits of having an ensemble of GPs and using them in the corresponding acquisition criteria. Further exploiting an ensemble of AFs in the adaptive EGP-multiAFs approach, significantly improves the prediction performance, and also effectively quantifies the prediction uncertainty, thus rendering it the best performing approach over all datasets.

B. Real datasets

All approaches here are tested on **California housing** [29] and **Diabetes** [8] real datasets. The latter deals with real medical data that are well motivated for AL because of the scarcity of labeled instances emanating from medical confidentiality. The description of the datasets is given below.

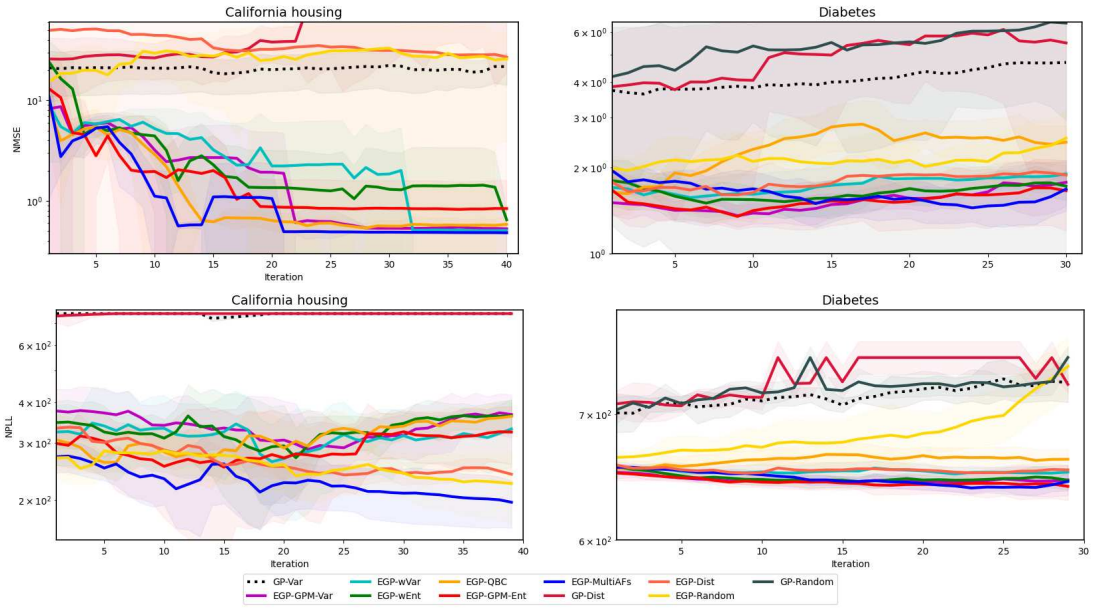

Fig. 3: NMSE (top) and NPLL (bottom) versus iterations for California housing and Diabetes datasets.

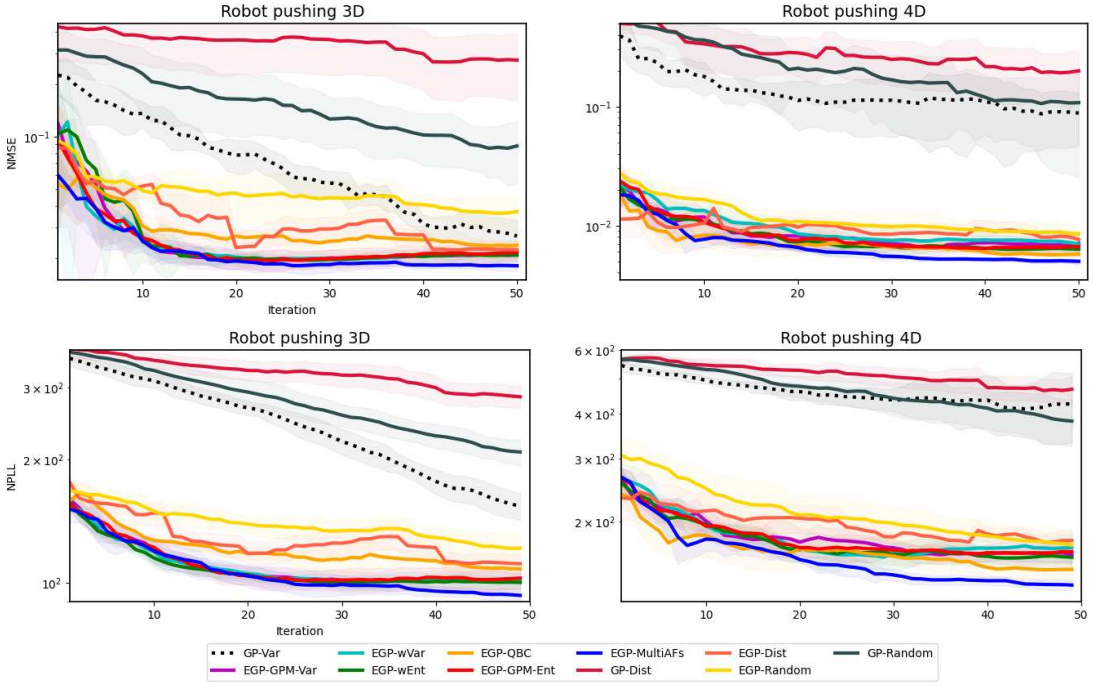

Fig. 4: NMSE (top) and NPLL (bottom) versus iterations for Robot Pushing 3D and 4D tasks.

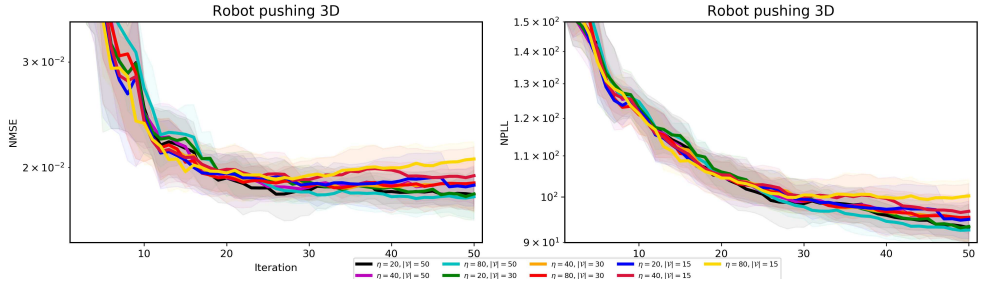

Fig. 5: Sensitivity analysis for $\eta, |\mathcal{V}|$ of the EGP-MultiAFs approach in the Robot Pushing 3D dataset.

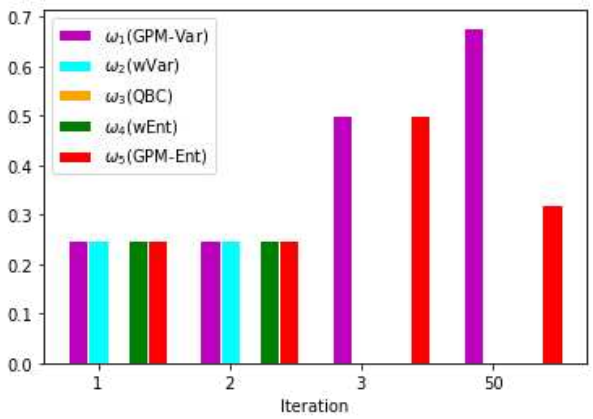

Fig. 6: AF weights of the EGP-MultiAFs approach in the Branin dataset.

California housing dataset. This dataset considers 8 features of districts in California, including not only demographic and location data, but also more general features such as average number of rooms and bedrooms per household, based on which a regression task is formed where the target variable is the median house price in these districts.

Diabetes dataset. This dataset considers 10 characteristics of diabetes patients, including age, sex, body mass index, average blood pressure, and six blood-related measurements. The target variable measures the disease progression in a single year.

It is evident in Fig. 3 that all EGP-based approaches markedly outperform the GP-Var, GP-Dist and GP-Random baselines in terms of NMSE and NPLL in the California housing and Diabetes datasets, showcasing the significance of adopting the EGP model to estimate the learning function, along with the corresponding AFs. In addition, all advocated EGP-based approaches outperform the EGP-Dist and EGP-Random baselines in terms of NMSE in the California housing and Diabetes datasets (except the EGP-QBC in the latter one), with the EGP-MultiAFs consistently being the best-performing method. It is worth mentioning that although most of the proposed EGP-based approaches are comparable in terms of prediction error in the California housing dataset, EGP-multiAFs outperforms all other methods in terms of NPLL. This illustrates the significance of properly adjusting AF weights in an online adaptive fashion.

C. Robotic tasks

The next experiments focus on a practical robotic task, where a robot pushes an object to a specific location [46]. Specifically, given as input the robot location (r_τ^x, r_τ^y) and pushing duration t_τ^p at slot τ, the object ends up in a location $\mathbf{o}_\tau := (o_\tau^x, o_\tau^y)$. We form a regression task where the goal per slot τ is to map the 3×1 feature vector $\mathbf{x}_\tau := [r_\tau^x, r_\tau^y, t_\tau^p]^\top$ to the target variable $y_\tau := ||\mathbf{o}_\tau - \mathbf{d}||_2$, with $\mathbf{d} := [d^x, d^y]$ denoting a pre-defined position vector, yielding the **Robot pushing 3D** dataset. This is of practical interest in various robotic problems such as obstacle avoidance, where y_τ is desired to be greater than a pre-defined threshold y^{th}. Augmenting the feature vector $\mathbf{x}_\tau$ with an additional pushing angle r_τ^θ, yields the **Robot pushing 4D** dataset.

Fig. 4 depicts the NMSE and NPLL at each iteration of all competing AL approaches for the Robot pushing 3D and 4D tasks respectively. It is evident that all EGP-based approaches enjoy lower NMSE and NPLL compared to the single GP based AL counterparts and the EGP-Dist, EGP-random baselines in both datasets, with the EGP-MultiAFs consistently being the best-performing one. This implies that in these practical robotic tasks, the function expressiveness offered by the advocated EGP model and the ensuing innovative acquisition criteria considerably improve the prediction performance providing also quantifiable prediction uncertainty.

D. Wireless communication tasks

The last experiments emphasize on a practical signal processing setting where given a small number of 5G signal measurements in different locations, the goal is to estimate 5G signal values at unmeasured locations. Specifically, the input feature vector $\mathbf{x}_\tau$ at location τ comprises the longitude and latitude of the location and the target variable y_τ to be estimated is the filtered beam reference signal received power [dBm] at this location. Details about the DeepMIMO dataset that was considered in our experimental setting can be found at [48]. Fig. 7 illustrates the NMSE and NPLL performance of all methods at each iteration of the AL process. It can be clearly seen that all advocated EGP-based AL approaches enjoy lower NMSE and (all except one) lower NPLL compared to the single GP-based counterparts, with the 'EGP-MultiAFs' approach being the best-performing one in this task too.

E. Additional experimental results

Additional ablation studies are presented here to further demonstrate the performance of the proposed EGP-MultiAFs approach.

Sensitivity analysis. In this ablation study, the aim is to gauge how sensitive the performance is to the size of the validation set $\mathcal{V}$ and the acquisition step size η. The NMSE and NPLL performance of the advocated EGP-MultiAFs approach is assessed on the Robot pushing 3D for different values of $|\mathcal{V}|$ and η. It is evident in Fig. 5 that when $|\mathcal{V}|$ is too small, the performance of EGP-MultiAFs is worse compared to that of a sufficiently larger validation set, which is as expected. The choice of η is also critical since it can lead to very good performance without the need for the validation set size to be the largest possible; see e.g the Robot pushing 3D dataset, where the performance of EGP-MultiAFs with $\eta = 20, |\mathcal{V}| = 30$ is comparable with that of $\eta = 80, |\mathcal{V}| = 50$ in terms of both NMSE and NPLL, as depicted in Fig. 5.

EGP-MultiAFs acquisition weights. In this ablation study, the goal is to demonstrate the role of the acquisition weights $\{\omega_k\}_{k=1}^K$ of the EGP-MultiAFs approach. Specifically, the acquisition weights of a single run are plotted as a function of the AL iteration index on the Branin dataset in Fig. 6, where it is evident that the weights of the GPM-Var and GPM-Ent AFs get larger values as more data are actively collected, which is intuitive since these AFs eventually enjoy the lowest

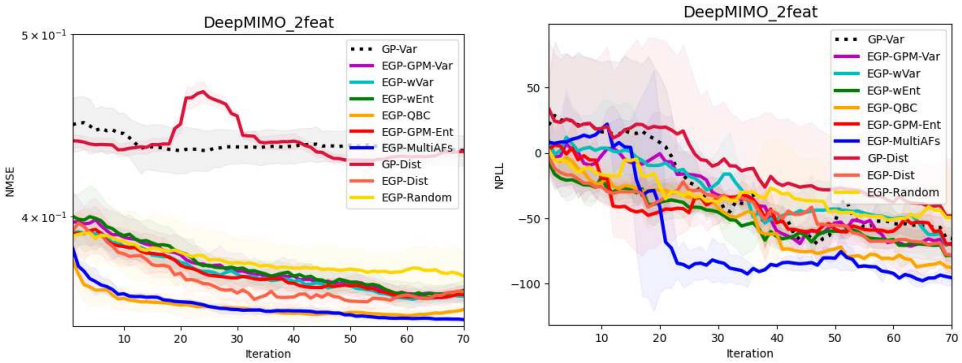

Fig. 7: NMSE (left) and NPLL (right) versus iterations for DeepMIMO_2feat dataset

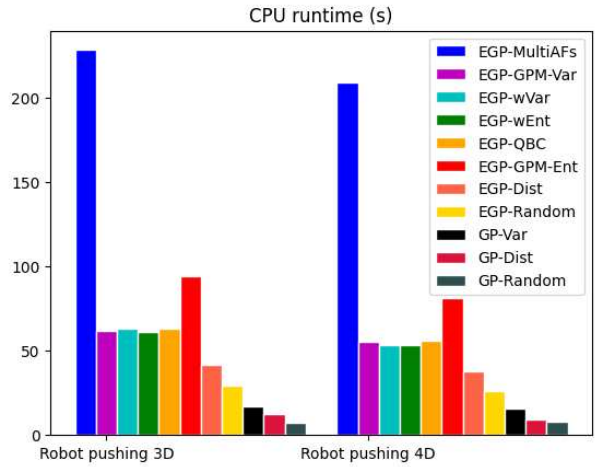

Fig. 8: CPU running time of all approaches in the robot pushing 3D and robot pushing 4D tasks.

NMSE compared to the other acquisition criteria. Therefore, the corresponding acquisition weights are properly adjusted as new data arrive on-the-fly.

CPU runtime. In this study, the runtime of all competing AL methods is assessed. For demonstration purposes, Fig. 8 illustrates the running time of all approaches in the robot pushing 3D and robot pushing 4D tasks. As expected, the EGP-based methods with a single pre-selected acquisition criterion require a small amount of extra runtime compared to the single GP counterparts since at each iteration the parameters of M GP models in the ensemble are updated. In addition, the EGP-MultiAFs approach requires the largest CPU runtime since an extra step is needed to adaptively learn the proper AF as new data are processed in an online fashion. Nonetheless, with the cost of some extra runtime, the advocated EGP-based AL approaches have superior performance over competing alternatives in most datasets, with the EGP-MultiAFs consistently being the best-performing one in terms of both NMSE and NPLL. It is also worth noticing that the reported runtime does not take into account the runtime needed to obtain a label.

Role of the parameters M, D, σ_n^2. In this ablation study the aim is to assess the role of the number of models M, the number of features D and the noise variance σ_n^2. Fig. 9 depicts the performance of the 'EGP-MultiAFs', 'EGP-QBC' and 'EGP-wEnt' approaches on the robot pushing 4D task. It can be clearly seen that when the number M of GP models in the ensemble is small, the prediction performance deteriorates. When the number of spectral features D is not sufficiently large, then the RFF approximation leads to larger prediction error. As expected, when the noise variance increases, the prediction error in all approaches also increases.

7. Discussion

Building on our novel EGP model, we have put forth five acquisition functions (AFs), that can be categorized into disagreement- and uncertainty-based ones. The former category derives from the so-termed "Query-by-Committee" (QBC) criterion in Sec.4.3.3, where each GP expert is a committee member, and the instance to be queried is the one that the committee members disagree the most. Albeit effective in several test cases (see e.g Ackley5D and Robot pushing 4D datasets in Fig. 1), this criterion does not account for the quantifiable uncertainty offered by the predictive variance of each GP expert, which can be of utmost importance for guiding the AL process in many cases; see, e.g., the Diabetes dataset in Fig. 2.

This uncertainty can be measured either directly by the variance or by the entropy. The "weighted variance" acquisition criterion in Sec.4.3.1 is given by a weighted combination of the predictive variances of all GP experts in the EGP model, which is intuitive because the variance of GP experts with larger weights should also weigh more in the acquisition step of the AL procedure. Although intuitive and simple, this approach considers only the predictive variance of the GP experts and does not account for the posterior mean of GP experts or any other interaction between the experts that may improve the prediction performance; see e.g the Ackley 5D dataset in Fig. 1. Combining the merits of the aforementioned approaches, we advocate the "variance of GP mixtures" criterion, which is the variance of the GP mixture in the EGP model given by the sum of the QBC and "weighted variance" criteria. The combination of these criteria in the "variance of GP mixtures", can significantly improve the prediction performance as corroborated in the Currin exponential, Diabetes, and Robot Pushing 3D datasets in Figs. 1-3.

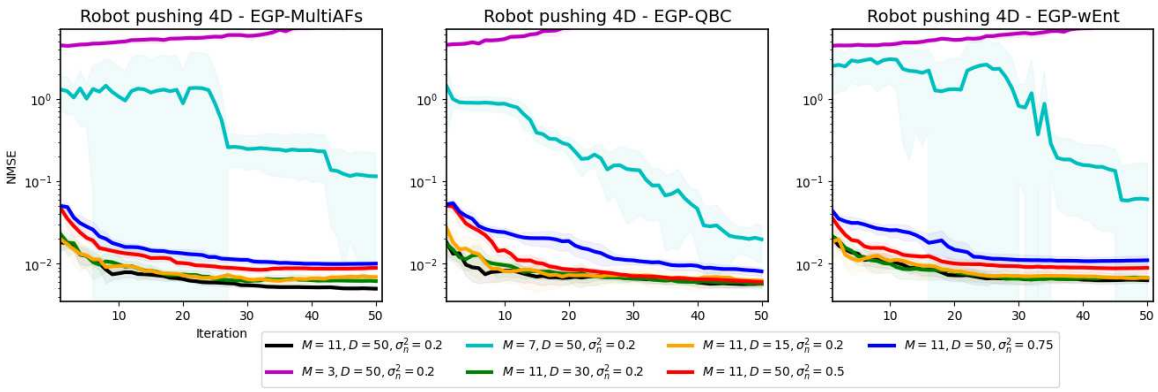

Fig. 9: NMSE versus iteration of 'EGP-MultiAFs', 'EGP-QBC' and 'EGP-wEnt' methods for different values of M, D, σ_n^2 on the robot pushing 4D task.

Relying alternatively on entropy as the uncertainty measure, we advocate the "weighted entropy" criterion in Sec.4.3.2, which is a weighted sum of the entropy values of the Gaussian predictive pdf of all GP experts in the EGP model. Although the maximum entropy criterion coincides with the maximum variance in the single GP case, this no longer holds for the EGP model. Adopting the 'weighted entropy' as an alternative uncertainty-based criterion to the 'weighted variance' one, can prove to be useful in several cases such as the Diabetes and Robot pushing 4D datasets in Figs. 2-3. Further allowing for interactions among individual GP models, one can develop an entropy measure based on the GP mixture pdf. Although this cannot be expressed in closed form, maximizing its analytic lower bound is tractable and yields the "Entropy of GP mixtures" criterion in Sec.4.3.5. Empirically, it is shown in Figs. 1-3 that neither the entropy-based nor the variance-based uncertainty criteria are always the best performing across all datasets, which is expected and well motivates the novel EGP-MultiAFs approach.

8. Conclusions and Future Directions

This work advocated a weighted ensemble of GPs as the statistical model in AL. By adapting the weights of individual GPs, the EGP model selects the appropriate kernel on-the-fly as new labeled data are included incrementally. Building on the novel EGP model, several AFs have been devised based on different criteria. Combining the candidate EGP-based AFs with weights being adjusted in an adaptive manner, further robustifies the AL performance. Tests on synthetic functions and real datasets showcase the merits of weighted ensembles of GPs and AFs in AL. Our future work includes development of EGP-based AFs for the classification task and theoretical analyses of the resultant approaches.

References

[1] Dimitris Berberidis and Georgios B Giannakis. Data-adaptive active sampling for efficient graph-cognizant classification. *IEEE Trans. Sig. Process.*, 66(19):5167–5179, 2018.

[2] Thang D Bui, Cuong Nguyen, and Richard E Turner. Streaming sparse gaussian process approximations. *Proc. Adv. Neural Inf. Process. Syst.*, 30, 2017.

[3] Robert Burbidge, Jem J Rowland, and Ross D King. Active learning for regression based on query by committee. In *Proc. Intl. Conf. Intel Data Engineering and Autom. Learn.*, pages 209–218. Springer, 2007.

[4] Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing expected model change for active learning in regression. In *Proc. Intl. Conf. Data Mining*, pages 51–60. IEEE, 2013.

[5] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[6] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[7] David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. Structure discovery in nonparametric regression through compositional kernel search. *Proc. Intl. Conf. Mach. Learn.*, pages 1166–1174, 2013.

[8] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. *Proc. Intl. Conf. Mach. Learn.*, pages 1183–1192, 2017.

[10] Matthew Hoffman, Eric Brochu, Nando de Freitas, et al. Portfolio allocation for Bayesian optimization. *Proc. Conf. Uncertainty in Artif. Intel.*, pages 327–336, 2011.

[11] Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Batch mode active learning and its application to medical image classification. In *Proc. Intl. Conf. Mach. Learn.*, pages 417–424, 2006.

[12] Wei-Ning Hsu and Hsuan-Tien Lin. Active learning by learning. In *Proc. AAAI Conf. Artif. Intel.*, 2015.

[13] Boshuang Huang, Sudeep Salgia, and Qing Zhao. Disagreement-based active learning in online settings. *IEEE Trans. Sig. Process.*, 70:1947–1958, 2022.

[14] Marco F Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D Hanebeck. On entropy approximation for Gaussian mixture random vectors. In *IEEE Intl. Conf. on Multisensor Fusion and Integ. for Intell. Syst.*, pages 181–188, 2008.

[15] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with Gaussian processes for object categorization. *Proc. Intl. Conf. Comp. Vision*, 2007.

[16] Hyunjik Kim and Yee Whye Teh. Scaling up the automatic statistician: Scalable structure discovery using Gaussian processes. *Proc. Intl. Conf. Artif. Intel. and Stats.*, pages 575–584, 2018.

[17] Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. *Proc. Adv. Neural Inf. Process. Syst.*, 30, 2017.

[18] Andreas Krause and Carlos Guestrin. Nonmyopic active learning of gaussian processes: an exploration-exploitation approach. In *Proc. Intl. Conf. Mach. Learn.*, page 449–456, 2007.

[19] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res.*, 9(2), 2008.

[20] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Proc. Adv. Neural Inf. Process. Syst.*, 7, 1994.

[21] Miguel Lázaro-Gredilla, Joaquin Quiñonero Candela, Carl Edward Rasmussen, and A. Figueiras-Vidal. Sparse spectrum Gaussian process regression. *J. Mach. Learn. Res.*, 11(Jun):1865–1881, 2010.

[22] John Lipor, Brandon P. Wong, Donald Scavia, Branko Kerkez, and Laura Balzano. Distance-penalized active learning using quantile search. *IEEE Trans. Sig. Process.*, 65(20):5453–5465, 2017.

[23] Ming Liu, Wray Buntine, and Gholamreza Haffari. Learning how to actively learn: A deep imitation learning approach. In *Proc. Annual Meet. Assoc. Comput. Linguistics*, pages 1874–1883, 2018.

[24] Qin Lu, Georgios Karanikolas, Yanning Shen, and Georgios B Giannakis. Ensemble Gaussian processes with spectral features for online interactive learning with scalability. *Proc. Intl. Conf. Artif. Intel. and Stats.*, pages 1910–1920, 2020.

[25] David JC MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[26] Wesley J Maddox, Samuel Stanton, and Andrew G Wilson. Conditioning sparse variational Gaussian processes for online decision-making. *Proc. Adv. Neural Inf. Process. Syst.*, 34, 2021.

[27] Gustavo Malkomes, Chip Schaff, and Roman Garnett. Bayesian optimization for automated model selection. *Proc. Adv. Neural Inf. Process. Syst.*, 2016.

[28] Duy Nguyen-Tuong, Jan Peters, and Matthias Seeger. Local gaussian process regression for real time online model learning. *Proc. Adv. Neural Inf. Process. Syst.*, 21, 2008.

[29] R Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics and Probability Letters*, 33(3):291–297, 1997.

[30] John Paisley, Xuejun Liao, and Lawrence Carin. Active learning and basis selection for kernel-based linear models: A bayesian perspective. *IEEE Trans. Sig. Process.*, 58(5):2686–2700, 2010.

[31] Edoardo Pasolli and Farid Melgani. Gaussian process regression within an active learning scheme. In *IEEE Intl. Geoscience and Remote Sensing Symp.*, pages 3574–3577, 2011.

[32] Konstantinos D Polyzos, Qin Lu, and Georgios B Giannakis. Active sampling over graphs for bayesian reconstruction with gaussian ensembles. In *Proc. Asilomar Conf. Sig., Syst., Comput.*, pages 58–64, 2022.

[33] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[34] Sami Remes, Markus Heinonen, and Samuel Kaski. Non-stationary spectral kernels. *Advances in neural information processing systems*, 30, 2017.

[35] Christoffer Riis, Francisco Antunes, Frederik Hüttel, Carlos Lima Azevedo, and Francisco Pereira. Bayesian active learning with fully bayesian gaussian processes. *Proc. Adv. Neural Inf. Process. Syst.*, 35:12141–12153, 2022.

[36] Christoffer Riis, Filipe Rodrigues, and Francisco Camara Pereira. Mixture of gaussian processes for bayesian active learning. *Authorea Preprints*, 2023.

[37] Walter Rudin. *Principles of Mathematical Analysis*, volume 3. McGraw-hill New York, 1964.

[38] Jens Schreiter, Duy Nguyen-Tuong, Mona Eberts, Bastian Bischoff, Heiner Markert, and Marc Toussaint. Safe exploration for active learning with Gaussian processes. In *European Conf. Mach. Learn. and Princ. and Practice of Knowledge Disc. in Databases*, 2015.

[39] Burr Settles. Active learning. *Synthesis Lectures on Artif. Intel. and Mach. Learn.*, 6(1):1–114, 2012.

[40] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proc. of the Workshop on Computational Learning Theory*, pages 287–294, 1992.

[41] Samuel Stanton, Wesley Maddox, Ian Delbridge, and Andrew Gordon Wilson. Kernel interpolation for scalable online gaussian processes. In *Proc. Intl. Conf. Artif. Intel. and Stats.*, pages 3133–3141. PMLR, 2021.

[42] Chuan Sun and Morteza Hashemi. Efficient user localization in wireless networks using active deep learning. In *Proc. Asilomar Conf. Sig., Syst., Comput.*, pages 1602–1606. IEEE, 2021.

[43] Annalisa T Taylor, Thomas A Berrueta, and Todd D Murphey. Active learning in robotics: A review of control principles. *Mechatronics*, 77:102576, 2021.

[44] Tong Teng, Jie Chen, Yehong Zhang, and Bryan Kian Hsiang Low. Scalable variational Bayesian kernel selection for sparse Gaussian process regression. *Proc. AAAI Conf. Artif. Intel.*, 34(04):5997–6004, 2020.

[45] Simon Tong. *Active Learning: Theory and Applications*. Stanford University, 2001.

[46] Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. *Proc. Intl. Conf. Mach. Learn.*, pages 3627–3635, 2017.

[47] Dongrui Wu, Chin-Teng Lin, and Jian Huang. Active learning for regression using greedy sampling. *Information Sciences*, 474:90–105, 2019.

[48] Wei Ye, Xinyue Hu, Tian Liu, Ruoyu Sun, Yanhua Li, and Zhi-Li Zhang. 5gnn: extrapolating 5g measurements through gnns. In *Proc. Intl. Workshop Graph Neural Netw.*, pages 36–41, 2022.

[49] Michael Minyi Zhang, Bianca Dumitrascu, Sinead A Williamson, and Barbara E Engelhardt. Sequential gaussian processes for online learning of nonstationary functions. *IEEE Trans. Sig. Process.*, 2023.

[50] Jing Zhao, Shiliang Sun, Huijuan Wang, and Zehui Cao. Promoting active learning with mixtures of Gaussian processes. *Knowledge-Based Systems*, 188:105044, 2020.

Konstantinos D. Polyzos obtained his Diploma (5-years degree) from the Department of Electrical Engineering and Computer Technology at the University of Patras, Greece in 2018. Currently, he is a PhD student at the Department of Electrical and Computer Engineering (ECE) at the University of Minnesota (UMN) – Twin Cities. He is a member of the SPiNCOM research group under the supervision of Prof. Georgios B. Giannakis. His research interests span the areas of machine learning, signal processing, network science and data science. Lately, he focuses on learning over graphs which can model complex networks including financial, social and biological ones to list a few. In the past, he has worked on the development of automatic aerial target recognition systems using passive Radar data. He has been awarded the UMN ECE Department fellowship (2019), Gerondelis Foundation scholarship (2020), Onassis Foundation scholarship (2021), the BEST PAPER AWARD at the International CIT&DS 2019 International Conference (2019) and the Outstanding REVIEWER AWARD (top 10 %) at the International Conference on Machine Learning (ICML 2022).

Qin Lu (M'18) is an assistant professor with the School of Electrical and Computer Engineering, College of Engineering, University of Georgia. Previously, she worked as a postdoctoral research associate at the University of Minnesota, Twin Cities. She received her B.S. and Ph.D. degrees in electrical engineering from the University of Electronic Science and Technology of China and the University of Connecticut (UConn) in 2013 and 2018, respectively. Her research interests span the areas of signal processing, machine learning, data science, and communications, with special focus on Gaussian processes, Bayesian optimization, spatio-temporal inference over graphs, and data association for multi-object tracking. She received the National Scholarship from China twice. She was awarded Summer Fellowship and Doctoral Dissertation Fellowship at UConn. She was also a recipient of the Women of Innovation Award in Collegian Innovation and Leadership by Connecticut Technology Council in March, 2018.

Georgios B. Giannakis (F'97) received his Diploma in Electrical Engr. from the Ntl. Tech. Univ. of Athens, Greece, 1981. From 1982 to 1986 he was with the Univ. of Southern California (USC), where he received his MSc. in Electrical Engineering, 1983, MSc. in Mathematics, 1986, and Ph.D. in Electrical Engr., 1986. He was a faculty member with the University of Virginia from 1987 to 1998, and since 1999 he has been a professor with the Univ. of Minnesota, where he holds an ADC Endowed Chair, a University of Minnesota McKnight Presidential Chair in ECE, and serves as director of the Digital Technology Center. His general interests span the areas of statistical learning, signal processing, communications, and networking - subjects on which he has published more than 480 journal papers, 780 conference papers, 25 book chapters, two edited books and two research monographs. Current research focuses on Data Science, and Network Science with applications to the Internet of Things, and power networks with renewables. He is the (co-) inventor of 34 issued patents, and the (co-) recipient of 10 best journal paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He also received the IEEE-SPS Norbert Wiener Society Award (2019); EURASIP's A. Papoulis Society Award (2020); Technical Achievement Awards from the IEEE-SPS (2000) and from EURASIP (2005); the IEEE ComSoc Education Award (2019); and the IEEE Fourier Technical Field Award (2015). He is a member of the Academia Europaea, and Fellow of the National Academy of Inventors, the European Academy of Sciences, IEEE and EURASIP. He has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SPS.