Active labeling for online ensemble learning

Konstantinos D. Polyzos*, Qin Lu‡, Georgios B. Giannakis*
*Dept. of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA
‡ School of Electrical and Computer Engineering, University of Georgia, Athens, GA, USA

***Abstract*—In many application domains including medical imaging, experimental design, as well as robotics, labeled data are expensive to acquire while unlabeled samples are abundant. Such high labeling costs well motivate the active learning (AL) paradigm that judiciously selects the most informative data instances to label. Specifically, this paper considers a streaming AL setting where unlabeled samples arrive sequentially and an *oracle* decides to label them or not based on a certain criterion. This active labeling process can benefit from a statistical function model, that provides well-calibrated uncertainty values to guide the oracle to make the informed labeling decision. Towards statistical modeling with adaptivity and robustness in the streaming setting, a recently developed ensemble Gaussian process (EGP) model is leveraged that has weights adapted to the labeled data collected incrementally. Building on this EGP model, this work advocates a novel labeling criterion where the oracle calculates the Kullback-Leibler divergence between the predictive pdfs of each unlabeled instance to make the labeling decision. Numerical tests on synthetic and real datasets in the regression task showcase the merits of the proposed EGP-AL approach relative to the competing alternatives.**

***Index terms*—** Active learning, Gaussian processes, online learning

I. Introduction

In machine learning (ML) and artificial intelligence, a gamut of learning tasks boil down to estimating a function. In supervised learning, given a budget of input-output data the goal is to learn a function that maps the input to the output data. Identifying such function may necessitate a sufficient number of input-output data. Although input data can be easily obtained, in several practical settings acquiring the output data (or labels) may be challenging due to privacy or high cost considerations. For example, in medical applications where labels can represent the medical condition of patients, obtaining these labels may require well trained experts or costly medical examinations, and in some cases cannot be revealed to preserve confidentiality. To cope with this challenge, active learning (AL) is a well motivated framework that provides principled methods to seek for a few yet informative input data to label, so that to effectively estimate the sought function even with fewer data at hand [1].

AL can be broadly categorized into *pool-based* [1] where the goal is to find the most informative unlabeled input data to label from a pool of unlabeled instances, and *stream-based* [2] where unlabeled instances arrive on-the-fly and the aim is to determine whether to label or disregard them. Emphasizing on the latter which arises in applications ranging from spam detection [3] to time series prediction [4], existing stream-based AL approaches capitalize on deep neural network architectures to devise acquisition strategies that decide whether to query or not new coming unlabeled instances [5], [6]. Recently, an ensemble multi-kernel stream-based AL approach was advocated in [7] that consists of a certain number of kernel-based learners and leverages the similarities between the function estimates of these learners so that to build an acquisition criterion to decide to label or not new unlabeled data. Albeit effective and appealing in several streaming scenarios, the aforementioned *deterministic* approaches fall short in inherently offering uncertainty quantification that can readily guide and aid the AL process. To that end, the Bayesian approaches in [8]–[10] rely on statistical models and acquisition criteria to further account for the model-associated uncertainty. Nonetheless, these approaches are tailored solely for the classification learning task, instead of the regression task that is the interest of the present work.

Focusing on the regression task, AL can benefit from the so-termed Gaussian process (GP) model that is capable of learning nonlinear functions with *uncertainty quantification* in a sample-efficient manner [11]. Given a typically small set of labeled data $\{\mathbf{x}_\tau, y_\tau\}_{\tau=1}^t$, learning with GPs yields the posterior probability density function (pdf) $p(f(\mathbf{x})|\mathcal{L}_t)$ of the sought function $f(\mathbf{x})$ for any unlabeled $\mathbf{x}$, which in regression is Gaussian with mean and variance given in closed-form. The model uncertainty captured by the posterior variance can guide the acquisition step of new unlabeled instances to be queried; see e.g., [12], [13]. Although interesting, these GP-based approaches are designed only for pool-based AL and not streaming AL that is the focus of the present work. In addition, their performance relies on a pre-selected kernel function, which significantly affects the performance of GP-based learning. How to properly select the fitted kernel function is a nontrivial task especially with only a few initially labeled data at hand. To deal with this challenge, the works in [14]–[17] advocate a Gaussian mixture (GM) posterior pdf that adaptively learns the proper kernel function as new data are processed online. Yet, this model is used only for conventional prediction-oriented tasks or pool-based AL and not streaming AL settings that the present work focuses on. Developing an adaptive GP-based framework for streaming AL settings, and devising acquisition criteria relying on this model to select which arriving unlabeled instances to label, are yet to be explored.

K. D. Polyzos and G. B. Giannakis were supported by NSF grants 2102312, 2103256, 1901134, 2126052, and 2128593. Q. Lu was supported by NSF grant 2340049. The work of K. D. Polyzos was also supported by the Onassis Foundation Scholarship. Emails: polyz003@umn.edu, Qin.Lu@uga.edu, georgios@umn.edu.

Contributions. To address the aforementioned challenges, the present work builds on an *ensemble* (E-) of GP models to prudently adapt to the appropriate GP model as new data arrive on-the-fly. Capitalizing on this EGP model, a novel acquisition critetion is advocated that leverages the statistical distance between the predictive pdfs of all GP models in the ensemble for each unlabeled instance, to determine whether to label it or not. Thorough tests on synthetic functions and real robotic-based regression problems demonstrate the impressive merits of the advocated EGP-based streaming AL method.

II. Preliminaries

In typical supervised learning problems, given a budget of labeled (or training) data $\mathcal{L}_0 := \{(\mathbf{x}_\tau, y_\tau)\}_{\tau=-L_0+1}^{0}$ [1] the goal is to estimate a function $f(\cdot)$ that maps each input feature vector $\mathbf{x}_\tau$ to the corresponding output y_τ that can be either a real number pertaining to a regression task or can belong to a finite alphabet in classification tasks; that is to learn $f(\cdot) : \mathbf{x}_\tau \to f(\mathbf{x}_\tau) \to y_\tau, \forall \tau$. A reliable estimate of $f(\cdot)$ may entail a sufficiently large number L_0 of initial labeled data. In several practical settings, although input data can be easily obtained, acquiring the corresponding labels may be expensive due to sampling costs or privacy considerations. In healthcare for instance, a label representing the medical condition of patients may not be revealed due to medical confidentiality or may need costly medical examinations to obtain. To cope with this challenge, one can rely on the AL paradigm that aims to prudently select the most *informative* input data to label so that to effectively estimate f in a data-efficient manner.

AL begins with the small-size initial set $\mathcal{L}_0$ of labeled samples and relying on the corresponding set $\mathcal{L}_t$ at time slot t, model-based AL typically adopts a probabilistic function model $p(f(\mathbf{x})|\mathcal{L}_t)$ to capture the sought function f for any input vector $\mathbf{x}$. Focusing on the streaming AL setting where data arrive sequentially, AL leverages the pdf $p(f(\mathbf{x})|\mathcal{L}_t)$ to form the so-termed *acquisition criterion* (AC) that determines whether to label or not each new coming input vector $\mathbf{x}_{t+1}$ [1]. If $\mathbf{x}_{t+1}$ is selected to be labeled, then upon querying an *oracle* for the associated label y_{t+1}, the labeled set is augmented as $\mathcal{L}_{t+1} := \mathcal{L}_t \cup \{(\mathbf{x}_{t+1}, y_{t+1})\}$, and if not $\mathcal{L}_{t+1} := \mathcal{L}_t$. Apparently, the two critical choices are the model for f, and the AC design. With emphasis on the regression task, next we will outline the GP-based model for f along with the associated AC.

A. GP-based streaming AL

GPs have been widely adopted to learn a nonparametric function estimate along with its associated uncertainty in a sample-efficient manner [11] that is appealing in AL settings. Learning with GPs starts with the assumption that a GP prior is postulated on f; that is $f \sim \mathcal{GP}(0, \kappa(\mathbf{x}, \mathbf{x}'))$ with $\kappa(\mathbf{x}, \mathbf{x}')$ denoting a positive-definite kernel function that measures the pairwise similarity between two distinct input vectors $\mathbf{x}$ and $\mathbf{x}'$. This implies that random vector $\mathbf{f}_t := [f(\mathbf{x}_1) \dots f(\mathbf{x}_t)]^\top$ (where $^\top$ denotes transposition) comprising the the function evaluations at inputs $\mathbf{X}_t := [\mathbf{x}_1 \dots \mathbf{x}_t]^\top$ 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 $\mathbf{K}_t$ is the $t \times t$ covariance (kernel) matrix whose (i, j) entry is $[\mathbf{K}_t]_{i,j} = \mathrm{cov}(f(\mathbf{x}_i), f(\mathbf{x}_j)) := \kappa(\mathbf{x}_i, \mathbf{x}_j)$ [11].

The next assumption that relates the random vector $\mathbf{f}_t$ with the (possibly noisy) output data $\mathbf{y}_t := [y_1 \cdots y_t]^\top$, is that the batch likelihood $p(\mathbf{y}_t|\mathbf{f}_t; \mathbf{X}_t)$ can be written as $p(\mathbf{y}_t|\mathbf{f}_t; \mathbf{X}_t) = \prod_{\tau=1}^{t} p(y_\tau|f(\mathbf{x}_\tau))$ where in the regression task the per-datum likelihood is $p(y_\tau|f(\mathbf{x}_\tau)) = \mathcal{N}(y_\tau; f(\mathbf{x}_\tau), \sigma_n^2)$. With the GP prior and batch likelihood at hand, it can be shown that the posterior pdf $p(f(\mathbf{x}_{t+1})|\mathbf{y}_t; \mathbf{X}_t)$ of a new unlabeled instance $\mathbf{x}_{t+1}$ at slot $t+1$ is [11]

$$p(f(\mathbf{x}_{t+1})|\mathbf{y}_t; \mathbf{X}_t) = \mathcal{N}(f(\mathbf{x}_{t+1}); \mu_t(\mathbf{x}_{t+1}), \sigma_t^2(\mathbf{x}_{t+1})) \quad (1)$$

where

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

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

and $\mathbf{k}_t(\mathbf{x}_{t+1}) := [\kappa(\mathbf{x}_1, \mathbf{x}_{t+1}), \dots, \kappa(\mathbf{x}_t, \mathbf{x}_{t+1})]^\top$.

Note that the mean in (2a) is a point estimate of $f(\mathbf{x}_{t+1})$, while the variance in (2b) quantifies the associated uncertainty, where it is intuitive that the larger the variance the more uncertain is the corresponding function estimate. Following the uncertainty-based criteria in [13], [18] that select instances to label with high uncertainty, the corresponding acquisition criterion in the streaming AL is expressed as

$$\sigma_t^2(\mathbf{x}_{t+1}) \geq \eta_\sigma \quad (3)$$

where η_σ is a pre-defined threshold. When the criterion in (3) is satisfied, the uncertainty associated with $f(\mathbf{x}_{t+1})$ is sufficiently large so that to query an oracle and obtain the label y_{t+1}.

Albeit interesting, the performance of GP-based AL hinges on a preselected kernel function, which may exhibit limited expressiveness. Finally, the acquisition criterion in (3) relies solely on the posterior variance without taking into account the information provided by the posterior mean. These limitations can be ameliorated via a novel ensemble approach, as delineated next.

III. Streaming AL with an ensemble (E) of GPs

To bypass the nontrivial task of pre-selecting a proper kernel, and to allow for a richer function space, we advocate an ensemble (E) of M distinct GPs, each relying on a kernel function selected from a given dictionary $\mathcal{K} := \{\kappa^1, \dots, \kappa^M\}$ that comprises kernels of different types and/or different hyperparameters. For each GP $m \in \mathcal{M} := \{1, \dots, M\}$, a GP prior is postulated as $f|m \sim \mathcal{GP}(0, \kappa^m(\mathbf{x}, \mathbf{x}'))$. Combining all GP priors yields the EGP prior of $f(\mathbf{x})$, expressed as

$$f(\mathbf{x}) \sim \sum_{m=1}^{M} w_0^m \mathcal{GP}(0, \kappa^m(\mathbf{x}, \mathbf{x}')), \quad \sum_{m=1}^{M} w_0^m = 1 \quad (4)$$

where the per GP weight $w_0^m := \Pr(i = m)$ is deemed as probability that assesses the contribution of GP model m in

[1] The negative instance index here is used for notational brevity as more labeled data will be added next.

the ensemble. Note that the *latent* variable i is introduced to indicate the contribution from GP model m. Although the Gaussian mixture (GM) prior in (4) has been adopted for online prediction-oriented tasks in [14], [19], the present work is the first to adapt it to the streaming AL setting of interest, where the design of an acquisition criterion is also needed.

A. EGP-based streaming AL with RF approximation

When the kernels in the dictionary $\mathcal{K}$ are shift-invariant, for each GP model m the RF approximation draws D independent and identically distributed (i.i.d) random vectors $\{\mathbf{v}_j^m\}_{j=1}^D$ from $\pi_{\bar{\kappa}}^m(\mathbf{v})$, which is the spectral density of the standardized kernel $\bar{\kappa}^m = \kappa^m/\sigma_{\theta^m}^2$. These are used to construct the per GP model RF vector defined as $\boldsymbol{\phi}_{\mathbf{v}^m}(\mathbf{x}) := \frac{1}{\sqrt{D}}\left[\sin(\mathbf{v}_1^{m\top}\mathbf{x}), \cos(\mathbf{v}_1^{m\top}\mathbf{x}) \cdots \sin(\mathbf{v}_D^{m\top}\mathbf{x}), \cos(\mathbf{v}_D^{m\top}\mathbf{x})\right]^\top$. Then a generative model that relates f and the noisy output y per GP m via the $2D \times 1$ vector $\boldsymbol{\theta}^m$ is [20]

$$\begin{aligned} p(\boldsymbol{\theta}^m) &= \mathcal{N}(\boldsymbol{\theta}^m; \mathbf{0}_{2D}, \sigma_{\theta^m}^2 \mathbf{I}_{2D}) \\ p(f(\mathbf{x})|i=m, \boldsymbol{\theta}^m) &= \delta(f(\mathbf{x}) - \boldsymbol{\phi}_{\mathbf{v}}^{m\top}(\mathbf{x})\boldsymbol{\theta}^m) \\ p(y|\boldsymbol{\theta}^m; \mathbf{x}) &= \mathcal{N}(y; \boldsymbol{\phi}_{\mathbf{v}}^{m\top}(\mathbf{x})\boldsymbol{\theta}^m, \sigma_n^2)\ . \end{aligned} \tag{5}$$

This parametric generative model allows each GP to characterize the learning function via the parametric posterior pdf $p(\boldsymbol{\theta}^m|\mathbf{y}_t; \mathbf{X}_t) = \mathcal{N}(\boldsymbol{\theta}^m; \hat{\boldsymbol{\theta}}_t^m, \boldsymbol{\Sigma}_t^m)$ along with the weight $w_t^m := \Pr(i = m|\mathbf{y}_t; \mathbf{X}_t)$. Next we will show how the ensemble posterior pdf is propagated at each time slot by updating the set $\{w_t^m, \boldsymbol{\theta}_t^m, \boldsymbol{\Sigma}_t^m, m \in \mathcal{M}\}$, and introduce the advocated EGP-based AC criterion that decides whether to label or not each arriving unlabeled input vector.

At the end of slot t each model keeps track of its posterior $p(\boldsymbol{\theta}^m|\mathbf{y}_t; \mathbf{X}_t) = \mathcal{N}(\boldsymbol{\theta}^m; \hat{\boldsymbol{\theta}}_t^m, \boldsymbol{\Sigma}_t^m)$, and upon the arrival of $\mathbf{x}_{t+1}$ at slot $t+1$ each GP model forms its predictive pdf of y_{t+1} as

$$p(y_{t+1}|i=m, \mathbf{y}_t; \mathbf{X}_{t+1}) = \mathcal{N}(y_{t+1}; \hat{y}_{t+1|t}^m, (\sigma_{t+1|t}^m)^2) \tag{6}$$

with $(\sigma_{t+1|t}^m)^2 = \boldsymbol{\phi}_{\mathbf{v}}^{m\top}(\mathbf{x}_{t+1})\boldsymbol{\Sigma}_t^m\boldsymbol{\phi}_{\mathbf{v}}^m(\mathbf{x}_{t+1}) + \sigma_n^2$ and $\hat{y}_{t+1|t}^m = \boldsymbol{\phi}_{\mathbf{v}}^{m\top}(\mathbf{x}_{t+1})\hat{\boldsymbol{\theta}}_t^m$. Thus, the ensemble pdf of y_{t+1} is given by

$$p(y_{t+1}|\mathbf{y}_t; \mathbf{X}_{t+1}) = \sum_{m=1}^M w_t^m \mathcal{N}(y_{t+1}; \hat{y}_{t+1|t}^m, (\sigma_{t+1|t}^m)^2). \tag{7}$$

Next, a pre-selected AC (that will be discussed in the next subsection) determines whether to obtain or not the label y_{t+1}. If the AC is satisfied, y_{t+1} is queried and the per-GP weight $w_{t+1}^m := \Pr(i = m|\mathbf{y}_{t+1}; \mathbf{X}_{t+1})$ is then updated via Bayes' rule as

$$w_{t+1}^m = \frac{w_t^m \mathcal{N}\left(y_{t+1}; \hat{y}_{t+1|t}^m, (\sigma_{t+1|t}^m)^2\right)}{\sum_{m'=1}^M w_t^{m'} \mathcal{N}\left(y_{t+1}; \hat{y}_{t+1|t}^{m'}, (\sigma_{t+1|t}^{m'})^2\right)}. \tag{8}$$

In addition, the posterior pdf of $\boldsymbol{\theta}^m$ is propagated in a recursive Bayes manner as

$$\begin{aligned} p(\boldsymbol{\theta}^m|\mathbf{y}_{t+1}; \mathbf{X}_{t+1}) &= \frac{p(\boldsymbol{\theta}^m|\mathbf{y}_t; \mathbf{X}_t) p(y_{t+1}|\boldsymbol{\theta}^m; \mathbf{x}_{t+1})}{p(y_{t+1}|, i=m, \mathbf{X}_{t+1}, \mathbf{y}_t)} \\ &= \mathcal{N}(\boldsymbol{\theta}^m; \hat{\boldsymbol{\theta}}_{t+1}^m, \boldsymbol{\Sigma}_{t+1}^m) \end{aligned} \tag{9}$$

with mean and covariance given by

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

Relying on the RF-based EGP model, the following subsection will introduce the EGP-based AC that decides if the label y_{t+1} at slot $t+1$ is queried or not.

B. EGP-based AC

At slot $t+1$, upon forming the per-GP predictive pdf (6) for the label y_{t+1} of $\mathbf{x}_{t+1}$, the following novel AC is assessed so as to make a decision on whether to query an oracle for the ground-truth label value y_{t+1} or not

$$\max_{m' \in \mathcal{M}} \sum_{m \in \mathcal{M}} w_t^m D_{KL}(\mathcal{N}(y_{t+1}; \hat{y}_{t+1|t}^m, (\sigma_{t+1|t}^m)^2) || \mathcal{N}(y_{t+1}; \hat{y}_{t+1|t}^{m'}, (\sigma_{t+1|t}^{m'})^2)) \leq \eta_k \tag{10}$$

where $\eta_k > 0$ is a pre-defined parameter and $D_{KL}(P||Q)$ denotes the Kullback–Leibler (KL) divergence that is a measure of statistical distance between pdfs P and Q. This criterion capitalizes on the pairwise similarities of the GP-based predictive pdfs. Specifically, when the M GP models have similar predictive pdfs of y_{t+1}, it is intuitive that obtaining the ground-truth label y_{t+1} will have a small influence on the GP models weights (c.f. (8)). In this case, the label y_{t+1} is not queried from the oracle so that to ensure a sufficient efficiency-accuracy trade-off. It is worth noticing that the AC in (10) is assessed using only the predictive pdfs of all GP models without knowledge of the true label y_{t+1}.

Remark 1. The variance-based criterion in (3) can be readily applied in the advocated EGP model by replacing the variance of the single GP model with the variance of the GP mixture of the function posterior (c.f. (7)) which is expressed as

$$(\sigma_{t+1|t}^{\text{ens}})^2 := \sum_{m=1}^M w_t^m ((\sigma_{t+1|t}^m)^2 + (\hat{y}_{t+1|t}^m - \hat{\bar{y}}_{t+1|t})^2)$$

where $\hat{\bar{y}}_{t+1|t} := \sum_{m=1}^M w_t^m \hat{y}_{t+1|t}^m$, yielding the criterion

$$(\sigma_{t+1|t}^{\text{ens}})^2 \geq \eta_\sigma\ . \tag{11}$$

Nonetheless, compared to the advocated AC in (10) the criterion in (11) focuses solely on the uncertainty offered by the GM variance without taking into account the similarities of the predictive pdfs of all GP models.

Remark 2. The deterministic RF-based ensemble multi-kernel acquisition criterion counterpart for streaming AL advocated in [7], relies on the similarities between the function estimates of the different kernel-based learners. This criterion can be equivalently written in the Bayesian EGP-based setting as follows

$$\max_{m' \in \mathcal{M}} \sum_{m \in \mathcal{M}} w_t^m (\hat{y}_{t+1|t}^m - \hat{y}_{t+1|t}^{m'})^2 \leq \eta_d\ . \tag{12}$$

Although interesting and effective in several streaming AL settings, the criterion in (12) does not consider the uncertainty offered by the variances of the M GP models that can markedly impact the streaming AL performance.

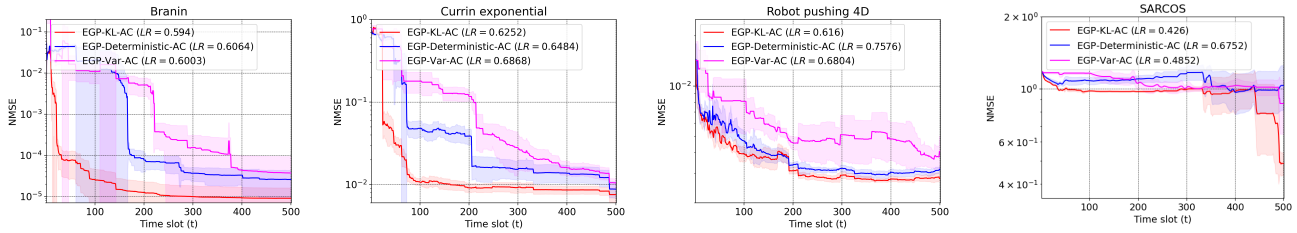

Fig. 1: Average NMSE performance vs time slot (t) of all competing streaming AL approaches

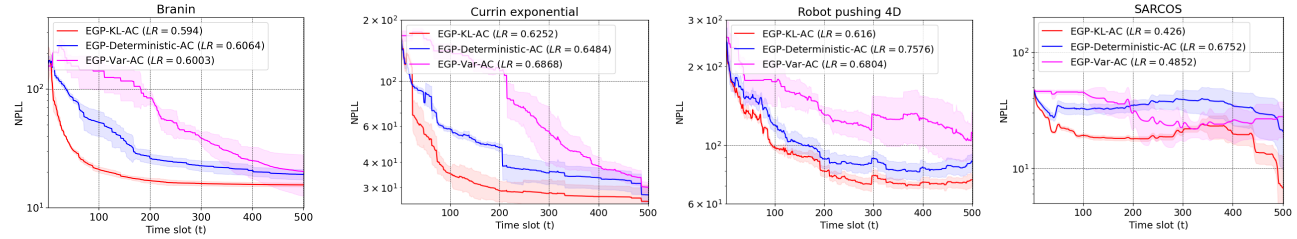

Fig. 2: Average NPLL performance vs time slot (t) of all competing streaming AL approaches

IV. Numerical Tests

In this section, the performance of the novel EGP-based streaming AL approach that leverages the AC in (10) (abbreviated as 'EGP-KL-AC') is evaluated on the Branin and Currin exponential standard synthetic functions, and real robotic problems whose description is given below

Robot pushing 4D. In this task, a robot pushes an object to a certain location [21]. With input vector $\mathbf{x}_\tau := [r^x_\tau, r^y_\tau, t^p_\tau, r^\theta_\tau]^\top$ at each slot τ consisting of the robot location (r^x_τ, r^y_τ), the pushing duration t^p_τ and pushing angle r^θ_τ, a regression task is formed where the aim is to map $\mathbf{x}_\tau$ to the output $y_\tau := ||\mathbf{o}_\tau - \mathbf{d}||_2$ with $\mathbf{o}_\tau := (o^x_\tau, o^y_\tau)$ representing the ending location of the object and $\mathbf{d} := [d^x, d^y]$ denoting a pre-defined position vector. This task is of practical interest in several robotic problems such as obstacle avoidance; see e.g [21], [22].

SARCOS. This dataset considers a seven degrees-of-freedom SARCOS anthropomorphic robot arm [11]. A regression task is formed with input vector comprising seven joint positions, seven joint velocities and seven joint accelerations, and output value to be predicted representing the first of the corresponding seven joint torques.

In our experimental setting, 10 and 50 initial labeled data are used for training on the synthetic functions and robotic problems respectively, and 1000 data are used for testing. We compare the performance of the advocated 'EGP-KL-AC' approach with the EGP-based approaches that utilize the criteria in (11) and (12) which will be abbreviated as 'EGP-Var-AC' and 'EGP-Deterministic-AC' respectively. In all competing approaches the kernel dictionary consists of $M = 11$ distinct RBF kernels with characteristic lengthscale selected from $\{10^c\}_{c=-4}^{6}$, and the weights of all GP models are initialized as $w_0^m = 1/M, \ \forall m \in \mathcal{M}$. For the RF approximation D is set to 50 and the kernel hyperparameters for each GP model are obtained maximizing the marginal log-likelihood using the initial labeled data. The performance of all competing approaches is evaluated utilizing the normalized mean square error (NMSE) and negative predictive log-likelihood (NPLL) metrics that are similarly defined as in [23]. To further assess the accuracy-efficiency trade-off, the NMSE and NPLL metrics are combined with the labeling rate (LR) that is defined as $LR = T_l/T$ where T_l denotes the number of labeled instances in the AL process and T the total number of new coming data that arrive sequentially.

The average NMSE and NPLL generalization performance of all competing approaches along with the corresponding standard deviation are reported for 10 independent runs. In Figs. 1-2, it is evident that the advocated EGP-KL-AC method consistently outperforms the EGP-Var-AC and EGP-Deterministic-AC baselines in all datasets. It is worth mentioning that the EGP-KL-AC enjoys the lowest NMSE and NPLL while also having the smallest LR value[2], as can be seen in Figs. 1-2, implying that the novel approach can achieve low prediction error with less labeled samples and hence have the best accuracy-efficiency trade-off compared to the baselines. This corroborates the benefits of coupling the advocated EGP model with the AC in (10) that considers the similarities of the predictive pdfs of all GP models in the ensemble.

V. Conclusions

This work considered a streaming AL setting where unlabeled data instances arrive sequentially and ask to be labeled. Building on an ensemble of GP models that adaptively selects the proper GP model on-the-fly, a novel AC was introduced that relied on the KL divergence between any two predictive pdfs for each unlabeled instance. Tests on synthetic and real datasets showcase the merits of the advocated AL approach.

[2]The values of η_k, η_σ and η_d are selected so that to have the LR values reported in Figs 1-2, and are omitted due to space limitations

REFERENCES

[1] Burr Settles, "Active learning," *Synthesis Lectures on Artif. Intel. and Mach. Learn.*, vol. 6, no. 1, pp. 1–114, 2012.

[2] Davide Cacciarelli and Murat Kulahci, "Active learning for data streams: a survey," *Machine Learning*, pp. 1–55, 2023.

[3] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker, "Identifying suspicious urls: an application of large-scale online learning," in *Proc. Intl. Conf. Mach. Learn.*, 2009, pp. 681–688.

[4] Cédric Richard, José Carlos M Bermudez, and Paul Honeine, "Online prediction of time series data with kernels," *IEEE Trans. Sig. Process.*, vol. 57, no. 3, pp. 1058–1067, 2008.

[5] Akanksha Saran, Safoora Yousefi, Akshay Krishnamurthy, John Langford, and Jordan T Ash, "Streaming active learning with deep neural networks," *arXiv preprint arXiv:2303.02535*, 2023.

[6] Kleanthis Malialis, Christos G Panayiotou, and Marios M Polycarpou, "Nonstationary data stream classification with online active learning and siamese neural networks," *Neurocomputing*, vol. 512, pp. 235–252, 2022.

[7] Songnam Hong and Jeongmin Chae, "Active learning with multiple kernels," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 7, pp. 2980–2994, 2021.

[8] Sanmin Liu, Shan Xue, Jia Wu, Chuan Zhou, Jian Yang, Zhao Li, and Jie Cao, "Online active learning for drifting data streams," *IEEE Trans. Neural Networks Learn. Syst.*, 2021.

[9] Xingquan Zhu, Peng Zhang, Xiaodong Lin, and Yong Shi, "Active learning from stream data using optimal weight classifier ensemble," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 6, pp. 1607–1621, 2010.

[10] Wei Chu, Martin Zinkevich, Lihong Li, Achint Thomas, and Belle Tseng, "Unbiased online active learning in data streams," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011, pp. 195–203.

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

[12] Andreas Krause, Ajit Singh, and Carlos Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies.," *J. Mach. Learn. Res.*, vol. 9, no. 2, 2008.

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

[14] 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.*, pp. 1910–1920, 2020.

[15] Qin Lu, Georgios V Karanikolas, and Georgios B Giannakis, "Incremental ensemble Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 45, no. 2, pp. 1876–1893, 2022.

[16] Konstantinos D Polyzos, Qin Lu, and Georgios B Giannakis, "Ensemble Gaussian processes for online learning over graphs with adaptivity and scalability," *IEEE Trans. Sig. Process.*, 2021.

[17] Konstantinos D. Polyzos, Qin Lu, and Georgios B. Giannakis, "Weighted ensembles for active learning with adaptivity," *arXiv:2206.05009*, 2022.

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

[19] Georgios V Karanikolas, Qin Lu, and Georgios B Giannakis, "Online unsupervised learning using ensemble Gaussian processes with random features," *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Process.*, pp. 3190–3194, 2021.

[20] Ali Rahimi and Benjamin Recht, "Random features for large-scale kernel machines," *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1177–1184, 2008.

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

[22] Qin Lu, Konstantinos D Polyzos, Bingcong Li, and Georgios B Giannakis, "Surrogate modeling for Bayesian optimization beyond a single Gaussian process," *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 45, no. 9, pp. 11283–11296, 2023.

[23] 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.*, 2022, pp. 58–64.