Active Sampling over Graphs for Bayesian Reconstruction with Gaussian Ensembles

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

Department of Electrical and Computer Engineering, University of Minnesota, USA

2022 56th Asilomar Conference on Signals, Systems, and Computers | 978-1-6654-5906-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/IEEECONF56349.2022.10051888

***Abstract*—Graph-guided semi-supervised learning (SSL) has gained popularity in several network science applications, including biological, social, and financial ones. SSL becomes particularly challenging when the available nodal labels are scarce, what motivates naturally the active learning (AL) paradigm. AL seeks the most informative nodes to label in order to effectively estimate the nodal values of unobserved nodes. It is also referred to as active sampling, and boils down to learning the sought function mapping, and an acquisition function (AF) to identify the next node(s) to sample. To learn the mapping, this work leverages an adaptive Bayesian model comprising an ensemble (E) of Gaussian Processes (GPs) with enhanced expressiveness of the function space. Unlike most alternatives, the EGP model relies only on the one-hop connectivity of each node. Capitalizing on this EGP model, a suite of novel and intuitive AFs are developed to guide the active sampling process. These AFs are then combined with weights that are adapted incrementally to further robustify performance. Numerical tests on real and synthetic datasets corroborate the merits of the novel methods.**

***Index Terms*—Gaussian processes, ensemble learning, active learning, semi-supervised learning over graphs**

I. Introduction

In the last decade, semi-supervised learning (SSL) over graphs has received growing attention from the scientific community at the crossroads of machine learning and network science, on the premise of its major impact in diverse fields such as medicine, biology, and financing [5]. Given observations from a subset of nodes, SSL on graphs can reconstruct unobserved nodal values by leveraging the connectivity of nodes. In practice however, only a few nodes can be observed due to privacy concerns or high sampling costs. In biomedical networks for instance, a medical doctor will not reveal a patient's record to respect confidentiality, while in protein networks some attributes may require expensive and time-consuming medical tests. The scarcity of nodal observations motivates the active learning (AL) paradigm.

In contrast to passive SSL approaches that rely on a given or randomly chosen set of observed nodes, the goal of AL is to prudently select which nodes to query and add to the observed set in order to improve prediction performance; that is, to sample the most informative nodes from a large set of unobserved nodes. AL over graphs is also known as active sampling, and requires (i) a model to learn the sought graph function that maps nodes to nodal values; and (ii) an acquisition function (AF) or sampling strategy to query nodes from the unlabeled set. Focusing on Bayesian models that offer uncertainty quantification compared to the deterministic ones, a Gauss-Markov random field (GMRF) model was adopted for (i) in [21], in conjunction with the so-termed 'Σ-optimality' AF for (ii) that minimizes the sum of the entries of the predictive covariance. Capitalizing on a GMRF model for the correlation of labels across neighboring nodes, [2] put forth an AF that selects to query the node causing the largest change on the GMRF model.

Belonging to the family of nonparametric Bayesian models, Gaussian processes (GPs) have been extensively used for AL because of their ability to learn the probability density function (pdf) of a random function in a sample-efficient manner; see e.g., [8], [9]. In the graph context, the AF in [8] uses the predictive mean and variance, and has also been adopted in [35] combined with a manifold-preserving reduction step that yields a sparse manifold graph. A GP-based approach using the graph Laplacian is coupled in [24] with a scalable variational inference scheme leveraging a 'Σ-optimal' AF [21]. Non-Gaussian models for (i) have relied on the Laplace approximation to obtain a Gaussian proxy of the non-Gaussian pdf, based on which a node is queried using the 'largest model change' criterion [23]. Albeit interesting, the aforementioned approaches rely on a single GP model that may exhibit limited expressiveness of the function space, and pertain only to the classification task, where the nodal values (or labels) are drawn from a finite alphabet.

Contributions. To cope with these limitations, we put forth a novel Bayesian approach that relies on an ensemble of Gaussian Processes (EGPs) for graph-guided AL. Besides allowing for a richer function space, the EGP model of the sought function uses only the one-hop connectivity vector of each node without requiring additional nodal features, and updates incrementally the model parameters with no need for retraining, a property that fits nicely the AL setup. Although EGPs were introduced in [18], [19], and also employed for graph-guided learning [25], [26], [27], reinforcement learning [13], [15], [29], Bayesian optimization [20] and conventional active learning [28], it is the first time to be utilized for active sampling over graphs. In addition, the advocated EGP model can readily accommodate a suite of novel and intuitive AFs that rely on the disagreement and uncertainty based rules. Further adopting a weighted ensemble of these AFs with weights properly adapted as data arrive incrementally, leads to enhanced robustness. Experimental tests on real and synthetic datasets showcase the benefits of both the EGP model and the

This work was supported by NSF grants 2126052, 2103256, 2102312, 2128593, and 1901134. The work of Konstantinos D. Polyzos was also supported by the Onassis Foundation Scholarship.

Emails: {*polyz003, qlu, georgios*}*@umn.edu*

accompanying novel acquisition criteria.

II. Problem formulation

Consider a graph that consists of N nodes collected in the vertex set $\mathcal{V} := \{1, \ldots, N\}$, and E edges connecting pairs of nodes. The connectivity of nodes is captured by the $N \times N$ adjacency matrix $\mathbf{A}$, whose (n, n')th entry $a_{nn'} := \mathbf{A}(n, n')$ represents a weighted edge connecting node n to node n'. A real-valued function $f(\cdot)$ on the graph is a mapping from a node $n \in \mathcal{V}$ to its noise-free nodal value f_n, which further yields the (possibly noisy) nodal observation y_n. For instance, f_n could be the age of user n in a social network.

If sampling nodal values incurs high cost, the training data is limited, and one deals with weak or semi-supervised learning (SSL) of the map from the observed nodal values $\mathcal{O}$ to the unobserved ones $\bar{\mathcal{O}}$. SSL has been carried out incrementally by utilizing the one-hop connectivity vector $\mathbf{a}_n := \mathbf{A}(:, n)$ of node n as the input to function f, that is, $f_n := f(\mathbf{a}_n)$ [25], [26], [34]. A reliable estimate of f requires sufficiently many observations $\{y_n\}$, which may not be always feasible in practice. This motivates well the AL paradigm that aims at selecting the few *most informative* nodes to sample, in order to efficiently and effectively estimate $f(\cdot)$.

AL begins with a small-size set of observed nodes $\mathcal{L}_0 := \{\mathbf{a}_n, y_n, n \in \mathcal{S}_0\}$ with $\mathcal{S}_0$ collecting initially sampled nodes, and a larger set of unobserved nodes $\mathcal{U}_0 := \{\mathbf{a}_n, n \in \bar{\mathcal{S}}_0\}$, where $\bar{\mathcal{S}}_0 := \mathcal{V} \setminus (\mathcal{S}_0 \cup \bar{\mathcal{O}})$. Relying on the sets $\mathcal{L}_t$ and $\mathcal{U}_t$ at time slot t, AL capitalizes on the function's probability density model $p(f(\mathbf{a})|\mathcal{L}_t)$ to build the so-termed *acquisition function* (AF) $\alpha(\cdot)$ that selects the one-hop connectivity vector $\mathbf{a}_{n_{t+1}} \in \mathcal{U}_t$ of node $n_{t+1} \in \bar{\mathcal{S}}_t$ at slot $t+1$ as

$$\mathbf{a}_{n_{t+1}} = \underset{\mathbf{a} \in \mathcal{U}_t}{\arg\max}\ \alpha(\mathbf{a}; \mathcal{L}_t)\ . \tag{1}$$

The AF looks for the most informative unobserved node to sample, leveraging the quantifiable uncertainty captured by $p(f(\mathbf{a})|\mathcal{L}_t)$ that aids function space exploration. Subsequently, an *oracle* is queried to reveal the associated value $y_{n_{t+1}}$ of node n_{t+1}, which can be either a real value in a regression task or a class label drawn from a finite alphabet in a classification task. With $y_{n_{t+1}}$ at hand, the observed (or labeled) set is augmented as $\mathcal{L}_{t+1} := \{\mathbf{a}_n, y_n, n \in \mathcal{S}_{t+1}\}$ with $\mathcal{S}_{t+1} = \mathcal{S}_t \cup \{n_{t+1}\}$, while the unlabeled set becomes $\mathcal{U}_{t+1} := \{\mathbf{a}_n, n \in \bar{\mathcal{S}}_{t+1}\}$ with $\bar{\mathcal{S}}_{t+1} = \bar{\mathcal{S}}_t \setminus \{n_{t+1}\}$. Thus, the critical choices for AL are the belief model for f and the AF α. In the next section, we will outline the GP-based Bayesian model for f along with the corresponding α that can quantitatively capture the associated uncertainty.

III. Active sampling with a single GP

GPs have well-documented merits in estimating the probability density function (pdf) of a nonparametric map in a sample-efficient manner [31]. This renders GPs a valuable model for AL; see e.g., [8], [9]. Learning with GPs starts with a GP prior on the sought function f; that is, $f \sim \mathcal{GP}(0, \kappa(\mathbf{a}, \mathbf{a}'))$, where $\kappa(\mathbf{a}, \mathbf{a}')$ is a kernel function that measures the pairwise similarity between the connectivity input vectors $\mathbf{a}$ and $\mathbf{a}'$. This $\mathcal{GP}$ definition implies that the random vector $\mathbf{f}_t := [f(\mathbf{a}_{n_1}) \ldots f(\mathbf{a}_{n_t})]^\top$ ($^\top$ for transposition) comprising all function values for inputs $\mathbf{A}_t := [\mathbf{a}_{n_1} \ldots \mathbf{a}_{n_t}]^\top$ with $\{n_\tau\}_{\tau=1}^t$ referring to the index of all labeled nodes up to slot t (including the $|\mathcal{L}_0|$ initially labeled ones), is Gaussian distributed as $p(\mathbf{f}_t|\mathbf{A}_t) = \mathcal{N}(\mathbf{f}_t; \mathbf{0}_t, \mathbf{K}_t)\ \forall t$, where $\mathbf{K}_t$ is the $t \times t$ covariance matrix whose (m, m') entry is $[\mathbf{K}_t]_{m,m'} = \mathrm{cov}(f(\mathbf{a}_{n_m}), f(\mathbf{a}_{n_{m'}})) := \kappa(\mathbf{a}_{n_m}, \mathbf{a}_{n_{m'}})$ [31].

The output data $\mathbf{y}_t := [y_1 \cdots y_t]^\top$[1] are linked with the function evaluations $\mathbf{f}_t$ through the likelihood $p(\mathbf{y}_t|\mathbf{f}_t; \mathbf{A}_t)$ that is supposed to be factored as $p(\mathbf{y}_t|\mathbf{f}_t; \mathbf{A}_t) = \prod_{\tau=1}^t p(y_\tau|f(\mathbf{a}_{n_\tau}))$ with known per-datum factors $p(y_\tau|f(\mathbf{a}_{n_\tau}))$. Focusing on the regression task, where $p(y_\tau|f(\mathbf{a}_{n_\tau})) = \mathcal{N}(y_\tau; f(\mathbf{a}_{n_\tau}), \sigma_n^2)$, the predictive pdf of the nodal value y_{t+1} of an unlabeled node with connectivity vector $\mathbf{a}$ is given by [31]

$$p(y_{t+1}|\mathcal{L}_t, \mathbf{a}) = \mathcal{N}(y_{t+1}; \hat{y}_{t+1|\mathbf{t}}(\mathbf{a}), \sigma^2_{t+1|\mathbf{t}}(\mathbf{a})). \tag{2}$$

The first two moments of the pdf in (2) are

$$\hat{y}_{t+1|\mathbf{t}}(\mathbf{a}) = \mathbf{k}_t^\top(\mathbf{a})(\mathbf{K}_t + \sigma_n^2\mathbf{I}_t)^{-1}\mathbf{y}_t \tag{3a}$$

$$\sigma^2_{t+1|\mathbf{t}}(\mathbf{a}) = \kappa(\mathbf{a}, \mathbf{a}) - \mathbf{k}_t^\top(\mathbf{a})(\mathbf{K}_t + \sigma_n^2\mathbf{I}_t)^{-1}\mathbf{k}_t(\mathbf{a}) + \sigma_n^2 \tag{3b}$$

with $\mathbf{k}_t(\mathbf{a}) := [\kappa(\mathbf{a}, \mathbf{a}_{n_1}), \ldots, \kappa(\mathbf{a}, \mathbf{a}_{n_t})]^\top$ and $t+1|\mathbf{t}$ signifying that all t nodes up until slot t have been employed to obtain $p(y_{t+1}|\mathcal{L}_t, \mathbf{a})$. It is worth mentioning that the mean in (3a) provides a label (or nodal value) prediction corresponding to $\mathbf{a}$, while the variance in (3b) offers quantification of the associated uncertainty. Most GP-based AL settings leverage this uncertainty to select the next node to be queried using the following acquisition function (AF)

$$\mathbf{a}_{n_{t+1}} = \underset{\mathbf{a} \in \mathcal{U}_t}{\arg\max}\ \sigma^2_{t+1|\mathbf{t}}(\mathbf{a}) \tag{4}$$

which for the Gaussian pdf is equivalent to maximizing the entropy [22].

Albeit interesting, the predictive mean and variance in (3) incur complexity $\mathcal{O}(t^3)$, which although affordable in AL settings when t is small, it can be further reduced. In addition, (3) entails direct access to the one-hop connectivity vector $\mathbf{a}$, thus discouraging privacy-sensitive scenarios. Furthermore, GP-based AL hinges on a pre-selected kernel that may confine the resultant function space expressiveness. To cope with these limitations, a novel ensemble approach is advocated that capitalizes on random spectral features, as described next.

IV. Modeling with an ensemble of GPs

Broadening the scope of the active sampling approach over graphs based on a single GP with a *pre-selected* kernel, this section deals with an ensemble (E) of M GP experts to learn the sought function model with richer expressiveness. Specifically, each GP expert $m \in \mathcal{M} := \{1, \ldots, M\}$ relies on a unique kernel chosen from a predefined kernel dictionary $\mathcal{K} := \{\kappa^m\}_{m=1}^M$, where the kernels have different hyperparameters, and may be of different type. Each expert postulates

[1]Note that the output data $\{y_{n_\tau}\}_{\tau=1}^t$ corresponding to nodes $\{n_\tau\}_{\tau=1}^t$ will be henceforth abbreviated as $\{y_\tau\}_{\tau=1}^t$ for notational brevity.

a unique GP prior on f as $f|m \sim \mathcal{GP}(0, \kappa^m(\mathbf{a}, \mathbf{a}'))$, and an EGP meta-learner combines them as

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

where each expert's weight $w_0^m := \Pr(i = m)$ measures its significance in the EGP model. As newly labeled data in AL become available incrementally, the EGP-based predictive pdf can be expressed using the sum-product rule as

$$p(y_{t+1}|\mathcal{L}_t, \mathbf{a}) = \sum_{m=1}^{M} \Pr(i=m|\mathcal{L}_t) p(y_{t+1}|i=m, \mathcal{L}_t, \mathbf{a}) \tag{6}$$

which is a Gaussian mixture (GM) with weights $\{w_t^m := \Pr(i = m|\mathcal{L}_t)\}_{m=1}^M$ measuring the contribution of experts, and enabling model adaptation online.

To further reduce complexity of the EGP model and allow for online updates, which is particularly appealing for AL, we will adopt a low-rank parametric function approximant based on the so-termed random features (RFs), as outlined next.

A. RF-based EGP parametric model

Capitalizing on a standardized shift-invariant kernel $\bar{\kappa}(\mathbf{a}, \mathbf{a}') = \bar{\kappa}(\mathbf{a} - \mathbf{a}')$ with $\bar{\kappa} = \kappa/\sigma_\theta^2$, RF-based approximation begins with expressing $\bar{\kappa}$ as the the inverse Fourier transform of a spectral density $\pi_{\bar{\kappa}}(\boldsymbol{\zeta})$ as [32]

$$\bar{\kappa}(\mathbf{a} - \mathbf{a}') = \int \pi_{\bar{\kappa}}(\boldsymbol{\zeta}) e^{j\boldsymbol{\zeta}^\top(\mathbf{a}-\mathbf{a}')} d\boldsymbol{\zeta} = \mathbb{E}_{\pi_{\bar{\kappa}}}\left[e^{j\boldsymbol{\zeta}^\top(\mathbf{a}-\mathbf{a}')}\right] \tag{7}$$

where $\pi_{\bar{\kappa}}(\boldsymbol{\zeta})$ integrates to 1, so that it can be viewed as pdf. Since $\bar{\kappa}$ is real, the imaginary part of (7) vanishes and the last expectation equals $\mathbb{E}_{\pi_{\bar{\kappa}}}\left[\cos(\boldsymbol{\zeta}^\top(\mathbf{a} - \mathbf{a}'))\right]$. Upon drawing a sufficient number D of independent and identically distributed (i.i.d.) samples $\{\boldsymbol{\zeta}_j\}_{j=1}^D$ from $\pi_{\bar{\kappa}}(\boldsymbol{\zeta})$, an estimate of $\bar{\kappa}$ is

$$\check{\bar{\kappa}}(\mathbf{a}, \mathbf{a}') := \frac{1}{D} \sum_{j=1}^{D} \cos\left(\boldsymbol{\zeta}_j^\top(\mathbf{a} - \mathbf{a}')\right) . \tag{8}$$

Let us now define the $2D \times 1$ RF vector [11]

$$\boldsymbol{\phi}_{\boldsymbol{\zeta}}(\mathbf{a}) := \frac{1}{\sqrt{D}}\left[\sin(\boldsymbol{\zeta}_1^\top \mathbf{a}), \cos(\boldsymbol{\zeta}_1^\top \mathbf{a}), \ldots, \sin(\boldsymbol{\zeta}_D^\top \mathbf{a}), \cos(\boldsymbol{\zeta}_D^\top \mathbf{a})\right]^\top \tag{9}$$

which can be used to express $\check{\bar{\kappa}}$ as $\check{\bar{\kappa}}(\mathbf{a}, \mathbf{a}') = \boldsymbol{\phi}_{\boldsymbol{\zeta}}^\top(\mathbf{a})\boldsymbol{\phi}_{\boldsymbol{\zeta}}(\mathbf{a}')$ that yields the *parametric linear* function approximant

$$\check{f}(\mathbf{a}) = \boldsymbol{\phi}_{\boldsymbol{\zeta}}^\top(\mathbf{a})\boldsymbol{\theta}, \quad \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}_{2D}, \sigma_\theta^2 \mathbf{I}_{2D}) . \tag{10}$$

This parametric model enables the propagation of the posterior $p(\boldsymbol{\theta}|\mathbf{y}_t; \mathbf{A}_t) = \mathcal{N}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_t, \boldsymbol{\Sigma}_t)$ per slot t using a recursive Bayesian iteration. Besides the online posterior update, the RF-based parametric model encourages applications where privacy needs to be preserved, because it does not require direct access to each node's one-hop connectivity vector $\mathbf{a}$, but relies on the RF vector in (9), which can be seen as an encrypted version of $\mathbf{a}$ because of its co-sinusoidals.

Having established the well-documented merits of the RF-based parametric model, next we will show how the RF-based EGP model is updated as new data arrive on-the-fly.

Algorithm 1 GradEGP-MultiAFs

Initialization: $\mathcal{L}_0$, $\mathcal{U}_0$, $\mathcal{V}$, $\mathcal{K}$;
$\boldsymbol{\omega}_0 = \frac{1}{K}[1, \ldots, 1]^\top$;
for $t = 0, 1, \ldots, T$ **do**
  Obtain EGP $\boldsymbol{\Xi}_t$ based on $\mathcal{L}_t$ using (16)-(18);
  for $k = 1, \ldots, K$ **do**
    Obtain instance $\tilde{\mathbf{a}}_{n_{t+1}}^k \in \mathcal{U}_t$ by (24);
    Obtain 'pseudo-label' $\tilde{y}_{t+1}^k$ by (25) using $\boldsymbol{\Xi}_t$;
    Obtain $\tilde{\boldsymbol{\Xi}}_{t+1}^k$ utilizing pseudo pair $\{\tilde{\mathbf{a}}_{n_{t+1}}^k, \tilde{y}_{t+1}^k\}$;
    Obtain error $\epsilon_{t+1}^{v,k}$ on $\mathcal{E}$ via (27);
  end for
  Update per AF weight using (29);
  Obtain $\mathbf{a}_{n_{t+1}} \in \mathcal{U}_t$ of node n_{t+1} via (30);
  Obtain label y_{t+1} upon querying the oracle;
  $\mathcal{L}_{t+1} = \{\mathbf{a}_n, y_n, n \in \mathcal{S}_{t+1}\}$, $\mathcal{S}_{t+1} = \mathcal{S}_t \cup \{n_{t+1}\}$;
  $\mathcal{U}_{t+1} := \{\mathbf{a}_n, n \in \bar{\mathcal{S}}_{t+1}\}$, $\bar{\mathcal{S}}_{t+1} = \bar{\mathcal{S}}_t \setminus \{n_{t+1}\}$;
end for

B. EGP model online updates

With the dictionary $\mathcal{K}$ consisting of distinct shift-invariant kernels, each GP expert $m \in \mathcal{M}$ constructs its RF vector $\boldsymbol{\phi}_{\boldsymbol{\zeta}}^m(\mathbf{a})$ upon drawing i.i.d. vectors $\{\boldsymbol{\zeta}_j^m\}_{j=1}^D$ from the power spectral density $\pi_{\bar{\kappa}}^m(\boldsymbol{\zeta})$ of the standardized kernel $\bar{\kappa}^m$ with $\bar{\kappa}^m = \kappa^m/\sigma_{\theta^m}^2$. Then, the generative parametric model describing the sought function f and the (possibly noisy) output y for expert m at slot t, is

$$\begin{aligned} p(\boldsymbol{\theta}^m) &= \mathcal{N}(\boldsymbol{\theta}^m; \mathbf{0}_{2D}, \sigma_{\theta^m}^2 \mathbf{I}_{2D}) \\ p(f(\mathbf{a})|i=m, \boldsymbol{\theta}^m) &= \delta(f(\mathbf{a}) - \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{a}_{n_\tau})\boldsymbol{\theta}^m) \\ p(y|\boldsymbol{\theta}^m, \mathbf{a}) &= \mathcal{N}(y; \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{a})\boldsymbol{\theta}^m, \sigma_n^2) . \end{aligned} \tag{11}$$

This generative model allows expert m to summarize all labeled samples $\mathcal{L}_t$ in the posterior pdf $p(\boldsymbol{\theta}^m|\mathcal{L}_t) = \mathcal{N}(\boldsymbol{\theta}^m; \hat{\boldsymbol{\theta}}_t^m, \boldsymbol{\Sigma}_t^m)$. Further accounting for per-expert weights, the RF-based EGP model updates the parameter set per slot t

$$\boldsymbol{\Xi}_t := \{w_t^m, \hat{\boldsymbol{\theta}}_t^m, \boldsymbol{\Sigma}_t^m, m \in \mathcal{M}\} . \tag{12}$$

Next, we will show how $\boldsymbol{\Xi}_t$ can form the predictive pdf, which will be used in the next section to design the AF, and how $\boldsymbol{\Xi}_t$ can be updated with the newly acquired pair.

RF-based EGP predictive pdf. Each expert m capitalizes on its posterior $p(\boldsymbol{\theta}^m|\mathcal{L}_t)$ to form the predictive pdf as

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

with

$$\hat{y}_{t+1|t}^m(\mathbf{a}) = \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{a})\hat{\boldsymbol{\theta}}_t^m \tag{13a}$$

$$(\sigma_{t+1|t}^m(\mathbf{a}))^2 = \boldsymbol{\phi}_{\boldsymbol{\zeta}}^{m\top}(\mathbf{a})\boldsymbol{\Sigma}_t^m \boldsymbol{\phi}_{\boldsymbol{\zeta}}^m(\mathbf{a}) + \sigma_n^2 . \tag{13b}$$

The EGP meta-learner combines the predictive pdfs of all M experts with the properly adjusted weights w_t^m to form its ensemble version, which is a GM given by

$$p(y_{t+1}|\mathcal{L}_t,\mathbf{a}) = \sum_{m=1}^{M} p(y_{t+1}|i=m,\mathcal{L}_t,\mathbf{a})p(i=m|\mathcal{L}_t)$$
$$= \sum_{m=1}^{M} w_t^m \mathcal{N}(y_{t+1};\hat{y}_{t+1|t}^m(\mathbf{a}),(\sigma_{t+1|t}^m(\mathbf{a}))^2)\ . \quad (14)$$

Considering the minimum mean-square error (MMSE) estimator of y_{t+1} yields the ensemble predictor along with the corresponding variance, which are available in closed form as

$$\hat{y}_{t+1|t}(\mathbf{a}) = \sum_{m=1}^{M} w_t^m \hat{y}_{t+1|t}^m(\mathbf{a}) \quad (15a)$$

$$\sigma_{t+1|t}^2(\mathbf{a}) = \sum_{m=1}^{M} w_t^m[(\sigma_{t+1|t}^m(\mathbf{a}))^2+(\hat{y}_{t+1|t}(\mathbf{a})-\hat{y}_{t+1|t}^m(\mathbf{a}))^2] \quad (15b)$$

where "$t+1|t$" signifies that only the model parameters and nodal observation of the previous slot t are involved in predicting y_{t+1}.

RF-based EGP model update. Based on the RF-based EGP predictive pdf (14), one can obtain the next query node n_{t+1} with input vector $\mathbf{a}_{n_{t+1}}$ by maximizing the AFs as elaborated in the next section. Upon evaluating $\mathbf{a}_{n_{t+1}}$ to obtain the label y_{t+1}, the posterior pdf of $\boldsymbol{\theta}^m$ is propagated as

$$p(\boldsymbol{\theta}^m|\mathcal{L}_{t+1}) = \frac{p(\boldsymbol{\theta}^m|\mathcal{L}_t)p(y_{t+1}|\boldsymbol{\theta}^m,\mathbf{a}_{n_{t+1}})}{p(y_{t+1}|\mathbf{a}_{n_{t+1}},i=m,\mathcal{L}_t)}$$
$$= \mathcal{N}(\boldsymbol{\theta}^m;\hat{\boldsymbol{\theta}}_{t+1}^m,\boldsymbol{\Sigma}_{t+1}^m) \quad (16)$$

with the mean $\hat{\boldsymbol{\theta}}_{t+1}^m$ and covariance matrix $\boldsymbol{\Sigma}_{t+1}^m$ given by

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

Meanwhile, each expert m updates the corresponding weight $w_{t+1}^m := \Pr(i=m|\mathcal{L}_{t+1})$ by applying Bayes' rule as

$$w_{t+1}^m = \frac{\Pr(i=m|\mathcal{L}_t)p(y_{t+1}|\mathbf{a}_{n_{t+1}},i=m,\mathcal{L}_t)}{p(y_{t+1}|\mathbf{a}_{n_{t+1}},\mathcal{L}_t)}$$
$$= \frac{w_t^m \mathcal{N}\left(y_{t+1};\hat{y}_{t+1|t}^m,(\sigma_{t+1|t}^m)^2\right)}{\sum_{m'=1}^{M} w_t^{m'}\mathcal{N}\left(y_{t+1};\hat{y}_{t+1|t}^{m'},(\sigma_{t+1|t}^{m'})^2\right)}\ . \quad (18)$$

V. EGP-based acquisition criteria

Building on the ensemble predictive pdf in (14), this section introduces a suite of intuitive AFs whose maximizers yield the next node to be sampled, based on different criteria.

A. Query-by-Committee (QBC)

The first AF hinges on the notion of 'disagreement' adopted in the so-termed 'QBC' acquisition criterion, which has been applied to both classification and regression tasks; see e.g., [33], [10] and [3]. With the M GP experts viewed as members of a committee, the EGP-based QBC rule is (cf. (13a))

$$\alpha^{\mathrm{QBC}}(\mathbf{a};\mathcal{L}_t) := \sum_{m=1}^{M} w_t^m(\hat{y}_{t+1|t}^m(\mathbf{a})-\hat{y}_{t+1|t}(\mathbf{a}))^2 \quad (19)$$

where $\hat{y}_{t+1|t}(\mathbf{a})$ in (15a) represents the consensus of the committee. Different from the standard QBC rule that has equal weights among committee members, the weights in both (19) and (15a) are different across m. Albeit interesting, this approach takes into account only the per-expert predictive mean in (13a) and disregards the predictive variance in (13b) that quantifies the associated uncertainty.

B. Weighted variance

Accounting for the uncertainty offered by each expert's predictive variance, the next AF combines the variances of all M experts with the properly adjusted weights w_t^m as follows

$$\alpha^{\mathrm{wVar}}(\mathbf{a};\mathcal{L}_t) := \sum_{m=1}^{M} w_t^m(\sigma_{t+1|t}^m(\mathbf{a}))^2\ . \quad (20)$$

Although intuitively simple, this AF does not leverage the valuable information provided by each expert's predictive mean in (13a).

C. Variance of GP mixture

Combining the merits of the last two AFs, one can directly build on the GM in (14) whose variance yields the AF

$$\alpha^{\mathrm{GPM-Var}}(\mathbf{a};\mathcal{L}_t) = \sigma_{t+1|t}^2(\mathbf{a}) \quad (21)$$

which is the sum of (19) and (20).

D. Weighted entropy

An alternative measure of uncertainty is provided by the entropy whose maximization is tantamount to maximizing the variance in the Gaussian pdf case. However, this does not apply in the GM of (14). Similar to the 'weighted variance' AF, one can consider instead a weighted combination of all experts' entropy as

$$\alpha^{\mathrm{wEnt}}(\mathbf{a};\mathcal{L}_t) := \frac{1}{2}\sum_{m=1}^{M} w_t^m \ln(2\pi((\sigma_{t+1|t}^m(\mathbf{a}))^2)\ . \quad (22)$$

E. Entropy of GP mixtures

Further accounting for each expert's predictive mean (cf. (13a)) besides the predictive variance (cf. (13b)), and allowing for interaction among GP experts, one can rely on the entropy of the GP mixture (cf. (14)). Although this is not available in closed form, one can fortunately exploit its analytic lower bound, which is given by [7]

$$-\sum_{m=1}^{M} w_t^m \int \mathcal{N}(y_{t+1}(\mathbf{a});\hat{y}_{t+1|t}^m(\mathbf{a}),(\sigma_{t+1|t}^m(\mathbf{a}))^2)$$
$$\times \log p(y_{t+1}(\mathbf{a})|\mathcal{L}_t)dy_{t+1}(\mathbf{a})$$
$$\overset{(a)}{\geq} -\sum_{m=1}^{M} w_t^m \log\Big(\int \mathcal{N}(y_{t+1}(\mathbf{a});\hat{y}_{t+1|t}^m(\mathbf{a}),(\sigma_{t+1|t}^m(\mathbf{a}))^2)$$
$$\times \log p(y_{t+1}(\mathbf{a})|\mathcal{L}_t)dy_{t+1}(\mathbf{a})\Big)$$

where (a) comes from Jensen's inequality. Since the term inside the logarithm is in analytic form, the last AF is

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

where $\psi^{m,m'}$ models the interaction of any two distinct GP models as

$$\begin{aligned}\psi_t^{m,m'} &:= \int \mathcal{N}(y_{t+1}(\mathbf{a});\hat{y}_{t+1|t}^m(\mathbf{a}),(\sigma_{t+1|t}^m(\mathbf{a}))^2) \\ &\quad\times \mathcal{N}(y_{t+1}(\mathbf{a});\hat{y}_{t+1|t}^{m'}(\mathbf{a}),(\sigma_{t+1|t}^{m'}(\mathbf{a}))^2)dy_{t+1}(\mathbf{a}) \\ &= \mathcal{N}(\hat{y}_{t+1|t}^m(\mathbf{a});\hat{y}_{t+1|t}^{m'}(\mathbf{a}),(\sigma_{t+1|t}^m(\mathbf{a}))^2+(\sigma_{t+1|t}^{m'}(\mathbf{a}))^2)\ .\end{aligned}$$

F. Ensembling EGP-based AFs

So far, we have devised a suite of novel AFs that the advocated graph-adaptive EGP model (abbreviated as 'GradEGP') can employ to sample an unlabeled node for query. Based on the relative Bayesian optimization context, it is shown that there does not exist a single AF that thrives in all different tasks [6]. Inspired by this observation that also applies in the AL context, it is intuitive that a proper combination of candidate AFs may exhibit robustness and improved performance. Similar to the GradEGP model, we assign each AF $k \in \{1\ldots K\}$ a weight $\omega_t^k \in [0,1]$ with $\sum_{k=1}^K \omega_t^k = 1$ so that ω_t^k can be thought of as probability measuring the significance of each expert. To properly adjust these weights, we rely on a validation set $\mathcal{E} := \{(\mathbf{a}_{n_\tau^v}, y_\tau^v)\}_{\tau=1}^V$ with $\{\mathbf{a}_{n_\tau^v}\}_{\tau=1}^V$ denoting the connectivity vectors of the observed nodes in $\mathcal{E}_n := \{n_\tau^v\}_{\tau=1}^V$, to assess the performance of different AFs.

Relying on the labeled set $\mathcal{L}_t$ at slot t, the RF-based EGP parameter set $\boldsymbol{\Xi}_t$ in (12) is estimated, and then each AF k selects the next query node n_{t+1}^k with connectivity vector $\tilde{\mathbf{a}}_{n_{t+1}}^k$ by maximizing the associated criterion as

$$\tilde{\mathbf{a}}_{n_{t+1}}^k = \underset{\mathbf{a}\in\mathcal{U}_t}{\arg\max}\ \alpha^k(\mathbf{a};\mathcal{L}_t)\ . \tag{24}$$

After receiving $\tilde{\mathbf{a}}_{n_{t+1}}^k$, AF k utilizes the EGP parameter set $\boldsymbol{\Xi}_t$ to form a *'pseudo-label'* corresponding to $\tilde{\mathbf{a}}_{n_{t+1}}^k$ as

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

Capitalizing on the pair $\{\tilde{\mathbf{a}}_{n_{t+1}}^k, \tilde{y}_{t+1}^k\}$, expert k relies on (16) – (18) to update the EGP parameter set

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

Using $\tilde{\boldsymbol{\Xi}}_{t+1}^k$, the performance of AF k is then evaluated based on the prediction error on the validation set as

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

where the predicted label for node n_τ^v in the validation set is

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

With $\{\epsilon_{t+1}^{v,k}\}_k$ at hand, the per-AF weight is 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. The weight update formula in (29) belongs to the exponentiated weight update in online learning with expert advice; see e.g., [4].

The updated weights are subsequently used to eventually query the next node n_{t+1} by optimizing the weighted ensemble of AFs as

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

This novel 'GradEGP-MultiAFs' approach that combines various acquisition criteria on-the-fly, is summarized in Alg. 1.

VI. Numerical tests

In this section, the performance of the proposed EGP-based AFs will be tested in both synthetic and real graph datasets. Relying on the advocated GradEGP statistical model to learn the sought function, the acquisition rules to be assessed are the ones described in Sec. V A-F, which from now on will be abbreviated as GradEGP with "QBC," "wVar," "GPM-Var," "wEnt," "GPM-Ent," and "MultiAFs," respectively. All these approaches will be compared against the GradEGP model that randomly selects new nodes to sample, abbreviated as "GradEGP-random," and the single GP model baseline that employs the maximum variance (entropy) criterion.

For fairness in comparison, the set of initially labeled nodes in $\mathcal{L}_0$ is common to all competing approaches, while the kernel hyperparameters for all GP experts in GradEGP and the single GP baseline, are obtained by maximizing the marginal likelihood. For all RF-based approaches, the number of RFs is $D = 50$. The kernel dictionary $\mathcal{K}$ comprises radial basis functions (RBFs) with lengthscales $\{10^c\}_{c=-4}^6$. Regarding the "GradEGP-MultiAFs" approach, each $\alpha^k(\mathbf{a};\mathcal{L}_t)$ in (30) is divided by its maximum value to range between 0 and 1.

The performance of all approaches is evaluated on a held-out test set $\bar{\mathcal{O}} := (\mathbf{a}_{n_\tau^e}, y_\tau^e)_{\tau=1}^{T^e}$ of nodes $\{n_\tau^e\}_{\tau=1}^{T^e}$ (where ^{e} stands for evaluation). As figure of merit, the normalized mean-square error (NMSE) at each iteration t is reported for all approaches, which is given by

$$\mathrm{NMSE}_t := \frac{1}{T^e}\sum_{\tau=1}^{T^e}(\hat{y}_{\tau|t}^e - y_\tau^e)^2/||\mathbf{y}_T^e||_2^2$$

where $\mathbf{y}_T^e := [y_1^e\ldots y_T^e]^\top$. All approaches are tested over 10 realizations, whose sample average NMSE performance along with the corresponding standard deviation are reported.

Synthetic dataset. A synthetic graph is constructed with $N = 100$ nodes utilizing a stochastic block model consisting of $C = 10$ communities, as in e.g. [30]. The output per node is the eigenvector corresponding to the lowest nonzero eigenvalue of the graph Laplacian. The number of initially labeled nodes for AL is $|\mathcal{L}_0| = 10$; the size of the unlabeled set is $|\mathcal{U}_0| = 60$; and, the test set $\bar{\mathcal{O}}$ consists of 20 nodes.

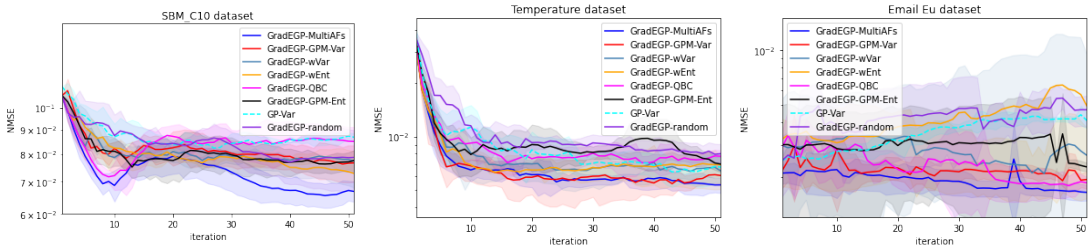

Fig. 1: NMSE performance on (a) "SBM_C10;" (b) "Temperature;" and (c) "Email Eu" and datasets.

Regarding the GradEGP-MultiAFs approach, the validation set comprises 10 nodes to evaluate each AF per iteration.

The NMSE performance of all competing approaches is depicted in Fig. 1a, where it is evident that all EGP-based approaches outperform the single GP-Var baseline, showcasing the merits of adopting an ensemble of GP learners with adaptive weights being properly adjusted as new data arrive on-the-fly. In addition, the superior performance of all (except one) GradEGP - based AL methods over the "GradEGP-random" approach demonstrates the benefits of the novel acquisition criteria. Further adopting a weighted ensemble of the candidate AFs in the "EGP-MultiAFs" approach significantly improves the prediction performance, with the latter being the best-performing approach in terms of NMSE.

Temperature dataset. This dataset comprises hourly temperature measurements offered by the National Climatic Data Center, at $N = 109$ measuring stations across the continental United States in 2010 [1]. A symmetric graph is constructed utilizing the geographic distances of these stations as in [17], [14], [16]. In the experimental setup, we choose $|\mathcal{L}_0| = 10$, $|\mathcal{U}_0| = 60$, $|\mathcal{E}| = 10$ and $|\bar{\mathcal{O}}| = 29$. As shown in Fig. 1b, all proposed approaches outperform the "GradEGP-random" baseline, corroborating the merits of adopting intuitive acquisition criteria to guide nodal sampling. Although the GP-Var baseline outperforms three out of the five advocated GradEGP-based single AF approaches, "GradEGP-MultiAFs" exhibits the lowest NMSE by properly combining the merits of all AFs using appropriate adaptive weights per iteration.

Email Eu dataset. In this dataset, a graph is constructed using email data from $N = 1,005$ individuals affiliated with a large European research institute. An edge (n, n') is nonzero only if person n sent person n' at least one email [12]. The sought nodal values are the ground-truth community memberships of the nodes, which are real with analog-amplitude as in a regression task; see e.g., [25]. For the experimental evaluation, we consider $|\mathcal{L}_0| = 50$, $|\mathcal{U}_0| = 700$, $|\mathcal{E}| = 50$, and $|\bar{\mathcal{O}}| = 205$. It can be clearly seen that all proposed approaches except "GradEGP-wEntr" significantly outperform the "GP-Var" and "GradEGP-random" baselines, with "GradEGP-MultiAFs" consistently performing best (except one iteration). Hence, utilizing an EGP model and combining a suite of intuitively effective criteria in an adaptive manner, it is possible to improve the prediction performance in AL settings, where the size of the unlabeled set is large.

VII. Conclusions

This contribution dealt with active node sampling for graph-based SSL. With a per-node one-hop connectivity vector as input to an EGP model, an incremental learning approach was developed to learn the graph function mapping adaptively. Building on this so-termed "GradEGP" model, a suite of novel AFs were devised to sequentially select unlabeled nodes based on different rules. Further combining the advocated single AFs on the fly with properly adaptive weights, yields a novel GradEGP-based ensemble acquisition approach. Tests on both synthetic and real graph datasets showcase the merits of the proposed methods relative to conventional passive sampling.

References

[1] "1981-2010 U.S. climate normals," https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/climate-normals/1981-2010-normals-data, [Online; accessed 29-April-2019].

[2] D. Berberidis and G. B. Giannakis, "Data-adaptive active sampling for efficient graph-cognizant classification," *IEEE Trans. Sig. Process.*, vol. 66, no. 19, pp. 5167–5179, 2018.

[3] R. Burbidge, J. J. Rowland, and R. D. King, "Active learning for regression based on query by committee," in *Int. Conf. Intelligent Data Engineering and Automated Learning*. Springer, 2007, pp. 209–218.

[4] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[5] G. B. Giannakis, Y. Shen, and G. V. Karanikolas, "Topology identification and learning over graphs: Accounting for nonlinearities and dynamics," *Proc. of the IEEE*, vol. 106, no. 5, pp. 787–807, May 2018.

[6] M. Hoffman, E. Brochu, N. de Freitas *et al.*, "Portfolio allocation for Bayesian optimization." *Proc. of Uncertainty in Artificial Intelligence*, pp. 327–336, 2011.

[7] M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck, "On entropy approximation for Gaussian mixture random vectors," in *IEEE Intl. Conf. on Multisensor Fusion and Integ. for Intell. Syst.*, 2008, pp. 181–188.

[8] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, "Active learning with Gaussian processes for object categorization," *Proc. Intl. Conf. Comp. Vision*, 2007.

[9] A. Krause and C. Guestrin, "Nonmyopic active learning of gaussian processes: an exploration-exploitation approach," in *Proc. Int. Conf. Mach. Learn.*, 2007, p. 449–456.

[10] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," *Proc. Advances Neural Inf. Process. Syst.*, vol. 7, 1994.

[11] M. Lázaro-Gredilla, J. Quiñonero Candela, C. E. Rasmussen, and A. Figueiras-Vidal, "Sparse spectrum Gaussian process regression," *J. Mach. Learn. Res.*, vol. 11, no. Jun, pp. 1865–1881, 2010.

[12] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. on Knowledge Discovery from Data*, vol. 1, no. 1, March 2007.

[13] Q. Lu and G. B. Giannakis, "Gaussian process temporal-difference learning with scalability and worst-case performance guarantees," *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Process.*, pp. 3485–3489, 2021.

[14] ——, "Probabilistic reconstruction of spatio-temporal processes over multi-relational graphs," *IEEE Trans. Sig. and Info. Process. over Net.*, vol. 7, pp. 166–176, 2021.
[15] ——, "Robust and adaptive temporal-difference learning using an ensemble of Gaussian processes," *arXiv preprint arXiv:2112.00882*, 2021.
[16] Q. Lu, V. Ioannidis, and G. B. Giannakis, "Semi-supervised tracking of dynamic processes over switching graphs," in *Proc. of IEEE Data Science Workshop*, Minneapolis, MN, Jun. 2019.
[17] Q. Lu, V. N. Ioannidis, and G. B. Giannakis, "Graph-adaptive semi-supervised tracking of dynamic processes over switching network modes," *IEEE Trans. Sig. Process.*, vol. 68, pp. 2586–2597, 2020.
[18] Q. Lu, G. Karanikolas, Y. Shen, and G. B. Giannakis, "Ensemble Gaussian processes with spectral features for online interactive learning with scalability," *Proc. Int. Conf. Artificial Intel. and Stats.*, June 2020.
[19] Q. Lu, G. V. Karanikolas, and G. B. Giannakis, "Incremental ensemble Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intel.*, 2022.
[20] Q. Lu, K. D. Polyzos, B. Li, and G. B. Giannakis, "Surrogate modeling for bayesian optimization beyond a single Gaussian process," *arXiv preprint arXiv:2205.14090*, 2022.
[21] Y. Ma, R. Garnett, and J. Schneider, "σ-optimality for active learning on gaussian random fields," *Proc. Advances Neural Inf. Process. Syst.*, vol. 26, 2013.
[22] D. J. MacKay, "Information-based objective functions for active data selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992.
[23] K. Miller, H. Li, and A. L. Bertozzi, "Efficient graph-based active learning with probit likelihood via gaussian approximations," *arXiv preprint arXiv:2007.11126*, 2020.
[24] Y. C. Ng, N. Colombo, and R. Silva, "Bayesian semi-supervised learning with graph gaussian processes," *Proc. Advances Neural Inf. Process. Syst.*, vol. 31, 2018.
[25] K. D. Polyzos, Q. Lu, and G. B. Giannakis, "Ensemble Gaussian processes for online learning over graphs with adaptivity and scalability," *IEEE Trans. Sig. Process.*, 2021.
[26] ——, "Graph-adaptive incremental learning using an ensemble of Gaussian process experts," *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Process.*, June 2021.
[27] ——, "Online graph-guided inference using ensemble gaussian processes of egonet features," in *Proc. Asilomar Conf. Sig., Syst., Comput.*, 2021, pp. 182–186.
[28] ——, "Weighted ensembles for active learning with adaptivity," *arXiv preprint arXiv:2206.05009*, 2022.
[29] K. D. Polyzos, Q. Lu, A. Sadeghi, and G. B. Giannakis, "On-policy reinforcement learning via ensemble Gaussian processes with application to resource allocation," *Proc. Asilomar Conf. Sig., Syst., Comput.*, pp. 1018–1022, 2021.
[30] K. D. Polyzos, C. Mavromatis, V. N. Ioannidis, and G. B. Giannakis, "Unveiling anomalous edges and nominal connectivity of attributed networks," *Proc. Asilomar Conf. Sig., Syst., Comput.*, Nov. 2020.
[31] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.
[32] W. Rudin, *Principles of Mathematical Analysis*. McGraw-hill New York, 1964, vol. 3.
[33] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. of the Workshop on Computational Learning Theory*, 1992, pp. 287–294.
[34] Y. Shen, G. Leus, and G. B. Giannakis, "Online graph-adaptive learning with scalability and privacy," *IEEE Trans. Sig. Process.*, vol. 67, no. 9, May 2019.
[35] J. Zhou and S. Sun, "Active learning of gaussian processes with manifold-preserving graph reduction," *Neural Computing and Applications*, vol. 25, no. 7, pp. 1615–1625, 2014.