Ensemble Gaussian Processes with Spectral Features for Online Interactive Learning with Scalability

Qin Lu[†] **Georgios V. Karanikolas**[†] **Yanning Shen**[‡] **Georgios B. Giannakis**[†]
[†]Dept. of ECE and DTC, University of Minnesota, Minneapolis, MN 55455
[‡]Dept. of EECS, University of California, Irvine, CA 92697

Abstract

Combining benefits of kernels with Bayesian models, Gaussian process (GP) based approaches have well-documented merits not only in learning over a rich class of nonlinear functions, but also quantifying the associated uncertainty. While most GP approaches rely on a single preselected prior, the present work employs a weighted ensemble of GP priors, each having a unique covariance (kernel) belonging to a prescribed kernel dictionary – which leads to a richer space of learning functions. Leveraging kernel approximants formed by spectral features for scalability, an online interactive ensemble (OI-E) GP framework is developed to jointly learn the sought function, and for the first time select interactively the EGP kernel on-the-fly. Performance of OI-EGP is benchmarked by the best fixed function estimator via regret analysis. Furthermore, the novel OI-EGP is adapted to accommodate dynamic learning functions. Synthetic and real data tests demonstrate the effectiveness of the proposed schemes.

1 Introduction

Gaussian processes (GPs) cross-fertilize merits of kernel methods and Bayesian models to benefit several learning tasks, including regression, classification, ranking, and dimensionality reduction (Rasmussen and Williams, 2006). In GP-based approaches, a Gaussian *prior* is assumed over a learning function $f(\cdot)$ with covariance (kernel) capturing similarities among $\{f(\mathbf{x}_t)\}$ dependent on inputs $\{\mathbf{x}_t\}$. Given observed outputs $\{y_t\}$ linked to the latent function $f(\cdot)$ via the conditionally independent per-datum likelihood $p(y_t|f(\mathbf{x}_t))$, Bayes rule produces the *posterior* distribution of $f(\cdot)$, based on which task-specific inference can be effected on the unseen data. Besides expressiveness of nonlinear functions that is also offered by deterministic kernel methods (Schölkopf et al., 2002), the Bayesian framework of GP-based approaches further quantifies uncertainty of the function estimate.

Past works. In spite of their documented merits, applicability of GPs is severely limited by the cubic complexity in the number T of the training samples. (Rasmussen and Williams, 2006). A scalable approach to overcome this hurdle is to summarize the training set via $m \ll T$ pseudo data that are employed for inference in the testing phase (Snelson and Ghahramani, 2006; Quiñonero-Candela and Rasmussen, 2005; Titsias, 2009). This global summary amounts to approximating the original GP prior with a kernel matrix having low rank m, thus reducing the complexity of batch computations to $\mathcal{O}(Tm^2)$. Another less explored approach to effect GP scalability leverages spectral components of shift-invariant kernels, leading to random feature (RF) based kernel approximation. Approximating the nonparametric GP prior by a parametric one, the resultant RF-based GP approaches can afford complexity comparable to the aforementioned low-rank approximants (Lázaro-Gredilla et al., 2010; Gal and Turner, 2015). While ensuring scalability, all these approaches rely on a *single* preselected GP kernel, which limits expressiveness of the sought learning function.

An *ensemble* of (*local* or distributed) GP experts, each relying on all or a subset of training samples, has been also considered to broaden the range of the function space (Rasmussen and Ghahramani, 2002; Tresp, 2001; Meeds and Osindero, 2006; Deisenroth and Ng, 2015). Each local GP utilizes a unique kernel to make predictions based on reduced-size local data, which not only lowers complexity but also accounts for nonstationarity of the learning function. Further, efforts have been

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

made to combine global approximants with local GPs; see, e.g., (Snelson and Ghahramani, 2007; Yuan and Neubauer, 2009). Unfortunately, these are *batch* approaches that can neither deal with time-critical applications welcoming online decision, nor handle massive data sets requiring enormous storage.

On the other hand, *online* GP-based approaches have been developed based on online variational inference or stochastic optimization (Bui et al., 2017; Cheng and Boots, 2016); but in both cases only for a single GP. Besides missing the online ensemble (E) GPs and their performance, there is an additional challenge facing online GP learners, namely that the data can be chosen adversarially, thus motivating regret-based performance analysis (Kakade and Ng, 2005). *Regret analysis* of online approaches has been carried out but only for a single GP (Kakade et al., 2006; Seeger et al., 2008; Nguyen et al., 2017). In a nutshell, scalable, online EGP approaches and their regret-based performance remain an uncharted research territory.

Parallel to the probabilistic GP-based learning, online scalable learning has been pursued also in the *deterministic* reproducing kernel Hilbert space (RKHS) setups both for a single (Wang et al., 2012; Lu et al., 2016) as well as for an ensemble of learners (Jin et al., 2010; Shen et al., 2019); see also (Micchelli and Pontil, 2005; Alvarez et al., 2012) for batch multi-kernel learning. Most recently, online RF-based approaches based on an ensemble of RKHS learners have been reported along with their regret-based performance for static and dynamic settings (Shen et al., 2019).

Contributions. Relative to prior GP and RHKS based approaches, the contributions of the present paper can be highlighted as follows.

c1) An online interactive (OI) approach is developed based on a weighted ensemble of GP (EGP) learners with scalable RF-based kernel approximations. The novel scheme learns an unknown stationary function and jointly adapts to the appropriate EGP kernel on-the-fly.

c2) To account for data being adversarially chosen in the online setting, the performance of the resultant algorithm is benchmarked by the best fixed function approximant with data in hindsight via static regret analysis. With $\mathcal{O}(\log T)$ regret over T slots, OI-EGP incurs no regret on average.

c3) With regards to online RF- and RKHS-based approaches (Shen et al., 2019), the proposed probabilistic GP-based methods offer extra uncertainty quantification of the sought function estimates. In addition, thanks to the second-order update, the regret bound is tighter than that of the first-order approach (Shen et al., 2019) in the static setting.

c4) Tracking of time-varying functions is enabled using a random walk (generally state space) model to capture the dynamics of the RF-based parameters, yielding a dynamic (D) OI-EGP approach.

Notation. Scalars are denoted by lowercase, column vectors by bold lowercase, and matrices by bold uppercase fonts. Superscripts $^\top$ and $^{-1}$ denote transpose, and matrix inverse, respectively; while $\mathbf{0}_N$ stands for the $N\times 1$ all-zero vector; and $\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\mathbf{K})$ for the probability density function (pdf) of a Gaussian random vector $\mathbf{x}$ with mean $\boldsymbol{\mu}$, and covariance matrix $\mathbf{K}$. Subscript "$t+1|\mathbf{t}$" signifies that prediction for slot $t+1$ relies on the *batch* of samples up to and including t, while "$t+1|t$" stands for a single-step predictor.

2 Preliminaries and background

As a prelude to our online EGP approach that will also introduce context and notation, this section deals with batch and online learning based on a single GP.

2.1 Non-scalable batch GP-based learning

Suppose we wish to estimate an unknown function f, which has a GP prior, denoted as $f\sim\mathcal{GP}(0,\kappa(\mathbf{x},\mathbf{x}'))$, with $\kappa(\cdot,\cdot)$ being a kernel measuring pairwise similarity of $d\times 1$ deterministic inputs $\mathbf{x}$ and $\mathbf{x}'$. For any number of inputs $\mathbf{X}_t:=[\mathbf{x}_1,\ldots,\mathbf{x}_t]$, the joint prior pdf of the function evaluations $\mathbf{f}_t:=[f(\mathbf{x}_1),\ldots,f(\mathbf{x}_t)]^\top$ is

$$p(\mathbf{f}_t;\mathbf{X}_t)=\mathcal{N}(\mathbf{f}_t;\mathbf{0}_t,\mathbf{K}_t)\quad\forall t \tag{1}$$

where $\mathbf{K}_t$ is a $t\times t$ covariance matrix with (i,j)th entry $[\mathbf{K}_t]_{ij}=\mathrm{cov}(f(\mathbf{x}_i),f(\mathbf{x}_j)):=\kappa(\mathbf{x}_i,\mathbf{x}_j)$.

To estimate f, we rely on the observed outputs $\mathbf{y}_t:=[y_1,\ldots,y_t]^\top$ that are linked with $\mathbf{f}_t$ via the conditional likelihood $p(\mathbf{y}_t|\mathbf{f}_t;\mathbf{X}_t)=\prod_{t'=1}^t p(y_{t'}|f(\mathbf{x}_{t'}))$ that is assumed known. Through Bayes rule, the latter will yield the posterior $p(\mathbf{f}_t|\mathbf{y}_t;\mathbf{X}_t)\propto p(\mathbf{f}_t;\mathbf{X}_t)p(\mathbf{y}_t|\mathbf{f}_t;\mathbf{X}_t)$. Function f pertains to either regression (analog amplitude y_t) or classification (y_t drawn from a finite alphabet). For Gaussian process regression (GPR) the conditional likelihood is assumed normal with mean $\mathbf{f}_t$ and covariance matrix $\tau\mathbf{I}_t$ as $p(\mathbf{y}_t|\mathbf{f}_t;\mathbf{X}_t)=\mathcal{N}(\mathbf{y}_t;\mathbf{f}_t,\tau\mathbf{I}_t)$, which along with the GP prior in (1) yields the Gaussian posterior $p(\mathbf{f}_t|\mathbf{y}_t;\mathbf{X}_t)$.

Prediction with a single GP. Given $\mathbf{X}_{t+1}$ and $\mathbf{y}_t$, we have from (1) that $p(f(\mathbf{x}_{t+1})|\mathbf{f}_t;\mathbf{X}_t)$ is Gaussian with known mean and covariance. Together with the known posterior $p(\mathbf{f}_t|\mathbf{y}_t;\mathbf{X}_t)$, the so-termed predictive pdf of $f(\mathbf{x}_{t+1})$ can be obtained as (Rasmussen and

Williams, 2006)

$$p(f(\mathbf{x}_{t+1})|\mathbf{y}_t;\mathbf{X}_t)=\int p(f(\mathbf{x}_{t+1})|\mathbf{f}_t;\mathbf{X}_t)p(\mathbf{f}_t|\mathbf{y}_t;\mathbf{X}_t)d\mathbf{f}_t \tag{2}$$

which is generally non-Gaussian if $p(\mathbf{f}_t|\mathbf{y}_t;\mathbf{X}_t)$ is non-Gaussian, and thus necessitates Monte Carlo (MC) sampling to estimate it. Alternatively, $p(\mathbf{f}_t|\mathbf{y}_t;\mathbf{X}_t)$ can be approximated by a Gaussian, which leads to a Gaussian $p(f(\mathbf{x}_{t+1})|\mathbf{y}_t;\mathbf{X}_t)$. Of course, $p(f(\mathbf{x}_{t+1})|\mathbf{y}_t;\mathbf{X}_t)$ is Gaussian for GPR, with its mean and covariance matrix available in closed form.

Using the pdf in (2) and the known $p(y_{t+1}|f(\mathbf{x}_{t+1}))$, it is also possible to find the predictive pdf of y_{t+1} as

$$\begin{aligned} &p(y_{t+1}|\mathbf{y}_t;\mathbf{X}_{t+1}) \\ &\quad=\int p(y_{t+1}|f(\mathbf{x}_{t+1}))p(f(\mathbf{x}_{t+1})|\mathbf{y}_t;\mathbf{X}_t)df(\mathbf{x}_{t+1}) \end{aligned} \tag{3}$$

which generally requires MC sampling or $p(f_{t+1}|\mathbf{y}_t;\mathbf{X}_{t+1})$ to be (at least approximately) Gaussian. Either way, (3) yields the data predictive pdf that fully quantifies the uncertainty of y_{t+1}. Clearly, the mean or the mode of $p(y_{t+1}|\mathbf{y}_t;\mathbf{X}_{t+1})$ provides readily a point prediction of y_{t+1}. In addition, its variance quantifies the uncertainty of this prediction.

Specifically for GPR, we have

$$p(y_{t+1}|\mathbf{y}_t;\mathbf{X}_{t+1})=\mathcal{N}(y_{t+1};\hat{y}_{t+1|\mathbf{t}},\sigma^2_{t+1|\mathbf{t}}) \tag{4}$$

where the mean (variance) yields the predictor (and its accuracy) as (Rasmussen and Williams, 2006)

$$\hat{y}_{t+1|\mathbf{t}}=\mathbf{k}_{t+1}^\top(\mathbf{K}_t+\tau\mathbf{I}_t)^{-1}\mathbf{y}_t \tag{5a}$$

$$\sigma^2_{t+1|\mathbf{t}}=\kappa(\mathbf{x}_{t+1},\mathbf{x}_{t+1})-\mathbf{k}_{t+1}^\top(\mathbf{K}_t+\tau\mathbf{I}_t)^{-1}\mathbf{k}_{t+1}+\tau \tag{5b}$$

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

Clearly, this GP predictor is not scalable, because the complexity of $\mathcal{O}(t^3)$ for inverting the $t\times t$ matrix in (5) will become prohibitively high as t grows.

2.2 Scalable RF learning with a single GP

Various attempts have been made to effect scalability in GP-based learning; see, e.g., (Quiñonero-Candela and Rasmussen, 2005; Titsias, 2009; Lázaro-Gredilla et al., 2010). Most existing approaches amount to summarizing the training data via a much smaller number (m) of pseudo data with inducing inputs, thereby obtaining a training-set-dependent low-rank approximant of $\mathbf{K}_t$ (Quiñonero-Candela and Rasmussen, 2005). Targeting a low-rank approximant that is not dependent on the training set, we rely here on a standardized shift-invariant $\bar{\kappa}(\cdot)$, whose inverse Fourier transform is

$$\begin{aligned} \bar{\kappa}(\mathbf{x},\mathbf{x}')=\bar{\kappa}(\mathbf{x}-\mathbf{x}')&=\int \pi_{\bar{\kappa}}(\mathbf{v})e^{j\mathbf{v}^\top(\mathbf{x}-\mathbf{x}')}d\mathbf{v} \\ &:=\mathbb{E}_{\pi_{\bar{\kappa}}}\left[e^{j\mathbf{v}^\top(\mathbf{x}-\mathbf{x}')}\right] \end{aligned} \tag{6}$$

where $\pi_{\bar{\kappa}}$ is the power spectral density (PSD), and the last equality follows after normalizing so that $\pi_{\bar{\kappa}}(\mathbf{v})$ integrates to 1, what allows one to view it as a pdf.

Upon drawing a sufficient number, say D, of independent and identically distributed (i.i.d.) samples (features) $\{\mathbf{v}_i\}_{i=1}^D$ from $\pi_{\bar{\kappa}}(\mathbf{v})$, the ensemble mean in (6) can be approximated by the sample average[1]

$$\check{\kappa}_c(\mathbf{x},\mathbf{x}'):=\frac{1}{D}\sum_{i=1}^D e^{j\mathbf{v}_i^\top(\mathbf{x}-\mathbf{x}')}. \tag{7}$$

Let us now define the real $2D\times 1$ random feature (RF) vector as (Lázaro-Gredilla et al., 2010)

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

which allows us to replace $\check{\kappa}_c$ in (7) with $\check{\kappa}(\mathbf{x},\mathbf{x}')=\boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x})\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}')$; and thus, the parametric approximant

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

can be viewed as coming from a realization of the Gaussian $\boldsymbol{\theta}$ combined with $\boldsymbol{\phi}_{\mathbf{v}}$ to yield the GP prior in (1) with $\kappa=\sigma_\theta^2\bar{\kappa}$, where σ_θ^2 is the magnitude of κ. Clearly, for any $\mathbf{X}_t$, the prior pdf of $\check{\mathbf{f}}_t$ is then

$$p(\check{\mathbf{f}}_t;\mathbf{X}_t)=\mathcal{N}(\check{\mathbf{f}}_t;\mathbf{0}_t,\check{\mathbf{K}}_t),\quad \check{\mathbf{K}}_t=\sigma_\theta^2\boldsymbol{\Phi}_t\boldsymbol{\Phi}_t^\top \tag{10}$$

where $\boldsymbol{\Phi}_t:=[\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}_1),\ldots,\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}_t)]^\top$, and $\check{\mathbf{K}}_t$ is then a low rank ($2D$) approximant of $\mathbf{K}_t$ in (1) for $t>2D$.

With the parametric form of $\check{f}(\mathbf{x})$ in (9), the likelihood $p(\mathbf{y}_t|\check{\mathbf{f}}_t;\mathbf{X}_t)$ is also parametrized by $\boldsymbol{\theta}$. This together with the Gaussian prior of $\boldsymbol{\theta}$ (cf. (9)), yields the posterior $p(\boldsymbol{\theta}|\mathbf{y}_t;\mathbf{X}_t)$, based on which we can predict f and y at each test input $\mathbf{x}$. Specifically, upon replacing $p(f(\mathbf{x}_{t+1})|\mathbf{f}_t;\mathbf{X}_t)$ and $p(\mathbf{f}_t|\mathbf{y}_t;\mathbf{X}_t)$ in (2) by $p(\check{f}(\mathbf{x}_{t+1})|\boldsymbol{\theta})=\delta(\check{f}(\mathbf{x}_{t+1})-\boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x}_{t+1})\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{y}_t;\mathbf{X}_t)$, respectively, we obtain the predictive pdf of the RF-based $\check{f}(\mathbf{x}_{t+1})$, which further leads to the predictive pdf of y_{t+1} in (3) after replacing $f(\mathbf{x}_{t+1})$ by $\check{f}(\mathbf{x}_{t+1})$. For GPR, the predictive pdf of y_{t+1} is

$$p(y_{t+1}|\mathbf{y}_t;\mathbf{X}_{t+1})=\mathcal{N}(y_{t+1};\hat{y}_{t+1|\mathbf{t}},\sigma^2_{t+1|\mathbf{t}}) \tag{11}$$

where

$$\hat{y}_{t+1|\mathbf{t}}=\boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x}_{t+1})\left(\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t+\frac{\tau}{\sigma_\theta^2}\mathbf{I}_{2D}\right)^{-1}\boldsymbol{\Phi}_t^\top\mathbf{y}_t \tag{12a}$$

$$\sigma^2_{t+1|\mathbf{t}}=\boldsymbol{\phi}_{\mathbf{v}}^\top(\mathbf{x}_{t+1})\left(\frac{\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t}{\tau}+\frac{\mathbf{I}_{2D}}{\sigma_\theta^2}\right)^{-1}\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}_{t+1})+\tau. \tag{12b}$$

This *batch* predictor incurs complexity $\mathcal{O}(t(2D)^2+(2D)^3)$, which is dominated by $\mathcal{O}(t(2D)^2)$ for $t\gg 2D$.

[1]Quantities with ˇ involve RF approximations.

This linear (in t) complexity is apparently much more affordable than the plain-vanilla GP predictor (5).

The RF-based function approximant $\check{f}$ easily accommodates *online* operation (Gijsberts and Metta, 2013), which is necessary in many time-critical applications, such as time series prediction (Richard et al., 2008), and robot localization (Xu et al., 2014). Our interest will be on *interactive learning*, where prediction of y_{t+1} is due upon receiving $\mathbf{x}_{t+1}$ at the beginning of slot $t+1$, and the pdf of $\boldsymbol{\theta}$ is then updated after receiving y_{t+1} at the end of slot $t+1$. Focusing on GPR for specificity, let $p(\boldsymbol{\theta}|\mathbf{y}_t;\mathbf{X}_t) = \mathcal{N}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_t,\boldsymbol{\Sigma}_t)$ be the posterior of $\boldsymbol{\theta}$ at slot t with mean $\hat{\boldsymbol{\theta}}_t$ and covariance matrix $\boldsymbol{\Sigma}_t$. Alternating between prediction and model update, online learning proceeds as follows.

s1. Find the predictive pdf $p(y_{t+1}|\mathbf{y}_t;\mathbf{X}_{t+1}) = \mathcal{N}(y_{t+1};\hat{y}_{t+1|t},\sigma^2_{t+1|t})$ with $\hat{y}_{t+1|t} = \boldsymbol{\phi}^\top_{\mathbf{v}}(\mathbf{x}_{t+1})\hat{\boldsymbol{\theta}}_t$ and $\sigma^2_{t+1|t} = \boldsymbol{\phi}^\top_{\mathbf{v}}(\mathbf{x}_{t+1})\boldsymbol{\Sigma}_t\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}_{t+1}) + \tau$; and

s2. Upon receiving y_{t+1}, propagate the posterior pdf of $\boldsymbol{\theta}$ using Bayes rule as

$$p(\boldsymbol{\theta}|\mathbf{y}_{t+1};\mathbf{X}_{t+1}) = \frac{p(\boldsymbol{\theta}|\mathbf{y}_t;\mathbf{X}_t)p(y_{t+1}|\boldsymbol{\theta};\mathbf{x}_{t+1})}{p(y_{t+1}|\mathbf{y}_t;\mathbf{X}_{t+1})} = \mathcal{N}(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}_{t+1},\boldsymbol{\Sigma}_{t+1}) \quad (13)$$

whose mean and covariance are given by

$$\hat{\boldsymbol{\theta}}_{t+1} = \hat{\boldsymbol{\theta}}_t + \sigma^{-2}_{t+1|t}\boldsymbol{\Sigma}_t\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}_{t+1})(y_{t+1} - \hat{y}_{t+1|t}) \quad (14a)$$

$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t - \sigma^{-2}_{t+1|t}\boldsymbol{\Sigma}_t\boldsymbol{\phi}_{\mathbf{v}}(\mathbf{x}_{t+1})\boldsymbol{\phi}^\top_{\mathbf{v}}(\mathbf{x}_{t+1})\boldsymbol{\Sigma}_t\,. \quad (14b)$$

Although the overall complexity over t slots is $\mathcal{O}(t(2D)^2)$, identical to its RF-based batch counterpart, online processing can significantly save data storage, which becomes prohibitive as t grows.

Next, we will broaden the scope of a single GP prior by an ensemble of GPs (EGP). Besides serving the role of a non-Gaussian prior, EGP will turn out to be scalable too, after adopting once again the RF approximation.

3 Online scalable ensemble GPs

While a scalable online approach is offered in the previous section, its performance hinges on a *preselected* kernel for the GP prior, which confines function space expressiveness. To alleviate this limitation and construct a richer function space, we employ an ensemble of GP experts (learners), each of which places a unique GP prior on f as $f|s \sim \mathcal{GP}(0,\kappa^s(\mathbf{x},\mathbf{x}'))$, where $s \in \mathcal{S} := \{1,\ldots,S\}$ is the expert index and κ^s is a shift-invariant kernel selected from a *known* kernel dictionary $\mathcal{K} := \{\kappa^1,\ldots,\kappa^S\}$. Here, $\mathcal{K}$ should be constructed as large as computational constraints allow, depending on resources and the learning task. Per expert s, the prior pdf of function values at $\mathbf{X}_t$ is

$$p(\mathbf{f}_t|s;\mathbf{X}_t) = \mathcal{N}(\mathbf{f}_t;\mathbf{0}_t,\mathbf{K}^s_t),\ [\mathbf{K}^s_t]_{ij} := \kappa^s(\mathbf{x}_i,\mathbf{x}_j) \quad (15)$$

The ensemble prior pdf of $\mathbf{f}_t$ over all GP experts is then given by the Gaussian mixture (GM)

$$p(\mathbf{f}_t;\mathbf{X}_t) = \sum_{s=1}^S w^s\mathcal{N}(\mathbf{f}_t;\mathbf{0}_t,\mathbf{K}^s_t)\,, \qquad \sum_{s=1}^S w^s = 1 \quad (16)$$

where the *unknown* weights $\{w^s\}_{s=1}^S$, viewed as probabilities of each GP expert to be present in the EGPs, are to be learned from data that arrive sequentially.

Seeking a scalable predictor, each expert s relies on the RF-based function approximant (9) with the per-expert parameter vector $\boldsymbol{\theta}^s$ and RF vector $\boldsymbol{\phi}^s_{\mathbf{v}}(\mathbf{x})$ constructed as in (8) using $\{\mathbf{v}^s_i\}_{i=1}^D$. Vectors $\{\mathbf{v}^s_i\}_{i=1}^D$ here are drawn i.i.d. from $\pi^s_{\bar{\kappa}}(\mathbf{v})$, which is the PSD of the standardized kernel $\bar{\kappa}^s$, relating to κ^s through the magnitude $\sigma^2_{\theta^s}$ as $\kappa^s = \sigma^2_{\theta^s}\bar{\kappa}^s$. The per expert s generative model for the function approximant $\check{f}$ is then

$$p(\boldsymbol{\theta}^s) = \mathcal{N}(\boldsymbol{\theta}^s;\mathbf{0}_{2D},\sigma^2_{\theta^s}\mathbf{I}_{2D}) \quad (17a)$$

$$p(\check{f}(\mathbf{x})|\boldsymbol{\theta}^s,s) = \delta(\check{f}(\mathbf{x}) - \boldsymbol{\phi}^{s\top}_{\mathbf{v}}(\mathbf{x})\boldsymbol{\theta}^s)\,. \quad (17b)$$

Still, our focus is on the online interactive (OI) setup, where the prediction of y_t is due upon receiving $\mathbf{x}_t$ at the beginning of slot t, and model parameters are updated after y_t arrives at the end of slot t. Each expert in OI-EGP predicts y_{t+1} based on the posterior $p(\boldsymbol{\theta}^s|\mathbf{y}_t,s;\mathbf{X}_t)$. To assess the per-expert contribution, we further rely on $w^s_t := \Pr(s|\mathbf{y}_t;\mathbf{X}_t)$, the posterior probability of expert s being active. We will see next that since the per-expert weights and posterior pdfs can be obtained sequentially, it will be possible to update the overall posterior $\{w^s_t, p(\boldsymbol{\theta}^s|\mathbf{y}_t,s;\mathbf{X}_t)\}_{s=1}^S$ from slot t to slot $t+1$, by proceeding in two steps, namely prediction and correction.

Algorithm 1 OI-EGP for GPR

Input: κ^s, $s = 1,\ldots,S$, and number of RFs D.
Initialization:
for s = 1, 2, ..., S **do**
  Draw D random vectors $\{\mathbf{v}^s_i\}_{i=1}^D$;
  $w^s_0 = 1/S$; $\hat{\boldsymbol{\theta}}^s_0 = \mathbf{0}_{2D}$; $\boldsymbol{\Sigma}^s_0 = \sigma^2_{\theta^s}\mathbf{I}_{2D}$;
end for

for t = 1, 2, ..., T **do**
  Receive datum $\mathbf{x}_t$;
  for s = 1, 2, ..., S **do**
    Construct RF $\boldsymbol{\phi}^s_{\mathbf{v}}(\mathbf{x}_t)$ via (8);
    Obtain per-expert pdf of y_t via (28);
    Update w^s_t via (30);
    Update per-expert pdf of $\boldsymbol{\theta}^s$ via (31);
  end for
end for

3.1 Prediction

Upon receiving $\mathbf{x}_{t+1}$, each expert s constructs the RF vector $\boldsymbol{\phi}_{\mathbf{v}}^s(\mathbf{x}_{t+1})$ using $\{\mathbf{v}_i^s\}_{i=1}^D$ as in (8). With $p(\boldsymbol{\theta}^s|\mathbf{y}_t, s; \mathbf{X}_t)$ available from slot t, the per-expert predictive pdf of $\check{f}(\mathbf{x}_{t+1})$ can be obtained by invoking Bayes rule and the total probability theorem (TPT)

$$p(\check{f}(\mathbf{x}_{t+1})|\mathbf{y}_t, s; \mathbf{X}_t) = \int p(\check{f}(\mathbf{x}_{t+1}), \boldsymbol{\theta}^s|\mathbf{y}_t, s; \mathbf{X}_t) d\boldsymbol{\theta}^s$$
$$= \int \delta(\check{f}(\mathbf{x}_{t+1}) - \boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_{t+1})\boldsymbol{\theta}^s) p(\boldsymbol{\theta}^s|\mathbf{y}_t, s; \mathbf{X}_t) d\boldsymbol{\theta}^s. \quad (18)$$

Consequently, the predictive pdf of y_{t+1} is

$$p(y_{t+1}|\mathbf{y}_t, s; \mathbf{X}_{t+1}) \quad (19)$$
$$= \int p(y_{t+1}|\check{f}(\mathbf{x}_{t+1})) p(\check{f}(\mathbf{x}_{t+1})|\mathbf{y}_t, s; \mathbf{X}_t) d\check{f}(\mathbf{x}_{t+1}) .$$

Leveraging again Bayes rule and the TPT leads to the ensemble predictive pdf

$$p(y_{t+1}|\mathbf{y}_t; \mathbf{X}_{t+1}) = \sum_{s=1}^S \Pr(s|\mathbf{y}_t; \mathbf{X}_t) p(y_{t+1}|\mathbf{y}_t, s; \mathbf{X}_{t+1})$$
$$= \sum_{s=1}^S w_t^s p(y_{t+1}|\mathbf{y}_t, s; \mathbf{X}_{t+1}) \quad (20)$$

which takes an intuitive form as a weighted ensemble of GP expert predictions from the previous slot. Having available the predictive pdf, we are ready to update the posterior pdf of the RF model parameter vector.

3.2 Correction

With the arrival of y_{t+1}, each expert s incurs the so-termed Bayesian loss at slot $t+1$ defined as (cf. (Kakade and Ng, 2005))

$$l_{t+1|t}^s := -\log p(y_{t+1}|\mathbf{y}_t, s; \mathbf{X}_{t+1}) . \quad (21)$$

For later use, its ensemble version is given by

$$\ell_{t+1|t} := -\log p(y_{t+1}|\mathbf{y}_t; \mathbf{X}_{t+1})$$
$$= -\log \sum_{s=1}^S w_t^s \exp\left(-l_{t+1|t}^s\right) . \quad (22)$$

Accordingly, the per-expert weight is updated as

$$w_{t+1}^s = \Pr(s|\mathbf{y}_{t+1}; \mathbf{X}_{t+1})$$
$$= \frac{\Pr(s|\mathbf{y}_t; \mathbf{X}_t) p(y_{t+1}|\mathbf{y}_t, s; \mathbf{X}_{t+1})}{p(y_{t+1}|\mathbf{y}_t; \mathbf{X}_{t+1})}$$
$$= w_t^s \exp(\ell_{t+1|t} - l_{t+1|t}^s) , \quad s = 1, \ldots, S \quad (23)$$

where w_t^s is available from slot t. Clearly, large $l_{t+1|t}^s$ implies small $\ell_{t+1|t} - l_{t+1|t}^s$, and thus w_{t+1}^s relative to the rest will be smaller than that at slot t.

At the same time, each expert s uses y_{t+1} to update the posterior pdf of $\boldsymbol{\theta}^s$ via Bayes rule as

$$p(\boldsymbol{\theta}^s|\mathbf{y}_{t+1}, s; \mathbf{X}_{t+1}) = \frac{p(\boldsymbol{\theta}^s|\mathbf{y}_t, s; \mathbf{X}_t) p(y_{t+1}|\boldsymbol{\theta}^s, s; \mathbf{x}_{t+1})}{p(y_{t+1}|\mathbf{y}_t, s; \mathbf{X}_{t+1})} \quad (24)$$

where $p(y_{t+1}|\boldsymbol{\theta}^s, s; \mathbf{x}_{t+1}) = p(y_{t+1}|\boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_{t+1})\boldsymbol{\theta}^s)$ is the known likelihood, $p(\boldsymbol{\theta}^s|\mathbf{y}_t, s; \mathbf{X}_t)$ is available from slot t, and $p(y_{t+1}|\mathbf{y}_t, s; \mathbf{X}_{t+1})$ is likewise known from (19).

Summarizing, our scalable OI-EGP algorithm for general likelihoods (and thus posteriors) relies on (18)-(24) to transition from slot t to slot $t+1$. The generally non-Gaussian pdfs and (possibly high-dimensional) integrals involved can be obtained using approximate inference techniques; see e.g., (Rasmussen and Williams, 2006) and references therein.

Next, we specialize our novel OI-EGP to GPRs that can afford closed-form pdf and weight updates.

3.3 Closed-form updates for GPR

For GPR, the likelihood per expert is given by $p(y_t|\boldsymbol{\theta}^s, s; \mathbf{x}_t) = \mathcal{N}(y_t; \boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_t)\boldsymbol{\theta}^s, \tau)$, which together with the per-expert Gaussian prior $p(\boldsymbol{\theta}^s|s)$, yields the Gaussian posterior at the end of slot t expressed as

$$p(\boldsymbol{\theta}^s|\mathbf{y}_t, s; \mathbf{X}_t) = \mathcal{N}(\boldsymbol{\theta}^s; \hat{\boldsymbol{\theta}}_t^s, \boldsymbol{\Sigma}_t^s) \quad (25)$$

with mean $\hat{\boldsymbol{\theta}}_t^s$ and covariance matrix $\boldsymbol{\Sigma}_t^s$ per expert s.

Building on (25) and (18), the predictive pdf of $\check{f}(\mathbf{x}_{t+1})$ for expert s is also Gaussian

$$p(\check{f}(\mathbf{x}_{t+1})|\mathbf{y}_t, s; \mathbf{X}_t) = \mathcal{N}\Big(\check{f}(\mathbf{x}_{t+1}); \hat{\check{f}}_{t+1|t}^s, \psi_{t+1|t}^s\Big) \quad (26)$$

where the predicted mean and variance are

$$\hat{\check{f}}_{t+1|t}^s = \boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_{t+1})\hat{\boldsymbol{\theta}}_t^s \quad (27a)$$
$$\psi_{t+1|t}^s = \boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_{t+1})\boldsymbol{\Sigma}_t^s\boldsymbol{\phi}_{\mathbf{v}}^s(\mathbf{x}_{t+1}) . \quad (27b)$$

Further, the predictive pdf of y_{t+1} in (19) becomes

$$p(y_{t+1}|\mathbf{y}_t, s; \mathbf{X}_{t+1}) = \mathcal{N}\Big(y_{t+1}; \hat{\check{f}}_{t+1|t}^s, \psi_{t+1|t}^s + \tau\Big) . \quad (28)$$

Thus, the ensemble predictive pdf of y_{t+1} in (20) specialized to GPR is a GM, based on which the minimum mean-square error (MMSE) predictor of y_{t+1} can be obtained together with the associated variance as

$$\hat{y}_{t+1|t} = \sum_{s=1}^S w_t^s \hat{\check{f}}_{t+1|t}^s \quad (29a)$$

$$\sigma_{t+1|t}^2 = \sum_{s=1}^S w_t^s [\psi_{t+1|t}^s + (\hat{y}_{t+1|t} - \hat{\check{f}}_{t+1|t}^s)^2] + \tau. \quad (29b)$$

When y_{t+1} becomes available, experts in GPR update their weights using (cf. (23) and (27))

$$w_{t+1}^s = \frac{w_t^s \mathcal{N}\left(y_{t+1}; \hat{\check{f}}_{t+1|t}^s, \psi_{t+1|t}^s + \tau\right)}{\sum_{s'=1}^S w_t^{s'} \mathcal{N}\Big(y_{t+1}; \hat{\check{f}}_{t+1|t}^{s'}, \psi_{t+1|t}^{s'} + \tau\Big)} . \quad (30)$$

With the per-expert Gaussian likelihood, the arrival of y_{t+1} also propagates Gaussianity to the posterior pdf of $\boldsymbol{\theta}^s$ from slot t to $t+1$, expressed as

$$p(\boldsymbol{\theta}^s|\mathbf{y}_{t+1}, s; \mathbf{X}_{t+1}) = \mathcal{N}(\boldsymbol{\theta}^s; \hat{\boldsymbol{\theta}}_{t+1}^s, \boldsymbol{\Sigma}_{t+1}^s) \quad (31)$$

where the per-expert mean $\hat{\boldsymbol{\theta}}^s_{t+1}$ and covariance matrix $\boldsymbol{\Sigma}^s_{t+1}$ are obtained using the update steps

$$\hat{\boldsymbol{\theta}}^s_{t+1} = \hat{\boldsymbol{\theta}}^s_t + \Big(\psi^s_{t+1|t}+\tau\Big)^{-1}\boldsymbol{\Sigma}^s_t\boldsymbol{\phi}^s_{\mathbf{v}}(\mathbf{x}_{t+1})(y_{t+1}-\hat{\bar{f}}^s_{t+1|t})$$

$$\boldsymbol{\Sigma}^s_{t+1} = \boldsymbol{\Sigma}^s_t - \Big(\psi^s_{t+1|t}+\tau\Big)^{-1}\boldsymbol{\Sigma}^s_t\boldsymbol{\phi}^s_{\mathbf{v}}(\mathbf{x}_{t+1})\boldsymbol{\phi}^{s\top}_{\mathbf{v}}(\mathbf{x}_{t+1})\boldsymbol{\Sigma}^s_t\,.$$

Accounting for all S expert updates, our scalable OI-EGP approach to GPR (see Algorithm 1) has per-iteration complexity of $\mathcal{O}(S(2D)^2)$; hence, scalability is not compromised by the ensemble approach that also offers a richer model for the learning function.

A couple of remarks are now in order.

Remark 1 (OI-EGP for classification). Our novel OI-EGP approach can accommodate classification in addition to regression. Although no closed-form expressions are possible for the predictive and correction pdfs (cf. (19) and (24)), we can readily resort to approximate inference techniques that have been employed in Bayesian logistic regression, namely Laplace approximation or MC sampling; see e.g.,(Rasmussen and Williams, 2006).

Remark 2 (Links with RKHS-based multi-kernel approach (Shen et al., 2019)). The deterministic RKHS online approach (termed "Raker" in (Shen et al., 2019)) relies on first-order gradient descent to update $\boldsymbol{\theta}$ at per-iteration complexity of $\mathcal{O}(SDd)$, which is lower than our second-order update in (14). At the expense of per-iteration complexity of $\mathcal{O}(S(2D)^2)$, our probabilistic OI-EGP approach offers numerically improved performance that is also analytically quantifiable through the predictor variance (29b) in (29a).

4 Regret analysis

The pdfs $\{p(\boldsymbol{\theta}^s|\mathbf{y}_t, s; \mathbf{X}_{t+1})\}_{s=1}^S$ in (24) provide an online performance metric for $\hat{y}_{t+1|t}$, from which its mean and variance can be also obtained (even in closed form, cf. (29b)). These metrics however, rely on the assumptions of knowing the prior pdf of f, and the conditional data likelihood. To guard against having *imperfect knowledge* of these pdfs (the norm in adversarial settings), regret analysis is well motivated along the lines of online convex optimization (Hazan, 2016). This is the subject of this section that aims to benchmark performance of our OI-EGP predictor relative to the best function estimator with data in hindsight.

To this end, let $\mathcal{L}(f(\mathbf{x}_t); y_t) := -\log p(y_t|f(\mathbf{x}_t))$ be the per-slot negative log-likelihood (NLL). For any fixed function estimator $\hat{f}^*(\cdot)$, the incurred loss over T slots is $\sum_{t=1}^T \mathcal{L}(\hat{f}^*(\mathbf{x}_t); y_t)$. In the static setting with the EGP prior (16), the best function estimate (benchmark) with data $\{\mathbf{X}_T, \mathbf{y}_T\}$ available in hindsight, are obtained with the optimal weights $\{w^s\}$ in the EGP prior by maximizing the batch function posterior, $p(\mathbf{f}_T|\mathbf{y}_T; \mathbf{X}_T) \propto p(\mathbf{f}_T; \mathbf{X}_T)p(\mathbf{y}_T|\mathbf{f}_T; \mathbf{X}_T)$, as

$$(\hat{\mathbf{f}}_T, \{\hat{w}^s\}) = \underset{\mathbf{f}_T, \{w^s\}, \sum_s w^s=1}{\arg\max}\; p(\mathbf{y}_T|\mathbf{f}_T; \mathbf{X}_T)\sum_{s=1}^S w^s p(\mathbf{f}_T|s; \mathbf{X}_T)$$

whose solution is $\hat{w}^{s^*} = 1$ and $\hat{w}^s = 0$ for $s \neq s^*$. This implies that only one GP expert s^* is active in the benchmark function estimate for $t = 1, \ldots, T$. The optimal estimate by expert s^* are then given by

$$\hat{\mathbf{f}}_T = \underset{\mathbf{f}_T}{\arg\max}\;\; p(\mathbf{f}_T|s^*; \mathbf{X}_T)p(\mathbf{y}_T|\mathbf{f}_T; \mathbf{X}_T)\,. \tag{33}$$

As every positive semidefinite kernel κ^s is associated with a unique RHKS $\mathcal{H}^s$, the optimal function estimator $\hat{f}^{s^*}(\cdot)$ is extracted from (33) as

$$s^* \in \underset{s=1,\ldots,S}{\arg\min} \sum_{t=1}^T \mathcal{L}(\hat{f}^s(\mathbf{x}_t); y_t) + \frac{1}{2}\|\hat{f}^s\|^2_{\mathcal{H}^s} \tag{34}$$

where the optimal function estimator per expert $\hat{f}^s(\cdot)$, $s = 1, \ldots, S$, is obtained as

$$\hat{f}^s(\cdot) \in \underset{f^s\in\mathcal{H}^s}{\arg\min} \sum_{t=1}^T \mathcal{L}(f^s(\mathbf{x}_t); y_t) + \frac{1}{2}\|f^s\|^2_{\mathcal{H}^s}\,.$$

With the best fixed function estimator $\hat{f}^{s^*}(\cdot)$ at hand, the static regret over T slots is then defined as (Kakade and Ng, 2005)

$$\mathcal{R}(T) := \sum_{t=1}^T \ell_{t|t-1} - \sum_{t=1}^T \mathcal{L}(\hat{f}^{s^*}(\mathbf{x}_t); y_t) \tag{35}$$

where $\ell_{t|t-1}$ is defined as in (22) and captures the ensemble online Bayesian loss incurred by OI-EGP.

Although the cumulative online loss in the first sum of (35) has different form than that of the benchmark, they are comparable by the data likelihood, where the function is nonrandom. In other words, the online Bayesian loss is obtained by taking the expectation of the likelihood wrt the online predictive pdf of the function, thus eliminating the randomness of the function in the likelihood.

To proceed, we will need the following assumptions.

- **(as1)** The NLL $\mathcal{L}(z_t; y_t)$ is continuously twice differentiable with $|\frac{d^2}{dz_t^2}\mathcal{L}(z_t; y_t)| \leq c, \forall z_t$;
- **(as2)** The NLL $\mathcal{L}(z_t; y_t)$ is convex and has bounded derivative wrt z_t; that is, $|\frac{d}{dz_t}\mathcal{L}(z_t; y_t)| \leq L$;
- **(as3)** Kernels $\{\bar{\kappa}^s\}_{s=1}^S$ are shift-invariant, standardized and bounded, that is $\bar{\kappa}^s(\mathbf{x}_t, \mathbf{x}_{t'}) \leq 1, \forall \mathbf{x}_t, \mathbf{x}_{t'}$;

Differentiability and convexity of the NLL in (as1)-(as2) are satisfied by most forms of likelihood in GP-based learning, including the Gaussian likelihood in GPR, and the logistic one for classification. Conditions in (as3) hold for a wide class of kernels including Gaussian, Laplace and Cauchy ones (Rahimi and

Recht, 2008). As the derivations rely on the general form of OI-EGP (cf. (18)-(24)) that corresponds to general likelihoods, the regret bound established here applies to general learning tasks.

For non-Gaussian likelihoods however, the (possibly high-dimensional) integrals involved in (18)-(24) must be evaluated using numerical integration or tractable techniques of approximate inference, thus rendering the ensuing regret analysis valid so long as these approximations are sufficiently accurate.

To establish the static regret bound of OI-EGP, we will need the following lemma.

Lemma 1: *Under (as1), with* (17a) *and* $\|\boldsymbol{\phi}_{\mathbf{v}}^{s}(\mathbf{x}_t)\|^2 \leq 1, \forall s,t$, *the following bound holds on the online Bayesian loss incurred by the OI-EGP and the loss from a single RF-based GP with fixed strategy* $\boldsymbol{\theta}_*^s$

$$\sum_{t=1}^{T} \ell_{t|t-1} \leq \sum_{t=1}^{T} \mathcal{L}(\boldsymbol{\phi}_{\mathbf{v}}^{s\top}(\mathbf{x}_t)\boldsymbol{\theta}_*^s; y_t) + \frac{\|\boldsymbol{\theta}_*^s\|^2}{2\sigma_{\theta^s}^2} + D\log\left(1+\frac{T\sigma_{\theta^s}^2}{2D}\right) + \log S\,. \quad (36)$$

Proof: See Appendix A.

Lemma 1 bounds the cumulative online Bayesian loss of OI-EGP relative to any single RF-based GP learner with a fixed strategy. Next, we will work towards the ultimate static regret by further bounding the loss of RF-based function estimator relative to the best function estimator in the original RKHS for each expert.

Theorem 1: *Under as(1)-as(3) and with* $\hat{f}^{s^*}$ *belonging to the RHKS* $\mathcal{H}^{s^*}$ *induced by* κ^{s^*}, *for a fixed* $\epsilon > 0$, *the following bound holds with probability at least* $1 - 2^8(\frac{\sigma_{s^*}}{\epsilon})^2 \exp\left(\frac{-D\epsilon^2}{4d+8}\right)$

$$\sum_{t=1}^{T} \ell_{t|t-1} - \sum_{t=1}^{T} \mathcal{L}(\hat{f}^{s^*}(\mathbf{x}_t); y_t) \leq \quad (37)$$
$$\frac{(1+\epsilon)C^2}{2\sigma_{\theta^{s^*}}^2} + D\log\left(1+\frac{Tc\sigma_{\theta^{s^*}}^2}{2D}\right) + \log S + \epsilon LTC$$

where C *is a constant, and* $\sigma_{s^*}^2 := \mathbb{E}_{\pi_{\tilde{\kappa}}^{s^*}}[\|\mathbf{v}^{s^*}\|^2]$ *is the second-order moment of* $\mathbf{v}^{s^*}$. *Setting* $\epsilon = \mathcal{O}(\log T/T)$, *the static regret in* (35) *boils down to*

$$\mathcal{R}(T) = \mathcal{O}(\log T)\,. \quad (38)$$

Proof: See Appendix B.

Theorem 1 asserts that OI-EGP incurs no regret on average with cumulative static regret $\mathcal{O}(\log T)$ over T slots, which is tighter than that of the deterministic RHKS-based online multi-kernel counterpart (Shen et al., 2019) with regret $\mathcal{O}(\sqrt{T})$ in the static setting.

5 DOI-EGP for dynamic learning

So far, each RF-based GP expert s in OI-EGP relies on a time-invariant $\boldsymbol{\theta}^s$ (cf. (17b)), and correspondingly $\check{f}$ entails no temporal dynamics. To handle a time-variant learning function $\check{f}_t$, this section outlines the dynamic (D) OI-EGP, where expert s adopts a time-varying $\boldsymbol{\theta}_t^s$ obeying the random walk model

$$\boldsymbol{\theta}_{t+1}^s = \boldsymbol{\theta}_t^s + \boldsymbol{\epsilon}_{t+1}^s \quad (39)$$

where the noise $\boldsymbol{\epsilon}_{t+1}^s$ is white and Gaussian distributed with mean zero and covariance matrix $\sigma_{\epsilon^s}^2\mathbf{I}_{2D}$.

Rather than updating $p(\boldsymbol{\theta}^s|\mathbf{y}_t, s; \mathbf{X}_t)$ as in OI-EGP, expert s in DOI-EGP propagates $p(\boldsymbol{\theta}_t^s|\mathbf{y}_t, s; \mathbf{X}_t)$ across slots. Taking into account (39), expert s first predicts the pdf of $\boldsymbol{\theta}_{t+1}^s$ at the beginning of slot $t+1$ as

$$p(\boldsymbol{\theta}_{t+1}^s|\mathbf{y}_t, s; \mathbf{X}_{t+1}) = \int p(\boldsymbol{\theta}_{t+1}^s|\boldsymbol{\theta}_t^s)\, p(\boldsymbol{\theta}_t^s|\mathbf{y}_t, s; \mathbf{X}_t)\, d\boldsymbol{\theta}_t^s \quad (40)$$

which replaces $p(\boldsymbol{\theta}^s|\mathbf{y}_t, s; \mathbf{X}_t)$ in (18) and (24) to obtain the predictive pdf $p(\check{f}_{t+1}(\mathbf{x}_{t+1})|\mathbf{y}_t, s; \mathbf{X}_t)$, and the posterior $p(\boldsymbol{\theta}_{t+1}^s|\mathbf{y}_{t+1}, s; \mathbf{X}_{t+1})$ in the dynamic setting. Specifically for GPR with per-expert Gaussian posterior $p(\boldsymbol{\theta}_t^s|\mathbf{y}_t, s; \mathbf{X}_t) = \mathcal{N}(\boldsymbol{\theta}_t^s; \hat{\boldsymbol{\theta}}_t^s, \boldsymbol{\Sigma}_t^s)$, the predictive pdf in (40) is $p(\boldsymbol{\theta}_{t+1}^s|\mathbf{y}_t, s; \mathbf{X}_{t+1}) = \mathcal{N}(\boldsymbol{\theta}_{t+1}^s; \hat{\boldsymbol{\theta}}_t^s, \boldsymbol{\Sigma}_t^s + \sigma_{\epsilon^s}^2\mathbf{I}_{2D})$.

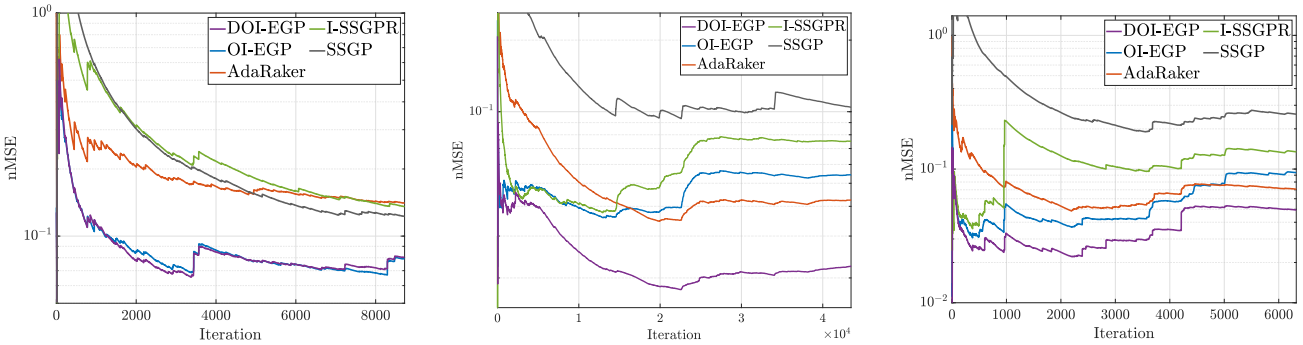
Remark 3. Time-varying state-space models of latent variables have been also considered for GP-based dynamic function learning in (Wang et al., 2006; Turner et al., 2010). The key novelty is that our RF-based OI-EGP can be broadened using the linear dynamic model in (39) or state-space generalizations thereof.

6 Numerical tests

To assess performance, tests are presented here for GPR; see also Appendix E for classification tests. Code for (D)OI-EGP can be found at `github.com/gkaranikolas/oiegp`

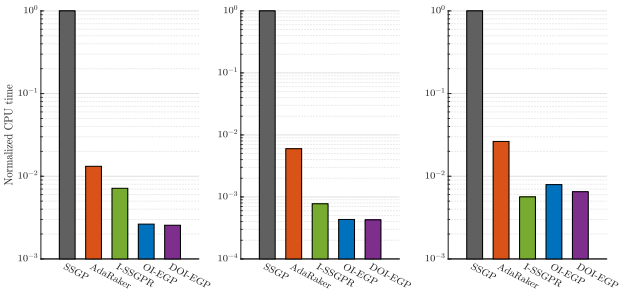
Tests on real data. We compared the proposed (D)OI-EGP approaches with AdaRaker (Shen et al., 2019), Incremental Sparse Spectrum Gaussian Process Regression (I-SSGPR) (Gijsberts and Metta, 2013), and the Streaming Sparse Gaussian Process (SSGP) approach (Bui et al., 2017), in terms of normalized mean-square error (nMSE) and running time. With s_y^2 denoting the sample variance of $\mathbf{y}_T$, the nMSE is defined as $\text{nMSE}_t := t^{-1}\sum_{t'=1}^{t}(y_{t'} - \hat{y}_{t'|t'-1})^2/s_y^2$.

Tests were performed on the SARCOS dataset (Rasmussen and Williams, 2006), widely used for evaluating GP-based approaches, as well as on the "Air quality" (De Vito et al., 2008), "Tom's hardware" and "Twitter" datasets (Kawala et al., 2013) from the UCI repository (Dua and Graff, 2017). The statistics of the datasets are summarized in Table 1 in Appendix C.



(a) (b) (c)

Figure 1: Log scale nMSE plots on (a) "Tom's hardware;" (b) SARCOS; and, (c) "Air quality" datasets.



(a) (b) (c)

Figure 2: Normalized running times on (a) "Tom's hardware", (b) SARCOS and (c) "Air quality" datasets. Notice the logarithmic scale.

For all RF-based approaches (namely (D)OI-EGP, AdaRaker and I-SSGPR) we used $2D = 100$ and the reported results correspond to the run which resulted in the median nMSE among 101 runs for the corresponding method. Finally, all reported runtimes include hyperparameter learning/model initialization computations performed on the first $1,000$ samples. If for some expert s and time instance t we have that $w_t^s = 0$, it follows that $w_{t'}^s = 0$ for all $t' > t$ (cf. (23)). Experts with $w_t^s < 10^{-16}$ were deemed inactive for $t' > t$; thus, we set $w_{t'}^s = 0$ for $t' > t$, and avoided unnecessary prediction/correction steps.

The kernel dictionary for (D)OI-EGP and AdaRaker comprised radial basis functions with variances from the set $\{10^k\}_{k=-4}^{6}$. The automatic relevance determination (ARD) kernel was used for I-SSGPR, as in (Gijsberts and Metta, 2013). The per kernel noise and prior variances (as well as ARD length scales for I-SSGPR), were estimated by maximizing the marginal likelihood of the first $1,000$ samples using the `minimize` function from the GPML toolbox (Rasmussen and Nickisch, 2010). The aforementioned samples were not used in the deployment phase. In DOI-EGP, $\sigma_{\epsilon^s}^2 = 0.001$ was used for all s and in all experiments. Regarding SSGP, the ARD kernel was used, the batch size was set to 300, the number of inducing points was 100 and the first $1,000$ samples were used for obtaining an initial model, all as per the original work (Bui et al., 2017).

The nMSE performance of the tested approaches on the "Tom's hardware" dataset is plotted in Fig. 1(a). The proposed OI-EGP and DOI-EGP approaches outperform the competing alternatives in terms of nMSE while also featuring the lowest running time, which corresponds to less than 0.3% of that of the most closely competing (in terms of nMSE) alternative (cf. Fig. 2(a)). The results on the SARCOS dataset are depicted in Fig. 1(b). Our OI-EGP remains competitive whereas the proposed dynamic variant (DOI-EGP) features the lowest nMSE, while also achieving both faster convergence as well as a runtime that is an order of magnitude lower than that of the second best (in terms of nMSE) approach (cf. Fig. 2(b)). These results further highlight the computational efficiency of the proposed approaches. Similar observations can be made on the "Air quality" (cf. Figs. 1,2(c)) and "Twitter" datasets (cf. Figs. 6, 7 in Appendix C). Additional test results in terms of predictive negative log-likelihood are presented in Appendix C.

Synthetic tests. Due to space limitations, Appendix D includes tests on synthetic data that validate the regret bound (cf. (38)), as well as illustrate the uncertainty quantification capabilities of OI-EGP.

7 Conclusions

This paper put forth an online interactive scheme that leverages an ensemble of scalable RF-based parametric GP learners to jointly infer the unknown function along with its performance, and a data-driven kernel combination. Regret analysis was conducted to benchmark even in adversarial settings the novel so-termed OI-EGP relative to the best fixed function estimator with data in hindsight. Time-varying learning was enabled through modeling the parameter dynamics. Experimental results with real data illustrate the superior performance of the novel EGP schemes.

Acknowledgement. We would like to thank the anonymous reviewers for their constructive feedback. We also gratefully acknowledge the support from NSF grants 1508993, 1711471 and 1901134.

References

[Alvarez et al. 2012] Alvarez, Mauricio A. ; Rosasco, Lorenzo ; Lawrence, Neil D. et al.: Kernels for vector-valued functions: A review. In: *Foundations and Trends® in Machine Learning* 4 (2012), Nr. 3, S. 195–266

[Bui et al. 2017] Bui, Thang D. ; Nguyen, Cuong ; Turner, Richard E.: Streaming sparse Gaussian process approximations. In: *Advances in Neural Information Processing Systems*, 2017, S. 3299–3307

[Cheng and Boots 2016] Cheng, Ching-An ; Boots, Byron: Incremental variational sparse Gaussian process regression. In: *Advances in Neural Information Processing Systems*, 2016, S. 4410–4418

[De Vito et al. 2008] De Vito, Saverio ; Massera, Ettore ; Piga, Marco ; Martinotto, Luca ; Di Francia, Girolamo: On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. In: *Sensors and Actuators B: Chemical* 129 (2008), Nr. 2, S. 750–757

[Deisenroth and Ng 2015] Deisenroth, Marc P. ; Ng, Jun W.: Distributed Gaussian processes. In: *Proc. of Intl. Conf. on Machine Learning*, 2015

[Dua and Graff 2017] Dua, Dheeru ; Graff, Casey: *UCI Machine Learning Repository*. 2017. – URL `http://archive.ics.uci.edu/ml`

[Gal and Turner 2015] Gal, Yarin ; Turner, Richard: Improving the Gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs. In: *Proc. of Intl. Conf. on Machine Learning*, 2015

[Gijsberts and Metta 2013] Gijsberts, Arjan ; Metta, Giorgio: Real-time model learning using incremental sparse spectrum Gaussian process regression. In: *Neural Networks* 41 (2013), S. 59–69

[Hazan 2016] Hazan, Elad: Introduction to online convex optimization. In: *Foundations and Trends® in Optimization* 2 (2016), Nr. 3-4, S. 157–325

[Jin et al. 2010] Jin, Rong ; Hoi, Steven C. ; Yang, Tianbao: Online multiple kernel learning: Algorithms and mistake bounds. In: *Proc. of Intl. Conf. on Algorithmic Learning Theory*, 2010, S. 390–404

[Kakade and Ng 2005] Kakade, Sham M. ; Ng, Andrew Y.: Online bounds for Bayesian algorithms. In: *Advances in Neural Information Processing Systems*, 2005, S. 641–648

[Kakade et al. 2006] Kakade, Sham M. ; Seeger, Matthias W. ; Foster, Dean P.: Worst-case bounds for Gaussian process models. In: *Advances in Neural Information Processing Systems*, 2006, S. 619–626

[Kawala et al. 2013] Kawala, François ; Douzal-Chouakria, Ahlame ; Gaussier, Eric ; Dimert, Eustache: Prédictions d'activité dans les réseaux sociaux en ligne. In: *4ième conférence sur les modèles et l'analyse des réseaux : Approches mathématiques et informatiques*. France, Oktober 2013, S. 16

[Lázaro-Gredilla et al. 2010] Lázaro-Gredilla, Miguel ; Candela, Joaquin Quiñonero ; Rasmussen, Carl E. ; Figueiras-Vidal, A.: Sparse spectrum Gaussian process regression. In: *Journal of Machine Learning Research* 11 (2010), Nr. Jun, S. 1865–1881

[Lu et al. 2016] Lu, Jing ; Hoi, Steven C. ; Wang, Jialei ; Zhao, Peilin ; Liu, Zhi-Yong: Large scale online kernel learning. In: *The Journal of Machine Learning Research* 17 (2016), Nr. 1, S. 1613–1655

[Meeds and Osindero 2006] Meeds, Edward ; Osindero, Simon: An alternative infinite mixture of Gaussian process experts. In: *Advances in Neural Information Processing Systems*, 2006, S. 883–890

[Micchelli and Pontil 2005] Micchelli, Charles A. ; Pontil, Massimiliano: Learning the kernel function via regularization. In: *Journal of machine learning research* 6 (2005), Nr. Jul, S. 1099–1125

[Nguyen et al. 2017] Nguyen, Cuong V. ; Bui, Thang D. ; Li, Yingzhen ; Turner, Richard E.: Online Variational Bayesian Inference: Algorithms for Sparse Gaussian Processes and Theoretical Bounds. In: *ICML Time Series Workshop*, 2017

[Quiñonero-Candela and Rasmussen 2005] Quiñonero-Candela, Joaquin ; Rasmussen, Carl E.: A unifying view of sparse approximate Gaussian process regression. In: *Journal of Machine Learning Research* 6 (2005), Nr. Dec, S. 1939–1959

[Rahimi and Recht 2008] Rahimi, Ali ; Recht, Benjamin: Random features for large-scale kernel machines. In: *Advances in Neural Information Processing Systems*, 2008, S. 1177–1184

[Rasmussen and Ghahramani 2002] Rasmussen, Carl E. ; Ghahramani, Zoubin: Infinite mixtures of Gaussian process experts. In: *Advances in Neural Information Processing Systems*, 2002, S. 881–888

[Rasmussen and Nickisch 2010] Rasmussen, Carl E. ; Nickisch, Hannes: Gaussian processes for machine learning (GPML) toolbox. In: *Journal of Machine Learning research* 11 (2010), Nr. Nov, S. 3011–3015

[Rasmussen and Williams 2006] Rasmussen, Carl E. ; Williams, Christopher K.: *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006

[Richard et al. 2008] Richard, Cédric ; Bermudez, José Carlos M ; Honeine, Paul: Online prediction of time series data with kernels. In: *IEEE Transactions on Signal Processing* 57 (2008), Nr. 3, S. 1058–1067

[Schölkopf et al. 2002] Schölkopf, Bernhard ; Smola, Alexander J. ; Bach, Francis et al.: *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, MA : MIT press, 2002

[Seeger et al. 2008] Seeger, Matthias W. ; Kakade, Sham M. ; Foster, Dean P.: Information consistency of nonparametric Gaussian process methods. In: *IEEE Transactions on Information Theory* 54 (2008), Nr. 5, S. 2376–2382

[Shen et al. 2019] Shen, Yanning ; Chen, Tianyi ; Giannakis, Georgios B.: Random feature-based online multi-kernel learning in environments with unknown dynamics. In: *The Journal of Machine Learning Research* 20 (2019), Nr. 1, S. 773–808

[Snelson and Ghahramani 2006] Snelson, Edward ; Ghahramani, Zoubin: Sparse Gaussian processes using pseudo-inputs. In: *Advances in Neural Information Processing Systems*, 2006, S. 1257–1264

[Snelson and Ghahramani 2007] Snelson, Edward ; Ghahramani, Zoubin: Local and global sparse Gaussian process approximations. In: *Artificial Intelligence and Statistics*, 2007, S. 524–531

[Titsias 2009] Titsias, Michalis: Variational learning of inducing variables in sparse Gaussian processes. In: *Proc. of Intl. Conf. on Artificial Intelligence and Statistics*, 2009, S. 567–574

[Tresp 2001] Tresp, Volker: Mixtures of Gaussian processes. In: *Advances in Neural Information Processing Systems*, 2001, S. 654–660

[Turner et al. 2010] Turner, Ryan ; Deisenroth, Marc ; Rasmussen, Carl: State-space inference and learning with Gaussian processes. In: *Proc. of Intl. Conf. on Artificial Intelligence and Statistics*, 2010, S. 868–875

[Wang et al. 2006] Wang, Jack ; Hertzmann, Aaron ; Fleet, David J.: Gaussian process dynamical models. In: *Advances in Neural Information Processing Systems*, 2006, S. 1441–1448

[Wang et al. 2012] Wang, Zhuang ; Crammer, Koby ; Vucetic, Slobodan: Breaking the curse of kernelization: Budgeted stochastic gradient descent for large-scale svm training. In: *Journal of Machine Learning Research* 13 (2012), Nr. Oct, S. 3103–3131

[Xu et al. 2014] Xu, Nuo ; Low, Kian H. ; Chen, Jie ; Lim, Keng K. ; Ozgul, Etkin B.: GP-Localize: Persistent mobile robot localization using online sparse Gaussian process observation model. In: *AAAI Conference on Artificial Intelligence*, 2014

[Yuan and Neubauer 2009] Yuan, Chao ; Neubauer, Claus: Variational mixture of Gaussian process experts. In: *Advances in Neural Information Processing Systems*, 2009, S. 1897–1904