Stochastic Variational Inference for Bayesian Sparse Gaussian Process Regression

Haibin Yu
National University of Singapore
Republic of Singapore
haibin@u.nus.edu

Trong Nghia Hoang
MIT-IBM Watson AI Lab
Cambridge, MA, USA
nghiaht@ibm.com

Bryan Kian Hsiang Low
National University of Singapore
Republic of Singapore
lowkh@comp.nus.edu.sg

Patrick Jaillet
MIT
Cambridge, MA, USA
jaillet@mit.edu

***Abstract*—This paper presents a novel variational inference framework for deriving a family of Bayesian *sparse Gaussian process regression* (SGPR) models whose approximations are variationally optimal with respect to the full-rank GPR model enriched with various corresponding correlation structures of the observation noises. Our *variational Bayesian SGPR* (VBSGPR) models jointly treat both the distributions of the inducing variables and hyperparameters as variational parameters, which enables the decomposability of the variational lower bound that in turn can be exploited for stochastic optimization. Such a stochastic optimization involves iteratively following the stochastic gradient of the variational lower bound to improve its estimates of the optimal variational distributions of the inducing variables and hyperparameters (and hence the predictive distribution) of our VBSGPR models and is guaranteed to achieve asymptotic convergence to them. We show that the stochastic gradient is an unbiased estimator of the exact gradient and can be computed in constant time per iteration, hence achieving scalability to big data. We empirically evaluate the performance of our proposed framework on two real-world, massive datasets.**

I. Introduction

A *Gaussian process regression* (GPR) model is a rich class of Bayesian non-parametric models that can exploit correlation of the data/observations for performing probabilistic non-linear regression by providing a Gaussian predictive distribution with formal measures of predictive uncertainty. Such a *full-rank GPR* (FGPR) model, though highly expressive, incurs cubic time in the data size to compute the predictive distribution and learn the hyperparameters (i.e., defining its correlation structure) via maximum likelihood estimation, specifically, in each iteration of gradient ascent to refine the hyperparameter estimates to improve the log-marginal likelihood. So, to learn the hyperparameters in reasonable time, only a very small subset of the data can be considered, which compromises the estimation accuracy: It is typically not representative of all the data in describing the underlying correlation structure due to its sparsity over the input space.

To improve its time efficiency, a number of *sparse GPR* (SGPR) models exploiting low-rank covariance matrix approximations [1], [2] have been proposed, many of which impose a common structural assumption of conditional independence (but of varying degrees) on the FGPR model based on the notion of *inducing variables* and can therefore be encompassed under a unifying view presented in [2]. As a result, they incur linear time in the data size that is still prohibitively expensive for training with big data (i.e., million-sized datasets). To scale up to big data, parallel [3]–[5] and online [6], [7] variants of several of these SGPR models have been developed for prediction (by assuming known hyperparameters) but not hyperparameter learning.

The chief concern with the unifying view of [2] is that it does not rigorously quantify the approximation quality of a SGPR model [8]. To address this concern, the work of [9] has proposed a principled variational inference framework that involves minimizing the *Kullback-Leibler* (KL) distance between distributions of some latent variables (including the inducing variables) induced by the variational SGPR approximation and the FGPR model given the data/observations or, equivalently, maximizing a lower bound of the log-marginal likelihood to yield the *deterministic training conditional* (DTC) approximation [10]. Hyperparameter learning is then achieved by maximizing this variational lower bound with respect to the hyperparameters via gradient ascent, which still incurs linear time in the data size per iteration but can be substantially reduced by means of parallelization [11] or stochastic optimization [12], [13]. Unifying frameworks of variational SGPR models and their stochastic and distributed variants are subsequently proposed in [14], [15] to, respectively, perform stochastic and distributed variational inference for any SGPR model (including DTC) spanned by the unifying view of [2]. The work of [16] has extended two SGPR models (i.e., DTC and *fully independent training conditional* (FITC) approximation [17]) to handle streaming data.

However, all the above-mentioned variational SGPR models and their stochastic and distributed variants suffer from the following critical issues: (a) The above equivalence only holds for the case of fixed hyperparameters; otherwise, since the log-marginal likelihood also depends on the same hyperparameters that are optimized to maximize its variational lower bound, the resulting KL distance, which quantifies the gap between the log-marginal likelihood and its lower bound, may not be minimized; (b) similar to variational expectation-maximization [18], the log-marginal likelihood does not necessarily increase in each iteration of gradient ascent to refine the hyperparameter

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme and the Singapore Ministry of Education Academic Research Fund Tier 2, MOE2016-T2-2-156.

estimates to improve its variational lower bound; and (c) they all find point estimates of the hyperparameters, which risk overfitting, especially when the number of hyperparameters is all but small.

To resolve these issues, the notable work of [19] has introduced a *variational Bayesian DTC* (VBDTC) approximation (Section IV) capable of learning a variational distribution of the hyperparameters. This learned distribution of hyperparameters is particularly desirable in conveying the uncertainty/confidence of the hyperparameter estimates and for use in Bayesian GP regression (Section VI), active learning [20]–[26], Bayesian optimization [27]–[30], among others. Unfortunately, such a VBDTC approximation cannot handle big data (e.g., million-sized datasets) because it incurs linear time in the data size per iteration of gradient ascent. The *variational Bayesian sparse spectrum GPR* (VSSGPR) model [31] overcomes this scalability issue by achieving constant time per iteration of stochastic gradient ascent. But, like VBDTC, VSSGPR imposes a highly restrictive assumption of conditional independence between the test outputs and the training data given the learned hyperparameters (i.e., in its test conditional in equation 4 therein), thus compromising its predictive performance as shown in our experiments (Section VII). This assumption is later relaxed in the work of [32]. It remains an open question whether more refined SGPR models as well as those others spanned by the unifying view of [2] (e.g., FITC, *partially independent training conditional* (PITC), *partially independent conditional* (PIC) [33] approximations) are amenable to the variational Bayesian treatment and achieve scalability through stochastic optimization.

To address this question, this paper presents a novel variational inference framework for deriving a family of Bayesian SGPR models (e.g., VBDTC, VBFITC, VBPIC) whose approximations are, interestingly, variationally optimal with respect to the FGPR model enriched with various corresponding correlation structures of the observation noises (Section IV). Our framework introduces a novel reparameterization of the GP model (Section III) for enabling a variational treatment of the distribution of hyperparameters. Unlike VBDTC, our framework does not need to assume independently distributed observation noises with constant variance and is thus more robust to different noise correlation structures, hence catering to more realistic applications of GP. Furthermore, instead of just considering the distribution of hyperparameters as variational parameters [19], [31], we jointly treat both the distributions of the inducing variables and hyperparameters as variational parameters, which enables the decomposability of the variational lower bound that in turn can be exploited for stochastic optimization (Section V). Such a stochastic optimization involves iteratively following the stochastic gradient of the variational lower bound to improve its estimates of the optimal variational distributions of the inducing variables and hyperparameters (and hence the predictive distribution (Section VI)) of our *variational Bayesian SGPR* (VBSGPR) models and is guaranteed to achieve asymptotic convergence to them. We show that the derived stochastic gradient is an unbiased estimator of the exact gradient and can be computed in constant time (i.e., independent of data size) per iteration, thus achieving scalability to big data. We empirically evaluate the performance of the stochastic variants of our VBSGPR models on two real-world datasets (Section VII).

II. Background and Notations

A. Full-Rank GP Regression (FGPR) with Correlated Noises

Let $\mathcal{X}$ denote a d-dimensional input feature space such that each input vector $\mathbf{x} \in \mathcal{X}$ is associated with a latent output variable $f_{\mathbf{x}}$. Let $\{f_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{X}}$ denote a *Gaussian process* (GP), that is, every finite subset of $\{f_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{X}}$ follows a multivariate Gaussian distribution. Then, the GP is fully specified by its *prior* mean $\mathbb{E}[f_{\mathbf{x}}]$ (i.e., assumed to be zero to ease notations) and covariance $k_{\mathbf{x}\mathbf{x}'} \triangleq \mathrm{cov}[f_{\mathbf{x}}, f_{\mathbf{x}'}]$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, the latter of which can be defined, for example, by the widely-used squared exponential covariance function $k_{\mathbf{x}\mathbf{x}'} \triangleq \sigma_f^2 \exp(-0.5\|\mathbf{\Lambda}\mathbf{x} - \mathbf{\Lambda}\mathbf{x}'\|_2^2)$ where $\mathbf{\Lambda} = \mathrm{diag}[\lambda_1, \ldots, \lambda_d]$ and σ_f^2 are its *inverted* length-scale and signal variance hyperparameters, respectively. Suppose that a column vector $\mathbf{y}_{\mathcal{D}} \triangleq (y_{\mathbf{x}})_{\mathbf{x}\in\mathcal{D}}^\top$ of noisy observed outputs $y_{\mathbf{x}} \triangleq f_{\mathbf{x}} + \varepsilon_{\mathbf{x}}$ (i.e., corrupted by an additive noise $\varepsilon_{\mathbf{x}}$) is available for some set $\mathcal{D} \subset \mathcal{X}$ of training inputs such that $\boldsymbol{\varepsilon}_{\mathcal{D}} \triangleq (\varepsilon_{\mathbf{x}})_{\mathbf{x}\in\mathcal{D}}^\top$ follows a multivariate Gaussian distribution $p(\boldsymbol{\varepsilon}_{\mathcal{D}}) \triangleq \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathcal{D}\mathcal{D}})$ where $\mathbf{C}_{\mathcal{D}\mathcal{D}}$ is a covariance matrix representing the correlation of observation noises $\boldsymbol{\varepsilon}_{\mathcal{D}}$. It follows that $p(\mathbf{y}_{\mathcal{D}}|\mathbf{f}_{\mathcal{D}}) = \mathcal{N}(\mathbf{f}_{\mathcal{D}}, \mathbf{C}_{\mathcal{D}\mathcal{D}})$ where $\mathbf{f}_{\mathcal{D}} \triangleq (f_{\mathbf{x}})_{\mathbf{x}\in\mathcal{D}}^\top$. Then, a FGPR model with correlated observation noises can perform probabilistic regression by providing a GP *posterior*/predictive distribution $p(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}}) = \mathcal{N}(\mathbf{K}_{\mathbf{x}^*\mathcal{D}}(\mathbf{K}_{\mathcal{D}\mathcal{D}} + \mathbf{C}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{y}_{\mathcal{D}}, k_{\mathbf{x}^*\mathbf{x}^*} - \mathbf{K}_{\mathbf{x}^*\mathcal{D}}(\mathbf{K}_{\mathcal{D}\mathcal{D}} + \mathbf{C}_{\mathcal{D}\mathcal{D}})^{-1}\mathbf{K}_{\mathcal{D}\mathbf{x}^*})$ of the latent output $f_{\mathbf{x}^*}$ for any test input $\mathbf{x}^* \in \mathcal{X}$ where $\mathbf{K}_{\mathbf{x}^*\mathcal{D}} \triangleq (k_{\mathbf{x}^*\mathbf{x}})_{\mathbf{x}\in\mathcal{D}}$, $\mathbf{K}_{\mathcal{D}\mathcal{D}} \triangleq (k_{\mathbf{x}\mathbf{x}'})_{\mathbf{x},\mathbf{x}'\in\mathcal{D}}$, and $\mathbf{K}_{\mathcal{D}\mathbf{x}^*} \triangleq \mathbf{K}_{\mathbf{x}^*\mathcal{D}}^\top$. Computing the GP predictive distribution incurs $\mathcal{O}(|\mathcal{D}|^3)$ time due to inversion of $\mathbf{K}_{\mathcal{D}\mathcal{D}} + \mathbf{C}_{\mathcal{D}\mathcal{D}}$. The FGPR hyperparameters $\boldsymbol{\theta} \triangleq (\lambda_1, \ldots, \lambda_d, \sigma_f)^\top$ can be learned using *maximum likelihood estimation* (MLE) by maximizing the log-marginal likelihood $\log p(\mathbf{y}_{\mathcal{D}}) = \log \mathcal{N}(\mathbf{0}, \mathbf{K}_{\mathcal{D}\mathcal{D}} + \mathbf{C}_{\mathcal{D}\mathcal{D}})$ with respect to $\boldsymbol{\theta}$ via gradient ascent, which incurs $\mathcal{O}(|\mathcal{D}|^3)$ time per iteration. So, the FGPR model with correlated noises scales poorly in data size $|\mathcal{D}|$. To improve its scalability, our key idea is to impose different sparsity structures on $\mathbf{C}_{\mathcal{D}\mathcal{D}}$ to yield a family of VBSGPR models, as shown in Section IV.

B. Sparse Gaussian Process Regression (SGPR)

To reduce the cubic time cost of the FGPR model, the SGPR models spanned by the unifying view of [2] exploit a vector $\mathbf{f}_{\mathcal{U}} \triangleq (f_{\mathbf{x}})_{\mathbf{x}\in\mathcal{U}}^\top$ of inducing output variables for some small set $\mathcal{U} \subset \mathcal{X}$ of inducing inputs (i.e., $|\mathcal{U}| \ll |\mathcal{D}|$) for approximating the GP predictive distribution $p(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}})$. In particular, they utilize a common structural assumption [33] that the joint distribution of $f_{\mathbf{x}^*}$ and $\mathbf{f}_{\mathcal{D}} \triangleq (f_{\mathbf{x}})_{\mathbf{x}\in\mathcal{D}}^\top$ given $\mathbf{f}_{\mathcal{U}}$ factorizes over a pre-defined partition of the input space $\mathcal{X}$ into B disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_B$ (i.e., $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \ldots \cup \mathcal{X}_B$): Formally, without loss of generality, supposing $\mathbf{x}^* \in \mathcal{X}_B$, then $p(f_{\mathbf{x}^*}, \mathbf{f}_{\mathcal{D}}|\mathbf{f}_{\mathcal{U}}) = p(f_{\mathbf{x}^*}|\mathbf{f}_{\mathcal{D}_B}, \mathbf{f}_{\mathcal{U}}) \prod_{i=1}^B p(\mathbf{f}_{\mathcal{D}_i}|\mathbf{f}_{\mathcal{U}})$ where $\mathbf{f}_{\mathcal{D}_i} \triangleq (f_{\mathbf{x}})_{\mathbf{x}\in\mathcal{D}_i}^\top$ is a column vector of latent outputs for the disjoint

subset $\mathcal{D}_i \triangleq (\mathcal{X}_i \cap \mathcal{D}) \subset \mathcal{D}$ for $i = 1, 2, \ldots, B$. Using this factorization, $p(f_{\mathbf{x}^*}|\mathbf{y}_\mathcal{D}) = \int p(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}_B}, \mathbf{f}_\mathcal{U})\, p(\mathbf{f}_\mathcal{U}|\mathbf{y}_\mathcal{D})\, \mathrm{d}\mathbf{f}_\mathcal{U} \simeq \int q(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}_B}, \mathbf{f}_\mathcal{U})\; q(\mathbf{f}_\mathcal{U})\; \mathrm{d}\mathbf{f}_\mathcal{U}$ where $\mathbf{y}_{\mathcal{D}_B} \triangleq (y_\mathbf{x})^\top_{\mathbf{x}\in\mathcal{D}_B}$ is a vector of noisy observed outputs for the subset $\mathcal{D}_B$ of training inputs, the equality is derived in Appendix C.1 of [14], and $p(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}_B}, \mathbf{f}_\mathcal{U})$ and $p(\mathbf{f}_\mathcal{U}|\mathbf{y}_\mathcal{D})$ are, respectively, approximated by $q(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}_B}, \mathbf{f}_\mathcal{U})$ and $q(\mathbf{f}_\mathcal{U})$ that can be appropriately defined to reproduce the predictive distribution of any SGPR model [14] spanned by the unifying view of [2], which can be computed in $\mathcal{O}(|\mathcal{D}||\mathcal{U}|^2)$ time. The SGPR hyperparameters can be learned using MLE by maximizing its corresponding log-marginal likelihood via gradient ascent, which incurs $\mathcal{O}(|\mathcal{D}||\mathcal{U}|^2)$ time per iteration. To scale up to big data, these linear time complexities can be significantly reduced using parallelization or stochastic optimization (Section I).

C. Bayesian SGPR Models

For the FGPR and SGPR models described above, point estimates of their hyperparameters are learned, which is vulnerable to overfitting, especially when the number of hyperparameters is all but small (Section I). To mitigate this issue of overfitting, a Bayesian approach to sparse GP regression can be employed by introducing priors $p(\boldsymbol{\theta}) \triangleq p(\boldsymbol{\Lambda})\; p(\sigma_f)$ over hyperparameters $\boldsymbol{\theta}$, thus yielding the predictive distribution:

$$\begin{aligned} p(f_{\mathbf{x}^*}|\mathbf{y}_\mathcal{D}) &= \int p(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}_B}, \mathbf{f}_\mathcal{U}, \boldsymbol{\theta})\; p(\mathbf{f}_\mathcal{U}, \boldsymbol{\theta}|\mathbf{y}_\mathcal{D})\; \mathrm{d}\mathbf{f}_\mathcal{U}\; \mathrm{d}\boldsymbol{\theta} \\ &\simeq \int q(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}_B}, \mathbf{f}_\mathcal{U}, \boldsymbol{\theta})\; q(\mathbf{f}_\mathcal{U}, \boldsymbol{\theta})\; \mathrm{d}\mathbf{f}_\mathcal{U}\; \mathrm{d}\boldsymbol{\theta} \qquad (1) \end{aligned}$$

where $p(\mathbf{f}_\mathcal{U}, \boldsymbol{\theta}|\mathbf{y}_\mathcal{D})$ is approximated by $q(\mathbf{f}_\mathcal{U}, \boldsymbol{\theta})$ which generalizes $q(\mathbf{f}_\mathcal{U})$ above by additionally and jointly considering the hyperparameters $\boldsymbol{\theta}$ as variational variables. Though (1), in principle, allows a Bayesian treatment of $\boldsymbol{\theta}$ to be incorporated into the existing SGPR models, computing the resulting predictive distribution is intractable because it involves integrating, over $\boldsymbol{\Lambda}$, probability terms in (1) containing the inverse of $\mathbf{K}_{\mathcal{U}\mathcal{U}} \triangleq (k_{\mathbf{x}\mathbf{x}'})_{\mathbf{x},\mathbf{x}'\in\mathcal{U}}$ that depends on $\boldsymbol{\Lambda}$ but without an analytical form with respect to $\boldsymbol{\Lambda}$. To resolve this, we introduce a reparameterization trick to make the prior distribution of inducing outputs independent of the hyperparameters $\boldsymbol{\theta}$, as discussed next.

III. Reparameterizing Bayesian SGPR Models

Let $\phi : \mathbb{R}^d \to \mathcal{H}$ denote a non-linear feature map from the input space $\mathbb{R}^d$ into a *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ whose inner product is defined as $\langle \phi(\mathbf{x}), \phi(\mathbf{x}')\rangle_\mathcal{H} \triangleq \exp(-0.5\|\mathbf{x} - \mathbf{x}'\|_2^2)$. Given ϕ, the GP covariance/kernel function can be interpreted as $k_{\mathbf{x}\mathbf{x}'} \triangleq \langle \sigma(\mathbf{x})\phi(\boldsymbol{\Lambda}\mathbf{x}), \sigma(\mathbf{x}')\phi(\boldsymbol{\Lambda}\mathbf{x}')\rangle_\mathcal{H} = \sigma(\mathbf{x})\sigma(\mathbf{x}')\exp(-0.5\|\boldsymbol{\Lambda}\mathbf{x} - \boldsymbol{\Lambda}\mathbf{x}'\|_2^2)$ where σ is an arbitrary function mapping from $\mathbb{R}^d$ to $\mathbb{R}$. This implies $k_{\mathbf{x}\mathbf{x}} = \sigma^2(\mathbf{x})$ which allows $\sigma^2(\mathbf{x})$ to be interpreted as the prior variance $k_{\mathbf{x}\mathbf{x}}$ of $f_\mathbf{x}$ (Section II-A).

We will now describe the reparameterization trick: Let $\mathcal{I} \triangleq \{\boldsymbol{\Lambda}\mathbf{x}\}_{\mathbf{x}\in\mathcal{U}}$. Intuitively, $\mathcal{I}$ can be interpreted as a set of *rotated inducing inputs* with the diagonal matrix $\boldsymbol{\Lambda}$ of inverted length-scales being the rotation matrix. Let each rotated inducing input $\mathbf{z} \in \mathcal{I}$ be associated with a latent output variable $s_\mathbf{z}$. Then, for all $\mathbf{z}, \mathbf{z}' \in \mathcal{I}$, $\mathrm{cov}[s_\mathbf{z}, s_{\mathbf{z}'}] \triangleq \langle \sigma(\mathbf{z})\phi(\mathbf{z}), \sigma(\mathbf{z}')\phi(\mathbf{z}')\rangle_\mathcal{H} = \sigma(\mathbf{z})\sigma(\mathbf{z}')\exp(-0.5\|\mathbf{z}-\mathbf{z}'\|_2^2)$, by definition of RKHS. By assuming that the prior variances of $s_\mathbf{z}$ for all $\mathbf{z} \in \mathcal{I}$ are identical and equal to some constant ζ^2 (i.e., $\sigma(\mathbf{z}) = \zeta > 0$), $\mathrm{cov}[s_\mathbf{z}, s_{\mathbf{z}'}] = \zeta^2\exp(-0.5\|\mathbf{z}-\mathbf{z}'\|_2^2)$ which is independent of $\boldsymbol{\theta}$. Consequently, the prior covariance matrix $\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}} \triangleq (\mathrm{cov}[s_\mathbf{z}, s_{\mathbf{z}'}])_{\mathbf{z},\mathbf{z}'\in\mathcal{I}}$ of the inducing output variables $\mathbf{s}_\mathcal{I} \triangleq (s_\mathbf{z})^\top_{\mathbf{z}\in\mathcal{I}}$ is independent of $\boldsymbol{\theta}$. Furthermore, the cross-covariance matrix $\mathbf{K}_{\mathcal{D}\mathcal{I}} \triangleq (\mathrm{cov}[f_\mathbf{x}, s_\mathbf{z}])_{\mathbf{x}\in\mathcal{D},\mathbf{z}\in\mathcal{I}}$ between the latent outputs $\mathbf{f}_\mathcal{D}$ for some set $\mathcal{D}$ of training inputs and the inducing outputs $\mathbf{s}_\mathcal{I}$ can be computed analytically using the definition of RKHS: $\mathrm{cov}[f_\mathbf{x}, s_\mathbf{z}] = \langle \sigma(\mathbf{x})\phi(\boldsymbol{\Lambda}\mathbf{x}), \sigma(\mathbf{z})\phi(\mathbf{z})\rangle_\mathcal{H} = \zeta\sigma(\mathbf{x})\exp(-0.5\|\boldsymbol{\Lambda}\mathbf{x} - \mathbf{z}\|_2^2)$. Like many existing GP models, the prior variances of $f_\mathbf{x}$ for all $\mathbf{x} \in \mathcal{X}$ are assumed to be identical and equal to a signal variance hyperparameter σ_f^2 (i.e., $\sigma(\mathbf{x}) = \sigma_f$) for tractable learning, hence circumventing the need to learn an infinite number of prior variance hyperparameters. The resulting representation of the GP model from the reparameterization trick will allow the optimal variational distributions of inducing outputs $\mathbf{s}_\mathcal{I}$ and hyperparameters $\boldsymbol{\theta}$ (hence the predictive distribution) to be tractably derived for a family of VBSGPR models, as discussed in Section IV.

Remark 1: The definition of $\mathcal{I}$ seems to suggest its construction by first selecting the inducing inputs $\mathcal{U}$ and then rotating them via $\boldsymbol{\Lambda}$, which is not possible since $\boldsymbol{\Lambda}$ is not known *a priori*. However, as shall be discussed in Section IV, it is possible to first select $\mathcal{I}$ and then optimize the variational distribution of $\boldsymbol{\Lambda}$, which has an effect of optimizing the distribution of inducing inputs $\mathcal{U}$ in original input space $\mathcal{X}$.

Remark 2: Let $\mathcal{Z} \triangleq \{\boldsymbol{\Lambda}\mathbf{x}\}_{\mathbf{x}\in\mathcal{X}}$. By setting the (identical) prior variances of $s_\mathbf{z}$ for all $\mathbf{z} \in \mathcal{Z}$ to unity (i.e., $\zeta = 1$), $\{s_\mathbf{z}\}_{\mathbf{z}\in\mathcal{Z}}$ denote a *standard* GP with unit signal variance and length-scales [19], which is a special case of our representation of the GP model here. Then, $f_\mathbf{x} = \sigma_f s_{\boldsymbol{\Lambda}\mathbf{x}}$ for all $\mathbf{x} \in \mathcal{X}$.

IV. Variational Bayesian SGPR Models

Using our representation of the GP model defined above (Section III), the predictive distribution (1) of a Bayesian SGPR model can be slightly modified to $p(f_{\mathbf{x}^*}|\mathbf{y}_\mathcal{D}) = \int p(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}_B}, \mathbf{s}_\mathcal{I}, \boldsymbol{\theta})\; p(\mathbf{s}_\mathcal{I}, \boldsymbol{\theta}|\mathbf{y}_\mathcal{D})\; \mathrm{d}\mathbf{s}_\mathcal{I}\; \mathrm{d}\boldsymbol{\theta}$ such that deriving the posterior $p(\mathbf{s}_\mathcal{I}, \boldsymbol{\theta}|\mathbf{y}_\mathcal{D}) = p(\mathbf{y}_\mathcal{D}, \mathbf{s}_\mathcal{I}, \boldsymbol{\theta})/p(\mathbf{y}_\mathcal{D})$ requires computing the likelihood:

$$p(\mathbf{y}_\mathcal{D}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\int p(\mathbf{y}_\mathcal{D}|\mathbf{f}_\mathcal{D})p(\mathbf{f}_\mathcal{D}|\mathbf{s}_\mathcal{I}, \boldsymbol{\theta})p(\mathbf{s}_\mathcal{I})\mathrm{d}\mathbf{f}_\mathcal{D}\mathrm{d}\mathbf{s}_\mathcal{I}\right] \qquad (2)$$

where $p(\boldsymbol{\theta}) \triangleq \mathcal{N}(\mathbf{1}, \mathrm{diag}[0.1])$, $p(\mathbf{s}_\mathcal{I}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}})$, $p(\mathbf{y}_\mathcal{D}|\mathbf{f}_\mathcal{D}) = \mathcal{N}(\mathbf{f}_\mathcal{D}, \mathbf{C}_{\mathcal{D}\mathcal{D}})$, and

$$p(\mathbf{f}_\mathcal{D}|\mathbf{s}_\mathcal{I}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{K}_{\mathcal{D}\mathcal{I}}\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1}\mathbf{s}_\mathcal{I}, \mathbf{K}_{\mathcal{D}\mathcal{D}} - \mathbf{K}_{\mathcal{D}\mathcal{I}}\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1}\mathbf{K}_{\mathcal{I}\mathcal{D}}) \qquad (3)$$

such that $\mathbf{K}_{\mathcal{D}\mathcal{I}}$ is previously defined in Section III and $\mathbf{K}_{\mathcal{I}\mathcal{D}} = \mathbf{K}_{\mathcal{D}\mathcal{I}}^\top$. However, the integration in (2) (and hence $p(\mathbf{s}_\mathcal{I}, \boldsymbol{\theta}|\mathbf{y}_\mathcal{D})$) cannot be evaluated in closed form. To resolve this, instead of using exact inference, we adopt variational inference to approximate the posterior distribution $p(\mathbf{f}_\mathcal{D}, \mathbf{s}_\mathcal{I}, \boldsymbol{\theta}|\mathbf{y}_\mathcal{D}) =$

$p(\mathbf{f}_\mathcal{D}|\mathbf{s}_\mathcal{I},\boldsymbol{\theta})\ p(\mathbf{s}_\mathcal{I},\boldsymbol{\theta}|\mathbf{y}_\mathcal{D})$ with a factorized variational distribution $q(\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta}) \triangleq p(\mathbf{f}_\mathcal{D}|\mathbf{s}_\mathcal{I},\boldsymbol{\theta})\ q(\mathbf{s}_\mathcal{I})\ q(\boldsymbol{\theta})$ where $p(\mathbf{f}_\mathcal{D}|\mathbf{s}_\mathcal{I},\boldsymbol{\theta})$ is the exact training conditional (3), $q(\mathbf{s}_\mathcal{I}) \triangleq \mathcal{N}(\mathbf{m},\mathbf{S})$, $q(\boldsymbol{\theta}) \triangleq q(\boldsymbol{\Lambda})\ q(\sigma_f)$, $q(\boldsymbol{\Lambda}) \triangleq \prod_{i=1}^{d}\mathcal{N}(\lambda_i|\nu_i,\xi_i)$ with $\boldsymbol{\nu} \triangleq (\nu_1,\ldots,\nu_d)^\top$ and $\boldsymbol{\Xi} \triangleq \mathrm{diag}[\xi_1,\ldots,\xi_d]$, and $q(\sigma_f) \triangleq \mathcal{N}(\alpha,\beta)$. Then, the log-marginal likelihood $\log p(\mathbf{y}_\mathcal{D})$ can be decomposed into a sum of its variational lower bound $\mathcal{L}(q)$ and the nonnegative KL distance between the variational distribution $q(\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta})$ and the posterior distribution $p(\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta}|\mathbf{y}_\mathcal{D})$, the latter of which quantifies the gap between $\log p(\mathbf{y}_\mathcal{D})$ and $\mathcal{L}(q)$, that is,

$$\log p(\mathbf{y}_\mathcal{D}) = \mathcal{L}(q) + \mathrm{KL}(q(\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta})||p(\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta}|\mathbf{y}_\mathcal{D})), \quad (4)$$

as derived in Appendix A of [34] where

$$\mathcal{L}(q) \triangleq \int q(\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta}) \log \frac{p(\mathbf{y}_\mathcal{D},\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta})}{q(\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta})}\ \mathrm{d}\mathbf{f}_\mathcal{D}\ \mathrm{d}\mathbf{s}_\mathcal{I}\ \mathrm{d}\boldsymbol{\theta}. \quad (5)$$

Remark 3: The likelihood term $p(\mathbf{y}_\mathcal{D})$ (2) in (4) is a constant with respect to $q(\mathbf{s}_\mathcal{I})$ and $q(\boldsymbol{\theta})$ (specifically, their parameters $\mathbf{m},\mathbf{S},\boldsymbol{\nu},\boldsymbol{\Xi},\alpha,\beta$). Consequently, maximizing $\mathcal{L}(q)$ with respect to $q(\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta})$ is equivalent to minimizing $\mathrm{KL}(q(\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta})||p(\mathbf{f}_\mathcal{D},\mathbf{s}_\mathcal{I},\boldsymbol{\theta}|\mathbf{y}_\mathcal{D}))$. This equivalence, however, does not hold for existing variational SGPR models and their stochastic and distributed variants optimizing point estimates of all hyperparameters, as discussed in Section I.

The variational inference framework of [19] is similar in spirit to the above. However, the framework of [19] assumes i.i.d. observation noises (i.e., $\mathbf{C}_{\mathcal{D}\mathcal{D}} = \sigma_n^2\mathbf{I}$ and $\zeta = 1$) and ignores their correlation, which consequently yields the VBDTC approximation (see Remark 4 later). The challenge remains in investigating whether the other more refined SGPR models spanned by the unifying view of [2] (e.g., FITC, PITC, PIC) are amenable to such a variational Bayesian treatment since they have been empirically demonstrated [14], [15] to give better predictive performance than DTC.

To address this challenge, our key idea is to relax the strong assumption of i.i.d. observation noises with constant variance σ_n^2 imposed by VBDTC and allow observation noises to be correlated with some structure across the input space, hence being robust to different noise correlation structures and in turn catering to more realistic applications of GP. Interestingly, this results in a noise-robust family of *variational Bayesian SGPR* (VBSGPR) models (e.g., VBDTC, VBFITC, VBPIC), which we will describe below.

Let $\mathbf{C}_{\mathcal{D}\mathcal{D}} \triangleq \mathrm{blkdiag}[\mathbf{K}^\varepsilon_{\mathcal{D}\mathcal{D}} - \mathbf{K}^\varepsilon_{\mathcal{D}\mathcal{U}}\mathbf{K}^{\varepsilon^{-1}}_{\mathcal{U}\mathcal{U}}\mathbf{K}^\varepsilon_{\mathcal{U}\mathcal{D}}] + \sigma_n^2\mathbf{I}$ be a block-diagonal noise covariance matrix constructed from the B diagonal blocks of $\mathbf{K}^\varepsilon_{\mathcal{D}\mathcal{D}} - \mathbf{K}^\varepsilon_{\mathcal{D}\mathcal{U}}\mathbf{K}^{\varepsilon^{-1}}_{\mathcal{U}\mathcal{U}}\mathbf{K}^\varepsilon_{\mathcal{U}\mathcal{D}} + \sigma_n^2\mathbf{I}$, each of which is a matrix $\mathbf{C}_{\mathcal{D}_i\mathcal{D}_i} \triangleq \mathbf{K}^\varepsilon_{\mathcal{D}_i\mathcal{D}_i} - \mathbf{K}^\varepsilon_{\mathcal{D}_i\mathcal{U}}\mathbf{K}^{\varepsilon^{-1}}_{\mathcal{U}\mathcal{U}}\mathbf{K}^\varepsilon_{\mathcal{U}\mathcal{D}_i} + \sigma_n^2\mathbf{I}$ for $i = 1,\ldots,B$, and $\mathbf{K}^\varepsilon_{\mathcal{D}\mathcal{D}} \triangleq (k^\varepsilon_{\mathbf{x}\mathbf{x}'})_{\mathbf{x},\mathbf{x}'\in\mathcal{D}}$, $\mathbf{K}^\varepsilon_{\mathcal{D}\mathcal{U}} \triangleq (k^\varepsilon_{\mathbf{x}\mathbf{x}'})_{\mathbf{x}\in\mathcal{D},\mathbf{x}'\in\mathcal{U}}$, $\mathbf{K}^\varepsilon_{\mathcal{U}\mathcal{U}} \triangleq (k^\varepsilon_{\mathbf{x}\mathbf{x}'})_{\mathbf{x},\mathbf{x}'\in\mathcal{U}}$, and $\mathbf{K}^\varepsilon_{\mathcal{U}\mathcal{D}} \triangleq \mathbf{K}^{\varepsilon^\top}_{\mathcal{D}\mathcal{U}}$ are matrices with components $k^\varepsilon_{\mathbf{x}\mathbf{x}'}$ defined by a covariance function similar to that used for $k_{\mathbf{x}\mathbf{x}'}$ (Section II-A) but with different hyperparameter values[1]. Our first major result ensues:

Theorem 1: $\mathcal{L}(q)$ in (4) can be analytically evaluated as

$$\begin{aligned}\mathcal{L}(q) = \frac{1}{2}\Big(&2\mathbf{m}^\top\boldsymbol{\Sigma}^{-1}_{\mathcal{I}\mathcal{I}}\boldsymbol{\Omega}_{\mathcal{I}\mathcal{D}}\mathbf{C}^{-1}_{\mathcal{D}\mathcal{D}}\mathbf{y}_\mathcal{D} - \mathbf{m}^\top\mathbf{Q}\mathbf{m} - \mathrm{Tr}[\mathbf{S}\mathbf{Q}]\\ &-\mathrm{Tr}[\mathbf{C}^{-1}_{\mathcal{D}\mathcal{D}}\boldsymbol{\Upsilon}_{\mathcal{D}\mathcal{D}}] + \mathrm{Tr}[\boldsymbol{\Sigma}^{-1}_{\mathcal{I}\mathcal{I}}\boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}}] + \log|\mathbf{S}|\\ &-\boldsymbol{\nu}^\top\boldsymbol{\nu} - \mathrm{Tr}[\boldsymbol{\Xi}] + \log|\boldsymbol{\Xi}| - \alpha^2 - \beta + \log\beta\Big) + \mathrm{const}\end{aligned} \quad (6)$$

where $\mathbf{Q} \triangleq \boldsymbol{\Sigma}^{-1}_{\mathcal{I}\mathcal{I}}\boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}}\boldsymbol{\Sigma}^{-1}_{\mathcal{I}\mathcal{I}} + \boldsymbol{\Sigma}^{-1}_{\mathcal{I}\mathcal{I}}$. More interestingly, using the above expression, it can be shown that $\mathcal{L}(q)$ is maximized at $q^*(\mathbf{s}_\mathcal{I}) = \mathcal{N}(\mathbf{m}^*,\mathbf{S}^*)$ where

$$\begin{aligned}\mathbf{m}^* &\triangleq \boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}(\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}} + \boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}})^{-1}\boldsymbol{\Omega}_{\mathcal{I}\mathcal{D}}\mathbf{C}^{-1}_{\mathcal{D}\mathcal{D}}\mathbf{y}_\mathcal{D}\ ,\\ \mathbf{S}^* &\triangleq \boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}(\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}} + \boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}})^{-1}\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}\end{aligned} \quad (7)$$

such that $\boldsymbol{\Omega}_{\mathcal{I}\mathcal{D}} \triangleq \mathbb{E}_{q(\boldsymbol{\theta})}[\mathbf{K}_{\mathcal{I}\mathcal{D}}]$, $\boldsymbol{\Upsilon}_{\mathcal{D}\mathcal{D}} \triangleq \mathbb{E}_{q(\boldsymbol{\theta})}[\mathbf{K}_{\mathcal{D}\mathcal{D}}]$, $\boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}} \triangleq \mathbb{E}_{q(\boldsymbol{\theta})}[\mathbf{K}_{\mathcal{I}\mathcal{D}}\mathbf{C}^{-1}_{\mathcal{D}\mathcal{D}}\mathbf{K}_{\mathcal{D}\mathcal{I}}]$, and const absorbs all terms indep. of $\mathbf{m}$, $\mathbf{S}$, $\boldsymbol{\nu}$, $\boldsymbol{\Xi}$, α, β.

Its proof is in Appendix B of [34]. Appendix C of [34] gives the closed-form expressions of $\boldsymbol{\Omega}_{\mathcal{I}\mathcal{D}}$, $\boldsymbol{\Upsilon}_{\mathcal{D}\mathcal{D}}$, and $\boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}}$.

Remark 4: Note that $q^*(\mathbf{s}_\mathcal{I})$ in Theorem 1 closely resembles that of PIC and PITC (see eqs. 64 and 65 in Appendix D.1.1 of [14]) except for the expectation over hyperparameters $\boldsymbol{\theta}$ due to the variational Bayesian treatment. So, we call them VBPIC and VBPITC, respectively. By setting $B = |\mathcal{D}|$, $\mathbf{C}_{\mathcal{D}\mathcal{D}}$ becomes a diagonal matrix to give VBFIC and VBFITC. When $\mathbf{C}_{\mathcal{D}\mathcal{D}} = \sigma_n^2\mathbf{I}$, $q^*(\mathbf{s}_\mathcal{I})$ (7) resembles that of DTC (see eqs. 68 and 69 in Appendix D.1.3 of [14]) except for the expectation over $\boldsymbol{\theta}$ due to the variational Bayesian treatment and coincides with that in Appendix B.1 of [19]. So, we refer to it as VBDTC.

Remark 5: In the non-Bayesian setting of the hyperparameters, it has been previously established that the predictive distribution of FITC can be reproduced as a direct result of applying either variational inference [8] with $\mathbf{C}_{\mathcal{D}\mathcal{D}} = \mathrm{diag}[\mathbf{K}_{\mathcal{D}\mathcal{D}} - \mathbf{K}_{\mathcal{D}\mathcal{U}}\mathbf{K}^{-1}_{\mathcal{U}\mathcal{U}}\mathbf{K}_{\mathcal{U}\mathcal{D}}] + \sigma_n^2\mathbf{I}$ or expectation propagation [35] on the FGPR model. Our derivation of VBFITC is in fact similar in spirit to that of [8] except for our variational Bayesian treatment of its hyperparameters. On the other hand, it is unclear whether FITC's equivalent EP derivation in [35] can be easily extended to incorporate a Bayesian treatment of its hyperparameters.

V. Stochastic Optimization

The VBDTC approximation [19] has explicitly plugged the optimal $q^*(\mathbf{s}_\mathcal{I})$ (see Theorem 1) into $\mathcal{L}(q)$ (6) and reduced it to $\mathcal{L}(q)$ (11) in Appendix B of [34]. Given $\mathcal{L}(q)$ (11), the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Xi}$ of $q(\boldsymbol{\Lambda})$ and α and β of $q(\sigma_f)$ can be optimized via gradient ascent. However, evaluating

[1]We do not assign any prior over the hyperparameters of $k^\varepsilon_{\mathbf{x}\mathbf{x}'}$ and the noise variance σ_n^2. Instead, they are treated as parameters optimized to maximize $\mathcal{L}(q)$ via stochastic gradient ascent [12]. In our experiments, we observe that even if we set the hyperparameters of $k^\varepsilon_{\mathbf{x}\mathbf{x}'}$ by hand, the predictive performance does not vary much and our VBPIC approximation can significantly outperform the state-of-the-art variational SGPR models and their stochastic and distributed variants. A Bayesian treatment of these hyperparameters is highly non-trivial due to a complication similar to that discussed in Section II-C and will be investigated in our future work.

the exact gradients $\partial\mathcal{L}/\partial\boldsymbol{\nu}$, $\partial\mathcal{L}/\partial\boldsymbol{\Xi}$, $\partial\mathcal{L}/\partial\alpha$ and $\partial\mathcal{L}/\partial\beta$ incur $\mathcal{O}(|\mathcal{D}||\mathcal{I}|^2)$ time, which scales poorly in the data size $|\mathcal{D}|$. To overcome the above issue of scalability, we utilize stochastic gradient ascent updates instead of exact ones, which requires the stochastic gradients to be unbiased estimators of the exact gradients to guarantee convergence. The key idea is to iteratively compute the stochastic gradients by randomly sampling a single or few mini-batches of data from the dataset (i.e., comprising B disjoint mini-batches) whose incurred time per iteration is independent of data size $|\mathcal{D}|$. To achieve this, an important requirement is the decomposability of $\mathcal{L}(q)$ (11) into a summation of B terms, each of which is associated with a mini-batch $(\mathcal{D}_i, \mathbf{y}_{\mathcal{D}_i})$ of data of size $|\mathcal{D}_i| = \mathcal{O}(|\mathcal{I}|)$ that can be exploited for computing the stochastic gradients. Unfortunately, $\mathcal{L}(q)$ (11) is not decomposable due to its $(\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}} + \boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}})^{-1}$ term. To remedy this, we do not plug $q^*(\mathbf{s}_\mathcal{I})$ (7) into $\mathcal{L}(q)$ (6) to yield (11) but instead jointly treat $q(\mathbf{s}_\mathcal{I})$, $q(\boldsymbol{\Lambda})$, and $q(\sigma_f)$ as variational parameters, which enables the decomposability of $\mathcal{L}(q)$ (6):

Corollary 1: $\mathcal{L}(q)$ (6) (Theorem 1) can be decomposed into

$$\mathcal{L}(q) = \sum_{i=1}^{B} \mathcal{L}_i(q) + \frac{1}{2}\Big(-\mathbf{m}^\top \boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1}\mathbf{m} - \mathrm{Tr}[\mathbf{S}\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1}] + \log|\mathbf{S}| - \boldsymbol{\nu}^\top\boldsymbol{\nu} - \mathrm{Tr}[\boldsymbol{\Xi}] + \log|\boldsymbol{\Xi}| - \alpha^2 - \beta + \log\beta\Big) + \mathrm{const}\ ,$$

$$\mathcal{L}_i(q) \triangleq \frac{1}{2}\Big(2\mathbf{m}^\top \boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1}\boldsymbol{\Omega}_{\mathcal{I}\mathcal{D}_i}\mathbf{C}_{\mathcal{D}_i\mathcal{D}_i}^{-1}\mathbf{y}_{\mathcal{D}_i} - \mathbf{m}^\top\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1}\boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}}^{i}\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1}\mathbf{m} - \mathrm{Tr}[\mathbf{S}\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1}\boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}}^{i}\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1}] - \mathrm{Tr}[\mathbf{C}_{\mathcal{D}_i\mathcal{D}_i}^{-1}\boldsymbol{\Upsilon}_{\mathcal{D}_i\mathcal{D}_i}] + \mathrm{Tr}[\boldsymbol{\Sigma}_{\mathcal{I}\mathcal{I}}^{-1}\boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}}^{i}]\Big)$$

where $\boldsymbol{\Psi}_{\mathcal{I}\mathcal{I}}^{i} \triangleq \mathbb{E}_{q(\boldsymbol{\theta})}[\mathbf{K}_{\mathcal{I}\mathcal{D}_i}\mathbf{C}_{\mathcal{D}_i\mathcal{D}_i}^{-1}\mathbf{K}_{\mathcal{D}_i\mathcal{I}}]$.

Our main result below exploits the decomposability of $\mathcal{L}(q)$ in Corollary 1 to derive stochastic gradients $\partial\widetilde{\mathcal{L}}/\partial\mathbf{m}$, $\partial\widetilde{\mathcal{L}}/\partial\mathbf{S}$, $\partial\widetilde{\mathcal{L}}/\partial\boldsymbol{\nu}$, $\partial\widetilde{\mathcal{L}}/\partial\boldsymbol{\Xi}$, $\partial\widetilde{\mathcal{L}}/\partial\alpha$, and $\partial\widetilde{\mathcal{L}}/\partial\beta$ that are unbiased estimators of their respective exact gradients, which is the key contribution of our work in this paper:

Theorem 2: Let $\mathcal{S}$ be a set of i.i.d. samples drawn from a uniform distribution over $\{1, 2, \ldots, B\}$. Construct the stochastic gradients $\partial\widetilde{\mathcal{L}}/\partial\mathbf{m}$, $\partial\widetilde{\mathcal{L}}/\partial\mathbf{S}$, $\partial\widetilde{\mathcal{L}}/\partial\boldsymbol{\nu}$, $\partial\widetilde{\mathcal{L}}/\partial\boldsymbol{\Xi}$, $\partial\widetilde{\mathcal{L}}/\partial\alpha$, and $\partial\widetilde{\mathcal{L}}/\partial\beta$ using the mini-batches $(\mathcal{D}_s, \mathbf{y}_{\mathcal{D}_s})$ for $s \in \mathcal{S}$ and current estimates of $(\mathbf{m}, \mathbf{S}, \boldsymbol{\nu}, \boldsymbol{\Xi}, \alpha, \beta)$ according to (12) in Appendix D of [34]. Then, $\mathbb{E}[\partial\widetilde{\mathcal{L}}/\partial\mathbf{m}] = \partial\mathcal{L}/\partial\mathbf{m}$, $\mathbb{E}[\partial\widetilde{\mathcal{L}}/\partial\mathbf{S}] = \partial\mathcal{L}/\partial\mathbf{S}$, $\mathbb{E}[\partial\widetilde{\mathcal{L}}/\partial\boldsymbol{\nu}] = \partial\mathcal{L}/\partial\boldsymbol{\nu}$, $\mathbb{E}[\partial\widetilde{\mathcal{L}}/\partial\boldsymbol{\Xi}] = \partial\mathcal{L}/\partial\boldsymbol{\Xi}$, $\mathbb{E}[\partial\widetilde{\mathcal{L}}/\partial\alpha] = \partial\mathcal{L}/\partial\alpha$, and $\mathbb{E}[\partial\widetilde{\mathcal{L}}/\partial\beta] = \partial\mathcal{L}/\partial\beta$.

Its proof is in Appendix D of [34].

Remark 6: The stochastic gradients (Theorem 2) can be computed in closed form in $\mathcal{O}(|\mathcal{S}||\mathcal{I}|^3)$ time per iteration that reduces to $\mathcal{O}(|\mathcal{I}|^3)$ time by setting $|\mathcal{S}| = 1$ in our experiments. So, if the number of iterations of stochastic gradient ascent needed for convergence is much smaller than $t\min(|\mathcal{D}|/|\mathcal{I}|, B)$ where t is the required number of iterations of exact gradient ascent, then our stochastic variants achieve a huge speedup over the corresponding VBSGPR models.

VI. Bayesian Prediction with VBSGPR Models

Recall that the predictive distribution $p(f_{\mathbf{x}^*}|\mathbf{y}_\mathcal{D})$ is computationally intractable. We thus approximate it by $q(f_{\mathbf{x}^*}|\mathbf{y}_\mathcal{D}) = \int q(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}_B}, \mathbf{s}_\mathcal{I}, \boldsymbol{\Lambda}, \sigma_f)\ q^+(\mathbf{s}_\mathcal{I})\ q^+(\boldsymbol{\Lambda})\ q^+(\sigma_f)\ \mathrm{d}\mathbf{s}_\mathcal{I}\ \mathrm{d}\boldsymbol{\Lambda}\ \mathrm{d}\sigma_\mathbf{f}$ where $q^+(\mathbf{s}_\mathcal{I}) \triangleq \mathcal{N}(\mathbf{m}^+, \mathbf{S}^+)$, $q^+(\boldsymbol{\Lambda}) \triangleq \prod_{i=1}^{d}\mathcal{N}(\nu_i^+, \xi_i^+)$ with $\boldsymbol{\nu}^+ \triangleq (\nu_1^+, \ldots, \nu_d^+)^\top$ and $\boldsymbol{\Xi}^+ \triangleq \mathrm{diag}[\xi_1^+, \ldots, \xi_d^+]$, and $q(\sigma_f) \triangleq \mathcal{N}(\alpha^+, \beta^+)$ are obtained from the stochastic gradient ascent updates (Section V). Note that $q(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}_B}, \mathbf{s}_\mathcal{I}, \boldsymbol{\Lambda}, \sigma_f)$ is set to $p(f_{\mathbf{x}^*}|\mathbf{s}_\mathcal{I}, \boldsymbol{\Lambda}, \sigma_f)$ for the VBPITC, VBFIC, VBFITC, and VBDTC models and to $p(f_{\mathbf{x}^*}|\mathbf{y}_{\mathcal{D}_B}, \mathbf{s}_\mathcal{I}, \boldsymbol{\Lambda}, \sigma_f)$ for the VBPIC model. Although the predictive distribution $q(f_{\mathbf{x}^*}|\mathbf{y}_\mathcal{D})$ is not Gaussian, its predictive mean $\mu_{\mathbf{x}^*|\mathcal{D}}$ and variance $\sigma^2_{\mathbf{x}^*|\mathcal{D}}$ can be computed analytically for VBPITC, VBFIC, VBFITC, and VBDTC and via sampling for VBPIC, as derived in Appendix F of [34]. Their respective predictive means $\mu_{\mathbf{x}^*|\mathcal{D}}$ closely resemble that of PITC, FIC, FITC, DTC, and PIC (see eqs. 84 and 86 in Appendix D.4 of [14]) except for the expectations over $\boldsymbol{\Lambda}$ and σ_f due to the variational Bayesian treatment. Their predictive variances $\sigma^2_{\mathbf{x}^*|\mathcal{D}}$ are also similar except for the expectations over $\boldsymbol{\Lambda}$ and σ_f and an additional positive term arising from the uncertainty of $\boldsymbol{\Lambda}$ and σ_f.

VII. Experiments and Discussion

This section empirically evaluates the predictive performance and time efficiency of the stochastic variants, denoted by VBDTC+, VBFITC+, and VBPIC+, of our VBSGPR models (respectively, VBDTC, VBFITC, and VBPIC). We will first use the small AIMPEAK dataset [3] on traffic speeds of size 41850 to evaluate the convergence of the variational distributions $q^+(\mathbf{s}_\mathcal{I})$ and $q^+(\boldsymbol{\Lambda}, \sigma_f)$ induced by our stochastic variants VBDTC+, VBFITC+, and VBPIC+ to, respectively, $q(\mathbf{s}_\mathcal{I})$ and $q(\boldsymbol{\Lambda}, \sigma_f)$ induced by VBDTC [19], VBFITC, and VBPIC performing exact gradient ascent updates via *scaled conjugate gradient* (SCG). To do this, we use the KL distance metric to measure the distance between the variational distributions obtained from the stochastic vs. exact gradient ascent.

Then, using the real-world TWITTER dataset on buzz events of size 583250 and AIRLINE dataset [12] on flight delays of size 2055733, we will compare the performance of the stochastic variants of our VBSGPR models with that of the state-of-the-art GP models such as the stochastic variants of variational DTC (SVIGP) [12] and variational PIC (PIC+) [14], distributed variational DTC (Dist-VGP) [11], and rBCM [36], all of which find point estimates of hyperparameters. Such a comparison will demonstrate the benefits of adopting a variational Bayesian treatment of the hyperparameters by our VBSGPR models. We will also compare the performance of our stochastic VBSGPR models with that of the stochastic variant of *variational Bayesian sparse spectrum GPR* (VSSGPR) model [31]. To evaluate their predictive performance, we use the *root mean square error* (RMSE) metric: $\sqrt{\sum_{\mathbf{x}^*\in\mathcal{T}}(y_{\mathbf{x}^*} - \mu_{\mathbf{x}^*|\mathcal{D}})^2/|\mathcal{T}|}$ and the *mean negative log probability* (MNLP) metric: $0.5\sum_{\mathbf{x}^*\in\mathcal{T}}\{(y_{\mathbf{x}^*} - \mu_{\mathbf{x}^*|\mathcal{D}})^2/\sigma^2_{x^*|\mathcal{D}} + \log(2\pi\sigma^2_{x^*|\mathcal{D}})\}/|\mathcal{T}|$ where $\mathcal{T}$ denotes a set of test inputs.

All datasets are modeled using GPs whose prior covariance is defined by the squared exponential covariance function defined in Section II-A. All experiments are run on a Linux

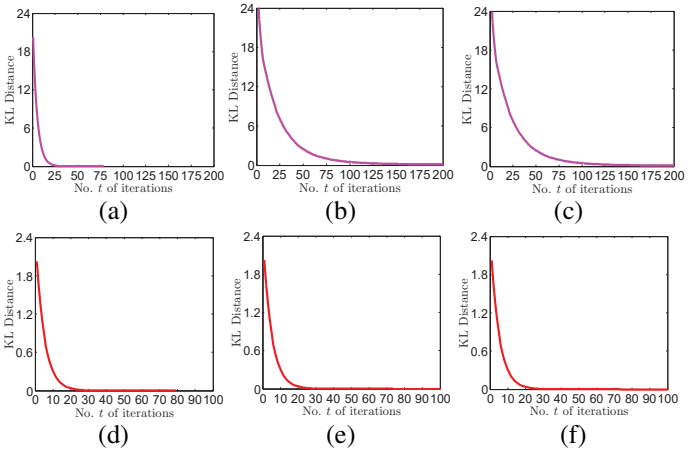

Fig. 1. Graphs of KL distance $\mathrm{KL}(q(\mathbf{s}_{\mathcal{I}})||q^{+}(\mathbf{s}_{\mathcal{I}}))$ of (a) VBDTC+ to VBDTC, (b) VBFITC+ to VBFITC, (c) VBPIC+ to VBPIC, and $\mathrm{KL}((q(\mathbf{\Lambda},\sigma_f)||q^{+}(\mathbf{\Lambda},\sigma_f))$ of (d) VBDTC+ to VBDTC, (e) VBFITC+ to VBFITC, (f) VBPIC+ to VBPIC vs. no. t of iterations for AIMPEAK dataset.

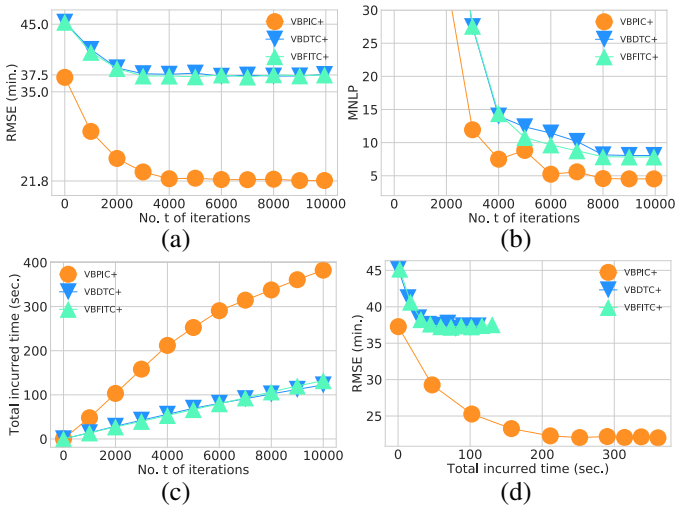

Fig. 2. Graphs of (a) RMSE, (b) MNLP, and (c) total incurred time vs. number t of iterations, and (d) graphs of RMSE vs. total incurred time of VBDTC+, VBFITC+, and VBPIC+ for the AIRLINE dataset.

system with Intel® Xeon® E5-2683 CPU at 2.1GHz with 256GB memory.

A. Empirical Convergence of Stochastic VBSGPR Models

The AIMPEAK dataset [3] of size 41850 comprises traffic speeds (km/h) along 775 road segments of an urban road network during morning peak hours on April 20, 2011. Each input (i.e., road segment) denotes a 5D feature vector of length, number of lanes, speed limit, direction, and time, the last of which comprises 54 five-minute time slots. The output corresponds to traffic speed. We randomly select training data of size 1000, which is partitioned into $B = 10$ mini-batches, and 50 inducing inputs from the inputs of the training data.

Figs. 1a-1c (Figs. 1d-1f) shows results of the KL distance $\mathrm{KL}(q(\mathbf{s}_{\mathcal{I}})||q^{+}(\mathbf{s}_{\mathcal{I}}))$ ($\mathrm{KL}(q(\mathbf{\Lambda},\sigma_f)||q^{+}(\mathbf{\Lambda},\sigma_f))$) of $q^{+}(\mathbf{s}_{\mathcal{I}})$ to $q(\mathbf{s}_{\mathcal{I}})$ ($q^{+}(\mathbf{\Lambda},\sigma_f)$ to $q(\mathbf{\Lambda},\sigma_f)$) averaged over 5 random selections of training data and mini-batch sequences with an increasing number t of iterations. It can be observed that the variational distributions $q^{+}(\mathbf{s}_{\mathcal{I}})$ and $q^{+}(\mathbf{\Lambda},\sigma_f)$ induced by VBDTC+, VBFITC+, and VBPIC+ converge rapidly to, respectively, $q(\mathbf{s}_{\mathcal{I}})$ and $q(\mathbf{\Lambda},\sigma_f)$ induced by VBDTC, VBFITC, and VBPIC, thus corroborating our theoretical results in Section V. From Figs. 1a-1c, it can also be observed that $q^{+}(\mathbf{s}_{\mathcal{I}})$ induced by VBDTC+ converges faster to $q(\mathbf{s}_{\mathcal{I}})$ than that by VBFITC+ and VBPIC+, which can be explained by its much simpler noise structure by assuming i.i.d. observation noises with constant variance σ_n^2.

B. Empirical Evaluation on AIRLINE and TWITTER Datasets

The TWITTER dataset contains 583250 instances of buzz events on Twitter. The input denotes a relatively large 77D feature vector described at http://ama.liglab.fr/datasets/buzz/, which makes this dataset suitable for evaluating robustness to overfitting. The output is the popularity of the instance's topic. The massive benchmark AIRLINE dataset [12] contains 2055733 records of information about every commercial flight in the USA from January to April 2008. The input denotes an 8D feature vector of age of the aircraft (no. of years in service), travel distance (km), airtime, departure and arrival time (min.) as well as day of the week, day of month, and month. The output is the delay time (min.) of the flight. For each dataset, 5% is randomly selected and set aside as test data. The remaining data is used as training data and partitioned into $B = 1000$ mini-batches using k-means (i.e., $k = B$). We randomly select 100 inducing inputs from the inputs of the training data.

Figs. 2a and 2b show results of RMSE and MNLP achieved by the stochastic variants of our VBSGPR models averaged over 5 random selections of 5% test data and mini-batch sequences with an increasing number t of iterations for the AIRLINE dataset. It can be observed that VBPIC+ (RMSE of 21.87 min. and MNLP of 4.53) achieves considerably better predictive performance than VBFITC+ (RMSE of 37.05 min. and MNLP of 7.84) and VBDTC+ (RMSE of 37.55 min. and MNLP of 8.06). To explain this, VBFITC+ and VBDTC+ have both imposed a strong assumption of independently distributed observation noises. In contrast, VBPIC+ caters to correlation of observation noises within each mini-batch of data (Sections IV and V), hence modeling and predicting real-world datasets with correlated noises better. Furthermore, unlike VBFITC+ and VBDTC+, VBPIC+ does not assume conditional independence between the training and test outputs given the inducing outputs in its test conditional.

Fig. 2c exhibits a near-linear increase in total incurred time with an increasing number t of iterations for VBDTC+, VBFITC+, and VBPIC+. Our experiments reveal that VBDTC+, VBFITC+, and VBPIC+ incur, respectively, an average of 0.0122, 0.0132, and 0.038 seconds per iteration of stochastic gradient ascent update. Fig. 2d shows that VBPIC+ can achieve a more superior trade-off between predictive performance vs. time efficiency than VBDTC+ and VBFITC+.

Figs. 3a and 3b show results of RMSE and MNLP achieved by the stochastic variants of our VBSGPR models averaged

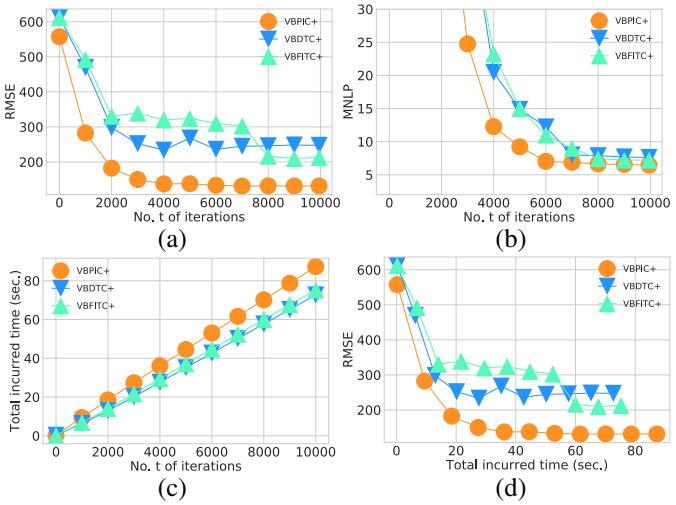

Fig. 3. Graphs of (a) RMSE, (b) MNLP, and (c) total incurred time vs. number t of iterations, and (d) graphs of RMSE vs. total incurred time of VBDTC+, VBFITC+, and VBPIC+ for the TWITTER dataset.

over 5 random selections of 5% test data and mini-batch sequences with an increasing number t of iterations for the TWITTER dataset. The observations are similar to that for the AIRLINE dataset: It can be observed that VBPIC+ (RMSE of 131.46 and MNLP of 6.45) achieves significantly better predictive performance than VBFITC+ (RMSE of 212.67 and MNLP of 7.21) and VBDTC+ (RMSE of 247.38 and MNLP of 7.69); this can be explained by the same reasons as that discussed previously for the AIRLINE dataset.

Fig. 3c also exhibits a linear increase in total incurred time with an increasing number t of iterations for VBDTC+, VBFITC+, and VBPIC+. Our experiments reveal that VBDTC+, VBFITC+, and VBPIC+ incur, respectively, an average of 0.0073, 0.0075, and 0.0087 seconds per iteration of stochastic gradient ascent update, which are shorter than that for the AIRLINE dataset due to a smaller mini-batch size. Fig. 3d reveals that VBPIC+ can similarly achieve the best trade-off between predictive performance vs. time efficiency.

Table I compares the predictive performance (RMSEs) achieved by state-of-the-art GP models for the AIRLINE and TWITTER datasets. It can be observed that our VBPIC+ significantly outperforms state-of-the-art SVIGP, Dist-VGP, rBCM, and PIC+, which find point estimates of hyperparameters, and VSSGPR due to its restrictive assumption, as discussed in Section I. In contrast, our VBPIC+ assumes a variational Bayesian treatment of its hyperparameters, thus achieving robustness to overfitting due to Bayesian model selection, as demonstrated later. Unlike VSSGPR, VBPIC+ does not assume conditional independence between the training and test outputs in its test conditional.

Fig. 4 shows results of RMSEs achieved by our VBPIC+ with an increasing number t of iterations and varying sample sizes for computing its predictive mean (Section VI). Note that a sample size of 1 reduces VBPIC to PIC that treats its sampled hyperparameters as a point estimate. By increasing the sample size, it can be observed that VBPIC+ converges

Dataset	SVIGP	Dist-VGP	rBCM	PIC+	VBPIC+	VSSGPR
AIRLINE	39.53	35.30	34.40	24.9	**21.87**	38.95
TWITTER	–	–	–	190.2	**131.4**	585.9

TABLE I
RMSE ACHIEVED BY VBPIC+ AND STATE-OF-THE-ART GP MODELS FOR AIRLINE AND TWITTER DATASETS. THE RESULTS OF PIC+ AND VSSGPR ARE OBTAINED USING THEIR GITHUB CODES. THE RESULTS OF DIST-VGP AND RBCM ARE TAKEN FROM THEIR RESPECTIVE PAPERS AND THAT OF SVIGP IS REPORTED IN [14]. THEY ARE ALL BASED ON THE SAME SETTINGS OF TRAINING/TEST DATA SIZES = 2M/100K (554K/29K) FOR THE AIRLINE (TWITTER) DATASET.

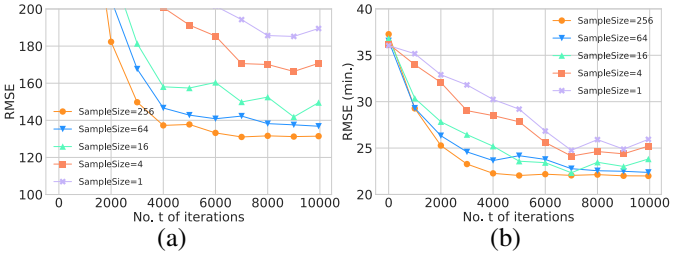

Fig. 4. Graphs of RMSEs of VBPIC+ vs. number t of iterations with varying sampling sizes for computing its predictive mean for the (a) TWITTER and (b) AIRLINE datasets.

faster to a lower RMSE using less iterations due to its Bayesian model selection/averaging, thus demonstrating its increasing robustness to overfitting.

Fig. 5 displays the 95% confidence intervals (mean $\nu_i^+ \pm 2 \times$ standard deviation $(\xi_i^+)^{1/2}$) for inverted length-scale hyperparameters λ_i for $i = 1, \ldots, d$ after $t = 10000$ iterations for the TWITTER ($d = 77$ normalized input dimensions) and AIRLINE ($d = 8$ normalized input dimensions) datasets. It can be observed that the confidence intervals are generally wider (i.e., larger uncertainty of $\lambda_1, \ldots, \lambda_d$) for the TWITTER dataset than for the AIRLINE dataset. To confirm this, we measure the *mean log variance* (MLV) $\sum_{i=1}^{d} \log \xi_i^+ / d$ of $\lambda_1, \ldots, \lambda_d$ and notice that the TWITTER dataset gives a higher MLV of -4.09 than that for the AIRLINE dataset (i.e., MLV $= -6.55$). So, with a larger uncertainty of $\lambda_1, \ldots, \lambda_d$, their point estimates have a greater tendency to overfit and hence yield a poorer predictive performance, as observed in Fig. 4 (compare the performance gap between sample sizes of 1 vs. 256).

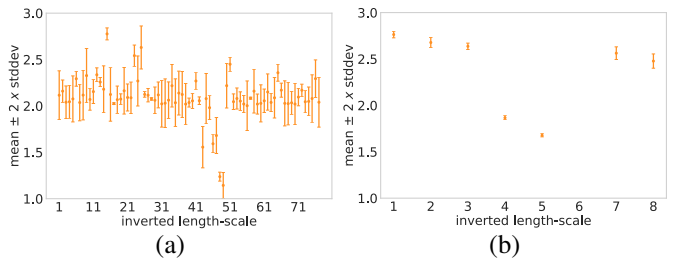

Fig. 5. 95% confidence intervals (mean $\nu_i^+ \pm 2 \times$ standard deviation $(\xi_i^+)^{1/2}$) for inverted length-scale hyperparameters λ_i for $i = 1, \ldots, d$ after $t = 10000$ iterations for the (a) TWITTER ($d = 77$ normalized input dimensions) and (b) AIRLINE ($d = 8$ normalized input dimensions) datasets.

VIII. Conclusion

This paper describes a novel variational inference framework for a family of VBSGPR models (e.g., VBDTC, VBFITC, VBPIC) whose approximations are variationally optimal with respect to the FGPR model enriched with various corresponding correlation structures of the observation noises. Our variational Bayesian treatment of hyperparameters enables our VBSGPR models to mitigate critical issues (e.g., overfitting) which plague existing variational SGPR models that optimize point estimates of hyperparameters (Section I). The stochastic variants of our VBSGPR models can yield good predictive performance fast and improve their predictive performance over time, thus achieving scalability to big data. Empirical evaluation on two real-world datasets reveals that the stochastic variant of our VBPIC can significantly outperform existing state-of-the-art GP models, thus demonstrating its robustness to overfitting due to Bayesian model selection while preserving scalability to big data through stochastic optimization. For our future work, we plan to integrate our proposed framework with that of decentralized/distributed data/model fusion [37]–[41] for collective online learning of a massive number of VBSGPR models.

References

[1] M. Lázaro-Gredilla, J. Quiñonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse spectrum Gaussian process regression," *JMLR*, vol. 11, pp. 1865–1881, 2010.

[2] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *JMLR*, vol. 6, pp. 1939–1959, 2005.

[3] J. Chen, N. Cao, K. H. Low, R. Ouyang, C. K.-Y. Tan, and P. Jaillet, "Parallel Gaussian process regression with low-rank covariance matrix approximations," in *Proc. UAI*, 2013, pp. 152–161.

[4] K. H. Low, J. Yu, J. Chen, and P. Jaillet, "Parallel Gaussian process regression for big data: Low-rank representation meets Markov approximation," in *Proc. AAAI*, 2015, pp. 2821–2827.

[5] K. H. Low, J. Chen, T. N. Hoang, N. Xu, and P. Jaillet, "Recent advances in scaling up Gaussian process predictive models for large spatiotemporal data," in *Proc. DyDESS*, S. Ravela and A. Sandu, Eds. LNCS 8964, Springer, 2015, pp. 167–181.

[6] L. Csató and M. Opper, "Sparse online Gaussian processes," *Neural Comput.*, vol. 14, pp. 641–669, 2002.

[7] N. Xu, K. H. Low, J. Chen, K. K. Lim, and E. B. Özgül, "GP-Localize: Persistent mobile robot localization using online sparse Gaussian process observation model," in *Proc. AAAI*, 2014, pp. 2585–2592.

[8] M. K. Titsias, "Variational model selection for sparse Gaussian process regression," School of Computer Science, University of Manchester, Tech. Rep., 2009.

[9] ——, "Variational learning of inducing variables in sparse Gaussian processes," in *Proc. AISTATS*, 2009, pp. 567–574.

[10] M. Seeger, C. Williams, and N. D. Lawrence, "Fast forward selection to speed up sparse Gaussian process regression," in *Proc. AISTATS*, 2003.

[11] Y. Gal, M. van der Wilk, and C. E. Rasmussen, "Distributed variational inference in sparse Gaussian process regression and latent variable models," in *Proc. NIPS*, 2014, pp. 3257–3265.

[12] J. Hensman, N. Fusi, and N. Lawrence, "Gaussian processes for big data," in *Proc. UAI*, 2013, pp. 282–290.

[13] C.-A. Cheng and B. Boots, "Incremental variational sparse Gaussian process regression," in *Proc. NIPS*, 2016, pp. 4410–4418.

[14] T. N. Hoang, Q. M. Hoang, and K. H. Low, "A unifying framework of anytime sparse Gaussian process regression models with stochastic variational inference for big data," in *Proc. ICML*, 2015, pp. 569–578.

[15] ——, "A distributed variational inference framework for unifying parallel sparse Gaussian process regression models," in *Proc. ICML*, 2016.

[16] T. D. Bui, C. Nguyen, and R. E. Turner, "Streaming sparse Gaussian process approximations," in *Proc. NIPS*, 2017, pp. 3301–3309.

[17] E. L. Snelson and Z. Gharahmani, "Sparse Gaussian processes using pseudo-inputs," in *Proc. NIPS*, 2005, pp. 1257–1264.

[18] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[19] M. K. Titsias and M. Lázaro-Gredilla, "Variational inference for Mahalanobis distance metrics in Gaussian process regression," in *Proc. NIPS*, 2013, pp. 279–287.

[20] N. Cao, K. H. Low, and J. M. Dolan, "Multi-robot informative path planning for active sensing of environmental phenomena: A tale of two algorithms," in *Proc. AAMAS*, 2013, pp. 7–14.

[21] T. N. Hoang, K. H. Low, P. Jaillet, and M. Kankanhalli, "Nonmyopic ϵ-Bayes-optimal active learning of Gaussian processes," in *Proc. ICML*, 2014, pp. 739–747.

[22] K. H. Low, J. M. Dolan, and P. Khosla, "Adaptive multi-robot wide-area exploration and mapping," in *Proc. AAMAS*, 2008, pp. 23–30.

[23] ——, "Information-theoretic approach to efficient adaptive path planning for mobile robotic environmental sensing," in *Proc. ICAPS*, 2009, pp. 233–240.

[24] ——, "Active Markov information-theoretic path planning for robotic environmental sensing," in *Proc. AAMAS*, 2011, pp. 753–760.

[25] R. Ouyang, K. H. Low, J. Chen, and P. Jaillet, "Multi-robot active sensing of non-stationary Gaussian process-based environmental phenomena," in *Proc. AAMAS*, 2014, pp. 573–580.

[26] Y. Zhang, T. N. Hoang, K. H. Low, and M. Kankanhalli, "Near-optimal active learning of multi-output Gaussian processes," in *Proc. AAAI*, 2016, pp. 2351–2357.

[27] E. Daxberger and K. H. Low, "Distributed batch Gaussian process optimization," in *Proc. ICML*, 2017, pp. 951–960.

[28] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani, "Predictive entropy search for efficient global optimization of black-box functions," in *Proc. NIPS*, 2014, pp. 918–926.

[29] T. N. Hoang, Q. M. Hoang, and K. H. Low, "Decentralized high-dimensional Bayesian optimization with factor graphs," in *Proc. AAAI*, 2018, pp. 3231–3238.

[30] C. K. Ling, K. H. Low, and P. Jaillet, "Gaussian process planning with Lipschitz continuous reward functions: Towards unifying Bayesian optimization, active learning, and beyond," in *Proc. AAAI*, 2016, pp. 1860–1866.

[31] Y. Gal and R. Turner, "Improving the Gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs," in *Proc. ICML*, 2015.

[32] Q. M. Hoang, T. N. Hoang, and K. H. Low, "A generalized stochastic variational Bayesian hyperparameter learning framework for sparse spectrum Gaussian process regression," in *Proc. AAAI*, 2017, pp. 2007–2014.

[33] E. L. Snelson and Z. Ghahramani, "Local and global sparse Gaussian process approximations," in *Proc. AISTATS*, 2007.

[34] H. Yu, T. N. Hoang, K. H. Low, and P. Jaillet, "Stochastic variational inference for Bayesian sparse Gaussian process regression," arXiv:1711.00221, 2017.

[35] M. Bauer, M. van der Wilk, and C. E. Rasmussen, "Understanding probabilistic sparse Gaussian process approximations," in *Proc. NIPS*, 2016, pp. 1533–1541.

[36] M. P. Deisenroth and J. W. Ng, "Distributed Gaussian processes," in *Proc. ICML*, 2015, pp. 1481–1490.

[37] J. Chen, K. H. Low, and C. K. Y. Tan, "Gaussian process-based decentralized data fusion and active sensing for mobility-on-demand system," in *Proc. RSS*, 2013.

[38] J. Chen, K. H. Low, P. Jaillet, and Y. Yao, "Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 3, pp. 901–921, 2015.

[39] J. Chen, K. H. Low, C. K.-Y. Tan, A. Oran, P. Jaillet, J. M. Dolan, and G. S. Sukhatme, "Decentralized data fusion and active sensing with mobile sensors for modeling and predicting spatiotemporal traffic phenomena," in *Proc. UAI*, 2012, pp. 163–173.

[40] T. N. Hoang, Q. M. Hoang, K. H. Low, and J. P. How, "Collective online learning of Gaussian processes in massive multi-agent systems," in *Proc. AAAI*, 2019.

[41] R. Ouyang and K. H. Low, "Gaussian process decentralized data fusion meets transfer learning in large-scale distributed cooperative perception," in *Proc. AAAI*, 2018, pp. 3876–3883.

[42] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," 2012.