A Generalized Stochastic Variational Bayesian Hyperparameter Learning Framework for Sparse Spectrum Gaussian Process Regression

Quang Minh Hoang[†] and **Trong Nghia Hoang**[*] and **Kian Hsiang Low**[†]

Department of Computer Science[†], Interactive Digital Media Institute[*]
National University of Singapore, Republic of Singapore
{hqminh, lowkh}@comp.nus.edu.sg[†], nghiaht@nus.edu.sg[*]

Abstract

While much research effort has been dedicated to scaling up sparse *Gaussian process* (GP) models based on inducing variables for big data, little attention is afforded to the other less explored class of low-rank GP approximations that exploit the sparse spectral representation of a GP kernel. This paper presents such an effort to advance the state of the art of sparse spectrum GP models to achieve competitive predictive performance for massive datasets. Our generalized framework of *stochastic variational Bayesian sparse spectrum GP* (*s*VBSSGP) models addresses their shortcomings by adopting a Bayesian treatment of the spectral frequencies to avoid overfitting, modeling these frequencies jointly in its variational distribution to enable their interaction *a posteriori*, and exploiting local data for boosting the predictive performance. However, such structural improvements result in a variational lower bound that is intractable to be optimized. To resolve this, we exploit a variational parameterization trick to make it amenable to stochastic optimization. Interestingly, the resulting stochastic gradient has a linearly decomposable structure that can be exploited to refine our stochastic optimization method to incur constant time per iteration while preserving its property of being an unbiased estimator of the exact gradient of the variational lower bound. Empirical evaluation on real-world datasets shows that *s*VBSSGP outperforms state-of-the-art stochastic implementations of sparse GP models.

1 Introduction

The machine learning community has recently witnessed the *Gaussian process* (GP) models gaining considerable traction in the research on kernel methods due to its expressive power and capability of performing probabilistic non-linear regression. However, a full-rank GP regression model incurs cubic time in the size of the data to compute its predictive distribution, hence limiting its use to small datasets. To lift this computational curse, a vast literature of sparse GP regression models (Quiñonero-Candela and Rasmussen 2005; Titsias 2009) have exploited a structural assumption of conditional independence based on the notion of *inducing variables* for achieving linear time in the data size. To scale up these sparse GP regression models further for performing real-time predictions necessary in many time-critical applications and decision support systems (e.g., environmental sensing (Cao, Low, and Dolan 2013; Dolan et al. 2009;

Ling, Low, and Jaillet 2016; Low, Dolan, and Khosla 2008; 2009; 2011; Low et al. 2012; Podnar et al. 2010; Zhang et al. 2016), traffic monitoring (Chen et al. 2012; Chen, Low, and Tan 2013; Chen et al. 2015; Hoang et al. 2014a; 2014b; Low et al. 2014a; 2014b; Ouyang et al. 2014; Xu et al. 2014; Yu et al. 2012)), (a) distributed (Chen et al. 2013; Hoang, Hoang, and Low 2016b; Low et al. 2015) and (b) stochastic (Hensman, Fusi, and Lawrence 2013; Hoang, Hoang, and Low 2015) implementations of such models have been developed to, respectively, (a) reduce their time to train with all the data by a factor close to the number of machines and (b) train with a small, randomly sampled subset of data in constant time per iteration of stochastic gradient ascent update and achieve asymptotic convergence to their predictive distributions.

On the other hand, there is a less well-explored, alternative class of low-rank GP approximations that exploit sparsity in the spectral representation of a GP kernel (Gal and Turner 2015; Lázaro-Gredilla et al. 2010) for gaining time efficiency and have empirically demonstrated competitive predictive performance for datasets of up to tens of thousands in size but, surprisingly, not received as much attention and research effort. In contrast to the above literature, such sparse spectrum GP regression models do not need to introduce an additional set of inducing inputs which is computationally challenging to be jointly optimized, especially with a large number of them that is necessary for accurate predictions. Unfortunately, the *sparse spectrum GP* (SSGP) model of Lázaro-Gredilla et al. (2010) does not scale well to massive datasets due to its linear time in the data size and also finds a point estimate of the spectral frequencies of its kernel that risks overfitting. The recent *variational SSGP* (VSSGP) model of Gal and Turner (2015) has attempted to address both shortcomings of SSGP with a stochastic implementation and a Bayesian treatment of the spectral frequencies, respectively. However, VSSGP has imposed the following highly restrictive structural assumptions to facilitate analytical derivations but potentially compromise its predictive performance: (a) The spectral frequencies are assumed to be fully independent *a posteriori* in its variational distribution, and (b) every test output assumes a deterministic relationship with the spectral frequencies in its test conditional and is thus conditionally independent of the training data, including the local data "close" to it (in the correla-

tion sense). As such, open questions remain whether VSSGP can still perform competitively well with these highly restrictive structural assumptions for massive, million-sized datasets and, more importantly, whether these assumptions can be relaxed to improve the predictive performance while preserving scalability to big data.

To tackle these questions, this paper presents a novel generalized framework of *stochastic variational Bayesian sparse spectrum GP* (sVBSSGP) regression models which (a) enables the spectral frequencies to interact *a posteriori* by modeling them jointly in its variational distribution, and (b) spans a spectrum of test conditionals that can trade off between the contributions of the degenerate test conditional of VSSGP vs. the *local* SSGP model trained with local data (Section 2). However, such proposed structural improvements over VSSGP and SSGP to boost the predictive performance result in a variational lower bound that is intractable to be optimized. To overcome this computational difficulty, we exploit a variational parameterization trick to make the variational lower bound amenable to stochastic optimization, which still incurs linear time in the data size per iteration of stochastic gradient ascent update (Section 3). Interestingly, we can exploit the linearly decomposable structure of this stochastic gradient to refine our stochastic optimization method to incur only constant time per iteration while preserving its property of being an unbiased estimator of the exact gradient of the variational lower bound (Section 4). We empirically evaluate the predictive performance and time efficiency of sVBSSGP on three real-world datasets, one of which is millions in size (Section 5).

2 A Generalized Bayesian Sparse Spectrum Gaussian Process Regression Framework

Let $\mathcal{X}$ be a d-dimensional input space such that each input vector $\mathbf{x} \in \mathcal{X}$ is associated with a latent output $f_{\mathbf{x}}$ and a noisy output $y_{\mathbf{x}} \triangleq f_{\mathbf{x}} + \epsilon$ generated by perturbing $f_{\mathbf{x}}$ with a random noise $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ where σ_n^2 is the noise variance. 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 0 for notational simplicity) and covariance $k(\mathbf{x}, \mathbf{x}') \triangleq \text{cov}[f_{\mathbf{x}}, f_{\mathbf{x}'}]$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, the latter of which can be defined by the commonly-used squared exponential kernel $k(\mathbf{x}, \mathbf{x}') \triangleq \sigma_s^2 \exp(-0.5(\mathbf{x} - \mathbf{x}')^\top \mathbf{\Delta}^{-1}(\mathbf{x} - \mathbf{x}'))$ where $\mathbf{\Delta} \triangleq \text{diag}[\ell_1^2, \ell_2^2, \ldots, \ell_d^2]$ and σ_s^2 are its squared length-scale and signal variance hyperparameters, respectively. Such a kernel can be expressed as the Fourier transform of a density function $p(\mathbf{r})$ over the domain of frequency vector $\mathbf{r}$ whose coefficients form a set of trigonometric basis functions (Lázaro-Gredilla et al. 2010):

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\mathbf{r} \sim p(\mathbf{r})}[\, \sigma_s^2 \cos(2\pi \mathbf{r}^\top (\mathbf{x} - \mathbf{x}')) \,] \qquad (1)$$

where $p(\mathbf{r}) \triangleq \mathcal{N}(\mathbf{0}, (4\pi^2 \mathbf{\Delta})^{-1})$. In the same spirit as that of Lázaro-Gredilla et al. (2010), we approximate the kernel in (1) by its unbiased estimator constructed from m i.i.d. sampled spectral frequencies $\mathbf{r}_i$ for $i = 1, \ldots, m$:

$$k(\mathbf{x}, \mathbf{x}') \simeq \frac{\sigma_s^2}{m} \sum_{i=1}^m \cos(2\pi \mathbf{r}_i^\top (\mathbf{x} - \mathbf{x}')) = \boldsymbol{\phi}_{\boldsymbol{\theta}}^\top(\mathbf{x}) \mathbf{\Lambda} \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{x}') \quad (2)$$

where $\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{x}) \triangleq [\phi_{\boldsymbol{\theta}}^1(\mathbf{x}), \phi_{\boldsymbol{\theta}}^2(\mathbf{x}), \ldots, \phi_{\boldsymbol{\theta}}^{2m}(\mathbf{x})]^\top$ denotes a vector of basis functions $\phi_{\boldsymbol{\theta}}^{2i-1}(\mathbf{x}) \triangleq \cos(2\pi \mathbf{r}_i^\top \mathbf{x})$ and $\phi_{\boldsymbol{\theta}}^{2i}(\mathbf{x}) \triangleq \sin(2\pi \mathbf{r}_i^\top \mathbf{x})$ for $i = 1, \ldots, m$, $\mathbf{\Lambda} \triangleq (\sigma_s^2/m)\mathbf{I}$, and $\boldsymbol{\theta} \triangleq \text{vec}(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_m)$. Learning the length-scales in $\mathbf{\Delta}$ of the original kernel is thus cast as optimizing $\boldsymbol{\theta}$ in this alternative representation (2) of the kernel. Then, the induced covariance matrix $\mathbf{K}(\mathbf{X}, \mathbf{X})$ for any finite subset $\mathbf{X} \triangleq \{\mathbf{x}_i\}_{i=1}^n$ of training inputs can be written as $\mathbf{K}(\mathbf{X}, \mathbf{X}) \triangleq \mathbf{\Phi}_{\boldsymbol{\theta}}^\top(\mathbf{X}) \mathbf{\Lambda} \mathbf{\Phi}_{\boldsymbol{\theta}}(\mathbf{X})$ where $\mathbf{\Phi}_{\boldsymbol{\theta}}(\mathbf{X}) \triangleq [\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{x}_1) \ldots \boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{x}_n)]$.

To learn the optimal parameters of the distribution over $\boldsymbol{\theta}$ (i.e., not known in advance) given by $\mathcal{N}(\mathbf{0}, (4\pi^2 \mathbf{\Delta}_{\boldsymbol{\theta}})^{-1})$ (i.e., derived from $p(\mathbf{r})$ below (1)) where $\mathbf{\Delta}_{\boldsymbol{\theta}} \triangleq \mathbf{I} \otimes \mathbf{\Delta}$, we adopt the standard Bayesian treatment for $\boldsymbol{\theta}$ by first imposing a prior distribution $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Theta})$ for some covariance $\mathbf{\Theta}$ designed *a priori* to reflect our knowledge about $\boldsymbol{\theta}$ and then using training data to infer its posterior which is expected to closely approximate the optimal distribution. This signifies a key difference between our generalized framework and the *sparse spectrum GP* (SSGP) model of Lázaro-Gredilla et al. (2010), the latter of which finds a point estimate of $\boldsymbol{\theta}$ via maximum likelihood estimation that risks overfitting.

As shall be elucidated later in this paper, the finite trigonometric representation of the kernel (2) can be used to efficiently and scalably compute the predictive distribution of our framework by exploiting some mild structural assumptions. Specifically, we assume that for any finite subset $\mathbf{X} \subset \mathcal{X}$, a vector $\mathbf{s}$ of *nuisance* variables exists for which the joint distribution of $\mathbf{f} \triangleq [f_{\mathbf{x}}]_{\mathbf{x} \in \mathbf{X}}^\top$ and $\mathbf{s}$ conditioned on $\boldsymbol{\theta}$ is

$$\begin{bmatrix} \mathbf{s} \\ \mathbf{f} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{\Lambda} & \mathbf{\Lambda}\mathbf{\Phi}_{\boldsymbol{\theta}}(\mathbf{X}) \\ \mathbf{\Phi}_{\boldsymbol{\theta}}^\top(\mathbf{X})\mathbf{\Lambda} & \mathbf{\Phi}_{\boldsymbol{\theta}}^\top(\mathbf{X})\mathbf{\Lambda}\mathbf{\Phi}_{\boldsymbol{\theta}}(\mathbf{X}) \end{bmatrix} \right). \quad (3)$$

Intuitively, $\mathbf{s}$ can be interpreted as the latent outputs of some imaginary inputs $\mathbf{U} \triangleq \{\mathbf{u}_j\}_{j=1}^{2m}$ such that $\phi_{\boldsymbol{\theta}}^i(\mathbf{u}_j) \triangleq \mathbb{I}(i = j)$. Using (2), it follows immediately that $\mathbf{K}(\mathbf{U}, \mathbf{U}) = \mathbf{\Lambda}$, $\mathbf{K}(\mathbf{U}, \mathbf{X}) = \mathbf{\Lambda}\mathbf{\Phi}_{\boldsymbol{\theta}}(\mathbf{X})$, and $\mathbf{K}(\mathbf{X}, \mathbf{U}) = \mathbf{\Phi}_{\boldsymbol{\theta}}^\top(\mathbf{X})\mathbf{\Lambda}$ which reproduce the covariance matrix in (3). Secondly, we assume that given $\boldsymbol{\alpha} \triangleq \text{vec}(\boldsymbol{\theta}, \mathbf{s})$, any latent test output depends on only a small subset of local training data of fixed size: Supposing the input space is partitioned into p disjoint subspaces (i.e., $\mathcal{X} = \bigcup_{i=1}^p \mathcal{X}_i$) which directly induce a partition on the training inputs $\mathbf{X} = \bigcup_{i=1}^p \mathbf{X}_i$ such that $\mathbf{X}_i \subset \mathcal{X}_i$, $p(f_{\mathbf{x}_*}|\mathbf{y}, \boldsymbol{\alpha}) = p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\alpha})$ for any test input $\mathbf{x}_* \in \mathcal{X}_k$ and $k = 1, \ldots, p$ where $\mathbf{y} \triangleq [y_{\mathbf{x}}]_{\mathbf{x} \in \mathbf{X}}^\top$ and $\mathbf{y}_k \triangleq [y_{\mathbf{x}}]_{\mathbf{x} \in \mathbf{X}_k}^\top$. Then, the predictive distribution can be computed using

$$p(f_{\mathbf{x}_*}|\mathbf{y}) = \mathbb{E}_{\boldsymbol{\alpha} \sim p(\boldsymbol{\alpha}|\mathbf{y})}[\, p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\alpha}) \,] \qquad (4)$$

which reveals that it can be evaluated by deriving posterior distribution $p(\boldsymbol{\alpha}|\mathbf{y})$ described later in Section 3, and the test conditional $p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\alpha})$ *consistent* with the above structural assumption of $\mathbf{s}$: Marginalizing out nuisance variables $\mathbf{s}$ from such a test conditional should yield $p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\theta})$, i.e., $p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\theta}) = \int_{\mathbf{s}} p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\alpha}) \, p(\mathbf{s}|\mathbf{y}_k, \boldsymbol{\theta}) \, d\mathbf{s}$ where both $p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\theta})$ and $p(\mathbf{s}|\mathbf{y}_k, \boldsymbol{\theta})$ can be derived from (3), as shown in (Hoang, Hoang, and Low 2016a).

Our first result below derives a spectrum of consistent test conditionals in our generalized framework that trade off between the use of global information $\boldsymbol{\alpha}$ vs. local data $(\mathbf{X}_k, \mathbf{y}_k)$, albeit to varying degrees controlled by γ:

Proposition 1 *For all* $\mathbf{x}_* \in \mathcal{X}_k$ *and* $|\gamma| \leq 1$*, define the test conditional* $p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\alpha}) \triangleq \mathcal{N}(\mu_{\mathbf{x}_*}(\boldsymbol{\alpha}), \sigma^2_{\mathbf{x}_*}(\boldsymbol{\alpha}))$ *where*

$$\mu_{\mathbf{x}_*}(\boldsymbol{\alpha}) \triangleq \gamma \boldsymbol{\phi}_{\boldsymbol{\theta}}^\top(\mathbf{x}_*)\mathbf{s} + (1-\gamma)\boldsymbol{\phi}_{\boldsymbol{\theta}}^\top(\mathbf{x}_*)\boldsymbol{\Gamma}_k^{-1}\boldsymbol{\Phi}_{\boldsymbol{\theta}}(\mathbf{X}_k)\mathbf{y}_k ,$$
$$\sigma^2_{\mathbf{x}_*}(\boldsymbol{\alpha}) \triangleq (1-\gamma^2)\sigma_n^2\boldsymbol{\phi}_{\boldsymbol{\theta}}^\top(\mathbf{x}_*)\boldsymbol{\Gamma}_k^{-1}\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{x}_*) ,$$

and $\boldsymbol{\Gamma}_k \triangleq \boldsymbol{\Phi}_{\boldsymbol{\theta}}(\mathbf{X}_k)\boldsymbol{\Phi}_{\boldsymbol{\theta}}^\top(\mathbf{X}_k) + \sigma_n^2\boldsymbol{\Lambda}^{-1}$*. Then,* $p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\alpha})$ *is consistent with the structural assumption of* $\mathbf{s}$ *in* (3).

Its proof is in (Hoang, Hoang, and Low 2016a).

Remark 1 The special case of $\gamma = 1$ recovers the degenerate test conditional $p(f_{\mathbf{x}_*}|\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\phi}_{\boldsymbol{\theta}}^\top(\mathbf{x}_*)\mathbf{s}, 0)$ induced by the *variational SSGP* (VSSGP) model of Gal and Turner (2015) (see equation 4 and Section 3 therein) which reveals that it imposes a highly restrictive deterministic relationship between $f_{\mathbf{x}_*}$ and $\boldsymbol{\alpha}$ and also fails to exploit the local data $(\mathbf{X}_k, \mathbf{y}_k)$ (i.e., due to conditional independence between $f_{\mathbf{x}_*}$ and $\mathbf{y}_k$ given $\boldsymbol{\alpha}$) that can potentially improve the predictive performance. Unfortunately, VSSGP cannot be trivially extended to span the entire spectrum since it relies on the induced deterministic relationship between $f_{\mathbf{x}_*}$ and $\boldsymbol{\alpha}$ to analytically derive its predictive distribution, which does not hold for $\gamma \neq 1$. On the other hand, when $\gamma = 0$, the test conditional in Proposition 1 becomes independent of the nuisance variables $\mathbf{s}$ and reduces to the predictive distribution $p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\theta})$ of the SSGP model of Lázaro-Gredilla et al. (2010) (see equation 7 therein) given its point estimate of the spectral frequencies $\boldsymbol{\theta}$ but restricted to the local data $(\mathbf{X}_k, \mathbf{y}_k)$ corresponding to input subspace $\mathcal{X}_k$. Hence, γ can also be perceived as a controlling parameter that trades off between the contributions of the degenerate test conditional of VSSGP vs. the *local* SSGP model to constructing a test conditional in our generalized framework. To investigate this trade-off, our experiments in Section 5 show that given the learned $\boldsymbol{\theta}$, the predictive performance is maximized at $\gamma = 0$ when the test conditional $p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\alpha})$ depends on local data $(\mathbf{X}_k, \mathbf{y}_k)$ but not $\mathbf{s}$, which justifies viewing $\mathbf{s}$ as a "nuisance" to prediction despite its crucial role in scalable learning of $\boldsymbol{\theta}$ via stochastic optimization (Section 4).

In the sections to follow, we will propose a novel stochastic optimization method for deriving a *variationally optimal* approximation to the posterior distribution $p(\boldsymbol{\alpha}|\mathbf{y})$ that is used to marginalize out the global information $\boldsymbol{\alpha}$ from any test conditional $p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\alpha})$ in Proposition 1 to yield an asymptotic approximation to the predictive distribution $p(f_{\mathbf{x}_*}|\mathbf{y})$ (4) regardless of the value of $\gamma \in [-1, 1]$.

3 Variational Inference for Bayesian Sparse Spectrum Gaussian Process Regression

This section presents a variational approximation $q(\boldsymbol{\alpha})$ of the posterior distribution $p(\boldsymbol{\alpha}|\mathbf{y})$ achieved by using variational inference which involves choosing a parameterization for $q(\boldsymbol{\alpha})$ (Section 3.1) and optimizing its defining parameters (Section 3.2) to minimize its *Kullback-Leibler* (KL) distance $D_{\mathrm{KL}}(q) \triangleq \mathrm{KL}(q(\boldsymbol{\alpha})\|p(\boldsymbol{\alpha}|\mathbf{y}))$ to $p(\boldsymbol{\alpha}|\mathbf{y})$. The optimized $q(\boldsymbol{\alpha})$ can then be used as a tractably cheap surrogate of $p(\boldsymbol{\alpha}|\mathbf{y})$ for marginalizing out $\boldsymbol{\alpha}$ from test conditional $p(f_{\mathbf{x}_*}|\mathbf{y}_k, \boldsymbol{\alpha})$ in Proposition 1 to derive an approximation to predictive distribution $p(f_{\mathbf{x}_*}|\mathbf{y})$ (4) efficiently (Section 4).

3.1 Variational Parameterization

Specifically, we parameterize $\boldsymbol{\alpha} = \mathrm{vec}(\boldsymbol{\theta}, \mathbf{s}) \triangleq \mathbf{Mz} + \mathbf{b}$ where the variational parameters $\boldsymbol{\eta} \triangleq \mathrm{vec}(\mathbf{M}, \mathbf{b})$ are independent of $(\boldsymbol{\theta}, \mathbf{s})$, and $\mathbf{z}$ is distributed by an analytically tractable user-specified distribution $\psi(\mathbf{z}) \simeq p(\mathbf{z}|\mathbf{y})$ that is straightforward to draw samples from (i.e., $\mathbf{z} \sim \psi(\mathbf{z})$). We assume that $\psi(\mathbf{z})$ is analytically differentiable with respect to $\mathbf{z}$ and the affine matrix $\mathbf{M}$ is invertible such that $\mathbf{z} = \mathbf{M}^{-1}(\boldsymbol{\alpha} - \mathbf{b})$ exists. Then, $q(\boldsymbol{\theta}, \mathbf{s})$ can be expressed in terms of $\mathbf{M}$, $\mathbf{b}$, and $\psi(\mathbf{z})$:

Lemma 1 *The variational distribution* $q(\boldsymbol{\theta}, \mathbf{s})$ *can be indirectly parameterized via* $\psi(\mathbf{z})$ *by* $q(\boldsymbol{\theta}, \mathbf{s}) = q(\boldsymbol{\alpha}) = \psi(\mathbf{M}^{-1}(\boldsymbol{\alpha} - \mathbf{b}))/|\mathbf{M}| = \psi(\mathbf{z})/|\mathbf{M}|$ *where* $|\mathbf{M}|$ *denotes the absolute value of* $\det(\mathbf{M})$.

Lemma 1 follows directly from the change of variables theorem. Using the above parameterization of $\boldsymbol{\alpha}$, we can also express the prior $p(\boldsymbol{\theta}, \mathbf{s})$ in terms of variational parameters $\mathbf{M}$ and $\mathbf{b}$. To see this, recall that $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta})$ and $p(\mathbf{s}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$ (see Section 2 and (3)). This implies $p(\boldsymbol{\alpha}) = p(\mathrm{vec}(\boldsymbol{\theta}, \mathbf{s})) = p(\boldsymbol{\theta})\,p(\mathbf{s}) = \mathcal{N}(\boldsymbol{\alpha}|\mathbf{0}, \mathrm{blkdiag}[\boldsymbol{\Theta}, \boldsymbol{\Lambda}])$. Plugging in the expression of $\boldsymbol{\alpha}$ yields $p(\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{Mz} + \mathbf{b}|\mathbf{0}, \mathrm{blkdiag}[\boldsymbol{\Theta}, \boldsymbol{\Lambda}])$[1]. At this time, the need to parameterize $q(\boldsymbol{\alpha})$ in Lemma 1 may not be obvious to a reader because it is motivated by a technical necessity to guarantee the asymptotic convergence of our stochastic optimization method (Remark 4) rather than a conceptual intuition.

Remark 2 Our generalized framework enables both the spectral frequencies $\boldsymbol{\theta}$ and the nuisance variables $\mathbf{s}$ to interact *a posteriori* by modeling them jointly in the variational distribution $q(\boldsymbol{\theta}, \mathbf{s})$, as detailed in Lemma 1, and still preserves scalability (Section 4). In contrast, the VSSGP model of Gal and Turner (2015) assumes $\mathbf{r}_1, \ldots, \mathbf{r}_m$, and $\mathbf{s}$ to be statistically independent *a posteriori* in its variational distribution (see section 4 therein). Relaxing this assumption will cause VSSGP to lose its scalability as its induced variational lower bound will inevitably become intractable.

Remark 3 Though the model of Titsias and Lázaro-Gredilla (2014) has adopted a similar parameterization but only for the original GP hyperparameters, it incurs cubic time in the data size per iteration of gradient ascent update, as shown in its supplementary materials and experiments. A factorized marginal likelihood has to be further assumed for this parameterization to achieve scalability. Our framework does not require such a strong assumption and still scales well to million-sized datasets (Section 5). This is, perhaps surprisingly, achieved through introducing the nuisance variables $\mathbf{s}$ (Section 2), which interestingly embeds a linearly decomposable structure within the log-likelihood of data $\log p(\mathbf{y}|\boldsymbol{\alpha})$ instead of $\log p(\mathbf{y}|\boldsymbol{\theta})$. As shown later in Theorem 1, such a structure is the key ingredient for developing our stochastic optimization method (Section 4).

3.2 Variational Optimization

To optimize $q(\boldsymbol{\alpha})$ (i.e., by minimizing $D_{\mathrm{KL}}(q)$), we first show that the log-marginal likelihood $\log p(\mathbf{y})$ can be de-

---

[1]Note that $\psi(\mathbf{z})$ is meant to approximate the posterior distribution of $\mathbf{z}$ given $\mathbf{y}$ instead of its prior distribution. So, it cannot be used to derive $p(\boldsymbol{\alpha})$ by applying the affine transformation on $\mathbf{z}$.

composed into a sum of a lower-bound functional $L(q)$ and $D_{\mathrm{KL}}(q)$: $\log p(\mathbf{y}) = L(q) + D_{\mathrm{KL}}(q)$ where $L(q) \triangleq \mathbb{E}_{\boldsymbol{\alpha} \sim q(\boldsymbol{\alpha})}[\log p(\mathbf{y}|\boldsymbol{\alpha}) - \log(q(\boldsymbol{\alpha})/p(\boldsymbol{\alpha}))]$, as detailed in (Hoang, Hoang, and Low 2016a). So, minimizing $D_{\mathrm{KL}}(q)$ is equivalent to maximizing $L(q)$ with respect to the variational parameter $\boldsymbol{\eta} = \mathrm{vec}(\mathbf{M}, \mathbf{b})$ of $q(\boldsymbol{\alpha})$ (Lemma 1) since $\log p(\mathbf{y})$ is a constant (i.e., independent of $\boldsymbol{\eta}$). In practice, this is usually achieved by setting the derivative $\partial L/\partial \boldsymbol{\eta} = 0$ and solving for $\boldsymbol{\eta}$, which is unfortunately intractable since its exact analytic expression is not known.

To sidestep this intractability issue, we instead adopt a stochastic optimization method that is capable of maximizing $L(q)$ via iterative stochastic gradient ascent updates. However, this also requires the stochastic gradient $\partial \widehat{L}/\partial \boldsymbol{\eta}$ (6) to be an *analytically tractable* and *unbiased* estimator of the exact gradient $\partial L/\partial \boldsymbol{\eta}$. To derive this, we first exploit Lemma 1 to re-express $L(q)$ as an expectation with respect to $\mathbf{z} \sim \psi(\mathbf{z})$: $L(q) = \mathbb{E}_{\mathbf{z} \sim \psi(\mathbf{z})}[\log p(\mathbf{y}|\boldsymbol{\alpha}) - \log(q(\boldsymbol{\alpha})/p(\boldsymbol{\alpha}))]$ where $\boldsymbol{\alpha} = \mathbf{Mz} + \mathbf{b}$, as shown in (Hoang, Hoang, and Low 2016a). Then, taking the derivatives with respect to $\boldsymbol{\eta}$ on both sides of the above equation yields

$$\partial L/\partial \boldsymbol{\eta} = \mathbb{E}_{\mathbf{z} \sim \psi(\mathbf{z})}[\, \partial(\log p(\mathbf{y}|\boldsymbol{\alpha}) - \log(q(\boldsymbol{\alpha})/p(\boldsymbol{\alpha})))/\partial \boldsymbol{\eta} \,] \tag{5}$$

which reveals a simple, analytically tractable (proven in (Hoang, Hoang, and Low 2016a) using Lemma 1) choice for our stochastic gradient

$$\partial \widehat{L}/\partial \boldsymbol{\eta} \triangleq \partial(\log p(\mathbf{y}|\boldsymbol{\alpha}) - \log(q(\boldsymbol{\alpha})/p(\boldsymbol{\alpha})))/\partial \boldsymbol{\eta} . \tag{6}$$

Thus, $\mathbb{E}_{\mathbf{z} \sim \psi(\mathbf{z})}[\partial \widehat{L}/\partial \boldsymbol{\eta}] = \partial L/\partial \boldsymbol{\eta}$ guarantees that $\partial \widehat{L}/\partial \boldsymbol{\eta}$ is indeed an unbiased estimator of $\partial L/\partial \boldsymbol{\eta}$.

Remark 4 (5) reveals that parameterizing $q(\boldsymbol{\alpha})$ indirectly through $\psi(\mathbf{z})$ in Lemma 1 is essential to enabling stochastic optimization in our generalized framework: Since $\psi(\mathbf{z})$ does not depend on the variational parameters $\boldsymbol{\eta} = \mathrm{vec}(\mathbf{M}, \mathbf{b})$, the derivative operator in (5) can be moved inside the expectation, which trivially reveals an unbiased stochastic estimate (6) of the exact gradient via sampling $\mathbf{z}$. Otherwise, suppose that we attempt to derive a stochastic gradient for $L(q)$ using the original expression $L(q) = \mathbb{E}_{\boldsymbol{\alpha} \sim q(\boldsymbol{\alpha})}[\log p(\mathbf{y}|\boldsymbol{\alpha}) - \log(q(\boldsymbol{\alpha})/p(\boldsymbol{\alpha}))]$ instead of (5) in the same manner with a direct parameterization of $q(\boldsymbol{\alpha})$ not exploiting $\psi(\mathbf{z})$. Then, after differentiating both sides of the above expression with respect to $\boldsymbol{\eta}$, the derivative operator on the RHS cannot be moved inside the expectation over $\boldsymbol{\alpha} \sim q(\boldsymbol{\alpha})$ since it depends on $\boldsymbol{\eta}$, which suggests that deriving an unbiased estimator for $\partial L/\partial \boldsymbol{\eta}$ has become non-trivial without using the parameterization of $q(\boldsymbol{\alpha})$ in Lemma 1.

To understand why $\partial \widehat{L}/\partial \boldsymbol{\eta}$ (6) is analytically tractable, it suffices to show that both its derivatives $\partial \log p(\mathbf{y}|\boldsymbol{\alpha})/\partial \boldsymbol{\eta}$ and $\partial \log(q(\boldsymbol{\alpha})/p(\boldsymbol{\alpha}))/\partial \boldsymbol{\eta}$ in (6) are analytically tractable, the latter of which is detailed in (Hoang, Hoang, and Low 2016a). For the former, since the trigonometric basis functions $\{\phi_{\boldsymbol{\theta}}^i(\mathbf{x})\}_{i=1}^{2m}$ (Section 2) are differentiable, it can be shown that $\partial \log p(\mathbf{y}|\boldsymbol{\alpha})/\partial \boldsymbol{\eta}$ is analytically tractable by exploiting the following result giving a closed-form expression of $\log p(\mathbf{y}|\boldsymbol{\alpha})$ in terms of $\boldsymbol{\theta}$ and $\mathbf{s}$:

Lemma 2 $\log p(\mathbf{y}|\boldsymbol{\alpha}) = -0.5\sigma_n^{-2}\mathbf{v}^\top \mathbf{v} - 0.5n \log(2\pi\sigma_n^2)$ where $\mathbf{v} \triangleq \mathbf{y} - \boldsymbol{\Phi}_{\boldsymbol{\theta}}^\top(\mathbf{X})\mathbf{s}$.

Its proof is in (Hoang, Hoang, and Low 2016a). The above choice of $\partial \widehat{L}/\partial \boldsymbol{\eta}$ (6) and Lemma 2 show that it is not only an unbiased estimator of the exact gradient but is also analytically tractable, which satisfies all the required conditions to guarantee the asymptotic convergence of our proposed stochastic optimization method. However, a critical issue remains that makes it scale poorly in practice: It can be derived from Lemma 2 that naively evaluating its derivative $\partial \log p(\mathbf{y}|\boldsymbol{\alpha})/\partial \boldsymbol{\eta}$ in (6) in a straightforward manner incurs linear time in data size n per iteration of stochastic gradient ascent update, which is expensive for massive datasets.

4 Stochastic Optimization

To overcome the issue of scalability in evaluating the stochastic gradient $\partial \widehat{L}/\partial \boldsymbol{\eta}$ (6) (Section 3.2), we will show in Theorem 1 below that $\partial \log p(\mathbf{y}|\boldsymbol{\alpha})/\partial \boldsymbol{\eta}$ is decomposable into a linear sum of analytically tractable terms, each of which depends on only a small subset of local data. Interestingly, since only the derivative $\partial \log p(\mathbf{y}|\boldsymbol{\alpha})/\partial \boldsymbol{\eta}$ in (6) involves the training data $(\mathbf{X}, \mathbf{y})$, Theorem 1 implies a similar decomposition of $\partial \widehat{L}/\partial \boldsymbol{\eta}$. As a result, we can derive a new stochastic estimate of the exact gradient $\partial L/\partial \boldsymbol{\eta}$ that can be computed efficiently and scalably using only one or a few randomly sampled subset(s) of local data of fixed size and still preserves the property of being its unbiased estimator (Section 4.1). Computing this new stochastic gradient incurs only constant time in the data size n per iteration of stochastic gradient ascent update, which, together with Proposition 1 in Section 2, constitute the foundation of our generalized framework of stochastic variational Bayesian SSGP regression models for big data (Section 4.2).

4.1 Stochastic Gradient Revisited

To derive a computationally scalable and unbiased stochastic gradient, we rely on our main result below showing the decomposability of $\partial \log p(\mathbf{y}|\boldsymbol{\alpha})/\partial \boldsymbol{\eta}$ in (6) into a linear sum of analytically tractable terms, each of which depends on only a small subset of local training data of fixed size:

Theorem 1 Let $\mathbf{v}_i \triangleq \mathbf{y}_i - \boldsymbol{\Phi}_{\boldsymbol{\theta}}^\top(\mathbf{X}_i)\mathbf{s}$ for $i = 1, \ldots, p$. Then, $\partial \log p(\mathbf{y}|\boldsymbol{\alpha})/\partial \boldsymbol{\eta} = \sum_{i=1}^{p} F_i(\boldsymbol{\eta}, \boldsymbol{\alpha})$ where $F_i(\boldsymbol{\eta}, \boldsymbol{\alpha}) \triangleq -0.5\sigma_n^{-2}\nabla_{\boldsymbol{\eta}}(\mathbf{v}_i^\top \mathbf{v}_i)$ is analytically tractable.

Its proof in (Hoang, Hoang, and Low 2016a) utilizes Lemma 2. Using Theorem 1, the following linearly decomposable structure of $\partial \widehat{L}/\partial \boldsymbol{\eta}$ (6) results:

$$\partial \widehat{L}/\partial \boldsymbol{\eta} = \mathbb{E}_{i \sim \mathcal{U}(1,p)}[\, pF_i(\boldsymbol{\eta}, \boldsymbol{\alpha}) - \partial \log(q(\boldsymbol{\alpha})/p(\boldsymbol{\alpha}))/\partial \boldsymbol{\eta} \,] \tag{7}$$

where i is treated as a discrete random variable uniformly distributed over the set of partition indices $\{1, 2, \ldots, p\}$. This interestingly reveals an unbiased estimator for the stochastic gradient $\partial \widehat{L}/\partial \boldsymbol{\eta}$ which can be constructed stochastically by sampling i:

$$\partial \widetilde{L}/\partial \boldsymbol{\eta} \triangleq pF_i(\boldsymbol{\eta}, \boldsymbol{\alpha}) - \partial \log(q(\boldsymbol{\alpha})/p(\boldsymbol{\alpha}))/\partial \boldsymbol{\eta}$$

such that substituting it into (7) yields $\mathbb{E}_{i \sim \mathcal{U}(1,p)}[\partial \widetilde{L}/\partial \boldsymbol{\eta}] = \partial \widehat{L}/\partial \boldsymbol{\eta}$. Then, taking the expectation over $\mathbf{z} \sim \psi(\mathbf{z})$ on both

sides of this equality gives $\mathbb{E}_{\mathbf{z} \sim \psi(\mathbf{z})}[\mathbb{E}_{i \sim \mathcal{U}(1,p)}[\partial \widetilde{L} / \partial \boldsymbol{\eta}]] = \mathbb{E}_{\mathbf{z} \sim \psi(\mathbf{z})}[\partial \widehat{L} / \partial \boldsymbol{\eta}] = \partial L / \partial \boldsymbol{\eta}$, which proves that $\partial \widetilde{L} / \partial \boldsymbol{\eta}$ is also an unbiased estimator of $\partial L / \partial \boldsymbol{\eta}$ that can be constructed by sampling both $i \sim \mathcal{U}(1, p)$ and $\mathbf{z} \sim \psi(\mathbf{z})$ (Section 3.1) independently. Replacing $\partial \widehat{L} / \partial \boldsymbol{\eta}$ with $\partial \widetilde{L} / \partial \boldsymbol{\eta}$ produces a highly efficient stochastic gradient ascent update that incurs constant time in the data size n per iteration since $\partial \widetilde{L} / \partial \boldsymbol{\eta}$ can be computed using only a single randomly sampled subset of local data $(\mathbf{X}_i, \mathbf{y}_i)$ of fixed size.

Instead of utilizing just a single pair $(i, \mathbf{z})$ of samples, the above stochastic gradient $\partial \widetilde{L} / \partial \boldsymbol{\eta}$ can be generalized to simultaneously process multiple pairs of independent samples and their corresponding sampled subsets of local data in one stochastic gradient ascent update (detailed in (Hoang, Hoang, and Low 2016a)), thereby improving the rate of convergence while preserving its property of an unbiased estimator of the exact gradient $\partial L / \partial \boldsymbol{\eta}$. Asymptotic convergence of the estimate of $\boldsymbol{\eta}$ (and hence the estimate of $q(\boldsymbol{\alpha})$) is guaranteed if the step sizes are scheduled appropriately.

4.2 Approximate Predictive Inference

In iteration t of stochastic gradient ascent update, an estimate $q_t(\boldsymbol{\alpha})$ of the variationally optimal approximation $q(\boldsymbol{\alpha})$ can be induced from the current estimate $\boldsymbol{\eta}_t = \text{vec}(\mathbf{M}_t, \mathbf{b}_t)$ of its variational parameters $\boldsymbol{\eta}$ using the parameterization in Lemma 1. As a result, using the law of iterated expectations, the predictive mean can be approximated by $\widehat{\mu}_{\mathbf{x}_*} = \mathbb{E}_{\boldsymbol{\alpha} \sim q_t(\boldsymbol{\alpha})}[\mathbb{E}[f_{\mathbf{x}_*} | \mathbf{y}_k, \boldsymbol{\alpha}]] = \mathbb{E}_{\boldsymbol{\alpha} \sim q_t(\boldsymbol{\alpha})}[\mu_{\mathbf{x}_*}(\boldsymbol{\alpha})] = \int_{\boldsymbol{\alpha}} q_t(\boldsymbol{\alpha}) \mu_{\mathbf{x}_*}(\boldsymbol{\alpha}) d\boldsymbol{\alpha}$ where $\boldsymbol{\alpha} = \text{vec}(\boldsymbol{\theta}, \mathbf{s})$ and $\mu_{\mathbf{x}_*}(\boldsymbol{\alpha})$ is previously defined in Proposition 1. However, since the integration over $\boldsymbol{\alpha}$ is not always analytically tractable, we approximate it by drawing i.i.d. samples $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r$ from $q_t(\boldsymbol{\alpha})$ to estimate $\widehat{\mu}_{\mathbf{x}_*} \simeq r^{-1} \sum_{i=1}^{r} \mu_{\mathbf{x}_*}(\boldsymbol{\alpha}_i)$. Similarly, using the variance decomposition formula and definition of variance, the predictive variance can be approximated by $\widehat{\sigma}_{\mathbf{x}_*}^2 = \mathbb{E}_{\boldsymbol{\alpha} \sim q_t(\boldsymbol{\alpha})}[\sigma_{\mathbf{x}_*}^2(\boldsymbol{\alpha})] + \mathbb{V}_{\boldsymbol{\alpha} \sim q_t(\boldsymbol{\alpha})}[\mu_{\mathbf{x}_*}(\boldsymbol{\alpha})] = \mathbb{E}_{\boldsymbol{\alpha} \sim q_t(\boldsymbol{\alpha})}[\sigma_{\mathbf{x}_*}^2(\boldsymbol{\alpha}) + \mu_{\mathbf{x}_*}^2(\boldsymbol{\alpha})] - \widehat{\mu}_{\mathbf{x}_*}^2 \simeq r^{-1} \sum_{i=1}^{r} (\sigma_{\mathbf{x}_*}^2(\boldsymbol{\alpha}_i) + \mu_{\mathbf{x}_*}^2(\boldsymbol{\alpha}_i)) - \widehat{\mu}_{\mathbf{x}_*}^2$ where $\sigma_{\mathbf{x}_*}^2(\boldsymbol{\alpha})$ is previously defined in Proposition 1 and the $\mathbb{V}_{\boldsymbol{\alpha} \sim q_t(\boldsymbol{\alpha})}[\mu_{\mathbf{x}_*}(\boldsymbol{\alpha})]$ term arises due to the uncertainty of $\boldsymbol{\alpha}$. The samples $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_r$ are in turn obtained by sampling $\mathbf{z}_1, \ldots, \mathbf{z}_r$ from $\psi(\mathbf{z})$ and applying the parametric transformation in Lemma 1 with respect to the current estimates $\mathbf{M}_t$ and $\mathbf{b}_t$, that is, $\text{vec}(\boldsymbol{\theta}_i, \mathbf{s}_i) = \boldsymbol{\alpha}_i = \mathbf{M}_t \mathbf{z}_i + \mathbf{b}_t$ for $i = 1, \ldots, r$. Computing the predictive mean $\widehat{\mu}_{\mathbf{x}_*}$ and variance $\widehat{\sigma}_{\mathbf{x}_*}^2$ thus incurs constant time in the data size n, hence achieving efficient approximate predictive inference.

5 Empirical Studies

This section empirically evaluates the predictive performance and time efficiency of our sVBSSGP model on three real-world datasets: (a) The AIMPEAK dataset (Chen et al. 2013) consists of 41800 traffic speed observations (km/h) along 775 urban road segments during the morning peak hours on April 20, 2011. Each observation features a 5-dimensional input vector of a road segment's length, number of lanes, speed limit, direction, and its recording time (i.e., discretized into 54 five-minute time slots), and a corresponding output measuring the traffic speed (km/h); (b) the bench-

mark AIRLINE dataset (Hensman, Fusi, and Lawrence 2013; Hoang, Hoang, and Low 2015) contains 2000000 information records of US commercial flights in 2008. Each record features a 8-dimensional input vector of the aircraft's age (year), travel distance (km), the flight's total airtime, departure time, arrival time (min), and the date given by day of week, day of month, and month, and a corresponding output of the flight's delay time (min); and (c) the BLOG feedback dataset (Buza 2014) contains 60000 instances of blog posts. Each blog post features a fairly large 60-dimensional input vector associated with its first 60 attributes described at https://archive.ics.uci.edu/ml/datasets/BlogFeedback, and a corresponding output measuring the number of comments in the next 24 hours. The BLOG dataset is used to evaluate the robustness of sVBSSGP to overfitting which usually occurs in training with datasets of high input dimensions. All datasets are modeled using GPs with prior covariance defined in Section 2 and split into 95% training data and 5% test data. All experimental results are averaged over 5 random splits. For the AIMPEAK, AIRLINE, and BLOG datasets, we use, respectively, $2m = 40, 40$, and 10 trigonometric basis functions to approximate the GP kernel (2) and sample $\mathbf{z}$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (Section 3.1). All experiments are run on a Linux system with Intel® Xeon® E5-2670 at 2.6GHz with 96 GB memory.

The performance of sVBSSGP is compared against the state-of-the-art VSSGP (Gal and Turner 2015) and stochastic implementations of sparse GP models based on inducing variables such as DTC+ and PIC+ (Hensman, Fusi, and Lawrence 2013; Hoang, Hoang, and Low 2015) (i.e., run with their GitHub codes) using the following metrics: (a) *Root mean square error* (RMSE) $\{|\overline{\mathbf{X}}|^{-1} \sum_{\mathbf{x}_* \in \overline{\mathbf{X}}} (y_{\mathbf{x}_*} - \widehat{\mu}_{\mathbf{x}_*})^2\}^{1/2}$ over the set $\overline{\mathbf{X}}$ of test inputs, (b) *mean negative log probability* (MNLP) $0.5 |\overline{\mathbf{X}}|^{-1} \sum_{\mathbf{x}_* \in \overline{\mathbf{X}}} \{(y_{\mathbf{x}_*} - \widehat{\mu}_{\mathbf{x}_*})^2 / \widehat{\sigma}_{\mathbf{x}_*}^2 + \log(2\pi \widehat{\sigma}_{\mathbf{x}_*}^2)\}$, and (c) training time.

AIMPEAK Dataset. This dataset is evenly partitioned into $p = 200$ disjoint subsets using k-means ($k = p$). Fig. 1 shows results of RMSEs of sVBSSGP for $\gamma = 0, 0.1, 0.2, 0.3$ that rapidly decrease by 4- to 5-fold over 30 iterations. It can also be observed that increasing γ from 0 results in higher converged RMSEs. To further investigate this, Fig. 2a reveals that sVBSSGP indeed achieves the lowest converged RMSE at $\gamma = 0$ among all tested values of γ. This confirms our hypothesis stated earlier in Remark 1 that given the learned spectral frequencies $\boldsymbol{\theta}$, the information carried by $\mathbf{s}$ becomes a "nuisance" to prediction despite its interaction with $\boldsymbol{\theta}$ in $q(\boldsymbol{\theta}, \mathbf{s}) = q(\boldsymbol{\alpha})$ during stochastic optimization. That is, when the influence of $\mathbf{s}$ on the test output is completely removed from the test conditional by setting
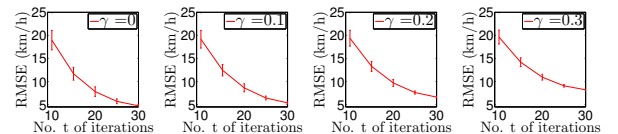

Figure 1: Graphs of RMSEs (with standard deviations) achieved by sVBSSGP vs. number t of iterations with varying values of γ for the AIMPEAK dataset.
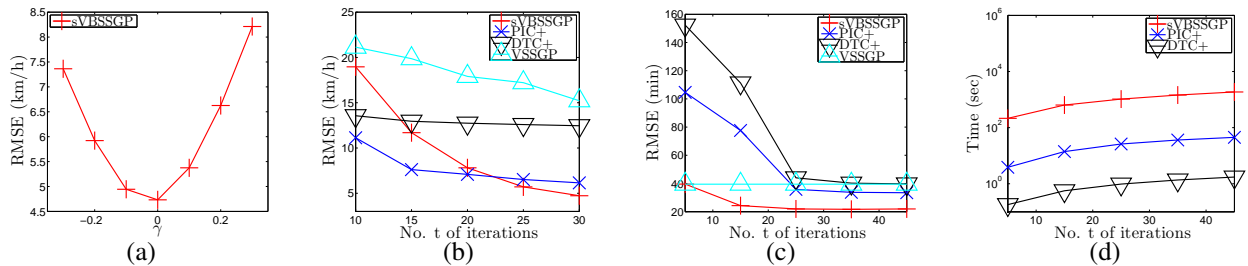
Figure 2: (a) Graph of RMSEs achieved by sVBSSGP vs. γ for the AIMPEAK dataset; graphs of RMSEs achieved by sVBSSGP, PIC+, DTC+, and VSSGP vs. number t of iterations for the (b) AIMPEAK and (c) AIRLINE datasets; and (d) graph of total training time incurred by sVBSSGP, PIC+, and DTC+ vs. number t of iterations for the AIRLINE dataset.

$\gamma = 0$ in Proposition 1, the predictions are no longer interfered by the nuisance information of s, hence explaining the lowest RMSE achieved by sVBSSGP at $\gamma = 0$.

Fig. 2b and Table 1 show that sVBSSGP ($\gamma = 0$ and $r = 20$) significantly outperforms VSSGP, DTC+, and PIC+ (250 inducing variables) in terms of RMSE after 30 iterations. To explain this, DTC+ and PIC+ find point estimates of the kernel hyperparameters, which may have resulted in their poorer performance. Though VSSGP also adopts a Bayesian treatment of the spectral frequencies, it uses a degenerate test conditional corresponding to the case of $\gamma = 1$ in Proposition 1. As a result, VSSGP imposes a highly restrictive deterministic relationship between the test output and spectral frequencies and also fails to exploit the local data for prediction (see Remark 1). The results of the MNLP metric are similar and reported in (Hoang, Hoang, and Low 2016a).

AIRLINE Dataset. This dataset is partitioned into $p = 2000$ disjoint subsets using k-means. Fig. 2c and Table 1 show that sVBSSGP ($\gamma = 0$ and $r = 5$) significantly outperforms VSSGP, DTC+, and PIC+ (512 inducing variables) in terms of RMSE after 45 iterations, as explained previously. Fig. 2d shows that the total training time of sVBSSGP increases linearly with the number t of iterations, which highlights a principled trade-off between its predictive performance and time efficiency. The training time of sVBSSGP, though longer than DTC+ and PIC+, is only 43 sec. per iteration; the training time of VSSGP is not included since its GitHub code runs on GPU instead of CPU. The results of MNLP metric are similar to that for the AIMPEAK dataset.

BLOG Dataset. This dataset is evenly partitioned into $p = 200$ disjoint subsets using k-means. Fig. 3 shows that with
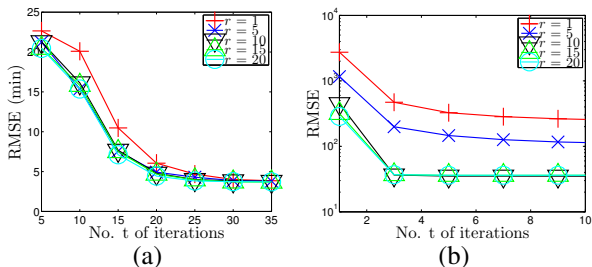

Figure 3: Graphs of RMSEs achieved by sVBSSGP vs. number t of iterations with varying number r of samples drawn from $q_t(\boldsymbol{\alpha})$ to compute predictive mean $\widehat{\mu}_{\mathbf{x}_*}$ (Section 4.2) for the (a) AIMPEAK and (b) BLOG datasets.

Dataset	sVBSSGP	DTC+	PIC+	SSGP	VSSGP
AIMPEAK	**4.73**	12.46	6.15	11.30	9.46
AIRLINE	**22.18**	39.60	33.40	n/a	39.49

Table 1: RMSEs achieved by sVBSSGP, DTC+, PIC+, SSGP (Lázaro-Gredilla et al. 2010), and VSSGP after final convergence for AIMPEAK and AIRLINE datasets.

more samples drawn from $q_t(\boldsymbol{\alpha})$ to compute predictive mean $\widehat{\mu}_{\mathbf{x}_*}$ (Section 4.2), sVBSSGP tends to converge faster and to a lower RMSE, which suggests a greater robustness to overfitting by exploiting a higher degree of Bayesian model averaging. More interestingly, the effect of overfitting appears to be more pronounced for the BLOG dataset with a much larger input dimension of 60: When $r = 1$, sVBSSGP effectively reduces to a local SSGP model utilizing the sampled spectral frequencies as its point estimate (see Proposition 1, Remark 1, and Section 4.2) and converges to the poorest performance that stops improving after 10 iterations. It can also be observed that the performance gap between sVBSSGP's with different number r of samples is much wider at early iterations, thus highlighting the practicality of our Bayesian treatment in the case of limited data.

6 Conclusion

This paper describes a novel generalized framework of sVBSSGP regression models that addresses the shortcomings of existing sparse spectrum GP models like SSGP (Lázaro-Gredilla et al. 2010) and VSSGP (Gal and Turner 2015) by adopting a Bayesian treatment of the spectral frequencies to avoid overfitting, modeling the spectral frequencies jointly in its variational distribution to enable their interaction *a posteriori*, and exploiting local data for improving the predictive performance while still being able to preserve its scalability to big data through stochastic optimization. As a result, empirical evaluation on real-world datasets (i.e., including the million-sized benchmark AIRLINE dataset) shows that our proposed sVBSSGP regression model can significantly outperform the existing sparse spectrum GP models like SSGP and VSSGP as well as the stochastic implementations of the sparse GP models based on inducing variables like DTC+ and PIC+ (Hensman, Fusi, and Lawrence 2013; Hoang, Hoang, and Low 2015).

References

Buza, K. 2014. Feedback prediction for blogs. In Spiliopoulou, M.; Schmidt-Thieme, L.; and Janning, R., eds., *Data Analysis, Machine Learning and Knowledge Discovery*. Springer International Publishing. 145–152.

Cao, N.; Low, K. H.; and Dolan, J. M. 2013. Multi-robot informative path planning for active sensing of environmental phenomena: A tale of two algorithms. In *Proc. AAMAS*.

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

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

Chen, J.; Low, K. H.; Jaillet, P.; and Yao, Y. 2015. Gaussian process decentralized data fusion and active sensing for spatiotemporal traffic modeling and prediction in mobility-on-demand systems. *IEEE T-ASE* 12(3):901–921.

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

Dolan, J. M.; Podnar, G.; Stancliff, S.; Low, K. H.; Elfes, A.; Higinbotham, J.; Hosler, J. C.; Moisan, T. A.; and Moisan, J. 2009. Cooperative aquatic sensing using the telesupervised adaptive ocean sensor fleet. In *Proc. SPIE Conference on Remote Sensing of the Ocean, Sea Ice, and Large Water Regions*, volume 7473.

Gal, Y., and Turner, R. 2015. Improving the Gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs. In *Proc. ICML*, 655–664.

Hensman, J.; Fusi, N.; and Lawrence, N. D. 2013. Gaussian processes for big data. In *Proc. UAI*, 282–290.

Hoang, T. N.; Low, K. H.; Jaillet, P.; and Kankanhalli, M. 2014a. Active learning is planning: Nonmyopic ϵ-Bayes-optimal active learning of Gaussian processes. In *Proc. ECML/PKDD Nectar Track*, 494–498.

Hoang, T. N.; Low, K. H.; Jaillet, P.; and Kankanhalli, M. 2014b. Nonmyopic ϵ-Bayes-Optimal Active Learning of Gaussian Processes. In *Proc. ICML*, 739–747.

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

Hoang, Q. M.; Hoang, T. N.; and Low, K. H. 2016a. A generalized stochastic variational Bayesian hyperparameter learning framework for sparse spectrum Gaussian process regression. arXiv:1611.06080.

Hoang, T. N.; Hoang, Q. M.; and Low, K. H. 2016b. A distributed variational inference framework for unifying parallel sparse Gaussian process regression models. In *Proc. ICML*, 382–391.

Lázaro-Gredilla, M.; Quiñonero-Candela, J.; Rasmussen, C. E.; and Figueiras-Vidal, A. R. 2010. Sparse spectrum Gaussian process regression. *JMLR* 1865–1881.

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

Low, K. H.; Chen, J.; Dolan, J. M.; Chien, S.; and Thompson, D. R. 2012. Decentralized active robotic exploration and mapping for probabilistic field classification in environmental sensing. In *Proc. AAMAS*, 105–112.

Low, K. H.; Chen, J.; Hoang, T. N.; Xu, N.; and Jaillet, P. 2014a. Recent advances in scaling up Gaussian process predictive models for large spatiotemporal data. In *Proc. DyDESS*.

Low, K. H.; Xu, N.; Chen, J.; Lim, K. K.; and Özgül, E. B. 2014b. Generalized online sparse Gaussian processes with application to persistent mobile robot localization. In *Proc. ECML/PKDD Nectar Track*, 499–503.

Low, K. H.; Yu, J.; Chen, J.; and Jaillet, P. 2015. Parallel Gaussian process regression for big data: Low-rank representation meets Markov approximation. In *Proc. AAAI*.

Low, K. H.; Dolan, J. M.; and Khosla, P. 2008. Adaptive multi-robot wide-area exploration and mapping. In *Proc. AAMAS*, 23–30.

Low, K. H.; Dolan, J. M.; and Khosla, P. 2009. Information-theoretic approach to efficient adaptive path planning for mobile robotic environmental sensing. In *Proc. ICAPS*.

Low, K. H.; Dolan, J. M.; and Khosla, P. 2011. Active Markov information-theoretic path planning for robotic environmental sensing. In *Proc. AAMAS*, 753–760.

Ouyang, R.; Low, K. H.; Chen, J.; and Jaillet, P. 2014. Multi-robot active sensing of non-stationary Gaussian process-based environmental phenomena. In *Proc. AAMAS*.

Podnar, G.; Dolan, J. M.; Low, K. H.; and Elfes, A. 2010. Telesupervised remote surface water quality sensing. In *Proc. IEEE Aerospace Conference*.

Quiñonero-Candela, J., and Rasmussen, C. E. 2005. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research* 6:1939–1959.

Titsias, M. K., and Lázaro-Gredilla, M. 2014. Doubly stochastic variational Bayes for non-conjugate inference. In *Proc. ICML*, 1971–1979.

Titsias, M. K. 2009. Variational learning of inducing variables in sparse Gaussian processes. In *Proc. AISTATS*.

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

Yu, J.; Low, K. H.; Oran, A.; and Jaillet, P. 2012. Hierarchical Bayesian nonparametric approach to modeling and learning the wisdom of crowds of urban traffic route planning agents. In *Proc. IAT*, 478–485.

Zhang, Y.; Hoang, T. N.; Low, K. H.; and Kankanhalli, M. 2016. Near-optimal active learning of multi-output Gaussian processes. In *Proc. AAAI*, 2351–2357.