A Unifying Framework of Anytime Sparse Gaussian Process Regression Models with Stochastic Variational Inference for Big Data

Trong Nghia Hoang[†] NGHIAHT@COMP.NUS.EDU.SG
Quang Minh Hoang[†] HQMINH@COMP.NUS.EDU.SG
Kian Hsiang Low[†] LOWKH@COMP.NUS.EDU.SG
[†]Department of Computer Science, National University of Singapore, Republic of Singapore

Abstract

This paper presents a novel unifying framework of anytime *sparse Gaussian process regression* (SGPR) models that can produce good predictive performance fast and improve their predictive performance over time. Our proposed unifying framework reverses the variational inference procedure to theoretically construct a non-trivial, concave functional that is maximized at the predictive distribution of any SGPR model of our choice. As a result, a stochastic natural gradient ascent method can be derived that involves iteratively following the stochastic natural gradient of the functional to improve its estimate of the predictive distribution of the chosen SGPR model and is guaranteed to achieve asymptotic convergence to it. Interestingly, we show that if the predictive distribution of the chosen SGPR model satisfies certain decomposability conditions, then the stochastic natural gradient is an unbiased estimator of the exact natural gradient and can be computed in constant time (i.e., independent of data size) at each iteration. We empirically evaluate the trade-off between the predictive performance vs. time efficiency of the anytime SGPR models on two real-world million-sized datasets.

1. Introduction

A *Gaussian process regression* (GPR) model is a Bayesian nonparametric model for performing nonlinear regression that provides a Gaussian predictive distribution with formal measures of predictive uncertainty. The expressivity of a *full-rank GPR* (FGPR) model, however, comes at a cost of cubic time in the size of the data, thus rendering it computationally impractical for training with massive datasets. To improve its scalability, a number of *sparse GPR* (SGPR) models (Lázaro-Gredilla et al., 2010; Quiñonero-Candela & Rasmussen, 2005; Snelson & Ghahramani, 2007; Titsias, 2009) exploiting low-rank approximate representations have been proposed, many of which share a similar structural assumption of conditional independence (albeit of varying degrees) based on the notion of inducing variables (Section 2) and consequently incur only linear time in the data size. The work of Quiñonero-Candela & Rasmussen (2005) has in fact presented a unifying view of such SGPR models, which include the *subset of regressors* (SoR) (Smola & Bartlett, 2001), *deterministic training conditional* (DTC) (Seeger et al., 2003), *fully independent training conditional* (FITC) (Snelson & Gharahmani, 2005), *fully independent conditional* (FIC), *partially independent training conditional* (PITC) (Schwaighofer & Tresp, 2003), and *partially independent conditional* (PIC) (Snelson & Ghahramani, 2007) approximations. To scale up these SGPR models further for performing real-time predictions necessary in many time-critical applications and decision support systems (e.g., ocean sensing (Cao et al., 2013; Dolan et al., 2009; Low et al., 2008; 2009; 2011; 2012; Podnar et al., 2010), traffic monitoring (Chen et al., 2012; 2013b; 2015; Hoang et al., 2014a;b; Low et al., 2014a;b; Ouyang et al., 2014; Xu et al., 2014; Yu et al., 2012)), the work of Gal et al. (2014) has parallelized DTC while that of Chen et al. (2013a) has parallelized FITC, FIC, PITC, and PIC to be run on multiple machines. The recent work of Low et al. (2015) has produced a spectrum of SGPR models with PIC and FGPR at the two extremes that are also amenable to parallelization on multiple machines. Ideally, these parallel SGPR models can reduce the incurred time of their centralized counterparts by a factor close to the number of machines. In practice, since the number of machines is limited due to budget constraints, their incurred time will still grow with an increasing size of data. Like their centralized counterparts, they can be trained using all the data.

A more affordable alternative is to instead train a SGPR model in an anytime fashion with a small, randomly sampled subset of the data at each iteration, which requires only a single machine. To the best of our knowledge, the

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

only notable anytime SGPR model (Hensman et al., 2013) exploits a result of Titsias (2009) that DTC can alternatively be obtained using variational inference by minimizing the *Kullback-Leibler* (KL) distance between the variational approximation and the GP posterior distribution of some latent variables given the data, from which a *stochastic natural gradient ascent* (SNGA) method can be derived to achieve an asymptotic convergence of its predictive performance to that of DTC while incurring constant time per iteration. This anytime variant of DTC promises a huge speedup if the number of sampled subsets of data needed for convergence is much smaller than the total number of possible disjoint subsets that can be formed and sampled from all the data. But, it can be observed in our experiments (Section 5) that DTC often does not predict as well as the other SGPR models (except SoR) encompassed by the unifying view of Quiñonero-Candela & Rasmussen (2005) because it imposes the most restrictive structural assumption (Snelson, 2007). This motivates us to consider the possibility of constructing an anytime variant of any SGPR model of our choice whose derived SNGA method can achieve an asymptotic convergence of its predictive performance to that of the chosen SGPR model while preserving constant time per iteration. However, no alternative formulation based on variational inference exists for any SGPR model other than DTC in order to derive such a SNGA method.

To address the above challenge, this paper presents a novel unifying framework of anytime SGPR models that can produce good predictive performance fast and improve their predictive performance over time. Our proposed unifying framework, perhaps surprisingly, reverses the variational inference procedure to theoretically construct a non-trivial, concave functional (i.e., of distributions) that is maximized at the predictive distribution of any SGPR model of our choice (Section 3). Consequently, a SNGA method can be derived that involves iteratively following the stochastic natural gradient of the functional to improve its estimate of the predictive distribution of the chosen SGPR model and is guaranteed to achieve asymptotic convergence to it. Interestingly, we show that if the predictive distribution of the chosen SGPR model satisfies certain decomposability conditions (e.g., DTC, FITC, PIC), then the stochastic natural gradient is an unbiased estimator of the exact natural gradient and can be computed in constant time (i.e., independent of data size) at each iteration (Section 4). We empirically evaluate the trade-off between the predictive performance vs. time efficiency of the anytime SGPR models spanned by our unifying framework (i.e., including state-of-the-art anytime variant of DTC (Hensman et al., 2013)) on two real-world million-sized datasets (Section 5).

2. Background and Notations

Full-Rank Gaussian Process Regression (FGPR). Let $\mathcal{X}$ be a set representing the input domain such that each d-dimensional input feature vector $\mathbf{x} \in \mathcal{X}$ is associated with a latent output variable $f_{\mathbf{x}}$. Let $\{f_{\mathbf{x}}\}_{\mathbf{x}\in\boldsymbol{\mathcal{X}}}$ denote a *Gaussian process* (GP), that is, every finite subset of $\{f_{\mathbf{x}}\}_{\mathbf{x}\in\boldsymbol{\mathcal{X}}}$ follows a multivariate Gaussian distribution. Then, the GP is fully specified by its *prior* mean $\mathbb{E}[f_{\mathbf{x}}]$, which we assume to be zero for notational simplicity, and covariance $k_{\mathbf{x}\mathbf{x}'} \triangleq \mathrm{cov}[f_{\mathbf{x}}, f_{\mathbf{x}'}]$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. Given a column vector $\mathbf{y}_{\mathcal{D}} \triangleq (y_{\mathbf{x}'})^{\top}_{\mathbf{x}'\in\mathcal{D}}$ of noisy observed outputs $y_{\mathbf{x}'} \triangleq f_{\mathbf{x}'} + \varepsilon$ for some set $\mathcal{D} \subset \mathcal{X}$ of training inputs where $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ and σ_n^2 is the noise variance, a FGPR model can perform probabilistic regression by providing a GP predictive distribution $p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}) = \mathcal{N}(K_{\mathbf{x}\mathcal{D}}(K_{\mathcal{D}\mathcal{D}} + \sigma_n^2 I)^{-1}\mathbf{y}_{\mathcal{D}}, k_{\mathbf{x}\mathbf{x}} - K_{\mathbf{x}\mathcal{D}}(K_{\mathcal{D}\mathcal{D}} + \sigma_n^2 I)^{-1}K_{\mathcal{D}\mathbf{x}})$ of the latent output $f_{\mathbf{x}}$ for any test input $\mathbf{x} \in \mathcal{X}$ where $K_{\mathbf{x}\mathcal{D}} \triangleq (k_{\mathbf{x}\mathbf{x}'})_{\mathbf{x}'\in\mathcal{D}}$, $K_{\mathcal{D}\mathcal{D}} \triangleq (k_{\mathbf{x}'\mathbf{x}''})_{\mathbf{x}',\mathbf{x}''\in\mathcal{D}}$, and $K_{\mathcal{D}\mathbf{x}} \triangleq K_{\mathbf{x}\mathcal{D}}^{\top}$. Computing the GP predictive distribution incurs $\mathcal{O}(|\mathcal{D}|^3)$ time due to inversion of $K_{\mathcal{D}\mathcal{D}}$, hence causing the FGPR model to scale poorly in the size $|\mathcal{D}|$ of data.

Sparse Gaussian Process Regression (SGPR). To improve the scalability of the FGPR model, the SGPR models encompassed by the unifying view of Quiñonero-Candela & Rasmussen (2005) exploit a vector $\mathbf{f}_{\mathcal{U}} \triangleq (f_{\mathbf{x}'})^{\top}_{\mathbf{x}'\in\mathcal{U}}$ of $|\mathcal{U}|$ 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 $p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})$. Specifically, they share a similar structural assumption (Snelson & Ghahramani, 2007) that the joint distribution of $f_{\mathbf{x}}$ and $\mathbf{f}_{\mathcal{D}} \triangleq (f_{\mathbf{x}'})^{\top}_{\mathbf{x}'\in\mathcal{D}}$ conditioned on $\mathbf{f}_{\mathcal{U}}$ factorizes across a pre-defined partition of the input domain $\mathcal{X}$ into P disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_P$That is, supposing $\mathbf{x} \in \mathcal{X}_P$,

$$p(f_{\mathbf{x}}, \mathbf{f}_{\mathcal{D}} \mid \mathbf{f}_{\mathcal{U}}) \;=\; p(f_{\mathbf{x}} \mid \mathbf{f}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}}) \prod_{i=1}^{P} p(\mathbf{f}_{\mathcal{D}_i} \mid \mathbf{f}_{\mathcal{U}}) \qquad (1)$$

where $\mathbf{f}_{\mathcal{D}_i} \triangleq (f_{\mathbf{x}'})^{\top}_{\mathbf{x}'\in\mathcal{D}_i}$ denotes a column vector of latent outputs for the disjoint subset $\mathcal{D}_i \triangleq \mathcal{X}_i \cap \mathcal{D} \subset \mathcal{D}$ of training inputs for $i = 1, \ldots, P$. Using (1), the GP predictive distribution $p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}})$ reduces to

$$\begin{aligned} p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}}) &= \int p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}})\, p(\mathbf{f}_{\mathcal{U}}|\mathbf{y}_{\mathcal{D}})\, \mathrm{d}\mathbf{f}_{\mathcal{U}} && (2) \\ &\simeq \int q^*(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}})\, q^*(\mathbf{f}_{\mathcal{U}})\, \mathrm{d}\mathbf{f}_{\mathcal{U}} && (3) \end{aligned}$$

where $\mathbf{y}_{\mathcal{D}_P} \triangleq (y_{\mathbf{x}'})^{\top}_{\mathbf{x}'\in\mathcal{D}_P}$ is a vector of noisy observed outputs for the subset $\mathcal{D}_P$ of training inputs, derivation of (2) is in Appendix C.1, and $p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}})$ and $p(\mathbf{f}_{\mathcal{U}}|\mathbf{y}_{\mathcal{D}})$ are, respectively, approximated by $q^*(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}})$ and $q^*(\mathbf{f}_{\mathcal{U}})$ in (3), as discussed in Remarks 1 and 2 below.

Remark 1. PIC sets $q^*(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}})$ as the exact test conditional $p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}})$. The other SGPR models have additionally assumed conditional independence of $f_{\mathbf{x}}$ and $\mathbf{f}_{\mathcal{D}_P}$ given $\mathbf{f}_{\mathcal{U}}$, thus resulting in $p(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}}) = p(f_{\mathbf{x}}|\mathbf{f}_{\mathcal{U}})$ and $q^*(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}}) \triangleq p(f_{\mathbf{x}}|\mathbf{f}_{\mathcal{U}})$ (e.g., see eq. 5 in (Titsias,

2009) for the case of DTC). Then, $q^*(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}})$ can be computed in $\mathcal{O}(|\mathcal{U}|^3)$ time and space (i.e., $|\mathcal{D}_P| = \mathcal{O}(|\mathcal{U}|)$ for the case of PIC), as shown in Appendix D.4.

Remark 2. The work of Titsias (2009) has approximated $p(\mathbf{f}_{\mathcal{U}}|\mathbf{y}_{\mathcal{D}})$ with a choice of $q^*(\mathbf{f}_{\mathcal{U}})$ whose resulting predictive distribution (3) coincides with that of DTC. Interestingly, other choices of $q^*(\mathbf{f}_{\mathcal{U}})$ can be derived to induce the predictive distributions (3) of the other SGPR models and computed in $\mathcal{O}(|\mathcal{D}||\mathcal{U}|^2)$ time, as shown in Appendix D.1.

Instead of computing $q^*(\mathbf{f}_{\mathcal{U}})$ directly that becomes prohibitively expensive (i.e., $\mathcal{O}(|\mathcal{D}||\mathcal{U}|^2)$ time) for massive datasets, we will derive a stochastic natural gradient ascent method that is guaranteed to achieve asymptotic convergence to $q^*(\mathbf{f}_{\mathcal{U}})$ of any SGPR model of our choice while incurring only $\mathcal{O}(|\mathcal{U}|^3)$ time per iteration (Section 4), thus producing an anytime variant of the chosen SGPR model.

Variational Inference for DTC. Variational inference can be used to derive $q^*(\mathbf{f}_{\mathcal{U}})$ of DTC as follows: A variational approximation to the posterior distribution of some latent output variables (e.g., $p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}|\mathbf{y}_{\mathcal{D}})$) can be derived analytically by minimizing their KL distance, provided that it factorizes in some way or has some parametric form that is inexpensive to evaluate (Bishop, 2006). The work of Titsias (2009) parameterizes the variational approximation $q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}})$ to the GP posterior distribution $p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}|\mathbf{y}_{\mathcal{D}})$ by

$$q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}) \quad \triangleq \quad p(\mathbf{f}_{\mathcal{D}} \mid \mathbf{f}_{\mathcal{U}})\, q(\mathbf{f}_{\mathcal{U}}) \tag{4}$$

where $p(\mathbf{f}_{\mathcal{D}}|\mathbf{f}_{\mathcal{U}})$ is the exact training conditional (8) and $q(\mathbf{f}_{\mathcal{U}}) \triangleq \mathcal{N}(\mu, \Sigma)$. The result below reveals how μ and Σ, which depend on training data $(\mathcal{D}, \mathbf{y}_{\mathcal{D}})$, can be selected to minimize the KL distance $D_{\text{KL}}(q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}})\|p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}|\mathbf{y}_{\mathcal{D}}))$:

Lemma 1 *For any PDF $q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}})$ and $p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}, \mathbf{y}_{\mathcal{D}})$,*

$$\log p(\mathbf{y}_{\mathcal{D}}) \quad = \quad L(q) \; + \; D_{\text{KL}}\left(q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}) \parallel p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}} \mid \mathbf{y}_{\mathcal{D}})\right)$$

where the functional $L(q)$ is defined as

$$L(q) \quad \triangleq \quad \int q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}) \log\left(\frac{p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}, \mathbf{y}_{\mathcal{D}})}{q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}})}\right) \mathrm{d}\mathbf{f}_{\mathcal{D}}\, \mathrm{d}\mathbf{f}_{\mathcal{U}} \; . \tag{5}$$

Its proof is in Appendix C.2. Lemma 1 implies that minimizing $D_{\text{KL}}(q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}})\|p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}|\mathbf{y}_{\mathcal{D}}))$ is equivalent to maximizing $L(q)$ since $p(\mathbf{y}_{\mathcal{D}})$ is constant with respect to $q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}})$. Using the parameterization in (4), $L(q)$ becomes a concave function in μ and Σ that is maximized when its gradient is zero. So, $q(\mathbf{f}_{\mathcal{U}}) \triangleq \mathcal{N}(\mu, \Sigma)$ can be optimized by solving for μ and Σ such that $\partial L/\partial\mu = 0$ and $\partial L/\partial\Sigma = 0$.

Remark 1. From Lemma 1, since $D_{\text{KL}}(.\|.)$ is non-negative, $\log p(\mathbf{y}_{\mathcal{D}}) \geq L(q)$, which recovers the variational lower bound of Titsias (2009) by setting $p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}, \mathbf{y}_{\mathcal{D}})$ as the GP joint distribution.

Remark 2. The work of Titsias (2009) is originally intended to jointly optimize $q(\mathbf{f}_{\mathcal{U}})$, inducing inputs $\mathcal{U}$, and hyperparameters of $k(.,.)$. In the context of our work here, by assuming $\mathcal{U}$ and the hyperparameters to be given, the optimal $q(\mathbf{f}_{\mathcal{U}}) \equiv q^*(\mathbf{f}_{\mathcal{U}})$ induces the predictive distribution (3) of DTC when $q^*(f_{\mathbf{x}}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_{\mathcal{U}}) \triangleq p(f_{\mathbf{x}}|\mathbf{f}_{\mathcal{U}})$ (Titsias, 2009).

3. Reverse Variational Inference

This section introduces a novel, interesting use of variational inference, which we term *reverse variational inference*, to theoretically construct a concave, differential functional $L(q)$ that is maximized at $q(\mathbf{f}_{\mathcal{U}}) \equiv q^*(\mathbf{f}_{\mathcal{U}})$ of any SGPR model of our choice; we call this requirement **R1**. The functional $L(q)$ allows us to derive a *stochastic natural gradient ascent* (SNGA) method (Section 4) that takes small, iterative steps in the direction of the stochastic natural gradient of $L(q)$ to improve its estimate $q(\mathbf{f}_{\mathcal{U}})$ of $q^*(\mathbf{f}_{\mathcal{U}})$ and is guaranteed to achieve asymptotic convergence of $q(\mathbf{f}_{\mathcal{U}})$ to $q^*(\mathbf{f}_{\mathcal{U}})$ if the step sizes are scheduled appropriately (Robbins & Monro, 1951). If the stochastic natural gradient of $L(q)$ can be computed in constant time (i.e., independent of data size $|\mathcal{D}|$) and we call this requirement **R2**, then such a SNGA method is desirable in practice due to its anytime behavior of improving the estimation of $q^*(\mathbf{f}_{\mathcal{U}})$ over time. In Section 4, we will establish sufficient conditions for $q^*(\mathbf{f}_{\mathcal{U}})$ to satisfy **R2**, thus entailing a unifying framework of anytime SGPR models.

Constructing $L(q)$ to Satisfy R1. Let $q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}})$ be factorized according to (4) and $q^*(\mathbf{f}_{\mathcal{U}}) \triangleq \mathcal{N}(\mu^*, \Sigma^*)$ where μ^* and Σ^* depend on training data $(\mathcal{D}, \mathbf{y}_{\mathcal{D}})$. Our key idea is to derive a joint distribution $p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}, \mathbf{y}_{\mathcal{D}})$ such that $L(q)$ is maximized at $q(\mathbf{f}_{\mathcal{U}}) \equiv q^*(\mathbf{f}_{\mathcal{U}})$ of any SGPR model of our choice, which remains largely unexplored except for DTC: A result of Titsias (2009) has established that $q^*(\mathbf{f}_{\mathcal{U}})$ of DTC (Appendix D.1.3) maximizes $L(q)$ when $p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}, \mathbf{y}_{\mathcal{D}})$ coincides with the GP joint distribution, hence satisfying **R1** for DTC only. Such a functional $L(q)$ is then shown by Hensman et al. (2013) to satisfy **R2** and can consequently be exploited for deriving a SNGA method to produce an anytime variant of DTC. However, this work neither extends nor discusses how $L(q)$ can be derived for other choices of $q^*(\mathbf{f}_{\mathcal{U}})$ and the conditions under which they will satisfy **R1** and **R2**. We will address both these issues in this section and the next, respectively. In addition, the predictive performance of their SNGA method is severely limited by the highly restrictive structural assumption of DTC. Finally, their anytime variant of DTC turns out to be a special case spanned by our unifying framework of anytime SGPR models (Section 4).

For the rest of this section, we will first evaluate $L(q)$ to a concave function in μ and Σ (i.e., Theorems 1, 2, and 3) subject to our factorization of $q(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}})$ in (4) and $p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}, \mathbf{y}_{\mathcal{D}})$ in (10). Then, we will show how the parameters defining $p(\mathbf{f}_{\mathcal{D}}, \mathbf{f}_{\mathcal{U}}, \mathbf{y}_{\mathcal{D}})$ can be appropriately selected such that the induced $L(q)$ (5) is maximized at

$q^*(\mathbf{f}_\mathcal{U}) \triangleq \mathcal{N}(\mu^*, \Sigma^*)$ of our choice (i.e., Theorem 4).

Theorem 1 *Let $q(\mathbf{f}_\mathcal{D}|\mathbf{f}_\mathcal{U}) \triangleq q(\mathbf{f}_\mathcal{D}, \mathbf{f}_\mathcal{U})/q(\mathbf{f}_\mathcal{U})$. Then,*

$$L(q) = \int q(\mathbf{f}_\mathcal{U})\ L_\mathcal{U}(q)\ \mathrm{d}\mathbf{f}_\mathcal{U} - D_{\mathrm{KL}}(q(\mathbf{f}_\mathcal{U}) \parallel p(\mathbf{f}_\mathcal{U})) \quad (6)$$

where the functional $L_\mathcal{U}(q)$ is defined as

$$L_\mathcal{U}(q) \ \triangleq\ \int q(\mathbf{f}_\mathcal{D} \mid \mathbf{f}_\mathcal{U})\ \log \frac{p(\mathbf{f}_\mathcal{D}, \mathbf{y}_\mathcal{D} \mid \mathbf{f}_\mathcal{U})}{q(\mathbf{f}_\mathcal{D} \mid \mathbf{f}_\mathcal{U})}\ \mathrm{d}\mathbf{f}_\mathcal{D}\ . \quad (7)$$

Its proof is in Appendix C.3. The parameterization of $q(\mathbf{f}_\mathcal{U}) \triangleq \mathcal{N}(\mu, \Sigma)$ and factorization of $q(\mathbf{f}_\mathcal{D}, \mathbf{f}_\mathcal{U})$ using (4) entail $q(\mathbf{f}_\mathcal{D}|\mathbf{f}_\mathcal{U})$ being set as the exact training conditional:

$$q(\mathbf{f}_\mathcal{D}|\mathbf{f}_\mathcal{U}) \triangleq p(\mathbf{f}_\mathcal{D} \mid \mathbf{f}_\mathcal{U}) = \mathcal{N}(P_{\mathcal{D}\mathcal{U}}\, \mathbf{f}_\mathcal{U}, K_{\mathcal{D}\mathcal{D}} - Q_{\mathcal{D}\mathcal{D}}) \quad (8)$$

where $P_{\mathcal{D}\mathcal{U}} \triangleq K_{\mathcal{D}\mathcal{U}}K_{\mathcal{U}\mathcal{U}}^{-1}$ and $Q_{\mathcal{D}\mathcal{D}} \triangleq K_{\mathcal{D}\mathcal{U}}K_{\mathcal{U}\mathcal{U}}^{-1}K_{\mathcal{U}\mathcal{D}}$. Using (8), $L_\mathcal{U}(q)$ is reduced to a quadratic function of $\mathbf{f}_\mathcal{U}$:

Theorem 2 *By substituting (8) into $L_\mathcal{U}(q)$ (7),*

$$L_\mathcal{U}(q) = -\frac{1}{2\sigma_n^2}\mathbf{f}_\mathcal{U}^\top P_{\mathcal{D}\mathcal{U}}^\top P_{\mathcal{D}\mathcal{U}}\ \mathbf{f}_\mathcal{U} + \frac{1}{\sigma_n^2}\mathbf{f}_\mathcal{U}^\top P_{\mathcal{D}\mathcal{U}}^\top \mathbf{y}_\mathcal{D} + C \quad (9)$$

where the constant C absorbs all terms independent of $\mathbf{f}_\mathcal{U}$.

Its proof is in Appendix B.1. Then, we factorize

$$p(\mathbf{f}_\mathcal{D}, \mathbf{f}_\mathcal{U}, \mathbf{y}_\mathcal{D}) \ \triangleq\ p(\mathbf{y}_\mathcal{D} \mid \mathbf{f}_\mathcal{D})\, p(\mathbf{f}_\mathcal{D} \mid \mathbf{f}_\mathcal{U})\, p(\mathbf{f}_\mathcal{U}) \quad (10)$$

where $p(\mathbf{y}_\mathcal{D}|\mathbf{f}_\mathcal{D}) = \mathcal{N}(\mathbf{f}_\mathcal{D}, \sigma_n^2 I)$, $p(\mathbf{f}_\mathcal{D}|\mathbf{f}_\mathcal{U})$ is the exact training conditional (8), and we let $p(\mathbf{f}_\mathcal{U}) \triangleq \mathcal{N}(\nu, \Lambda^{-1})$ (instead of defining $p(\mathbf{f}_\mathcal{U})$ as a GP prior by Titsias (2009) that makes $p(\mathbf{f}_\mathcal{D}, \mathbf{f}_\mathcal{U}, \mathbf{y}_\mathcal{D})$ a GP joint distribution) where Λ denotes a precision matrix. Then, by applying Theorems 1 and 2, $L(q)$ becomes a concave function in both μ and Σ:

Theorem 3 *By substituting $q(\mathbf{f}_\mathcal{U}) \triangleq \mathcal{N}(\mu, \Sigma)$, $p(\mathbf{f}_\mathcal{U}) \triangleq \mathcal{N}(\nu, \Lambda^{-1})$, and $L_\mathcal{U}(q)$ (9) into $L(q)$ (6),*

$$\begin{aligned} L(q) \ =\ & -\frac{1}{2}\mu^\top \Psi \mu - \frac{1}{2}\mathrm{tr}(\Psi\Sigma)\ +\ \frac{1}{2}\log|\Sigma|\ + \\ & \mu^\top \left((1/\sigma_n^2)P_{\mathcal{D}\mathcal{U}}^\top\, \mathbf{y}_\mathcal{D} + \Lambda\nu\right)\ +\ C' \end{aligned} \quad (11)$$

where $\Psi \triangleq (1/\sigma_n^2)P_{\mathcal{D}\mathcal{U}}^\top P_{\mathcal{D}\mathcal{U}} + \Lambda$ and the constant C' absorbs all terms independent of μ and Σ.

Its proof is in Appendix B.2. Using Theorem 3, the conditions for the parameters ν and Λ defining $p(\mathbf{f}_\mathcal{U})$ (or, equivalently, $p(\mathbf{f}_\mathcal{D}, \mathbf{f}_\mathcal{U}, \mathbf{y}_\mathcal{D})$) can be determined such that $L(q)$ is maximized at $q(\mathbf{f}_\mathcal{U}) \equiv q^*(\mathbf{f}_\mathcal{U})$ by making its derivatives with respect to μ and Σ go to zero at $(\mu, \Sigma) = (\mu^*, \Sigma^*)$:

Theorem 4 *If ν and Λ satisfy the following conditions:*

$$\Lambda\nu + \frac{1}{\sigma_n^2}P_{\mathcal{D}\mathcal{U}}^\top\, \mathbf{y}_\mathcal{D} \ =\ \left(\frac{1}{\sigma_n^2}P_{\mathcal{D}\mathcal{U}}^\top P_{\mathcal{D}\mathcal{U}} + \Lambda\right)\mu^* \quad (12)$$

$$\textit{and} \quad \Lambda \ =\ \Sigma^{*-1} - \frac{1}{\sigma_n^2}P_{\mathcal{D}\mathcal{U}}^\top P_{\mathcal{D}\mathcal{U}}\ , \quad (13)$$

then $L(q)$ is maximized at $q(\mathbf{f}_\mathcal{U}) \equiv q^*(\mathbf{f}_\mathcal{U})$.

Its proof is in Appendix B.3.

Remark. (12) and (13) define the space of feasible pairs (ν, Λ) guaranteeing that $L(q)$ is maximized at $(\mu, \Sigma) = (\mu^*, \Sigma^*)$. Interestingly, it is not necessary to explicitly solve for (ν, Λ) in order to construct $L(q)$ that is maximized at $q(\mathbf{f}_\mathcal{U}) \equiv q^*(\mathbf{f}_\mathcal{U})$, as shown in (34) in Appendix B.3.

4. Anytime Sparse GP Regression Models

Using Theorem 4, a gradient ascent method that is guaranteed to achieve asymptotic convergence of (μ, Σ) to (μ^*, Σ^*) can now be derived. Specifically, it starts with randomly initialized $(\mu, \Sigma) = (\mu^0, \Sigma^0)$ and iterates the following gradient ascent update until convergence:

$$\mu^{t+1} = \mu^t + \rho_t\, \frac{\partial L}{\partial \mu}(\mu^t, \Sigma^t),\ \Sigma^{t+1} = \Sigma^t + \rho_t\, \frac{\partial L}{\partial \Sigma}(\mu^t, \Sigma^t) \quad (14)$$

where ρ_t is the step size and $\frac{\partial L}{\partial \mu}(\mu^t, \Sigma^t)$ and $\frac{\partial L}{\partial \Sigma}(\mu^t, \Sigma^t)$ denote, respectively, $\partial L/\partial \mu$ and $\partial L/\partial \Sigma$ (i.e., see (35) and (36) in Appendix B.3 for their expressions) being evaluated at $(\mu, \Sigma) = (\mu^t, \Sigma^t)$. This method is guaranteed to converge if (a) $\sum_t \rho_t = +\infty$ and (b) $\sum_t \rho_t^2 < +\infty$, which is a well-known result in optimization. For example, one possible schedule is $\rho_t = \rho_0/(1 + \tau\rho_0 t)^\kappa$ where τ, κ, and ρ_0 are determined empirically. However, evaluating the exact gradient $(\partial L/\partial \mu, \partial L/\partial \Sigma)$ requires computing $q^*(\mathbf{f}_\mathcal{U})$ directly that incurs $\mathcal{O}(|\mathcal{D}||\mathcal{U}|^2)$ time (Appendix D.1), which is prohibitively expensive for massive datasets.

Stochastic Gradient Ascent (SGA). To sidestep the above scalability issue, we adopt the *stochastic gradient ascent* (SGA) method (Robbins & Monro, 1951) that replaces the exact gradient in (14) with its stochastic gradient $(\partial\widehat{L}/\partial\mu, \partial\widehat{L}/\partial\Sigma)$. The key idea is to iteratively compute $(\partial\widehat{L}/\partial\mu, \partial\widehat{L}/\partial\Sigma)$ in an efficient manner by randomly sampling a small block of data of size $|\mathcal{U}|$ whose incurred time per iteration is independent of the data size $|\mathcal{D}|$. We will prove in Theorem 5 later that such a stochastic gradient is an unbiased estimator of the exact gradient. As a result, (14) is also guaranteed to converge using the above schedule of $\{\rho_t\}_t$. To derive $(\partial\widehat{L}/\partial\mu, \partial\widehat{L}/\partial\Sigma)$, the following decomposability conditions for (μ^*, Σ^*) are necessary:

Decomposability Conditions. *Let $F'(\mathcal{U})$ and $G'(\mathcal{U})$ ($F(\mathcal{U},\ \mathbf{y}_{\mathcal{D}_i})$ and $G(\mathcal{U},\ \mathbf{y}_{\mathcal{D}_i})$) denote arbitrary functions depending on $\mathcal{U}$ ($\mathcal{U}$ and $(\mathcal{D}_i,\ \mathbf{y}_{\mathcal{D}_i})$) only. The decomposability conditions for (μ^*, Σ^*) are*

$$\Sigma^{*-1} \ =\ F'(\mathcal{U}) + \sum_{i=1}^{P} F(\mathcal{U},\ \mathbf{y}_{\mathcal{D}_i})\ , \quad (15)$$

$$\Sigma^{*-1}\mu^* \ =\ G'(\mathcal{U}) + \sum_{i=1}^{P} G(\mathcal{U},\ \mathbf{y}_{\mathcal{D}_i})\ . \quad (16)$$

Remark 1. Though (15) and (16) may appear rather awkward when viewed using the moment parameterization

$q^*(\mathbf{f}_\mathcal{U}) \triangleq \mathcal{N}(\mu^*, \Sigma^*)$, they can alternatively be perceived as simple additive decomposability of the natural parameters $\theta_1 \triangleq \Sigma^{*-1}\mu^*$ and $\theta_2 \triangleq -(1/2)\Sigma^{*-1}$, which define the canonical parameterization of $q^*(\mathbf{f}_\mathcal{U})$ (Appendix E).

Remark 2. Interestingly, this canonical view reveals a systematic way to construct new SGPR models from existing ones that satisfy (15) and (16): Given a set $\{q^*_m(\mathbf{f}_\mathcal{U})\}_{m=1}^M$ of M SGPR models specified by their respective canonical parameterizations $\{(\theta_{1,m}, \theta_{2,m})\}_{m=1}^M$ satisfying (15) and (16), if they share the same test conditional $q^*(f_\mathbf{x}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_\mathcal{U})$ in (3), then any new SGPR model constructed with $\widehat{\theta}_1 \triangleq \sum_{m=1}^M \alpha_m \theta_{1,m}$ and $\widehat{\theta}_2 \triangleq \sum_{m=1}^M \alpha_m \theta_{2,m}$ (i.e., $\{\alpha_m\}_{m=1}^M$ is a set of linear coefficients) also satisfies (15) and (16).

In practice, the decomposability conditions (15) and (16) are satisfied by many SGPR models that share a similar structural assumption of conditional independence in (1) such as SoR, DTC, FITC, FIC, PITC, and PIC. For interested readers, $\{F(\mathcal{U}, \mathbf{y}_{\mathcal{D}_i})\}_{i=1}^P$, $\{G(\mathcal{U}, \mathbf{y}_{\mathcal{D}_i})\}_{i=1}^P$, $F'(\mathcal{U})$, and $G'(\mathcal{U})$ of these SGPR models are derived in Appendix D.2. If our choice of (μ^*, Σ^*) (i.e., $q^*(\mathbf{f}_\mathcal{U})$) satisfies (15) and (16), then the stochastic gradient $(\partial\widehat{L}/\partial\mu, \partial\widehat{L}/\partial\Sigma)$ is an unbiased estimator of the exact gradient $(\partial L/\partial\mu, \partial L/\partial\Sigma)$:

Theorem 5 *Let $\mathcal{S}$ be a set of i.i.d. samples (i.e., $|\mathcal{S}| > 0$) drawn from a uniform distribution over $\{1, 2, \ldots, P\}$ and*

$$\frac{\partial\widehat{L}}{\partial\mu} \triangleq G'(\mathcal{U}) - F'(\mathcal{U})\mu + \frac{P}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} G(\mathcal{U}, \mathbf{y}_{\mathcal{D}_s}) - F(\mathcal{U}, \mathbf{y}_{\mathcal{D}_s})\mu\ , \tag{17}$$

$$\frac{\partial\widehat{L}}{\partial\Sigma} \triangleq \frac{1}{2}\Sigma^{-1} - \frac{1}{2}F'(\mathcal{U}) - \frac{P}{2|\mathcal{S}|}\sum_{s\in\mathcal{S}} F(\mathcal{U}, \mathbf{y}_{\mathcal{D}_s})\ . \tag{18}$$

If (μ^*, Σ^*) satisfies (15) and (16), then $\mathbb{E}[\partial\widehat{L}/\partial\mu] = \partial L/\partial\mu$ and $\mathbb{E}[\partial\widehat{L}/\partial\Sigma] = \partial L/\partial\Sigma$.

Its proof is in Appendix B.4.

Remark. By assuming $|\mathcal{D}_i| = \mathcal{O}(|\mathcal{U}|)$ for $i = 1, \ldots, P$, computing stochastic gradient $(\partial\widehat{L}/\partial\mu, \partial\widehat{L}/\partial\Sigma)$ (i.e., (17) and (18)) incurs time independent of data size $|\mathcal{D}|$, in particular, $\mathcal{O}(|\mathcal{S}||\mathcal{U}|^3)$ time for SoR, DTC, FITC, FIC, PITC, and PIC (Appendix D.3) that reduces to $\mathcal{O}(|\mathcal{U}|^3)$ time by setting $|\mathcal{S}| = 1$ in our experiments. Also, since (μ, Σ) (i.e., $q(\mathbf{f}_\mathcal{U})$) is readily available from the SGA update (14), the prediction time (i.e., time incurred to analytically integrate $q^*(f_\mathbf{x}|\mathbf{y}_{\mathcal{D}_P}, \mathbf{f}_\mathcal{U})$ with $q^*(\mathbf{f}_\mathcal{U}) \equiv q(\mathbf{f}_\mathcal{U})$ in (3)) is independent of $|\mathcal{D}|$ (Appendix D.4). So, if the number t of iterations of SGA update is much smaller than $\min(|\mathcal{D}|/|\mathcal{U}|, P)$, then the anytime variants spanned by our unifying framework achieve a huge computational gain (i.e., $\mathcal{O}(t|\mathcal{U}|^3)$ time) over their corresponding SGPR models that incur $\mathcal{O}(|\mathcal{D}||\mathcal{U}|^2)$ time (Appendix D.1).

Stochastic Natural Gradient Ascent (SNGA). As the standard gradient of a function (e.g., $L(q)$) only points in the direction of the steepest ascent when the space of its parameters (e.g., (μ, Σ)) is Euclidean (Amari, 1998), the SGA update (14) has implicitly defined the parameter space of $q(\mathbf{f}_\mathcal{U})$ using the Euclidean distance between two candidate parameters, which unfortunately appears to be a poor dissimilarity measure between their corresponding distributions (Hoffman et al., 2013). To capture a more meaningful notion of dissimilarity, the parameter space of $q(\mathbf{f}_\mathcal{U})$ is redefined using the symmetrized KL distance, which is a natural dissimilarity measure between two probability distributions (Hoffman et al., 2013). This motivates the use of the natural gradient of $L(q)$ in the Euclidean space that can be equivalently considered its standard gradient in the redefined parameter space implementing the symmetrized KL distance (Amari, 1998). Such a natural gradient of $L(q)$ will be used to derive the *stochastic natural gradient ascent* (SNGA) method. Intuitively, SNGA can be regarded as another version of SGA that operates in a different parameter space defined with a different distance metric. Therefore, both converge to the same optimal parameters, although SNGA is empirically demonstrated to converge faster than SGA (Amari, 1998) when an objective function $L(q)$ is optimized with respect to a parameterized distribution $q(\mathbf{f}_\mathcal{U})$. This is expected since the symmetrized KL distance is more accurate than the Euclidean distance in measuring the dissimilarity between parameterized distributions.

To derive the natural gradient of $L(q)$, the moment parameterization of $q(\mathbf{f}_\mathcal{U}) \triangleq \mathcal{N}(\mu, \Sigma)$ is first replaced by its canonical counterpart $q(\mathbf{f}_\mathcal{U}|\theta)$:

$$q(\mathbf{f}_\mathcal{U} \mid \theta) \triangleq \mathcal{N}(\mu, \Sigma) = h(\mathbf{f}_\mathcal{U}) \exp(\theta^\top \mathbf{T}(\mathbf{f}_\mathcal{U}) - A(\theta))$$

where $\mathbf{T}(\mathbf{f}_\mathcal{U}) \triangleq \left(\mathbf{f}_\mathcal{U}; \mathrm{vec}(\mathbf{f}_\mathcal{U}\mathbf{f}_\mathcal{U}^\top)\right)$, $h(\mathbf{f}_\mathcal{U}) \triangleq (2\pi)^{-|\mathcal{U}|/2}$, $A(\theta)$ is simply a normalizing function guaranteeing that $q(\mathbf{f}_\mathcal{U}|\theta)$ integrates to unity, and the natural parameters $\theta \triangleq (\theta_1; \mathrm{vec}(\theta_2))$ where $\theta_1 = \Sigma^{-1}\mu$ and $\theta_2 = -(1/2)\Sigma^{-1}$. In particular, the metric distance defining the parameter space is given by the Riemannian metric tensor $H(\theta)$ (Amari, 1998) that corresponds to the identity matrix when the Euclidean metric is used. Otherwise, when the parameter space implements the symmetrized KL distance, the work of Hoffman et al. (2013) has shown that $H(\theta)$ is defined by the Fisher information matrix (Amari, 1998):

$$H(\theta) \triangleq -\mathbb{E}_{\mathbf{f}_\mathcal{U}|\theta}\left[\frac{\partial^2 \log q(\mathbf{f}_\mathcal{U} \mid \theta)}{\partial\theta\partial\theta^\top}\right] = \frac{\partial^2 A(\theta)}{\partial\theta\partial\theta^\top}\ . \tag{19}$$

The last equality is formally verified in Appendix E.2. Let $\partial L/\partial\theta$ be the standard gradient of $L(q)$ with respect to θ. Then, its natural gradient is defined as $\partial\overline{L}/\partial\theta \triangleq H(\theta)^{-1}\partial L/\partial\theta$. To express $\partial\overline{L}/\partial\theta$ in terms of μ and Σ, let $\eta \triangleq [\eta_1; \mathrm{vec}(\eta_2)]$ where $\eta_1 \triangleq \mu$ and $\eta_2 \triangleq \mu\mu^\top + \Sigma$. It can be verified that $\mathbb{E}[\mathbf{T}(\mathbf{f}_\mathcal{U})] = \eta$ (Appendix E.1), which implies $\partial\eta/\partial\theta = H(\theta)$ (Appendix E.3). Using this result,

$$\frac{\partial \overline{L}}{\partial \theta} \triangleq H(\theta)^{-1}\frac{\partial L}{\partial \theta} = H(\theta)^{-1}\frac{\partial \eta}{\partial \theta}\frac{\partial L}{\partial \eta} = \frac{\partial L}{\partial \eta} . \quad (20)$$

The last equality is due to $\partial\eta/\partial\theta = H(\theta)$. So, the natural gradient can be evaluated by taking the derivative of $L(q)$ with respect to η (20). To simplify the derivation, the partial derivatives of $L(q)$ are taken with respect to η_1 and η_2 instead of differentiating it with η directly. To achieve this, $L(q)$ is first represented as a function of η_1 and η_2:

$$\begin{aligned} L(q) &= \frac{1}{2}\log|\eta_2 - \eta_1\eta_1^\top| + \eta_1^\top\left(\frac{1}{\sigma_n^2}P_{\mathcal{DU}}^\top \mathbf{y}_{\mathcal{D}} + \Lambda\nu\right) \\ &- \frac{1}{2}\eta_1^\top\Psi\eta_1 - \frac{1}{2}\mathrm{tr}(\Psi\eta_2 - \Psi\eta_1\eta_1^\top) + C' \end{aligned}$$

which can be straightforwardly verified using (11) and η's definition. The natural gradient of $L(q)$ is then given by

$$\frac{\partial L}{\partial \eta_1} = -(\eta_2 - \eta_1\eta_1^\top)^{-1}\eta_1 + \frac{1}{\sigma_n^2}P_{\mathcal{DU}}^\top \mathbf{y}_{\mathcal{D}} + \Lambda\nu , \quad (21)$$

$$\frac{\partial L}{\partial \eta_2} = \frac{1}{2}\left(\left(\eta_2 - \eta_1\eta_1^\top\right)^{-1} - \Psi\right) . \quad (22)$$

Finally, note that if (ν, Λ) is chosen to satisfy (12) and (13) to guarantee that $L(q)$ is maximized at $q(\mathbf{f}_\mathcal{U}) \equiv q^*(\mathbf{f}_\mathcal{U})$ (Theorem 4), then $\Psi = \Sigma^{*-1}$ and $(1/\sigma_n^2)P_{\mathcal{DU}}^\top \mathbf{y}_{\mathcal{D}} + \Lambda\nu = \Sigma^{*-1}\mu^*$. In addition, by definition, $\theta_1 = (\eta_2 - \eta_1\eta_1^\top)^{-1}\eta_1$ and $\theta_2 = -(1/2)(\eta_2 - \eta_1\eta_1^\top)^{-1}$. Hence, (21) and (22) can be rewritten as

$$\frac{\partial L}{\partial \eta_1} = \Sigma^{*-1}\mu^* - \theta_1 \text{ and } \frac{\partial L}{\partial \eta_2} = -\theta_2 - \frac{1}{2}\Sigma^{*-1} . \quad (23)$$

So, if (μ^*, Σ^*) satisfies the decomposability conditions (15) and (16), then it is possible to derive a stochastic natural gradient that is an unbiased estimator of the exact natural gradient (23), as formalized in the result below:

Theorem 6 *Let $\mathcal{S}$ be a set of i.i.d. samples (i.e., $|\mathcal{S}| > 0$) drawn from a uniform distribution over $\{1, 2, \ldots, P\}$ and*

$$\frac{\partial \widehat{L}}{\partial \eta_1} \triangleq G'(\mathcal{U}) - \theta_1 + \frac{P}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} G(\mathcal{U}, \mathbf{y}_{\mathcal{D}_s}) , \quad (24)$$

$$\frac{\partial \widehat{L}}{\partial \eta_2} \triangleq -\theta_2 - \frac{1}{2}F'(\mathcal{U}) - \frac{P}{2|\mathcal{S}|}\sum_{s\in\mathcal{S}} F(\mathcal{U}, \mathbf{y}_{\mathcal{D}_s}) . \quad (25)$$

If (μ^*, Σ^*) satisfies (15) and (16), then $\mathbb{E}[\partial\widehat{L}/\partial\eta_1] = \partial L/\partial\eta_1$ and $\mathbb{E}[\partial\widehat{L}/\partial\eta_2] = \partial L/\partial\eta_2$.

Its proof is in Appendix B.5. The gradient ascent update in (14) can now be revised to

$$\theta_1^{t+1} = \theta_1^t + \rho_t\frac{\partial \widehat{L}}{\partial \eta_1}(\theta_1^t, \theta_2^t),\ \theta_2^{t+1} = \theta_2^t + \rho_t\frac{\partial \widehat{L}}{\partial \eta_2}(\theta_1^t, \theta_2^t) \quad (26)$$

such that the parameters (μ, Σ) of $q(\mathbf{f}_\mathcal{U})$ can be recovered from its natural parameters θ by setting $(\mu^t, \Sigma^t) = (-(1/2)(\theta_2^t)^{-1}\theta_1^t, -(1/2)(\theta_2^t)^{-1})$. Therefore, if $q^*(\mathbf{f}_\mathcal{U}) \triangleq \mathcal{N}(\mu^*, \Sigma^*)$ is selected as that of DTC, then (26) recovers the SNGA method of Hensman et al. (2013) to produce an anytime variant of DTC, which is a special case spanned by our unifying framework of anytime SGPR models.

5. Experiments and Discussion

This section empirically evaluates the predictive performance and time efficiency of anytime SGPR models[1] such as the anytime variants of PIC and FITC, which we, respectively, call PIC+ and FITC+, and the state-of-the-art anytime variant of DTC (Hensman et al., 2013), which we name DTC+[2], spanned by our unifying framework on two real-world datasets of a few million in size:

(a) The *EMULATE mean sea level pressure* (EMSLP) dataset (Ansell et al., 2006) of size 1278250 spans a 5° lat.-lon. grid bounded within lat. 25-70N and lon. 70W-50E from 1900 to 2003. Each input denotes a 6-dimensional feature vector of latitude, longitude, year, month, day, and incremental day count (starting from 0 on first day). The output is the mean sea level pressure (Pa).

(b) The AIRLINE dataset contains 2055733 records of information about every commercial flight in the USA from January to April 2008. The input denotes a 8-dimensional feature vector of the age of the aircraft (i.e., no. of years in service), travel distance (km), airtime, departure and arrival time (min.) as well as day of the week, day of the month, and month. The output is the delay time (min.) of the flight.

Both datasets are modeled using GPs whose prior covariance is defined by the squared exponential covariance function $k_{\mathbf{xx}'} \triangleq \sigma_s^2\exp(-0.5(\mathbf{x}-\mathbf{x}')^\top\Upsilon^{-2}(\mathbf{x}-\mathbf{x}'))$ with a diagonal matrix Υ of d length-scale components and signal variance σ_s^2 being its defining hyperparameters. These hyperparameters together with the noise variance σ_n^2 are learned by generalizing the distributed, variational DTC-like learning framework of Gal et al. (2014) to account for the more relaxed structural assumptions of PIC and FITC. Such a generalization can then handle massive datasets by distributing the computational load of learning the hyperparameters of PIC and FITC among parallel computing nodes; its details are deferred to a separate paper since the focus of our work here is on scaling up the existing SGPR models while assuming that the hyperparameters are learned in advance. On a separate note, learning these hyperparameters in an anytime fashion is highly non-trivial and beyond the scope of this paper, which we intend to pursue in the future as a continuation of our current work.

For each dataset, 5% is randomly selected and set aside as test data $\mathcal{S}$. The remaining data (i.e., training data) is partitioned into P blocks using k-means (i.e., $k = P$). All experiments are run on a Linux system with Intel® Xeon® E5-2670 at 2.6GHz with 96 GB memory.

[1] In the case of performing multiple predictions for single test inputs, the predictive means of FITC and DTC coincide with that of FIC and SoR, respectively.

[2] DTC+ coincides with stochastic variational inference for GPs in (Hensman et al., 2013), as discussed in Section 3.

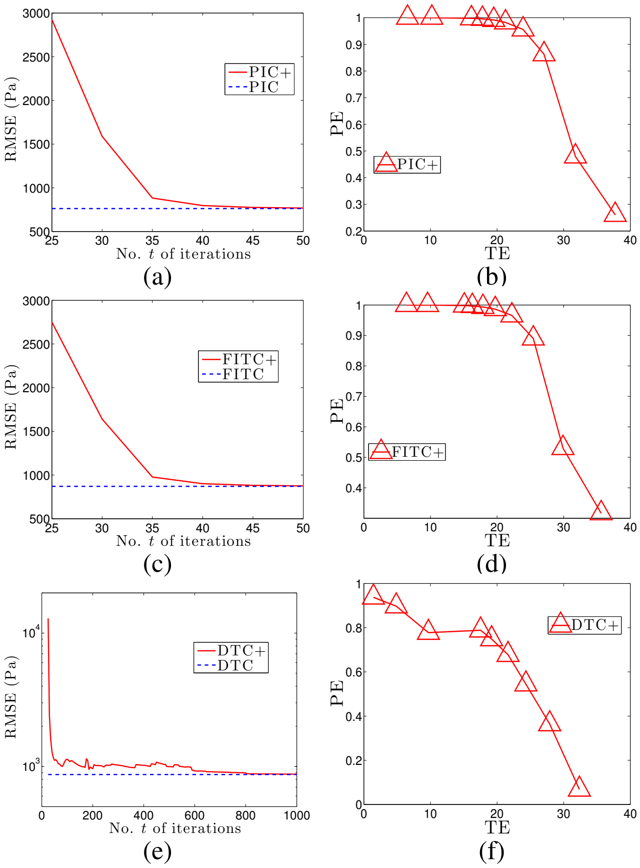

Figure 1. Graphs of RMSEs achieved by (a) PIC+, (c) FITC+, and (e) DTC+ vs. number t of iterations, and graphs of predictive efficiency (PE) vs. time efficiency (TE) showing the anytime efficiencies of (b) PIC+, (d) FITC+, and (f) DTC+ with $|\mathcal{U}| = 512$ inducing outputs and $P = 1000$ blocks for EMSLP dataset.

Four performance metrics are used to evaluate the anytime SGPR models: (a) *Root mean square error* (RMSE): $\sqrt{|\mathcal{S}|^{-1}\sum_{\mathbf{x}\in\mathcal{S}}(y_{\mathbf{x}} - \mu_{\mathbf{x}|\mathcal{D}})^2}$, (b) *mean negative log probability* (MNLP): $0.5|\mathcal{S}|^{-1}\sum_{\mathbf{x}\in\mathcal{S}}((y_{\mathbf{x}} - \mu_{\mathbf{x}|\mathcal{D}})^2/\sigma_{\mathbf{xx}|\mathcal{D}} + \log(2\pi\sigma_{\mathbf{xx}|\mathcal{D}}))$, (c) incurred time, and (d) anytime efficiency demonstrating the trade-off between *time efficiency* (TE) vs. *predictive efficiency* (PE). Formally, TE (PE) is defined as the incurred time (RMSE) of the SGPR model divided by that of its anytime variant. Intuitively, increasing TE (i.e., by decreasing the number of iterations of SNGA update) reduces the incurred time of an anytime variant of the SGPR model at the cost of degrading its PE.

EMSLP Dataset. Figs. 1 and 2 show results of RMSEs, incurred times, and anytime efficiencies of PIC+, FITC+, and DTC+ averaged over 5 random instances with varying number t of iterations. It can be observed from Figs. 1a, 1c, and 1e that the RMSEs of PIC+, FITC+, and DTC+ consistently converge to within 0.75% of that of PIC (RMSE of 762.263 Pa), FITC (RMSE of 870.857 Pa), and DTC (RMSE of 870.878 Pa), respectively. The results of their MNLPs show similar convergence behavior, as detailed in Appendix A. This corroborates our theoretical results in Section 4 that the anytime variants spanned by our unifying framework can achieve asymptotic convergence to the predictive distributions of their corresponding SGPR models. In particular, the RMSEs of PIC+ and FITC+ decrease quickly with an increasing number t of iterations of SNGA update and converge after 50 iterations, which demonstrate their scalability to massive datasets. In fact, during these first 50 iterations, the RMSEs achieved by PIC+ and FITC+ are significantly lower than that achieved by DTC+, as observed in Fig. 2a. On the other hand, the RMSE of DTC+ decreases more gradually and can only converge after 1000 iterations. This inferior predictive performance of DTC+ may be caused by its more restrictive structural assumption of deterministic relation between the training and inducing outputs (Appendix D.1), thus making it perform less robustly among heterogeneous datasets.

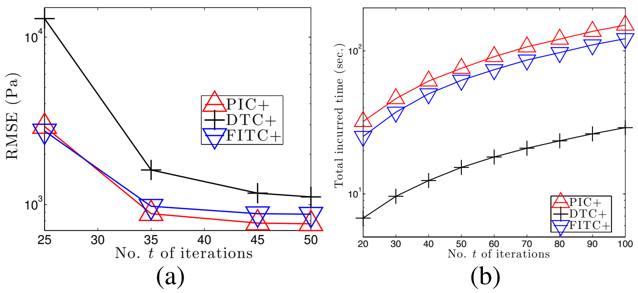

Figure 2. Graphs of (a) RMSEs and (b) total incurred times of PIC+, FITC+, and DTC+ vs. number t of iterations with $|\mathcal{U}| = 512$ inducing outputs and $P = 1000$ blocks for EMSLP dataset.

It can also be observed from Fig. 2a that the superior predictive performance (i.e., lower RMSE) of PIC+ over FITC+ becomes more pronounced with an increasing number t of iterations, which is expected: PIC+ imposes a more relaxed structural assumption of conditional independence than FITC+. For example, unlike FITC+, PIC+ does not assume conditional independence between the test and training outputs given the inducing outputs. Fig. 2b shows linear increases of total incurred time in the number t of iterations for PIC+, FITC+, and DTC+. Our experiments reveal that PIC+, FITC+, and DTC+ incur, respectively, an average of 1.53, 1.15, and 0.32 seconds per update iteration. So, PIC+ and FITC+ take $\sim$76.5 and $\sim$57.5 seconds to converge after 50 iterations while DTC+ takes $\sim$320 seconds to converge after 1000 iterations.

Figs. 1b, 1d, and 1f reveal how the predictive efficiencies of the anytime SGPR models can be traded off to improve their time efficiencies to meet the real-time requirement in time-critical applications. It can be observed that both PIC+ and FITC+ can achieve a speedup of 22-24 (i.e., TE = 25) while preserving 96% of the predictive efficiencies of PIC and FITC (i.e., PE = 0.95); in other words, the RMSEs achieved by PIC+ and FITC+ are only $1/0.95 \approx 1.05$ times larger than that achieved by PIC and FITC. On the

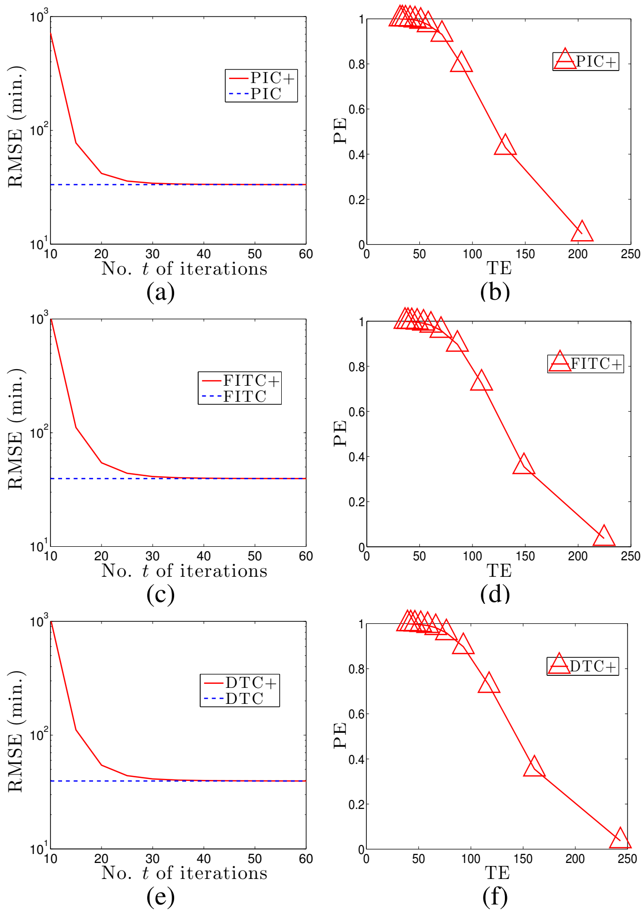

Figure 3. Graphs of RMSEs achieved by (a) PIC+, (c) FITC+, and (e) DTC+ vs. number t of iterations, and graphs of predictive efficiency (PE) vs. time efficiency (TE) showing the anytime efficiencies of (b) PIC+, (d) FITC+, and (f) DTC+ with $|\mathcal{U}| = 100$ inducing outputs and $P = 2000$ blocks for AIRLINE dataset.

other hand, with a speedup of 22, DTC+ can only reach 68% of the predictive efficiency of DTC.

AIRLINE Dataset. Figs. 3 and 4 show results of RMSEs, incurred times, and anytime efficiencies of PIC+, FITC+, and DTC+ averaged over 5 random instances with varying number t of iterations. The observations are mostly similar to that of the EMSLP dataset: From Figs. 3a, 3c, and 3e, the RMSEs of PIC+, FITC+, and DTC+ converge to within 0.04% of that of PIC (RMSE of 33.3515 min.), FITC (RMSE of 39.5302 min.), and DTC (RMSE of 39.5310 min.), respectively. The same observation can be made regarding the results of their MNLPs, as detailed in Appendix A. The RMSEs of PIC+, FITC+, and DTC+ decrease rapidly with an increasing number t of iterations and converge after 60 iterations. During these first 60 iterations, the RMSE achieved by PIC+ is much lower than that achieved by FITC+ and DTC+, as observed in Fig. 4a; this was previously explained in the discussion on the experimental results for EMSLP dataset. Fig. 4b shows linear increases of total incurred time in the number t of iterations for PIC+, FITC+, and DTC+. Our experiments reveal that PIC+, FITC+, and DTC+ incur an average of 0.97, 0.08, and 0.04 seconds per iteration of SNGA update. So, it takes less than 1 minute for PIC+, FITC+, and DTC+ to converge after 60 iterations. Figs. 3b, 3d, and 3f reveal that PIC+, FITC+, and DTC+ can achieve a speedup of 50 (i.e., TE = 50) while preserving almost 100% of the predictive efficiencies of PIC, FITC, and DTC (i.e., PE = 1). But, as observed in Fig. 4a, PIC+ outperforms FITC+ and DTC+ by a huge margin; the same observation can be made for the EMSLP dataset (Fig. 2a).Hence, PIC+ offers the best predictive performance, anytime efficiency, and robustness in both EMSLP and AIRLINE datasets.

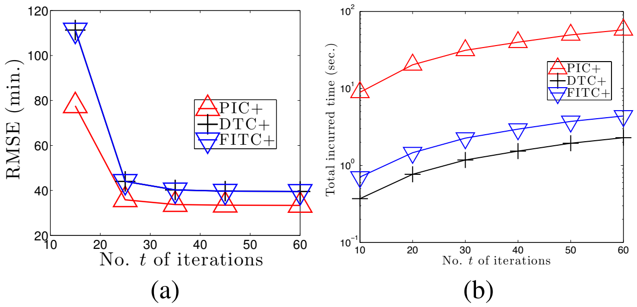

Figure 4. Graphs of (a) RMSEs and (b) total incurred times of PIC+, FITC+, and DTC+ vs. number t of iterations with $|\mathcal{U}| = 100$ and $P = 2000$ for the AIRLINE dataset.

6. Conclusion and Future Work

This paper describes a novel unifying framework of anytime SGPR models (e.g., PIC+, FITC+, DTC+) that can produce good predictive performance fast and trade off between predictive performance vs. time efficiency. After applying our reverse variational inference procedure, a stochastic natural gradient ascent method can be derived that is guaranteed to achieve asymptotic convergence to the predictive distribution of any SGPR model of our choice. We prove that if the predictive distribution of the chosen SGPR model satisfies certain decomposability conditions, then the stochastic natural gradient is an unbiased estimator of the exact natural gradient and can be computed in constant time at each iteration. Empirical evaluation on two real-world million-sized datasets show that PIC+ outperforms FITC+ and state-of-the-art DTC+ (Hensman et al., 2013) in terms of predictive performance and anytime efficiency. A limitation of our unifying framework is that though it can produce the anytime variants of many existing SGPR models (Quiñonero-Candela & Rasmussen, 2005), it does not cover some recent ones like (Lázaro-Gredilla et al., 2010; Low et al., 2015). So, in our future work, we will extend our framework to address this limitation as well as to learn the hyperparameters in an anytime fashion.

Acknowledgments. This research work was carried out at the NUS-ZJU Sensor-Enhanced Social Media (SeSaMe) Centre and supported by Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the Interactive Digital Media Programme Office.

References

Amari, S. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

Ansell et al., T. J. Daily mean sea level pressure reconstructions for the European-North Atlantic region for the period 1850-2003. *J. Climate*, 19(12):2717–2742, 2006.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

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

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

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

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

Chen, J., Low, K. H., Jaillet, P., and Yao, Y. 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*, 2015.

Dolan, J. M., Podnar, G., Stancliff, S., Low, K. H., Elfes, A., Higinbotham, J., Hosler, J. C., Moisan, T. A., and Moisan, J. 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, 2009.

Gal, Y., van der Wilk, M., and Rasmussen, C. E. Distributed variational inference in sparse Gaussian process regression and latent variable models. In *Proc. NIPS*, 2014.

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

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

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

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, pp. 1303–1347, 2013.

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, pp. 1865–1881, 2010.

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

Low, K. H., Dolan, J. M., and Khosla, P. Information-theoretic approach to efficient adaptive path planning for mobile robotic environmental sensing. In *Proc. ICAPS*, pp. 233–240, 2009.

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

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

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

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

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

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

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

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

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Robbins, H. and Monro, S. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 1951.

Schwaighofer, A. and Tresp, V. Transductive and inductive methods for approximate Gaussian process regression. In *Proc. NIPS*, pp. 953–960, 2003.

Seeger, M., Williams, C. K. I., and Lawrence, N. D. Fast forward selection to speed up sparse Gaussian process regression. In *Proc. AISTATS*, 2003.

Smola, A. J. and Bartlett, P. L. Sparse greedy Gaussian process regression. In *Proc. NIPS*, pp. 619–625, 2001.

Snelson, E. and Gharahmani, Z. Sparse Gaussian processes using pseudo-inputs. In *Proc. NIPS*, pp. 1259–1266, 2005.

Snelson, E. L. *Flexible and efficient Gaussian process models for machine learning*. Ph.D. Thesis, University College London, London, UK, 2007.

Snelson, E. L. and Ghahramani, Z. Local and global sparse Gaussian process approximation. In *Proc. AISTATS*, 2007.

Titsias, M. K. Variational learning of inducing variables in sparse Gaussian processes. In *Proc. AISTATS*, pp. 567–574, 2009.

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

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