Incremental Variational Sparse Gaussian Process Regression

Ching-An Cheng
Institute for Robotics & Intelligent Machines
Georgia Institute of Technology
Atlanta, GA 30332
cacheng@gatech.edu

Byron Boots
Institute for Robotics & Intelligent Machines
Georgia Institute of Technology
Atlanta, GA 30332
bboots@cc.gatech.edu

Abstract

Recent work on scaling up Gaussian process regression (GPR) to large datasets has primarily focused on sparse GPR, which leverages a small set of basis functions to approximate the full Gaussian process during inference. However, the majority of these approaches are batch methods that operate on the entire training dataset at once, precluding the use of datasets that are streaming or too large to fit into memory. Although previous work has considered incrementally solving variational sparse GPR, most algorithms fail to update the basis functions and therefore perform suboptimally. We propose a novel incremental learning algorithm for variational sparse GPR based on stochastic mirror ascent of probability densities in reproducing kernel Hilbert space. This new formulation allows our algorithm to update basis functions online in accordance with the manifold structure of probability densities for fast convergence. We conduct several experiments and show that our proposed approach achieves better empirical performance in terms of prediction error than the recent state-of-the-art incremental solutions to variational sparse GPR.

1 Introduction

Gaussian processes (GPs) are nonparametric statistical models widely used for probabilistic reasoning about functions. Gaussian process regression (GPR) can be used to infer the distribution of a latent function f from data. The merit of GPR is that it finds the *maximum a posteriori* estimate of the function while providing the profile of the remaining uncertainty. However, GPR also has drawbacks: like most nonparametric learning techniques the time and space complexity of GPR scale polynomially with the amount of training data. Given N observations, inference of GPR involves inverting an $N \times N$ covariance matrix which requires $O(N^3)$ operations and $O(N^2)$ storage. Therefore, GPR for large N is infeasible in practice.

Sparse Gaussian process regression is a pragmatic solution that trades accuracy against computational complexity. Instead of parameterizing the posterior using *all* N observations, the idea is to approximate the full GP using the statistics of finite $M \ll N$ function values and leverage the induced low-rank structure to reduce the complexity to $O(M^2N + M^3)$ and the memory to $O(M^2)$. Often sparse GPRs are expressed in terms of the distribution of $f(\tilde{x}_i)$, where $\tilde{X} = \{\tilde{x}_i \in \mathcal{X}\}_{i=1}^M$ are called *inducing points* or *pseudo-inputs* [13, 14, 11, 15]. A more general representation leverages the information about the *inducing function* $(L_i f)(\tilde{x}_i)$ defined by indirect measurement of f through a bounded linear operator L_i (e.g. integral) to more compactly capture the full GP [16, 5]. In this work, we embrace the general notion of inducing functions, which trivially includes $f(\tilde{x}_i)$ by choosing L_i to be identity. With abuse of notation, we reuse the term inducing points $\tilde{X}$ to denote the parameters that define the inducing functions.

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Learning a sparse GP representation in regression can be summarized as inference of the hyperparameters, the inducing points, and the statistics of inducing functions. One approach to learning is to treat all of the parameters as hyperparameters and find the solution that maximizes the marginal likelihood [13, 14, 11]. An alternative approach is to view the inducing points and the statistics of inducing functions as variational parameters of a class of full GPs, to approximate the true posterior of f, and solve the problem via variational inference, which has been shown robust to over-fitting [15, 1].

All of the above methods are designed for the batch setting, where all of the data is collected in advance and used at once. However, if the training dataset is extremely large or the data are streaming and encountered in sequence, we may want to *incrementally* update the approximate posterior of the latent function f. Early work by Csató and Opper [4] proposed an online version of GPR, which greedily performs moment matching of the true posterior given *one* sample instead of the posterior of *all* samples. More recently, several attempts have been made to modify variational batch algorithms to incremental algorithms for learning sparse GPs [1, 6, 7]. Most of these methods rely on the fact that variational sparse GPR with fixed inducing points and hyperparameters is equivalent to inference of the conjugate exponential family: Hensman et al. [6] propose a stochastic approximation of the variational sparse GPR problem [15] based on stochastic natural gradient ascent [8]; Hoang et al. [7] generalizes this approach to the case with general Gaussian process priors. Unlike the original variational algorithm for sparse GPR [15], which finds the optimal inducing points and hyperparameters, these algorithms only update the statistics of the inducing functions $f_{\tilde{X}}$.

In this paper, we propose an incremental learning algorithm for variational sparse GPR, which we denote as *iVSGPR*. Leveraging the dual formulation of variational sparse GPR in reproducing kernel Hilbert space (RKHS), *iVSGPR* performs stochastic mirror ascent in the space of probability densities [10] to update the approximate posterior of f, and stochastic gradient ascent to update the hyperparameters. Stochastic mirror ascent, similar to stochastic natural gradient ascent, considers the manifold structure of probability functions and therefore converges faster than the naive gradient approach. In each iteration, *iVSGPR* solves a variational sparse GPR problem of the size of a minibatch. As a result, *iVSGPR* has constant complexity per iteration and can learn all the hyperparameters, the inducing points, and the associated statistics online.

2 Background

Gaussian Processes Regression A GP is a distribution of functions f such that, for any finite index set X, $\{f(x)|x \in X\}$ is Gaussian distributed $\mathcal{N}(f(x)|m(x), k(x,x'))$, where, $m(x)$ and $k(x,x')$ represent the mean of $f(x)$ and the covariance between $f(x)$ and $f(x')$ for $x, x' \in X$. In shorthand, we write $f \sim \mathcal{GP}(m,k)$. The objective of GPR is to infer the posterior probability of the function f given data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$. It treats the function value $f(x_i)$ as a latent variable and assumes that $y_i = f(x_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(\epsilon|0, \sigma^2)$. Let $X = \{x_i\}_{i=1}^N$. We have $p(f|y) = \mathcal{GP}(m_{|\mathcal{D}}, k_{|\mathcal{D}})$:

$$m_{|\mathcal{D}}(x) = k_{x,X}(K_X + \sigma^2 I)^{-1} y \tag{1}$$

$$k_{|\mathcal{D}}(x,x') = k_{x,x'} - k_{x,X}(K_X + \sigma^2 I)^{-1} k_{X,x'} \tag{2}$$

where $y = (y_i)_{i=1}^N \in \mathbb{R}^N$, $k_{x,X} \in \mathbb{R}^{1\times N}$ denotes the cross-covariance, and $K_X \in \mathbb{R}^{N\times N}$ denotes the empirical covariance matrix on X. The hyperparameters θ in the GP are learned by maximizing the log-likelihood of the observation y

$$\max_\theta \log p(y) = \max_\theta \log \mathcal{N}(y|0, K_X + \sigma^2 I). \tag{3}$$

Variational Sparse Gaussian Processes Regression Variational sparse GPR approximates the posterior $p(f|y)$ by a full GP parameterized by inducing points and the statistics of inducing functions [1, 15]. Let f_X and $f_{\tilde{X}}$ denote the function values on X and the inducing points $\tilde{X}$. Specifically, Titsias [15] proposes to use

$$q(f_X, f_{\tilde{X}}) = p(f_X|f_{\tilde{X}}) q(f_{\tilde{X}}) \tag{4}$$

to approximate $p(f_X, f_{\tilde{X}}|y)$, where $q(f_{\tilde{X}}) = \mathcal{N}(f_{\tilde{X}}|\tilde{m}, \tilde{S})$ is the Gaussian approximation of $p(f_{\tilde{X}}|y)$ and $p(f_X|f_{\tilde{X}})$ is a conditional GP. The novelty here is that $q(f_X, f_{\tilde{X}})$, despite parametrization by finite parameters, is still a full GP, which, unlike its predecessor [13], can be infinite-dimensional.

The inference problem of variational sparse GPR is solved by minimizing the KL-divergence $\mathrm{KL}[q(f_X, f_{\tilde{X}})||p(f_X, f_{\tilde{X}}|y)]$. In practice, the minimization problem is transformed into the maximization of the lower bound of the log-likelihood [15]:

$$\begin{aligned}\max_{\theta} \log p(y) &\geq \max_{\theta,\tilde{X},\tilde{m},\tilde{S}} \int q(f_X, f_{\tilde{X}}) \log \frac{p(y|f_X)p(f_X|f_{\tilde{X}})p(f_{\tilde{X}})}{q(f_X, f_{\tilde{X}})} \mathrm{d}f_X \mathrm{d}f_{\tilde{X}} \\ &= \max_{\theta,\tilde{X},\tilde{m},\tilde{S}} \int p(f_X|f_{\tilde{X}})q(f_{\tilde{X}}) \log \frac{p(y|f_X)p(f_{\tilde{X}})}{q(f_{\tilde{X}})} \mathrm{d}f_X \mathrm{d}f_{\tilde{X}} \\ &= \max_{\theta,\tilde{X}} \log \mathcal{N}(y|0, \hat{K}_X + \sigma^2 I) - \frac{1}{2\sigma^2}\mathrm{Tr}(K_X - \hat{K}_X) \end{aligned} \tag{5}$$

Compared with previous literature[11], the variational approach in (5) regularizes the learning with penalty $\mathrm{Tr}(K_X - \hat{K}_X)$ and therefore exhibits better generalization performance.

3 Incremental Variational Sparse Gaussian Process Regression

Despite leveraging sparsity, the batch solution to the variational objective in (5) requires $O(M^2N)$ operations and access to all of the training data during each optimization step [15], which means that learning from large datasets is still infeasible. Recently, several attempts have been made to *incrementally* solve the variational sparse GPR problem in order to learn better models from large datasets [1, 6, 7]. The key idea is to rewrite (5) explicitly into the sum of individual observations:

$$\begin{aligned}&\max_{\theta,\tilde{X},\tilde{m},\tilde{S}} \int p(f_X|f_{\tilde{X}})q(f_{\tilde{X}}) \log \frac{p(y|f_X)p(f_{\tilde{X}})}{q(f_{\tilde{X}})} \mathrm{d}f_X \mathrm{d}f_{\tilde{X}} \\ =& \max_{\theta,\tilde{X},\tilde{m},\tilde{S}} \int q(f_{\tilde{X}}) \left(\sum_{i=1}^{N} \mathbb{E}_{p(f_{x_i}|f_{\tilde{X}})}[\log p(y_i|f_{x_i})] + \log \frac{p(f_{\tilde{X}})}{q(f_{\tilde{X}})} \right) \mathrm{d}f_{\tilde{X}} \end{aligned} \tag{6}$$

The objective function in (6), with fixed $\tilde{X}$, becomes identical to the problem of stochastic variational inference [8] of conjugate exponential families. Hensman et al. [6] exploit this idea to incrementally update the statistics $\tilde{m}$ and $\tilde{S}$ via stochastic *natural* gradient ascent, which can consider the manifold structure of probability distribution derived from KL divergence and is known to be Fisher efficient [2]. Though the optimal inducing points $\tilde{X}$, like the statistics $\tilde{m}$ and $\tilde{S}$, should be updated accordingly as new observations are made, it is hard to design natural gradient ascent for online learning of the inducing points $\tilde{X}$. Because $p(f_X|f_{\tilde{X}})$ in (6) depends on all the observations, evaluating the divergence with respect to $p(f_X|f_{\tilde{X}})q(f_{\tilde{X}})$ over iterations becomes infeasible.

We propose a novel approach, *iVSGPR*, to incremental variational sparse GPR that works by reformulating (5) in its RKHS dual form as

$$\max_{q(f)} \int q(f) \left(\sum_{i=1}^{N} \log p(y_i|f) + \log \frac{p(f)}{q(f)} \right) \mathrm{d}f, \tag{7}$$

where $p(f)$ and $q(f)$ are the Gaussian measures of the prior and the approximate posterior GPs. In particular, $q(f)$ is parametrized by $\tilde{X}$, $\tilde{m}$, and $\tilde{S}$. This avoids the issue of using $p(f_X|f_{\tilde{X}})q(f_{\tilde{X}})$ which refers to all observations. As a result, we can perform stochastic approximation of (5) while monitoring the KL divergence between the posterior approximates due to the change of $\tilde{X}$, $\tilde{m}$, and $\tilde{S}$ across iterations. Specifically, we use stochastic mirror ascent [10] in the space of probability densities in RKHS, which was recently proven as efficient as stochastic natural gradient ascent [12]. In each iteration, *iVSGPR* solves a subproblem of fractional Bayesian inference, which we show can be formulated into a standard variational sparse GPR of the size of a minibatch in $O(M^2N_m + M^3)$ operations, where N_m is the size of a minibatch. See [3] for details.

4 Experiments

We compare our method *iVSGPR* with *VSGPR*$_{\mathrm{svi}}$ the state-of-the-art variational sparse GPR based on stochastic variational inference [6], in which *i.i.d.* data are sampled from the training dataset

	VSGPR$_{\text{svi}}$	*iVSGPR*$_5$	*iVSGPR*$_{10}$	*iVSGPR*$_{\text{ada}}$
kin40k	0.0959	0.0648	0.0608	**0.0607**
SARCOS J$_1$	0.0247	0.0228	0.0214	**0.0210**
SARCOS J$_2$	0.0193	0.0176	0.0159	**0.0156**
SARCOS J$_3$	0.0125	0.0112	0.0104	**0.0103**
SARCOS J$_4$	0.0048	0.0044	0.0040	**0.0038**
SARCOS J$_5$	0.0267	0.0243	0.0229	**0.0226**
SARCOS J$_6$	0.0300	0.0259	0.0235	**0.0229**
SARCOS J$_7$	0.0101	0.0090	0.0082	**0.0081**

(a) *kin40k* and *SARCOS*

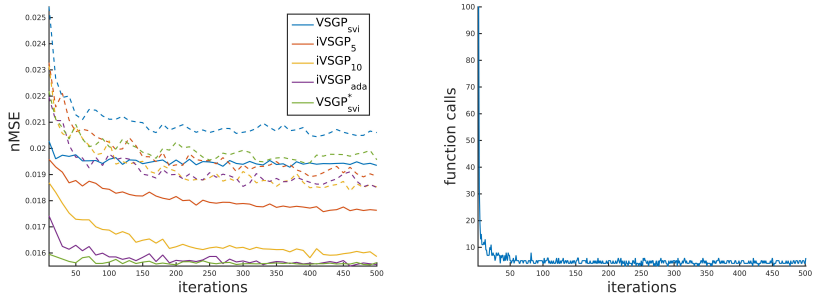
	VSGPR$_{\text{svi}}$	*iVSGPR*$_5$	*iVSGPR*$_{10}$	*iVSGPR*$_{\text{ada}}$
J$_1$	0.1699	0.1455	0.1257	**0.1176**
J$_2$	0.1530	0.1305	0.1221	**0.1138**
J$_3$	0.1873	0.1554	0.1403	**0.1252**
J$_4$	0.1376	0.1216	0.1151	**0.1108**
J$_5$	0.1955	0.1668	0.1487	**0.1398**
J$_6$	0.1766	0.1645	0.1573	**0.1506**
J$_7$	0.1374	0.1357	0.1342	**0.1333**

(b) KUKA1

	VSGPR$_{\text{svi}}$	*iVSGPR*$_5$	*iVSGPR*$_{10}$	*iVSGPR*$_{\text{ada}}$
J$_1$	0.1737	0.1452	0.1284	**0.1214**
J$_2$	0.1517	0.1312	0.1183	**0.1081**
J$_3$	0.2108	0.1818	0.1652	**0.1544**
J$_4$	0.1357	0.1171	0.1104	**0.1046**
J$_5$	0.2082	0.1846	0.1697	**0.1598**
J$_6$	0.1925	0.1890	0.1855	**0.1809**
J$_7$	0.1329	0.1309	0.1287	**0.1275**

(c) KUKA2

Table 1: Testing error (nMSE) after 500 iterations. $N_m = 2048$; J_i denotes the ith joint.



(a) Test error (b) Functions calls of *iVSGPR*$_{\text{ada}}$

Figure 1: Online learning results of *SARCOS* joint 2. (a) nMSE evaluated on the held out test set; the dash lines and the solid lines denote the results with $N_m = 512$ and $N_m = 2048$, respectively. (b) Number of function calls used by *iVSGPR*$_{\text{ada}}$ in solving each subproblem (A maximum of 100 calls is imposed)

to update the models. The experiments of variants of *iVSGPR*[1] are conducted on three real-world robotic datasets datasets, *kin40k* [14] *SARCOS*[2], *KUKA*[9]. For example, Figure 1a shows the change of test error over iterations in learning joint 2 of the *SARCOS* dataset. In general, the adaptive scheme *iVSGPR*$_{\text{ada}}$ performs the best. For all methods, the convergence rate improves with a larger minibatch. In addition, from Figure 1b, we observe that the required number of steps *iVSGPR*$_{\text{ada}}$ needed to solve each subproblem decays with the number of iterations; only a small number of line searches is required after the first few iterations.

5 Conclusion

We propose a stochastic approximation of variational sparse GPR [15], *iVSGPR*. By reformulating the variational inference in RKHS, the update of the statistics of the inducing functions and the inducing points can be unified as stochastic mirror ascent on probability densities to consider the manifold structure. In our experiments, *iVSGPR* shows better performance than the direct adoption of stochastic variational inference to solve variational sparse GPs. As *iVSGPR* executes a fixed number of operations for each minibatch, it is suitable for applications where training data is abundant, e.g. sensory data in robotics. In future work, we are interested in applying *iVSGPR* to extensions of sparse Gaussian processes such as GP-LVMs and dynamical system modeling.

[1]The subscript denotes the number of function calls allowed in each subproblems and ada denotes solving the subproblem until the relative function change is less than a threshold.

[2] http://www.gaussianprocess.org/gpml/data/

References

[1] Ahmed H Abdel-Gawad, Thomas P Minka, et al. Sparse-posterior gaussian processes for general likelihoods. *arXiv preprint arXiv:1203.3507*, 2012.

[2] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[3] Ching-An Cheng and Byron Boots. Incremental variational sparse gaussian process regression. In *Advances in Neural Information Processing Systems*, 2016.

[4] Lehel Csató and Manfred Opper. Sparse on-line gaussian processes. *Neural computation*, 14(3):641–668, 2002.

[5] Anibal Figueiras-vidal and Miguel Lázaro-gredilla. Inter-domain gaussian processes for sparse inference using inducing features. In *Advances in Neural Information Processing Systems*, pages 1087–1095, 2009.

[6] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*, 2013.

[7] Trong Nghia Hoang, Quang Minh Hoang, and Kian Hsiang Low. A unifying framework of anytime sparse gaussian process regression models with stochastic variational inference for big data. In *Proc. ICML*, pages 569–578, 2015.

[8] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[9] Franziska Meier, Philipp Hennig, and Stefan Schaal. Incremental local gaussian regression. In *Advances in Neural Information Processing Systems*, pages 972–980, 2014.

[10] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[11] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[12] Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *Information Theory, IEEE Transactions on*, 61(3):1451–1457, 2015.

[13] Matthias Seeger, Christopher Williams, and Neil Lawrence. Fast forward selection to speed up sparse gaussian process regression. In *Artificial Intelligence and Statistics 9*, number EPFL-CONF-161318, 2003.

[14] Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2005.

[15] Michalis K Titsias. Variational learning of inducing variables in sparse gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.

[16] Christian Walder, Kwang In Kim, and Bernhard Schölkopf. Sparse multiscale gaussian process regression. In *Proceedings of the 25th international conference on Machine learning*, pages 1112–1119. ACM, 2008.