Direct Loss Minimization for Sparse Gaussian Processes

Yadi Wei
Indiana University

Rishit Sheth
Microsoft Research New England

Roni Khardon
Indiana University

Abstract

The paper provides a thorough investigation of Direct Loss Minimization (DLM), which optimizes the posterior to minimize predictive loss, in sparse Gaussian processes. For the conjugate case, we consider DLM for log-loss and DLM for square loss showing a significant performance improvement in both cases. The application of DLM in non-conjugate cases is more complex because the logarithm of expectation in the log-loss DLM objective is often intractable and simple sampling leads to biased estimates of gradients. The paper makes two technical contributions to address this. First, a new method using product sampling is proposed, which gives unbiased estimates of gradients (uPS) for the objective function. Second, a theoretical analysis of biased Monte Carlo estimates (bMC) shows that stochastic gradient descent converges despite the biased gradients. Experiments demonstrate empirical success of DLM. A comparison of the sampling methods shows that, while uPS is potentially more sample-efficient, bMC provides a better tradeoff in terms of convergence time and computational efficiency.

1 Introduction

Bayesian models provide an attractive approach for learning from data. Assuming that model assumptions are correct, given the data and prior one can calculate a posterior distribution that compactly captures all our knowledge about the problem. Then, given a prediction task with an associated loss for wrong predictions, we can pick the best action according to our posterior. This is less clear, however, when exact inference is not possible or when the model is misspecified. Variational inference, which is widely used, chooses the approximate posterior that minimizes the KL-divergence to the true Bayesian posterior, and equivalently maximizes a lower bound on the marginal likelihood. While these properties provide some intuitive justification, they do not immediately guarantee that the resulting approximation has good performance.

To address this, prior work, which is discussed in more details below, showed that (under some technical conditions) variational inference or some variants converge to the "best parameter setting" in the class considered, or has loss comparable to that parameter. These are strong guarantees but they do not show performance competitive with the "prediction resulting from the best posterior over parameters" in the class considered. The latter might use a broad posterior whose predictions are much better in some cases, so there is room for improvement either in better analysis of variational inference or in alternative algorithms.

As argued by several authors (e.g., Lacoste-Julien et al. (2011); Stoyanov et al. (2011)), when exact inference is not possible, it makes sense to optimize the choice of approximate posterior so as to minimize the expected loss of the learner in the future. This requires using the loss function directly during training of the model. We call this approach *direct loss minimization* (DLM). Exploring this idea, Sheth and Khardon (2019) have shown theoretically that (under some technical conditions) DLM does provide performance competitive with the "prediction resulting from the best posterior over parameters" in the class considered.

1.1 DLM and Our Contributions

Motivated by these observations, in this paper we explore the potential of DLM to improve performance in practice, in the context of sparse Gaussian Processes (sGP), and in the process make technical contributions to the problem of gradient estimation for log-expectation terms.

To ground the discussion consider a model with latent variables z and observations y, generating examples via

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

$p(z)\prod p(y_i|z_i)$. When calculating the posterior $p(z|y)$ is hard, variational inference finds an approximation $q(z)$ by maximizing the evidence lower bound (ELBO) or minimizing its negation:

$$\begin{aligned}&-\log p(y)\\&\leq -\int q(z)\log\left(\frac{p(z)}{q(z)}\prod_i p(y_i|z_i)\right)dz\\&=\sum_i E_{q(z_i)}[-\log p(y_i|z_i)]+\beta\ d_{KL}(q(z)\|p(z))\end{aligned}$$

where d_{KL} is the Kullback-Leibler divergence, and $\beta=1$ (but we discuss other values of β below). From this perspective variational inference is seen to perform regularized loss minimization, with d_{KL} as the regularizer. But viewed in this manner the loss on example i is assumed to be $E_{q(z_i)}[-\log p(y_i|z_i)]$ which is not the intended process for a Bayesian predictor. Instead, given a posterior, $q(z)$, the Bayesian algorithm first calculates its predictive distribution $q(y_i)=E_{q(z_i)}[p(y_i|z_i)]$, potentially calculates a prediction $\hat{y}_i$, and then suffers a loss that depends on the context in which the algorithm is used. For the case of log-loss, where $\hat{y}_i$ is not used, the loss term is $-\log q(y_i)$. This suggests a new regularized direct-loss objectve:

$$\begin{aligned}&\text{LogLoss DLM objective}\\&=\sum_i -\log E_{q(z_i)}[p(y_i|z_i)]+\beta\ d_{KL}(q(z),p(z)).\end{aligned}$$

Comparing LogLoss DLM to the ELBO we see that the main difference is the log term which is applied before the expectation. On the other hand, if we care about square loss in the case of regression, the training criterion becomes

$$\begin{aligned}&\text{squareLoss DLM objective}\\&=\sum_i(\hat{y}_i-y_i)^2+\beta\ d_{KL}(q(z),p(z)).\end{aligned}$$

Other losses will similarly lead to different objectives, and hence different posteriors even when trained on the same dataset. This distinction is important. It is not required when performing exact Bayesian inference in correct models but it has practical implications with approximate inference. One of the contributions of this paper is to investigate this issue empirically and our experimental evaluation shows that this distinction is important in practice.

Applying log-loss DLM raises the difficulty of optimizing objectives including $\log E_{q(z_i)}[p(y_i|z_i)]$ terms in cases when the expectation is not analytically tractable. The standard Monte Carlo estimate of the objective, $\log\frac{1}{L}\sum_k p(y_i|z_i^{(k)})$, where $z_i^{(k)}\sim q(z_i)$ (or its reparameterized version) is biased leading to biased gradients — we call this approach bMC. We make two technical contributions in this context. The first is a new method, uPS, for unbiased estimates of gradients for objectives with log-expectation terms through Product Sampling. The method is general and we develop a practical version for the case when $q(z_i)$ is Gaussian. Our second contribution is a theoretical analysis of bMC, showing that (under some technical conditions) stochastic gradient descent using bMC gradients converges despite the bias. bMC has been used in some prior work either explicitly or implicitly and therefore the result may be of independent interest.

An empirical evaluation in sGP for regression, classification and count prediction compares log-loss DLM, to ELBO, as well as β-ELBO (which explicitly optimizes the regularization parameter for ELBO). The evaluation shows that DLM is an effective approach which in some cases matches and in some cases significantly improves over the performance of variational inference and β-ELBO. Results comparing the sampling methods show that uPS is potentially more sample-efficient but bMC provides a better tradeoff in terms of convergence time and computational efficiency.

To summarize, the paper develops new analysis for sampling methods and optimization with log-expectation terms, shows how this can be incorporated in DLM for sGP, and shows empirically that DLM has the potential for significant performance improvements over ELBO.

2 ELBO and DLM for Sparse GP

In this section we review sGP and the development of ELBO and DLM for this model. The GP (Rasmussen and Williams, 2006) is a flexible Bayesian model capturing functions over arbitrary spaces but the complexity of inference in GP is cubic in the number of examples n. Sparse GP solutions reduce this complexity to $O(M^2n)$ where M is the number of pseudo inputs which serve as an approximate sufficient statistic for prediction. The two approaches most widely used are FITC (Snelson and Ghahramani, 2006) and the variational solution of Titsias (2009). The variational solution has been extended for large datasets and general likelihoods and is known as SVGP (Hensman et al., 2013, 2015; Sheth et al., 2015).

In sGP, the GP prior jointly generates the pseudo values u and the latent variables f which we write as $p(u)p(f|u)$ and the observations $y=\{y_i\}$ are generated from the likelihood model $p(y_i|f_i)$. Most previous works use a restricted form for the posterior $q(u,f)=q(u)p(f|u)$ where $q(u)=\mathcal{N}(m,V)$ is Gaussian and where the conditional $p(f|u)$ remains fixed from the prior. Although sGP is slightly more general than the model discussed in the introduction a simi-

lar derivation yields the same forms for ELBO and DLM as above, where the loss term in the ELBO is $E_{q(u)p(f_i|u)}[-\log p(y_i|f_i)] = E_{q(f_i)}[-\log p(y_i|f_i)]$. β-SVGP optimizes the objective with regularizer $\beta d_{KL}(q(u), p(u))$ where sampling through reparameterization is used when exact computation of the objective is not tractable. The collapsed form (Titsias, 2009) for the regression case uses the fact that $E_{q(f_i)}[-\log p(y_i|f_i)]$ has an analytic solution and through it derives an analytic solution for m, V so that only hyperparameters need to be optimized explicitly. FITC (Snelson and Ghahramani, 2006) is not specified using the same family of objective functions but has a related collapsed form which is used in our experiments.

The log-loss term for DLM is $-\log E_{q(u)p(f_i|u)}[p(y_i|f_i)] = -\log E_{q(f_i)}[p(y_i|f_i)] = -\log q(y_i)$. Since both $q(u)$ and $p(f_i|u)$ are Gaussian distributions, the marginal $q(f_i)$ is also Gaussian with mean $\mu_i = K_{iu}K_{uu}^{-1}m$ and variance $v_i = K_{ii} + K_{iu}K_{uu}^{-1}(V - K_{uu})K_{uu}^{-1}K_{ui}$ where $K_{uu} = K(u,u)$, $K_{iu} = K(x_i, u)$ etc.

In the following we consider log loss for regression, binary prediction through Probit regression and count prediction through Poisson regression. For regression we have $p(y_i|f_i) = \mathcal{N}(f_i, \sigma_n^2)$ and the loss term is $-\log q(y_i) = -\log \mathcal{N}(y_i|\mu_i, v_i + \sigma_n^2)$. For probit regression $p(y_i = 1|f_i) = \Phi(f_i)$ where $\Phi(f)$ is the CDF of the standard normal distribution. Here we have for $y_i \in \{0, 1\}$, $-\log q(y_i) = -\log \Phi\left(\frac{(2y_i-1)\mu_i}{\sqrt{v_i+1}}\right)$. In both cases we can calculate derivatives directly through $-\log q(y_i)$. For Poisson regression (with log link function) we have $p(y_i|f_i) = e^{-e^{f_i}}e^{y_i f_i}/y_i!$ and we do not have a closed form for $q(y_i)$. In this case we must resort to sampling when optimizing the DLM objective.

For square loss, $q(y_i)$ is the same as in the regression case, but calculating the loss requires optimal prediction $\hat{y}_i$. In this case, the optimal prediction is the mean of the predictive distribution, that is $\hat{y}_i = K_{iu}K_{uu}^{-1}m$. Therefore the loss term in square loss DLM is $\frac{1}{2}(K_{iu}K_{uu}^{-1}m - y_i)^2$. It is easy to show that the the optimization criterion simplifies into an objective that depends only on m, and the regularized square loss DLM objective for sparse GP is $\frac{1}{2}\sum_i(K_{iu}K_{uu}^{-1}m - y_i)^2 + \frac{\beta}{2}m^T K_{uu}^{-1}m$.

To summarize, both ELBO and DLM include a loss term and KL regularization term. When the loss term is analytically tractable optimization can be performed as usual. When it is not, solutions use sampling where ELBO can use unbiased estimates of derivatives through reparameterization, but log-loss DLM has to compute derivatives for log-expectation terms which are more difficult.

3 Unbiased Gradient Estimates

In this section we develop a new approach for gradients of log-expectation terms. In particular, we describe an extension of a standard technique from the Reinforce algorithm (Williams, 1992) that yields unbiased gradient estimates, by sampling from a product of distributions. The following proposition describes the technique.

Proposition 1. *The estimate*

$$\hat{G}(\theta) = \nabla_\theta \log q(f^{(l)}|\theta), \tag{1}$$

where $f^{(l)} \sim \tilde{q}(f^{(l)}|\theta)$ *and* $\tilde{q}(f|\theta) = \frac{q(f|\theta)p(y|f)}{\mathrm{E}_{q(f|\theta)}\, p(y|f)}$*, is an unbiased estimate of* $\nabla_\theta \log \mathrm{E}_{q(f|\theta)}\, p(y|f)$.

Proof. The true derivative $G(\theta) = \nabla_\theta \log \mathrm{E}_{q(f|\theta)}\, p(y|f)$ is given by

$$\frac{\nabla_\theta \mathrm{E}_{q(f|\theta)}\, p(y|f)}{\mathrm{E}_{q(f|\theta)}\, p(y|f)} = \frac{G_n(\theta)}{\mathrm{E}_{q(f|\theta)}\, p(y|f)}. \tag{2}$$

We next observe using (Williams, 1992) that $G_n(\theta)$ can be written as

$$G_n(\theta) = \mathop{\mathrm{E}}_{q(f|\theta)}\Big[p(y|f)\nabla_\theta \log q(f|\theta)\Big]. \tag{3}$$

The expectation of (1) with respect to the sample $f^{(l)}$ is given by

$$\begin{aligned}
&\mathop{\mathrm{E}}_{\tilde{q}(f^{(l)}|\theta)} \nabla_\theta \log q(f^{(l)}|\theta) \\
&= \int_{f^{(l)}} \Big[\nabla_\theta \log q(f^{(l)}|\theta)\Big] \frac{q(f^{(l)}|\theta)p(y|f^{(l)})}{C} \mathrm{d}f^{(l)} \\
&= \frac{1}{C} \mathop{\mathrm{E}}_{q(f^{(l)}|\theta)} \Big[p(y|f^{(l)})\nabla_\theta \log q(f^{(l)}|\theta)\Big] = \frac{G_n(\theta)}{C} = G(\theta),
\end{aligned}$$

where $C = \mathrm{E}_{q(f|\theta)}\, p(y|f)$, and the second-to-last equality follows from the identity (3). □

The derivation in the lemma is general and does not depend on the form of f. However, the estimate can have high variance and in addition the process of sampling can be expensive. In this paper we develop an effective rejection sampler for the case where f is 1-dimensional and $q(f) = \mathcal{N}(\mu, \sigma^2)$. We provide a sketch here and full details are given in the supplement. Let $\ell(f) = p(y|f)$. To avoid a high rejection rate we sample from $h_2(f) = \mathcal{N}(\mu, n\sigma^2)$ with the same mean as $q()$ but larger variance. We optimize the width multiplier n to balance rejection rate in the region between intersection points of $q()$ and $h_2()$ (where $q()$ is larger) and outside this region ($q()$ is smaller). It is easy to show that this gives a valid rejection sampler with $K = \max_f \ell(f)$, that is, $h_2(f)K \geq q(f)\ell(f)$. This construction requires separate sampling for each example in the dataset (or a minibatch) and significant speedup can be obtained by partly vectorizing the individual samples.

4 Convergence with Biased Gradients

This section shows that biased Monte Carlo estimates can be used to optimize the DLM objective. For presentation clarity, in this section we scale the objective by the number of examples n to get $-\frac{1}{n}\sum_i \log E_{q(f_i)}[p(y_i|f_i)] + \beta\frac{1}{n}\ d_{KL}(q(), p())$.[1] Let $r := (m, V)$ and consider the univariate distribution $q(f_i|r) := \mathcal{N}(f_i|a_{i,1}^\top m + b_{i,1}, a_{i,2}^\top V a_{i,2} + b_{i,2})$ for known vector $a_{i,1}, a_{i,2}$ and scalar constants $b_{i,1}, b_{i,2}$. This form includes many models including sGP. In the following, references to the parameter V and gradients w.r.t. it should be understood as appropriately vectorized. We consider the reparameterized objective $h_i(r) = -\log \mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\, p(y_i|f_i = g_i(r,\epsilon))$ and its gradient

$$\begin{aligned}\nabla_r h_i(r) &= -\frac{\nabla_r \mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\, p(y_i|f_i = g_i(r,\epsilon))}{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\, p(y_i|f_i = g_i(r,\epsilon))} \\ &= -\frac{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\left[\frac{\partial}{\partial f_i}\left[p(y_i|f_i = g_i(r,\epsilon))\right]\nabla_r g_i(r,\epsilon)\right]}{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\, p(y_i|f_i = g_i(r,\epsilon))},\end{aligned} \tag{4}$$

where $g_i(r,\epsilon) = \sqrt{a_{i,2}^\top V a_{i,2} + b_{i,2}}\epsilon + a_{i,1}^\top m + b_{i,1}$. Letting $\phi_i(r,\epsilon) := p(y_i|f_i = g_i(r,\epsilon))$, $\phi_i'(r,\epsilon) := \frac{\partial}{\partial f_i} p(y_i|f_i = g_i(r,\epsilon))$, and $\phi_i''(r,\epsilon) := \frac{\partial^2}{\partial f_i^2} p(y_i|f_i = g_i(r,\epsilon))$, the components of the gradient in (4) are

$$\nabla_m h_i(r) = -\frac{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\left[\phi_i'(r,\epsilon)\right]}{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\left[\phi_i(r,\epsilon)\right]} a_{i,1}, \tag{5}$$

$$\begin{aligned}\nabla_V h_i(r) &= -\frac{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\left[\phi_i'(r,\epsilon)\epsilon\right]}{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\left[\phi_i(r,\epsilon)\right]} \frac{a_{i,2}a_{i,2}^\top}{2\sqrt{a_{i,2}^\top V a_{i,2} + b_{i,2}}} \\ &= -\frac{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\left[\phi_i''(r,\epsilon)\right]}{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\left[\phi_i(r,\epsilon)\right]} \frac{a_{i,2}a_{i,2}^\top}{2},\end{aligned} \tag{6}$$

where the final equality holds under various conditions (Opper and Archambeau, 2009; Rezende et al., 2014; Sheth et al., 2015).

We consider the bMC procedure that replaces the fraction in the true gradients of the loss term with $(\sum_{\ell=1}^L \nabla_r p(y_i|f_i^{(\ell)}))/(\sum_{\ell=1}^L p(y_i|f_i^{(\ell)}))$ where $f_i^{(\ell)} \sim q(f_i|r), 1 \le \ell \le L$. The corresponding bMC estimates of the gradients are

$$d_{i,m}(r) := \frac{\sum_{\ell=1}^L \phi_i'(r,\epsilon^{(\ell)})}{\sum_{\ell=1}^L \phi_i(r,\epsilon^{(\ell)})} a_{i,1} \tag{7}$$

$$d_{i,V}(r) := \frac{\sum_{\ell=1}^L \phi_i''(r,\epsilon^{(\ell)})}{\sum_{\ell=1}^L \phi_i(r,\epsilon^{(\ell)})} \frac{a_{i,2}a_{i,2}^\top}{2} \tag{8}$$

[1] For the sparse GP case, the KL term is over the inducing inputs, whereas for the simpler model in the introduction, the KL term is over f.

where $\{\epsilon^{(\ell)}\}_{\ell=1}^L$ are drawn i.i.d. from $\mathcal{N}(\epsilon|0,1)$.

The main result of this section, given in Corollary 5, shows that it is safe to use (7), (8) instead of (5), (6), with a gradient descent procedure.

Our proof uses the following result from Bertsekas and Tsitsiklis (1996) establishing conditions under which deterministic gradient descent with errors converges:

Proposition 2 (Proposition 3.7 of Bertsekas and Tsitsiklis (1996))**.** *Let r_t be a sequence generated by a gradient method $r_{t+1} = r_t + \gamma_t d_t$, where $d_t = (s_t + w_t)$ and s_t and w_t satisfy (i) $c_1\|\nabla h(r_t)\|^2 \le -\nabla h(r_t)^\top s_t$, (ii) $\|s_t\| \le c_2\|\nabla h(r_t)\|$, and (iii) $\|w_t\| \le \gamma_t(c_3 + c_4\|\nabla h(r_t)\|)$ for some positive constants c_1,c_2, c_3, c_4. If $\nabla h()$ is Lipschitz and $\sum_{t=0}^\infty \gamma_t^2 < \infty$ and $\sum_{t=0}^\infty \gamma_t = \infty$, then either $h(r_t) \to -\infty$ or else $h(r_t)$ converges to a finite value and $\lim_{t\to\infty} \nabla h(r_t) = 0$.*

Intuitively, condition (i) guarantees that the step is in roughly opposite direction of the true gradients and thus the objective is decreasing; condition (ii) bounds the relative magnitude of the step with respect to true gradients; condition (iii) bounds the norm of the error and thus bounds the negative impact of the errors so that the objective can converge to a stationary point.

The next condition is needed for the proof, and as shown by the following proposition it is easy to satisfy.

Assumption 3. *There exist finite constants B, b', B', b'', B'' such that $B \ge \phi_i(r,\epsilon) \ge 0$, $B' \ge \phi_i'(r,\epsilon) \ge b'$ and $B'' \ge \phi_i''(r,\epsilon) \ge b''$. Further, denote $B^* = max\{B, |B'|, |B''|, |b'|, |b''|\}$.*

Proposition 4. *Assumption 3 holds for the likelihood models shown in Table 1.*

The proof of the proposition is given in the supplement. We can now state the main result of this section:

Corollary 5. *Suppose Assumption 3 holds. If for every t and i, $\mathrm{E}_{q(f_i|r)}\, p(y_i|f_i) \ge \zeta > 0$ and*

$$L > \frac{\log(6n/\delta_t)}{2\gamma_t^2} M \quad \textit{where} \tag{9}$$

$$M = \max\left\{\frac{B^2}{|\,\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i(r,\epsilon)|^2}, \frac{(B'-b')^2}{P^2}, \frac{(B''-b'')^2}{Q^2}\right\}, \tag{10}$$

$$P = \begin{cases} |\,\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i'(r,\epsilon)|, & \textit{if } \mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i'(r,\epsilon) \neq 0 \\ 1, & \textit{otherwise} \end{cases},$$

$$Q = \begin{cases} |\,\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i''(r,\epsilon)|, & \textit{if } \mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i''(r,\epsilon) \neq 0 \\ 1, & \textit{otherwise} \end{cases},$$

and $\sum_t \delta_t = \delta$, then with probability at least $1-\delta$, bMC satisfies the conditions of the Proposition 2 and hence converges.

Table 1: Derivative bounds for different models

Likelihood	B	b'	B'	b''	B''
Logistic, $\sigma(yf)$	1	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$
Gaussian, $e^{-(y-f)^2/2\sigma^2}, c=\frac{1}{\sqrt{2\pi}\sigma}$	c	$-\frac{c}{\sqrt{e}\sigma}$	$\frac{c}{\sqrt{e}\sigma}$	$-\frac{c}{\sigma^2}$	$\frac{2c}{\sigma^2 e^{3/2}}$
Probit, $\Phi(yf)$, Φ is cdf of Gaussian	1	$-1/\sqrt{2\pi}$	$1/\sqrt{2\pi}$	$-1/\sqrt{2\pi e}$	$1/\sqrt{2\pi e}$
Poisson, $\frac{g(f)^y e^{g(f)}}{y!}, g(f)=\log(e^f+1)$	1	-1	1	-2.25	2.25
Poisson, $\frac{g(f)^y e^{g(f)}}{y!}, g(f)=e^f$	1	$-y-1$	y	$-y-1/4$	$2y^2+3y+2$
Student's t, $c(1+\frac{(y-f)^2}{\sigma^2\nu})^{-\frac{\nu+1}{2}}, c=\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\sigma}$	c	$-\frac{\frac{c}{\sigma}\frac{\nu+1}{\nu}\sqrt{\frac{\nu}{\nu+2}}}{(\frac{\nu+3}{\nu+2})^{(\nu+3)/2}}$	$\frac{\frac{c}{\sigma}\frac{\nu+1}{\nu}\sqrt{\frac{\nu}{\nu+2}}}{(\frac{\nu+3}{\nu+2})^{(\nu+3)/2}}$	$-\frac{c}{\sigma^2}\frac{\nu+1}{\nu}$	$2\frac{c}{\sigma^2}\frac{\nu+1}{\nu}(\frac{\nu+2}{\nu+5})^{(\nu+5)/2}$

The first condition for convergence requires a uniform lower bound ζ on the overall "agreement" between $q(f_i|r)$ and $p(y_i|f_i)$. Intuitively this is reasonable because we expect the agreement to improve with the training process (albeit not simultaneously for all examples). The second condition requires that the number of samples L is sufficiently large. We first introduce the following lemma.

Lemma 6 (Two-sided relative Hoeffding bound)**.** *Consider i.i.d. draws $\{x^{(\ell)}\}$ from a random variable with mean $\mu \neq 0$ and support $[a,b]$. For $\delta,\alpha \in (0,1)$, if $L > \frac{1}{2}\frac{(b-a)^2}{(\alpha\mu)^2}\log\frac{2}{\delta}$, then, w.p. at least $1-\delta$, $(1/L)\sum_\ell x^{(\ell)}$ and μ have the same sign and $0 < 1-\alpha \leq \frac{(1/L)\sum_\ell x^{(\ell)}}{\mu} \leq 1+\alpha$.*

Proof. First, assume $\mu > 0$. From Hoeffding's inequality, we know that if the condition on L is met, then w.p. $\geq 1-\delta$, we have $\mu-\alpha\mu \leq (1/L)\sum_\ell x^{(\ell)} \leq \mu+\alpha\mu$ from which the result follows. If $\mu < 0$, apply the same argument to the negation of the random variable. □

The main ideas of the proof are as follows. Let s_t be the true gradient, then conditions (i),(ii) hold trivially with $c_1 = c_2 = 1$. Then we aim to bound the error w_t. According to Lemma 6, for sufficiently large sample size L, with high probability, $\frac{1}{L}\sum_\ell \phi_i'(r,\epsilon^{(\ell)})$ and $\frac{1}{L}\sum_\ell \phi_i(r,\epsilon^{(\ell)})$ are close to $\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i'(r,\epsilon)$ and $\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i'(r,\epsilon)$, up to a factor of α. Looking back at (7) and (8) and using Assumption 3, we show that the error term can be bounded linearly in α. Finally we set the value of α at iteration t to be γ_t to guarantee that the error decreases with iterations.

Proof of Corollary 5. We first show that h has Lipshitz gradients. This follows from a generalization of the mean-value theorem applied to continuous and differentiable vector-valued functions (see e.g., Theorem 5.19 of Rudin (1976)). The Lipschitz constant will be equal to the maximum norm of the gradient over the domain and, in our case, will be finite when $\mathrm{E}_{q(f_i|r)}\,p(y_i|f_i) \geq \zeta > 0$. Note that it is always the case that the expectation is > 0 but we must assume a uniform bound for all t, i.

Let $d_t = s_t + w_t$ where $s_t = -\nabla h(r_t)$ so that conditions (i),(ii) hold trivially with $c_1 = c_2 = 1$. We next develop the expression for w_t to show that condition (iii) holds. Now $w_t = \sum_i w_{t,i}$ where m's portion of $w_{t,i}$ is

$$w_{t,i,m} = \frac{a_{i,1}}{n}\left(\frac{(1/L)\sum_\ell \phi_i'(r,\epsilon^{(\ell)})}{(1/L)\sum_\ell \phi_i(r,\epsilon^{(\ell)})} - \frac{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i'(r,\epsilon)}{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i(r,\epsilon)}\right) \tag{11}$$

and a similar expression holds for V's portion.

Our claim follows from three conditions that hold with high probability. When $E\phi' \neq 0$ and $E\phi'' \neq 0$ the conditions require the averages $(1/L)\sum_\ell \phi_i(r,\epsilon^{(\ell)})$, $(1/L)\sum_\ell \phi_i'(r,\epsilon^{(\ell)})$, $(1/L)\sum_\ell \phi_i''(r,\epsilon^{(\ell)})$ to be close to their expectations, i.e. within $1\pm\alpha$ relative error, w.p. $\geq 1-\delta/(3n)$. Using Lemma 6 these are accomplished by assuming that $L > \frac{\log(6n/\delta)}{2\alpha^2}M$.

From (11) we have

$$\|w_{t,i,m}\|^2 = \frac{\|a_{i,1}\|^2}{n^2}\left|\frac{(1/L)\sum_\ell \phi_i'(r,\epsilon^{(\ell)})}{(1/L)\sum_\ell \phi_i(r,\epsilon^{(\ell)})} - \frac{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i'(r,\epsilon)}{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i(r,\epsilon)}\right|^2. \tag{12}$$

Considering the portion with absolute value, if $\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i'(r,\epsilon) \neq 0$ then both fractions have the same sign. Then since $\frac{1+\alpha}{1-\alpha} - 1 > 1 - \frac{1-\alpha}{1+\alpha}$ we have

$$\|w_{t,i,m}\|^2 \leq \frac{\|a_{i,1}\|^2}{n^2}\left((\frac{(1+\alpha)}{(1-\alpha)}-1)\frac{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i'(r,\epsilon)}{\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i(r,\epsilon)}\right)^2 \leq \frac{\|a_{i,1}\|^2}{n^2}\left(\frac{2\alpha}{1-\alpha}\right)^2\left(\frac{B^*}{\zeta}\right)^2, \tag{13}$$

and using $\alpha \leq 0.5$ we get $\|w_{t,i,m}\|^2 \leq \left(\frac{4B^*\alpha}{\zeta}\frac{\|a_{i,1}\|}{n}\right)^2$.

When $\mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i'(r,\epsilon) = 0$, we use the standard Hoeffding bound and $L > \frac{(B'-b')^2\log(6n/\delta)}{2\alpha^2}$ to guarantee that $|(1/L)\sum_\ell \phi_i'(r,\epsilon^{(\ell)})| \leq \alpha$ w.p. $\geq 1-\delta/(3n)$. We also have $(1/L)\sum_\ell \phi_i(r,\epsilon^{(\ell)}) \geq \mathrm{E}_{\mathcal{N}(\epsilon|0,1)}\,\phi_i(r,\epsilon)(1-\alpha) \geq \zeta(1-\alpha)$, and therefore, for m's portion we have

$$\|w_{t,i,m}\| \leq \frac{\|a_{i,1}\|}{n}\frac{\alpha}{\zeta(1-\alpha)} \leq \frac{\|a_{i,1}\|}{n}\frac{2\alpha}{\zeta}. \tag{14}$$

Therefore we can bound m's portion by the sum of bounds from the two cases:

$$\|w_{t,i,m}\| \leq \frac{\|a_{i,1}\|}{n}\frac{2\alpha}{\zeta}(2B^*+1). \tag{15}$$

Similar expressions for both cases hold simultaneously for V, replacing $a_{i,1}$ with $a_{i,2}a_{i,2}^\top$, and, therefore, combining bounds for m, V we have

$$\|w_{t,i}\| \leq \frac{\|a_i\|}{n}\frac{2\alpha}{\zeta}(2B^*+1) \tag{16}$$

where a_i is the concatenation of $a_{i,1}$ and the vectorization of $a_{i,2}a_{i,2}^\top$.

Summing over all examples, we see that

$$\begin{aligned}\|w_t\| = \|\sum_i w_{t,i}\| &\leq \sqrt{\sum_i \sum_j \|w_{t,i}\|\ \|w_{t,j}\|} \\ &\leq A\frac{2\alpha}{\zeta}(2B^*+1)\end{aligned} \tag{17}$$

where $A = \max_i \|a_i\|$. Using the union bound we see that this holds w.p. $\geq 1-\delta$.

To complete the analysis we need to make sure that the above holds for all iterations simultaneously. For this let δ_t be such that $\sum_t \delta_t = \delta$. For example, $\delta_t = \frac{6}{\pi^2}\frac{\delta}{t^2}$. Use δ_t in the definition of L above to obtain the result.

This satisfies condition (iii) if we set α for step t to be $\alpha_t = \gamma_t$ and set $c_3 = A\frac{2\sqrt{2}}{\zeta}(2B^*+1)$ and $c_4 = 0$. □

The implication of the choices of α_t and δ_t is that the number of samples L increases with t. Specifically, for $\gamma_t = 1/t$ this implies $L \propto t^2\log(nt)$. While this is a strong condition, we are not aware of any other analysis for a procedure like bMC. In practice, we use a fixed sample size L in our experiments, and as shown there, the procedure is very effective.

Notice that Corollary 5 only guarantees convergence with high probability. By adding a smoothing factor to the denominator of (7) and (8), we can strengthen the result and prove convergence w.p. 1. However, smoothing did not lead to a significant difference in results of our experiments. Details of the proof and experimental results are provided in the supplement.

5 Related Work

DLM is not a new idea and it can be seen as regularized empirical risk minimization (ERM), a standard approach in the frequentist setting. An intriguing line of work in the frequentist setting follows McAllester et al. (2010) to develop DLM algorithms for non-differentiable losses. Extending the ideas in this paper to develop Bayesian DLM for non-differentiable losses is an important challenge for future work. In the following we disucss related work along several dimensions.

Approximating the Bayesian objective function: In the Bayesian context, DLM can be seen as part of a larger theme which modifies the standard ELBO objective to change the loss term, change the regularization term, and allow for a regularization parameter, as captured by the GVI framework (Knoblauch et al., 2019; Knoblauch, 2019) which is a view strongly connected to regularized loss minimization. For example, the robustness literature, e.g., Knoblauch et al. (2019); Chérief-Abdellatif and Alquier (2019); Bissiri et al. (2016); Futami et al. (2018); Knoblauch (2019), aims to optimize log loss but changes the training loss function in order to be robust to outliers or misspecification and the safe-Bayesian approach of Grünwald (2012); Grünwald and van Ommen (2017) selects β in order to handle misspecification. However, in all these papers the loss term is the *Gibbs loss*, $E_{q()}[\ell()]$, where $\ell()$ is the training loss. In contrast, DLM uses the *loss of the Bayesian predictor* with the motivation that this makes sense as an empirical risk minimization algorithm.

Another interesting connection arises w.r.t. power-EP and α-divergence minimization and their approximations in the BB-α and AEPEP objectives (Hernandez-Lobato et al., 2016; Li and Gal, 2017; Villacampa-Calvo and Hernández-Lobato, 2020) where the latter optimizes $\frac{1}{\alpha}\sum_i -\log E_{q(z_i)}[p(y_i|z_i)^\alpha] + d_{KL}(q(z), p(z))$. The two objectives are identical when $\alpha = \beta = 1$. However, for other values of β, LogLoss DLM cannot be replaced by this objective because AEPEP uses α also as a power of the likelihood. In practice, LogLoss DLM tends to pick small β values but the α-divergence criterion tends to pick α closer to 1 showing the difference is important. On the other hand the DLM perspective can be seen to provide a theoretical motivation for BB-α and AEPEP.

Convergence analysis for Bayesian approximations: A range of approaches have also been used from a theoretical perspective. Some prior analysis of Bayesian algorithms aims to show that the approximations recover exact inference under some conditions. This includes, for example, consistency results for variational inference (Wang and Blei, 2019a,b) and the Laplace approximation (Dehaene, 2017). For sparse GP, Burt et al. (2019) shows that this holds when using the RBF kernel, and when the number and location of pseudo inputs are carefully selected. The work of Alquier et al. (2016) uses PAC Bayes theory and formulates conditions under which the variational approximation is close to the true posterior. In contrast to these, Alquier et al. (2016) and Sheth and Khardon (2017, 2019) analyze variational and DLM al-

gorithms bounding their prediction loss relative to the "best approximate pseudo posterior". Our paper further elaborates algorithmic details of DLM and provides an empirical evaluation.

Sparse GPs: sGP have received significant attention in the last few years. Bauer et al. (2016) investigates the performance of the variational and FITC approximations and provide many insights. Their observations on difficulties in the optimization of hyperparameters in FITC might have parallels in DLM. Our experimental setup explicitly evaluates joint optimization of hyperparameters with DLM as well as a hybrid algorithm to address these difficulties. Reeb et al. (2018) develops a new sGP algorithm by optimizing a PAC-Bayes bound. The output of their algorithm is chosen in a manner that provides better upper bound guarantees on its true error, but the actual test error is not improved over SVGP. The work of Salimbeni et al. (2018) develops a novel variant of SVGP that uses different pseudo locations for m and V. In contrast with these works our paper emphasizes the DLM objective and evaluates its potential to improve performance.

DLM: Several works have explored the idea of DLM for Bayesian algorithms. Sheth and Khardon (2017) demonstrated the success of DLM in topic models. The work of Sheth and Khardon (2016); Jankowiak et al. (2020b,a) applied log loss DLM and variants for regression showing competitive performance with ELBO. Our work significantly improves over this work by exhibiting the differences between square-loss DLM and log-loss DLM for regression, and by developing extensions, sampling methods and analysis for the non-conjugate case of log-loss DLM, which are stated as open questions by Jankowiak et al. (2020a). Finally, Masegosa (2020) motivates DLM as the right procedure, but then identifies a novel alternative objective which is sandwiched between ELBO and DLM. This offers an interesting alternative to DLM with the potential advantage that its loss term is the Gibbs loss (i.e., does not have log-expectation issues), but the disadvantage that it is an approximation to true DLM.

Overall, the space of loss terms, regularizers, and the balance between them offer a range of choices and identifying the best choice in any application is a complex problem. We believe that DLM is an important contribution in this space.

6 Experimental Evaluation

Our experiments[2] have two goals, the first is to evaluate whether DLM provides advantages over variational inference in practice, and the second is to explore the properties of the sampling methods, including efficiency, accuracy and stability. Due to space constraints, we summarize the main results here, and full details are provided in the supplement.

6.1 Details of Algorithms and Experiments

Preliminary experiments with joint optimization of variational parameters and hyperparameters in DLM showed that it is successful in many problems but that in some specific cases the optimization is not stable. We suspect that this is due to interaction between optimization of variational parameters and hyperparameters which complicates an experimental comparison. We therefore run two variants of DLM. The first performs joint optimization of variational parameters and hyperparameters. The second uses fixed hyperparameters, fixing them to the values learned by SVGP. This also allows us to compare the variational posterior of SVGP and DLM on the same hyperparameters.

Prior theoretical results do not have a clear recommendation for setting the regularization parameter β where some analysis uses $\beta = 0$ (no regularization), $\beta = 1$ (the standard setting), and $\beta = \Theta(\sqrt{n})$. Here we use grid search with a validation set on an exponentially-spaced grid, i.e., $\beta = [n, n/2, n/4, n/8, ..., 0.01]$. In some experiments below we diverge from this and present results for specific values of β. To facilitate a fair comparison, we include ELBO with $\beta = 1$ and a variant of ELBO that selects β in exactly the same manner as DLM.

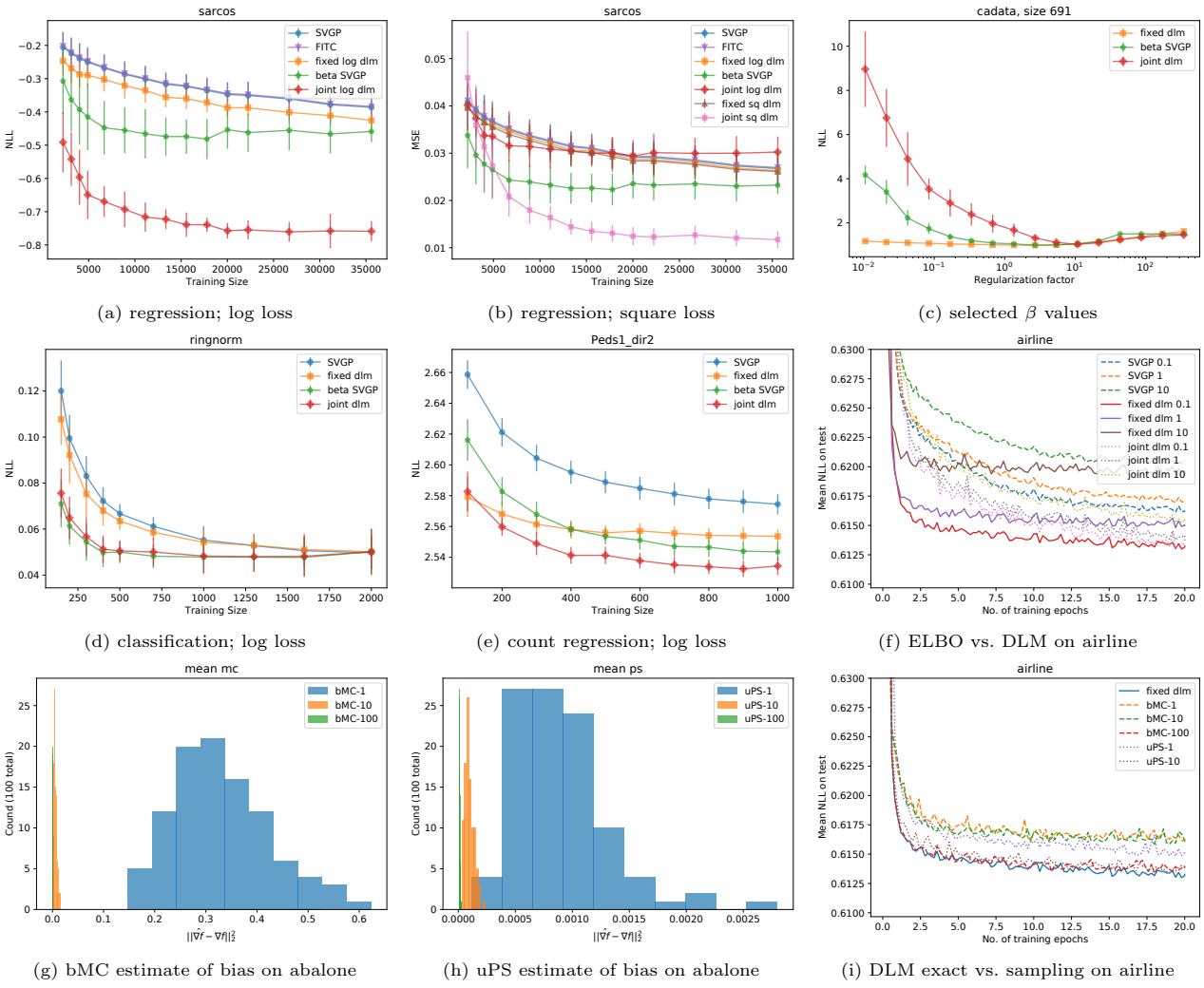
We selected 4 moderate size datasets for each of the likelihoods, giving 16 test cases including regression, square error, classification, and count prediction. In addition, we selected one large classification dataset that has been used before for evaluating sparse GP.

All algorithms are trained with the Adam optimizer. Isotropic RBF kernels are used except for the *airline* dataset where an ARD RBF kernel was used. Evaluations are performed on held-out test data and 5 repetitions are used to generate error bars. Full details of the experiments are given in the supplement.

6.2 Results

Our first set of experiments aims to evaluate the merit of the DLM objective as compared to ELBO. To achieve this, we fix the number of pseudo points and then each point in Figure 1 (a-e), shows the final test set loss score *when the algorithm has converged on the corresponding sample size*. That is, we compare the quality that results from optimizing the objective, and not the optimization algorithm or convergence speed. This allows a cleaner separation of the objectives.

[2]The code used in our experiments is available at `https://github.com/weiyadi/dlm_sgp`.



(a) regression; log loss (b) regression; square loss (c) selected β values

(d) classification; log loss (e) count regression; log loss (f) ELBO vs. DLM on airline

(g) bMC estimate of bias on abalone (h) uPS estimate of bias on abalone (i) DLM exact vs. sampling on airline

Figure 1: Selected results. Description of individual plots is given in the text.

Log-loss and sq-loss in sGP Regression: Figure 1(a) shows the result for log-loss regression on the *sarcos* dataset where log-loss DLM has a significant advantage. Figure 1(b) shows the result for square-loss on the same dataset. Here we see that square-loss DLM has a significant advantage over other algorithms (including log-loss DLM). This illustrates the point made in the introduction, that optimizing DLM for a specific loss can have an advantage over methods that aim for a generic posterior. We can also observe that β-ELBO shows a clear improvement over ELBO, which suggests that selection of β should be adopted more generally in variational inference. The supplement includes results for 3 additional datasets with similar trends.

β-values: It is interesting to consider the β values selected by the algorithms. For most datasets and most training set sizes a small value of $\beta < 1$ is often a good choice. However, this is not always the case. Figure 1(c) shows a plot of log-loss as a function of β for a small (691) training set size on the *cadata* dataset. We observe that the optimal β is larger than 1 for all methods. For larger size data (see supplement) joint DLM selects $\beta < 1$ but other methods do not.

These values can be seen in the context of the safe-Bayes algorithm Grünwald (2012); Grünwald and van Ommen (2017) that selects $\eta = 1/\beta$, but does so using Gibbs loss in a sequential Bayesian prediction. The theoretical analysis leading to safe-Bayes suggests using $\eta = 1/\beta \ll 1$ and similarly PAC-Bayes analysis yields the choice $\beta = \sqrt{n} \gg 1$ both leading to more regularization. However, the best choice might depend on the relation between dataset size and the difficulty of the problem and in practice might mean less regularization.

Log-loss DLM in non-conjugate sGP: Figure 1 (d-e) show log-loss results for classification on the *ringnorm* dataset and for count regression on the *peds1* dataset, where DLM for count regression uses bMC

sampling with 10 samples. We observe that log-loss DLM is comparable to or better than ELBO and β-ELBO. The supplement includes results for 3 additional datasets for each likelihood with similar trends. In some cases hyperparameter optimization in joint-DLM is sensitive, but taken together the two DLM variants are either comparable to or significantly better than ELBO and β-ELBO. In addition, results from the same experiments which are included in the supplement show that DLM achieves better calibration in the non-conjugate cases without sacrificing classification error or count mean relative error.

Non-conjugate DLM on a large dataset: We next consider whether DLM is applicable on large datasets and whether it still shows an advantage over ELBO. For this we use the *airline* dataset (Hensman et al., 2015) which has been used before to evaluate sGP for classification. Due to the size of the dataset we do not perform β selection and instead present results for values 0.1, 1, and 10. In contrast with previous plots, Figure 1(f) is a learning curve, showing log-loss *as a function of training epochs*.[3] We observe that *for all values of β in the experiment* both variants of β-DLM significantly improve over β-ELBO and they significantly improve over ELBO ($\beta = 1$).

Evaluation of the sampling algorithms: We first explore the quality of samples regardless of their effect on learning. Figure 1 portions (g,h) show estimates of bias for bMC and uPS on the *abalone* count prediction dataset (where the true gradient is estimated from 10000 bMC samples). The statistics for the gradients are collected immediately after the initialization of the algorithm. Additional plots in the supplement show estimates for the direction of the update step d_t and its norm relative to the true gradient (similar to conditions (i) and (ii) of Proposition 2 but for d_t and similar to conditions in Proposition 4.1 in Bertsekas and Tsitsiklis (1996)). The plots show that uPS indeed has lower bias as expected (note the scale in x-axis in the plots).

We next compare the quality of predictions when learning using the sampling methods, to each other and to the results of exact computations. Learning curves for *airline* for $\beta = 0.1$ are shown in Figure 1(i) and plots for $\beta = 1, 10$ are given in the supplement. We observe that with enough samples both algorithms can recover the performance of the exact algorithm. We also observe in plot 1(i) that to achieve this uPS can use 10 samples and bMC needs 100 samples. Similarly, uPS with 1 sample is better than bMC with 10 samples. This suggests that uPS makes better use of samples and has a potential advantage. The supplement shows learning curves comparing uPS and bMC for count prediction on two datasets. In this case even one sample of bMC yields good results and there are no significant differences between bMC and uPS in terms of log-loss. Finally, learning curves for log-loss in regression given in the supplement show that bMC can recover the results of exact gradients with ≥ 10 samples. Overall, uPS is unbiased and might make more efficient use of samples. However, despite the speedup developed for uPS, it is significantly slower in practice due to the cost of generating the samples, and bMC provides a better tradeoff in practice.

[3]The plot shows the result of one run but the result is robust. We repeated the experiment 5 times and the learning curves look similar. We chose to show one run to avoid clutter in the plot with error bars.

7 Conclusion

The paper explores the applicability and utility of DLM in sparse GP. We make two technical contributions for sample based estimates of gradients of log-expectation terms: uPS provides unbiased samples and bMC is biased but is proved to lead to convergence nonetheless. An extensive experimental evaluation shows that DLM for sparse GP is competitive and in some cases significantly better than the variational approach and that bMC provides a better time-accuracy tradeoff than uPS in practice. While we have focused on sGP, DLM is at least in principle generally applicable. As mentioned above, this has already been demonstrated for the correlated topic model, where the hidden variable is not 1-dimensional, but where equations simplify and gradients can be efficiently estimated through sampling. We believe that variants of the methods in this paper will enable applicability in probabilistic matrix factorization, GPLVM (through its reparameterized objective), and the variational auto-encoder and we leave these for future work. Extending the analysis of bMC to provide finite time bounds is another important direction for future work.

Acknowledgments

This work was partly supported by NSF under grant IIS-1906694. Some of the experiments in this paper were run on the Big Red 3 computing system at Indiana University, supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

References

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *JMLR*, 17:1–41.

Bauer, M., van der Wilk, M., and Rasmussen, C. E. (2016). Understanding probabilistic sparse Gaussian process approximations. In *Advances in neural information processing systems*, pages 1533–1541.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130.

Burt, D. R., Rasmussen, C. E., and van der Wilk, M. (2019). Rates of convergence for sparse variational Gaussian process regression. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97, pages 862–871.

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. *arXiv 1909.13339*.

Dehaene, G. P. (2017). Computing the quality of the Laplace approximation. *arXiv 1711.08911*.

Futami, F., Sato, I., and Sugiyama, M. (2018). Variational inference based on robust divergences. *arXiv 1710.06595*.

Grünwald, P. (2012). The safe Bayesian - learning the learning rate via the mixability gap. In *Algorithmic Learning Theory*, volume 7568 of *Lecture Notes in Computer Science*, pages 169–183.

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. In *Proceedings of the 29th UAI Conference*, pages 282–290.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable variational Gaussian process classification. *JMLR*.

Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernandez-Lobato, D., and Turner, R. (2016). Black-box alpha divergence minimization. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1511–1520, New York, New York, USA. PMLR.

Jankowiak, M., Pleiss, G., and Gardner, J. R. (2020a). Deep sigma point processes. In *Proceedings of UAI*.

Jankowiak, M., Pleiss, G., and Gardner, J. R. (2020b). Parametric gaussian process regressors. In *ICML*.

Knoblauch, J. (2019). Robust deep Gaussian processes. *arXiv 1904.02303*.

Knoblauch, J., Jewson, J., and Damoulas, T. (2019). Generalized variational inference: Three arguments for deriving new posteriors. *arXiv 1904.02063*.

Lacoste-Julien, S., Huszar, F., and Ghahramani, Z. (2011). Approximate inference for the loss-calibrated bayesian. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 15, pages 416–424.

Li, Y. and Gal, Y. (2017). Dropout inference in Bayesian neural networks with alpha-divergences. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2052–2061, International Convention Centre, Sydney, Australia. PMLR.

Masegosa, A. R. (2020). Learning under model misspecification: Applications to variational and ensemble methods. In *Advances in Neural Information Processing Systems*.

McAllester, D. A., Hazan, T., and Keshet, J. (2010). Direct loss minimization for structured prediction. In *Advances in Neural Information Processing Systems 23*, pages 1594–1602.

Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural Computation*, pages 786–792.

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

Reeb, D., Doerr, A., Gerwinn, S., and Rakitsch, B. (2018). Learning gaussian processes by minimizing pac-bayesian generalization bounds. In *Advances in Neural Information Processing Systems*, pages 3341–3351.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286.

Rudin, W. (1976). *Principles of mathematical analysis (3rd ed.)*. McGraw-hill New York.

Salimbeni, H., Cheng, C.-A., Boots, B., and Deisenroth, M. (2018). Orthogonally decoupled variational gaussian processes. In *Proceedings of Advances in Neural Information Processing Systems 32 (NeurIPS)*.

Sheth, R. and Khardon, R. (2016). Monte carlo structured svi for two-level non-conjugate models. *arXiv 1612.03957*.

Sheth, R. and Khardon, R. (2017). Excess risk bounds for the Bayes risk using variational inference in latent Gaussian models. In *NIPS*, pages 5151–5161.

Sheth, R. and Khardon, R. (2019). Pseudo-Bayesian learning via direct loss minimization with applications to sparse Gaussian process models. In *Symposium on Advances in Approximate Bayesian Inference (AABI)*.

Sheth, R., Wang, Y., and Khardon, R. (2015). Sparse variational inference for generalized Gaussian process models. In *ICML*, pages 1302–1311.

Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264.

Stoyanov, V., Ropson, A., and Eisner, J. (2011). Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 15, pages 725–733.

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *AISTATS*, pages 567–574.

Villacampa-Calvo, C. and Hernández-Lobato, D. (2020). Alpha divergence minimization in multi-class gaussian process classification. *Neurocomputing*, 378:210–227.

Wang, Y. and Blei, D. M. (2019a). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114:1147–1161.

Wang, Y. and Blei, D. M. (2019b). Variational Bayes under model misspecification. In *Advances in Neural Information Processing Systems*, pages 13357–13367.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.