2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–8

Variational Gaussian Process Models without Matrix Inverses

Mark van der Wilk MARK@PROWLER.IO
ST John ST@PROWLER.IO
Artem Artemev ARTEM@PROWLER.IO
James Hensman JAMES@PROWLER.IO
PROWLER.io

1. Introduction

One major obstacle to the wider adoption of Gaussian Process (GP) (Rasmussen and Williams, 2006) based models is their computational cost, which is mainly caused by matrix inverses and determinants. Advances in variational approximate inference methods have reduced the size of the matrices on which expensive operations need to be performed, leading to $O(NM^2)$ time costs instead of $O(N^3)$ (Titsias, 2009), with approximations arbitrarily good with $M \ll N$ (Burt et al., 2019). Minibatches of size $B \ll N$ can be used for training at a cost of $O(BM^2 + M^3)$ per iteration (Hensman et al., 2013).

The usefulness of training with small minibatches is hampered by the iteration cost being dominated by $O(M^3)$, which again comes from an inverse and determinant. The computation is usually done using the Cholesky decomposition, which requires serial operations and high-precision arithmetic. So in addition to being an asymptotically expensive operation, it is also poorly suited to modern deep learning hardware. Removing these per-iteration matrix operations therefore seems necessary to speed up training.

In this work, we provide a variational lower bound that can be computed without expensive matrix operations like inversion. Our bound can be used as a drop-in replacement to the existing variational method of Hensman et al. (2013, 2015), and can therefore directly be applied in a wide variety of models, such as deep GPs (Damianou and Lawrence, 2013). We focus on the theoretical properties of this new bound, and show some initial experimental results for optimising this bound. We hope to realise the full promise in scalability that this new bound has in future work.

2. Variational inference for Gaussian process models

We will consider a straightforward model where we want to learn some relation $f : \mathcal{X} \to \mathbb{R}$ with a GP prior through an arbitrary factorised likelihood. We write the model as

$$p(f(\cdot)) = \mathcal{GP}(\mu(\cdot), k(\cdot, \cdot')), \qquad p(\mathbf{y} \mid f(\cdot)) = \prod_{n=1}^{N} p(y_n \mid f(\cdot)), \tag{1}$$

using some abuse of notation for denoting the GP prior. We need approximate inference to deal with a) the non-conjugate likelihood which prevents a closed-form solution, and b) the $O(N^3)$ matrix operations that come from the Gaussian prior. Our starting point is Hensman et al. (2013, 2015), who propose a solution based on variational inference. An

inducing variable posterior is used, which is constructed by conditioning on M random variables $\mathbf{u} \in \mathbb{R}^M$, and then specifying a free-form Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{S})$ for them:

$$q(f(\cdot)) = \int p(f(\cdot) \,|\, \mathbf{u}) q(\mathbf{u}) \mathrm{d}\mathbf{u} = \mathcal{GP}\left(\mathbf{k}_{\mathbf{u}\cdot}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1} \boldsymbol{\mu}, k_{\cdot\cdot} - \mathbf{k}_{\mathbf{u}\cdot}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{k}_{\mathbf{u}\cdot} + \mathbf{k}_{\mathbf{u}\cdot}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{S} \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{k}_{\mathbf{u}\cdot}\right), \quad (2)$$

where $k_{\cdot\cdot}$, $\mathbf{k}_{\mathbf{u}\cdot}$, and $\mathbf{K}_{\mathbf{uu}}$ are covariances between some the function value at some point $\cdot$ or inducing variables $\mathbf{u}$. The inducing variables are commonly taken to be $\mathbf{u} = \{f(\mathbf{z}_m)\}_{m=1}^M$, making the covariances simple evaluations of the kernel $k(\cdot, \cdot')$. We can minimise the KL divergence between $q(f(\cdot))$ and $p(f(\cdot) \,|\, \mathbf{y})$ (Matthews et al., 2016) by maximising the bound from Hensman et al. (2015) (stochastic variational, "sv"):

$$\mathcal{L}_{\text{sv}} = \sum_{n=1}^N \mathbb{E}_{q(f(\mathbf{x}_n))}[\log p(y_n \,|\, f(\mathbf{x}_n))] - \mathrm{KL}[q(\mathbf{u}) || p(\mathbf{u})], \quad (3)$$

where $q(f(\mathbf{x}_n)) = \mathcal{N}\left(f_n; \mu_n, \sigma_n^2\right)$, with $f_n = f(\mathbf{x}_n)$, $\mu_n = \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1} \boldsymbol{\mu}$, and $\sigma_n^2 = k_{f_n f_n} - \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{k}_{\mathbf{u}f_n} + \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{S} \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{k}_{\mathbf{u}f_n}$ (eq. 2). The expensive $O(M^3)$ operations are the inverse $\mathbf{K}_{\mathbf{uu}}$ in the approximate posterior (eq. 2) and KL term, and the log-determinant in the KL term. In sections 3 and 4 we remove them from the approximate posterior and KL term respectively.

3. Inverse-free variational posteriors

We begin by eliminating inverses from the expected log-likelihood term in eq. 3, which stem from the inverses in the predictive mean μ_n and variance σ_n^2 (eq. 2). By reparameterising $\mathbf{K}_{\mathbf{uu}}^{-1} \boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$ and $\mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{S} \mathbf{K}_{\mathbf{uu}}^{-1} = \hat{\mathbf{S}}$, we can trivially get rid of all inverses, except for the term $k_{f_n f_n} - \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{k}_{\mathbf{u}f_n}$, which we denote as $\sigma_{n|\mathbf{u}}^2$ and call the residual variance.

3.1. Log-concave likelihoods

While we cannot similarly remove the inverse in $\sigma_{n|\mathbf{u}}^2$, we note in lemma 7 (see appendix A for lemmas, proofs and details) that we can lower-bound the Gaussian expectation by using an upper bound for $\sigma_{n|\mathbf{u}}^2$. Upper bounds to $\sigma_{n|\mathbf{u}}^2$ were investigated by Gibbs and MacKay (1997) and in follow-on work by Davies (2015) in the context of conjugate gradient (CG) approximations to matrix inversion. They note that for all values of $\mathbf{a}_n$ we have

$$\sigma_{n|\mathbf{u}}^2 = k_{f_n f_n} - \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{k}_{\mathbf{u}f_n} \leq k_{f_n f_n} + \mathbf{a}_n \mathbf{K}_{\mathbf{uu}} \mathbf{a}_n - 2\mathbf{a}_n^\intercal \mathbf{k}_{\mathbf{u}f_n}, \quad (4)$$

with equality when $\mathbf{a}_n = \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{k}_{\mathbf{u}f_n}$ (lemma 8). We parameterise this upper bound as

$$\bar{\sigma}_{n|\mathbf{u}}^2 = k_{f_n f_n} + \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{T} \mathbf{K}_{\mathbf{uu}} \mathbf{T} \mathbf{k}_{\mathbf{u}f_n} - 2\mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{T} \mathbf{k}_{\mathbf{u}f_n} \quad (5)$$

and use it to construct a lower bound to $\mathcal{L}_{\text{sv}}$ without inverses in the expectation terms:

$$\mathcal{L}_{\text{lc}} = \sum_{n=1}^N \mathbb{E}_{\mathcal{N}(f_n; \mu_n, \bar{\sigma}_n^2)}[\log p(y_n \,|\, f_n)] - \mathrm{KL}[q(\mathbf{u}) || p(\mathbf{u})], \quad (6)$$

$$\mu_n = \mathbf{k}_{\mathbf{u}f_n}^\intercal \hat{\boldsymbol{\mu}}, \qquad \bar{\sigma}_n^2 = \bar{\sigma}_{n|\mathbf{u}}^2 + \mathbf{k}_{\mathbf{u}f_n}^\intercal \hat{\mathbf{S}} \mathbf{k}_{\mathbf{u}f_n}. \quad (7)$$

Proposition 1 *For log-concave ("lc") likelihoods, $\mathcal{L}_{lc}$ is a valid lower bound to the log marginal likelihood, as $\log p(\mathbf{y}) \geq \mathcal{L}_{sv} \geq \mathcal{L}_{lc}$. We have the equality $\mathcal{L}_{lc} = \mathcal{L}_{sv}$ when $\mathbf{T} = \mathbf{K}_{\mathbf{uu}}^{-1}$.*

Remark 2 *Since we are predicting with a different distribution than is in the KL,* $\log p(\mathbf{y}) - \mathcal{L}_{lc}$ *is* not *the KL gap between the approximate and true posteriors.*

With $\mathcal{L}_{\text{lc}}$ we have a bound on the marginal likelihood that we can optimise with respect to the parameters of $\mathcal{L}_{\text{sv}}$, with the addition of $\mathbf{T}$. Section 3.3 discusses more properties.

3.2. General likelihoods

To create a proper variational method which also works with any likelihood, we need to use the same distribution in the predictions as in the KL term. To do this, we find the $q(\mathbf{u})$ in $\mathcal{L}_{\text{sv}}$ that would give the μ_n and $\bar{\sigma}_n^2$ from eq. 7. We solve for $\mathbf{S}$:

$$\mathcal{N}\big(f_n;\ \mathbf{k}_{\mathbf{u}f_n}^\intercal \hat{\boldsymbol{\mu}},\ k_{f_nf_n} + \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{T}\mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{k}_{\mathbf{u}f_n} - 2\mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{T}\mathbf{k}_{\mathbf{u}f_n} + \mathbf{k}_{\mathbf{u}f_n}^\intercal \hat{\mathbf{S}}\mathbf{k}_{\mathbf{u}f_n}\big) = \tag{8}$$

$$\mathcal{N}\big(f_n;\ \mathbf{k}_{\mathbf{u}f_n}^\intercal \hat{\boldsymbol{\mu}},\ k_{f_nf_n} - \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f_n} + \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{S}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f_n}\big) \tag{9}$$

$$\implies \mathbf{S}^* = \mathbf{K}_{\mathbf{uu}} + \mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{K}_{\mathbf{uu}} - 2\mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{K}_{\mathbf{uu}} + \mathbf{K}_{\mathbf{uu}}\hat{\mathbf{S}}\mathbf{K}_{\mathbf{uu}} \tag{10}$$

This shows that we can obtain inverse-free predictions using a simple reparameterisation of $\mathbf{S}$ in $\mathcal{L}_{\text{sv}}$. This gives a new *fully relaxed* ("fr") bound $\mathcal{L}_{\text{fr}}$, as we are relaxing the optimisation by adding $\mathbf{T}$ as an additional variational parameter:

$$\mathcal{L}_{\text{fr}} = \sum_{n=1}^{N} \mathbb{E}_{\mathcal{N}(f(\mathbf{x}_n);\mu_n,\bar{\sigma}_n^2)}[\log p(y_n \mid f(\mathbf{x}_n)] - \mathrm{KL}[\mathcal{N}(\mathbf{K}_{\mathbf{uu}}\hat{\boldsymbol{\mu}}, \mathbf{S}^*)||\mathcal{N}(0, \mathbf{K}_{\mathbf{uu}})]\,. \tag{11}$$

Remark 3 *The fully relaxed bound* $\mathcal{L}_{fr}$ *is simply a reparameterisation of* $\mathcal{L}_{sv}$ *from Hensman et al. (2013, 2015). Each setting of* $\mathbf{T}$ *has a setting for* $\mathcal{L}_{sv}$ *that is exactly equivalent. This means that any model that relies on a variant of* $\mathcal{L}_{sv}$ *for inference can be trivially adapted to use the inverse-free fully relaxed bound, by reparameterising the predictions and KL.*

3.3. Properties

We may worry that additionally optimising over $\mathbf{T}$ prevents us from recovering the same result as from $\mathcal{L}_{\text{sv}}$, due to additional local optima or gradient variance. The following two propositions show that this is not the case.

Proposition 4 *The local optima of* $\mathcal{L}_{lc}$ *and* $\mathcal{L}_{fr}$ *(*$\mathbf{T} = \mathbf{K}_{\mathbf{uu}}^{-1}$*) are identical to those of* $\mathcal{L}_{sv}$*.*

Proposition 5 *The variance of* $\nabla_{\mathbf{T}}\mathcal{L}_{lc}$ *is zero when* $\mathbf{T}$ *is at its optimum* $\mathbf{T} = \mathbf{K}_{\mathbf{uu}}^{-1}$*.*

4. Inverse-free KL estimators

Here, we remove costly matrix operations from the KL term through unbiased estimation. We start by highlighting the $O(M^3)$ terms needed in $\mathcal{L}_{\text{lc}}$ ($\mathcal{L}_{\text{fr}}$ is similar):

$$\mathrm{KL}[q(\mathbf{u})||p(\mathbf{u})] = \frac{1}{2}\mathrm{Tr}\left[\mathbf{K}_{\mathbf{uu}}\hat{\mathbf{S}}\right] + \frac{1}{2}\hat{\boldsymbol{\mu}}^\intercal\mathbf{K}_{\mathbf{uu}}\hat{\boldsymbol{\mu}} - \frac{1}{2}M - \frac{1}{2}\log|\mathbf{K}_{\mathbf{uu}}| - \frac{1}{2}\log\left|\hat{\mathbf{S}}\right|. \tag{12}$$

The trace term requires full matrix multiplications as we parameterise $\hat{\mathbf{S}} = \mathbf{L}\mathbf{L}^\intercal$, and computing the determinant typically requires a costly decomposition.

The trace term is dealt with using the Hutchinson trace estimator, which uses random vectors with $\mathbb{E}_{\mathbf{r}}[\mathbf{r}\mathbf{r}^\intercal] = \mathbf{I}$ to estimate $\widehat{\mathrm{Tr}[\mathbf{K}_{\mathbf{uu}}\hat{\mathbf{S}}]} = \sum_{h=1}^{H} \mathbf{r}_h^\intercal\mathbf{K}_{\mathbf{uu}}\hat{\mathbf{S}}\mathbf{r}_h$, which can be evaluated

with matrix-vector multiplications only. With $H \ll M$ probe vectors this has a cost of $O(HM^2)$.

The $\log|\mathbf{K_{uu}}|$ term is more challenging. Since we use gradient-based optimisation, we will focus on obtaining an unbiased estimate of its gradient. We follow Filippone and Engler (2015) by starting with the unbiased estimator $\widehat{\mathbf{K}^{-1}} = \mathbf{s}\mathbf{r}^\intercal$, where $\mathbf{s} = \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{r}$ and $\mathbb{E}_{\mathbf{r}}[\mathbf{r}\mathbf{r}^\intercal] = \mathbf{I}$. We then use the Unbiased LInear Systems SolvEr (ULISSE), a randomly truncated CG run, to compute an unbiased estimate of $\mathbf{s}$. The key insight of ULISSE is that the conjugate gradient method expresses the solution $\mathbf{s} = \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{r}$ as a sum of separate terms. The sum is randomly truncated, and each term is re-weighted by the probability of its inclusion to keep an unbiased estimate:

$$\mathbf{s} = \mathbf{s}_1 + \mathbf{s}_2 + \ldots\,, \quad \hat{\mathbf{s}} = \sum_{i=1}^{I} w_i \mathbf{s}_i\,, \quad \text{with } I \sim p(i)\,, \quad \text{and } w_i \text{ s.t. } \mathbb{E}_I[\hat{\mathbf{s}}] = \mathbf{s} = \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{r}\,. \tag{13}$$

The cost of ULISSE is $O(H\bar{I}M^2)$, where $\bar{I} = \mathbb{E}[I]$, so we need $\bar{I}$ to be small for the method to be practical, while keeping small gradient variance. If we parameterise $\mathbf{T} = \mathbf{L_T}\mathbf{L_T}^\intercal$ (with $\mathbf{L_T}$ lower triangular), we can use it as a preconditioner with the following property:

Proposition 6 *When using $\mathbf{L_T}$ as a preconditioner, ULISSE will converge in a single step when $\mathbf{T} = \mathbf{K}_{\mathbf{uu}}^{-1}$. This allows $\bar{I} = 1$ without adding additional variance.*

The hope is that during optimisation $\mathbf{T}$ is updated quickly enough to remain close to the current $\mathbf{K}_{\mathbf{uu}}^{-1}$, which would allow us to choose a small $\bar{I}$ without adding significant variance.

5. Initial results

We show the inverse free GP using the log-concave bound in fig. 1, and a deep GP based on Salimbeni and Deisenroth (2017) and the fully relaxed bound in fig. 2, both optimising $\mathbf{T}$. We see that in fig. 1 the correct solution is completely recovered, while in fig. 2 a similar fit is achieved with a somewhat lower ELBO. See appendix B for additional details.

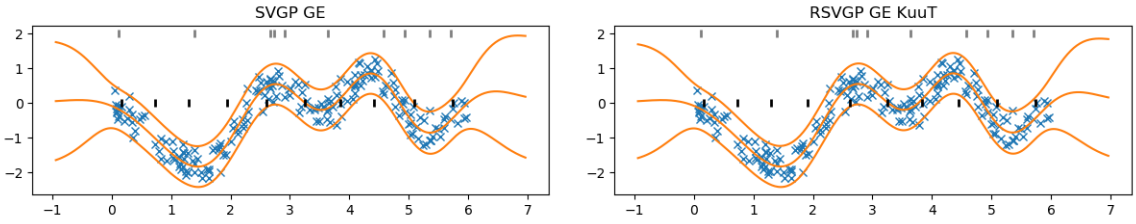

Figure 1: Solutions of $\mathcal{L}_{\text{sv}}$ (left) and $\mathcal{L}_{\text{lc}}$ (right) to a toy dataset. The initialisation and final inducing inputs are shown at the top and middle of the image respectively.

6. Discussion

We presented new variational bounds for GP models that function as drop-in replacements to those developed by Hensman et al. (2013, 2015), but without needing to compute expensive matrix operations each iteration. We prove their properties and show that they behave

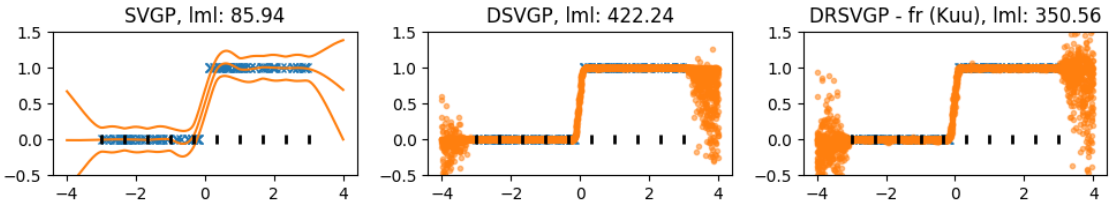

Figure 2: Deep GP solutions to a step function, based on Salimbeni and Deisenroth (2017). Left: shallow GP, middle: deep GP based on $\mathcal{L}_{\text{sv}}$, right: deep GP based on $\mathcal{L}_{\text{fr}}$.

as expected in simple experiments using a single layer and deep GP. We believe this method to be promising, as it removes the most frequently cited impediment against the scaling of GP models. However, more improvements are needed to obtain the full practical benefits.

References

David R. Burt, Carl E. Rasmussen, and Mark van der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *Proceedings of the 36st International Conference on Machine Learning (ICML 2019)*, 2019.

Andreas C. Damianou and Neil D. Lawrence. Deep Gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*, 2013.

Alexander Davies. *Effective implementation of Gaussian process regression for machine learning*. PhD thesis, University of Cambridge, 2015.

Maurizio Filippone and Raphael Engler. Enabling scalable stochastic gradient-based inference for Gaussian processes by employing the Unbiased LInear System SolvEr (ULISSE). In *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, 2015.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, pages 7576–7586, 2018.

Mark N. Gibbs and David J.C. MacKay. Efficient implementation of Gaussian processes. Technical report, University of Cambridge, 1997.

James Hensman, Nicoló Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, 2013.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2015)*, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander G. de G. Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016)*, 2016.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning. 2006.* The MIT Press, Cambridge, MA, USA, 2006.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems 30 (NIPS 2017).* 2017.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*, 2009.

Appendix A. Proofs

Lemma 7 *The Gaussian expectation of a concave function $\phi(\cdot)$ is lower-bounded by a Gaussian expectation with the same mean and a larger variance, $\mathbb{E}_{\mathcal{N}(x;\mu,\sigma^2)}[\phi(x)] \geq \mathbb{E}_{\mathcal{N}(\tilde{x};\mu,\sigma^2+\tilde{\sigma}^2)}[\phi(\tilde{x})]$.*

Proof We write $\tilde{x} = x + \epsilon$, where $\epsilon \sim \mathcal{N}\big(0, \tilde{\sigma}^2\big)$ and $x \sim \mathcal{N}\big(\mu, \sigma^2\big)$. We can then write the right-hand side of the inequality as a nested expectation:

$$\mathbb{E}_{\mathcal{N}(\tilde{x};\mu,\sigma^2+\tilde{\sigma}^2)}[\phi(\tilde{x})] = \mathbb{E}_{\mathcal{N}(x;\mu,\sigma^2)}\big[\mathbb{E}_{\mathcal{N}(\epsilon;0,\tilde{\sigma}^2)}[\phi(x+\epsilon)]\big].$$

We can move the inner expectation over ϵ into the argument of $\phi(\cdot)$ by applying Jensen's inequality, $\mathbb{E}_\epsilon[\phi(\epsilon)] \leq \phi(\mathbb{E}_\epsilon[\epsilon])$:

$$\mathbb{E}_{\mathcal{N}(x;\mu,\sigma^2)}\big[\mathbb{E}_{\mathcal{N}(\epsilon;0,\tilde{\sigma}^2)}[\phi(x+\epsilon)]\big] \leq \mathbb{E}_{\mathcal{N}(x;\mu,\sigma^2)}\big[\phi(\mathbb{E}_{\mathcal{N}(\epsilon;0,\tilde{\sigma}^2)}[x+\epsilon])\big] = \mathbb{E}_{\mathcal{N}(x;\mu,\sigma^2)}[\phi(x)],$$

and so

$$\mathbb{E}_{\mathcal{N}(\tilde{x};\mu,\sigma^2+\tilde{\sigma}^2)}[\phi(\tilde{x})] \leq \mathbb{E}_{\mathcal{N}(x;\mu,\sigma^2)}[\phi(x)].$$

■

Lemma 8 *For any $\mathbf{a}_n$, we have*

$$k_{f_n f_n} - \mathbf{k}_{\mathbf{u}f_n}^T \mathbf{K}_{\mathbf{uu}}^{-1} \mathbf{k}_{\mathbf{u}f_n} \leq k_{f_n f_n} + \mathbf{a}_n^T \mathbf{K}_{\mathbf{uu}} \mathbf{a}_n - 2\mathbf{a}_n^T \mathbf{k}_{\mathbf{u}f_n}\,, \tag{14}$$

with equality when $\mathbf{a}_n = \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f_n}$.

Proof This follows directly from

$$\mathbf{a}_n^\mathsf{T}\mathbf{K}_{\mathbf{uu}}\mathbf{a}_n - 2\mathbf{k}_{\mathbf{u}f_n}^\mathsf{T}\mathbf{a}_n + \mathbf{k}_{\mathbf{u}f_n}^\mathsf{T}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f_n} = \big(\mathbf{a}_n - \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f_n}\big)^\mathsf{T}\mathbf{K}_{\mathbf{uu}}\big(\mathbf{a}_n - \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f_n}\big) \geq 0. \tag{15}$$

■

Proposition 1

Proof We consider a single term in the sum in eq. 6, $\mathcal{L}_{\text{lc}}^{(n)} = \mathbb{E}_{\mathcal{N}(f_n;\mu_n,\bar{\sigma}_n^2)}[\log p(y_n \mid f_n)]$. By lemma 7, for a concave $\log p(y_n \mid f_n)$, we have

$$\mathcal{L}_{\text{lc}}^{(n)} = \mathbb{E}_{\mathcal{N}(f_n;\mu_n,\bar{\sigma}_n^2)}[\log p(y_n \mid f_n)] \leq \mathbb{E}_{\mathcal{N}(f_n;\mu_n,\sigma_n^2)}[\log p(y_n \mid f_n)]$$

if $\bar{\sigma}_n^2 \geq \sigma_n^2$, which is ensured by $\bar{\sigma}_n^2 - \sigma_n^2 = \bar{\sigma}_{n|\mathbf{u}}^2 - \sigma_{n|\mathbf{u}}^2$ and eq. 4.

At the optimum $\mathbf{T} = \mathbf{K}_{\mathbf{uu}}^{-1}$, the variance upper bound (5) is tight: $\bar{\sigma}_{n|\mathbf{u}}^2 = k_{f_n f_n} + \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uu}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f_n} - 2\mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f_n} = k_{f_n f_n} - \mathbf{k}_{\mathbf{u}f_n}^\intercal \mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{k}_{\mathbf{u}f_n} = \sigma_{n|\mathbf{u}}^2$, and we have equality $\mathcal{L}_{\text{lc}} = \mathcal{L}_{\text{sv}}$. ■

Proposition 4

Proof We consider the interesting case where $N > M$ and where our kernels are non-degenerate to avoid underdetermined linear systems.

We first consider the case for $\mathcal{L}_{\text{lc}}$.

From proposition 1 we know that $\mathcal{L}_{\text{lc}} = \mathcal{L}_{\text{sv}}$ when $\mathbf{T} = \mathbf{K}_{\mathbf{uu}}^{-1}$. Here we additionally show that whenever $\mathbf{T} \neq \mathbf{K}_{\mathbf{uu}}^{-1}$, there is a non-zero gradient. We begin by showing that the $\nabla_{\mathbf{T}}\mathcal{L}_{\text{lc}} \neq 0$ unless $\mathbf{T} = \mathbf{K}_{\mathbf{uu}}^{-1}$.

Here, we denote $\epsilon_n \sim \mathcal{N}(0,1)$, and write the expected likelihood term of $\mathcal{L}_{\text{lc}}$ as

$$\mathbb{E}_{f(\mathbf{x}_n)}[\mathbb{E}_{\epsilon_n}[\log p(y_n \mid f(\mathbf{x}_n) + \alpha_n\epsilon_n)]] = \mathbb{E}_{f(\mathbf{x}_n)}[\mathbb{E}_{\epsilon_n}[\phi_n((\mathbf{x}_n) + \alpha_n\epsilon_n)]]\,, \tag{16}$$

where $\alpha_n = \sqrt{\bar{\sigma}_n^2 - \sigma_n^2} = (\mathbf{k}_{\mathbf{u}f_n}^\intercal(\mathbf{T} - \mathbf{K}_{\mathbf{uu}}^{-1})\mathbf{K}_{\mathbf{uu}}(\mathbf{T} - \mathbf{K}_{\mathbf{uu}}^{-1})\mathbf{k}_{\mathbf{u}f_n})^{\frac{1}{2}}$. We set the gradient w.r.t. $\mathbf{T}$ to zero:

$$\nabla_{\mathbf{T}}\mathbb{E}_{f(\mathbf{x}_n)}[\mathbb{E}_{\epsilon_n}[\phi(f(\mathbf{x}_n) + \alpha_n\epsilon_n)]] = \mathbb{E}_{f(\mathbf{x}_n)}\left[\mathbb{E}_{\epsilon_n}\left[\phi'(f(\mathbf{x}_n) + \alpha_n\epsilon_n)\epsilon_n\right]\right]\nabla_{\mathbf{T}}\alpha_n = 0\,. \tag{17}$$

We write $\nabla_{\mathbf{T}}\alpha_n = \frac{1}{2}\alpha_n^{-1}\nabla_{\mathbf{T}}(\alpha_n^2)$, as α_n^2 is a quadratic function of $\mathbf{T}$. Taking the sum over all N data points, and given that $N > M$, this quadratic has a unique optimum at $\mathbf{T} = \mathbf{K}_{\mathbf{uu}}^{-1}$, and so a unique point at which the gradients will be zero, at which $\mathcal{L}_{\text{lc}} = \mathcal{L}_{\text{sv}}$.

For $\mathcal{L}_{\text{fr}}$, we additionally need to consider the KL term, which is a convex function of $\mathbf{S}$. So if $\nabla_{\mathbf{T}}\mathbf{S} = 0$, then the KL is at a stationary point as well. In the $\mathcal{L}_{\text{fr}}$ bound, we have

$$\mathbf{S} = \mathbf{K}_{\mathbf{uu}} + \mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{K}_{\mathbf{uu}} - 2\mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{K}_{\mathbf{uu}} + \mathbf{K}_{\mathbf{uu}}\hat{\mathbf{S}}\mathbf{K}_{\mathbf{uu}}\,, \tag{18}$$

$$\nabla_{\mathbf{T}}\mathbf{S} = \mathbf{K}_{\mathbf{uu}} \otimes (\mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{K}_{\mathbf{uu}}) + (\mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{K}_{\mathbf{uu}}) \otimes \mathbf{K}_{\mathbf{uu}} - 2\mathbf{K}_{\mathbf{uu}} \otimes \mathbf{K}_{\mathbf{uu}} = 0 \iff \mathbf{T} = \mathbf{K}_{\mathbf{uu}}^{-1}\,.$$

■

Proposition 5

Proof We evaluate unbiased estimates of the gradients using the reparameterisation trick, with samples $\epsilon_n \sim \mathcal{N}(0,1)$, and $f_n = \mu_n + (\bar{\sigma}_n^2)^{\frac{1}{2}}\epsilon_n$ (where $\mu_n = \mathbf{k}_{\mathbf{u}f_n}^\intercal\hat{\boldsymbol{\mu}}$ does not depend on $\mathbf{T}$). For the gradient of a single term $\mathcal{L}_{\text{lc}}^{(n)} = \mathbb{E}_{f_n}[\phi(f_n)]$ this gives the estimator:

$$\hat{g}_n = \nabla_{\mathbf{T}}\phi(f_n) = \phi'(f_n)\frac{1}{2}(\bar{\sigma}_n^2)^{-\frac{1}{2}}\epsilon_n\nabla_{\mathbf{T}}\left(\mathbf{k}_{\mathbf{u}f_n}^\intercal\mathbf{T}\mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{k}_{\mathbf{u}f_n} - 2\mathbf{k}_{\mathbf{u}f_n}^\intercal\mathbf{T}\mathbf{k}_{\mathbf{u}f_n}\right). \tag{19}$$

The term $\left(\mathbf{k}_{\mathbf{u}f_n}^\intercal\mathbf{T}\mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{k}_{\mathbf{u}f_n} - 2\mathbf{k}_{\mathbf{u}f_n}^\intercal\mathbf{T}\mathbf{k}_{\mathbf{u}f_n}\right)$ is a quadratic in $\mathbf{T}$, with an optimum at $\mathbf{T} = \mathbf{K}_{\mathbf{uu}}^{-1}$, at which point $\nabla_{\mathbf{T}}\left(\mathbf{k}_{\mathbf{u}f_n}^\intercal\mathbf{T}\mathbf{K}_{\mathbf{uu}}\mathbf{T}\mathbf{k}_{\mathbf{u}f_n} - 2\mathbf{k}_{\mathbf{u}f_n}^\intercal\mathbf{T}\mathbf{k}_{\mathbf{u}f_n}\right) = 0$, which makes $\hat{g}_n = 0$, irrespective of n. This implies that the reparameterisation gradient w.r.t. $\mathbf{T}$ is zero whenever $\mathbf{T}$ is at its optimum, regardless of which minibatch or ϵ_n is sampled. ■

Proposition 6

Proof We have $\mathbf{K_{uu}^{-1}} = \mathbf{T} = \mathbf{L_T L_T^\intercal}$, so $\mathbf{I} = \mathbf{K_{uu} K_{uu}^{-1}} = \mathbf{K_{uu} L_T L_T^\intercal}$. Left-multiplying with $\mathbf{L_T^\intercal}$ and right-multiplying with $\mathbf{L_T^{-\intercal}}$ we obtain $\mathbf{I} = \mathbf{L_T^\intercal K_{uu} L_T}$. Conjugate gradients solves against the identity matrix in one step, and ULISSE$(\mathbf{I}, \mathbf{L_T^\intercal r}) = \mathbf{L_T^\intercal r}$ exactly, and $\hat{\mathbf{s}}' = \mathbf{L_T L_T} = \mathbf{K_{uu}^{-1} r}$. ■

Appendix B. Additional empirical results

For the examples in section 5 our methods can recover solutions similar to $\mathcal{L}_{\text{sv}}$, which computes inverses exactly. The current set-up optimises all hyperparameters and variational parameters (including $\mathbf{T}$) jointly using Adam (Kingma and Ba, 2014). We see in figs. 3 and 5 that the inverse-free methods do require more iterations to achieve a similar ELBO, although the difference in ELBO tends to exaggerate the visual difference in fit quality.

In fig. 4 we visualise the quality of the inverse that is obtained from optimising $\mathcal{L}_{\text{lc}}$ for the single layer experiment. We see that all initialisations recover $\mathbf{T} = \mathbf{K_{uu}^{-1}}$ almost perfectly.

Future improvements will focus on improved optimisation behaviour, and software improvements. The optimisation surface becomes more challenging for larger datasets causing Adam to become unstable, so perhaps a different optimisation routine can be used to take advantage of structure in the objective functions. Software improvements to allow taking advantage of using only matrix-vector products (Gardner et al., 2018) will also be needed to gain the full per-iteration speed-up that this method allows.

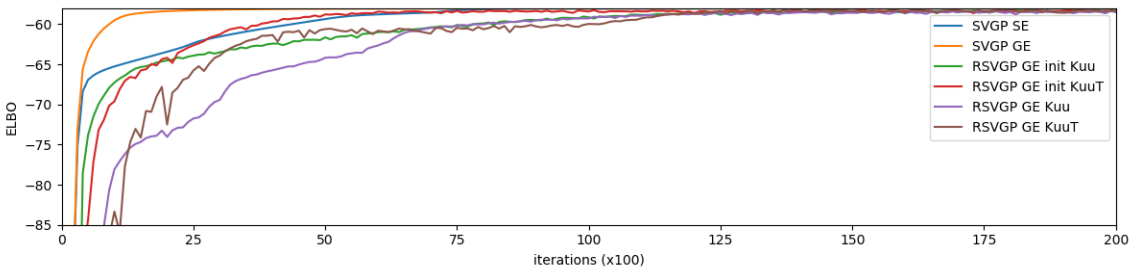

Figure 3: Optimisation of the single-layer ELBOs against number of iterations, for various initialisations of the log-concave bound (RSVGP), with a run of $\mathcal{L}_{\text{sv}}$ for comparison (SVGP GE).

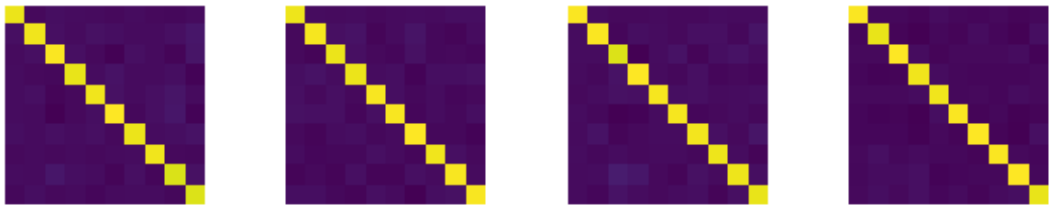

Figure 4: Visualisation of $\mathbf{K_{uu}T}$ at end of the single layer experiments for different initialisations. If the recovery of the inverse through optimisation works, the image should show an exact identity matrix.

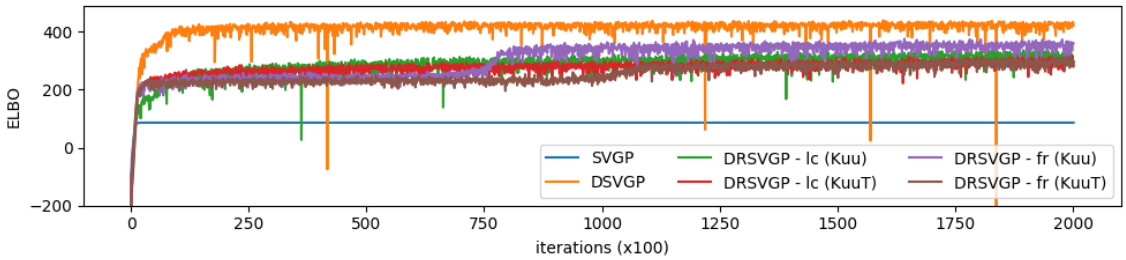

Figure 5: Optimisation of the deep GP ELBOs against number of iterations, for various initialisations of the deep GP variant of $\mathcal{L}_{\text{fr}}$, with a single layer model (SVGP) and a deep GP based on $\mathcal{L}_{\text{sv}}$ (DSVGP) for comparison.