Sparse Gaussian Processes: Structured Approximations and Power-EP Revisited

Thang D. Bui
Australian National University
thang.bui@anu.edu.au

Michalis K. Titsias
Google DeepMind
mtitsias@google.com

Abstract

Inducing-point-based sparse variational Gaussian processes have become the standard workhorse for scaling up GP models. Recent advances show that these methods can be improved by introducing a diagonal scaling matrix to the conditional posterior density given the inducing points. This paper first considers an extension that employs a block-diagonal structure for the scaling matrix, provably tightening the variational lower bound. We then revisit the unifying framework of sparse GPs based on Power Expectation Propagation (PEP) and show that it can leverage and benefit from the new structured approximate posteriors. Through extensive regression experiments, we show that the proposed block-diagonal approximation consistently performs similarly to or better than existing diagonal approximations while maintaining comparable computational costs. Furthermore, the new PEP framework with structured posteriors provides competitive performance across various power hyperparameter settings, offering practitioners flexible alternatives to standard variational approaches.

1 Introduction

Gaussian processes (GPs) provide a principled framework for modelling functions that offer calibrated uncertainty and safeguard against overfitting, among many other benefits (see e.g., Rasmussen & Williams, 2006). However, their computational requirement, cubic in the number of training data N, is prohibitive for many practical applications. This bottleneck motivates the development of a plethora of scalable approximation methods (Quiñonero-Candela & Rasmussen, 2005; Liu et al., 2020), with sparse variational methods using inducing points arguably the most popular (Titsias, 2009; Hensman et al., 2013).

The key idea behind sparse variational GPs (SVGPs) is to approximate the posterior process using a small set of $M \ll N$ inducing points, reducing the computational complexity to $\mathcal{O}(NM^2)$ or $\mathcal{O}(M^3)$ in the batch and stochastic settings, respectively. A key assumption in the standard SVGP approximation is the prior distribution of the non-inducing function values conditioned on the inducing points remains unchanged in the approximate posterior, that is, $q(f_{\neq \boldsymbol{u}}|\boldsymbol{u}) = p(f_{\neq \boldsymbol{u}}|\boldsymbol{u})$. Titsias (2025); Bui et al. (2025) recently showed that relaxing this assumption yields provably tighter variational bounds. In particular, the key innovation is slightly adjusting covariance of $q(f_{\neq \boldsymbol{u}}|\boldsymbol{u})$ by a diagonal scaling matrix $\mathbf{M}$, leading to improved predictive performance while maintaining computational tractability. This approach has the original SVGP approach as a special case when $\mathbf{M} = \mathbf{I}$. Such improvement begs the question: can we achieve even better approximations by considering more expressive structures for $\mathbf{M}$ while preserving efficient computation?

To this end, we propose using block-diagonal structures for $\mathbf{M}$ and show that this choice leads to provably tighter variational bounds compared to existing diagonal approximations while maintaining the same computational complexity and ease of implementation. We then show that these structured approximations can also help with other inference schemes beyond variational inference. Specifically,

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

certain structural choices for $\mathbf{M}$ lead to tractable Power Expectation Propagation (PEP) updates and approximate log marginal likelihood. This greatly extends and improves over the unifying framework of Bui et al. (2017).

The remainder of this paper is organised as follows. Section 2 reviews sparse variational GPs, recent advances in structured approximations, and the PEP framework for sparse GPs. Section 3 presents the proposed block-diagonal variational approximation. Section 4 extends the existing PEP framework with various structured posteriors. Section 5 evaluates the proposed methods on a suite of tasks. We then discuss related work in section 6 and conclude with a discussion of future directions in section 7.

2 Background

We first provide a summary of inducing-point sparse variational Gaussian processes (SVGP; Titsias, 2009; Hensman et al., 2013, 2015; Matthews et al., 2016), a recently proposed tighter bound (Bui et al., 2025; Titsias, 2025), and a power-EP based approach (Bui et al., 2017). Consider the supervised learning setting with an unknown input-output mapping f, a GP prior over this function $p(f|\gamma) = \mathcal{GP}(f; 0, k_\gamma)$, and a pointwise likelihood $p(\boldsymbol{y}|f, \mathbf{X}, \omega) = \prod_n p(y_n|f(\boldsymbol{x}_n), \omega)$, where $\mathbf{X} \in \mathbb{R}^{N\times D}$ and $\boldsymbol{y} \in \mathbb{R}^N$ are the training inputs and outputs, k_γ is the covariance function governed by hyperparameters γ, and ω is the likelihood hyperparameters. In what follows, we will use θ to denote these hyperparameters and, when clear, drop the dependence on θ for brevity. Inference and learning in this model are computationally challenging for large-scale datasets due to the $\mathcal{O}(N^3)$ complexity; thus, efficient approximations are required. Sparse variational methods parameterise an approximate posterior based on M inducing points, $\{\boldsymbol{z} \in \mathbb{R}^{M\times D}, \boldsymbol{u} \in \mathbb{R}^M\}$, with $M \ll N$, as follows,

$$q(f) = p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f}, \boldsymbol{u})q(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u}), \tag{1}$$

where $\boldsymbol{f} = [f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_N)]$. Note that the factorisation here mirrors that in the prior, $p(f) = p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f}, \boldsymbol{u})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})$, where $p(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}; \mathbf{0}, \mathbf{K_{uu}})$, $p(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{f}; \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\boldsymbol{u}, \mathbf{D_{ff}})$, $\mathbf{D_{ff}} = \mathbf{K_{ff}} - \mathbf{Q_{ff}}$, $\mathbf{Q_{ff}} = \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}}$, $\mathbf{K_{ff}} = k(\mathbf{X}, \mathbf{X})$, $\mathbf{K_{fu}} = k(\mathbf{X}, \mathbf{z})$, and $\mathbf{K_{uu}} = k(\mathbf{z}, \mathbf{z})$. Note that we use f to denote the function and $\boldsymbol{f}$ to denote the function values at the training inputs. The resulting variational lower bound to the log marginal likelihood is

$$\mathcal{F}_0 = -\mathrm{KL}[q(\boldsymbol{u})||p(\boldsymbol{u})] - \int q(\boldsymbol{u})\mathrm{KL}[q(\boldsymbol{f}|\boldsymbol{u})||p(\boldsymbol{f}|\boldsymbol{u})] + \sum_n \int q(\boldsymbol{u})q(f(\boldsymbol{x}_n)|\boldsymbol{u}) \log p(y_n|f(\boldsymbol{x}_n)).$$

When $q(\boldsymbol{f}|\boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{f}; \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\boldsymbol{u}, \mathbf{D_{ff}})$, the bound above becomes,

$$\mathcal{F}_1(q(\boldsymbol{u}), \theta) = -\mathrm{KL}[q(\boldsymbol{u})||p(\boldsymbol{u})] + \sum_n \int q(\boldsymbol{u})p(f(x_n)|\boldsymbol{u}) \log p(y_n|f(x_n)), \tag{2}$$

commonly known as the uncollapsed SVGP bound (Hensman et al., 2015; Titsias, 2009). This bound conveniently allows both (i) tractable computation [$\mathcal{O}(NM^2)$ in the batch setting or $\mathcal{O}(BM^2 + M^3)$ where B is the batch size in the mini-batch setting] and (ii) tractably handling of non-Gaussian likelihoods using quadrature or Monte Carlo estimation for the expected log-likelihood terms. For the Gaussian likelihood, the bound can be simplified to

$$\mathcal{F}_{1,r}(q(\boldsymbol{u}), \theta) = -\mathrm{KL}[q(\boldsymbol{u})||p(\boldsymbol{u})] + \sum_n \left[\int q(\boldsymbol{u}) \log \mathcal{N}(y_n; \mathbf{k}_{f_n\mathbf{u}}\mathbf{K_{uu}^{-1}}\boldsymbol{u}, \sigma^2) - \frac{d_{nn}}{2\sigma^2}\right], \tag{3}$$

where σ^2 is the observation noise and $d_{nn} = [\mathbf{D_{ff}}]_{nn}$. Furthermore, an optimal form for $q(\boldsymbol{u})$ can be found, $q(\boldsymbol{u}) \propto p(\boldsymbol{u})\mathcal{N}(\boldsymbol{y}; \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\boldsymbol{u}, \sigma^2\mathbf{I}_N)$, yielding the following analytic collapsed bound (Titsias, 2009),

$$\mathcal{F}_{1,rc}(\theta) = \log \mathcal{N}(\boldsymbol{y}; \mathbf{0}, \mathbf{Q_{ff}} + \sigma^2\mathbf{I}_N) - \frac{1}{2\sigma^2}\sum_n d_{nn}. \tag{4}$$

The SVGP approach above has arguably been the most popular scalable GP approach in the literature. More recently, Bui et al. (2025); Titsias (2025) show that this approach can be improved by relaxing the $q(\boldsymbol{f}|\boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})$ assumption. Specifically, when $q(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{f}; \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\boldsymbol{u}, \mathbf{D_{ff}^{1/2}}\mathbf{M}\mathbf{D_{ff}^{1/2}})$,

where $\mathbf{M} = \text{diag}([m_1, \ldots, m_N])$, the uncollapsed and collapsed bounds in the regression case are:

$$\mathcal{F}_{2,r}(q(\boldsymbol{u}), \theta) = -\text{KL}[q(\boldsymbol{u})||p(\boldsymbol{u})] + \sum_n \left[\int q(\boldsymbol{u}) \log \mathcal{N}(y_n; \mathbf{k}_{f_n \mathbf{u}} \mathbf{K}_{\mathbf{uu}}^{-1} \boldsymbol{u}, \sigma^2) - \frac{1}{2} \log \left(1 + \frac{d_{nn}}{\sigma^2} \right) \right] \quad (5)$$

$$\mathcal{F}_{2,rc}(\theta) = \log \mathcal{N}(\boldsymbol{y}; \mathbf{0}, \mathbf{Q}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_N) - \frac{1}{2} \sum_n \log \left(1 + \frac{d_{nn}}{\sigma^2} \right). \quad (6)$$

Note that the optimal form for m_n is $m_n = \sigma^2/(\sigma^2 + d_{nn}) < 1$; and eqs. (5) and (6) are tighter than eqs. (3) and (4) for fixed θ and $q(\boldsymbol{u})$ since $\log\left(1 + d_{nn}/\sigma^2\right) \leq d_{nn}/\sigma^2$.

The posterior approximation in eq. (1) can also be used in other deterministic inference strategies. For example, in the regression case, for $q(\boldsymbol{f}|\boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})$, Bui et al. (2017) showed that Power-Expectation Propagation (PEP) yields an analytic collapsed approximate marginal likelihood,

$$\mathcal{F}_{3,rc}(\theta) = \log \mathcal{N}(\boldsymbol{y}; \mathbf{0}, \mathbf{Q}_{\mathbf{ff}} + \alpha \mathbf{D}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_N) - \frac{1-\alpha}{2\alpha} \sum_n \log \left(1 + \alpha \frac{d_{nn}}{\sigma^2} \right), \quad (7)$$

and a closed form $q(\boldsymbol{u})$, $q(\boldsymbol{u}) \propto p(\boldsymbol{u}) \mathcal{N}(\boldsymbol{y}; \mathbf{K}_{\mathbf{fu}} \mathbf{K}_{\mathbf{uu}}^{-1} \boldsymbol{u}, \alpha \mathbf{D}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_N)$, where α is the power hyperparameter in PEP. This framework encompasses a multitude of approximations, such as the SVGP approximation (as $\alpha \to 0$) and FITC (Snelson & Ghahramani, 2005; Qi et al., 2010) ($\alpha = 1$).

3 A block-diagonal structured variational approximation

We first consider the following posterior approximation:

$$q(f) = p(f_{\neq \boldsymbol{f}, \boldsymbol{u}} | \boldsymbol{f}, \boldsymbol{u}) q(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u}), \quad q(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{f}; \mathbf{K}_{\mathbf{fu}} \mathbf{K}_{\mathbf{uu}}^{-1} \boldsymbol{u}, \mathbf{C}),$$

where we have not posited a form for the covariance $\mathbf{C}$. Interestingly, this leads to the familiar optimal form for $q(\boldsymbol{u})$, $q(\boldsymbol{u}) \propto p(\boldsymbol{u}) \mathcal{N}(\boldsymbol{y}; \mathbf{K}_{\mathbf{fu}} \mathbf{K}_{\mathbf{uu}}^{-1} \boldsymbol{u}, \sigma^2 \mathbf{I}_N)$. The resulting collapsed bound is,

$$\mathcal{F}(\theta) = \log \mathcal{N}(\boldsymbol{y}; 0, \mathbf{Q}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_N) - \frac{1}{2} \text{trace}[(\sigma^{-2} \mathbf{I}_N + \mathbf{D}_{\mathbf{ff}}^{-1}) \mathbf{C}] - \frac{1}{2} \log |\mathbf{C}^{-1} \mathbf{D}_{\mathbf{ff}}| + \frac{N}{2}.$$

Except for some special cases, the bound above is as expensive as the original log marginal likelihood to compute. Specifically, as shown in the background, $\mathbf{C} = \mathbf{D}_{\mathbf{ff}}^{1/2} \mathbf{M} \mathbf{D}_{\mathbf{ff}}^{1/2}$ with $\mathbf{M} = \mathbf{I}_N$ (Titsias, 2009) or $\mathbf{M} = m\mathbf{I}_N$ (Artemev et al., 2021) or $\mathbf{M} = \text{diag}(\{m_n\}_{n=1}^N)$ (Titsias, 2025; Bui et al., 2025) admit tractability, and each move (from $\mathbf{I}_N$ to $m\mathbf{I}_N$, and from $m\mathbf{I}_N$ to $\text{diag}(\{m_n\}_{n=1}^N)$) makes the bound tighter. It is thus natural to enquire what structure to encode in $\mathbf{M}$ to further improve the bound, retain tractable computation, and potentially improve predictive performance.

We now consider one such structure, a block-diagonal $\mathbf{M}$, $\mathbf{M} = \text{blkdiag}(\{\mathbf{m}_b\}_{b=1}^B)$, where B is the number of blocks and $\mathbf{m}_b \in \mathbb{R}^{N_b \times N_b}$. Substituting this into the bound above gives

$$\mathcal{F}_4(\theta) = \log \mathcal{N}(\boldsymbol{y}; 0, \mathbf{Q}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_N) - \frac{1}{2} \sum_b \left[\frac{1}{\sigma^2} \text{trace}[\mathbf{m}_b \mathbf{D}_{\mathbf{f}_b \mathbf{f}_b}] + \text{trace}[\mathbf{m}_b] - \log |\mathbf{m}_b| - N_b \right].$$

We can obtain the optimal $\mathbf{m}_b$, $\mathbf{m}_b = (\mathbf{I}_b + \sigma^{-2} \mathbf{D}_{\mathbf{f}_b \mathbf{f}_b})^{-1}$, leading to the following collapsed bound,

$$\mathcal{F}_{4,rc}(\theta) = \log \mathcal{N}(\boldsymbol{y}; 0, \mathbf{Q}_{\mathbf{ff}} + \sigma^2 \mathbf{I}_N) - \frac{1}{2} \sum_b \log |\mathbf{I}_b + \sigma^{-2} \mathbf{D}_{\mathbf{f}_b \mathbf{f}_b}|. \quad (8)$$

Due to the Hadamard's inequality, $|\mathbf{I}_b + \sigma^{-2} \mathbf{D}_{\mathbf{f}_b \mathbf{f}_b}| < \prod_i (1 + \sigma^{-2} [\mathbf{D}_{\mathbf{f}_b \mathbf{f}_b}]_{ii})$, and thus $\log |\mathbf{I}_b + \sigma^{-2} \mathbf{D}_{\mathbf{f}_b \mathbf{f}_b}| < \sum_i \log(1 + \sigma^{-2} [\mathbf{D}_{\mathbf{f}_b \mathbf{f}_b}]_{ii})$. In other words, the bound in eq. (8) [$\mathbf{M}$ is block-diagonal] is provably tighter than the bound in eq. (6) [$\mathbf{M}$ is diagonal].

Similar to the standard SVGP approach, for large datasets, it is more convenient to work with the following uncollapsed bound that supports stochastic optimisation,

$$\mathcal{F}_{4,r}(\cdot) = -\text{KL}[q(\boldsymbol{u})||p(\boldsymbol{u})] + \sum_{b=1}^B \left[\int q(\boldsymbol{u}) \log \mathcal{N}(\boldsymbol{y}_b; \mathbf{K}_{\mathbf{f}_b \mathbf{u}} \mathbf{K}_{\mathbf{uu}}^{-1} \boldsymbol{u}, \sigma^2 \mathbf{I}_b) - \frac{1}{2} \log |\mathbf{I}_b + \sigma^{-2} \mathbf{D}_{\mathbf{f}_b \mathbf{f}_b}| \right]. \quad (9)$$

If the B blocks are of roughly equal size, computing the bound in eq. (9) using the entire training set takes $\mathcal{O}(M^3 + NM^2 + B[N/B]^3)$. However, in practice, we perform stochastic optimisation,

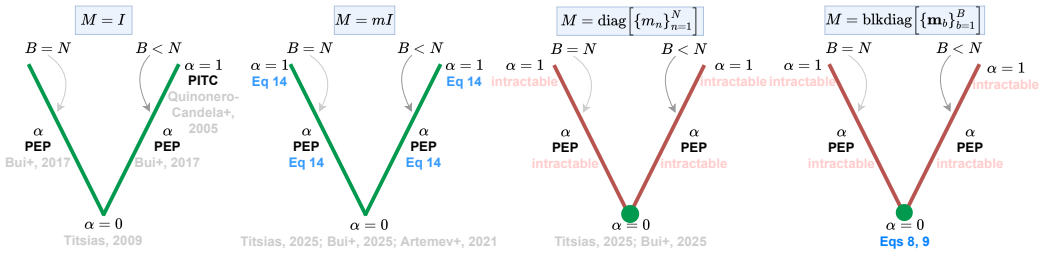

Figure 1: Connections between the sparse GP regression methods from the Power-EP perspective. Green means computationally tractable, red means intractable, and blue represents the new methods presented in this paper. $B < N$ means the training points are partitioned into B disjoint blocks. $B = N$ means having the same number of blocks as training points, i.e., block size equal to 1.

where we unbiasedly approximate the sum over blocks in eq. (9) using one random block to obtain the stochastic bound

$$\widetilde{\mathcal{F}}_{4,r}(\cdot) = -\mathrm{KL}[q(\boldsymbol{u})||p(\boldsymbol{u})] + B\left[\int q(\boldsymbol{u}) \log \mathcal{N}(\boldsymbol{y}_b; \mathbf{K}_{\mathbf{f}_b\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\boldsymbol{u}, \sigma^2\mathbf{I}_b) - \frac{1}{2}\log|\mathbf{I}_b + \sigma^{-2}\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}|\right], \quad (10)$$

based on which we perform stochastic gradient updates by cycling over the B blocks. If we judiciously choose the block size $\frac{N}{B}$ to be M (i.e., block size equals to the number of inducing points), the computational requirement per iteration is only $\mathcal{O}(M^3)$. Therefore, eq. (10) has a small implementation overhead compared to standard stochastic sparse GP objectives. The precise extra overhead involves taking the Cholesky decomposition of $\mathbf{I}_b + \sigma^{-2}\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}$, needed when computing the log-determinant regularisation term.

We will next consider a special case. When we let all $\mathbf{m}_b$ matrices to be the same, $\mathbf{m}_b = \mathbf{m}$, we arrive at the optimal $\mathbf{m}$, $\mathbf{m} = (\mathbf{I}_b + B^{-1}\sigma^{-2}\sum_b \mathbf{D}_{\mathbf{f}_b\mathbf{f}_b})^{-1}$, and the resulting collapsed bound,

$$\mathcal{F}_5(\theta) = \log \mathcal{N}(\boldsymbol{y}; 0, \mathbf{Q}_{\mathbf{f}\mathbf{f}} + \sigma^2\mathbf{I}_N) - \frac{B}{2}\log|\mathbf{I}_b + \frac{1}{B\sigma^2}\sum_b \mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}|. \quad (11)$$

Since the log-determinant is a concave function on the cone of positive definite matrices, we can apply Jensen's inequality to show that the bound above is less tight compared to eq. (8). As the block size equals one, this becomes the *spherical* bound in Titsias (2025); Artemev et al. (2021).

A disadvantage of diagonal and block diagonal structures in $\mathbf{M}$ is the expensive predictive covariance. However, we can approximate it by reverting to using $q(\boldsymbol{f}|\boldsymbol{u}) \approx p(\boldsymbol{f}|\boldsymbol{u})$ at test time. Bui et al. (2025) noted that this approximation does not degrade the performance compared to the expensive exact predictive distribution. In other words, in practice, we only use the new structured posterior in training, and therefore, any improvement in predictive performance at test time will come from better $q(\boldsymbol{u})$ and hyperparameters.

4 A more general approximation based on Power Expectation Propagation

Although the variational sparse GP approach has captured the spotlight in the sparse GP literature, Bui et al. (2017) showed various variants of PEP can be as competitive or better. We will now revisit the framework of Bui et al. (2017) and explore how it can be improved by leveraging the recent innovation in structured posterior approximations (Titsias (2025); Bui et al. (2025) and section 3) originally developed in the variational inference setting. We first write down the joint density of the exact model and the approximate posterior,

$$p(f, \boldsymbol{y}) = p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f}, \boldsymbol{u})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})\prod_{n=1}^N p(y_n|\boldsymbol{f}_n) \quad (12)$$

$$q(f) \propto p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f}, \boldsymbol{u})q(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})\prod_{b=1}^B t_b(\boldsymbol{u}) \quad (13)$$

where the N training points are partitioned into B disjoint blocks, and the factors $t_b(\boldsymbol{u})$ are assumed to be Gaussian. Instead of using $q(\boldsymbol{f}|\boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})$ as in Bui et al. (2017), we consider $q(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{f}; \mathbf{K_{fu}}\mathbf{K_{uu}}^{-1}\boldsymbol{u}; \mathbf{D_{ff}}^{1/2}\mathbf{M}\mathbf{D_{ff}}^{1/2})$ where $\mathbf{M} = \text{blkdiag}(\{\mathbf{m}_b\}_{b=1}^B)$, that is, the blocks in $\mathbf{M}$ match that of the likelihood partitions.

The PEP procedure (Minka, 2004) iteratively updates $t_b(\boldsymbol{u})$ by (i) first remove an α-fraction of $t_b(\boldsymbol{u})$ from $q(f)$ to form the cavity distribution, $q^{\backslash b}(f) = q(f)/t_b^\alpha(\boldsymbol{u})$, (ii) incorporate an α-fraction of the likelihood for the b-th block $p(\boldsymbol{y}_b|\boldsymbol{f}_b) = \prod_{n=1}^{N_b} p(y_n|f_n)$ to form the tilted distribution, $\tilde{q}(f) = q^{\backslash b}(f)p^\alpha(\boldsymbol{y}_b|\boldsymbol{f}_b)$, (iii) find a new approximation $q(f)$ that minimises $\text{KL}[\tilde{q}(f)||q(f)]$, and (iv) adjust $t_b(\boldsymbol{u})$ based on the new posterior using $t_b(\boldsymbol{u}) = [q(f)/q^{\backslash b}(f)]^{1/\alpha}$ *or* $t_b(\boldsymbol{u}) \leftarrow t_b^{1-\alpha}(\boldsymbol{u})[q(f)/q^b(f)]$. These steps are repeated for all blocks until convergence. Readers might have noticed that step (iii) is a daunting task as it involves moment matching for the entire Gaussian processes; however, due to the structure of the approximate posterior $q(f)$, it is sufficient to perform moment matching for the finite function values $\boldsymbol{u}$ (Bui et al., 2017). In addition, this procedure returns an estimate of the log marginal likelihood that can be used for hyperparameter optimisation.

Mirroring the derivation in Bui et al. (2017), we can show the optimal form for $t_b(\boldsymbol{u})$ has rank $N_b = |\boldsymbol{f}_b|$, $t_b(\boldsymbol{u}) = \mathcal{N}(\mathbf{K}_{\mathbf{f}_b\mathbf{u}}\mathbf{K_{uu}}^{-1}\boldsymbol{u}; \boldsymbol{g}_b, \boldsymbol{v}_b)$. In the regression case, $\boldsymbol{g}_b = \boldsymbol{y}_b$ and $\boldsymbol{v}_b = \alpha[\mathbf{D_{ff}}^{1/2}\mathbf{M}\mathbf{D_{ff}}^{1/2}]_{bb} + \sigma^2\mathbf{I}_b$. The full derivation is rather lengthy and is included in the appendix; however, one can verify that this is a stable fixed point of the procedure by noting that the α-fraction of $t_b(\boldsymbol{u})$ is identical to the contribution of $p^\alpha(y_b|\boldsymbol{f}_b)$ to the posterior at $\boldsymbol{u}$, $\int d\boldsymbol{f}_b q(\boldsymbol{f}_b|\boldsymbol{u})p^\alpha(\boldsymbol{y}_b|\boldsymbol{f}_b)$. The optimal $q(\boldsymbol{u})$ is thus $q(\boldsymbol{u}) \propto p(\boldsymbol{u})\mathcal{N}(\boldsymbol{y}; \mathbf{K_{fu}}\mathbf{K_{uu}}^{-1}\boldsymbol{u}, \alpha\text{blkdiag}(\{[\mathbf{D_{ff}}^{1/2}\mathbf{M}\mathbf{D_{ff}}^{1/2}]_{bb}\}_{b=1}^B) + \sigma^2\mathbf{I}_N)$. Furthermore, for the regression case, we can further derive the approximate marginal likelihood,

$$\mathcal{F}_{5,rc}(\theta, \mathbf{M}) = \log\mathcal{N}(\boldsymbol{y}; \mathbf{0}, \mathbf{Q_{ff}} + \alpha\text{blkdiag}(\{[\mathbf{D_{ff}}^{1/2}\mathbf{M}\mathbf{D_{ff}}^{1/2}]_{bb}\}_{b=1}^B) + \sigma^2\mathbf{I}_N) + \sum_b \left[-\frac{1-\alpha}{2\alpha}\log\left|\mathbf{I}_b + \alpha\frac{[\mathbf{D_{ff}}^{1/2}\mathbf{M}\mathbf{D_{ff}}^{1/2}]_{bb}}{\sigma^2}\right| - \frac{1}{2\alpha}\log|\mathbf{I}_b + \alpha(\mathbf{m}_b - \mathbf{I}_b)| + \frac{1}{2}\log|\mathbf{m}_b| \right].$$

We note that, for a general α and $\mathbf{M}$, including diagonal and block-diagonal cases, the PEP procedure above as well the approximate marginal likelihood for the regression case is computationally intractable (i.e., cubic in N) due to the need to find the (block-)diagonal of $\mathbf{D_{ff}}^{1/2}\mathbf{M}\mathbf{D_{ff}}^{1/2}$. We will now discuss the tractable special cases.

Remark 1 *When* $\mathbf{M}$ *is diagonal or block-diagonal, the approximate marginal likelihood and posterior approximation above are only tractable as* $\alpha \to 0$*. Specifically, when* $\mathbf{M} = \text{diag}(\{m_n\}_{n=1}^N)$*, the objective becomes the variational bound of Titsias (2025); Bui et al. (2025), and when* $\mathbf{M} = \text{blkdiag}(\{\mathbf{m}_b\}_{b=1}^B)$*, the objective matches the variational bound in eq.* (8)*.*

Remark 2 *When* $\mathbf{M} = m\mathbf{I}_N$*, the approximate marginal likelihood and posterior approximation are computationally tractable for all* α*'s. In particular, the optimal* $q(\boldsymbol{u})$ *is* $q(\boldsymbol{u}) \propto p(\boldsymbol{u})\mathcal{N}(\boldsymbol{y}; \mathbf{K_{fu}}\mathbf{K_{uu}}^{-1}\boldsymbol{u}, m\alpha\text{blkdiag}(\{\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}\}_{b=1}^B) + \sigma^2\mathbf{I}_N)$*, and the approximate marginal likelihood becomes,*

$$\mathcal{F}_6(\theta, m) = \log\mathcal{N}(\boldsymbol{y}; 0, \mathbf{Q_{ff}} + m\alpha\text{blkdiag}(\{\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}\}_{b=1}^B) + \sigma^2\mathbf{I}_N) - \frac{1-\alpha}{2\alpha}\sum_b\left[\log\left|\mathbf{I}_b + \frac{\alpha m\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}}{\sigma^2}\right|\right] - \frac{N}{2\alpha}\log(1+\alpha(m-1)) + \frac{N}{2}\log(m). \quad (14)$$

In this special case, we note the following. First, as a sanity check, we can see that when $m = 1$, we recover the Power-EP approximate marginal likelihood of Bui et al. (2017):

$$\mathcal{F}_{6,m=1}(\theta) = \log\mathcal{N}(\boldsymbol{y}; 0, \mathbf{Q_{ff}} + \alpha\text{blkdiag}(\mathbf{D_{ff}}) + \sigma^2\mathbf{I}_N) - \frac{1-\alpha}{2\alpha}\sum_n \log\left|\mathbf{I}_b + \alpha\frac{\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}}{\sigma^2}\right|.$$

Second, when $\alpha = 1$, the objective in eq. (14) becomes $\mathcal{F}_{6,\alpha=1}(\theta, m) = \log\mathcal{N}(\boldsymbol{y}; 0, \mathbf{Q_{ff}} + m\text{blkdiag}(\mathbf{D_{ff}}) + \sigma^2\mathbf{I}_N)$. When $m = 1$, this becomes the PITC marginal likelihood and for block size equal to one it further reduces to FITC (Quiñonero-Candela & Rasmussen, 2005).

Third, as $\alpha \to 0$, we recover the *spherical* bound in (Bui et al., 2025; Titsias, 2025; Artemev et al., 2021). Only in this setting, we can derive the optimal $m = (1 + N^{-1}\sum_n d_n/\sigma^2)^{-1}$.

Finally, inspired by the uncollapsed variational bound, we can optimise an uncollapsed version of eq. (14) that supports stochastic optimisation as follows,

$$\mathcal{F}_{6,r}(q(\boldsymbol{u}), \theta, \mathbf{M}) = -\mathrm{KL}[q(\boldsymbol{u})||p(\boldsymbol{u})] + \sum_b \left[\int q(\boldsymbol{u}) \log \mathcal{N}(\boldsymbol{y}_b; \mathbf{K}_{\mathbf{f}_b\mathbf{u}}\mathbf{K}_{\mathbf{u}\mathbf{u}}^{-1}\boldsymbol{u}, m\alpha\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b} + \sigma^2\mathbf{I}_b) \right]$$
$$- \frac{1-\alpha}{2\alpha} \sum_b \left[\log \left| \mathbf{I}_b + \frac{\alpha m \mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}}{\sigma^2} \right| \right] - \frac{N}{2\alpha} \log\left(1 + \alpha(m-1)\right) + \frac{N}{2} \log(m).$$

That is, instead of running the PEP procedure, we can optimise the objective above to yield the same fixed point as PEP. We attempt to visualise the connections between the methods, the special cases and the broader literature in fig. 1.

5 Experiments

Having described the new block-diagonal structure in sparse variational GPs and revisited the unified work of Bui et al. (2017) in light of the new approximate posteriors, we will detail the experiments to qualitatively investigate (i) if the proposed block-diagonal approximation in section 3 yields better performance and, if yes, how, and (ii) whether having $m \neq 1$ benefits power expectation propagation in section 4 the same way it does to variational inference. All experiments were done on either a V100 GPU or a MacBook laptop. We provide an implementation here `https://github.com/thangbui/tighter_sparse_gp`.

5.1 1-D regression and biases in hyperparameter estimation

We first illustrate the difference between the proposed and existing methods on a simple 1D regression problem (Snelson & Ghahramani, 2005). In particular, we compare Titsias' collapsed bound in eq. (4) [SGPR], the bound of Titsias (2025); Bui et al. (2025) in eq. (6) [T-SGPR], the bound with block diagonal $\mathbf{M}$ in eq. (8) with 10 and 20 blocks [20 and 10 data points per block, respectively, BT-SGPR], the PEP approach of Bui et al. (2017) with $\alpha = 0.5$ [PEP], and the PEP approach in eq. (14) with $B = N$ and $\alpha = 0.5$ [T-PEP]. We used 5 inducing points in this experiment. The key results are summarised in fig. 2. It can be observed that (i) the block-diagonal approximation improves over the diagonal one in this example, (ii) increasing the number of training points in each block tightens the bound, (iii) the structured posterior approximation also helps in PEP, and (iv) hyperparameter optimisation using a more structured approximation tend to result in a smaller noise variance and a larger kernel variance. We note that the PEP approximate marginal likelihood is not guaranteed to be a lower bound and therefore optimising it can result in pathological behaviours, for example, when $\alpha = 1$, the noise variance can be severely underestimated (Bauer et al., 2016).

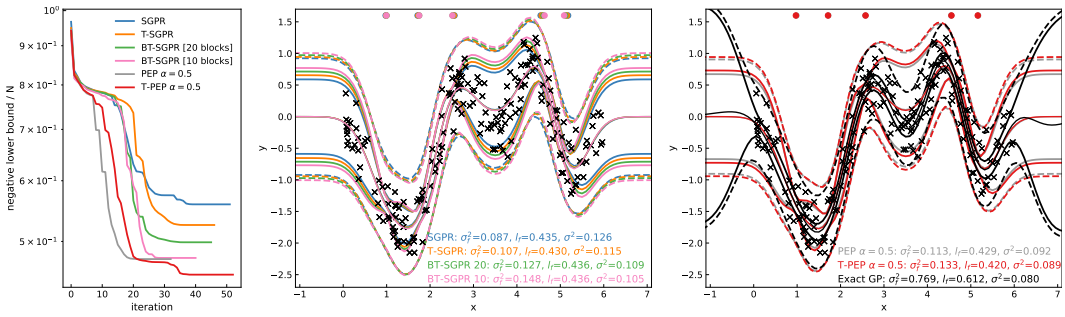

Figure 2: *Left:* Variational bounds during training on the Snelson dataset. *Middle and right:* Predictive mean and intervals using various methods and the final hyperparameter values.

To further investigate point (iv), we picked a subset of the KIN40K dataset with 5,000 data points, and ran an experiment to compare the sparse approximations with exact GP. For each method, we recorded in table 1 the exact or approximate log marginal likelihood, the predictive performance measured by root mean squared error (RMSE) and log likelihood (LL), the noise standard deviation σ and the kernel lengthscales. Similar to the observation in the Snelson dataset above, the noise estimate is smaller when moving from $\mathbf{M} = \mathbf{I}_N$ to increasingly more structured $\mathbf{M}$, translating

to better predictions. The trend seems to be consistent across two numbers of inducing points. In addition, there is no notable difference in the lengthscales between PEP, T-PEP, and the structured variational approximations; however, these methods tend to *leverage* more dimensions than SGPR for $M = 256$.

Table 1: Exact/approximate marginal likelihoods, predictive performance, and lengthscales given by various methods on 5,000 samples from the KIN40K dataset.

	M = 256					M = 512				
Method	Obj.	RMSE	LL	σ	lengthscales	Obj.	RMSE	LL	σ	lengthscales
Exact	-0.66	0.12	0.80	0.00		-0.66	0.12	0.80	0.00	
SGPR	0.88	0.26	-0.14	0.30		0.66	0.22	0.02	0.25	
T-SGPR	0.78	0.22	-0.06	0.26		0.51	0.18	0.11	0.21	
BT-SGPR [50]	0.75	0.22	-0.05	0.25		0.50	0.18	0.12	0.20	
BT-SGPR [10]	0.66	0.20	-0.03	0.23		0.44	0.17	0.13	0.19	
PEP [0.5]	0.66	0.23	-0.02	0.22		0.42	0.20	0.14	0.19	
T-PEP [0.5]	0.48	0.20	0.02	0.18		0.18	0.16	0.19	0.14	

5.2 Block-diagonal structured variational approximation

We next ran an experiment to validate the utility of the proposed block-structured approximation in section 3 on four real-world regression datasets[1]. For each dataset and each inducing point configuration ($M = 256$ or $M = 512$), we compare the uncollapsed variational bounds of Titsias (2009); Hensman et al. (2015) [eq. (3), SVGP], Titsias (2025); Bui et al. (2025) [eq. (5), T-SVGP], and the proposed bound in eq. (9) [BT-SVGP], corresponding to $\mathbf{M} = \mathbf{I}_N$, $\mathbf{M} = \mathrm{diag}(\{m_n\}_{n=1}^N])$, and $\mathbf{M} = \mathrm{blkdiag}(\{\mathbf{m}_b\}_{b=1}^B)$, respectively. We repeated the experiment 10 times, each using a random train/test split, a batch size of 500 (also the block size), random partitioning of the training data into blocks, and 300 epochs for training. The average variational bound (ELBO) and test performance after training are shown in fig. 3. Similar to the earlier experiments, the benefit of the block-structured approximation is also clearly demonstrated here: it tightens the variational bound compared to that of the diagonal $\mathbf{M}$ and consistently yields comparable or better predictive performance. We note again that (i) the estimated observation noise tends to be smaller when employing the new bound (see the appendix), and (ii) there is a minimal implementation overhead compared to Titsias (2009, 2025); Bui et al. (2025) to result in these gains.

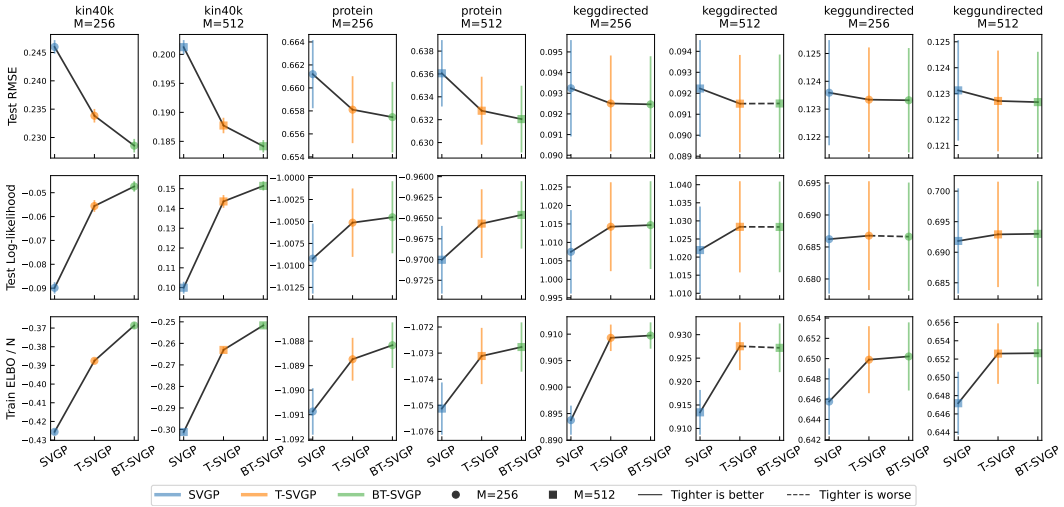

Figure 3: Lower bounds (ELBO) and predictive performance of various variational methods with $\mathbf{M} = \mathbf{I}_N$ [SVGP], $\mathbf{M} = \mathrm{diag}(\{m_n\}_{n=1}^N)$ [T-SVGP], and $\mathbf{M} = \mathrm{blkdiag}(\{\mathbf{m}_b\}_{b=1}^B)$ [BT-SVGP].

[1]We used the splits available in this repository `https://github.com/treforevans/uci_datasets`.

5.3 Power-EP with a structured approximate posterior [$\mathbf{M} = m\mathbf{I}_N$]

As shown in section 4, the structured approximate posterior considered by Titsias (2025) can be utilised in PEP and in the regression case, the approximate posterior and marginal likelihood are analytically available. To evaluate its practical utility, we ran an experiment on five small regression datasets, comparing the PEP approach of Bui et al. (2017) [$\mathbf{M} = \mathbf{I}_N$] to the proposed approach in section 4 [$\mathbf{M} = m\mathbf{I}_N$]. The typical performance across various inducing point configurations is shown in fig. 4, with the full results included in the appendix. It is noticeable that the Power-EP scheme with $m \neq 1$ tends to outperform the corresponding setting when $m = 1$. To elucidate the trend, we plot the difference between the performance of $\mathbf{M} = \mathbf{I}_N$ and $\mathbf{M} = m\mathbf{I}_N$ in fig. 5. We note that $m \neq 1$ outperforms $m = 1$ on all datasets in terms of RMSE, but log-likelihood performance degrades when α is closer to 1. These results suggest that for $m \neq 1$, intermediate α values such as 0.5 are most competitive in terms of both RMSE and LL, in line with recommendations from Bui et al. (2017) when $m = 1$.

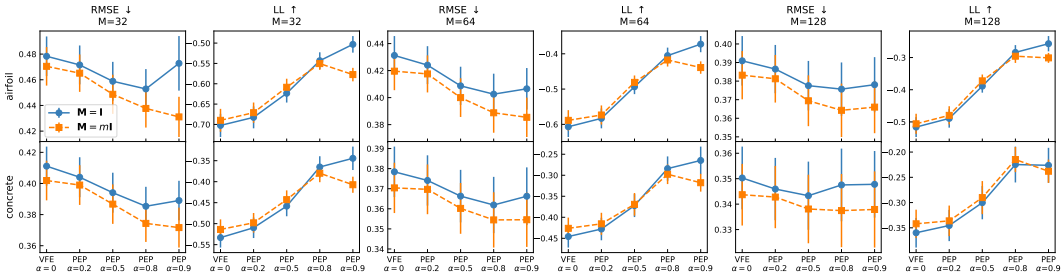

Figure 4: Predictive performance of power expectation propagation with $\mathbf{M} = \mathbf{I}_N$ and $\mathbf{M} = m\mathbf{I}_N$ on two UCI datasets. Results for other datasets are in the appendix.

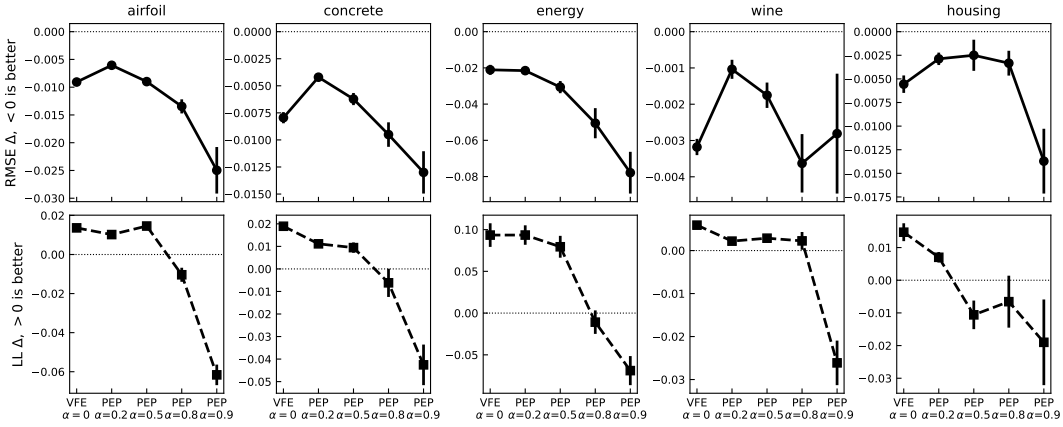

Figure 5: Difference in PEP performance between $\mathbf{M} = \mathbf{I}_N$ and $\mathbf{M} = m\mathbf{I}_N$ on five UCI datasets.

6 Related work

The use of inducing points for sparse approximations in Gaussian processes has a rich history, to name a few approaches, sparse online GPs (Csató & Opper, 2002), DTC (Seeger et al., 2003), FITC approximation (Snelson & Ghahramani, 2005), and PITC (Quiñonero-Candela & Rasmussen, 2005). The most notable was the variational approach of Titsias (2009), who introduced a principled method for selecting inducing points by optimising a variational lower bound. Hensman et al. (2013, 2015) extended this approach to enable stochastic optimisation and non-Gaussian likelihoods, significantly broadening the applicability of sparse GPs to large datasets. Other work on inducing point methods have exploited Kronecker products (Wilson & Nickisch, 2015), nearest neighbour structures (Tran et al., 2021; Wu et al., 2022) and inter-domain inducing points (Lázaro-Gredilla &

Figueiras-Vidal, 2009; Hensman et al., 2018). Also, recent theoretical work (Burt et al., 2020) studied the approximation convergence with respect to the number of inducing points.

Our work is most closely related to the recent advances by Titsias (2025); Bui et al. (2025), who showed that relaxing the standard assumption with diagonal scaling matrices improves the variational bound. Our block-diagonal extension naturally builds upon this line of work, showing practical benefits. Similarly, our extension of the PEP framework builds directly on Bui et al. (2017), expanding their unifying perspective by incorporating structured posterior approximations. Note that our work is distinct from the PITC approximation. PITC is derived from the prior modification perspective, where the prior is modified so that the blocks of function values are conditionally independent given the inducing points. Bui et al. (2017) showed that this is equivalent to EP when retaining the prior conditional in the approximate posterior, which differs from our proposed structured conditional distribution.

A key component in the sparse GP approximate posterior is $q(\boldsymbol{u})$, and imposing additional structures for this object will likely lead to improvement. For example, Shi et al. (2020) showed that $q(\boldsymbol{u})$ can be parameterised by two sets of inducing points, *orthogonal* to each other, leading to better predictive performance at a much lower compute cost compared to doubling up the inducing points in the standard SVGP approximation. This line of work is complementary to our work here, as it focuses on a different aspect of the posterior, and thus, the two approaches can be combined.

A well-known pathology of variational sparse GP regression is the large estimated observation noise variance (Bauer et al., 2016). It can be partially alleviated by changing the objective function (Jankowiak et al., 2019) or mixing separate schemes for learning and inference (Li et al., 2023). Our work shows that principled structured variational approximations can also partly address this issue.

7 Summary

Approximation schemes using inducing points are the method of choice for scaling GP models to large datasets. We show that (i) these methods can be improved by introducing additional structures in the approximate posterior and (ii) these new structures can be applied to various inference strategies, including PEP and variational inference. The resulting methods show comparable or better predictive performance and smaller hyperparameter estimation biases in many standard regression tasks.

There are several potential future directions. First, we have assumed that the size of the data blocks in a dataset is the same and the data partitioning in the experiments was random, but these can be adjusted based on the data characteristics, potentially tightening the variational objective further. Second, the power hyperparameter α in PEP can be made private per block; this will require an understanding of when variational or EP might work best and how to dynamically select α. Third, a full discussion for non-Gaussian likelihoods and models beyond GP regression (e.g., deep GPs, GP latent variable models) and how they benefit from structured approximations is a promising exploratory direction.

Acknowledgments

We would like to thank the anonymous reviewers for the feedback. TDB would like to thank the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government, for the computing resources.

References

Artem Artemev, David R Burt, and Mark van der Wilk. Tighter bounds on the log marginal likelihood of Gaussian process regression using conjugate gradients. In *International Conference on Machine Learning*, pp. 362–372, 2021.

Matthias Bauer, Mark van der Wilk, and Carl Edward Rasmussen. Understanding probabilistic sparse Gaussian process approximations. In *Advances in Neural Information Processing Systems*, pp. 1533–1541, 2016.

Thang D. Bui, Josiah Yan, and Richard E. Turner. A unifying framework for Gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research*, 18(104):1–72, 2017.

Thang D. Bui, Matthew Ashman, and Richard E. Turner. Tighter sparse variational Gaussian processes, 2025.

David R. Burt, Carl Edward Rasmussen, and Mark van der Wilk. Convergence of sparse variational inference in Gaussian processes regression. *Journal of Machine Learning Research*, 21(131):1–63, 2020.

Lehel Csató and Manfred Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3): 641–668, 03 2002.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Conference on Uncertainty in Artificial Intelligence*, pp. 282–290, 2013.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 351–360, 2015.

James Hensman, Nicolas Durrande, and Arno Solin. Variational fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52, 2018.

Jose Miguel Hernández-Lobato, Yingzhen Li, Mark Rowland, Daniel Bui, Thang D. and Hernández-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *International Conference on Machine Learning*, pp. 1511–1520, 2016.

Martin Jankowiak, Geoff Pleiss, and Jacob R Gardner. Sparse Gaussian process regression beyond variational inference. 2019.

Miguel Lázaro-Gredilla and Aníbal Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

Rui Li, ST John, and Arno Solin. Improving hyperparameter learning under approximate inference in Gaussian process models. In *International Conference on Machine Learning*, 2023.

Yingzhen Li, Jose Miguel Hernández-Lobato, and Richard E. Turner. Stochastic expectation propagation. In *Advances in Neural Information Processing Systems*, pp. 2323–2331, 2015.

Haitao Liu, Yew Ong, Xiaobo Shen, and Jianfei Cai. When Gaussian process meets big data: A review of scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–19, 01 2020.

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 *International Conference on Artificial Intelligence and Statistics*, pp. 231–239, 2016.

Thomas Minka. Power EP. Technical report, Microsoft Research, 2004.

Yuan Qi, Ahmed H Abdel-Gawad, and Thomas P Minka. Sparse-posterior Gaussian processes for general likelihoods. In *Conference on Uncertainty in Artificial Intelligence*, pp. 450–457, 2010.

Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(65):1939–1959, 2005.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Matthias W. Seeger, Christopher K. I. Williams, and Neil D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In *International Workshop on Artificial Intelligence and Statistics*, pp. 254–261, 2003.

Jiaxin Shi, Michalis K. Titsias, and Andriy Mnih. Sparse orthogonal variational inference for Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 1932–1942, 2020.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, 18, 2005.

Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 567–574, 2009.

Michalis K. Titsias. New bounds for sparse variational Gaussian processes, 2025.

Gia-Lac Tran, Dimitrios Milios, Pietro Michiardi, and Maurizio Filippone. Sparse within sparse Gaussian processes using neighbor information. In *International Conference on Machine Learning*, pp. 10369–10378, 2021.

Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *International Conference on Machine Learning*, pp. 1775–1784, 2015.

Luhuan Wu, Geoff Pleiss, and John P Cunningham. Variational nearest neighbor Gaussian process. In *International Conference on Machine Learning*, pp. 24114–24130, 2022.

A Full derivation of the block-diagonal variational bound

We start with a general posterior approximation of the form:

$$q(f) = p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f},\boldsymbol{u})q(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u}) \tag{15}$$

$$q(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(f; \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\boldsymbol{u}, \mathbf{C}) \tag{16}$$

where we have not specified the form of the covariance matrix $\mathbf{C}$. The variational lower bound to the log marginal likelihood is

$$\mathcal{F}(q,\theta) = -\mathrm{KL}[q(\boldsymbol{u})||p(\boldsymbol{u})] - \int q(\boldsymbol{u})\mathrm{KL}[q(\boldsymbol{f}|\boldsymbol{u})||p(\boldsymbol{f}|\boldsymbol{u})] + \int q(\boldsymbol{u})q(\boldsymbol{f}|\boldsymbol{u})\log p(\boldsymbol{y}|\boldsymbol{f}) \tag{17}$$

Setting the gradient wrt $q(\boldsymbol{u})$ to zeros gives, $q(\boldsymbol{u}) \propto p(\boldsymbol{u})\exp[\int q(\boldsymbol{f}|\boldsymbol{u})\log p(\boldsymbol{y}|\boldsymbol{f})]$. In the regression case, $p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \sigma^2\mathbf{I}_N)$ and thus,

$$\int q(\boldsymbol{f}|\boldsymbol{u})\log p(\boldsymbol{y}|\boldsymbol{f}) = \int \mathcal{N}(\boldsymbol{f}; \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\boldsymbol{u}, \mathbf{C})\log \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \sigma^2\mathbf{I}_N) \tag{18}$$

$$= \log \mathcal{N}(\boldsymbol{y}; \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\boldsymbol{u}, \sigma^2\mathbf{I}_N) - \frac{1}{2\sigma^2}\mathrm{trace}(\mathbf{C}). \tag{19}$$

The middle term in the bound can be simplified to,

$$\int q(\boldsymbol{u})\mathrm{KL}[q(\boldsymbol{f}|\boldsymbol{u})||p(\boldsymbol{f}|\boldsymbol{u})] = \frac{1}{2}\mathrm{trace}(\mathbf{D_{ff}^{-1}}\mathbf{C}) + \frac{1}{2}\log|\mathbf{D_{ff}}| - \frac{1}{2}\log|\mathbf{C}| - \frac{N}{2}. \tag{20}$$

Substituting this and the optimal $q(\boldsymbol{u})$ back to the bound gives,

$$\mathcal{F} = \log \mathcal{N}(\boldsymbol{y}; \mathbf{0}, \mathbf{Q_{ff}} + \sigma^2\mathbf{I}_N) - \frac{1}{2}\mathrm{trace}[(\mathbf{D_{ff}^{-1}} + \sigma^{-2}\mathbf{I}_N)\mathbf{C}] - \frac{1}{2}\log|\mathbf{D_{ff}}| + \frac{1}{2}\log|\mathbf{C}| + \frac{N}{2}.$$

When $\mathbf{C} = \mathbf{D_{ff}^{1/2}}\mathbf{M}\mathbf{D_{ff}^{1/2}}$, the collapsed bound above becomes,

$$\mathcal{F}(\theta) = \log \mathcal{N}(\boldsymbol{y}; 0, \mathbf{Q_{ff}} + \sigma^2\mathbf{I}_N) - \frac{1}{2}\sum_b \left[\frac{1}{\sigma^2}\mathrm{trace}[\mathbf{m}_b\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}] + \mathrm{trace}[\mathbf{m}_b] - \log|\mathbf{m}_b| - N_b\right].$$

We can find the gradient of the bound wrt $\mathbf{m}_b$,

$$\frac{\partial}{\partial \mathbf{m}_b}\mathcal{F} = \frac{1}{2}\left[\sigma^{-2}\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b} + \mathbf{I}_b - \mathbf{m}_b^{-1}\right] \tag{21}$$

Setting this to zero gives $\mathbf{m}_b = (\mathbf{I}_b + \sigma^{-2}\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b})^{-1}$, and the resulting $\mathbf{m}$-collapsed bound:

$$\mathcal{F} = \log \mathcal{N}(\boldsymbol{y}; 0, \mathbf{Q_{ff}} + \sigma^2\mathbf{I}_N) - \frac{1}{2}\sum_b \log|\mathbf{I}_b + \sigma^{-2}\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}|. \tag{22}$$

When the block size is 1, the above bounds become the bounds presented in Titsias (2025); Bui et al. (2025).

We now consider a special case when we let all $\mathbf{m}_b$ matrices to be the same, $\mathbf{m}_b = \mathbf{m}$. The gradient wrt $\mathbf{m}$ in this case is,

$$\frac{\partial}{\partial \mathbf{m}}\mathcal{F} = \frac{1}{2}\sum_b \left[\sigma^{-2}\mathbf{D}_{\mathbf{f}_b\mathbf{f}_b} + \mathbf{I}_b - \mathbf{m}^{-1}\right]. \tag{23}$$

This leads to the optimal $\mathbf{m}$, $\mathbf{m} = (\mathbf{I}_b + B^{-1}\sigma^{-2}\sum_b \mathbf{D}_{\mathbf{f}_b\mathbf{f}_b})^{-1}$, and the corresponding $\mathbf{m}$-collapsed bound,

$$\mathcal{F} = \log \mathcal{N}(\boldsymbol{y}; 0, \mathbf{Q_{ff}} + \sigma^2\mathbf{I}_N) - \frac{B}{2}\log|\mathbf{I}_b + \frac{1}{B\sigma^2}\sum_b \mathbf{D}_{\mathbf{f}_b\mathbf{f}_b}|. \tag{24}$$

A special case is when the block size is only 1, we arrive at the spherical diagonal approximation $\mathbf{M} = m\mathbf{I}$ (Titsias, 2025; Artemev et al., 2021). Note that, since the log-determinant is a concave function on the cone of positive definite matrices, we can apply Jensen's inequality to show that the bound above (when all $\mathbf{m}$ blocks are the same) is less tight compared to the bound when all blocks are different.

B Power-EP posterior and approximate marginal likelihood

B.1 Power EP steps

Given a data set of N input-output pairs $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, we use M pseudo-points $\boldsymbol{y}$ at locations $\boldsymbol{z}$ to approximate the exact posterior. Power-EP posits the following approximation to the joint:

$$p(f, \boldsymbol{y}) = p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f},\boldsymbol{u})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})\prod_b p(\boldsymbol{y}_b|\boldsymbol{f}_b) \approx p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f},\boldsymbol{u})q(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})\prod_b t_b(\boldsymbol{u}) = q(f)$$

where we have partitioned the data into B disjoint blocks, b indexes blocks of data and $t_b(\boldsymbol{u})$ are the approximate factors. Crucially, we employ a *structured* conditional approximate posterior $q(\boldsymbol{f}|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{f}; \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\boldsymbol{u}, \mathbf{D_{ff}^{1/2}}\mathbf{M}\mathbf{D_{ff}^{1/2}})$. The Power-EP procedure with power α iteratively updates the factors $\{t_b\}_{b=1}^B$ as follows:

1. **Deletion step**: Compute the cavity distribution by removing a fraction α of one approximate factor:

$$q^{\backslash i}(f) \propto \frac{q(f)}{t_i^\alpha(\boldsymbol{u})} = p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f},\boldsymbol{u})q(\boldsymbol{f}|\boldsymbol{u})\frac{q(\boldsymbol{u})}{t_i^\alpha(\boldsymbol{u})} = p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f},\boldsymbol{u})q(\boldsymbol{f}|\boldsymbol{u})q^{\backslash i}(\boldsymbol{u}), \quad (25)$$

   where $q(\boldsymbol{u}) = p(\boldsymbol{u})\prod_b t_b(\boldsymbol{u})$ and $q^{\backslash i}(\boldsymbol{u}) = q(\boldsymbol{u})/t_i^\alpha(\boldsymbol{u})$

2. **Projection step**: First, compute the tilted distribution by incorporating a corresponding fraction of the true likelihood factor:

$$\tilde{p}(f) = q^{\backslash i}(f)p^\alpha(\boldsymbol{y}_i|\boldsymbol{f}_i) = p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f},\boldsymbol{u})q(\boldsymbol{f}|\boldsymbol{u})q^{\backslash i}(\boldsymbol{u})p^\alpha(\boldsymbol{y}_i|\boldsymbol{f}_i) \quad (26)$$

   Second, project the tilted distribution onto the new approximate posterior using KL divergence:

$$q(f) \leftarrow \arg\min_{q(f)} \mathrm{KL}[\tilde{p}(f)||q(f)] \quad (27)$$

   Due to the structure of the approximate posterior, this minimisation is achieved when the moments at the pseudo-inputs are matched: $\mathbb{E}_{\tilde{p}(f)}[\phi(\boldsymbol{u})] = \mathbb{E}_{q(f)}[\phi(\boldsymbol{u})]$, where $\phi(\boldsymbol{u}) = \{\boldsymbol{u}, \boldsymbol{u}\boldsymbol{u}^T\}$ are the sufficient statistics (Bui et al., 2017). In practice, this can be done by using the moment-matching shortcut involving the gradients of the log-normalising constant of the tilted distribution.

3. **Update step**: Compute the new fraction by dividing the new approximate posterior by the cavity:

$$t_{i,\text{new}}^\alpha(\boldsymbol{u}) = \frac{q(f)}{q\backslash i(f)} \quad (28)$$

   The factor then is updated using $t_i(\boldsymbol{u}) = t_{i,\text{new}}(\boldsymbol{u})$ or with damping, $t_i(\boldsymbol{u}) = t_{i,\text{old}}^{1-\alpha}(\boldsymbol{u}) \cdot t_{i,\text{new}}^\alpha(\boldsymbol{u})$.

B.2 Optimal factors

The factors are parameterised as follows,

$$t_b(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}; z_b, \mathbf{T}_{1,b}, \mathbf{T}_{2,b}) = z_b \exp(\boldsymbol{u}^T\mathbf{T}_{1,b} - \frac{1}{2}\boldsymbol{u}^T\mathbf{T}_{2,b}\boldsymbol{u}) \quad (29)$$

The posterior distribution over $\boldsymbol{u}$ is therefore $q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}; \mathbf{m}, \mathbf{S})$, where

$$\mathbf{S}^{-1} = \mathbf{K_{uu}^{-1}} + \sum_b \mathbf{T}_{2,b} \quad (30)$$

$$\mathbf{S}^{-1}\mathbf{m} = \sum_b \mathbf{T}_{1,b}. \quad (31)$$

Similarly, the cavity distribution over $\boldsymbol{u}$ is $q^{\setminus i}(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}; \mathbf{m}^{\setminus i}, \mathbf{S}^{\setminus i})$, where

$$\mathbf{S}^{\setminus i,-1} = \mathbf{K}_{\mathbf{uu}}^{-1} + \sum_{b \neq i} \mathbf{T}_{2,b} + (1-\alpha)\mathbf{T}_{2,i} = \mathbf{S}^{-1} - \alpha \mathbf{T}_{2,i} \tag{32}$$

$$\mathbf{S}^{\setminus i,-1}\mathbf{m}^{\setminus i} = \sum_{b \neq i} \mathbf{T}_{1,b} + (1-\alpha)\mathbf{T}_{1,i} = \mathbf{S}^{-1}\mathbf{m} - \alpha \mathbf{T}_{1,i}. \tag{33}$$

The moments of the tilted distribution (and the new posterior) can be computed efficiently using the following shortcuts,

$$\mathbf{m} = \mathbf{m}^{\setminus i} + \mathbf{V}^{\setminus i}_{\boldsymbol{u}\boldsymbol{f}_i} \frac{d \log \tilde{Z}_i}{d \mathbf{m}^{\setminus i}_{\boldsymbol{f}_i}}, \tag{34}$$

$$\mathbf{V} = \mathbf{V}^{\setminus i} + \mathbf{V}^{\setminus i}_{\boldsymbol{u}\boldsymbol{f}_i} \frac{d^2 \log \tilde{Z}_i}{d(\mathbf{m}^{\setminus i}_{\boldsymbol{f}_i})^2} \mathbf{V}^{\setminus i}_{\boldsymbol{f}_i\boldsymbol{u}} \tag{35}$$

where $\tilde{Z}_i = \int q^{\setminus i}(\boldsymbol{f}_i) p^{\alpha}(\boldsymbol{y}_i|\boldsymbol{f}_i) d\boldsymbol{f}_i$ is the normaliser of the tilted distribution.

At convergence, the optimal form of $\mathbf{T}_{2,b}$ is rank-N_b, $\mathbf{T}_{2,b} = \mathbf{w}_b \mathbf{v}_b^{-1} \mathbf{w}_b^T$, where $\mathbf{w}_b = \mathbf{V}^{\setminus b,-1}_{\boldsymbol{u}\boldsymbol{u}} \mathbf{V}^{\setminus b}_{\boldsymbol{u}\boldsymbol{b}} = \mathbf{K}^{-1}_{\mathbf{uu}} \mathbf{K}_{\mathbf{u}\mathbf{f}_b}$, $\mathbf{v}_b = -\boldsymbol{d}_2^{-1} - \mathbf{V}^{\setminus b}_{\boldsymbol{b}\boldsymbol{u}} \mathbf{V}^{\setminus b,-1}_{\boldsymbol{u}\boldsymbol{u}} \mathbf{V}^{\setminus b}_{\boldsymbol{u}\boldsymbol{b}}$, and $\boldsymbol{d}_2 = \frac{d^2 \log \tilde{Z}_b}{d(\boldsymbol{m}_{\boldsymbol{b}}^{\setminus b})^2}$.

In the regression case, at convergence, $t_b(\boldsymbol{u}) = \mathcal{N}(\mathbf{K}_{\mathbf{f}_b\mathbf{u}}\mathbf{K}^{-1}_{\mathbf{uu}}\boldsymbol{u}; \boldsymbol{y}_b, \alpha[\mathbf{D}^{1/2}_{\mathbf{ff}}\mathbf{M}\mathbf{D}^{1/2}_{\mathbf{ff}}]_{bb} + \sigma^2\mathbf{I}_b)$. We can check this by computing the contribution of an α fraction of the exact likelihood to the posterior $q(\boldsymbol{u})$,

$$\int q(\boldsymbol{f}_b|\boldsymbol{u}) p^{\alpha}(\boldsymbol{y}_b|\boldsymbol{f}_b) d\boldsymbol{f}_b = \int \mathcal{N}(\boldsymbol{f}_b; \mathbf{K}_{\mathbf{f}_b\mathbf{u}}\mathbf{K}^{-1}_{\mathbf{uu}}\boldsymbol{u}, [\mathbf{D}^{1/2}_{\mathbf{ff}}\mathbf{M}\mathbf{D}^{1/2}_{\mathbf{ff}}]_{bb} \mathcal{N}^{\alpha}(\boldsymbol{y}_b; \boldsymbol{f}_b, \sigma^2\mathbf{I}_b) d\boldsymbol{f}_b \tag{36}$$

$$\propto \mathcal{N}(\boldsymbol{y}_b; \mathbf{K}_{\mathbf{f}_b\mathbf{u}}\mathbf{K}^{-1}_{\mathbf{uu}}\boldsymbol{u}, [\mathbf{D}^{1/2}_{\mathbf{ff}}\mathbf{M}\mathbf{D}^{1/2}_{\mathbf{ff}}]_{bb} + \sigma^2\mathbf{I}_b/\alpha), \tag{37}$$

which is exactly an α-fraction of the optimal factor listed above.

B.3 Power-EP approximate marginal likelihood

After convergence, Power EP provides an approximate log marginal likelihood:

$$\log Z_{\text{PEP}} = \log \int p(f_{\neq \boldsymbol{f},\boldsymbol{u}}|\boldsymbol{f},\boldsymbol{u}) q(\boldsymbol{f}|\boldsymbol{u}) p(\boldsymbol{u}) \prod_b t_b(\boldsymbol{u}) df \tag{38}$$

$$= \mathcal{G}(q(\boldsymbol{u})) - \mathcal{G}(p(\boldsymbol{u})) + \frac{1}{\alpha}\sum_b \left[\log \tilde{Z}_b + \mathcal{G}(q^{\setminus b}(\boldsymbol{u})) - \mathcal{G}(q(\boldsymbol{u}))\right] + \frac{1}{\alpha}\log \tilde{Z}_q, \tag{39}$$

where

$$\log \tilde{Z}_b = \log \int q(\boldsymbol{f}_b|\boldsymbol{u}) q^{\setminus b}(\boldsymbol{u}) p^{\alpha}(\boldsymbol{y}_b|\boldsymbol{f}_b) d\boldsymbol{f}_b d\boldsymbol{u} \tag{40}$$

$$\log \tilde{Z}_q = \log \int q^{1-\alpha}(\boldsymbol{f}_b|\boldsymbol{u}) q^{\setminus b}(\boldsymbol{u}) p^{\alpha}(\boldsymbol{f}_b|\boldsymbol{u}) d\boldsymbol{f}_b d\boldsymbol{u} \tag{41}$$

$$\mathcal{G}(q(\boldsymbol{u})) = \frac{M}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{V}| + \frac{1}{2}\mathbf{m}^{\intercal}\mathbf{V}^{-1}\mathbf{m} \tag{42}$$

$$\mathcal{G}(p(\boldsymbol{u})) = \frac{M}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{K}_{\mathbf{uu}}| \tag{43}$$

$$\mathcal{G}(q^{\setminus b}(\boldsymbol{u})) = \frac{M}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{V}^{\setminus b}| + \frac{1}{2}\mathbf{m}^{\setminus b,\intercal}\mathbf{V}^{\setminus b,-1}\mathbf{m}^{\setminus b} \tag{44}$$

In the regression case, following closely the steps in (Bui et al., 2017), we can derive the closed-form approximate log marginal likelihood

$$\log Z_{\text{PEP}} = \log \mathcal{N}(\boldsymbol{y}; \mathbf{0}, \mathbf{Q}_{\mathbf{ff}} + \alpha \text{blkdiag}(\{[\mathbf{D}^{1/2}_{\mathbf{ff}}\mathbf{M}\mathbf{D}^{1/2}_{\mathbf{ff}}]_{bb}\}_{b=1}^{B}) + \sigma^2\mathbf{I}_N)$$
$$+ \sum_b \left[-\frac{1-\alpha}{2\alpha}\log\left|\mathbf{I}_b + \alpha\frac{[\mathbf{D}^{1/2}_{\mathbf{ff}}\mathbf{M}\mathbf{D}^{1/2}_{\mathbf{ff}}]_{bb}}{\sigma^2}\right| - \frac{1}{2\alpha}\log|\mathbf{I}_b + \alpha(\mathbf{m}_b - \mathbf{I}_b)| + \frac{1}{2}\log|\mathbf{m}_b|\right] \tag{45}$$

B.4 Extension to classification

Instead of working with individual factors, we can use the *stochastic* Power-EP parameterisation (Li et al., 2015) , i.e., assuming contributions from all blocks to the posterior are equal $t_b(\boldsymbol{u}) = t(\boldsymbol{u})$. In addition, instead of running stochastic Power-EP iteration, we can directly work with $q(\boldsymbol{u}) \propto p(\boldsymbol{u})t^B(\boldsymbol{u})$ and optimise the Power-EP energy, also known as the black-box α-divergence objective (Hernández-Lobato et al., 2016). We will explore this direction in future work.

C Additional experimental results

C.1 Experimental set-up

In addition to the details in the main text, we provide additional information here. For all experiments involving the block-diagonal matrix M, we randomly partitioned the training data into B blocks. In the Snelson, kin40k, and Power-EP experiments, we optimised the collapsed bound using the L-BFGS optimiser. In the block-diagonal experiments with medium-scale datasets, we used the Adam optimiser with a learning rate of 0.005. To initialise the inducing point locations, we picked M random training inputs in the Snelson experiment, and employed k-means clustering for all other experiments. For the later datasets, we used the median distance between the data points to initialise the lengthscales and set the initial observation noise variance to 0.1.

C.2 Snelson dataset

We compared several sparse variational GP variants, including SGPR, T-SGPR, and BT-SGPR, with $M = 10$ to exact GP regression, and the objective and hyperparameters collected during optimisation are included in fig. 6. We note that, by using structured approximations, (i) the variational bound that is provably tighter for fixed hyperparameters indeed is tighter in practice, and (ii) the observation noise variance (the kernel variance) is smaller (larger).

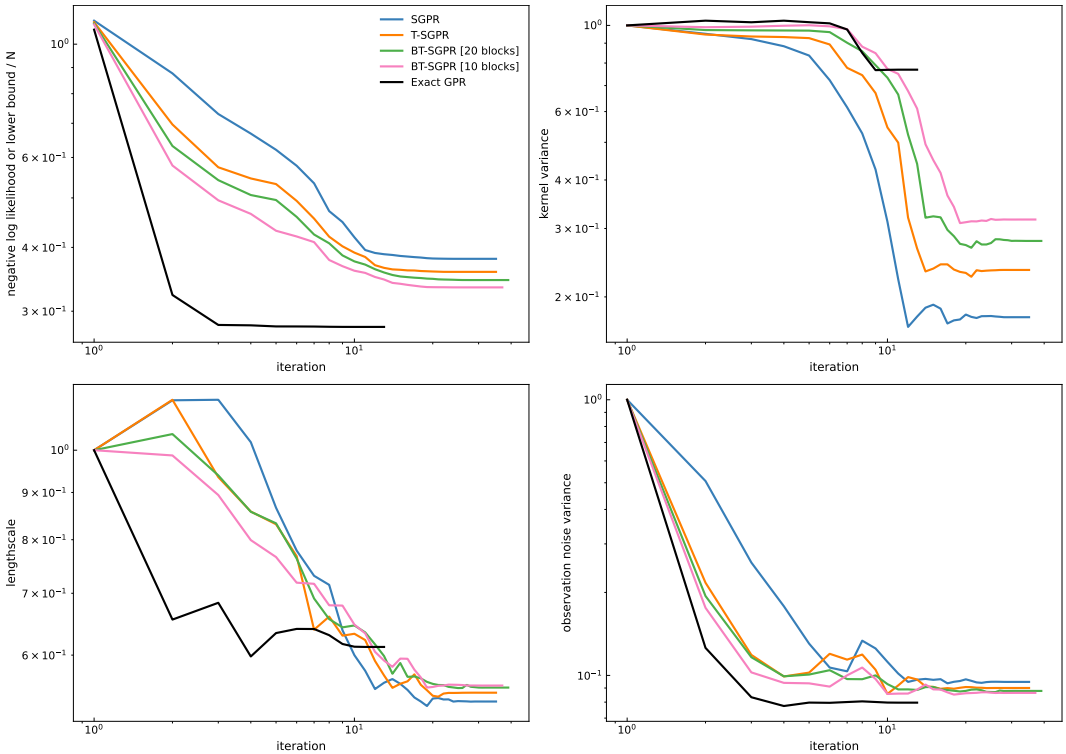

Figure 6: Objectives and hyperparameters provided by sparse variational and exact methods.

C.3 KIN40K hyperparameters

We include the full results, including the standard errors, for the KIN40K experiment in table 2.

C.4 Block-diagonal structured variational approximation

In addition to the predictive performance metrics in the main text, we also recorded the estimated hyperparameters when using the new structured variational approximations. These results are included in fig. 7, and agree with observations in smaller datasets (Snelson and kin40k): kernel variance and observation noise variance tend to be larger and smaller, respectively, when using the improved bounds.

C.5 Power-EP

We include the full results for all five datasets considered in the main text in fig. 8.

Table 2: Exact/approximate marginal likelihoods, predictive performance, and lengthscales given by various methods on 5,000 samples from the KIN40K dataset, including the standard errors across three repeats

	M = 256					M = 512				
Method	Obj.	RMSE	LL	σ	lengthscales	Obj.	RMSE	LL	σ	lengthscales
Exact	-0.656 $\pm$ 0.005	0.117 $\pm$ 0.000	0.796 $\pm$ 0.004	0.001 $\pm$ 0.000		-0.656 $\pm$ 0.005	0.117 $\pm$ 0.000	0.796 $\pm$ 0.004	0.001 $\pm$ 0.000	
SGPR	0.883 $\pm$ 0.006	0.256 $\pm$ 0.001	-0.136 $\pm$ 0.002	0.299 $\pm$ 0.001		0.660 $\pm$ 0.006	0.215 $\pm$ 0.001	0.022 $\pm$ 0.002	0.252 $\pm$ 0.001	
T-SGPR	0.779 $\pm$ 0.006	0.223 $\pm$ 0.001	-0.057 $\pm$ 0.002	0.259 $\pm$ 0.001		0.515 $\pm$ 0.005	0.184 $\pm$ 0.000	0.115 $\pm$ 0.001	0.206 $\pm$ 0.001	
BT-SGPR $B = 50$	0.752 $\pm$ 0.006	0.217 $\pm$ 0.000	-0.045 $\pm$ 0.002	0.250 $\pm$ 0.001		0.499 $\pm$ 0.006	0.181 $\pm$ 0.000	0.120 $\pm$ 0.002	0.201 $\pm$ 0.001	
BT-SGPR $B = 10$	0.659 $\pm$ 0.006	0.200 $\pm$ 0.001	-0.032 $\pm$ 0.002	0.227 $\pm$ 0.001		0.437 $\pm$ 0.006	0.173 $\pm$ 0.000	0.133 $\pm$ 0.002	0.186 $\pm$ 0.001	
PEP $\alpha = 0.5$	0.661 $\pm$ 0.006	0.235 $\pm$ 0.001	-0.015 $\pm$ 0.002	0.225 $\pm$ 0.001		0.422 $\pm$ 0.005	0.200 $\pm$ 0.001	0.140 $\pm$ 0.002	0.187 $\pm$ 0.000	
T-PEP $\alpha = 0.5$	0.480 $\pm$ 0.005	0.200 $\pm$ 0.000	0.024 $\pm$ 0.002	0.182 $\pm$ 0.001		0.184 $\pm$ 0.006	0.164 $\pm$ 0.000	0.190 $\pm$ 0.003	0.138 $\pm$ 0.001	

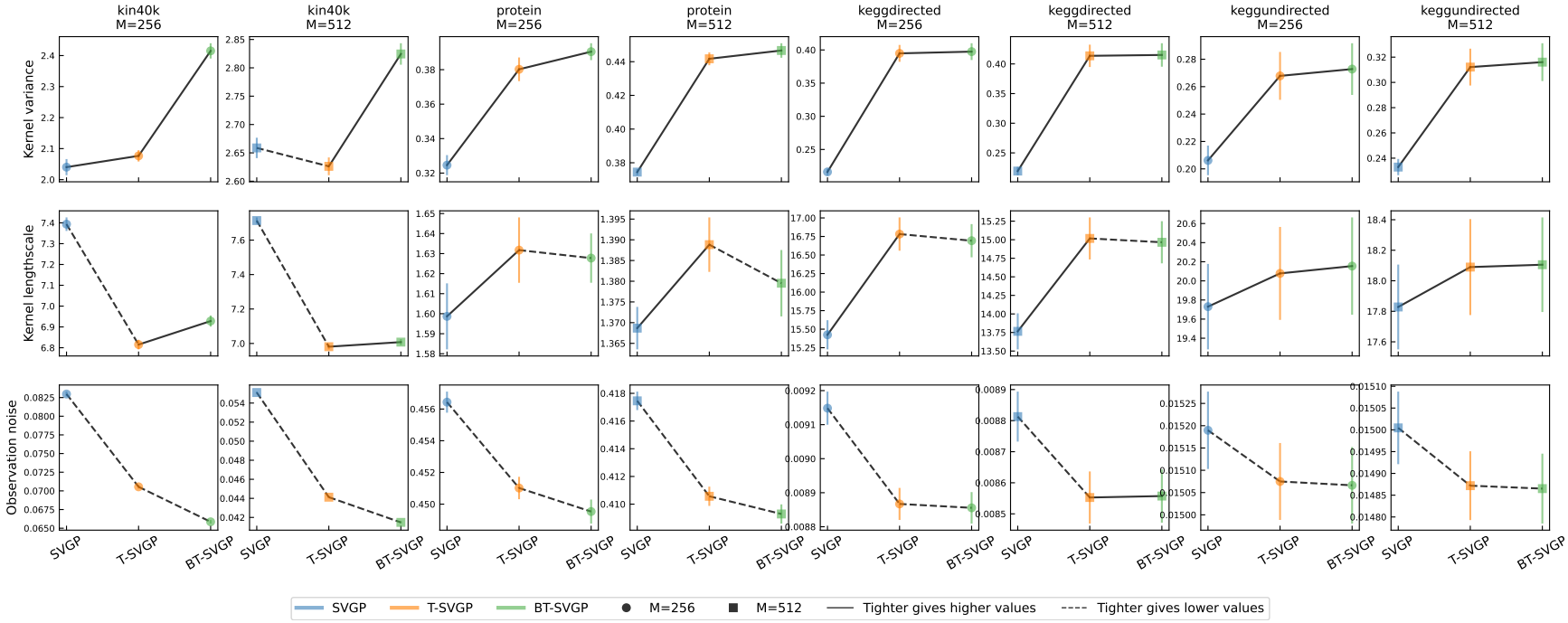

Figure 7: Estimated hyperparameters by using SVGP, T-SVGP and BT-SVGP on four UCI datasets.

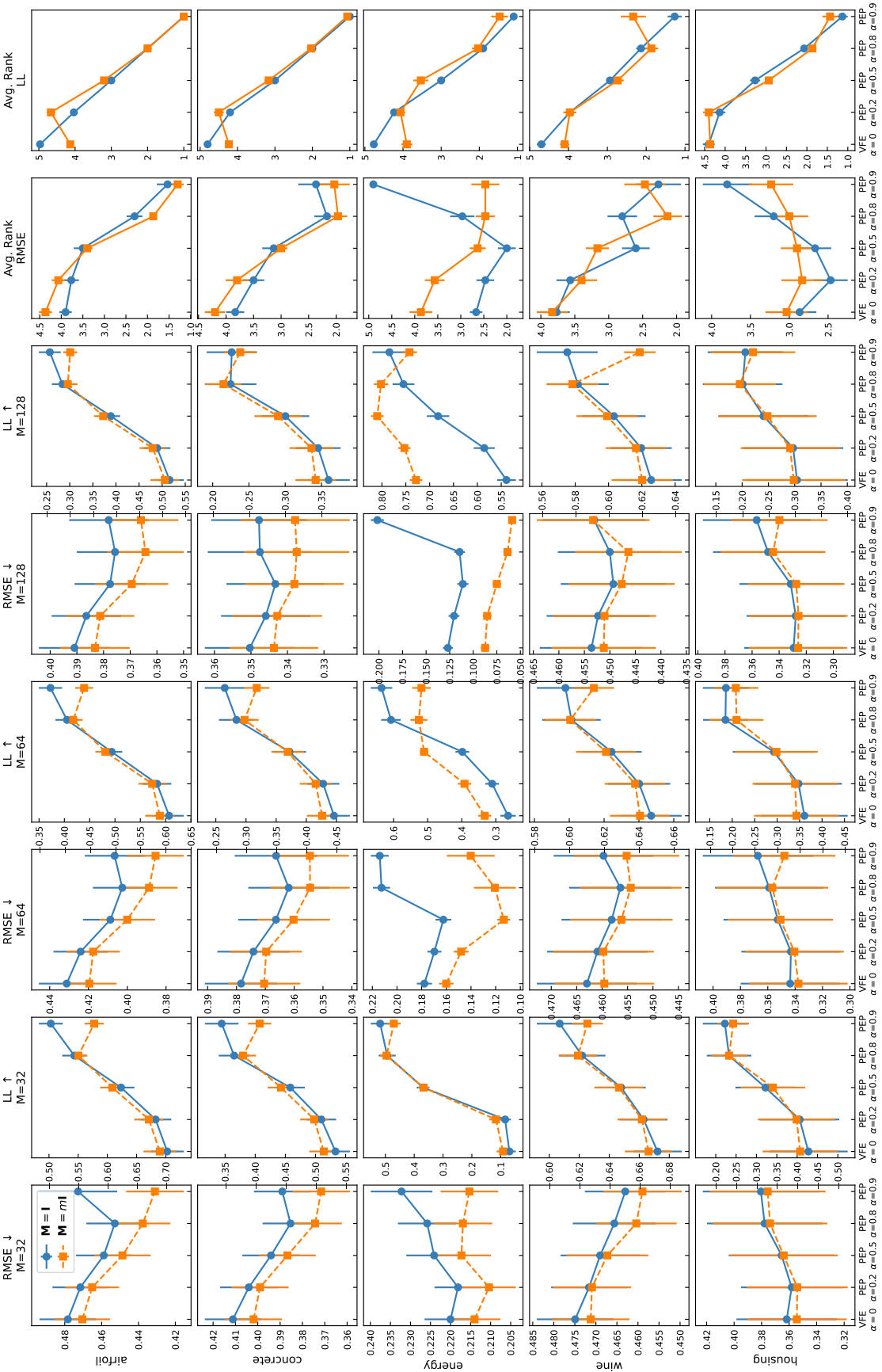

Figure 8: A comparison between $\mathbf{M} = \mathbf{I}$ and $\mathbf{M} = m\mathbf{I}$ for Power Expectation Propagation.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims: a new block-diagonal structured variational approximation and an improved Power-EP framework for sparse GPs. These are presented in sections 3 and 4 and experiments are provided in section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We provided several limitations (and potential future directions) at the end of section 7.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The posterior approximation and bounds are provided in the main text, with the full derivations in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided all necessary information (number of epochs, batch size, block size, number of splits, source of data) to repeat the experiments. Code will be provided upon acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We used publicly available datasets, typically used in the sparse GP literature. The implementation was built on GPflow, and released here `https://github.com/thangbui/tighter_sparse_gp`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided sufficient information to replicate.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have reported all error bars in the main text and the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have added a comment about the compute workers in section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We, at all times, follow all general research ethics and the NeurIPS code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The proposed methods will benefit probabilistic ML practitioners. We do not expect any near-term societal impacts of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We did not use any models or datasets that require safeguards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We used open-source data sets and software packages.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing or human subjects involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: We did not use LLMs for this submission.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.