# Variational Sparse Inverse Cholesky Approximation for Latent Gaussian Processes via Double Kullback-Leibler Minimization

**Jian Cao** [*1] **Myeongjong Kang** [*2] **Felix Jimenez** [2] **Huiyan Sang** [2] **Florian Schäfer** [3] **Matthias Katzfuss** [1]

## Abstract

To achieve scalable and accurate inference for latent Gaussian processes, we propose a variational approximation based on a family of Gaussian distributions whose covariance matrices have sparse inverse Cholesky (SIC) factors. We combine this variational approximation of the posterior with a similar and efficient SIC-restricted Kullback-Leibler-optimal approximation of the prior. We then focus on a particular SIC ordering and nearest-neighbor-based sparsity pattern resulting in highly accurate prior and posterior approximations. For this setting, our variational approximation can be computed via stochastic gradient descent in polylogarithmic time per iteration. We provide numerical comparisons showing that the proposed double-Kullback-Leibler-optimal Gaussian-process approximation (DKLGP) can sometimes be vastly more accurate for stationary kernels than alternative approaches such as inducing-point and mean-field approximations at similar computational complexity.

## 1. Introduction

Gaussian process (GP) priors are popular models for unknown functions in a variety of settings, including geostatistics (e.g., Stein, 1999; Banerjee et al., 2004; Cressie & Wikle, 2011), computer model emulation (e.g., Sacks et al., 1989; Kennedy & O'Hagan, 2001; Gramacy, 2020), and machine learning (e.g., Rasmussen & Williams, 2006; Deisenroth, 2010). Latent GP (LGP) models, such as generalized GPs, assume a Gaussian or non-Gaussian distribution for the data conditional on a GP (e.g., Diggle et al., 1998; Chan & Dong, 2011). LGPs extend GPs to a large class of settings, including noisy, categorical, and count data. However, LGP inference is generally analytically intractable and hence requires approximations. In addition, direct GP inference is prohibitive for large datasets due to cubic scaling in the data size. There are two main challenges for (L)GPs in many applications: One is to specify or learn a suitable kernel for the GP, and the other is carrying out fast inference for a given kernel. In this paper, we make no contributions to the former and instead focus on the latter challenge: We assume that a parametric kernel form is given and propose an efficient approximation method for LGP inference via structured variational learning.

Many approaches to scaling GPs to large datasets were reviewed in Heaton et al. (2019) and Liu et al. (2020), including low-rank approaches with a small number of pseudo points that are popular in machine learning. Such low-rank GP approximations have been combined with variational inference for GPs (e.g., Titsias, 2009; Hensman et al., 2013) and LGPs (e.g., Hensman et al., 2015; Leibfried et al., 2020).

A highly promising approach to achieve GP scalability is given by nearest-neighbor Vecchia approximations from spatial statistics (e.g., Vecchia, 1988; Stein et al., 2004; Datta et al., 2016; Katzfuss & Guinness, 2021), which are optimal with respect to forward Kullback-Leibler (KL) divergence under the restriction of sparse inverse Cholesky (SIC) factors of the covariance matrix (Schäfer et al., 2021a). Such SIC approximations have several attractive properties (e.g., as reviewed by Katzfuss et al., 2022). They result in a valid joint density function given by the product of univariate conditional Gaussians, each of which can be independently computed in cubic complexity in the number of neighbors. This allows straightforward mini-batch subsampling with unbiased gradient estimators (Cao et al., 2022). For the ordering and sparsity pattern used here, the number of neighbors needs to grow only polylogarithmically with the data size to achieve $\epsilon$-accurate approximations for Matérn-type kernels up to boundary effects (Schäfer et al., 2021a) due to the screening effect (Stein, 2011). Many existing GP approximations, including low-rank and partially-independent conditional approaches, can be viewed as special cases of SIC approximations corresponding to particular orderings and sparsity patterns (Katzfuss & Guinness, 2021). SIC

[*]Equal contribution [1]Department of Statistics and Institute of Data Science, Texas A&M University, College Station, TX, USA [2]Department of Statistics, Texas A&M University, College Station, TX, USA [3]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: Matthias Katzfuss <katzfuss@gmail.com>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

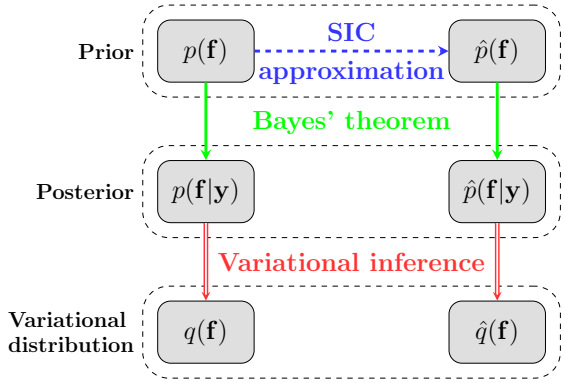



*Figure 1.* Double KL minimization for approximating the posterior distribution of a latent Gaussian field $\mathbf{f}$ given data $\mathbf{y}$: Based on a forward-KL-optimal SIC approximation $\hat{p}(\mathbf{f})$ of the prior, we obtain an SIC-restricted reverse-KL-optimal variational approximation $\hat{q}(\mathbf{f})$ to the posterior.

approximation using our ordering and sparsity pattern does not exhibit the same limitations as low-rank approximations (Stein, 2014) and can hence be significantly more accurate for non-latent (i.e., directly observed) GPs (Cao et al., 2022).

SIC approximations of LGPs are more challenging. For LGPs with Gaussian noise, applying SIC approximations to the noisy responses reduces accuracy, and SIC approximations of the latent field may not be scalable (e.g., Katzfuss & Guinness, 2021). Existing approaches addressing this challenge (Datta et al., 2016; Katzfuss & Guinness, 2021; Schäfer et al., 2021a; Geoga & Stein, 2022) do not consider estimation using stochastic gradient descent (SGD). For non-Gaussian LGPs, Laplace SIC approximations (Zilber & Katzfuss, 2021) are straightforward but can be inaccurate. Liu & Liu (2019) combined an SIC-type approximation to the prior with variational inference based on a variational family of Gaussians with a sparse Cholesky factor of the covariance matrix, but we are not aware of results guaranteeing that the covariance-Cholesky factor exhibits (approximate) sparsity under random ordering. Wu et al. (2022) combined SIC-type approximations of LGPs with mean-field variational inference, but the latter may be inaccurate when there are strong correlations in the GP posterior (MacKay, 1992).

To achieve scalable and accurate inference for LGPs, we propose a variational family of SIC Gaussian distributions and combine it with a SIC approximation to the GP prior (see Figure 1). Our approach is double-KL-optimal in the sense that variational approximation is reverse-KL-optimal for a given log normalizer (i.e., evidence) and our prior SIC approximation, which is available in closed form, is forward-KL-optimal for a given sparsity pattern (Schäfer et al., 2021a). Within our double-Kullback-Leibler-optimal Gaussian-process framework (DKLGP), we then focus on a particular ordering and nearest-neighbor-based sparsity pattern resulting in highly accurate prior and posterior approximations. We adopt a novel computational trick based on the concept of reduced ancestor sets for achieving efficient and scalable LGP inference. For this setting, our variational approximation can be computed via SGD in polylogarithmic time per iteration. While inducing-point methods assume that unobserved points depend on data only through inducing points (e.g., Frigola et al., 2014; Hensman et al., 2015), our method allows fast and accurate KL-optimal prediction based on the screening effect. Our numerical comparisons show that DKLGP can be vastly more accurate than state-of-the-art alternatives such as inducing-point and mean-field approximations at a similar computational complexity.

## 2. Methodology

### 2.1. Model

Assume we have a vector $\mathbf{y} = (y_1, \dots, y_n)^\top$ of noisy observations of a latent GP $f(\cdot) \sim \mathcal{GP}(\mu, K)$ at inputs $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, such that $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^n p(y_i|f_i)$, where

$$\mathbf{f} = (f_1, \dots, f_n)^\top \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K}) \tag{1}$$

with $\boldsymbol{\mu}_i = \mu(\mathbf{x}_i)$ and $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. Throughout, we view the inputs $\mathbf{x}_i$ as fixed (i.e., non-random) and hence do not explicitly condition on them.

Unless $\mathbf{y}|\mathbf{f}$ follows a Gaussian distribution, inference (such as computing the posterior $p(\mathbf{f}|\mathbf{y})$) generally cannot be carried out in closed form. In addition, even for Gaussian likelihoods, direct inference scales as $\mathcal{O}(n^3)$ and is thus computationally infeasible for large $n$. To address these challenges, we propose an approximation based on double KL minimization.

### 2.2. Variational Sparse Inverse Cholesky Approximation

Consider a lower-triangular sparsity pattern $\mathcal{S}^q \subset \{1, \dots, n\}^2$, with $\{(i,i) : i = 1, \dots, n\} \subset \mathcal{S}^q$ and such that $i \geq j$ for all $(i,j) \in \mathcal{S}^q$. Our preferred choice of $\mathcal{S}^q$ will be discussed in Section 2.5, but typically we will have $(i,j) \in \mathcal{S}^q$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close." Corresponding to $\mathcal{S}^q$, define the family of distributions $\mathcal{Q} = \{\mathcal{N}_n(\boldsymbol{\nu}, (\mathbf{V}\mathbf{V}^\top)^{-1}) : \boldsymbol{\nu} \in \mathbb{R}^n, \mathbf{V} \in \mathbb{R}^{n \times n}, \mathbf{V} \in \mathcal{S}^q\}$, where we write $\mathbf{V} \in \mathcal{S}^q$ if $(i,j) \in \mathcal{S}^q$ for all $\mathbf{V}_{ij} \neq 0$. It is straightforward to show that any $q \in \mathcal{Q}$ can be represented in ordered conditional form as $q(\mathbf{f}) = \prod_{i=1}^n q(f_i|\mathbf{f}_{s_i^q})$, where $s_i^q = \{j > i : (j,i) \in \mathcal{S}^q\}$ for $i = 1, \dots, n-1$ and $s_n^q = \emptyset$.

We approximate the posterior $p(\mathbf{f}|\mathbf{y})$ by the closest distribution in $\mathcal{Q}$ in terms of reverse KL divergence:

$$\hat{q}(\mathbf{f}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q(\mathbf{f}) \middle\| p(\mathbf{f}|\mathbf{y})\right).$$

We have $\mathrm{KL}(q(\mathbf{f}) \| p(\mathbf{f}|\mathbf{y})) = \log p(\mathbf{y}) - \mathrm{ELBO}(q)$, where

$p(\mathbf{y})$ does not depend on $q$, and so $\hat{q}$ satisfies

$$\hat{q}(\mathbf{f}) = \arg\max_{q \in \mathcal{Q}} \mathrm{ELBO}(q). \tag{2}$$

**Proposition 2.1.** *The ELBO in* (2) *can be written up to an additive constant of* $n/2$ *as*

$$\mathrm{ELBO}(q) = \sum_{i=1}^{n} \big( \mathbb{E}_q \log p(y_i|f_i) - ((\boldsymbol{\nu} - \boldsymbol{\mu})^\top \mathbf{L}_{:,i})^2/2 + \log(\mathbf{V}_{ii}^{-1}\mathbf{L}_{ii}) - \|\mathbf{V}^{-1}\mathbf{L}_{:,i}\|^2/2 \big), \tag{3}$$

*where* $\mathbf{L}$ *is the inverse Cholesky factor of* $\mathbf{K}$ *such that* $\mathbf{K}^{-1} = \mathbf{L}\mathbf{L}^\top$*, and* $\mathbf{L}_{:,i}$ *denotes its* $i$*th column.*

All proofs can be found in Appendix C.

### 2.3. Approximating the Prior via a Second KL Minimization

Even for a sparse $\mathbf{V}$, computing the ELBO in (3) is prohibitively expensive for large $n$, because computing $\mathbf{L}$ (or any of its columns) from $\mathbf{K}$ generally requires $\mathcal{O}(n^3)$ time. To avoid this, we replace the prior $p(\mathbf{f})$ defined in (1) by a Gaussian distribution that minimizes a second KL divergence under an SIC constraint.

Specifically, consider a second lower-triangular sparsity pattern $\mathcal{S}^p \subset \{1, \dots, n\}^2$, which may be the same as $\mathcal{S}^q$. We define the corresponding set of distributions $\mathcal{P} = \{\mathcal{N}_n(\tilde{\boldsymbol{\mu}}, (\tilde{\mathbf{L}}\tilde{\mathbf{L}}^\top)^{-1}) : \tilde{\boldsymbol{\mu}} \in \mathbb{R}^n, \tilde{\mathbf{L}} \in \mathbb{R}^{n \times n}, \tilde{\mathbf{L}} \in \mathcal{S}^p\}$. We approximate the prior $p(\mathbf{f})$ by the closest approximation in $\mathcal{P}$ in terms of forward KL divergence:

$$\hat{p}(\mathbf{f}) = \arg\min_{\tilde{p} \in \mathcal{P}} \mathrm{KL}\left(p(\mathbf{f}) \middle\| \tilde{p}(\mathbf{f})\right). \tag{4}$$

By a slight extension of Schäfer et al. (2021a, Thm. 2.1), we can show that this optimization problem has an efficient closed-form solution.

**Proposition 2.2.** *The solution to* (4) *is* $\hat{p}(\mathbf{f}) = \mathcal{N}_n(\mathbf{f}|\boldsymbol{\mu}, (\hat{\mathbf{L}}\hat{\mathbf{L}}^\top)^{-1})$*, where the nonzero entries of the* $i$*th column of* $\hat{\mathbf{L}}$ *can be computed in* $\mathcal{O}(|\mathcal{S}_i^p|^3)$ *time as*

$$\hat{\mathbf{L}}_{\mathcal{S}_i^p, i} = \mathbf{b}_i(\mathbf{b}_{i,1})^{-1/2}, \quad \textit{with } \mathbf{b}_i = \mathbf{K}^{-1}_{\mathcal{S}_i^p, \mathcal{S}_i^p}\mathbf{e}_1, \tag{5}$$

*and* $\mathcal{S}_i^p = \{j : (j, i) \in \mathcal{S}^p\}$ *is an ordered set with elements in increasing order (i.e., the first element is* $i$*).*

Throughout, we denote by $\mathbf{e}_i$ a vector whose $i$th entry is one and all others are zero, and we index matrices before inverting so that $\mathbf{K}^{-1}_{\mathcal{S}_i^p, \mathcal{S}_i^p} := (\mathbf{K}_{\mathcal{S}_i^p, \mathcal{S}_i^p})^{-1}$.

The approximation in Proposition 2.2 is equivalent to an ordered conditional approximation (Vecchia, 1988) of the prior density $p(\mathbf{f}) = \prod_{i=1}^n p(f_i|\mathbf{f}_{(i+1):n})$ by:

$$\hat{p}(\mathbf{f}) = \textstyle\prod_{i=1}^n p(f_i|\mathbf{f}_{s_i^p}) = \prod_{i=1}^n \mathcal{N}(f_i|\eta_i, \sigma_i^2),$$

where $\eta_i = \boldsymbol{\mu}_i - \hat{\mathbf{L}}^\top_{s_i^p, i}(\mathbf{f}_{s_i^p} - \boldsymbol{\mu}_{s_i^p})/\hat{\mathbf{L}}_{i,i}$ and $\sigma_i^2 = \hat{\mathbf{L}}_{i,i}^{-2}$, with $s_i^p = \mathcal{S}_i^p \setminus \{i\}$.

### 2.4. Computing the ELBO based on Ancestor Sets

Plugging $\hat{p}(\mathbf{f})$ into (2), the ELBO in (3) becomes

$$\mathrm{ELBO}(q) = \sum_{i=1}^{n} \big( \mathbb{E}_q \log p(y_i|f_i) - ((\boldsymbol{\nu} - \boldsymbol{\mu})^\top \hat{\mathbf{L}}_{:,i})^2/2 + \log(\mathbf{V}_{ii}^{-1}\hat{\mathbf{L}}_{ii}) - \|\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}\|^2/2 \big), \tag{6}$$

with the $i$th summand depending on $\hat{\mathbf{L}}$ only via its $i$th column $\hat{\mathbf{L}}_{:,i}$, whose nonzero entries can be computed in $\mathcal{O}(|\mathcal{S}_i^p|^3)$ time using (5).

We need to compute $\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}$ and $\mathbf{V}^{-1}\mathbf{e}_i$, the latter of which appears in $\mathbb{E}_q \log p(y_i|f_i)$ (see Section 2.6). The nonzero entry of $\mathbf{e}_i$ (i.e., $\{i\}$) is a subset of the nonzero entries of $\hat{\mathbf{L}}_{:,i}$ (i.e., $\mathcal{S}_i^p$), and hence we focus our discussion on computing $\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}$. Solving this sparse triangular system in principle requires $\mathcal{O}(|\mathcal{S}^q|)$ time.

However, it is possible to speed up computation by omitting rows and columns of $\mathbf{V}$ that do not correspond to the ancestor set $\mathcal{A}_i$ of $\mathcal{S}_i^p$ with respect to $\mathcal{S}^q$, which is defined as $\mathcal{A}_i = \big\{j \geq i :$ there exists a path $\mathcal{L} = \{(j, l_1), (l_1, l_2), \dots, (l_{a-1}, l_a), (l_a, l)\} \subset \mathcal{S}^q$ for some $l \in \mathcal{S}_i^p\big\}$. Ancestor sets are properties of the directed acyclic graphs that can be used to represent our triangular sparsity structures, as illustrated in Appendix B.

**Proposition 2.3.** $(\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i})_j = 0$ *for all* $j \notin \mathcal{A}_i$.

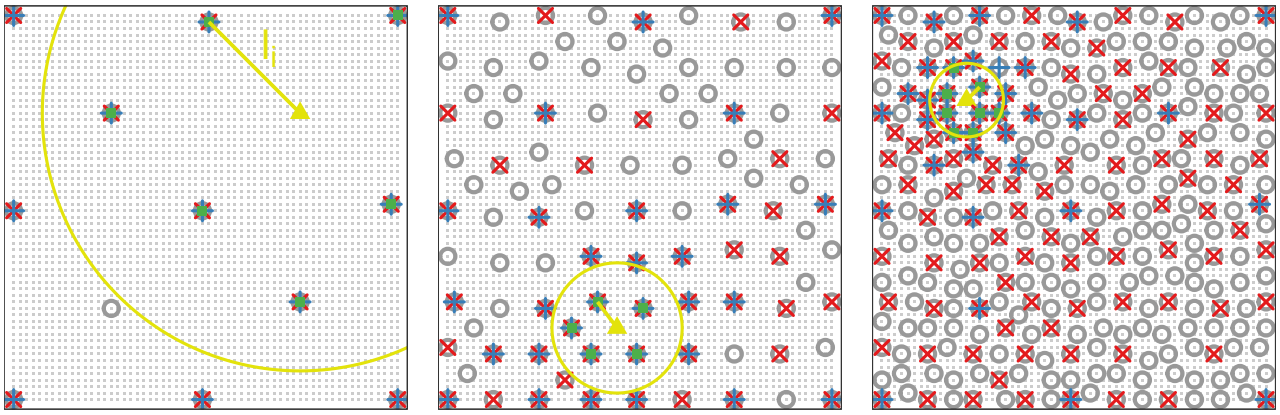
Thus, we have

$$\|\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}\| = \|\mathbf{V}^{-1}_{\mathcal{A}_i, \mathcal{A}_i}\hat{\mathbf{L}}_{\mathcal{A}_i, i}\|, \tag{7}$$

where $\mathbf{V}^{-1}_{\mathcal{A}_i, \mathcal{A}_i}\hat{\mathbf{L}}_{\mathcal{A}_i, i}$ can be computed in $\mathcal{O}(|\mathcal{A}_i||\mathcal{S}_i^q|)$ time.

### 2.5. Maximin Ordering and Nearest-neighbor Sparsity

Schäfer et al. (2021a) proposed a sparsity pattern $\mathcal{S}$ based on reverse-maximum-minimum-distance (r-maximin) ordering (see Figure 2 for an illustration). R-maximin ordering picks the last index $i_n$ arbitrarily (often in the center of the input domain), and then the previous indices are sequentially selected for $k = n-1, n-2, \dots, 1$ as $i_k = \arg\max_{i \notin \mathcal{I}_k} \min_{j \in \mathcal{I}_k} \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathcal{I}_k = \{i_{k+1}, \dots, i_n\}$. Throughout, we assume that our indexing follows r-maximin ordering (e.g., $f_k = f_{i_k}$). We can then define the sparsity pattern by $\mathcal{S}_i = \{j \geq i : \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j) \leq \rho\ell_i\}$, for some fixed $\rho \geq 1$, where $\ell_i = \min_{j>i} \mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)$. We can compute $\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)$ as Euclidean distance between the inputs, potentially in a transformed input space (see Section 2.6 for more details). The conditioning sets are all of approximately size $|\mathcal{S}_i| = \mathcal{O}(\rho^d) \approx m = |\mathcal{S}|/n$ under mild assumptions on the regularity of the inputs. Schäfer et al. (2021a) proved that an $\epsilon$-accurate approximation of the prior can be obtained using $\mathcal{S}^p = \mathcal{S}$ with $\rho = \mathcal{O}(\log(n/\epsilon))$ for



(a) $i = n - 12$ (b) $i = n - 100$ (c) $i = n - 289$

*Figure 2.* Reverse maximin ordering on a grid (small gray dots) of size $n = 60 \times 60 = 3{,}600$ on a square. For three different indices $i$, we show the $i$th ordered input (▲), the subsequently ordered $n - i$ inputs (○), the distance $\ell_i$ to the nearest neighbor (—), the neighboring subsequent inputs $\mathcal{S}_i$ (■) within a (yellow) circle of radius $\rho\ell_i$ (here, $\rho = 2$), the reduced ancestors $\tilde{\mathcal{A}}_i$ (+), and the ancestors $\mathcal{A}_i$ (×).

kernels $K$ that are Green's functions of elliptic boundary-value problems (similar to Matérn kernels up to boundary effects) and demonstrated high numerical accuracy of the posterior using $\mathcal{S}^q = \mathcal{S}$ for Gaussian likelihoods. For non-Gaussian likelihoods, this implies highly accurate approximations to the posterior when a second-order Taylor expansion can adequately approximate the posterior.

While this means that our DKLGP can achieve high accuracy by choosing $\mathcal{S}^p = \mathcal{S}^q = \mathcal{S}$, the resulting ancestor sets can grow roughly linearly with $n$ (e.g., see Figure 3a). Hence, even evaluating the ELBO based on the ancestor sets would often be prohibitively expensive for large $n$. However, it is possible to ignore most ancestors in (7) and only incur a small approximation error. Specifically, consider reduced ancestor sets $\tilde{\mathcal{A}}_i = \{j \geq i : \operatorname{dist}(\mathbf{x}_i, \mathbf{x}_j) \leq \rho\ell_j\}$, where the last subscript is now a $j$, not an $i$. As illustrated in Figure 2, we have $\mathcal{S}_i \subset \tilde{\mathcal{A}}_i$ (because $\ell_j \geq \ell_i$ for $j \geq i$) and approximately $\tilde{\mathcal{A}}_i \subset \mathcal{A}_i$. The reduced ancestor sets are of size $|\tilde{\mathcal{A}}_i| = \mathcal{O}(\rho^d \log n) = \mathcal{O}(m \log n)$ and can all be computed together in $\mathcal{O}(nm \log^2 n)$ time (Schäfer et al., 2021b). Hence, reduced ancestor sets can be orders of magnitude smaller than full ancestor sets (see Figures 3a and 6).

**Claim 2.4.** *For Matérn-type LGPs with exponential-family likelihoods,* $(\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i})_j \approx 0$ *for all* $j \notin \tilde{\mathcal{A}}_i$*, where* $\mathbf{V}$ *minimizes the ELBO in* (6)*, under mild conditions.*

We provide a non-rigorous justification for this claim in Appendix C. Together, Proposition 2.3 and Claim 2.4 imply that $\|\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}\| \approx \|\mathbf{V}^{-1}_{\tilde{\mathcal{A}}_i,\tilde{\mathcal{A}}_i}\hat{\mathbf{L}}_{\tilde{\mathcal{A}}_i,i}\|$ (as illustrated in Figure 3b), and so replacing the former by the latter in the ELBO causes negligible error (Figure 3c). Similar numerical results were obtained for two other popular kernels in Figures 7 and 8 in Appendix A, suggesting that our approach is applicable to beyond the Matérn family.

### 2.6. Optimization of the ELBO

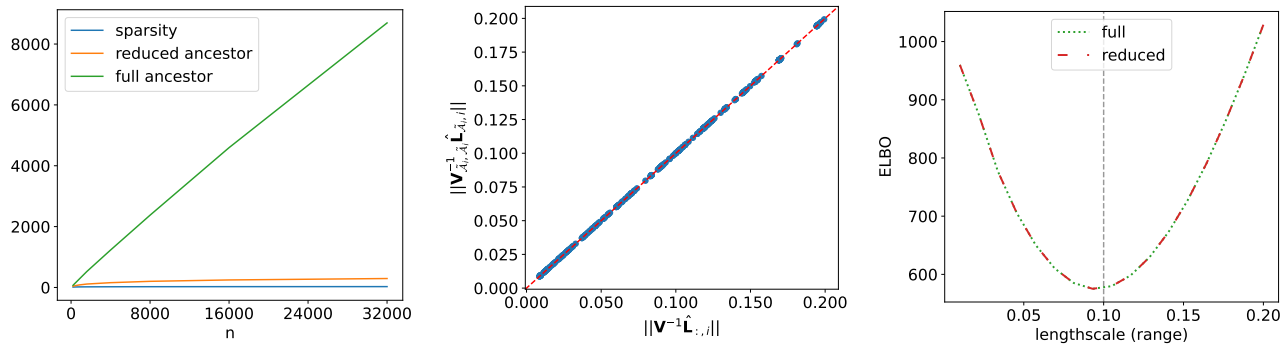
The class of distributions $\mathcal{Q} = \{\mathcal{N}_n(\boldsymbol{\nu}, (\mathbf{V}\mathbf{V}^\top)^{-1}) : \boldsymbol{\nu} \in \mathbb{R}^n, \mathbf{V} \in \mathbb{R}^{n\times n}, \mathbf{V} \in \mathcal{S}^q\}$ has $n$ parameters in $\boldsymbol{\nu}$ and $|\mathcal{S}|$ parameters in $\mathbf{V}$. We propose to find the optimal $\hat{q} \in \mathcal{Q}$ by minimizing our approximation of $-\operatorname{ELBO}(q)$ with respect to these $\mathcal{O}(nm)$ unknown parameters via minibatch stochastic gradient descent. For each minibatch $\mathcal{B}$, this requires computing the gradient of

$$\sum_{i\in\mathcal{B}} \Big( \mathop{\mathbb{E}}_{q} \log p(y_i|f_i) - ((\boldsymbol{\nu} - \boldsymbol{\mu})^\top \hat{\mathbf{L}}_{:,i})^2/2 + \log(\mathbf{V}_{ii}^{-1}\hat{\mathbf{L}}_{ii}) - \|\mathbf{V}^{-1}_{\tilde{\mathcal{A}}_i,\tilde{\mathcal{A}}_i}\hat{\mathbf{L}}_{\tilde{\mathcal{A}}_i,i}\|^2/2 \Big) \tag{8}$$

using automatic differentiation.

For Gaussian observations with $y_i|f_i \sim \mathcal{N}(f_i, \tau_i^2)$, we have $-2\,\mathbb{E}_q \log p(y_i|f_i) = \big((y_i - \boldsymbol{\nu}_i)^2 + \|\mathbf{V}^{-1}\mathbf{e}_i\|^2\big)/\tau_i^2 + \log\tau_i^2 + \log 2\pi$. For more general distributions $p(y_i|f_i)$, we can use the Monte Carlo gradient estimator (Kingma & Welling, 2014) and approximate $\mathbb{E}_q \log p(y_i|f_i) \approx (1/L)\sum_{l=1}^{L} p(y_i|f_i^{(l)})$, where $f_i^{(l)} = \boldsymbol{\nu}_i + (\mathbf{V}^{-1}\mathbf{e}_i)^\top \mathbf{z}^{(l)}$, $\mathbf{z}^{(l)} \overset{iid}{\sim} \mathcal{N}_n(\mathbf{0}, \mathbf{I}_n)$, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Evaluating each summand in (8) requires $\mathcal{O}(|\mathcal{S}_i|^3) = \mathcal{O}(m^3)$ time for obtaining $\hat{\mathbf{L}}_{:,i}$ and $\mathcal{O}(m^2 \log n)$ time for solving $\mathbf{V}^{-1}_{\tilde{\mathcal{A}}_i,\tilde{\mathcal{A}}_i}\hat{\mathbf{L}}_{\tilde{\mathcal{A}}_i,i}$, because $|\tilde{\mathcal{A}}_i| = \mathcal{O}(m \log n)$. The $\mathcal{O}(m^3)$ cost dominates, as we typically need $m = \mathcal{O}(\log^d n)$ for accurate approximations (Schäfer et al., 2021a); for example, in Figure 3a, $|\tilde{\mathcal{A}}_i||\mathcal{S}_i|$ is smaller than $|\mathcal{S}_i|^3$. Also, $\hat{\mathbf{L}}$ does not need to be pre-computed and stored, as each column $\hat{\mathbf{L}}_{:,i}$ can be computed "on-the-fly"; this is especially useful for hyperparameter estimation, for which $p(\mathbf{f})$ and hence $\hat{\mathbf{L}}$ changes with the hyperparameters at each gradient-descent iteration.




(a) Average sizes of $\mathcal{S}_i$, $\tilde{\mathcal{A}}_i$, and $\mathcal{A}_i$

(b) $\|\mathbf{V}^{-1}_{\tilde{\mathcal{A}}_i,\tilde{\mathcal{A}}_i}\hat{\mathbf{L}}_{\tilde{\mathcal{A}}_i,i}\|$ vs $\|\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}\|$

(c) ELBO using reduced ancestors

*Figure 3.* Reduced ancestor sets are much smaller than full ancestor sets, as shown in (a), and hence greatly reduce computational cost, but result in negligible approximation error in the ELBO, as shown in (b) and (c). Specifically, (a) shows average sizes of the sparsity sets $\mathcal{S}_i$, reduced ancestor sets $\tilde{\mathcal{A}}_i$, and full ancestor sets $\mathcal{A}_i$ as a function of $n$ with $d = 5$; for $n = 32{,}000$, we have $|\mathcal{S}_i| = 30$, $|\tilde{\mathcal{A}}_i| = 293$, and $|\mathcal{A}_i| = 8{,}693$. (b) compares $\|\mathbf{V}^{-1}_{\tilde{\mathcal{A}}_i,\tilde{\mathcal{A}}_i}\hat{\mathbf{L}}_{\tilde{\mathcal{A}}_i,i}\|$ with reduced ancestor sets versus $\|\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}\|$ for $i = 1, \dots, n$, where $n = 500$ and $d = 2$. (c) compares ELBO curves based on full (6) and reduced (8) ancestor sets, as functions of the range parameter with true value 0.1, for $n = 500$ and $d = 2$. In all plots, we set $\rho = 2$ and the $n$ inputs are sampled uniformly on $[0, 1]^d$.

We initialize the optimization using an estimate of $\boldsymbol{\nu}$ and $\mathbf{V}$ based on a Vecchia-Laplace approximation (Zilber & Katzfuss, 2021) of $p(\mathbf{f}|\mathbf{y})$ combined with an efficient incomplete Cholesky (IC0) approximation (Schäfer et al., 2021a) of the posterior SIC factor. While this initialization itself provides a reasonable approximation to the posterior, hyperparameter estimation for this approach is more difficult, and it is less accurate than DKLGP even for known hyperparameters as shown in Figure 9 in Appendix A.

The ordering and sparsity pattern in Section 2.5 depend on a distance metric, $\text{dist}(\mathbf{x}_i, \mathbf{x}_j)$, between inputs. We have found that the accuracy of the resulting approximation can be improved substantially by computing the Euclidean distance between inputs in a transformed input space in which the GP kernel is isotropic, as suggested by Katzfuss et al. (2022); Kang & Katzfuss (2023). For example, consider an automatic relevance determination (ARD) kernel of the form $K(\mathbf{x}_i, \mathbf{x}_j) = K_o(q(\mathbf{x}_i, \mathbf{x}_j))$, where $K_o$ is an isotropic kernel (e.g., a Matérn kernel with smoothness 1.5 is used throughout this paper) and $q(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i^{\boldsymbol{\lambda}} - \mathbf{x}_j^{\boldsymbol{\lambda}}\|$ is a Euclidean distance based on scaled inputs $\mathbf{x}^{\boldsymbol{\lambda}} = (x_1/\lambda_1, \dots, x_d/\lambda_d)$ with individual ranges or length-scales $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d)$ for the $d$ input dimensions. In this example, we take $\text{dist}(\mathbf{x}_i, \mathbf{x}_j) = q(\mathbf{x}_i, \mathbf{x}_j)$ when computing the sparsity pattern. When the scaled distance and hence the sparsity pattern depend on unknown hyperparameters (e.g., $\boldsymbol{\lambda}$ in the ARD case), we carry out a two-step optimization procedure: First, we run our ELBO optimization for a few epochs based on the sparsity pattern obtained using an initial guess of $\boldsymbol{\lambda}$ to obtain a rough estimate of $\boldsymbol{\lambda}$, which we then use to obtain the final ordering and sparsity pattern and warm-start our ELBO optimization.

### 2.7. Prediction

An important task for (L)GP models is prediction at unobserved inputs, meaning that we want to obtain the posterior distribution of latent GP variables $\mathbf{f}^*$ at new inputs $\mathbf{x}_1^*, \dots, \mathbf{x}_{n^*}^*$ given the data $\mathbf{y}$. To do so, we consider the joint posterior distribution of $\tilde{\mathbf{f}} = (\mathbf{f}^*, \mathbf{f})$, from which any desired marginal distribution can be computed. Since working with the joint covariance matrix $\tilde{\mathbf{K}}$ is again computationally prohibitive, we make a joint SIC assumption on the posterior distribution of $\tilde{\mathbf{f}}$ (with the prediction variables ordered first) that naturally extends the SIC assumption for $\mathbf{f}$ in $q(\mathbf{f})$. For the exact posterior, we have

$$p(\tilde{\mathbf{f}}|\mathbf{y}) = p(\mathbf{f}^*|\mathbf{f}, \mathbf{y})p(\mathbf{f}|\mathbf{y}) = p(\mathbf{f}^*|\mathbf{f})p(\mathbf{f}|\mathbf{y}).$$

Similarly, we assume $q(\tilde{\mathbf{f}}) = q(\mathbf{f}^*|\mathbf{f})q(\mathbf{f})$, where $q(\mathbf{f}) = \mathcal{N}_n(\mathbf{f}|\boldsymbol{\nu}, (\mathbf{V}\mathbf{V}^\top)^{-1})$ was obtained as described in previous sections, and $q(\mathbf{f}^*|\mathbf{f})$ is a sparse approximation of $p(\mathbf{f}^*|\mathbf{f})$. For $i = 1, \dots, n^*$, let $\mathcal{S}_i^* \subset \{i, i+1, \dots, n^* + n\}$ denote the $i$th sparsity set relative to the joint posterior.

We define the approximation to the joint posterior by the minimizer of the expected forward-KL divergence between $p(\mathbf{f}^*|\mathbf{f})$ and $q(\mathbf{f}^*|\mathbf{f})$ for given $\boldsymbol{\nu}$ and $\mathbf{V}$, that is,

$$\hat{q}(\tilde{\mathbf{f}}) = \underset{q(\tilde{\mathbf{f}}) \in \tilde{\mathcal{Q}}(\boldsymbol{\nu}, \mathbf{V})}{\arg\min} \; \underset{p}{\mathbb{E}} \Big[ \text{KL}\left(p(\mathbf{f}^*|\mathbf{f}) \big\| q(\mathbf{f}^*|\mathbf{f})\right) \Big],$$

where

$$\tilde{\mathcal{Q}}(\boldsymbol{\nu}, \mathbf{V}) = \{\mathcal{N}_{n^*+n}((\boldsymbol{\nu}^{*\top}, \boldsymbol{\nu}^\top)^\top, (\mathbf{V}^*, (\mathbf{0}, \mathbf{V}^\top)^\top)) : \boldsymbol{\nu}^* \in \mathbb{R}^{n^*}, \mathbf{V}^* \in \mathbb{R}^{(n^*+n)\times n^*}, \mathbf{V}^* \in \mathcal{S}^*\}$$

and $\mathcal{S}^* = \bigcup_{i=1}^{n^*} \{(j, i) : j \in \mathcal{S}_i^*\}$. The resulting approximation can be obtained efficiently:

**Proposition 2.5.** *For given* $\boldsymbol{\nu}$, $\mathbf{V}$, *and* $\mathcal{S}^*$, $\hat{q}(\tilde{\mathbf{f}}) = \mathcal{N}_{n^*+n}(\tilde{\mathbf{f}}|\tilde{\boldsymbol{\nu}}, (\tilde{\mathbf{V}}\tilde{\mathbf{V}}^\top)^{-1})$, *where* $\tilde{\boldsymbol{\nu}} = (\hat{\boldsymbol{\nu}}^{*\top}, \boldsymbol{\nu}^\top)^\top$, $\tilde{\mathbf{V}} = (\hat{\mathbf{V}}^*, (\mathbf{0}, \mathbf{V}^\top)^\top)$, $\hat{\mathbf{V}}^* = (\hat{\mathbf{V}}^{**\top}, \hat{\mathbf{V}}^{o*\top})^\top$,

$$\hat{\mathbf{V}}^*_{\mathcal{S}^*_i,i} = \mathbf{c}_i(\mathbf{c}_{i,1})^{-1/2}, \quad \textit{with } \mathbf{c}_i = K(\mathcal{S}^*_i, \mathcal{S}^*_i)^{-1}\mathbf{e}_1,$$

$$\hat{\boldsymbol{\nu}}^* = \boldsymbol{\mu}^* - (\hat{\mathbf{V}}^{**})^{-\top}\hat{\mathbf{V}}^{o*\top}(\boldsymbol{\nu} - \boldsymbol{\mu}),$$

*and* $\boldsymbol{\mu}^* = (\mu(\mathbf{x}^*_1), \ldots, \mu(\mathbf{x}^*_{n^*}))^\top$.

The posterior distribution of a desired summary, say $\mathbf{a}^\top\tilde{\mathbf{f}}$ can then be computed as $q(\mathbf{a}^\top\tilde{\mathbf{f}}) = \mathcal{N}(\mathbf{a}^\top\tilde{\boldsymbol{\nu}}, \|\tilde{\mathbf{V}}^{-1}\mathbf{a}\|^2)$. In particular, the marginal posterior of $f^*_i$ can be obtained using $\mathbf{a} = \mathbf{e}_i$ as $q(\mathbf{e}_i^\top\tilde{\mathbf{f}}) = \mathcal{N}(\boldsymbol{\nu}^*_i, \|\tilde{\mathbf{V}}^{-1}\mathbf{e}_i\|^2)$.

We again consider an r-maximin ordering and nearest-neighbor sparsity pattern similar to above, but now conditioned on the prediction points being ordered first, and the training points ordered after (in the same ordering as before). Once the prediction points are in this conditional r-maximin ordering, we can define

$$\ell^*_i = \min_{i<j\le n^*} \text{dist}(\mathbf{x}^*_i, \mathbf{x}^*_j) \wedge \min_{1\le j\le n} \text{dist}(\mathbf{x}^*_i, \mathbf{x}_j)$$

and

$$\mathcal{S}^*_i = \{j \ge i : \text{dist}(\mathbf{x}^*_i, \mathbf{x}^*_j) \le \rho\ell^*_i\} \cup \{j + n^* : \text{dist}(\mathbf{x}^*_i, \mathbf{x}_j) \le \rho\ell^*_i\}.$$

This ordering and sparsity pattern can be computed rapidly and was shown to lead to highly accurate approximations; more details can be found in Schäfer et al. (2021a, Section 4.2.1). Note that while computing the prediction variances can be expensive, we can again approximate $\|\tilde{\mathbf{V}}^{-1}\mathbf{e}_i\| \approx \|\tilde{\mathbf{V}}^{-1}_{\tilde{\mathcal{A}}^*_i,\tilde{\mathcal{A}}^*_i}\mathbf{e}_{i;\tilde{\mathcal{A}}^*_i}\|$ using a reduced ancestor set

$$\tilde{\mathcal{A}}^*_i = \{j \ge i : \text{dist}(\mathbf{x}^*_i, \mathbf{x}^*_j) \le \rho\ell^*_j\} \cup \{j + n^* : \text{dist}(\mathbf{x}^*_i, \mathbf{x}_j) \le \rho\ell^*_j\},$$

where the last subscript is a $j$, not an $i$.

## 3. Numerical Comparisons

### 3.1. Experimental Setup

We compared the following approaches:

**DKLGP:** Our method with r-maximin ordering and nearest-neighbor sparsity pattern

**DKL-G:** Same as DKLGP but with global sparsity pattern $\mathcal{S}^p_i = \mathcal{S}^q_i = \{1, \ldots, m\}$

**DKL-D:** Same as DKLGP but with diagonal sparsity pattern $\mathcal{S}^q_i = \{i\}$

**SVIGP:** Stochastic variational GP proposed by Hensman et al. (2013)

**VNNGP:** Variational nearest neighbor GP proposed by Wu et al. (2022)

In figures and tables, we use abbreviated acronyms DKL, SVI, and VNN to save space. SVIGP and VNNGP are two state-of-the-art variational GP methods, while DKL-G and DKL-D are variants of our DKLGP that resemble SVIGP and VNNGP, respectively. SVIGP assumes independence in $\mathbf{f}$ conditional on $m$ global inducing variables. VNNGP scales up the inducing points to be equal to the observed input locations, ensuring computational feasibility by assuming that each conditions only on $m$ others a priori, combined with a mean-field approximation to the posterior. We used the GPyTorch (Gardner et al., 2018) implementations of SVIGP and VNNGP. For DKL-G and DKL-D, one can easily see that $\mathcal{A}_i = \mathcal{S}^p_i$, and so reduced ancestor sets are not necessary. For all methods, computing a term in the ELBO requires $\mathcal{O}(m^3)$ time per sample. (Reusing Cholesky factors for all samples in a minibatch is straightforward for SVIGP; similar savings may also be possible for the other methods based on the supernode ideas suggested by Schäfer et al., 2021a.) Hence, $m$ can be viewed as a comparable complexity parameter that trades off computational speed (for small $m$) against accuracy (large $m$). Thus, for our numerical comparison, we aligned the $m$ for all methods with the average size of $\mathcal{S}_i$ for a given $\rho$.

Throughout, we assumed $f(\cdot) \sim \mathcal{GP}(0, K)$, where $K$ is a Matérn1.5 ARD kernel whose variance (set to one for simulations) and range (i.e., length-scale) parameters $\boldsymbol{\lambda}$ were estimated. We considered three different likelihoods $p(y_i|f_i)$:

**Gaussian:** $y_i|f_i \sim \mathcal{N}(f_i, \sigma^2_\epsilon)$
**Student-$t$:** $y_i|f_i \sim \mathcal{T}_2(f_i, \sigma^2_\epsilon)$ with 2 degrees of freedom
**Bernoulli-logit:** $y_i|f_i \sim \mathcal{B}((1 + e^{-f_i})^{-1})$

The noise variance $\sigma^2_\epsilon$ was estimated from the data; for simulations, we used $\sigma^2_\epsilon = 0.1^2$ except where specified otherwise.

For estimation of hyperparameters, the initial values for $\boldsymbol{\lambda}$, $\sigma^2_\epsilon$, and the variance in $K$ were all $0.25$. DKLGP and its variants ran the Adam optimizer for $35$ epochs. SVIGP and VNNGP used natural gradient descent and Adam, respectively, as their optimizer for $500$ epochs as suggested in Wu et al. (2022). The minibatch size was $128$ and a multi-step scheduler with a scaling factor of $0.1$ was used for all methods.

### 3.2. Visual Comparison in One Dimension

Figure 4 provides a visual comparison of SVIGP, VNNGP, and DKLGP predictions for a toy example in one dimension. We also included predictions from the exact GP (DenseGP) which cannot be obtained for large $n$. DKLGP approximated the DenseGP most closely, especially in terms of

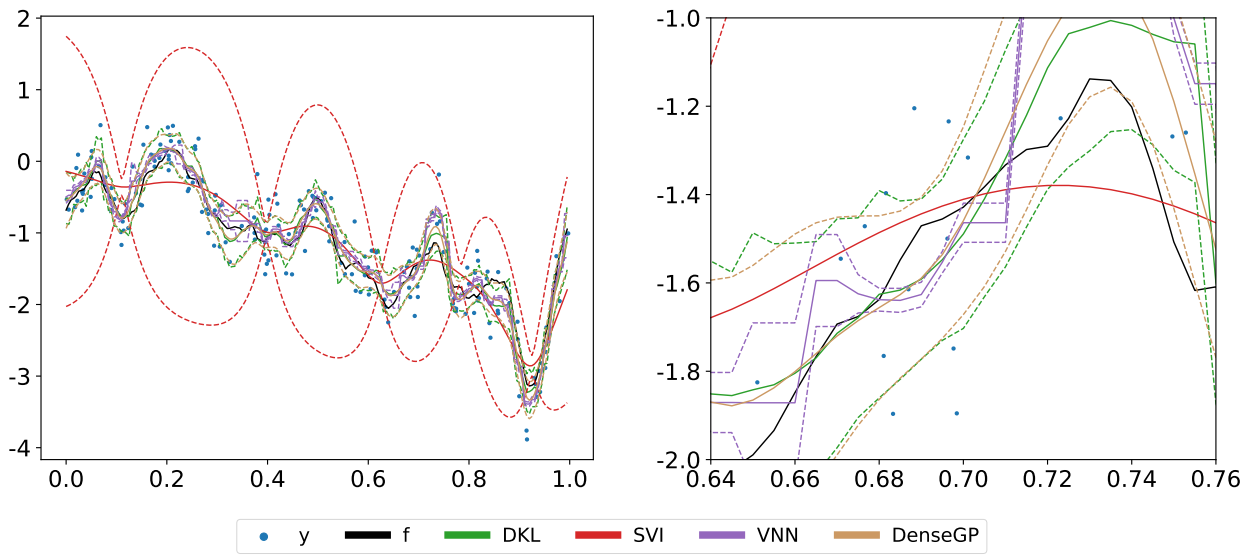



*Figure 4.* Comparison of exact GP predictions (DenseGP) to three variational GP approximations for simulated data with Gaussian noise at $n = 200$ randomly sampled training inputs on $[0, 1]$ with $\sigma_\epsilon = 0.3$ and true range $\lambda = 0.1$. We show the means (solid lines) and 95% pointwise intervals of the posterior predictive distribution $\mathbf{f}^*|\mathbf{y}$ at 200 regularly spaced test inputs. The right panel zooms into a smaller region of the left panel to highlight the differences.

the prediction intervals. SVIGP oversmoothed heavily and produced very wide prediction intervals. VNNGP assumes a diagonal covariance in the variational distribution $q(\mathbf{f})$, which appears to have caused sharply fluctuating predictions and narrow prediction intervals. Figure 11 in Appendix A shows similar comparisons for Student-$t$ and Bernoulli likelihoods.

### 3.3. Results on Synthetic Data

We also carried out a more comprehensive comparison for 10,000 inputs randomly distributed in the unit hypercube, $[0, 1]^5$, with true range parameters $\boldsymbol{\lambda} = (0.25, 0.50, 0.75, 1.00, 1.25)$. We used $n = 8{,}000$ inputs for training and 2,000 for testing. Performance was measured in terms of the variational inference of the latent field $f(\cdot)$ at training and test inputs. For each scenario, results over five replicates were produced and averaged.

Figure 5 compares root mean squared error (RMSE) and negative log-likelihood (NLL) at test inputs. For the Gaussian and Student-$t$ likelihoods, DKLGP produced the most accurate predictions, while for the Bernoulli-logit likelihood, SVIGP and DKLGP appeared to be similarly accurate in terms of RMSE. DKLGP outperformed the competing methods in terms of NLL. While DKLGP, DKL-G, and SVIGP improved with increasing $\rho$ as expected, the mean-field approximations (VNNGP and DKL-D) generally did not. We performed the same comparison for the squared exponential and rational quadratic kernels in Figures 12 and 13 in Appendix A, which resulted in the same rankings as for the Matérn kernel, except that DKLGP was marginally outperformed by SVIGP in terms of RMSE for the Bernoulli-logit likelihood at $\rho = 2.0$.

We also computed RMSE and NLL scores at training inputs for the methods we considered in Figure 5, as presented in Figure 10 in Appendix A. Consistent with the results from Figure 5, DKLGP generally performed best, which is consistent with the results from Figure 5. VNNGP performed similarly to DKLGP for the Gaussian and Student-$t$ likelihoods, but underestimated the variance at test inputs and so led to poor NLL scores. Note that variational methods are generally known to underestimate the posterior variance (Blei et al., 2017).

### 3.4. Results on UCI Data

To provide a more comprehensive comparison of SVIGP, VNNGP and DKLGP, we considered datasets from the UCI data repository widely used for benchmarking purposes. For the UCI datasets we considered in this section, covariates were first standardized to $[0, 1]$ and removed from analysis if the standard deviation after standardization was smaller than $0.01$. Furthermore, inputs were filtered to ensure that the minimum distance between inputs was greater than $0.001$ to prevent numerical singularity. Approximately 20% of each dataset was used for testing. We chose different $\rho$ for different datasets and computed the corresponding $m$. Since Section 3.3 demonstrated the advantage of DKLGP over its variants DKL-G and DKL-D, we excluded the two variants here for ease of presentation. We included SVIGP with

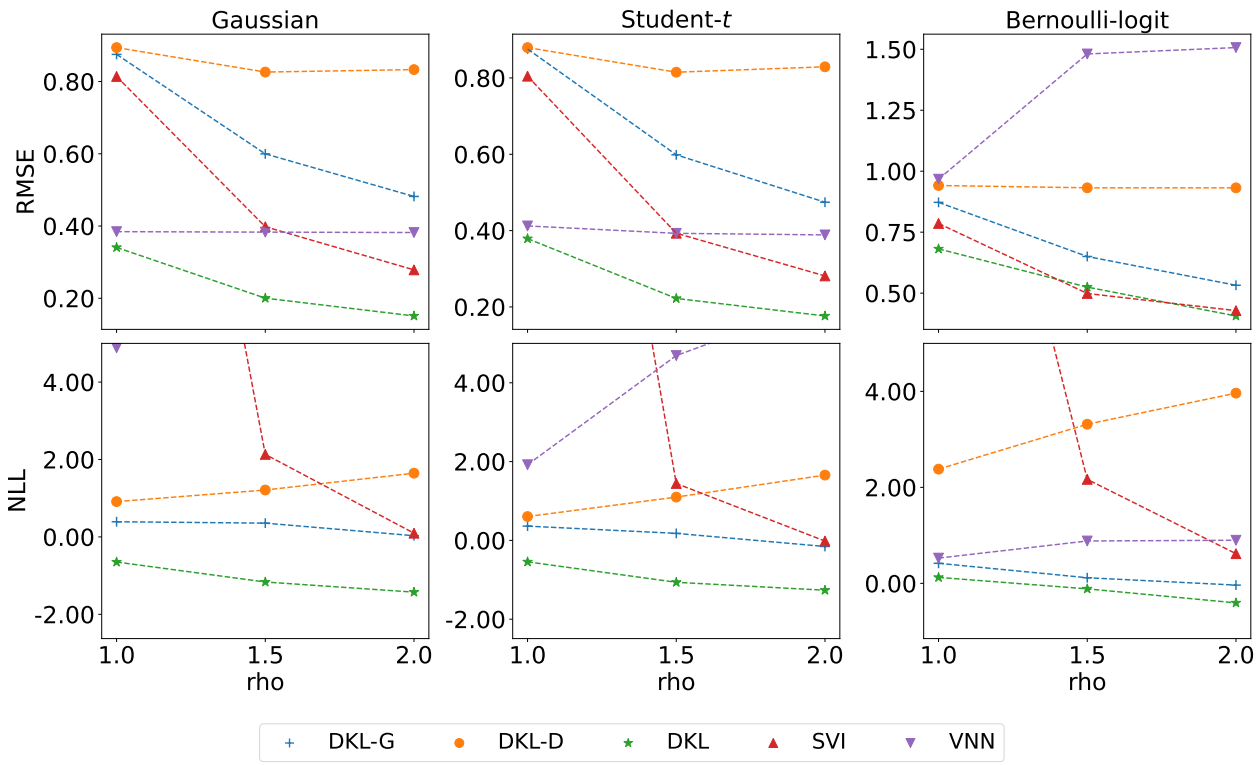



*Figure 5.* RMSE (top) and NLL (bottom) for predicting the latent field at test inputs for simulated data in a five-dimensional input domain, as a function of the complexity parameter $\rho$, with Gaussian (left), Student-$t$ (center) and Bernoulli-logit (right) likelihoods

$m = 32$ and $m = 512$ inducing points as benchmarks for easier comparison with relevant works in the literature.

Table 1 summarizes the performance of the three methods across nine UCI datasets. DKLGP had better scores than VNNGP for all datasets except for COVTYPE, for which VNNGP ran out of memory on a 64GB node despite having reduced the data size to a subset of size 100K. Compared to the SVIGP with similar computation cost, DKLGP provided substantially better performance for the binary response data COVTYPE and for low-dimensional ($d < 10$) settings, and roughly similar performance for most high-dimensional datasets except the KEGGU data, for which SVIGP produced much lower RMSE than DKLGP. However, this does not appear to be due to DKLGP providing a less accurate approximation to the exact GP, but rather it appears to be due to the exact GP (with its simple ARD kernel) being severely misspecified for KEGGU. To explore this further, we fitted the exact GP (DenseGP) to KEGGU. The DenseGP's RMSE was 0.14 (same as for DKLGP), and the root average squared distance between the DenseGP predictions and the DKLGP and SVIGP predictions was 0.05 and 0.13, respectively, which implies that the DKLGP predictions were a much better approximation of the exact-GP predictions than the SVIGP predictions. VNNGP provided better point predictions than SVIGP for the low-dimensional datasets, which is consistent with the results in Wu et al. (2022); however, VNNGP's NLL was high due to its underestimation of posterior variance.

Table 2 summarizes the wall-clock times on an Intel Xeon E5-2680 v4 CPU with 14 cores and 28 threads for the methods under comparison, where the computation of sparsity and ancestor sets is only applicable to DKLGP or DKL. The DKL computation times were closer to those of SVI than to those of $\text{SVI}_{32}$, indicating that DKLGP and SVIGP should be compared at the same $m$ on the basis of comparable computation times. While increasing $m$ significantly improved SVIGP's performance on low-dimensional datasets, even $m = 512$ inducing points made the training of SVIGP challenging on the workstation we used for comparison. The performance of DKLGP can also be improved by using a larger $\rho$; for example, DKLGP's RMSE for the KIN40K data was reduced to 0.27 for $m = 21$.

## 4. Conclusion

We have introduced a variational approach using a variational family and approximate prior based on SIC restrictions. The (r-)maximin ordering, nearest-neighbor sparsity pattern, and a computational trick based on reduced ancestor sets together result in efficient and accurate inference and

*Table 1.* RMSE and NLL at held-out test points averaged over five splits for several UCI datasets, ordered from low to high dimension $d$. The Student-$t$ and Bernoulli-logit likelihoods were used for PRECIP and COVTYPE, respectively; a Gaussian likelihood was used for the other datasets. The average sparsity-set size for DKL is denoted by $m$. SVI used $m$ inducing points, while $\text{SVI}_{32}$ and $\text{SVI}_{512}$ used 32 and 512 points, respectively. While $\text{SVI}_{32}$ and $\text{SVI}_{512}$ are included for reference, they exhibit substantially higher computational complexity and training time than the other approaches and are hence colored in grey.

| | 3DROAD | | PRECIP | | KIN40K | | PROTEIN | | BIKE | | ELEVATORS | | KEGG | | KEGGU | | COVTYPE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n, d$ | 65K, 3 | | 85K, 3 | | 40K, 8 | | 44K, 9 | | 17K, 17 | | 17K, 18 | | 16K, 20 | | 18K, 26 | | 100K, 53 | |
| $m$ | 2 | | 5 | | 7 | | 8 | | 12 | | 22 | | 19 | | 21 | | 3 | |
| SVI | .80 | .28 | .91 | .43 | .61 | .01 | .81 | 0.29 | .09 | -1.85 | .39 | -.43 | .08 | -2.05 | .06 | -2.30 | .50 | NA |
| $\text{SVI}_{32}$ | .59 | -.02 | .83 | .34 | .37 | -.47 | .75 | 0.22 | .06 | -2.21 | .39 | -.45 | .07 | -2.16 | **.06** | **-2.29** | .50 | NA |
| $\text{SVI}_{512}$ | .38 | -.44 | .64 | .11 | **.17** | **-1.2** | .67 | 0.10 | **.03** | **-2.69** | **.37** | **-.49** | **.07** | **-2.22** | .06 | -2.28 | .50 | NA |
| VNN | .28 | 2.16 | .49 | 4.35 | .56 | 24.19 | .69 | 5.59 | .49 | 7.60 | .65 | 1.26 | .13 | 1.22 | .14 | 3.85 | NA | NA |
| DKL | **.27** | **-.83** | **.41** | **-.38** | .37 | -.55 | **.56** | **-.19** | .11 | -1.63 | .43 | -.37 | .09 | -1.97 | .11 | -2.08 | **.28** | NA |

*Table 2.* Comparison of wall-clock time (in seconds) for the datasets and methods in Table 1. $\mathcal{S}\&\mathcal{A}$ refers to computing the r-maximin ordering, sparsity pattern and ancestor sets.

| | 3DROAD | PRECIP | KIN40K | PROTEIN | BIKE | ELEVATORS | KEGG | KEGGU | COVTYPE |
|---|---|---|---|---|---|---|---|---|---|
| SVI | 3,283 | 3,965 | 3,305 | 3,589 | 1,487 | 1,386 | 1,329 | 1,594 | 5,945 |
| $\text{SVI}_{32}$ | 8,879 | 9,207 | 5,460 | 5,941 | 3,082 | 2,952 | 3,159 | 3,387 | 7,722 |
| $\text{SVI}_{512}$ | 25,409 | 26,223 | 21,710 | 22,179 | 12,232 | 9,518 | 9,988 | 10,642 | 44,839 |
| VNN | 2,788 | 3,332 | 1,696 | 2,081 | 568 | 454 | 487 | 595 | NA |
| DKL | 1,591 | 3,948 | 1,859 | 2,736 | 3,129 | 807 | 1,285 | 1,536 | 4,932 |
| $\mathcal{S}\&\mathcal{A}$ | 90 | 268 | 2,866 | 440 | 170 | 577 | 171 | 207 | 1,277 |

prediction for LGPs. While the time complexity is cubic in the number of neighbors, quadratic complexity for the prior approximation can be achieved by grouping observations and re-using Cholesky factors (Schäfer et al., 2021a); we will investigate an extension of this idea to computing the ELBO in our variational setting. Although we here assume that the input domain is Euclidean, our method can be applied more generally; using a correlation-based distance instead of Euclidean distance (Kang & Katzfuss, 2023), one can use our method to perform LGP inference for large data on complex domains (cf. Tibo & Nielsen, 2022). We will also explore extensions to deep GPs (cf. Sauer et al., 2022). An implementation of our method, along with code to reproduce all results, is publicly available at `https://github.com/katzfuss-group/DKL-GP`.

Our approach is applicable to irregularly spaced observations and in principle to any desired covariance structure. Our method provides state-of-the-art performance when fine-scale structure in the function of interest can be discerned from the data; in contrast, if the data are highly noisy or sparse or the covariance model is severely misspecified, inducing-point methods such as SVIGP that produce smooth predictions and wide uncertainty intervals may be competitive with our approach.

## Acknowledgments

Jian Cao was partially supported by the Texas A&M Institute of Data Science (TAMIDS) Postdoctoral Project program, Jian Cao and Matthias Katzfuss by National Science Foundation (NSF) Grant DMS–1654083, and Felix Jimenez and Matthias Katzfuss by NSF Grant DMS–1953005. We would like to thank Luhuan Wu for helpful comments and discussions.

## References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, 2004.

Bao, J. Y., Ye, F., and Yang, Y. Screening effect in isotropic Gaussian processes. *Acta Mathematica Sinica, English Series*, 36(5):512–534, 2020.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Cao, J., Guinness, J., Genton, M. G., and Katzfuss, M. Scalable Gaussian-process regression and variable selection using Vecchia approximations. *Journal of Machine Learning Research*, 23(348):1–30, 2022.

Chan, A. B. and Dong, D. Generalized Gaussian process models. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2681–2688, 2011.

Cressie, N. and Wikle, C. K. *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ, 2011.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812, 2016.

Deisenroth, M. P. *Efficient reinforcement learning using Gaussian processes*, volume 9. KIT Scientific Publishing, 2010.

Diggle, P., Tawn, J., and Moyeed, R. Model-based geostatistics. *Journal of the Royal Statistical Society, Series C*, 47 (3):299–350, 1998.

Frigola, R., Chen, Y., and Rasmussen, C. E. Variational Gaussian process state-space models. *Advances in Neural Information Processing Systems*, 27, 2014.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. volume 31, 2018.

Geoga, C. J. and Stein, M. L. A scalable method to exploit screening in gaussian process models with noise. *arXiv:2208.06877*, 2022.

Gramacy, R. B. *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Chapman and Hall/CRC, 2020.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D. M., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F., and Zammit-Mangion, A. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological, and Environmental Statistics*, 24(3):398–425, 2019.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *The 29th Conference on Uncertainty in Artificial Intelligence*, pp. 282–290, 2013.

Hensman, J., Matthews, A., and Ghahramani, Z. Scalable variational Gaussian process classification. In *The 18th International Conference on Artificial Intelligence and Statistics*, pp. 351–360. PMLR, 2015.

Kang, M. and Katzfuss, M. Correlation-based sparse inverse Cholesky factorization for fast Gaussian-process inference. *Statistics and Computing*, 33(3):56, 2023.

Katzfuss, M. and Guinness, J. A general framework for Vecchia approximations of Gaussian processes. *Statistical Science*, 36(1):124–141, 2021.

Katzfuss, M., Guinness, J., and Lawrence, E. Scaled Vecchia approximation for fast computer-model emulation. *SIAM/ASA Journal on Uncertainty Quantification*, 10(2): 537–554, 2022.

Kennedy, M. C. and O'Hagan, A. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, 63(3):425–464, 2001.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *The 2nd International Conference on Learning Representations*, 2014.

Leibfried, F., Dutordoir, V., John, S., and Durrande, N. A tutorial on sparse Gaussian processes and variational inference. *arXiv:2012.13962*, 2020.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. When Gaussian process meets big data: A review of scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Liu, L. and Liu, L. Amortized variational inference with graph convolutional networks for Gaussian processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2291–2300. PMLR, 2019.

MacKay, D. J. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Nickisch, H. and Rasmussen, C. E. Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078, 2008.

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

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. Design and analysis of computer experiments. *Statistical Science*, 4 (4):409–435, 1989.

Sauer, A., Cooper, A., and Gramacy, R. B. Vecchia-approximated deep Gaussian processes for computer experiments. *Journal of Computational and Graphical Statistics*, pp. 1–14, 2022.

Schäfer, F., Katzfuss, M., and Owhadi, H. Sparse Cholesky factorization by Kullback-Leibler minimization. *SIAM Journal on Scientific Computing*, 43(3):A2019–A2046, 2021a.

Schäfer, F., Sullivan, T. J., and Owhadi, H. Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. *Multiscale Modeling & Simulation*, 19(2):688–730, 2021b.

Stein, M. L. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, NY, 1999.

Stein, M. L. When does the screening effect hold?. *Annals of Statistics*, 39(6):2795–2819, 12 2011.

Stein, M. L. Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8: 1–19, 5 2014.

Stein, M. L., Chi, Z., and Welty, L. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 66(2):275–296, 2004.

Tibo, A. and Nielsen, T. D. Inducing gaussian process networks. *arXiv:2204.09889*, 2022.

Titsias, M. K. Variational learning of inducing variables in sparse Gaussian processes. In *The 12th International Conference on Artificial Intelligence and Statistics*, volume 5, pp. 567–574, 2009.

Vecchia, A. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society, Series B*, 50(2):297–312, 1988.

Wu, L., Pleiss, G., and Cunningham, J. P. Variational nearest neighbor Gaussian process. In *The 39th International Conference on Machine Learning*, pp. 24114–24130. PMLR, 2022.

Zilber, D. and Katzfuss, M. Vecchia-Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data. *Computational Statistics & Data Analysis*, 153:107081, 2021.

## A. Additional numerical results

This section contains additional figures not shown in the main paper. Complementing Figure 3a, Figure 6 shows that reduced ancestor sets $\tilde{\mathcal{A}}_i$ are much smaller than full ancestor sets $\mathcal{A}_i$ across a range of $\rho$ values.

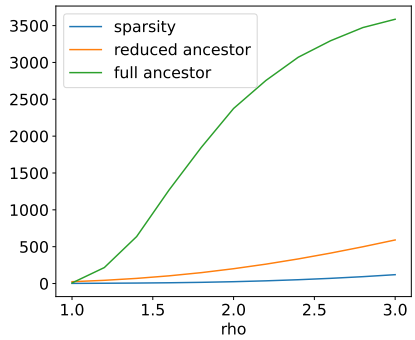
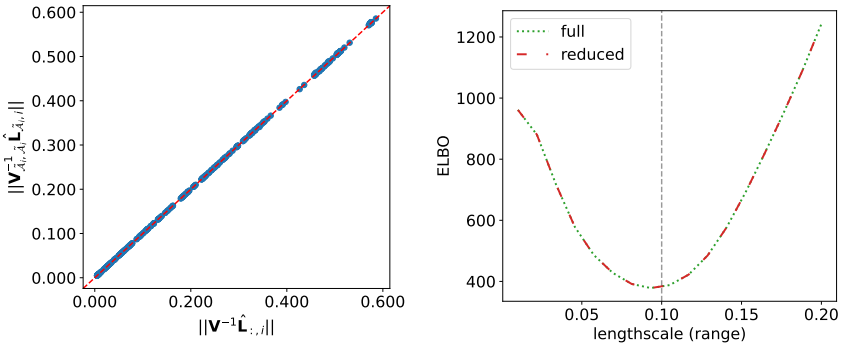



*Figure 6.* Average sizes of the sparsity sets $\mathcal{S}_i$, reduced ancestor sets $\tilde{\mathcal{A}}_i$, and full ancestor sets $\mathcal{A}_i$ as a function of $\rho$ with $n = 8{,}000$. The inputs are sampled uniformly on $[0, 1]^5$.

Complementing Figure 3, Figures 7 and 8 show that the approximation error in computing the ELBO caused by using reduced ancestor sets is negligible even for the squared-exponential and rational-quadratic kernels, respectively.




(a) $\|\mathbf{V}^{-1}_{\tilde{\mathcal{A}}_i,\tilde{\mathcal{A}}_i}\hat{\mathbf{L}}_{\tilde{\mathcal{A}}_i,i}\|$ vs $\|\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}\|$ (b) ELBO using reduced ancestors

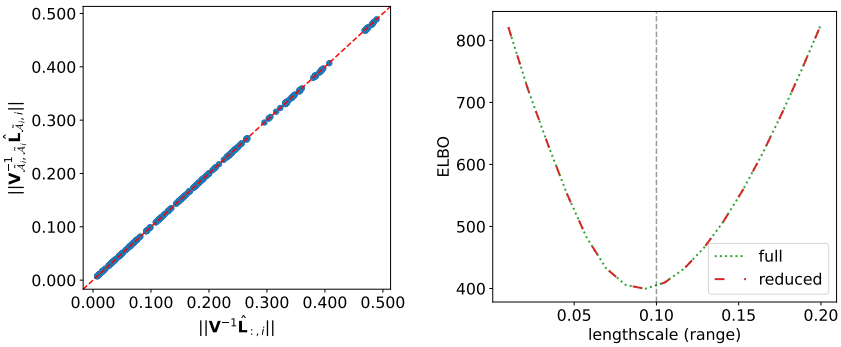
*Figure 7.* The squared-exponential-kernel versions of Figures 3b (left) and 3c (right): The left figure compares $\|\mathbf{V}^{-1}_{\tilde{\mathcal{A}}_i,\tilde{\mathcal{A}}_i}\hat{\mathbf{L}}_{\tilde{\mathcal{A}}_i,i}\|$ with reduced ancestor sets versus $\|\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}\|$ for $i = 1, \ldots, n$, where $n = 500$ and $d = 2$. The right figure compares ELBO curves based on full (6) and reduced (8) ancestor sets, as functions of the range parameter with true value 0.1, for $n = 500$ and $d = 2$. In all plots, we set $\rho = 2$ and the $n$ inputs are sampled uniformly on $[0, 1]^d$.

Figure 9 suggests that the initialization of $\boldsymbol{\nu}$ using Vecchia-Laplace approximation and incomplete Cholesky (IC0) approximation provides reasonable starting values for $\boldsymbol{\nu}$, which can be further refined by optimizing the ELBO.

Figure 10 shows a comparison of RMSE and NLL scores for the posterior marginals of the entries of $\mathbf{f}$ at training inputs. In contrast to Figure 5, VNNGP performed similarly to DKLGP and outperformed SVIGP for Gaussian and Student-t likelihoods. Furthermore, the Vecchia-Laplace approximation with IC0 (used as the initialization for DKLGP) was usually the third best model, indicating an advantage of using the SIC restriction for $\mathbf{L}$ and $\mathbf{V}$.

Figure 11 shows one-dimensional toy examples for Student-$t$ and Bernoulli-logit likelihoods. Note that the DenseGP is available only for the Gaussian likelihood.

Similar to Figure 5, Figures 12 and 13 provide RMSE and NLL scores at test inputs but for the squared-exponential and rational-quadratic kernels, respectively.




(a) $\|\mathbf{V}^{-1}_{\tilde{\mathcal{A}}_i,\tilde{\mathcal{A}}_i}\hat{\mathbf{L}}_{\tilde{\mathcal{A}}_i,i}\|$ vs $\|\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}\|$ (b) ELBO using reduced ancestors

*Figure 8.* The rational-quadratic-kernel versions of Figures 3b (left) and 3c (right): The left figure compares $\|\mathbf{V}^{-1}_{\tilde{\mathcal{A}}_i,\tilde{\mathcal{A}}_i}\hat{\mathbf{L}}_{\tilde{\mathcal{A}}_i,i}\|$ with reduced ancestor sets versus $\|\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}\|$ for $i = 1, \ldots, n$, where $n = 500$ and $d = 2$. The right figure compares ELBO curves based on full (6) and reduced (8) ancestor sets, as functions of the range parameter with true value 0.1, for $n = 500$ and $d = 2$. In all plots, we set $\rho = 2$ and the $n$ inputs are sampled uniformly on $[0, 1]^d$.

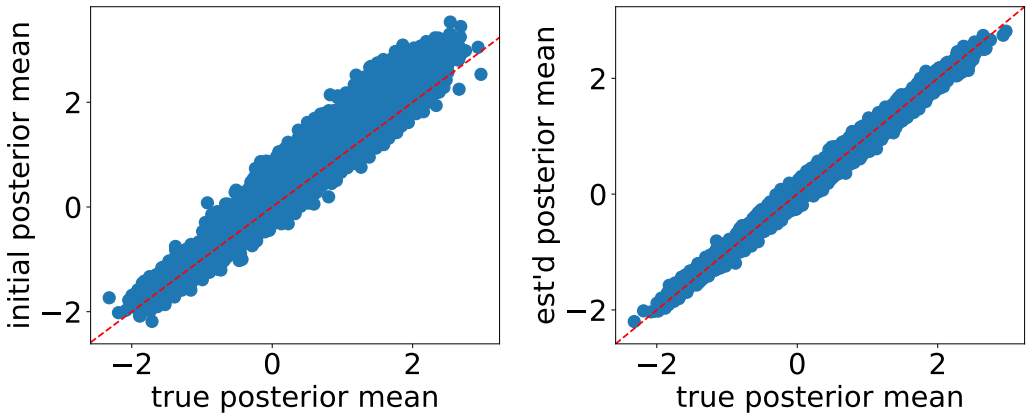



*Figure 9.* Comparison of estimated and true posterior means: The left panel uses the estimated posterior means at the initialization step (i.e., the IC0 solution) and right panel used the estimated posterior means after ELBO optimization for simulated Gaussian data. The simulation setting is the same as in Figure 10.

## B. Graph representation of sparsity patterns and ancestor sets

We here illustrate the sparsity and ancestor sets, using their graph representations. As pointed out by Katzfuss & Guinness (2021), the sparsity patterns can be represented by directed acyclic graphs (DAGs), which also allows straightforward visualization of ancestor sets. Figure 14 presents sparsity and ancestor sets for three selected points ($i = 12, 4, 1$) of 16 grid points in the unit square. For example, $\mathbf{x}_1 = \left(\frac{1}{3}, 1\right)$ and $\mathbf{x}_{16} = \left(\frac{2}{3}, \frac{2}{3}\right)$. One can easily see that $\ell_{16} = \infty$, $\ell_{15} = \frac{2\sqrt{2}}{3}$, $\ell_{14} = \ell_{13} = \sqrt{\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2}$, $\ell_{12} = \ell_{11} = \frac{\sqrt{2}}{3}$ and $\ell_{10} = \cdots = \ell_1 = \frac{1}{3}$. The edges of the graphs corresponding to the ancestor sets $\mathcal{A}_{12}$, $\mathcal{A}_4$ and $\mathcal{A}_1$ are denoted by the black curved arrows. Specifically, the sparsity set $\mathcal{S}_1 = \{2, 7, 13\}$, the reduced ancestor set $\tilde{\mathcal{A}}_1 = \mathcal{S}_1 \cup \{9, 11, 12\}$ and the (full) ancestor set $\mathcal{A}_1 = \tilde{\mathcal{A}}_1 \cup \{15, 16\}$. Note that $\mathcal{A}_1$ contains $\tilde{\mathcal{A}}_1$, which is a desirable property for leveraging the screening effect in GPs (Stein, 2011; Bao et al., 2020). This is not always the case for small-scale problems and it depends on distribution of the points, as shown in Figure 14b. Specifically, $\mathcal{A}_4 = \{10, 11, 14, 15, 16\}$, but $\tilde{\mathcal{A}}_4 \setminus \mathcal{A}_4 = \{13\} \neq \emptyset$. But our numerical studies suggest that $\tilde{\mathcal{A}}_i \setminus \mathcal{A}_i$ are typically empty or very small for large-scale problems, for which computational issues are severe and hence our method is most likely to be used. For relatively large $i = 12$, $\mathcal{S}_{12} = \tilde{\mathcal{A}}_{12} = \mathcal{A}_{12} = \{16\}$. As illustrated here, all the reduced ancestor sets include $\mathbf{x}_{16}$, since $\ell_{16} = \infty$. Otherwise, unlike $\tilde{\mathcal{A}}_4$ and $\tilde{\mathcal{A}}_1$, $\tilde{\mathcal{A}}_{12}$ does not include $\mathbf{x}_{15}$ since $\text{dist}(\mathbf{x}_{15}, \mathbf{x}_{12}) = \sqrt{2}$ is larger than $\rho\ell_{15} \simeq 1.226$.

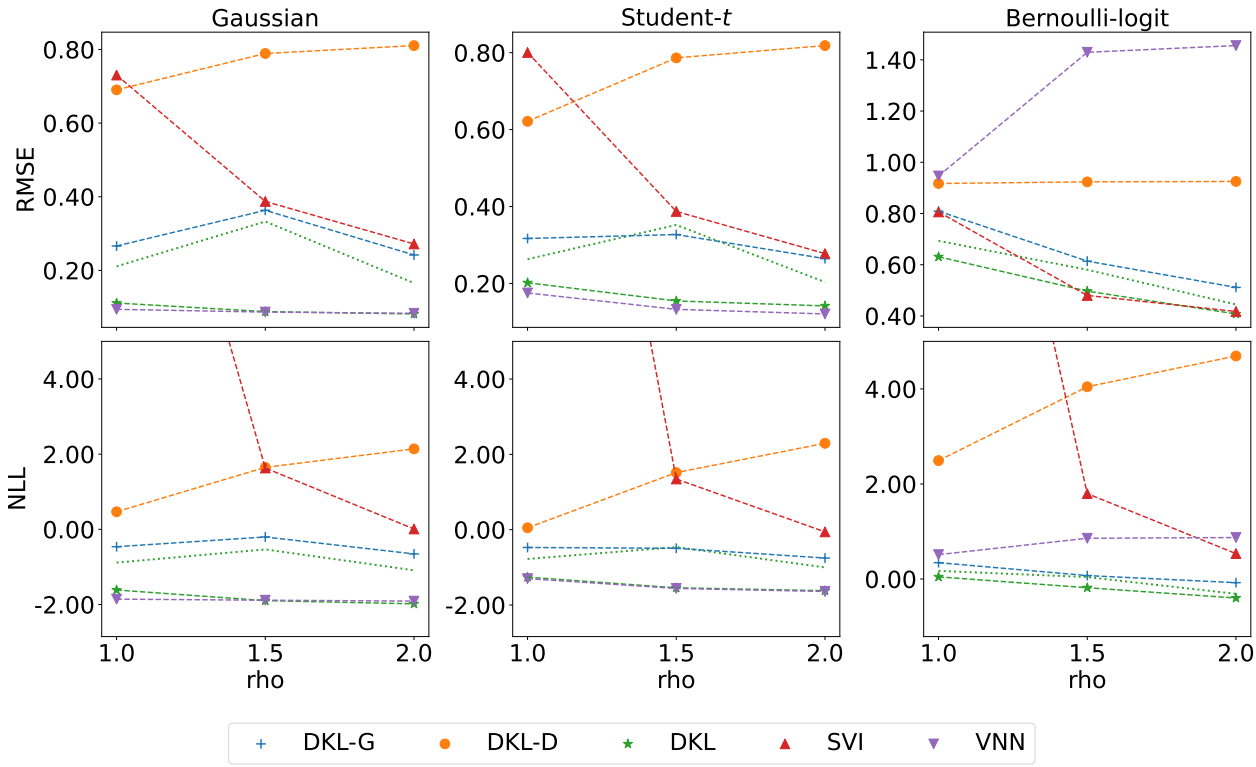



*Figure 10.* RMSE (top) and NLL (bottom) for predicting the latent field at training inputs, as a function of the complexity parameter $\rho$, with Gaussian (left), Student-$t$ (center) and Bernoulli-logit (right) likelihoods, under the same experimental setting in Figure 5. The green dotted lines are the scores of the model obtained only by initialization using Vecchia-Laplace approximation and IC0.

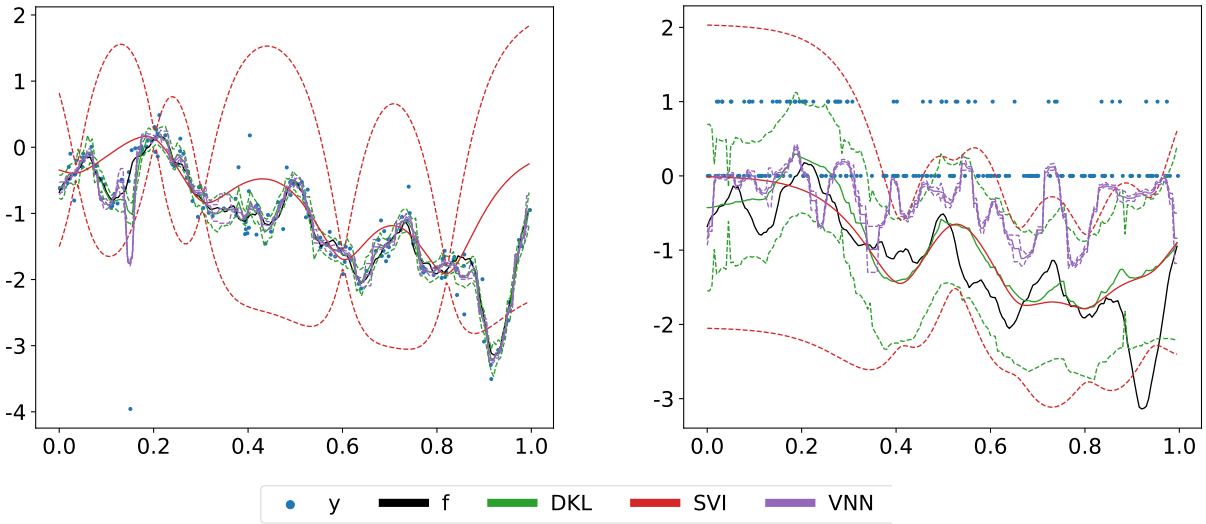



*Figure 11.* Comparison of three variational approximations to the predictive GP posteriors for the Student-$t$ (left) and Bernoulli-logit likelihoods (right). We show the means (solid lines) and 95% pointwise intervals of the posterior predictive distribution $\mathbf{f}^*|\mathbf{y}$ at 200 regularly spaced test inputs. Note that the noise variance $\sigma_\epsilon = 0.3$ and range (or length-scale) $\lambda = 0.1$ are used. The exact-GP result (DenseGP) is available only for the Gaussian likelihood.

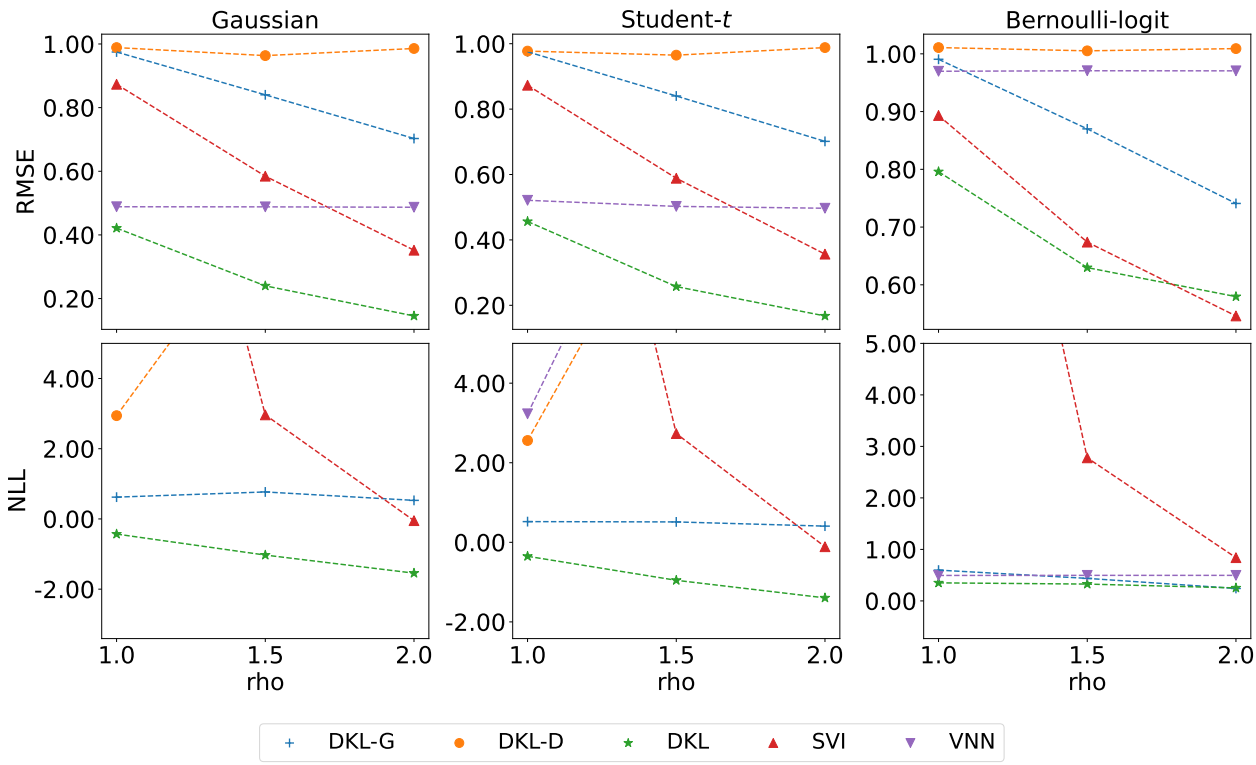



*Figure 12.* RMSE (top) and NLL (bottom) for predicting the latent field at test inputs for simulated data with the squared exponential kernel in a five-dimensional input domain, as a function of the complexity parameter $\rho$, with Gaussian (left), Student-$t$ (center) and Bernoulli-logit (right) likelihoods. In the bottom panels, some lines are truncated for clearer comparison.

## C. Proofs

This section contains the postponed proofs of technical statements in the main paper. A non-rigorous justification for Claim 2.4 can be also found here.

*Proof of Proposition 2.1.* We have

$$\mathrm{ELBO}(q) = \mathop{\mathbb{E}}_{q} \log p(\mathbf{y}|\mathbf{f}) - \mathrm{KL}(q(\mathbf{f})\|p(\mathbf{f})),$$

where $\mathbb{E}_q \log p(\mathbf{y}|\mathbf{f}) = \sum_{i=1}^n \mathbb{E}_q \log p(y_i|f_i)$. Using a well-known expression for the KL divergence between two Gaussian distributions, we have

$$2\,\mathrm{KL}(q(\mathbf{f})\|p(\mathbf{f})) = tr\big((\mathbf{L}\mathbf{L}^\top)(\mathbf{V}\mathbf{V}^\top)^{-1}\big) + (\boldsymbol{\nu}-\boldsymbol{\mu})^\top(\mathbf{L}\mathbf{L}^\top)(\boldsymbol{\nu}-\boldsymbol{\mu}) + \log|\mathbf{V}\mathbf{V}^\top| - \log|\mathbf{L}\mathbf{L}^\top| - n, \tag{9}$$

where $\log|\mathbf{V}\mathbf{V}^\top| = 2\sum_{i=1}^n \log \mathbf{V}_{ii}$, $\log|\mathbf{L}\mathbf{L}^\top| = 2\sum_{i=1}^n \log \mathbf{L}_{ii}$, $(\boldsymbol{\nu}-\boldsymbol{\mu})^\top(\mathbf{L}\mathbf{L}^\top)(\boldsymbol{\nu}-\boldsymbol{\mu}) = \sum_{i=1}^n((\boldsymbol{\nu}-\boldsymbol{\mu})^\top\mathbf{L}_{:,i})^2$, $\mathbf{L}_{:,i}$ denotes the $i$th column of $\mathbf{L}$, and

$$tr\big((\mathbf{L}\mathbf{L}^\top)(\mathbf{V}\mathbf{V}^\top)^{-1}\big) = tr\big((\mathbf{V}^{-1}\mathbf{L})^\top(\mathbf{V}^{-1}\mathbf{L})\big) = \sum_{i=1}^n(\mathbf{V}^{-1}\mathbf{L}_{:,i})^\top(\mathbf{V}^{-1}\mathbf{L}_{:,i}) = \sum_{i=1}^n\|\mathbf{V}^{-1}\mathbf{L}_{:,i}\|^2.$$

□

*Proof of Proposition 2.2.* Using a well-known formula for the KL divergence between two Gaussian distributions (e.g., see (9)), we have

$$\mathrm{KL}\left(p(\mathbf{f})\middle\|\tilde{p}(\mathbf{f})\right) = (\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top(\tilde{\mathbf{L}}\tilde{\mathbf{L}}^\top)(\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu})/2 + \mathrm{KL}\left(\mathcal{N}_n(\mathbf{0},\mathbf{K})\middle\|\mathcal{N}_n(\mathbf{0},(\tilde{\mathbf{L}}\tilde{\mathbf{L}}^\top)^{-1})\right),$$

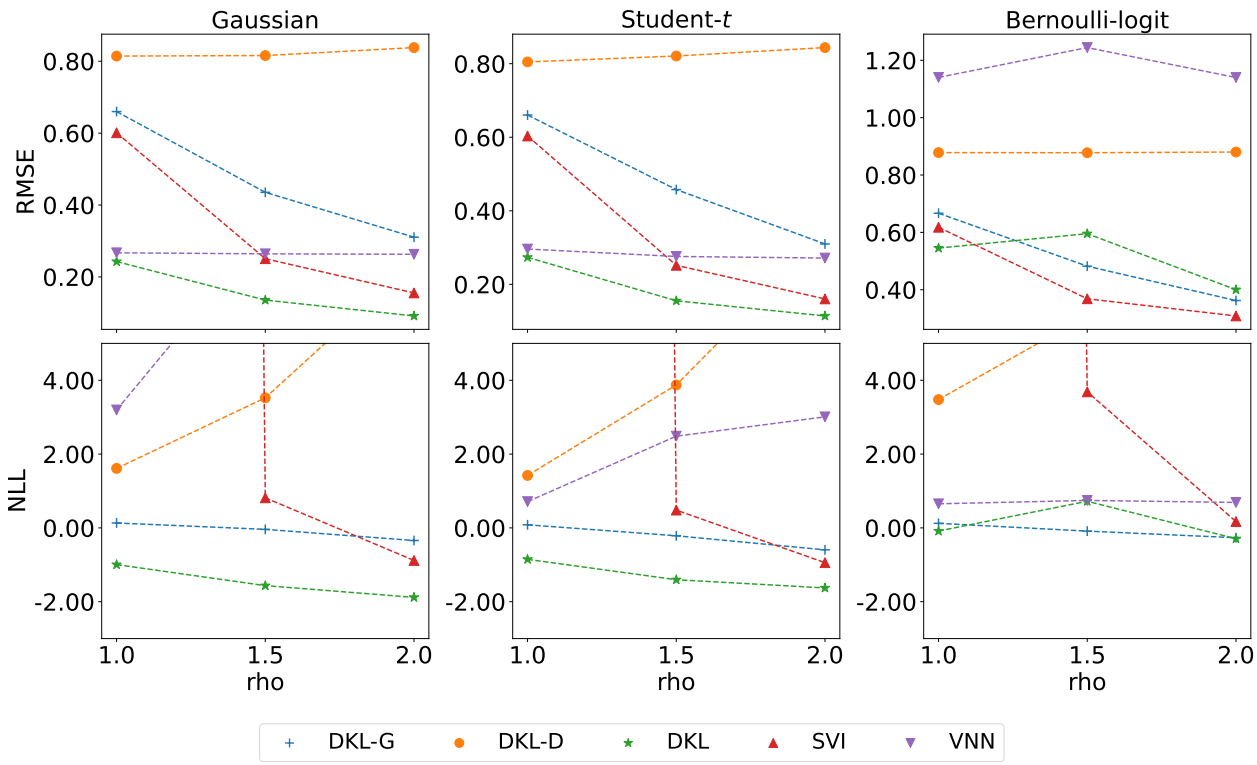



*Figure 13.* RMSE (top) and NLL (bottom) for predicting the latent field at test inputs for simulated data with the rational quadratic kernel in a five-dimensional input domain, as a function of the complexity parameter $\rho$, with Gaussian (left), Student-$t$ (center) and Bernoulli-logit (right) likelihoods

which is minimized with respect to $\tilde{\boldsymbol{\mu}}$ by $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}$, the exact prior mean. Plugging this in, the first summand is zero and the second summand was shown in Schäfer et al. (2021a, Thm. 2.1) to be minimized by an inverse Cholesky factor $\hat{\mathbf{L}}$ whose $i$th column can be computed in parallel for $i = 1, \dots, n$ as

$$\hat{\mathbf{L}}_{\mathcal{S}_i^p,i} = \mathbf{b}_i / \sqrt{\mathbf{b}_{i,1}}, \qquad \text{with } \mathbf{b}_i = \mathbf{K}^{-1}_{\mathcal{S}_i^p,\mathcal{S}_i^p} \mathbf{e}_1.$$

□

*Proof of Proposition 2.3.*

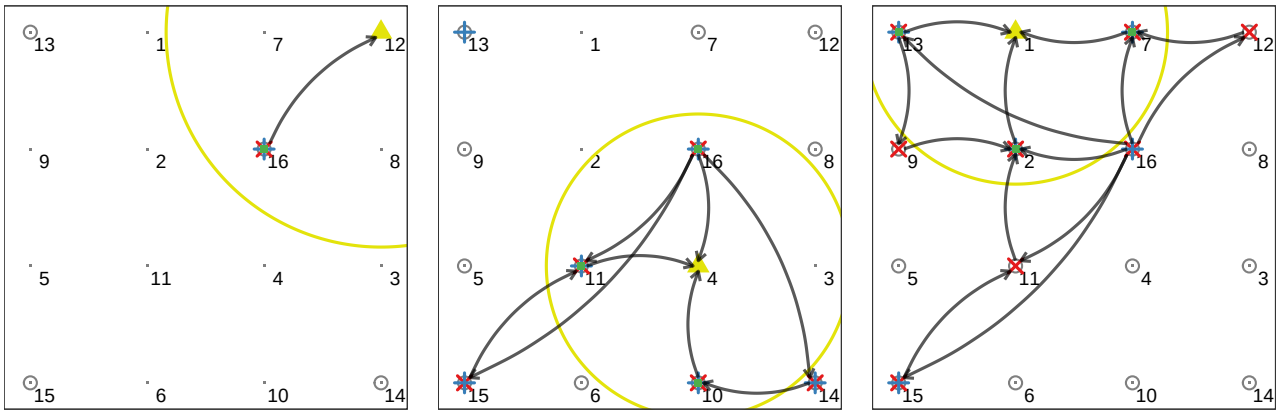
$$\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i} = \begin{bmatrix} \mathbf{V}_{1:i-1,1:i-1} & \mathbf{0} \\ \mathbf{V}_{i:n,1:i-1} & \mathbf{V}_{i:n,i:n} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{L}}_{i:n,i} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{V}^{-1}_{i:n,i:n} \hat{\mathbf{L}}_{i:n,i} \end{bmatrix}$$

Let $\mathbf{X}$ be the inverse of $\mathbf{V}_{i:n,i:n}$. Then,

$$(\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i})_j = \frac{1}{\mathbf{V}_{j,j}} \left[ \hat{\mathbf{L}}_{j,i} - \hat{\mathbf{L}}_{j-1,i} \sum_{r=j-1}^{j-1} \mathbf{V}_{j,r}\mathbf{X}_{r-i+1,j-i} - \dots - \hat{\mathbf{L}}_{i,i} \sum_{r=j-1}^{i} \mathbf{V}_{j,r}\mathbf{X}_{r-i+1,1} \right]$$

Since $\mathcal{S}_i^p \subset \mathcal{A}_i$, $\hat{\mathbf{L}}_{j,i} = 0$ for $j \notin \mathcal{A}_i$. Also, from the definition of $\mathcal{A}_i$, it can be shown for $j \notin \mathcal{A}_i$ that $\hat{\mathbf{L}}_{j-1,i} \sum_{r=j-1}^{j-1} \mathbf{V}_{j,r}\mathbf{X}_{r-i+1,j-i} = \dots = \hat{\mathbf{L}}_{i,i} \sum_{r=j-1}^{i} \mathbf{V}_{j,r}\mathbf{X}_{r-i+1,1} = 0$. For instance, suppose $j = i+1 \notin \mathcal{A}_i$. Then, $(\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i})_{i+1} = \frac{1}{\mathbf{V}_{i+1,i+1}} \left[ \hat{\mathbf{L}}_{i+1,i} - \hat{\mathbf{L}}_{i,i}\mathbf{V}_{i+1,i}\mathbf{X}_{1,1} \right] = 0$, since $\hat{\mathbf{L}}_{i+1,i} = \mathbf{V}_{i+1,i} = 0$. Therefore, $(\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i})_j = 0$ for all $j \notin \mathcal{A}_i$. □

*Justification for Claim 2.4.* We now provide theoretical justification for our claim that the entries of the vector $\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i}$ are small outside of $\tilde{\mathcal{A}}_i$ with magnitudes that decay exponentially as a function of $\rho$ for each $i = 1, \dots, n$. In other words, our




(a) $i = 12$ (b) $i = 4$ (c) $i = 1$

*Figure 14.* Reverse maximin ordering on a grid (small gray points) of size $n = 4 \times 4 = 16$ on a unit square, $[0, 1]^d$ with $d = 2$. The $i$th ordered input (▲), the subsequently ordered $n - i$ inputs (○), the distance $\ell_i$ to the nearest neighbor (—), the neighboring subsequent inputs $\mathcal{S}_i$ (■) within a (yellow) circle of radius $\rho\ell_i$, with $\rho = 1.3$, the reduced ancestors $\tilde{\mathcal{A}}_i$ (+), and the ancestors $\mathcal{A}_i$ (×). The directed acyclic graphs of the sparsity patterns are denoted by arrows (↷).

claim is that for $j \geq i$,

$$\log\left(\left|(\mathbf{V}^{-1}\hat{\mathbf{L}}_{:,i})_j\right|\right) \lessapprox \log(n) - \operatorname{dist}(\mathbf{x}_j, \mathbf{x}_i)/\ell_j.$$

By the results on exponential screening in Schäfer et al. (2021b), the matrix $\hat{\mathbf{L}}$ satisfies the above decay property for covariances that are Green's functions of elliptic PDEs. It satisfies even the stronger property with $\ell_j$ replaced by $\ell_i$.

For a Gaussian likelihood, the matrix $\mathbf{V}$ satisfies

$$\mathbf{V}\mathbf{V}^\top = \hat{\mathbf{L}}\hat{\mathbf{L}}^\top + \mathbf{R}^{-1} =: \mathbf{\Sigma}^{-1}, \tag{10}$$

where $\mathbf{R}$ is a diagonal covariance matrix of the likelihood. Interpreted as a PDE, the diagonal matrix $\mathbf{R}^{-1}$ corresponds to a zero-order term. Thus, the associated covariance matrix $(\hat{\mathbf{L}}\hat{\mathbf{L}}^\top)^{-1}$ behaves like a discretized elliptic Green's function and is therefore subject to an exponential screening effect (Schäfer et al., 2021a, Section 4.1). Let $\mathbf{P}^\updownarrow$ denote the permutation matrix that reverts the order of the degrees of freedom. Since $\mathbf{P}^\updownarrow\mathbf{V}^{-\top}\mathbf{P}^\updownarrow$ is lower triangular and

$$\mathbf{P}^\updownarrow\mathbf{\Sigma}\mathbf{P}^\updownarrow = \mathbf{P}^\updownarrow\mathbf{V}^{-\top}\mathbf{P}^\updownarrow\mathbf{P}^\updownarrow\mathbf{V}^{-1}\mathbf{P}^\updownarrow = \left(\mathbf{P}^\updownarrow\mathbf{V}^{-\top}\mathbf{P}^\updownarrow\right)\left(\mathbf{P}^\updownarrow\mathbf{V}^{-\top}\mathbf{P}^\updownarrow\right)^\top,$$

the matrix $\mathbf{P}^\updownarrow\mathbf{V}^{-\top}\mathbf{P}^\updownarrow$ is the Cholesky factor of $\mathbf{\Sigma}$ in the maximin (as opposed to the reverse maximin) ordering. In Schäfer et al. (2021b), it is shown that the Cholesky factors of discretized Green's functions of elliptic PDEs in the maximin ordering have exponentially decaying Cholesky factors. In particular, the results of Schäfer et al. (2021b) suggest that

$$\begin{aligned}\forall j \geq i:\ \log\left(\left|\left(\mathbf{P}^\updownarrow\mathbf{V}^{-\top}\mathbf{P}^\updownarrow\right)_{ji}\right|\right) &\lessapprox \log(n) - \operatorname{dist}(\mathbf{x}_j, \mathbf{x}_i)/\ell_i \\ \Rightarrow \forall j \geq i:\ \log\left(\left|\left(\mathbf{V}^{-1}\right)_{ji}\right|\right) &\lessapprox \log(n) - \operatorname{dist}(\mathbf{x}_j, \mathbf{x}_i)/\ell_j.\end{aligned}$$

As shown, for instance, in Schäfer et al. (2021b, Lemma 5.19), products of matrices that decay rapidly with respect to a distance function $\operatorname{dist}(\cdot,\cdot)$ on its index set, inherit this decay property. To this end, assume that lower triangular matrices $\mathbf{A}$ and $\mathbf{B}$ satisfy this property. We then have

$$\begin{aligned}\log\left(\left|(\mathbf{AB})_{ji}\right|\right) = \log\left(\left|\sum_k \mathbf{A}_{jk}\mathbf{B}_{ki}\right|\right) &\leq \log(n) + \log\left(\max_k |\mathbf{A}_{jk}\mathbf{B}_{ki}|\right) \\ &\lessapprox \log(n) - \max_k\left(\operatorname{dist}(\mathbf{x}_j, \mathbf{x}_k)/\ell_j - \operatorname{dist}(\mathbf{x}_j, \mathbf{x}_k)/\ell_k\right).\end{aligned}$$

By the triangle inequality, we have $\text{dist}(\mathbf{x}_j, \mathbf{x}_k) + \text{dist}(\mathbf{x}_k, \mathbf{x}_i) \geq \text{dist}(\mathbf{x}_j, \mathbf{x}_i)$. Since the right hand is $-\infty$ unless $j > i$ and thus $\ell_j \geq \ell_i$, we have thus

$$\log\left(\left|(\mathbf{AB})_{ji}\right|\right) = \log\left(\left|\sum_k \mathbf{A}_{jk}\mathbf{B}_{ki}\right|\right) \lessapprox \log(n) - \text{dist}(\mathbf{x}_j, \mathbf{x}_i)/\ell_j,$$

proving the the result.

For a general exponential family likelihood, the matrix $\mathbf{V}$ does not necessarily satisfy (10). Instead, according to Nickisch & Rasmussen (2008), a quadratic approximation to the log-likelihood under mild conditions implies that

$$\mathbf{V}\mathbf{V}^\top = \hat{\mathbf{L}}\hat{\mathbf{L}}^\top + \mathbf{W}^{-1},$$

where $\mathbf{W}$ is the covariance of the *effective likelihood* obtained by dividing the approximate posterior by the prior. Assuming that $\mathbf{W}^{-1}$ corresponds to a zero-order term in the context of a PDE, one can also obtain the result from the justification for the Gaussian likelihood case above. □

*Proof of Proposition 2.5.* Note that $p(\mathbf{f}^*|\mathbf{f}) = p(\tilde{\mathbf{f}})/p(\mathbf{f}) = \mathcal{N}_{n^*}\left(\boldsymbol{\mu}^* + \mathbf{K}^{*o}\mathbf{K}^{-1}(\mathbf{f} - \boldsymbol{\mu}), \mathbf{K}_{*|o}\right)$, where $\mathbf{K}_{*|o} = \mathbf{K}^{**} - \mathbf{K}^{*o}\mathbf{K}^{-1}\mathbf{K}^{o*}$, and $q(\mathbf{f}^*|\mathbf{f}) = q(\tilde{\mathbf{f}})/q(\mathbf{f}) = \mathcal{N}_{n^*}\left(\boldsymbol{\nu}^* - (\mathbf{V}^{**})^{-\top}\mathbf{V}^{o*\top}(\mathbf{f} - \boldsymbol{\nu}), (\mathbf{V}^{**}\mathbf{V}^{**\top})^{-1}\right)$. Then, since $\text{KL}\left(p(\mathbf{f}^*|\mathbf{f})\middle\|q(\mathbf{f}^*|\mathbf{f})\right)$ is a KL divergence between two Gaussian distributions, we have

$$2\,\text{KL}\left(p(\mathbf{f}^*|\mathbf{f})\middle\|q(\mathbf{f}^*|\mathbf{f})\right) = (\mathbf{G}\mathbf{f} + \mathbf{h})^\top(\mathbf{V}^{**}\mathbf{V}^{**\top})(\mathbf{G}\mathbf{f} + \mathbf{h}) + 2\,\text{KL}\left(\mathcal{N}_{n^*}\left(\mathbf{0}, \mathbf{K}_{*|o}\right)\middle\|\mathcal{N}_{n^*}\left(\mathbf{0}, (\mathbf{V}^{**}\mathbf{V}^{**\top})^{-1}\right)\right)$$

where $\mathbf{G} = -(\mathbf{V}^{**})^{-\top}\mathbf{V}^{o*\top} - \mathbf{K}^{*o}\mathbf{K}^{-1}$ and $\mathbf{h} = \boldsymbol{\nu}^* + (\mathbf{V}^{**})^{-\top}\mathbf{V}^{o*\top}\boldsymbol{\nu} - \boldsymbol{\mu}^* + \mathbf{K}^{*o}\mathbf{K}^{-1}\boldsymbol{\mu}$. Using the fact that the first term is quadratic in form, one can show that

$$\mathop{\mathbb{E}}_{p}\left[(\mathbf{G}\mathbf{f} + \mathbf{h})^\top(\mathbf{V}^{**}\mathbf{V}^{**\top})(\mathbf{G}\mathbf{f} + \mathbf{h})\right] = (\mathbf{G}\boldsymbol{\mu} + \mathbf{h})^\top(\mathbf{V}^{**}\mathbf{V}^{**\top})(\mathbf{G}\boldsymbol{\mu} + \mathbf{h}) + tr\Big((\mathbf{V}^{**}\mathbf{V}^{**\top})(\mathbf{G}\mathbf{K}\mathbf{G}^\top)\Big).$$

Then, we can see that $\text{KL}\left(p(\mathbf{f}^*|\mathbf{f})\middle\|q(\mathbf{f}^*|\mathbf{f})\right)$ is minimized with respect to $\boldsymbol{\nu}^*$ by $\mathbf{G}\boldsymbol{\mu} + \mathbf{h} = \mathbf{0}$. This implies that $\hat{\boldsymbol{\nu}}^* = \boldsymbol{\mu}^* - (\mathbf{V}^{**})^{-\top}\mathbf{V}^{o*\top}(\boldsymbol{\nu} - \boldsymbol{\mu})$. Plugging this in, we have

$$\begin{aligned}\underset{\mathbf{V}^* \in \mathcal{S}^*}{\arg\min}\, \mathop{\mathbb{E}}_{p}\left[\text{KL}\left(p(\mathbf{f}^*|\mathbf{f})\middle\|q(\mathbf{f}^*|\mathbf{f})\right)\right] &= \underset{\mathbf{V}^* \in \mathcal{S}^*}{\arg\min}\left[tr\Big(\mathbf{V}^{*\top}\tilde{\mathbf{K}}\mathbf{V}^*\Big) - \log\det(\mathbf{V}^{**}\mathbf{V}^{**\top})\right] \\ &= \underset{\mathbf{V}^* \in \mathcal{S}^*}{\arg\min}\sum_{i=1}^{n^*}\Big(\mathbf{V}^*_{\mathcal{S}^*_i,i}{}^\top\tilde{\mathbf{K}}_{\mathcal{S}^*_i,\mathcal{S}^*_i}\mathbf{V}^*_{\mathcal{S}^*_i,i} - 2\log\mathbf{V}^*_{i,i}\Big)\end{aligned}$$

Taking the first derivative of the summation with respect to the column vector $\mathbf{V}^*_{\mathcal{S}^*_i,i}$ and setting it to zero, one can show that $\hat{\mathbf{V}}^*_{\mathcal{S}^*_i,i} = \tilde{\mathbf{K}}^{-1}_{\mathcal{S}^*_i,\mathcal{S}^*_i}\mathbf{e}_1/\mathbf{V}^*_{i,i}$. Since $\mathbf{V}^*_{i,i}$ is the first entry of $\hat{\mathbf{V}}^*_{\mathcal{S}^*_i,i}$, we can have $\hat{\mathbf{V}}^*_{\mathcal{S}^*_i,i} = \mathbf{c}_i/\sqrt{\mathbf{c}_{i,1}}$ where $\mathbf{c}_i = \tilde{\mathbf{K}}^{-1}_{\mathcal{S}^*_i,\mathcal{S}^*_i}\mathbf{e}_1$. From the definition of $\mathcal{S}^*_i$, it can be easily shown that $\tilde{\mathbf{K}}^{-1}_{\mathcal{S}^*_i,\mathcal{S}^*_i} = K(\mathcal{S}^*_i, \mathcal{S}^*_i)^{-1}$. □