Posterior contraction rates for banded approximations of Gaussian processes

Omar Kassi
BIDSA, Department of Decision Sciences
Bocconi University
Milano, MI, Italy
omar.kassi@unibocconi.it

Bernhard Stankewitz
Department of Mathematics
University of Potsdam
Potsdam, Germany
bernhard.stankewitz@uni-potsdam.de

Botond Szabo
BIDSA, Department of Decision Sciences
Bocconi University
Milano, MI, Italy
botond.szabo@unibocconi.it

Abstract

Gaussian process regression models are widely used in modern statistics and machine learning due to their flexibility, interpretability, built-in uncertainty quantification, and strong theoretical foundations. However, training Gaussian processes (GPs) has a computational cost of $O(n^3)$, while the memory and prediction costs are also $O(n^2)$. For large datasets commonly encountered in practice, this becomes infeasible and necessitates the use of approximate posterior methods. In this work, we propose a class of approximations based on sparse covariance matrices, where the kernel matrix is sparsified according to the spatial proximity of the design points. We analyze the accuracy of the resulting approximate posterior distribution and show that the proposed approach preserves optimal statistical inference properties, provided that the dependencies among a sufficient number of neighboring points for each feature are retained. We derive explicit sufficient and necessary conditions on the number of neighbors required in a general setting and apply these results to standard kernels, including the squared exponential and Matérn kernels. Our theoretical findings are supported by numerical experiments.

1 Introduction

Gaussian Processes (GPs) form arguably one of the standard and most popular Bayesian nonparametric approaches for regression and classification tasks. They have a broad spectrum of applications, ranging from disease risk mapping in the Malaria Atlas Project (2) through understanding properties of molecules and chemicals (7) to aircraft designs (1). Their high popularity stems from the built-in uncertainty quantification, their effectiveness even with small data sizes, their flexibility, and improved interpretability compared to black box procedures.

In this paper, we focus on the random design nonparametric regression model, in which we observe pairs of random variables $(X_i, Y_i)_{i=1,\dots,n}$ satisfying

$$Y_i = f(X_i) + \varepsilon_i, \quad \text{with} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d., } X_i \sim G \text{ i.i.d.}, \qquad i = 1, \dots, n, \tag{1}$$

where $f : \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subset \mathbb{R}^d$, denotes the unknown regression function of interest and G denotes the distribution of each design point. Also $\varepsilon = (\varepsilon_i)_{i \le n}$ and $X = (X_i)_{i \le n}$ are independent. In the Bayesian approach, f is endowed with a flexible nonparametric prior. Gaussian Processes provide a

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

convenient and highly efficient choice for the prior measure. Due to conjugacy, the corresponding posterior is also a Gaussian Process and can be explicitly derived, see e.g. (18). Denoting by $k(\cdot,\cdot)$ the covariance kernel of a centered GP used as the prior, the mean and covariance functions of the corresponding posterior take the form

$$\begin{aligned} x &\mapsto k(X,x)^\top (K+\sigma^2 I_n)^{-1} Y, && x \in \mathcal{X}, \\ (x,x') &\mapsto k(x,x') - k(X,x)^\top (K+\sigma^2 I_n)^{-1} k(X,x'), && x,x' \in \mathcal{X}, \end{aligned} \tag{2}$$

where $K := (k(X_i,X_j))_{i,j=1,..,n} \in \mathbb{R}^{n\times n}$ denotes the empirical prior covariance matrix, I_n the n-dimensional identity matrix, $Y = (Y_1,\dots,Y_n)^\top \in \mathbb{R}^n$ the observation vector and $k(X,x) = (k(X_1,x),\dots,k(X_n,x))^\top \in \mathbb{R}^n$. Note that the inversion of the matrix $K_\sigma := K + \sigma^2 I_n \in \mathbb{R}^{n\times n}$ has computational complexity $O(n^3)$. Hence the computation of the GP posterior has training cost of $O(n^3)$, memory cost $O(n^2)$ and prediction cost $O(n^2)$. These render the computation infeasible or at least undesirable for large sample sizes. Since in modern applications the number of available data can be extensive this hinders the practical applicability of GPs.

To overcome this issue often scalable alternatives are considered, where the GP posterior (or prior) is approximated with a low rank/sparse process. There is a wide range and ever increasing number of such approaches, including distributed GPs (13; 11; 28; 20), variational Bayes methods using inducing variables (27), covariance tapering (10; 15; 23), and Vecchia approximation (31; 5; 14). These methods typically rely on carefully tuning various sparsity, low-rank, and localization parameters, trading off accuracy against computational complexity. The theoretical understanding of these approaches and the optimal tuning has been investigated in recent years, e.g. for distributed GPs (25; 12; 24), Vecchia GPs (26) variational GPs (4; 16), but our fundamental understanding is still limited.

In our paper, we consider a novel approximation method inducing sparsity to the prior covariance matrix K and hence substantially speeding up the inversion of the matrix and as a consequence reducing the cost of training, prediction and storage.[1] To preserve the statistical properties of the posterior and minimize information loss, for each feature X_i, we retain its dependence on its closest neighbors within a radius $r > 0$ while setting all other entries of the covariance matrix K to zero. The idea of preserving the information available in the closest neighbours while discarding the less relevant peripheral data has been widely used in other approximation methods in various manners. For instance, in the nearest neighbour Vecchia approximation, the precision matrix of the prior was sparsified in a similarl way, in covariance tapering the covariance matrix is multiplied by a sparse, compactly supported correlation function, forcing small, distant covariances to zero while maintaining a positive definite matrix. While our approach is closely connected to the latter one it has substantial algorithmic and conceptual differences, see the next section discussing them in detail.

Our paper investigates both the methodological and theoretical aspects of the proposed approximation approach. We derive general theoretical guarantees and limitations of the method for statistical inference, focusing on recovering the functional parameter of interest in the regression model and on the approximation accuracy of the posterior. We connect our results to the general theory of nonparametric Bayesian methods using Gaussian Process priors and obtain sufficient and necessary conditions for the approximation to maintain the same level of accuracy as the original posterior while substantially reducing the computational complexity. Our general, abstract results are then applied to standard covariance kernels revealing the trade-off between computational speed and approximation accuracy in concrete examples. Our theoretical results then can be used as a guideline in practice for tuning the approximations in an optimal manner instead of following an ad hoc empirical approach.

Organization. In Section 2 we introduce the approximation approach rigorously and discuss its methodological aspects. We provide general theoretical guarantees for approximation accuracy and statistical inference in Section 3 and apply these results for the popular squared exponential and Matérn kernels in Section 4. Our theoretical findings are demonstrated with numerical examples considering synthetic data sets. The proof of the results together with additional examples and numerical studies are deferred to the appendices material.

Notations. For two positive sequences (a_n) and (b_n), we write $a_n \lesssim b_n$ if there exists a constant $C > 0$, independent of n, such that $a_n \le C b_n$ for all n. We further write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and

[1] While writing up our manuscript we learned about the recently published conference paper (8) proposing a similar method in case of squared exponential covariance kernel. The authors investigate some practical aspects of this procedure, but no statistical guarantees were derived and extension to other GP kernels were also not considered.

$b_n \lesssim a_n$ hold. When needed, we explicitly use constants $c, C > 0$ that do not depend on n and may vary from line to line. For a function $f \in L^2(G)$, its evaluation at the design points $(X_1, \dots, X_n)$ is denoted by $\mathbf{f} := (f(X_1), \dots, f(X_n)) \in \mathbb{R}^n$. Furthermore, we denote by $K = k(X, X) \in \mathbb{R}^{n \times n}$ the empirical covariance matrix and by $K_\sigma = K + \sigma^2 I_n \in \mathbb{R}^{n \times n}$ the matrix appearing the in GP posterior. For two probability distributions $\mathbb{P}$, $\mathbb{Q}$, $\mathrm{KL}(\mathbb{P}, \mathbb{Q}) = \int \log(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q})\mathrm{d}\mathbb{P}$ denotes the Kullback-Leibler (KL) divergence. Given a probability measure $\mathbb{P}$, the distribution of a random variable X under $\mathbb{P}$ is denoted by $\mathbb{P}_X$ and $\mathbb{E}_X$ denotes the associated expectation. Likewise, for another random variable Y, the conditional distributions of $Y \mid X$ and $Y \mid X = x$ are denoted by $\mathbb{P}^{Y|X}$ and $\mathbb{P}^{Y|X=x}$, respectively.

2 Algorithms

For many commonly used kernels, correlations between observations decay rapidly as the distance between design points increases. This suggests that entries of the kernel matrix corresponding to distant points contribute little to the posterior and may be neglected. Therefore, a straightforward idea to reduce the computational complexity of GP regression is to sparsify the kernel matrix K by setting small correlations to zero when design points are far apart. This leads to the following approximation $\widehat{K} = (\widehat{K})_{i,j \le n}$ of K with

$$\widehat{K}_{ij} = \begin{cases} k(X_i, X_j), & \|X_i - X_j\| \le r, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $r > 0$ is a fixed radius. In dimension $d = 1$, when the design points can be assumed to be ordered on the real line, this construction reduces to a banded matrix with maximal bandwidth $m = m_r$ depending on r. In higher dimensions, where no natural ordering is available, the approach is more closely related to covariance tapering methods (10). Covariance tapering typically constructs a new kernel by multiplying the original kernel with a compactly supported function, and then performs full Bayesian inference under this modified prior. In contrast, our perspective retains the original kernel as the underlying model and views $\widehat{K}$ as a computational approximation of K. As a consequence, the resulting posterior should be interpreted as an approximation to the true posterior rather than the exact posterior under a modified prior. This viewpoint is particularly useful when analyzing how the sparsity parameter, i.e. the radius r, should scale with the sample size. Defining $\widehat{K}_\sigma := \widehat{K} + \sigma^2 I_n$, we obtain the approximate posterior distribution $\Psi_r(\cdot \mid X, Y)$, whose mean and covariance functions are given by

$$x \mapsto k(X, x)^\top \widehat{K}_\sigma^{-1} Y, \qquad (x, x') \mapsto k(x, x') - k(X, x)^\top \widehat{K}_\sigma^{-1} k(X, x'), \qquad x, x' \in \mathcal{X}. \tag{4}$$

Ψ_r is a well defined posterior as long as $\|\widehat{K} - K\|_{\mathrm{op}} < \sigma^2$, see Appendix D.1.

Setting, analogously to the one-dimensional case, $m = m_r$ equal to the maximal number of non-zero entries in a row of $\widehat{K}$, storing $\widehat{K}_\sigma$ scales as $\mathcal{O}(mn)$ in terms of memory compared to $\mathcal{O}(n^2)$ for the full kernel matrix. Importantly, the full matrix K does not need to be stored explicitly; only local kernel evaluations are required to construct $\widehat{K}$ and to evaluate terms such as $k(X, x)$.

The main advantage of the sparsified matrix $\widehat{K}_\sigma$ is that it contains at most $\mathcal{O}(mn)$ nonzero entries. This structure is particularly well-suited for iterative linear solvers such as the conjugate gradient (CG) algorithm. Each iteration of CG only requires matrix-vector multiplications of the form $\widehat{K}_\sigma u$, along with vector operations and inner products. When $\widehat{K}_\sigma$ is dense, such multiplications cost $\mathcal{O}(n^2)$ operations. In contrast, the sparsity of $\widehat{K}_\sigma$ reduces this cost to $\mathcal{O}(mn)$, since only the nonzero entries contribute. Then, the total computational cost of solving the system is therefore $\mathcal{O}(mn \cdot N_{\mathrm{iter}})$, where N_{iter} is the number of iterations required for convergence.

A key question is how to choose the parameter $r > 0$. Larger values of r improve the approximation of the true kernel matrix but also increase computational cost, while smaller values enhance scalability at the expense of approximation accuracy. In particular, one expects that $m = m_r$ should grow with n to ensure that the approximation error vanishes asymptotically, while remaining as small as possible to retain computational gains. This interplay between r, m and n will be reflected in the theoretical results developed in subsequent sections and analyzed also numerically in a range of examples.

3 Main results

We now provide theoretical guarantees for the approximation method described in the previous section. In our theoretical analysis, we consider the asymptotic regime where the sample size n tends to infinity. The finite sample behaviour of the procedure is investigated numerically in the next sections. We derive how well the proposed scalable alternative approximates the original posterior distribution and how much it preserves its statistical inference properties. To analyze the accuracy of the method, we consider the standard frequentist framework where we assume that there exists a deterministic true regression function f_0 in the model from Equation (1). Then, we derive how fast the approximate posterior recovers the underlying function f_0 as the sample size increases. This is measured via the speed of contraction of the approximate posterior around f_0.

First we consider the approximation accuracy of the proposed scalable alternative Ψ_r to the GP posterior $\Pi_n(\cdot|X,Y)$. The geometry of the prior Gaussian Process W is described by the Reproducing Kernel Hilbert Space (RKHS) $\mathbb{H}$ associated to its covariance kernel k. Indeed the RKHS $\mathbb{H}$ of W is obtained simply as the completion of the linear span of functions $\{k(x,\cdot) : x \in \mathcal{X}\}$ endowed with the inner product

$$\Big\langle \sum_{i=1}^{p} a_i k(t_i,\cdot), \sum_{j=1}^{q} b_j k(s_j,\cdot) \Big\rangle_{\mathbb{H}} = \sum_{i=1}^{p}\sum_{j=1}^{q} a_i b_j k(t_i,s_j), \tag{5}$$

for arbitrary $a \in \mathbb{R}^p, b \in \mathbb{R}^q, t_1, .., t_p, s_1, ..., s_q \in \mathcal{X}$. The associated RKHS norm is then defined as $\|g\|_{\mathbb{H}}^2 = \langle g, g\rangle_{\mathbb{H}}$ for $g \in \mathbb{H}$. For detailed introduction to RKHSs and their analytical properties we refer the reader to Section D.2 in the Appendix. Then, the statistical properties of the posterior corresponding to the GP priors is encoded in the behaviour of the so called concentration function

$$\varphi_{f_0}(\eta) = \inf_{h\in\mathbb{H}:\, \|h-f_0\|<\eta} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}\big(\|W\| < \eta\big), \tag{6}$$

see (29). The first term determines the approximation accuracy of the function f_0 with the RKHS. This term can be associated to the bias or approximation error of the procedure. The second term in the previous display is the log small ball probability. This term can be associated to the spread of the Bayesian procedure. Then, in view of (29), the posterior corresponding to the GP prior contracts around the true parameter f_0 with rate η_n, if the inequality

$$\varphi_{f_0}(\eta_n) \le C_\varphi n\eta_n^2 \tag{7}$$

holds for some constant $C_\varphi > 0$. Typically the optimal contraction rate can be expressed as the fastest rate η_n that satisfies the previous inequality. This is a standard condition describing the statistical inference properties of the GP posterior and hence is a natural assumption in our case as well.

Next we express the approximation error of the proposed scalable approach of the original GP posterior with respect to the estimation error between the sparsified covariance matrix $\widehat{K}$ and the true covariance matrix K. This latter quantity naturally appears in the approximation error and can be controlled based on the spectral and spatial properties of the prior covariance kernel k.

Theorem 3.1 (Kullback-Leibler bound)**.** *Assume that the concentration function satisfies condition* (7) *for some sequence η_n. Moreover, suppose that the threshold radius $r = r_n$ is chosen such that*

$$\|\widehat{K} - K\|_{\mathrm{op}} \le \bar{c} < \sigma^2/2 \tag{8}$$

with $\mathbb{P}_{f_0}^{\otimes n}$-probability converging to one. Then there exists a constant $\mathfrak{C} > 0$ depending on σ^2, d, and $\bar{c}$ such that, for any sequence $M_n \to \infty$, the approximate posterior Ψ_{r_n} satisfies

$$\mathrm{KL}(\Psi_{r_n}, \Pi_n(\cdot \mid X, Y)) \le \mathfrak{C}\, M_n \|\widehat{K} - K\|_{\mathrm{op}}^2 \left\{\mathrm{tr}(KK_\sigma^{-1}) + n\eta_n^2\right\}, \tag{9}$$

with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$ as $n \to \infty$.

The above theorem, which is proven in Appendix A, shows that in case that $\|\widehat{K} - K\|_{\mathrm{op}}^2 = o\big((\mathrm{tr}(KK_\sigma^{-1}) + n\eta_n^2)^{-1}\big)$ with probability converging to one, the Kullback-Leibler divergence between the approximate and true posteriors tends to zero. The operator norm $\|\widehat{K} - K\|_{\mathrm{op}}$ decreases as the threshold radius $r > 0$ in the approximation $\widehat{K}$ increases. The exact decay depends both on the

spatial and spectral properties of the covariance kernel k. We will investigate this interplay in case of several examples, including both polynomially and exponentially decaying eigenvalues and various spatial behaviors of the covariance kernel.

Controlling the Kullback-Leibler approximation error of the posterior provides us the means to derive similar contraction rate results for the approximation Ψ_{r_n} as the original posterior $\Pi_n(\cdot|X,Y)$, see Theorem 5 of (19). Formally, the sequence $(\Psi_{r_n})_{n\in\mathbb{N}}$ of distributions on $L^2(G)$ contracts around the truth f_0 with respect to a metric d if, for arbitrary sequence $M_n \to \infty$

$$\Psi_{r_n}(d(\cdot, f_0) \geq M_n\eta_n) \to 0 \qquad \text{in probability under } \mathbb{P}_{f_0}^{\otimes n} \text{ as } n \to \infty. \tag{10}$$

Intuitively, this means that asymptotically, Ψ_{r_n} does not put any mass outside of a ball of size $O(\eta_n)$. In the following, we will consider the metric $d = d_H$ given by

$$d_H(f,g)^2 = \int (\sqrt{p_f} - \sqrt{p_g})^2 dG \otimes \lambda, \qquad f, g \in L^2(G), \tag{11}$$

which is induced by the Hellinger distance between the two Gaussian densities p_f and p_g with $p_f(x,y) = \exp(-(y-f(x))^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$. The next theorem provides contraction rates for the approximate posterior Ψ_{r_n}. Its proof is in Appendix A.

Theorem 3.2 (Contraction rates)**.** *In the situation of Theorem 3.1, assume that additionally* $\operatorname{tr} KK_\sigma^{-1} = O_{\mathbb{P}_{f_0}}(n\eta_n^2)$*. Then, for any sequence* $M_n \to \infty$*, the approximate posterior satisfies*

$$\Psi_{r_n}(d_H(\cdot, f_0) \geq M_n\eta_n) \to 0 \tag{12}$$

in probability under $\mathbb{P}_{f_0}^{\otimes n}$ *as* $n \to \infty$*.*

Note that the condition $\operatorname{tr}(KK_\sigma^{-1}) = O_{\mathbb{P}_{f_0}}(n\eta_n^2)$ ensures that the Kullback-Leibler divergence between the approximation and the true posterior is of the order $O_{\mathbb{P}_{f_0}}(\|\widehat{K} - K\|_{\mathrm{op}}^2 n\eta_n^2)$. Then, under the assumption $\|\widehat{K} - K\|_{\mathrm{op}} < \sigma^2/2$, the Kullback-Leibler divergence is bounded by a multiple of $n\eta_n^2$, which in view of (19), is sufficient to derive the contraction rate η_n for Ψ_{r_n}. Note that to obtain this rate, it is not required that the approximation error tends to zero in Kullback-Leibler divergence. In fact, it can diverge to infinity. We discuss this disparity phenomenon in our examples. In the next section, we demonstrate the above general results on concrete examples, providing both theoretical and numerical justifications.

4 Examples

In this section, we present two explicit examples and provide a guide for selecting the neighbourhood size r in the approximation. The derived bounds depend on the smoothness of the underlying regression function f_0, the regularity of the GP, the spatial decay of the covariance kernel, and the dimension d of the process. To make this precise, we recall two classical smoothness classes. The Hölder space $C^\alpha(\mathbb{R}^d)$, for $\alpha > 0$, consists of functions whose partial derivatives up to order $\lfloor\alpha\rfloor$ exist and are bounded, and whose derivatives of order $\lfloor\alpha\rfloor$ are Hölder continuous with exponent $\alpha - \lfloor\alpha\rfloor$. The Sobolev space $H^\alpha(\mathcal{X})$ is defined as the class of restrictions $h|_{\mathcal{X}}$ of functions $h : \mathbb{R}^d \to \mathbb{R}$ whose Fourier transform $\hat{h}$ satisfies

$$\int_{\mathbb{R}^d} (1 + \|y\|^2)^\alpha |\hat{h}(y)|^2 \, dy < \infty.$$

It is well known that, in contrast to regular parametric models, in a nonparametric setting the choice of the prior plays a crucial role in the asymptotic behaviour of the posterior. Centered Gaussian processes described by their covariance kernel typically involve one or more tuning parameters. These parameters can determine the smoothness of the process (e.g., in the case of Matérn, Riemann–Liouville, or fractional Brownian motion kernels) or rescale the temporal or spatial behavior of the process. For example, set

$$W^{\tau_n}(x) = W(x/\tau_n), \quad x \in \mathcal{X},$$

where τ_n is a time re-scaling parameter possibly depending on the sample size. Then, one can tune the scale parameter τ_n to balance the bias and variance terms of the process, thereby preventing over- and underfitting.

Below, we consider the rescaled version of two standard GPs, the squared exponential and Matérn processes. We investigate the approximation error of the corresponding posteriors based on our approach, derive contraction rate results and study numerically the small to moderate sample size behaviour of the approximation procedure.

4.1 The squared exponential kernel

In this section, we focus on the rescaled squared exponential (SE) kernel in dimension d:

$$k_\tau(x, x') = \exp\big(-\tau^{-2}\|x' - x\|^2\big), \qquad x, x' \in \mathbb{R}^d, \tag{13}$$

where $\|\cdot\|$ denotes the Euclidean norm. For $\tau = 1$ the sample paths from the SE kernel are almost surely analytical. Infinitely smooth GPs are not suitable to approximate a less regular regression function $f_0 \in C^\alpha$, $\alpha > 0$, as the bias of the procedure will be sub-optimally large. Therefore, we consider a rescaled version of the SE GP with scale parameter taken to be $\tau = \tau_n = n^{-1/(2\alpha+d)}$. This optimal choice will shrink the sample paths of the SE GP and allows for estimating α-smooth functions f_0, see for instance (6) for 1-dimensional case and Lemma 8 of (16) for the d-dimensional case. We work with this optimally rescaled process and approximate it via the proposed nearest neighbour technique. We show that, by retaining a sufficiently large number of neighbours, our approximation is accurate for the posterior with the KL divergence tending to zero. Furthermore, the procedure preserves optimal statistical performance: the resulting posterior achieves minimax-optimal contraction up to a slowly varying term.

Proposition 4.1 (Result for the squared exponential kernel)**.** *In model* (1)*, assume that the design distribution G has a centered Gaussian Lebesgue density g on $\mathcal{X} = \mathbb{R}^d$ and that $f_0 \in C^\alpha(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ for some $\alpha > d/2$. Then for the squared exponential kernel k_τ with $\tau = \tau_n = n^{-1/(2\alpha+d)}$, the approximate posterior with truncation radius $r = r_n = \sqrt{2\tau_n^2 \log n}$ satisfies that for arbitrary $M_n \to \infty$, both*

$$\Psi_{r_n}(d_H(\cdot, f_0) \geq M_n n^{-\alpha/(d+2\alpha)} \log^{2d} n) \to 0 \quad \text{and} \quad \mathrm{KL}(\Psi_{r_n}, \Pi_n(\cdot|X, Y) \to 0 \tag{14}$$

in probability with respect to $\mathbb{P}_{f_0}^{\otimes n}$ for $n \to \infty$. Further, the number of non-zero entries per row of $\widehat{K}$ is at most of order $m_n = n r_n^d \asymp n^{2\alpha/(2\alpha+d)} \log^{d/2} n$ with probability converging to one.

Proposition 4.1, which is proven in Appendix B, shows that if the truncation radius is as above then the approximate posterior attains the optimal contraction rate, and the KL divergence between the approximate and true posteriors converges to zero, i.e., the approximate posterior shares the essential statistical properties of the true posterior. Regarding the computational complexity, the result guarantees that any row of the matrix $\widehat{K}$ contains at most of order $m_n = n r_n^d$ non-zero entries. In the context of the discussion in Section 2, this reduces storage to $O(m_n n)$ and the computational cost to $O(m_n n N_{\text{iter}})$.

Recall from Section 3 that to obtain the same contraction rate as the original posterior one does not require that the KL divergence between the two measures vanishes. In the present example, however, the gap between the optimal choices of r vanishes (up to log factors), see also Corollary 4.2. This is due to the squared exponential tail decay of the covariance kernel $r_{\tau_n}(\|t\|)$ resulting in a substantial increase in the accuracy of the covariance matrix approximation $\widehat{K}$ in operator norm when increasing r. Furthermore, note that as α increases to infinity, the exponent $2\alpha/(2\alpha + d)$ approaches 1, prohibiting any meaningful banding. This reflects the fact that smoother functions exhibit stronger long-range dependence, and thus require incorporating more points to be accurately approximated. On the other hand, as the dimension d increases, the exponent decreases, which indicates that fewer neighbors are needed. Next we show that the required size of the truncation radius in Proposition 4.1 is sharp.

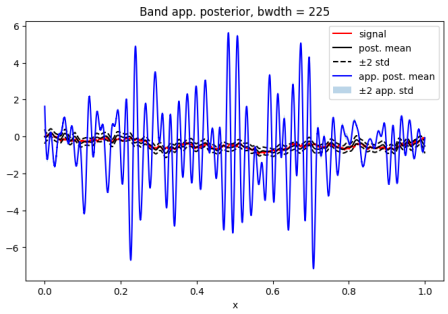
Corollary 4.2 (Lower bound for the squared exponential kernel)**.** *In model* (1)*, assume that the design distribution G is the standard Gaussian distribution with density g on $\mathcal{X} = \mathbb{R}^d$ and let $\alpha > d/2 + 1$. Then, for the squared exponential kernel k_τ with $\tau = \tau_n = n^{-1/(2\alpha+d)}$, the truncated kernel matrix $\widehat{K}$ with truncation radius r smaller than $r_n = \tau_n / \log^{1/(2d)} n$ satisfies that*

$$\inf_{r < r_n} \sup_{f_0 \in C^\alpha(\mathbb{R}^d)} \|\mathbb{E}_\varepsilon K \widehat{K}_\sigma^{-1} Y - f_0\|_n \gtrsim \sqrt{\log n}\, n^{-\frac{d/2}{2\alpha+d}}, \tag{15}$$

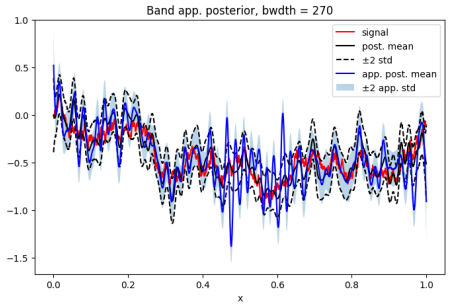
with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$.

The above corollary states that choosing even a logarithmic multiplicative factor less neighbours in the approximation, the posterior mean has sub-optimally large bias $n^{-d/(2(2\alpha+d))} \gg n^{-\alpha/(2\alpha+d)}$, for $\alpha > d/2 + 1$, with respect to the empirical L_2-norm, which results in a polynomial factor loss in the convergence. Its proof in Appendix B follows from a more general lower bound in Proposition C.6.

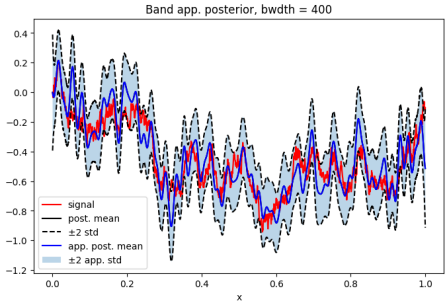
Next, we demonstrate the performance of the approximation method numerically on synthetic data. We take a draw from a Brownian path as the true function f_0 and set $\sigma = 1$ in the regression model from Equation (1). For simplicity, we illustrate our simulations based on the choice of the number of neigbours m instead of the truncation radius r noting that essentially $m \asymp nr^d$. We generate $n = 10^4$ samples and consider the number neighbours m ranging from 225 to 400, see Figure 1. Note that for a too small choice of m the approximation is very poor, the process might not even be well defined (the corresponding covariance kernel is not positive definite), see Figure 1a. As m increases the resulting approximation becomes a valid Gaussian Process, but the approximation accuracy might not be acceptable, see Figure 1b. Then after reaching a threshold the approximating GP, see Figure 1c becomes almost indistinguishable from the true posterior and further increasing m does not improve the approximation but only adds to the computational complexity. The log-computation time of the true posterior and the approximation is reported in Figure 1d. The results are further illustrated in Table 1, which reports the average and standard deviation of the squared L_2-error over 100 draws for different bandwidths m.



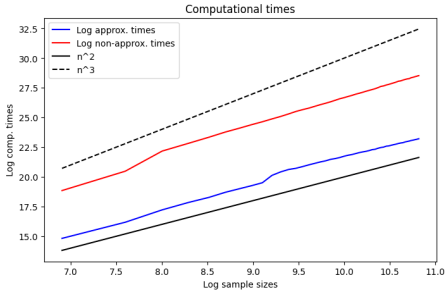
(a) Approx. posterior with insufficient bandwidth. Approx. posterior covariance is numerically not well defined.



(b) Approx. posterior with insufficient bandwidth. Approx. posterior is well defined but deviates from the true one.



(c) Approx. posterior with good bandwidth.



(d) Log-log plot of the computation times.

Figure 1: Squared exponential kernel simulations. The true function f_0 was set to a draw from a Brownian motion and the covariates are iid uniformly distributed. In the simulations $n = 10^4$, $\sigma = 1$, and $\tau_n = n^{-1/(2\alpha+d)}$ were taken and the bandwidth m was set to $225, 270$ and 400, respectively.

Bandwidth	225	270	400	10^4 (full posterior)
Avg. MSE	159	32	0.01	0.01
Std. Dev.	813	230	0.001	0.001

Table 1: Squared error of the true and approximate posterior based on SE kernel. Average and standard deviation over 100 repetition of the experiment.

4.2 The Matérn kernel

In this subsection, we discuss the Matérn kernel. The Matérn process is the centered GP with covariance kernel

$$k_{\nu,\tau}(x,x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\Big(\frac{\sqrt{2\nu}}{\tau}\|x'-x\|_2\Big)^{\nu} B_\nu\Big(\frac{\sqrt{2\nu}}{\tau}\|x'-x\|_2\Big), \quad x,x' \in \mathbb{R}^d, \tag{16}$$

where B_ν denotes the modified Bessel function of the second kind and $\Gamma(\nu)$ is the Gamma function. When ν is half-integer, i.e. $\nu = \ell + 1/2$, for $\ell \in \mathbb{N}$, the Matérn kernel has an explicit expression

$$k_{\nu,\tau}(x,x') = \exp\left(\frac{-\sqrt{2\ell+1}\|x'-x\|_2}{\tau}\right)\frac{\Gamma(\ell+1)}{\Gamma(2\ell+1)}\sum_{q=0}^{\ell}\frac{(\ell+q)!}{q!(\ell-q)!}\left(\frac{\sqrt{8\ell+4}\|x'-x\|_2}{\tau}\right)^{\ell-q}, \tag{17}$$

see Equation (4.16) in (18). The sample paths from the process are ν-Hölder smooth. A special case of the Matérn process with $\nu = 1/2$ is the Ornstein-Uhlenbeck, process also referred to as the GP with exponential kernel.

In the regression model, one can recover the unknown α-smooth function f_0 optimally by choosing a Matérn process with parameter $\nu = \alpha$. In case the regularity of the process is not matching the regularity of f_0 one has to rescale the process for optimal recovery. For $f_0 \in C^\alpha(\mathbb{R}^d)$ with $0 < \alpha \le \nu = \ell + 1/2$ the optimal rescaling is given by $\tau_n = n^{-(2\ell+1-2\alpha)/((2\alpha+d)(2\ell+1))}$, see for instance (9; 24). Below we consider the Matérn process prior with the above optimal rescaling and derive guarantees for the proposed approximation procedure.

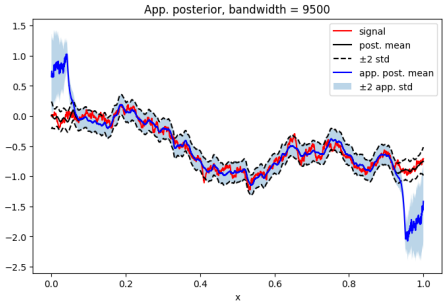
Proposition 4.3 (Result for the Matérn kernel)**.** *In model* (1)*, assume that $\mathcal{X} \subset \mathbb{R}^d$ is bounded and the design observations are distributed according to G on $\mathcal{X}$ with bounded Lebesgue density g. Let $f_0 \in C^\alpha(\mathcal{X}) \cap H^\alpha(\mathcal{X})$ for $\alpha > d/2$ and let k_{ν,τ_n} be the rescaled Matérn kernel with smoothness $\nu = \ell + 1/2$ for $\ell \in \mathbb{N}_0$ and lengthscale $\tau_n = n^{-(\nu-\alpha)/(2\alpha\nu+d\nu)}$. Then, the approximate posterior with truncation radius $r_n = 2\tau_n \log n$ satisfies that for arbitrary $M_n \to \infty$, both*

$$\Psi_{r_n}(d_H(\cdot,f_0) \ge M_n n^{-\alpha/(2\alpha+d)}) \to 0 \quad \textit{and} \quad \mathrm{KL}(\Psi_{r_n}, \Pi_n(\cdot|X,Y)) \to 0 \tag{18}$$

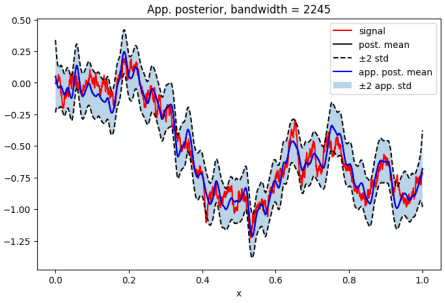
in probability with respect to $\mathbb{P}_{f_0}^{\otimes n}$ for $n \to \infty$. Further, the number of non-zero entries per row of $\widehat{K}$ is at most of order $nr_n^d \asymp n^{(2\alpha+\alpha d/\nu)/(2\alpha+d)}\log^d n$ with probability converging to one.

Proposition 4.3, which is proven in Appendix B, shows that if the truncation radius is chosen at least of order $n^{-(\nu-\alpha)/(2\alpha\nu+d\nu)}$, up to logarithmic factors, then the approximate posterior attains the optimal contraction rate, and the Kullback-Leibler divergence between the approximate and true posteriors converges to zero. Similarly to the squared exponential kernel, there is no substantial gap between the minimal choice of r_n required for vanishing approximation of the posterior measure in Kullback-Leibler divergence and obtaining the minimax optimal contraction rate. This is again due to the exponential decay of the Matérn kernel in the spatial domain. It is also worth emphasizing that the exponent $-(\nu-\alpha)/(2\alpha\nu+d\nu)$ is decreasing in ν and increasing in α. Consequently, smoother kernels (larger ν) allow for a smaller truncation radius r_n. Hence, considering smoother processes and rescaling them would allow to imply faster approximations. In contrast, as the smoothness α of the true function f_0 increases, a larger neighborhood size is required. Finally, as ν increases, the Matérn kernel approaches the squared exponential kernel, and the corresponding minimal choice of r_n becomes comparable to that obtained in the SE setting up to logarithmic factors. The same is true for the row-wise number $m_n \asymp nr_n^d$ of non-zero entries in $\widehat{K}$. Again, storage is reduced to $O(m_n n)$ and computational cost to $O(m_n n N_{\text{iter}})$.

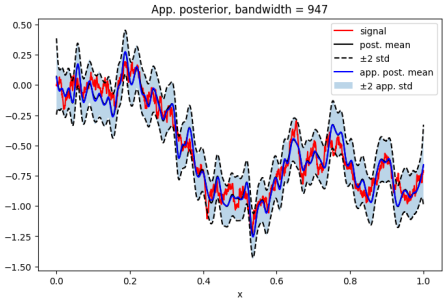
Next we investigate the numerical performance of the approximation over a synthetic data set. We study the influence of the regularity of the Matérn process and the optimal choice of the neighbourhood size used in the approximation. As in Section 4.1, we fix m explicitly instead of r. We set the true function f_0 to be a random draw of a Brownian motion, take sample size $n = 10^4$ and set $\sigma = 1$ in the regression model from Equation (1). In Figure 2, we consider the rescaled Matérn process with regularities $\nu = 1/2$, $\nu = 3/2$ and $\nu = 5/2$. For $\nu = 0.5$, the optimal rescaling is $\tau_n = 1$. Consequently, according to our theory structural computational gains via banding should not be possible. This is illustrated by the fact that even when keeping $m = 9500$ neighbours, the approximate and full posteriors are still substantially different. For $\nu = 1.5$ and the corresponding rescaling $\tau_n = n^{-1/3}$, we take $m = 2245$, the smallest neighbourhood size after which the approximation didn't improve further in our simulations (and became indistinguishable from the true posterior). Finally, for $\nu = 2.5$ with rescaling $\tau_n = n^{-2/5}$ this effect intensifies and the true and approximate posteriors visually coincide already at $m = 947$. The results are further confirmed by comparing the MSEs in Table (d).



(a) Regularity $\nu = 0.5$ with $\tau_n = 1$.



(b) Regularity $\nu = 1.5$ with $\tau_n = n^{-1/3}$.



(c) Regularity $\nu = 2.5$ with $\tau_n = n^{-2/5}$.

(ν, bwdth)	$(0.5, 9500)$	$(1.5, 2245)$	$(2.5, 947)$
App. MSE	0.08	0.007	0.008
App. StD	0.5	0.001	0.001
True MSE	0.005	0.007	0.007
True StD	0.0007	0.001	0.001

(d) Average mean squared error of the approximate and true posteriors with their standard deviations over 100 draws.

Figure 2: Simulations for the Matérn kernel. The true function f_0 was set to a draw from a Brownian path. In the simulation $n = 10^4$ and $\sigma = 1$ was taken and the number of neighbours m was set to $9500, 2245$ and 947, respectively.

5 Discussion

The results for both the squared exponential and the Matérn kernels exhibit a notable alignment between statistical optimality and approximation accuracy. The order of the minimal truncation radius $r = r_n$ required to achieve the optimal contraction rate and the one ensuring vanishing Kullback-Leibler divergence do not seem to differ substantially. However, this phenomenon is not generic. As shown in Appendix D.3, for kernels with polynomial decay, the two requirements can differ at a polynomial scale in n, showing that optimal recovery of the underlying function can be achieved with lower computational complexity than the one required for accurate approximation of the posterior.

We note that as the dimension d increases, the required number of neighbours m_n in the approximation decreases in the above examples (for the Matérn case, this is true as long as $\nu > \alpha$). In contrast, many other approximation methods exhibit complementary behaviour: as d increases the scalability

of the procedures reduces. For instance, in the inducing variable variational Bayes method, one has to involve $n^{d/(d+2\alpha)}$ inducing variables (16; 17), the conjugate gradient descent and Lanczos approximations need to run for $O(n^{d/(d+2\alpha)})$ iterations (22), and in the Vecchia approximation one requires parent sets of size $\binom{\lfloor\alpha\rfloor+d}{d}$ see (26). At the same time as the regularity α of the true function f_0 increases, one needs to keep more neighbours hence resulting in a less scalable procedure.

It is worth noting that the proposed banding strategy fundamentally relies on the kernel exhibiting a notion of spatial decay, typically of the form $k(x, x') = r(\|x' - x\|)$, where r denotes a radial function. In this case, Euclidean proximity is aligned with correlation decay, which justifies the nearest-neighbour truncation. For general non-stationary kernels that do not admit such a representation, such as Brownian motion kernel, spatial distance may therefore be inappropriate.

Finally, the optimal choice of the truncation radius r depends on the unknown regularity α of the true function. Since this quantity is not available in practice, an important direction for future work is to develop adaptive procedures that select r in a data-driven manner, for instance using early stopping.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their helpful comments. Co-funded by the European Union (ERC, BigBayesUQ, project number: 101041064). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The research of Bernhard Stankewitz. has been partially funded by the Deutsche Forschungsgemeinschaft (DFG) – Project-ID 318763901-SFB1294.

References

[1] ARNSON, M. G., ALJABER, R., AND CINAR, G. Predicting aircraft design parameters using gaussian process regressions on historical data. In *AIAA SciTech 2025 Forum* (2025), p. 1287.

[2] BHATT, S., CAMERON, E., FLAXMAN, S. R., WEISS, D. J., SMITH, D. L., AND GETHING, P. W. Improved prediction accuracy for disease risk mapping using gaussian process stacked generalization. *Journal of The Royal Society Interface 14*, 134 (2017).

[3] BOUCHERON, S., LUGOSI, G., AND MASSART, P. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, Oxford, 2013.

[4] BURT, D. R., RASMUSSEN, C. E., AND VAN DER WILK, M. Rates of convergence for sparse variational Gaussian process regression. In *Proceedings of the 36th International Conference on Machine Learning* (2019), vol. 97 of *Proceedings of Machine Learning Research*, PMLR, pp. 862–871.

[5] 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 (2016), 800–812.

[6] DER VAART, A. V., AND VAN ZANTEN, H. Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat. 1* (2007), 433–448.

[7] DERINGER, V. L., BARTÓK, A. P., BERNSTEIN, N., WILKINS, D. M., CERIOTTI, M., AND CSÁNYI, G. Gaussian process regression for materials and molecules. *Chemical reviews 121*, 16 (2021), 10073–10141.

[8] EHRHARDT, E. C., AND TOBAR, F. Accelerated training of Gaussian processes using banded square exponential covariances. In *Proceedings of the 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2026), IEEE, pp. 5856–5860.

[9] FANG, X., AND BHADRA, A. Posterior concentration for Gaussian process priors under rescaled and hierarchical Matérn and confluent hypergeometric covariance functions. *Electronic Journal of Statistics 19*, 2 (2025), 3333–3369.

[10] FURRER, R., GENTON, M. G., AND NYCHKA, D. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics 15*, 3 (2006), 502–523.

[11] GRAMACY, R. B., AND LEE, H. K. H. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association 103*, 483 (2008), 1119–1130.

[12] GUHANIYOGI, R., LI, C., SAVITSKY, T. D., AND SRIVASTAVA, S. Distributed Bayesian varying coefficient modeling using a Gaussian process prior. *Journal of Machine Learning Research 23*, 1 (2022), 3642–3700.

[13] JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J., AND HINTON, G. E. Adaptive mixtures of local experts. *Neural Computation 3*, 1 (1991), 79–87.

[14] KATZFUSS, M., GUINNESS, J., GONG, W., AND ZILBER, D. Vecchia approximations of Gaussian-process predictions. *Journal of Agricultural, Biological and Environmental Statistics 25* (2020), 383–414.

[15] KAUFMAN, C. G., SCHERVISH, M. J., AND NYCHKA, D. W. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association 103*, 484 (2008), 1545–1555.

[16] NIEMAN, D., SZABÓ, B., AND VAN ZANTEN, H. Contraction rates for sparse variational approximations in gaussian process regression. *Journal of Machine Learning Research 23* (2022), 1–26.

[17] NIEMAN, D., SZABÓ, B., AND VAN ZANTEN, H. Uncertainty quantification for sparse spectral variational approximations in Gaussian process regression. *Electronic Journal of Statistics 17*, 2 (2023), 2250–2288.

[18] RASMUSSEN, C. E., AND WILLIAMS, C. K. I. *Gaussian processes for machine learning*. Cambridge, MA: MIT Press, 2006.

[19] RAY, K., AND SZABÓ, B. Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association 117*, 539 (2022), 1270–1281.

[20] SCOTT, S. L., BLOCKER, A. W., BONASSI, F. V., CHIPMAN, H. A., GEORGE, E. I., AND MCCULLOCH, R. E. Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management 11*, 2 (2016), 78–88.

[21] SEEGER, M. Addendum to: Information consistency of nonparametric Gaussian process methods, 2007. Technical report, Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

[22] STANKEWITZ, B., AND SZABÓ, B. Contraction rates for conjugate gradient and Lanczos approximate posteriors in Gaussian process regression, 2024.

[23] STEIN, M. L. Statistical properties of covariance tapers. *Journal of Computational and Graphical Statistics 22*, 4 (2013), 866–885.

[24] SZABÓ, B., HADJI, A., AND VAN DER VAART, A. W. Adaptation using spatially distributed Gaussian processes. *Journal of the American Statistical Association 120*, 552 (2025), 2784–2795.

[25] SZABÓ, B., AND VAN ZANTEN, H. An asymptotic analysis of distributed nonparametric methods. *Journal of Machine Learning Research 20*, 87 (2019), 1–30.

[26] SZABÓ, B., AND ZHU, Y. Vecchia Gaussian processes: on probabilistic and statistical properties. *The Annals of Statistics 54*, 3 (2026), 1289–1314.

[27] TITSIAS, M. Variational learning of inducing variables in sparse gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* (2009), vol. 5, PMLR, pp. 567–574.

[28] TRESP, V. The generalized Bayesian committee machine. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2000), Association for Computing Machinery, pp. 130–139.

[29] VAN DER VAART, A. W., AND VAN ZANTEN, J. H. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Stat. 36*, 3 (2008), 1435–1463.

[30] VAN DER VAART, A. W., AND VAN ZANTEN, J. H. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics 37*, 5B (2009), 2655–2675.

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

[32] WILLIAMS, C. K. I., AND SHAWE-TAYLOR, J. The stability of kernel principal components analysis and its relation to the process eigenspectrum. In *Advances in Neural Information Processing Systems* (2002), vol. 15.

A Proofs of the main results

Proof of Theorem 3.1. The proof is based on the technical bound in Proposition A.1. It suffices to bound the terms on the right-hand side with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$ as $n \to \infty$. From Markov's inequality, in this sense, we have $\|f_0 - h\|_n^2 \leq M_n \|f_0 - h\|_{L^2}^2$ for any sequence $M_n \to \infty$. Now, under the concentration function assumption in Equation (8), h can be choseen such that $\|h\|_{\mathbb{H}}^2 \lesssim n\eta_n^2$ and $\|f_0 - h\|_{L^2}^2 \lesssim \eta_n^2$. From there, we obtain

$$\frac{\|h\|_{\mathbb{H}}^2 \|E\|_{\text{op}}^2}{\sigma^2(\sigma^2 - \|E\|_{\text{op}})} + \frac{n\|f_0 - h\|_n^2 \|E\|_{\text{op}}^2}{\sigma^4(\sigma^2 - \|E\|_{\text{op}})} \leq \mathfrak{C} M_n \|E\|_{\text{op}}^2 n\eta_n^2. \tag{19}$$

Moreover,

$$\sum_{j=1}^{n} \Big(\frac{\widehat{\mu}_j}{\widehat{\mu}_j + \sigma^2}\Big)^2 \leq \sum_{j=1}^{n} \frac{\widehat{\mu}_j}{\widehat{\mu}_j + \sigma^2} = \operatorname{tr} K K_\sigma^{-1}, \tag{20}$$

which finishes the proof. ■

Proposition A.1 (Technical Kullback-Leibler bound)**.** *For any $h \in \mathbb{H}$ and any sequence $M_n \to \infty$, the approximate posterior satisfies*[2]

$$\begin{aligned} \mathrm{KL}(\Psi_{n,m}, \Pi(\cdot|\mathcal{D}^n)) \mathbf{1}\{\|E\|_{op} < \sigma^2/2\} \leq{}& \frac{2\|h\|_{\mathbb{H}}^2 \|E\|_{op}^2}{\sigma^2(\sigma^2 - \|E\|_{op})} + \frac{2n\|f_0 - h\|_n^2 \|E\|_{op}^2}{\sigma^4(\sigma^2 - \|E\|_{op})} \\ &+ \frac{M_n \|E\|_{op}^2}{\sigma^4} \sum_{j=1}^{n} \frac{\widehat{\mu}_j}{\widehat{\mu}_j + \sigma^2} + \frac{2\|E\|_{op}^2}{(\sigma^2 - 2\|E\|_{op})^2} \sum_{j=1}^{n} \Big(\frac{\widehat{\mu}_j}{\widehat{\mu}_j + \sigma^2}\Big)^2 \end{aligned} \tag{21}$$

with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$ and $n \to \infty$.

Proof. In view of Lemma D.1, $\mathrm{KL}(\Psi_{n,m}, \Pi(\cdot|\mathcal{D}^n)) = \mathrm{KL}(N_1, N_2)$, where $N_i \sim \mathcal{N}(m_i, \Sigma_i)$ with $m_1 = K\widehat{K}_\sigma^{-1} Y$, $m_2 = K K_\sigma^{-1} Y$, $\Sigma_1 = (K - K\widehat{K}_\sigma^{-1} K)$ and $\Sigma_2 = (K - K K_\sigma^{-1} K)$. N_1 and N_2 are always absolutely continuous with respect to each other. Indeed, when K is non-singular, both Σ_1 and Σ_2 are non-singular as well. Otherwise, a short calculation shows that $\operatorname{ran}\Sigma_1 = \operatorname{ran}\Sigma_2 = \operatorname{ran} K$ and absolute continuity follows from the fact that borth m_1 and m_2 are also elements of $\operatorname{ran} K$. Therefore, the Kullback-Leibler divergence between the finite dimensional Gaussians can be expressed as

$$\mathrm{KL}(N_1, N_2) = \frac{1}{2}\big(\underbrace{\operatorname{tr}\Sigma_2^{-1}\Sigma_1 - n}_{\text{(I)}} + \underbrace{(m_2 - m_1)^\top \Sigma_2^{-1}(m_2 - m_1)}_{\text{(II)}} + \underbrace{\log\det \Sigma_1^{-1}\Sigma_2}_{\text{(III)}} \big), \tag{22}$$

where "$(\cdot)^{-1}$" denotes the generalized inverse when K is singular. We deal with the three terms on the right hand side separately.

Step 1: bound for the mean term (II). By substitution and recalling that $Y = \mathbf{f}_0 + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\mathbf{f}_0 = (f(X_1), \ldots, f(X_n))^\top \in \mathbb{R}^n$, and the inequality $\|a + b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2$,

$$\begin{aligned} \text{(II)} &= (K(K_\sigma^{-1} - \widehat{K}_\sigma^{-1})Y)^\top (K - K K_\sigma^{-1} K)^{-1} (K(K_\sigma^{-1} - \widehat{K}_\sigma^{-1})Y) \\ &\leq 2(K(K_\sigma^{-1} - \widehat{K}_\sigma^{-1})\varepsilon)^\top (K - K K_\sigma^{-1} K)^{-1} (K(K_\sigma^{-1} - \widehat{K}_\sigma^{-1})\varepsilon) \\ &\quad + 2(K(K_\sigma^{-1} - \widehat{K}_\sigma^{-1})\mathbf{f}_0)^\top (K - K K_\sigma^{-1} K)^{-1} (K(K_\sigma^{-1} - \widehat{K}_\sigma^{-1})\mathbf{f}_0). \end{aligned} \tag{23}$$

[2] 2026-09-19-BS-TODO: Formualte conditions if any.

Now, we bound each term separately. In view of Markov's inequality for arbitrary $M_n \to \infty$ it holds with probability tending to one that

$$\begin{aligned}
&\varepsilon^\top (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) K (K - K K_\sigma^{-1} K)^{-1} K (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) \varepsilon \\
&\le M_n \mathbb{E}_\varepsilon \operatorname{tr} \varepsilon\varepsilon^\top (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) K (K - K K_\sigma^{-1} K)^{-1} K (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) \\
&= M_n \sigma^2 \operatorname{tr} (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) K (K - K K_\sigma^{-1} K)^{-1} K (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) \\
&= M_n \sigma^2 \operatorname{tr} \widehat{K}_\sigma^{-1} E K_\sigma^{-1} K (K - K K_\sigma^{-1} K)^{-1} K K_\sigma^{-1} E \widehat{K}_\sigma^{-1} \\
&= M_n \sigma^2 \| (K - K K_\sigma^{-1} K)^{-1/2} K K_\sigma^{-1} E \widehat{K}_\sigma^{-1} \|_F^2 \\
&\le M_n \sigma^2 \| (K - K K_\sigma^{-1} K)^{-1/2} K K_\sigma^{-1} \|_F^2 \|E\|_{\mathrm{op}}^2 \|\widehat{K}_\sigma^{-1}\|_{\mathrm{op}}^2 \\
&\le M_n \sigma^{-2} \|E\|_{\mathrm{op}}^2 \sum_{j=1}^n \Big\{ \frac{\hat\mu_j}{\hat\mu_j + \sigma^2} \times \Big(\hat\mu_j - \frac{\hat\mu_j^2}{\hat\mu_j + \sigma^2} \Big)^{-1/2} \Big\}^2 \\
&= M_n \sigma^{-4} \|E\|_{\mathrm{op}}^2 \sum_{j=1}^n \frac{\hat\mu_j}{\hat\mu_j + \sigma^2}.
\end{aligned} \tag{24}$$

For the second term in the bound (23), we split $\mathbf{f}_0 = (\mathbf{f}_0 - \mathbf{h}) + \mathbf{h}$, where for a function $h \in \mathbb{H}$, $\mathbf{h} = (h(X_1), \dots, h(X_n))^\top$. Then, noting that

$$K - K K_\sigma^{-1} K = K(I - K_\sigma^{-1} K) = \sigma^2 K K_\sigma^{-1} \tag{25}$$

and consequently

$$K (K - K K_\sigma^{-1} K)^{-1} K = \sigma^{-2} K (K K_\sigma^{-1})^{-1} K = \sigma^{-2} K K_\sigma, \tag{26}$$

we obtain

$$\begin{aligned}
&\mathbf{h}^\top (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) K (K - K K_\sigma^{-1} K)^{-1} K (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) \mathbf{h} \\
&= \sigma^{-2} \mathbf{h}^\top (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) K K_\sigma (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) \mathbf{h} \le \sigma^{-2} \|K_\sigma (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) \mathbf{h}\|^2 \\
&= \sigma^{-2} \|(I - K_\sigma \widehat{K}_\sigma^{-1}) \mathbf{h}\|^2 = \sigma^{-2} \|E \widehat{K}_\sigma^{-1} \mathbf{h}\|^2 \le \frac{\|E\|^2}{\sigma^2} \|\widehat{K}_\sigma^{-1} \mathbf{h}\|^2.
\end{aligned} \tag{27}$$

From $\widehat{K}_\sigma = K_\sigma + E$ it follows that $\widehat{K}_\sigma^{-1} \le (1 - \|E\|/\sigma^2)^{-1} K_\sigma^{-1}$ and therefore

$$\begin{aligned}
&\mathbf{h}^\top (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) K (K - K K_\sigma^{-1} K)^{-1} K (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) \mathbf{h} \\
&\le \frac{\|E\|_{\mathrm{op}}^2}{\sigma^2 - \|E\|_{\mathrm{op}}} \|K_\sigma^{-1} \mathbf{h}\|^2 \le \frac{\|E\|_{\mathrm{op}}^2}{\sigma^2 (\sigma^2 - \|E\|_{\mathrm{op}})} \mathbf{h}^\top K_\sigma^{-1} \mathbf{h} \le \frac{\|E\|_{\mathrm{op}}^2 \|h\|_{\mathbb{H}}^2}{\sigma^2 (\sigma^2 - \|E\|_{\mathrm{op}})},
\end{aligned} \tag{28}$$

where we have used that $\mathbf{h}^\top K_\sigma^{-1} \mathbf{h} \le \|h\|_{\mathbb{H}}^2$ in the last inequality. If K is non-singular, this follows from the fact that $\mathbf{h}^\top K^{-1} \mathbf{h}$ is the squared RKHS norm of the projection of h onto $\operatorname{span}\{k(X_i, \cdot) : i \le n\}$. Otherwise a simple additional calculation provides the same result.

Dealing with the term $\mathbf{f}_0 - \mathbf{h}$, by similar arguments,

$$\begin{aligned}
&(\mathbf{f}_0 - \mathbf{h})^\top (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) K (K - K K_\sigma^{-1} K)^{-1} K (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) (\mathbf{f}_0 - \mathbf{h}) \\
&\le \frac{n \|f_0 - h\|_n^2 \|E\|_{\mathrm{op}}^2}{\sigma^4 (\sigma^2 - \|E\|_{\mathrm{op}})},
\end{aligned} \tag{29}$$

where for a function g, $\|g\|_n^2 = n^{-1} \sum_{i=1}^n g(X_n)^2$. By combining the above displays,

$$\text{(II)} \le \frac{2 M_n \|E\|_{\mathrm{op}}^2}{\sigma^4} \sum_{j=1}^n \frac{\hat\mu_j}{\hat\mu_j + \sigma^2} + \frac{4 \|h\|_{\mathbb{H}}^2 \|E\|_{\mathrm{op}}^2}{\sigma^2 (\sigma^2 - \|E\|_{\mathrm{op}})} + \frac{4n \|f_0 - h\|_n^2 \|E\|_{\mathrm{op}}^2}{\sigma^4 (\sigma^2 - \|E\|_{\mathrm{op}})}. \tag{30}$$

Step 2: Decomposition of the log-determinant term (III). Note that

$$\begin{aligned}
&(K - K K_\sigma^{-1} K)^{-1} (K - K \widehat{K}_\sigma^{-1} K) \\
&= I + (K - K K_\sigma^{-1} K)^{-1} K (K_\sigma^{-1} - \widehat{K}_\sigma^{-1}) K \\
&= I + (K - K K_\sigma^{-1} K)^{-1} K K_\sigma^{-1} E \widehat{K}_\sigma^{-1} K = I + A,
\end{aligned} \tag{31}$$

for $A = (K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\widehat{K}_\sigma^{-1}K$. By Lemma C.1 and the fact that $\|E\|_{\mathrm{op}} < \sigma^2/2$, we have $\|A\|_{\mathrm{op}} \le \|E\|_{\mathrm{op}}/(\sigma^2 - \|E\|_{\mathrm{op}}) < 1$. Thus, the matrix logarithm $\log(I_n + A)$ is well-defined and we obtain

$$(\mathrm{III}) = -\log\det(I_n + A) = -\operatorname{tr}\log(I_n + A) \tag{32}$$

The matrix logarithm admits the convergent (in operator norm) series

$$\log(I_n + A) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} A^k. \tag{33}$$

Taking traces term by term yields

$$(\mathrm{III}) = -\operatorname{tr}\log(I_n + A) = -\operatorname{tr} A + \frac{1}{2}\operatorname{tr} A^2 - \sum_{k=3}^{\infty} \frac{(-1)^{k+1}}{k} \operatorname{tr} A^k. \tag{34}$$

In the same notation, the first term (I) can be expressed as

$$(\mathrm{I}) = \operatorname{tr}(K - KK_\sigma^{-1}K)^{-1}(K - K\widehat{K}_\sigma^{-1}K) - n = \operatorname{tr} A \tag{35}$$

such that

$$(\mathrm{I}) + (\mathrm{III}) = \frac{1}{2}\operatorname{tr} A^2 - \sum_{k=3}^{\infty} \frac{(-1)^{k+1}}{k} \operatorname{tr} A^k. \tag{36}$$

Step 3: The square trace term. For the square trace term, we obtain

$$\begin{aligned} |\operatorname{tr} A^2| &= |\operatorname{tr}(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\widehat{K}_\sigma^{-1}K \cdot (K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\widehat{K}_\sigma^{-1}K| \\ &= |\operatorname{tr}\widehat{K}_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E \cdot \widehat{K}_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E| \\ &\le \|\widehat{K}_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\|_{\mathrm{F}}^2 \\ &\le \|\widehat{K}_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}\|_{\mathrm{F}}^2\|E\|_{\mathrm{op}}^2, \end{aligned} \tag{37}$$

where we have used elementary estimates for the Frobenius norm and the fact that all matrices involved are symmetric. Next, we use the identity $\widehat{K}_\sigma^{-1} = (K_\sigma + E)^{-1} = (I_n + K_\sigma^{-1}E)^{-1}K_\sigma^{-1}$, which yields the final estimate

$$\begin{aligned} |\operatorname{tr} A^2| &\le \|\widehat{K}_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}\|_{\mathrm{F}}^2\|E\|_{\mathrm{op}}^2 \\ &\le \|(I_n + K_\sigma^{-1}E)^{-1}\|_{\mathrm{op}}^2\|K_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}\|_{\mathrm{F}}^2\|E\|_{\mathrm{op}}^2 \\ &= \|(I_n + K_\sigma^{-1}E)^{-1}\|_{\mathrm{op}}^2 \sum_{j=1}^{n} \Big(\frac{\widehat{\mu}_j^2}{(\widehat{\mu}_j + \sigma^2)^2}\frac{\widehat{\mu}_j + \sigma^2}{\widehat{\mu}_j\sigma^2}\Big)^2\|E\|_{\mathrm{op}}^2 \\ &\le \frac{\|E\|_{\mathrm{op}}^2}{(\sigma^2 - \|E\|_{\mathrm{op}})^2}\sum_{j=1}^{n}\Big(\frac{\widehat{\mu}_j}{\widehat{\mu}_j + \sigma^2}\Big)^2 \end{aligned} \tag{38}$$

with the estimate $\|(I_n + K_\sigma^{-1}E)^{-1}\|_{\mathrm{op}}^2 \le 1/(1 - \|E\|_{\mathrm{op}}/\sigma^2)$.

Step 4: Controlling the remainder. Let $k \ge 3$ be an integer. Then, we have by Cauchy-Schwartz inequality that

$$\begin{aligned} &|\operatorname{tr} A^k| \\ &= \big|\operatorname{tr}\big((K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\widehat{K}_\sigma^{-1}K\big)^k\big| \\ &= \big|\operatorname{tr}\big(\widehat{K}_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\big)^k\big| \\ &\le \big\|\big(\widehat{K}_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\big)^{k-1}\big\|_{\mathrm{F}}\big\|\big(\widehat{K}_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\big)\big\|_{\mathrm{F}} \\ &\le \big\|\big(\widehat{K}_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\big)^{k-2}\big\|_{\mathrm{op}}\big\|\big(\widehat{K}_\sigma^{-1}K(K - KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\big)\big\|_{\mathrm{F}}^2 \end{aligned} \tag{39}$$

As in Step 3, we have

$$\big\|\big(\widehat{K}_\sigma^{-1}K(K-KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\big)\big\|_{\mathrm{F}}^2 \leq \frac{\|E\|_{\mathrm{op}}^2}{(\sigma^2-\|E\|_{\mathrm{op}})^2}\sum_{j=1}^n\Big(\frac{\widehat{\mu}_j}{\widehat{\mu}_j+\sigma^2}\Big)^2 \tag{40}$$

Further, again using the identity $\widehat{K}_\sigma^{-1} = (I_n + K_\sigma^{-1}E)^{-1}K_\sigma^{-1}$,

$$\begin{aligned}
&\big\|\big(\widehat{K}_\sigma^{-1}K(K-KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\big)^{k-2}\big\|_{\mathrm{op}} \leq \big\|\widehat{K}_\sigma^{-1}K(K-KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}E\big\|_{\mathrm{op}}^{k-2}\\
&\leq \|E\|_{\mathrm{op}}^{k-2}\|(I_n+K_\sigma^{-1}E)^{-1}\|_{\mathrm{op}}^{k-2}\big\|K_\sigma^{-1}K(K-KK_\sigma^{-1}K)^{-1}KK_\sigma^{-1}\big\|_{\mathrm{op}}^{k-2}\\
&\leq \frac{\|E\|_{\mathrm{op}}^{k-2}}{(1-\|E\|_{\mathrm{op}}/\sigma^2)^{k-2}\sigma^{2(k-2)}}\max_{j\leq n}\Big(\frac{\widehat{\mu}_j}{\widehat{\mu}_j+\sigma^2}\Big)^{k-2} \leq \Big(\frac{\|E\|_{\mathrm{op}}}{(\sigma^2-\|E\|_{\mathrm{op}})}\Big)^{k-2}.
\end{aligned} \tag{41}$$

For the whole remainder in (I) + (III), we therefore obtain

$$\begin{aligned}
\sum_{k=3}^\infty\frac{1}{k}|\operatorname{tr}A^k| &\leq \frac{\|E\|_{\mathrm{op}}^2}{(\sigma^2-\|E\|_{\mathrm{op}})^2}\sum_{j=1}^n\Big(\frac{\widehat{\mu}_j}{\widehat{\mu}_j+\sigma^2}\Big)^2\sum_{k=3}^\infty\frac{1}{k}\Big(\frac{\|E\|_{\mathrm{op}}}{\sigma^2-\|E\|_{\mathrm{op}}}\Big)^{k-2}\\
&\leq \frac{\|E\|_{\mathrm{op}}}{(\sigma^2-\|E\|_{\mathrm{op}})}\sum_{j=1}^n\Big(\frac{\widehat{\mu}_j}{\widehat{\mu}_j+\sigma^2}\Big)^2\sum_{k=1}^\infty\frac{1}{k}\Big(\frac{\|E\|_{\mathrm{op}}}{\sigma^2-\|E\|_{\mathrm{op}}}\Big)^{k}\\
&= \frac{-\|E\|_{\mathrm{op}}}{(\sigma^2-\|E\|_{\mathrm{op}})}\sum_{j=1}^n\Big(\frac{\widehat{\mu}_j}{\widehat{\mu}_j+\sigma^2}\Big)^2\log\Big(1-\frac{\|E\|_{\mathrm{op}}}{\sigma^2-\|E\|_{\mathrm{op}}}\Big)\\
&\leq \frac{\|E\|_{\mathrm{op}}^2}{(\sigma^2-2\|E\|_{\mathrm{op}})^2}\sum_{j=1}^n\Big(\frac{\widehat{\mu}_j}{\widehat{\mu}_j+\sigma^2}\Big)^2,
\end{aligned} \tag{42}$$

where we have used the inequality $-\log(1-x) \leq x/(1-x)$ for $x \in [0,1)$ in the last step. Combining the results yields the final bound. ■

Proof of Theorem 3.2. By Theorem 3.2 and the additional assumption on $\operatorname{tr}KK_\sigma^{-1}$ there is an event with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$, such that for any $M_n \to \infty$,

$$\mathrm{KL}(\Psi_{n,m}(\cdot), \Pi_n(\cdot|\mathcal{D}^n)) \lesssim M_n n\eta_n^2. \tag{43}$$

The stattement of the theorem now follows by applying Proposition 6.1 in (22), see also Theorem 5 in (19). ■

B Proofs for the Examples

Proof of Proposition 4.1 (Result for the squared exponential kernel). Initially, we note that it is a well known fact that under the assumptions of Proposition 4.1, the concentration function inequality (7) is satisfied for $\eta_n = n^{-\alpha/(2\alpha+d)}\log^{(1/2+3d/4)}n$, see, e.g., Lemma 8 in (16).

Consequently, in order to apply Theorems 3.2 and 3.2, it suffices to bound $\|E\|_{\mathrm{op}} = \|\widehat{K}-K\|_{\mathrm{op}}$ and $\operatorname{tr}KK_\sigma^{-1}$ with probability converging to one. For $r = r_n = \sqrt{2\tau_n^2\log n}$, we immediately have that

$$\|\widehat{K}-K\|_{\mathrm{op}} \leq \max_{i\leq n}\sum_{j:\|X_j-X_i\|>r}k_{\tau_n}(X_i,X_j) \leq n\exp\Big(\frac{-r_n^2}{\tau_n^2}\Big) \leq \frac{1}{n} \to 0. \tag{44}$$

In particular $\|\widehat{K}-K\|_{\mathrm{op}} < \sigma^2/2$ eventually for all $n \in \mathbb{N}$. Further, by Lemma C.2, for any $M_n \to \infty$,

$$\operatorname{tr}KK_\sigma^{-1} \lesssim M_n n^{\frac{d}{2\alpha+d}}\log^d n \lesssim M_n n\eta_n^2 \tag{45}$$

with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$ and $n \to \infty$. Both the contraction rate and the convergence of the Kullback-Leibler divergence follow from combining these bounds.

Finally, the statement about the non-zero row entries follows from applying Lemma C.4 with $nr^d = nr_n^d = 2^{d/2}n\tau_n^d\log^{d/2}n = 2^{d/2}n^{2\alpha/(2\alpha+d)}\log^{d/2}n$ and $\delta = \|g\|_\infty r^d\mathrm{Vol}(B_1(0))$ together with a union bound. Note that $\alpha > d/2$ guarantees that $n\delta^2 \to \infty$ in the exponential bound. ■

Proof of Corollary 4.2 (Lower bound for the squared exponential kernel). We first consider the case $d = 1$. Assume that $G = \mathbb{P}^X$ has density $g(x) \propto e^{-x^2/2}$. As in Lemma C.2, the eigenvalues of the covariance operator then satisfy

$$\lambda_j = \sqrt{1/(2A_n)} \left(\frac{1}{A_n \tau_n^2} \right)^{j-1}, \qquad j \geq 1, \tag{46}$$

where $A_n = 1/4 + \tau_n^{-2} + \sqrt{1/16 + \tau_n^{-2}/2}$. In particular $\lambda_1 = \sqrt{1/(2A_n)} \asymp \tau_n$. For $d > 1$, by tensorisation $\lambda_1 \asymp \tau_n^d$. Noting that the first eigenfunction of the kernel operator is given by $\phi_1(x) = (1+4a_\tau)^{d/4} \exp(-a_\tau \|x\|^2)$ with $a_\tau = (\sqrt{1+4/\tau^2} - 1)/(4(1+\tau^2))$, see (18), we obtain that $\varphi_1 = \lambda_1^{1/2} \phi_1 \in C^\alpha(\mathbb{R}^d)$ for any fixed $\alpha > 0$. Further, it is a well known fact that in this setting the squared exponential kernel is strictly positive definite in the sense of Proposition C.6.

Next, for $r_n = n^{-1/(2\alpha+d)} / \log^{1/(2d)} n$, by Lemma C.4 and the boundedness of the kernel,

$$\|\widehat{K}\|_{\mathrm{op}} \leq \max_{1 \leq i \leq n} \sum_{j : \|X_j - X_i\| \leq r_n} k_{\tau_n}(X_i, X_j) \lesssim n r_n^d \tag{47}$$

with probability converging to one. Taking $m_n := n r_n^d \asymp n^{2\alpha/(2\alpha+d)} / \sqrt{\log n}$ and using the fact that $\tau_n = n^{-1/(2\alpha+d)}$, we obtain

$$\inf_{r \leq r_n} \sqrt{\lambda_1} \left(\frac{n\lambda_1}{2\|\widehat{K}\|_{\mathrm{op}}} - 1 \right) \gtrsim \sqrt{\tau_n^d} \left(\frac{n\tau_n^d}{m_n} \right) \gtrsim n^{\frac{-d/2}{2\alpha+d}} \sqrt{\log n}. \tag{48}$$

This controls the first tern in the bound from Proposition C.6.

We now control the second term using

$$\mathbb{E}\|k(X_1, \cdot)\|_{\mathbb{H}}^4 \lesssim 1, \qquad \|\Sigma\|_{\mathrm{HS}}^2 = \sum_{j \geq 1} \lambda_j^2 \asymp \tau_n^d, \tag{49}$$

where the second statement follows from Equation (65). Hence, $\mathbb{E}\|k(X_1, \cdot)\|_{\mathbb{H}}^4 - \|\Sigma\|_{\mathrm{HS}}^2 \lesssim 1$. Moreover, using the exact formulas for λ_1 and λ_2, we obtain $\lambda_1 - \lambda_2 \asymp \tau_n^{d+1}$. Therefore,

$$\frac{\sqrt{\mathbb{E}\|k(X_1, \cdot)\|_{\mathbb{H}}^4 - \|\Sigma\|_{\mathrm{HS}}^2}}{(\lambda_1 - \lambda_2)\sqrt{n/\log n}} \lesssim \frac{1}{\tau_n^{d+1}\sqrt{n/\log n}}. \tag{50}$$

Multiplying by $\sqrt{\lambda_1} \asymp \tau_n^{d/2}$ and using $\tau_n = n^{-1/(2\alpha+d)}$, we get

$$\sqrt{\lambda_1} \frac{\sqrt{\mathbb{E}\|k(X_1, \cdot)\|_{\mathbb{H}}^4 - \|\Sigma\|_{\mathrm{HS}}^2}}{(\lambda_1 - \lambda_2)\sqrt{n/\log n}} \lesssim \sqrt{\log(n)} n^{\frac{-\alpha+1}{2\alpha+d}}. \tag{51}$$

Combining both results and applying Proposition C.6, we conclude that with probability tending to 1, the approximate posterior is either not well defined or

$$\inf_{r \leq r_n} \sup_{f_0 \in C^\alpha(\mathbb{R}^d)} \|\mathbb{E}_\varepsilon K \widehat{K}_\sigma^{-1} Y - f_0\|_n \gtrsim \left(\sqrt{\log(n)}\, n^{\frac{-d}{2(2\alpha+d)}} - \sqrt{\log(n)}\, n^{\frac{-\alpha+1}{2\alpha+d}} \right)^+. \tag{52}$$

The result follows since $\alpha > d/2 + 1$. ■

Proof of Proposition 4.3 (Result for the Matérn kernel). From Lemmas 3.2 and 3.3 in (9), it follows that the concentration function inequality (7) is satisfied with $\eta_n = n^{-\alpha/(2\alpha+d)}$.

As for the squared exponential kernel, it is then sufficient to bound $\|E\|_{\mathrm{op}} = \|\widehat{K} - K\|_{\mathrm{op}}$ and $\operatorname{tr} K K_\sigma^{-1}$ with probability converging to one. We use the fact that for $\nu = \ell + 1/2$, we have an

explicit representation of the Matérn kernel such that with $r = r_n = 2\tau_n \log n$, we have

$$\begin{aligned}
\|\widehat{K} - K\|_{\mathrm{op}} &\leq \max_{i \leq n} \sum_{j : \|X_j - X_i\| > r_n} k_{\tau_n}(X_i, X_j) \qquad (53)\\
&= \max_{i \leq n} \sum_{j : \|X_j - X_i\| > r_n} \frac{\ell!}{(2\ell)!} \exp\Big(\frac{-\sqrt{2\ell+1} r_n}{\tau_n}\Big) \sum_{q=0}^{\ell} \frac{(\ell+q)!}{q!(\ell-q)!} \Big(\frac{2\sqrt{2\ell+1} r_n}{\tau_n}\Big)^{\ell-q} \\
&\leq n \exp\big(-2\sqrt{2\ell+1}\log n\big) \sum_{q=0}^{\ell} \binom{\ell}{q} \big(4\sqrt{2\ell+1}\log n\big)^{\ell-q} \\
&= n \exp\big(-2\sqrt{2\ell+1}\log n\big) \big(4\sqrt{2\ell+1}\log n + 1\big)^{\ell} \\
&\lesssim n^{1-\sqrt{8\ell+4}} \log^{\ell} n \to 0
\end{aligned}$$

where we have used the binomial theorem in the last equality. The required bound on $\operatorname{tr} K K_\sigma^{-1}$ is provided in Lemma C.5.

Finally, the statement about the non-zero row entries follows from applying Lemma C.4 with $nr^d = nr_n^d = 2^d n \tau_n^d \log^d n = 2^d n^{(2\alpha\nu + d\alpha)/(2\alpha\nu + d\nu)} \log^d n$ and $\delta = \|g\|_\infty r^d \mathrm{Vol}(B_1(0))$ together with a union bound. Note that $\alpha > d/2$ guarantees that $n\delta^2 \to \infty$ in the exponential bound for any choice of $\nu = \ell + 1/2$ with $\ell \in \mathbb{N}_0$. ■

C Auxiliary results

Lemma C.1 (Norm estimate for the log-determinant term)**.** *Set $E = \widehat{K} - K$. Then*

$$\|(K - K K_\sigma^{-1} K)^{-1} K K_\sigma^{-1} E \widehat{K}_\sigma^{-1} K\|_{\mathrm{op}} \leq \frac{\|E\|_{\mathrm{op}}}{\sigma^2 - \|E\|_{\mathrm{op}}}. \qquad (54)$$

Proof. Since $\widehat{K}_\sigma = K_\sigma + E$, we have

$$\begin{aligned}
&\|(K - K K_\sigma^{-1} K)^{-1} K K_\sigma^{-1} E \widehat{K}_\sigma^{-1} K\|_{\mathrm{op}} \qquad (55)\\
&= \|(K - K K_\sigma^{-1} K)^{-1} K K_\sigma^{-1} E (I_n + K_\sigma^{-1} E)^{-1} K_\sigma^{-1} K\|_{\mathrm{op}} \\
&\leq \|(K - K K_\sigma^{-1} K)^{-1} K K_\sigma^{-1}\|_{\mathrm{op}} \, \|E\|_{\mathrm{op}} \, \|(I_n + K_\sigma^{-1} E)^{-1}\|_{\mathrm{op}} \, \|K_\sigma^{-1} K\|_{\mathrm{op}}.
\end{aligned}$$

We now bound each factor individually. First, $\|K_\sigma^{-1} K\|_{\mathrm{op}} \leq 1$.

Next, since $\|K_\sigma^{-1} E\|_{\mathrm{op}} < 1$, the Neumann series gives

$$\|(I_n + K_\sigma^{-1} E)^{-1}\|_{\mathrm{op}} \leq \frac{1}{1 - \|K_\sigma^{-1} E\|_{\mathrm{op}}}. \qquad (56)$$

Moreover, $\|K_\sigma^{-1} E\|_{\mathrm{op}} \leq \|K_\sigma^{-1}\|_{\mathrm{op}} \|E\|_{\mathrm{op}} \leq \|E\|_{\mathrm{op}} / \sigma^2$. Hence,

$$\|(I_n + K_\sigma^{-1} E)^{-1}\|_{\mathrm{op}} \leq \frac{1}{1 - \|E\|_{\mathrm{op}}/\sigma^2}. \qquad (57)$$

Finally, since K and K_σ are simultaneously diagonalizable, we obtain

$$\|(K - K K_\sigma^{-1} K)^{-1} K K_\sigma^{-1}\|_{\mathrm{op}} = \max_{j=1,\dots,n} \Big\{ \Big(\hat{\mu}_j - \frac{\hat{\mu}_j^2}{\hat{\mu}_j + \sigma^2}\Big)^{-1} \frac{\hat{\mu}_j}{\hat{\mu}_j + \sigma^2} \Big\} = \sigma^{-2}. \qquad (58)$$

Combining the previous bounds yield the result. ■

Lemma C.2 (Trace term for the squared exponential kernel)**.** *Assume that $\mathcal{X} = \mathbb{R}^d$ and that $G = N(0, \gamma^2 I_d)$ for some $\gamma^2 > 0$. Let k_{τ_n} be the squared exponential kernel with length scale $\tau_n = n^{-1/(2\alpha+d)}$, where $\alpha > 0$. Then, for any sequence $M_n \to \infty$,*

$$\operatorname{tr}(K K_\sigma^{-1}) \lesssim M_n n^{\frac{d}{2\alpha+d}} \log^d n, \qquad (59)$$

with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$ as $n \to \infty$.

Proof. First we give an upper bound for the eigenvalues for the squared exponential covariance kernel. This derivation is standard, but for completeness we provide a proof. For $d = 1$ case, let $(\phi_j)_{j\geq 1}$ and $(\lambda_{1,j})_{j\geq 1}$ denote the eigenfunctions and eigenvalues of the one-dimensional kernel k_{τ_n}, respectively. The distribution G has a density $g(x) \propto e^{-2ax^2}$ for some $a = 1/(2\gamma^2)$. In this case, the eigenvalues admit the explicit form

$$\lambda_{1,j} = \sqrt{2a/A_n}\left(\frac{1}{A_n\tau_n^2}\right)^{j-1}, \qquad j \geq 1, \tag{60}$$

where $A_n = a + \tau_n^{-2} + \sqrt{a^2 + 2a\tau_n^{-2}}$, see (18). With elementary derivations, it can then be shown that

$$\lambda_{1,j} \lesssim \tau_n \exp(-D_{\exp}\tau_n j), \qquad j \geq 1, \tag{61}$$

for any $D_{\exp} < \sqrt{2a}$, see, e.g., the proof of Corollary 9 in (16).

Next, the eigenfunctions and eigenvalues of the general d-dimensional kernel k_{τ_n} are given by

$$\Phi_{\mathbf{j}}(x) = \prod_{k=1}^{d} \phi_{j_k}(x_k), \quad \Lambda_{\mathbf{j}} = \prod_{k=1}^{d} \lambda_{1,j_k}, \qquad \mathbf{j} = (j_1, \ldots, j_d) \in \mathbb{N}^d. \tag{62}$$

Using the one-dimensional bound in Equation (61), we obtain that for every $\mathbf{j} \in \mathbb{N}^d$,

$$\Lambda_{\mathbf{j}} = \prod_{k=1}^{d} \lambda_{1,j_k} \lesssim \tau_n^d \exp\big(-D_{\exp}\tau_n \|\mathbf{j}\|_1\big), \text{ where } \|\mathbf{j}\|_1 = \sum_{k=1}^{d} j_k. \tag{63}$$

We now consider the non-increasing single-index rearrangement $(\lambda_j)_{j\geq 1}$ of the multi-index eigenvalues $(\Lambda_{\mathbf{j}})_{\mathbf{j}\in\mathbb{N}^d}$. For the counting function

$$N(q) := \#\{\mathbf{j} \in \mathbb{N}^d : \|\mathbf{j}\|_1 \leq q\}, \qquad q \in \mathbb{N}, \tag{64}$$

it is well known that $N(q) \asymp q^d$ for $q \to \infty$. Hence, the relation between the rank j and the level q satisfies $j \asymp q^d$, which is equivalent to $q \asymp j^{1/d}$. Since all the eigenvalues corresponding to multi-indices $\mathbf{j}$ with $\|\mathbf{j}\|_1 \asymp q$ are bounded by $\exp(-D_{\exp}\tau_n q)$, it follows that the j-th largest eigenvalue satisfies

$$\lambda_j \lesssim \tau_n^d \exp\big(-D_{\exp}\tau_n q\big) \asymp \tau_n^d \exp\big(-D_{\exp}\tau_n j^{1/d}\big) \quad \text{with } q \asymp j^{1/d}. \tag{65}$$

where we potentially adjusted the constant $D_{\exp}$. Next, let $J = J_n \in \mathbb{N}$ with $J \leq n$ and introduce the notation $\widehat{\lambda}_j = \widehat{\mu}_j/n$ $j \leq n$. Then, from Proposition 2 in (32) combined with Markov's inequality, it follows that

$$\begin{aligned}\operatorname{tr}\big(KK_\sigma^{-1}\big) = \sum_{j=1}^{n} \frac{\widehat{\mu}_j}{\widehat{\mu}_j + \sigma^2} = \sum_{j=1}^{n} \frac{n\widehat{\lambda}_j}{n\widehat{\lambda}_j + \sigma^2} &\leq J + \sum_{j=J+1}^{n} \frac{n\widehat{\lambda}_j}{n\widehat{\lambda}_j + \sigma^2}. \\ &\leq J + \frac{n}{\sigma^2}\sum_{j=J+1}^{n} \widehat{\lambda}_j \leq J + \frac{M_n n}{\sigma^2}\sum_{j=J+1}^{n} \lambda_j\end{aligned} \tag{66}$$

with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$ and any $M_n \to \infty$. Restricting to this event and using the inequality in Equation (65), we obtain

$$\operatorname{tr}\big(KK_\sigma^{-1}\big) \lesssim M_n\Big(J_n + \frac{n\tau_n^d}{\sigma^2}\sum_{j=J_n+1}^{n} \exp\big(-D_{\exp}\tau_n j^{1/d}\big)\Big). \tag{67}$$

Noting that by Lemma C.3,

$$\begin{aligned}\sum_{j=J_n+1}^{n} \exp\big(-D_{\exp}\tau_n j^{1/d}\big) &\leq \int_{J_n}^{\infty} \exp\big(-D_{\exp}\tau_n t^{1/d}\big)\, dt \\ = d\int_{J_n^{1/d}}^{\infty} u^{d-1}\exp\big(-D_{\exp}\tau_n u\big)\, du &\lesssim \tau_n^{-1} J_n^{1-1/d}\exp\Big(-D_{\exp}\tau_n J_n^{1/d}\Big).\end{aligned} \tag{68}$$

Finally, choosing $J_n = \tau_n^{-d}\log^d(n)/D_{\exp}$ and adjusting the sequence M_n to account for constants yields the results. ∎

Lemma C.3 (Bound for the upper incomplete Γ-function)**.** *For $p > 0$, let*

$$\Gamma(p, x) = \int_x^\infty t^{p-1} e^{-t}\, dt, \qquad x > 0 \tag{69}$$

be the upper incomplete Γ-function. Then,

$$\Gamma(p, x) \leq \max\{1, 2^{p-2}\}\big(x^{p-1} + \Gamma(p)\big)e^{-x}, \qquad x > 0. \tag{70}$$

Proof. We have that

$$\begin{aligned}\int_x^\infty t^{p-1} e^{-t}\, dt &= e^{-x} \int_0^\infty (x+s)^{p-1} e^{-s}\, ds \leq \max\{1, 2^{p-2}\} e^{-x} \int_0^\infty (x^{p-1} + s^{p-1}) e^{-s}\, ds \\ &= \max\{1, 2^{p-2}\} e^{-x}\big(x^{p-1} + \Gamma(p)\big),\end{aligned} \tag{71}$$

where for $p - 1 \geq 1$, we have used the convexity estimate $(x + s)^{p-1} \leq 2^{p-2}(x^{p-1} + s^{p-1})$ and $(x + s)^{p-1} \leq x^{p-1} + s^{p-1}$ for $p - 1 < 1$. ■

Lemma C.4 (Number of neighbours bound)**.** *Let $X_1, \ldots, X_n \sim$ i.i.d. be distributed according to G on $\mathcal{X}$ with Lebesgue density g which is continuous and strictly positive at $x \in \mathcal{X}$. For a radius $r > 0$, let*

$$N_i(r) = \sum_{j \neq i} \mathbf{1}\{\|X_j - X_i\| \leq r\} \tag{72}$$

be the number of neighbours of X_i not further away than r. Then, conditionally on $X_i = x$, with probability at least $1 - 2\exp(-(n-1)\delta^2/4)$, we have that

$$|N_i(r) - (n-1)p_r(x)| \leq (n-1)\delta \tag{73}$$

where $p_r(x) = \int_{B_r(x)} g(u)du \asymp g(x)r^d Vol(B_1(0))$ for $r \to 0$.

Proof. For $r \to 0$, conditionally on $X_i = x \in \mathcal{X}$, $N_i(r)$ is distributed as a binomial $\mathrm{Bin}(n-1, p_r(x))$ with

$$p_r(x) = \int_{B_r(x)} g(u)du = \int_{B_1(0)} g(x + ru)r^d du = g(x)r^d\, \mathrm{Vol}(B_1(0))(1 + o(1)) \tag{74}$$

for $r \to 0$. By Chernoff's inequality, c.f. Exercise 2.10 in (3), for $\delta > 0$,

$$\mathbb{P}\big(|N_i(r) - (n-1)p_r(x)| \geq (n-1)\delta\big) \leq \exp\Big(\frac{-(n-1)\delta^2}{4}\Big). \tag{75}$$

Note that compared to Exercise 2.10, the bound can be formulated to avoid the appearance of $p_r(x)$ on the right-hand side. ■

Lemma C.5 (Trace term for the Matérn kernel)**.** *Assume that $\mathcal{X} \subset \mathbb{R}^d$ is bounded and the design distribution G on $\mathcal{X}$ has a bounded Lebesgue density g. Let k_{τ_n} be the Matérn kernel with smoothness ν and length scale $\tau_n = n^{-(\nu-\alpha)/(2\alpha\nu+d\nu)}$. Then, for any sequence $M_n \to \infty$,*

$$\operatorname{tr} K K_\sigma^{-1} \lesssim M_n n^{d/(2\alpha+d)} \tag{76}$$

with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$ for $n \to \infty$.

Proof. As in Lemma C.2, we have that for any $J \in \mathbb{N}$ and any $M_n \to \infty$,

$$\operatorname{tr} K K_\sigma^{-1} \leq J + \frac{M_n n}{\sigma^2} \sum_{j=J+1}^{n} \lambda_j \tag{77}$$

with probability converging to one under $\mathbb{P}_{f_0}^{\otimes n}$.

Now, under the assumptions of Theorem 1 in (21) together with the definition of the spectral density of the Matérn prior guarantees that

$$\lambda_j \lesssim \tau_n^{-2\nu} j^{-1-2\nu/d}, \qquad j \geq 1. \tag{78}$$

Plugging into the bound in Equation (77) yields

$$\operatorname{tr} K K_\sigma^{-1} \lesssim M_n\Big(J + n\tau_n^{-2\nu} \sum_{j \geq J+1} j^{-1-2\nu/d}\Big) \lesssim M_n\Big(J + n\tau_n^{-2\nu} J^{-2\nu/d}\Big). \tag{79}$$

Setting $J = J_n = n^{d/(2\alpha+d)}$ finishes the proof. ■

Proposition C.6 (Lower bound)**.** *Let k be a strictly positive definite and bounded kernel, i.e. $K > 0$ a.e. and $\sup k(x,x) < \infty$. Further assume that $f_0 = \varphi_1 = \lambda_1^{1/2}\phi_1$ is the first eigenfunction of the covariance operator $\Sigma = T_k|_{\mathbb{H}}$ of $k(X_1,.)$, see Appendix D.2. Then, there is an event with probability converging to 1 on which the approximate posterior is either not well defined in the sense that $0 \leq K - K\widehat{K}_\sigma^{-1}K \leq K$ or*

$$\|\mathbb{E}_\varepsilon K\widehat{K}_\sigma^{-1}Y - f_0\|_n \geq c\sqrt{\lambda_1}\Big(\Big(\frac{n\lambda_1}{2\|\widehat{K}\|_{op}} - 1\Big) - \frac{\sqrt{\mathbb{E}\|k(X_1,\cdot)\|_{\mathbb{H}}^4 - \|\Sigma\|_{HS}^2}}{(\lambda_1-\lambda_2)\sqrt{n/\log n}}\Big)^+ \tag{80}$$

The well definedness condition can be interpreted such that the approximate posterior has a positive semi-definite covariance structure that reduces the variance of the prior.

Proof. **Step 1: Deterministic analysis of kernel matrix terms.** The condition $K - K\widehat{K}_\sigma^{-1}K \leq K$ requires that $K\widehat{K}_\sigma^{-1}K \geq 0$. By Sylvester's law of inertia and the fact that k is strictly positive definite, this is equivalent to $K^{1/2}\widehat{K}_\sigma^{-1}K^{1/2} \geq 0$, which in turn is equivalent to $K\widehat{K}_\sigma^{-1} \geq 0$ by the similarity of matrices. With the same arguments this is also equivalent to $\widehat{K}_\sigma \geq 0$.

With a similar argument $K - K\widehat{K}_\sigma^{-1}K \geq 0$ is equivalent to $I - K^{1/2}\widehat{K}_\sigma^{-1}K^{1/2} \geq 0$. This implies that that $K^{1/2}\widehat{K}_\sigma^{-1}K^{1/2}$ may only have eigenvalues smaller than 1 and by similarity of the matrices the condition becomes $KK_\sigma^{-1} \leq I$. In the following we therefore only consider realisations of the design for which $0 \leq KK_\sigma^{-1} \leq I$.

Now, let

$$S_n : \mathbb{H} \to \mathbb{R}^n, \qquad h \mapsto (h(X_1), \dots, h(X_n))^\top \tag{81}$$

be the sampling operator and let $(\widehat{\varphi}_j)_{j\leq n}$ be the eigenfunctions of the empirival covariance operator

$$\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^n k(X_i,\cdot)\otimes k(X_i,\cdot) = S_n^* S_n, \tag{82}$$

where S_n^* is the adjoint of S_n with respect to the space $(\mathbb{R}^n, \langle\cdot,\cdot\rangle_n)$. Noting that $S_nS_n^* = n^{-1}K$, the operators $\widehat{\Sigma}$ and $n^{-1}K$ have the same eigenvalues $(\widehat{\lambda}_j)_{j\leq n}$.

We may now write

$$\begin{aligned}\|\mathbb{E}_\varepsilon K\widehat{K}_\sigma^{-1} - f_0\|_n &= \|K\widehat{K}_\sigma^{-1}S_nf_0 - S_nf_0\|_n = \|K\widehat{K}_\sigma^{-1}S_n\varphi_1 - S_n\varphi_1\|_n \\ &\geq \|K\widehat{K}_\sigma^{-1}S_n\widehat{\varphi}_1 - S_n\widehat{\varphi}_1\|_n - \|(K\widehat{K}_\sigma^{-1} - I)S_n(\varphi_1 - \widehat{\varphi}_1)\|_n.\end{aligned} \tag{83}$$

We treat the two remaining terms individually.

Davis-Kahan theorem implies that, by potentially adjusting the sign of $\widehat{\varphi}_1$,

$$\|\varphi_1 - \widehat{\varphi}_1\|_{\mathbb{H}} \leq \frac{\|\widehat{\Sigma} - \Sigma\|_{\text{HS}}}{\lambda_1 - \lambda_2}. \tag{84}$$

Therefore, under the assumption that $0 \leq K\widehat{K}_\sigma^{-1} \leq I$, we may estimate

$$\begin{aligned}&\|(K\widehat{K}_\sigma^{-1} - I)S_n(\varphi_1 - \widehat{\varphi}_1)\|_n \leq \|(K\widehat{K}_\sigma^{-1} - I)\|_{\text{op}}\|S_n(\varphi_1 - \widehat{\varphi}_1)\|_n \\ &\leq C\sqrt{\widehat{\lambda}_1}\|\widehat{\varphi}_1 - \varphi_1\|_{\mathbb{H}} \leq C\sqrt{\widehat{\lambda}_1}\frac{\|\widehat{\Sigma} - \Sigma\|_{\text{HS}}}{\lambda_1 - \lambda_2}.\end{aligned} \tag{85}$$

For the first term, we use the fact that $S_n\widehat{\varphi}_1 = \widehat{\lambda}_1^{1/2}\widehat{v}_1$ where $\widehat{v}_1$ is an eigenvevtor of K with $\|\widehat{v}_1\|_n = 1$. Therefore,

$$\|K\widehat{K}_\sigma^{-1}S_n\widehat{\varphi}_1 - S_n\widehat{\varphi}_1\|_n \geq \langle \widehat{v}_1, (K\widehat{K}_\sigma^{-1} - I)S_n\widehat{\varphi}_1\rangle_n = \widehat{\lambda}_1^{1/2}\langle \widehat{v}_1, (K\widehat{K}_\sigma^{-1} - I)\widehat{v}_1\rangle_n \tag{86}$$

$$= \widehat{\lambda}_1^{1/2}\big(\langle \widehat{v}_1, KK_\sigma^{-1}\widehat{v}_1\rangle_n - 1\big) = \widehat{\lambda}_1^{1/2}\big(n\widehat{\lambda}_1\langle \widehat{v}_1, (I + K_\sigma^{-1}E)^{-1}K_\sigma^{-1}\widehat{v}_1\rangle_n - 1\big)$$

$$= \widehat{\lambda}_1^{1/2}\Big(\frac{n\widehat{\lambda}_1}{n\widehat{\lambda}_1 + \sigma^2}\langle \widehat{v}_1, (I + K_\sigma^{-1}E)^{-1}\widehat{v}_1\rangle_n - 1\Big)$$

$$= \widehat{\lambda}_1^{1/2}\Big(\frac{n\widehat{\lambda}_1}{n\widehat{\lambda}_1 + \sigma^2}\langle \widehat{v}_1, K_\sigma^{-1/2}(I + K_\sigma^{-1/2}EK_\sigma^{-1/2})^{-1}K_\sigma^{1/2}\widehat{v}_1\rangle_n - 1\Big)$$

$$= \widehat{\lambda}_1^{1/2}\Big(\frac{n\widehat{\lambda}_1}{n\widehat{\lambda}_1 + \sigma^2}\langle \widehat{v}_1, (I + K_\sigma^{-1/2}EK_\sigma^{-1/2})^{-1}\widehat{v}_1\rangle_n - 1\Big).$$

We note that $I + K_\sigma^{-1/2}EK_\sigma^{-1/2} > 0$ is equivalent to $K_\sigma + E > 0$, which is in turn equivalent to $\widehat{K}_\sigma > 0$. We can then use the inequality

$$\langle v, Av\rangle_n\langle v, A^{-1}v\rangle_n \geq \langle v, v\rangle_n^2, \tag{87}$$

which holds for all $A > 0$ and $v \in \mathbb{R}^n$. This yields the estimate

$$\|K\widehat{K}_\sigma^{-1}S_n\widehat{\varphi}_1 - S_n\widehat{\varphi}_1\|_n \geq \widehat{\lambda}_1^{1/2}\Big(\frac{n\widehat{\lambda}_1}{n\widehat{\lambda}_1 + \sigma^2}\langle \widehat{v}_1, (I + K_\sigma^{-1/2}EK_\sigma^{-1/2})\widehat{v}_1\rangle_n^{-1} - 1\Big) \tag{88}$$

$$= \widehat{\lambda}_1^{1/2}\Big(\frac{n\widehat{\lambda}_1}{n\widehat{\lambda}_1 + \sigma^2}\Big[1 + \frac{\langle \widehat{v}_1, E\widehat{v}_1\rangle_n}{n\widehat{\lambda}_1 + \sigma^2}\Big]^{-1} - 1\Big)$$

$$= \widehat{\lambda}_1^{1/2}\Big(\frac{n\widehat{\lambda}_1}{n\widehat{\lambda}_1 + \sigma^2}\Big[\frac{n\widehat{\lambda}_1 + \sigma^2 + \langle \widehat{v}_1, \widehat{K}\widehat{v}_1\rangle_n}{n\widehat{\lambda}_1 + \sigma^2}\Big]^{-1} - 1\Big) \geq \widehat{\lambda}_1^{1/2}\Big(\frac{n\widehat{\lambda}_1}{\|\widehat{K}\|_{\text{op}}} - 1\Big).$$

Step 2: Contructing the event. We have

$$\mathbb{E}\|\widehat{\Sigma} - \Sigma\|_{\text{HS}}^2 = \mathbb{E}\|\widehat{\Sigma}\|_{\text{HS}}^2 - \|\Sigma\|_{\text{HS}}^2 = \mathbb{E}\sum_{j=1}^n\langle \widehat{\Sigma}\widehat{\varphi}_j, \widehat{\Sigma}\widehat{\varphi}_j\rangle - \|\Sigma\|_{\text{HS}}^2 \tag{89}$$

$$= \frac{1}{n^2}\mathbb{E}\Big(\sum_{j=1}^n\sum_{i,i'=1}^n \big\langle\langle k(X_i,\cdot), \widehat{\varphi}_j\rangle k(X_i,\cdot), \langle k(X_{i'},\cdot), \widehat{\varphi}_j\rangle k(X_{i'},\cdot)\big\rangle\Big) - \|\Sigma\|_{\text{HS}}^2$$

$$\leq \frac{1}{n^2}\big(n\mathbb{E}\|k(X_1,\cdot)\|_{\mathbb{H}}^4 + (n^2 - n)\|\Sigma\|_{\text{HS}}^2\big) - \|\Sigma\|_{\text{HS}}^2 = \frac{\mathbb{E}\|k(X_1,\cdot)\|_{\mathbb{H}}^4 - \|\Sigma\|_{\text{HS}}^2}{n}.$$

By an applying Markov inequality, there is an event Ω_1 on which $\|\widehat{\Sigma} - \Sigma\|_{\text{HS}}^2 \leq (\mathbb{E}\|k(X_1,\cdot)\|_{\mathbb{H}}^4 - \|\Sigma\|_{\text{HS}}^2)\log(n)/n$.

Further, by Proposition 2 in (32) and the fact that k is bounded, there is an event Ω_2 on which $\widehat{\lambda}_1 \geq \lambda_1/2$ with probability converging to 1. The result now follows by restricting to $\Omega_1 \cap \Omega_2$. ■

D Complementary results and examples

D.1 Well definedness of the approximate posterior

We show that $\Psi_{r_n}(\cdot)$ is a well defined Gaussian process whenever $\|\widehat{K} - K\| < \sigma^2$. This follows from the fact that it arises as the posterior of the following perturbed observation model:

$$\tilde{Y}_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{90}$$

where the noise vector $\xi = (\xi_i)_{i\leq n}$ satisfies $\xi \sim \mathcal{N}(0, \widehat{K} - K + \sigma^2 I_n)$ and is independent of the prior process W on f (although it may depend on the design X). The condition $\|\widehat{K} - K\|_{\text{op}} < \sigma^2$ ensures that the covariance matrix $\widehat{K} - K + \sigma^2 I_n$ is symmetric positive definite, and therefore defines a valid Gaussian noise model.

Let $\tilde{n} \in \mathbb{N}$ and $\tilde{\mathbf{x}} = (\tilde{x}_1, \dots, \tilde{x}_{\tilde{n}}) \in \mathcal{X}^{\tilde{n}}$ be some evaluation points. Under the perturbed model (90), the joint distribution of the observations $\tilde{Y}$ and the vector $\tilde{\mathbf{f}} = (f(\tilde{x}_1), \cdots, f(\tilde{x}_{\tilde{n}}))^\top$ is Gaussian,

$$\begin{pmatrix} \tilde{Y} \\ \tilde{\mathbf{f}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \widehat{K}_\sigma & K(X, \tilde{x}) \\ K(X, \tilde{x})^\top & K(\tilde{x}, \tilde{x}) \end{pmatrix} \right), \tag{91}$$

where $K(\mathbf{x}, \tilde{\mathbf{x}}) = (k(x_i, \tilde{x}_j))_{i \le n, j \le \tilde{n}}$. The assumption $\|\widehat{K} - K\|_{\mathrm{op}} < \sigma^2$ implies that $\widehat{K}_\sigma$ is strictly positive definite and therefore invertible. By the standard formula for conditional distributions of multivariate Gaussian vectors, the conditional distribution of $\mathbf{f}(\mathbf{x})$ given $(X, \tilde{Y})$ is

$$\tilde{\mathbf{f}} \mid X, \tilde{Y} \sim \mathcal{N}\Big(K(X, \tilde{\mathbf{x}})^\top \widehat{K}_\sigma^{-1} \tilde{Y},\ K(\mathbf{x}, \mathbf{x}) - K(X, \mathbf{x})^\top \widehat{K}_\sigma^{-1} K(X, \mathbf{x}) \Big). \tag{92}$$

Since a Gaussian process is uniquely determined by its finite-dimensional distributions, this shows that $\Psi_{r_n}(\cdot | (X, Y) = (x, y))$ is a well-defined Gaussian process and coincides with the posterior distribution $W | (X, \tilde{Y}) = (x, y)$ in the perturbed model. Based on this, we can show that the Kullback-Leibler divergence between the processes Ψ_{r_n} and $\Pi_n(\cdot | X, Y)$ coincides with the Kullback-Leibler divergence between the processes evaluated at the design points.

Lemma D.1 (Process KL vs. finite dimensional KL)**.** *Conditional on the design X and the observations Y, the KL divergence $\mathrm{KL}(\Psi_{r_n}, \Pi_n(\cdot | X, Y))$ between the banded posterior GP and the true posterior GP is equal to the KL divergence $\mathrm{KL}(\Psi_{r_n}(\mathbf{f} \in \cdot), \Pi_n(\mathbf{f} \in \cdot | X, Y))$ between the finite dimensional Gaussians given by $\mathbf{f} = (f(X_1), \dots f(X_n))^\top$.*

Proof. Initially, we introduce some helpful notation. Let $(\Omega, \mathscr{A}, \mathbb{P})$ denote the probability space on which both regression models from Equations (1) and (90) are defined conditional on the design X. In the following, superscripts will denote push forward measures.

Assuming all quantities exist, we may now write the Kullback-Leibler divergence in terms of Radon-Nikodym derivatives, i.e.,

$$\begin{aligned} &\mathrm{KL}(\Psi_{r_n}(\cdot | Y = y), \Pi_n(\cdot | Y = y)) \\ &= \int \log \frac{d\Psi_{r_n}(\cdot | Y = y)}{d\Pi} \, d\Psi_{r_n}(\cdot | Y = y) + \int \log \frac{d\Pi}{d\Pi_n(\cdot | Y = y)} \, d\Psi_{r_n}(\cdot | Y = y) \\ &= \int \log \frac{d\mathbb{P}^{W | \tilde{Y} = y}}{d\mathbb{P}^W} \, d\mathbb{P}^{W | \tilde{Y} = y} + \int \log \frac{d\mathbb{P}^W}{d\mathbb{P}^{W | Y = y}} \, d\mathbb{P}^{W | \tilde{Y} = y} \end{aligned} \tag{93}$$

and analogously

$$\begin{aligned} &\mathrm{KL}(\Psi_{r_n}(\mathbf{f} \in \cdot | Y = y), \Pi_n(\mathbf{f} \in \cdot | Y = y)) \\ &= \int \log \frac{d\mathbb{P}^{\vec{W} | \tilde{Y} = y}}{d\mathbb{P}^{\vec{W}}} \, d\mathbb{P}^{\vec{W} | \tilde{Y} = y} + \int \log \frac{d\mathbb{P}^{\vec{W}}}{d\mathbb{P}^{\vec{W} | Y = y}} \, d\mathbb{P}^{\vec{W} | \tilde{Y} = y}, \end{aligned} \tag{94}$$

where $\vec{W} := (W(X_1), \dots, W(X_n))$ denotes the finite dimensional evaluation of the process W. Therefore, it is sufficient to show that the integrands in Equations (93) and (94) exist and coincide respectively.

Since any posterior is almost surely absolutely continuous with respect to the prior, we have that the Radon-Nikodym derivatives $d\mathbb{P}^{W | Y = y} / d\mathbb{P}^W$, $d\mathbb{P}^{W | \tilde{Y} = y} / d\mathbb{P}^W$ exist and are well defined. Further, via Bayes-theorem,

$$\frac{d\mathbb{P}^{W | Y = y}}{d\mathbb{P}^W} = \frac{d(p(Y = y | W = f) / p(Y = y) \pi(df))}{d\Pi} = \frac{p(Y | \vec{W} = \mathbf{f})}{p(Y = y)} > 0 \tag{95}$$

due to the exponential form of the likelihood p. Therefore,

$$\frac{d\mathbb{P}^W}{d\mathbb{P}^{W | Y = y}} = \Big(\frac{d\mathbb{P}^{W | Y = y}}{d\mathbb{P}^W} \Big)^{-1} = \frac{p(Y = y)}{p(Y = y | W)} = \frac{p(Y = y)}{p(Y = y | \vec{W})}. \tag{96}$$

We conclude that the right-most terms in Equations (93) and (94) exist and coincide.

An analogous argument as above also shows that $d\mathbb{P}^{W | \tilde{Y} = y} / d\mathbb{P}^W$ only depends on $\vec{W}$. This finishes the proof, since therefore also the left-most terms in Equation (93) and (94) exist and coincide. ■

D.2 Reproducing kernel Hilbert spaces

Let $\mathcal{X} \subset \mathbb{R}^d$ and let $k : \mathcal{X}^2 \to \mathbb{R}$ be a continuous positive definite kernel. The associated reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ can be characterized in several equivalent ways.

Definition and reproducing property. The RKHS $\mathbb{H}$ is defined as the completion of the linear span of functions $\{k(x, \cdot) : x \in \mathcal{X}\}$, equipped with the canonical inner product

$$\left\langle \sum_{i=1}^{m} a_i k(x_i, \cdot), \sum_{j=1}^{n} b_j k(s_j, \cdot) \right\rangle_{\mathbb{H}} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j k(x_i, s_j), \qquad a \in \mathbb{R}^m,\ b \in \mathbb{R}^n. \tag{97}$$

In particular, for every $s \in \mathcal{X}$, the function $k(s, \cdot)$ belongs to $\mathbb{H}$ and satisfies the reproducing property

$$f(s) = \langle f, k(s, \cdot) \rangle_{\mathbb{H}}, \qquad f \in \mathbb{H}. \tag{98}$$

Consequently, pointwise evaluation is a continuous linear functional on $\mathbb{H}$. Moreover, by the Cauchy–Schwarz inequality,

$$|f(x)| \le \|f\|_{\mathbb{H}} \sqrt{k(x, x)}, \qquad f \in \mathbb{H}. \tag{99}$$

Stationary kernels and Bochner's theorem. When the kernel is stationary, i.e.

$$k(x, x') = r(x' - x), \qquad x, x' \in \mathcal{X}, \tag{100}$$

for some function $r : \mathbb{R}^d \to \mathbb{R}$, a more explicit characterization is available via Fourier analysis. A continuous function $r : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite non-negative measure μ, namely

$$r(h) = \int_{\mathbb{R}^d} e^{i\langle \omega, h \rangle}\, \mu(d\omega). \tag{101}$$

If μ admits a density $S(\omega)$ with respect to Lebesgue measure, then S is called the spectral density of the kernel. In this case, the associated RKHS consists of the real parts of functions of the form

$$f(t) = \int_{\mathbb{R}^d} e^{i\langle t, \omega \rangle} \psi(\omega)\, \mu(d\omega), \qquad \psi \in L^2(\mu), \tag{102}$$

with RKHS norm

$$\|f\|_{\mathbb{H}}^2 = \inf_{\psi:\, f = \mathcal{F}\psi} \int_{\mathbb{R}^d} |\psi(\omega)|^2\, \mu(d\omega). \tag{103}$$

Covariance operator and spectral decomposition. Let

$$(T_k f)(x) = \int_{\mathcal{X}} k(x, y) f(y)\, d\mathbb{G}(dy), x \in \mathcal{X} \tag{104}$$

denote the covariance operator associated with k. Assume that k is measurable, symmetric and positive definite, and that

$$\int_{\mathcal{X}} k(x, x)\, d\mathbb{P}_X(x) < \infty. \tag{105}$$

Then, T_k defines a positive trace-class operator on $L^2(\mathbb{P}_X)$. Consequently, there exist eigenpairs $(\lambda_j, \phi_j)_{j \ge 1}$ satisfying

$$T_k \phi_j = \lambda_j \phi_j, \qquad j \ge 1, \tag{106}$$

such that

$$k(x, y) = \sum_{j \ge 1} \lambda_j \phi_j(x) \phi_j(y), \qquad x, y \in \mathcal{X} \tag{107}$$

in $L^2(G \otimes G)$.

Moreover, the associated RKHS admits the characterization

$$\mathbb{H} = \left\{ f = \sum_{j\geq 1} a_j \phi_j : \sum_{j\geq 1} \frac{a_j^2}{\lambda_j} < \infty \right\}, \tag{108}$$

with RKHS norm

$$\|f\|_{\mathbb{H}}^2 = \sum_{j\geq 1} \frac{a_j^2}{\lambda_j}. \tag{109}$$

The decay of the eigenvalues $(\lambda_j)_{j\geq 1}$ reflects the smoothness of the kernel and the complexity of the associated RKHS. Exponential eigenvalue decay, as for squared exponential kernels, corresponds to very smooth functions, whereas polynomial decay, as for Matérn-type kernels, yields Sobolev-type space.

Karhunen–Loève expansion. If W is a centered Gaussian process with covariance kernel k, then under the above assumptions,

$$W = \sum_{j\geq 1} \sqrt{\lambda_j}\, \xi_j \phi_j, \tag{110}$$

where $(\xi_j)_{j\geq 1}$ are i.i.d. standard Gaussian random variables, with convergence in $L^2(G)$. This representation is known as the Karhunen–Loève expansion.

D.3 A polynomial kernel

We consider the stationary polynomial-type kernel of order $p > 0$, defined as

$$k_{\tau,p}(x, x') = (1 + \tau^{-2}\|x' - x\|^2)^{-p}, \qquad x, x' \in \mathbb{R}^d, \tag{111}$$

where $\tau > 0$ is a rescaling parameter. This kernel induces a Gaussian process with sample paths that are smoother than those of finite-order Matérn processes, but generally less regular than those of the squared exponential kernel. Initially, we show that the polynomial kernel is well defined.

Lemma D.2 (Spectral density of the polynomial kernel)**.** *Let $p > d/2$. Then, the polynomial kernel $k_{\tau,p}$ has the spectral density*

$$f_{\tau,p}(x) = \frac{\Gamma(p - d/2)}{\Gamma(p)(4\pi)^{d/2}} \tau^d \kappa_{p-d/2,\tau^{-1}\sqrt{2p-d}}(x), \qquad x \in \mathbb{R}^d, \tag{112}$$

where

$$\kappa_{\nu,\sigma}(x) = \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\frac{\sqrt{2\nu}}{\sigma} \|x\|_2 \Big)^{\nu} B_\nu \Big(\frac{\sqrt{2\nu}}{\sigma} \|x\|_2 \Big), \qquad x \in \mathbb{R}^d, \tag{113}$$

is the radial function of the Matérn kernel. In particular, the polynomial kernel is a well defined covariance kernel.

Proof. It is a well known fact that the spectral density of the Matérn process with smoothness ν and scaling σ is given by

$$g_{\nu,\sigma}(x) = \frac{\Gamma(\nu + d/2)}{\pi^{d/2}\Gamma(\nu)} \frac{\sigma^d}{(2\nu)^{d/2}} \Big(1 + \frac{\sigma^2 \|x\|^2}{2\nu} \Big)^{-(\nu+d/2)}, \qquad x \in \mathbb{R}^d, \tag{114}$$

which, up to constants, is a version of the radial function underlying the polynomial kernel. By the Fourier inversion theorem for $L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$, then

$$g_{\nu,\sigma}(x) = (2\pi)^{-d} \int e^{-i\langle x,u\rangle} \kappa_{\nu,\sigma}(u) du = (2\pi)^{-d} \int e^{i\langle x,u\rangle} \kappa_{\nu,\sigma}(u) du, \qquad x \in \mathbb{R}^d. \tag{115}$$

By Bochner's theorem, it then follows that the polynomial kernel is a well defined covariance kernel. The explicit formula for the spectral density then follows from setting $\sigma = \sqrt{2\nu}/\tau$ and $\nu = p - d/2$. ∎

Similarly to the squared exponential case, using a fixed scale parameter τ leads to suboptimal performance when estimating a regression function $f_0 \in C^\alpha(\mathbb{R}^d)$ due to a mismatch between the prior regularity and that of f_0. To address this issue, we set $\tau = \tau_n$, which shrinks the kernel as n increases. In the next Lemma, we show that this kernel fits in the framework of (30) which yields the optimal choice $\tau_n = n^{-1/(2\alpha+d)}$.

Lemma D.3 (Concentration function inequality for the polynomial kernel)**.** *Let $\mathcal{X} = [0,1]^d$, $f_0 \in C^\alpha([0,1]^d)$ and $p = d/2 + \ell + 1/2$ for some $\ell \in \mathbb{N}$. Then, the polynomial kernel $k_{\tau_n,p}$ with lengthscale $\tau_n = n^{-1/(2\alpha+d)}$ satisfies the concentration function inequality in Equation* (6) *with $\eta_n = n^{-\alpha/(2\alpha+d)} \log^{1+d/2} n$.*

Proof. By Lemma D.2, the spectral density of the

$$
\begin{aligned}
f_{\tau,p}(x) &= \frac{\Gamma(p-d/2)}{\Gamma(p)(4\pi)^{d/2}} \tau_n^d \kappa_{p-d/2, \tau_n^{-1}\sqrt{2p-d}}(x), \qquad (116)\\
&= \frac{\Gamma(\ell+1/2)}{\Gamma(d/2+\ell+1/2)(4\pi)^{d/2}} \tau_n^d \kappa_{\ell+1/2, \tau_n^{-1}\sqrt{2\ell+1}}(x)\\
&= \frac{\Gamma(\ell+1/2)}{\Gamma(d/2+\ell+1/2)(4\pi)^{d/2}} \frac{\Gamma(\ell+1)}{\Gamma(2\ell+1)} \tau_n^d \exp(-\tau_n \|x\|_2) \sum_{q=0}^{\ell} \frac{(\ell+q)!}{q!(\ell-q)!} \Big(\frac{\sqrt{8\ell+4}\tau_n \|x\|_2}{\sqrt{2\ell+1}} \Big)^{\ell-q},
\end{aligned}
$$

where we have used the fact that for $p = d/2 + \ell + 1/2$, the radial function of the Matérn kernel has an explicit form.

It is easy to check that setting $\tau = \tau_n = 1$ and $\delta < 1$,

$$
\int e^{\delta \|x\|_2} f_{1,p}(x)\, dx < \infty, \qquad (117)
$$

which is condition (3.3) in (30). Consequently, Lemmas 4.3 and 4.6 from the same paper yield the inequalities

$$
\tau_n^{-d} \log^{1+d}(1/(\tau_n \eta_n)) \lesssim n\eta_n^2, \qquad \tau_n^{-d} \lesssim n\eta_n^2 \qquad \text{and} \qquad \tau_n^\alpha \lesssim \eta_n, \qquad (118)
$$

for the concentration function inequality (7) to be satisfied with rate η_n. This is true, however, for the choices $\tau_n = n^{-1/(2\alpha+d)}$ and the rate $\eta_n = n^{-\alpha/(2\alpha+d)} \log^{1+d/2} n$, which finishes the proof. ■

In similar fashion to the examples in Section 4, we obtain a contraction rate result for the approximate posterior. Note, however, that in this case, the required radius for the contraction regime does not coincide with the one required for the Kullback-Leibler divergence to converge to zero.

Proposition D.4 (Result for the polynomial kernel)**.** *In model* (1)*, assume that $\mathcal{X} = [0,1]^d$ and the desgin observations are distributed according to G on $\mathcal{X}$ with bounded Lebesgue density g. Let $f_0 \in C^\alpha(\mathcal{X})$ and let $k_{\tau_n,p}$ be the rescaled polynomial kernel with lengthscale $\tau_n = n^{-1/(2\alpha+d)}$ and $p > d/2 + 2\alpha$. Then, the approximate posterior with truncation radius $r_n = n^{-(p-2\alpha-d/2)/((2\alpha+d)p)}$ satisfies that for arbitrary $M_n \to \infty$*

$$
\Psi_{r_n}(d_H(\cdot, f_0) \geq M_n n^{-\alpha/(2\alpha+d)}) \to 0 \qquad (119)
$$

in probability with respect to $\mathbb{P}_{f_0}^{\otimes n}$ for $n \to \infty$. Further, the number of non-zero entries per row of $\widehat{K}$ is at most of order $m_n = n r_n^d = n^{((2\alpha+d/2)d/p+2\alpha)/(2\alpha+d)}$ with probability converging to one.

Moreover, if $p > d + 2\alpha$ and $r_n = n^{-(p-2\alpha-d)/((2\alpha+d)p)} \log^{1/(2p)} n$, then also

$$
\mathrm{KL}(\Psi_{r_n}, \Pi_n(\cdot, |X, Y)) \to 0 \qquad (120)
$$

in probability with respect to $\mathbb{P}_{f_0}^{\otimes n}$ for $n \to \infty$. The number of non-zero entries per row of $\widehat{K}$ then is at most of order $m_n = n r_n^d = n^{((2\alpha+d)d/p+2\alpha)/(2\alpha+d)}$ with probability converging to one.

Proof. We initially consider the first statement. As in the examples before, we have the estimate

$$\begin{aligned}\|\widehat{K} - K\|_{\mathrm{op}} &\le \max_{i\le n} \sum_{X_j : \|X_j - X_i\|_2 > r_n} k_{\tau_n,p}(X_i, X_j) \\ &\le \max_{i\le n} \sum_{X_j : \|X_j - X_i\|_2 > r_n} \Big(1 + \frac{r_n^2}{\tau_n^2}\Big)^{-p} \lesssim n\Big(\frac{r_n}{\tau_n}\Big)^{-2p} \\ &= n^{-2\alpha/(2\alpha+d)} \to 0.\end{aligned} \tag{121}$$

In particular, $\|\widehat{K} - K\|_{\mathrm{op}} < \sigma^2/2$ eventually for all $n \in \mathbb{N}$. Further, by Theorem 3.1,

$$\begin{aligned}\mathrm{KL}(\Psi_{r_n}, \Pi_n(\cdot|X,Y)) &\lesssim M_n \|\widehat{K} - K\|_{\mathrm{op}} (\operatorname{tr} K K_\sigma^{-1} + n\eta_n^2) \\ &\lesssim M_n n n^{-2\alpha/(2\alpha+d)} = M_n n \eta_n^2.\end{aligned} \tag{122}$$

As in the proof of Theorem 3.2, we now obtain the contraction rate η_n around f_0. For the statement about the non-zero row entries, we consider again Lemma C.4 with $n r_n^d \asymp n^{((2\alpha+d/2)d/p+2\alpha)/(2\alpha+d)}$ and $\delta = \|g\|_\infty r_n^d \mathrm{Vol}(B_1(0))$.

The second statement follows from the same bounds with the new choice of r_n. ■

Remark D.5 (An improved choice of the radius). In the proof of Proposition D.4, we have used the straigt forward estimate $\operatorname{tr} K K_\sigma^{-1} \le n$. As in the other examples, the trace term should also satisfy the bound $\operatorname{tr} K K_\sigma^{-1} \lesssim n\eta_n^2$. However, proving the required bounds for the population eigenvalues goes beyond the scope of this paper. Note that the result from (21) that was used in Lemma C.5 does not apply here.

For completeness, we note that under the assumption $\operatorname{tr} K K_\sigma^{-1} \lesssim n\eta_n^2$, the required radius for contraction could be improved to $r_n = n^{-(p-\alpha-d/2)/((2\alpha+d)p)} \log^{1/(2p)} n$ resulting in $m_n = n r_n^d = n^{((\alpha+d/2)d/p+2\alpha)/(2\alpha+d)} \log^{d/(2p)} n$ of non-zero row entries. Analogously, for the convergence of the Kullback-Leibler divergence, we would need $r_n = n^{-(p-\alpha-d)/((2\alpha+d)p)}$ $\log^{1/(2p)} n$ resulting in $m_n = n r_n^d = n^{((\alpha+d)d/p+2\alpha)/(2\alpha+d)} \log^{d/(2p)} n$ of non-zero row entries.

D.4 Additional numerical results

We consider one-dimensional nonparametric regression with observations

$$Y_i = f_0(X_i) + \sigma\varepsilon_i, \qquad i = 1, \dots, n, \tag{123}$$

where the design points X_i and errors ε_i are mutually independent standard normal random variables. The noise standard deviation is $\sigma = 0.2$. The true regression function is given by

$$f_0(x) = |x - 0.4|^\beta - |x - 0.2|^\beta, \qquad \beta = 0.6. \tag{124}$$

This function has a smoothness equal to $\beta = 0.6$. We use the rescaled squared exponential kernel as a prior.

We compare the full Gaussian process posterior mean with the banded approximation, the conjugate gradient (CG) approximation, and distributed Method for $n \in \{10\,000, 20\,000, 30\,000\}$. Each configuration is evaluated over $R = 100$ repetitions. For a test grid $x_1, \dots, x_L$, the error in repetition b is

$$\mathrm{MSE}^{(b)} = \frac{1}{L} \sum_{\ell=1}^{L} \big(\widehat{f}^{(b)}(x_\ell) - f_0(x_\ell)\big)^2. \tag{125}$$

We report its average and empirical standard deviation across repetitions.

The banding radii are $r \in \{0.75 r_n, r_n, 1.25 r_n\}$, where $r_n = \tau_n \sqrt{2 \log n}$ and $\tau_n = n^{-1/2.2}$. For distributed Method, we consider the number of local machines $m \in \{2, 5, 50\}$. For CG, we compare a maximum budgets of 100 and 150 iterations with the reference budgets $q_n = 303, 447, 559$, respectively.

Method and setting	Mean MSE (SD)	S_{fit}	S_{total}
$n = 10\,000$			
Full GP	2.610 (0.315)	1.00	1.00
Banded ($r = 0.75 r_n$)	2.610 (0.315)	5.53	4.92
Banded ($r = r_n$)	2.610 (0.315)	3.37	3.17
Banded ($r = 1.25 r_n$)	2.610 (0.315)	2.25	2.18
CG ($q_{\max} = 100$)	2.642 (0.315)	1.10	1.11
CG ($q_{\max} = 150$)	2.611 (0.315)	0.79	0.80
CG ($q_{\max} = 303$)	2.610 (0.315)	0.41	0.42
Distributed ($m = 2$)	2.670 (0.293)	2.38	2.29
Distributed ($m = 5$)	3.091 (0.353)	10.48	8.24
Distributed ($m = 50$)	6.301 (0.706)	142.13	28.61
$n = 20\,000$			
Full GP	1.858 (0.209)	1.00	1.00
Banded ($r = 0.75 r_n$)	1.858 (0.209)	9.16	8.54
Banded ($r = r_n$)	1.858 (0.209)	5.49	5.28
Banded ($r = 1.25 r_n$)	1.858 (0.209)	3.66	3.58
CG ($q = 100$)	1.968 (0.215)	1.87	11.87
CG ($q = 150$)	1.863 (0.209)	1.33	1.33
CG ($q = 447$)	1.858 (0.209)	0.53	0.53
Distributed ($m = 2$)	1.899 (0.191)	3.24	3.18
Distributed ($m = 5$)	2.144 (0.202)	12.63	11.44
Distributed ($m = 50$)	6.174 (0.557)	307.06	81.87
$n = 30\,000$			
Full GP	1.516 (0.168)	1.00	1.00
Banded ($r = 0.75 r_n$)	1.516 (0.168)	13.21	12.57
Banded ($r = r_n$)	1.516 (0.168)	7.81	7.59
Banded ($r = 1.25 r_n$)	1.516 (0.168)	5.07	4.98
CG ($q = 100$)	1.727 (0.194)	2.63	2.62
CG ($q = 150$)	1.525 (0.166)	1.88	1.87
CG ($q = 559$)	1.516 (0.168)	0.67	0.67
Distributed ($m = 2$)	1.569 (0.154)	2.94	2.92
Distributed ($m = 5$)	1.763 (0.152)	15.07	14.22
Distributed ($m = 50$)	5.556 (0.392)	403.85	149.39

Table 2: Accuracy and computational performance over 100 repetitions. Both mean MSE and its standard deviation, reported in parentheses, are multiplied by 10^3. The CG parameter q denotes the iteration budget; iterations may stop earlier.

Fitting and prediction times are measured separately, and their sum defines total runtime. the average paired speedups

$$S_{\text{fit}} = \frac{1}{R}\sum_{b=1}^{R} \frac{T^{(b)}_{\text{full,fit}}}{T^{(b)}_{\text{method,fit}}}, \qquad S_{\text{total}} = \frac{1}{R}\sum_{b=1}^{R} \frac{T^{(b)}_{\text{full,total}}}{T^{(b)}_{\text{method,total}}}. \tag{126}$$

T_{fit} is the time needed to calculate the posterior mean and $T_{\text{prediction}}$is the time needed to calculate the mean in new 2000 test points. A speedup larger than one therefore indicates a computational gain relative to the full GP.

NeurIPS Paper Checklist

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 abstract and introduction accurately summarize the main contributions and theoretical results of the paper, and these claims are consistent with the analysis and scope developed in the manuscript.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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 provide also lower bound on the number of neighbours one has to select for accurate inference and approximation. In certain cases this is proportional to n, hence no speed up is possible. We also discuss that our paper does not considers adaptation (i.e. data driven choice of m.)

   Guidelines:

   - The answer [N/A] 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: All theoretical results are stated with explicit assumptions, and complete and correct proofs are provided in the main text or supplementary material.

Guidelines:

- The answer [N/A] 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 provide all the details needed to reproduce the numerical results in the Examples section.

Guidelines:

- The answer [N/A] 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: The code to produce our simulations is available as supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: All these information are given in the numerical experiment section.

   Guidelines:

   - The answer [N/A] 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: Yes, the tables report the standard deviations of the estimators.

   Guidelines:

   - The answer [N/A] 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 report the run times of the algorithms. The simulations were run on a Intel(R) Xeon(R) Silver 4110 CPU and an nvidia A40 unit.

   Guidelines:

   - The answer [N/A] 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 follow the code of ethics.

   Guidelines:

   - The answer [N/A] 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: [No]

    Justification: This paper focuses on theoretical aspects of Gaussian process approximation and does not explicitly discuss societal impacts. While such methods may be used in applications involving decision-making and uncertainty quantification, analyzing their societal implications is beyond the scope of this work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The experiments are conducted on synthetic data and do not involve the release of pretrained models or sensitive datasets. As such, no specific safeguards for misuse are required beyond standard responsible use of research code and experimental procedures.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: The experiments rely on standard open-source scientific computing and Gaussian process software packages, including NumPy, SciPy, PyTorch, GPyTorch, and pandas.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: No new assets are introduced.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM are used only for editing (e.g., grammar, spelling, word choice).

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.