

Uncertainty quantification for sparse spectral variational approximations in Gaussian process regression

Dennis Nieman

Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands
e-mail: d.nieman@vu.nl

Botond Szabo

Department of Decision Sciences and BIDSa, Bocconi University, Italy
e-mail: botond.szabo@unibocconi.it

Harry van Zanten

Department of Mathematics, Vrije Universiteit Amsterdam, The Netherlands
e-mail: j.h.van.zanten@vu.nl

Abstract: We investigate the frequentist guarantees of the variational sparse Gaussian process regression model. In the theoretical analysis, we focus on the variational approach with spectral features as inducing variables. We derive guarantees and limitations for the frequentist coverage of the resulting variational credible sets. We also derive sufficient and necessary lower bounds for the number of inducing variables required to achieve minimax posterior contraction rates. The implications of these results are demonstrated for different choices of priors. In a numerical analysis we consider a wider range of inducing variable methods and observe similar phenomena beyond the scope of our theoretical findings.

MSC2020 subject classifications: Primary 62G20; secondary 62G05, 62G08, 62G15.

Keywords and phrases: Variational Bayes, uncertainty quantification, nonparametric regression, inducing variables method, Bayesian asymptotics.

Received December 2022.

1. Introduction

One of the key challenges in Bayesian statistics is to approximate intractable or computationally infeasible posterior distributions. This problem is becoming even more pronounced in applications where the amount of available information is rapidly growing, further increasing the complexity of the posterior. Variational methods provide a convenient way to overcome such computational issues in Bayesian statistics. The variational approximation starts by selecting an appropriate class of distributions, referred to as the variational class. Then

the complex posterior distribution is approximated by its projection on this variational class with respect to the Kullback-Leibler divergence. The challenge in choosing the variational class is twofold: firstly, the variational posterior should reduce computational complexity and if possible increase the interpretability of the distribution; secondly, for meaningful inference it is crucial that the variational posterior has good statistical properties. For an overview of variational Bayes methods we refer to the review article [5].

Although variational Bayes approximations are routinely used in practice, up to recently they were considered black box procedures with very limited theoretical underpinning. In the last few years the asymptotic properties of the variational posterior have been investigated. Abstract results were derived for posterior contraction rates and applied to various high-dimensional and non-parametric models, considering typically mean-field variational classes; see for instance [37, 36, 1, 19, 20]. However, almost all of these results focus on the recovery of the underlying true parameter of interest and do not address the quality of uncertainty quantification.

In fact, one of the main appeals and strengths of the Bayesian paradigm is that it provides a probabilistic solution to the statistical problem. The posterior distribution can be used to quantify the remaining uncertainty about the parameter of interest. In practice this uncertainty is usually visualised by plotting credible regions. These are subsets of the parameter space with prescribed posterior probability (typically 95%). In parametric models the celebrated Bernstein-von Mises theorem [16, 29] provides asymptotic frequentist coverage guarantees for credible sets under mild assumptions, meaning that credible sets can be interpreted as frequentist confidence sets. In high-dimensional and non-parametric frameworks such a strong guarantee does not automatically hold in general (see e.g. [8]). Nevertheless, by now we have a relatively good understanding of how to tune the prior to achieve asymptotic confidence guarantees, see for instance [15, 7, 25, 21].

Despite its importance, so far hardly any results are available on the frequentist reliability of the variational Bayesian uncertainty quantification, where the real credible sets are replaced by approximate credible sets derived from the variational posterior. In fact, many of the available results are rather negative, showing that (mean-field) variational methods are often over-confident in the sense that they substantially underestimate the uncertainty of the procedure, see for instance [4, 5] for some standard examples. There are only a few positive results available. In [34] a variational version of the Bernstein-von Mises theorem was derived in parametric models, while in [11] a correction was proposed using linear response methods to recover the original posterior covariance structure. However, both of these results consider only parametric models and it is in general unclear how variational credible sets behave in high- and infinite-dimensional settings. A related result we are aware of can be found in the recent paper [27], where confidence statements for a variational posterior are obtained in a special bandit-like regression setting similar to the results of [23] for the true posterior. Although interesting, these results and the underlying techniques do not transfer to the nonparametric regression setting we consider in this paper.

In our analysis we focus on the popular and routinely used Gaussian process (GP) regression model. Exact computation of the posterior quickly becomes infeasible in practice since the computational cost scales cubically in the sample size. To overcome this problem, a sparse variational approximation method was proposed in [26]. The variational class is parametrized by so called inducing variables which are fitted to the posterior. This approach has become increasingly popular in the machine learning community and has been applied in various settings including deep Gaussian processes and solving inverse problems. Recently theoretical guarantees were also derived for it. In [6] the average Kullback-Leibler distance between the variational and true posterior was studied, while in [17] minimax posterior contraction rates were derived. However, none of these results provided guarantees for the frequentist validity of the resulting uncertainty quantification, which is arguably one of the main aims in the Bayesian analysis.

We focus in our theoretical analysis on a specific choice of inducing variables which we call population spectral features, allowing (relatively) tractable mathematical analysis. In our numerical analysis we also show that with other choices of inducing variables similar behaviour is obtained. We observe that, in contrast to the simple parametric examples using mean-field variational approximations, in the nonparametric GP regression framework the variational posterior provides, from a frequentist perspective, reliable uncertainty statements for appropriately tuned priors. In fact, the good coverage property does not depend on the number of inducing variables used in the procedure. Besides the coverage of variational credible sets we also derive lower bounds for the number of inducing variables one has to use to achieve minimax contraction rates. This complements the contraction rate guarantees given in [17], where the sharpness of the lower bound was conjectured, but not verified. To achieve this we use a different proof technique than in [17]. We apply a variational version of the standard kernel ridge regression method [22]. This direct approach provides more control of posterior properties. Finally, we apply our abstract results to two GP priors with polynomially and exponentially decaying eigenvalues, respectively.

Contributions We summarize our contributions below:

- We give an explicit formula for the contraction rate of the variational posterior in terms of the prior and the true regression function. This gives a condition on the prior and the minimal number of inducing variables in the variational approximation needed to obtain minimax contraction rates.
- If the number of inducing variables is too low, the contraction rate is sub-optimal, regardless of the choice of prior.
- Irrespective of the number of inducing variables, variational credible sets cover the truths that are at least as smooth as the prior, whereas coverage may be bad if the prior over-smoothes the true regression function.

Outline In the next section, we describe the Gaussian process regression model studied in this paper. We recall the details of the variational procedure

with inducing variables and derive a connection with kernel ridge regression, used in the proofs. Lastly we introduce the specific choice of inducing variables considered in the theoretical analysis. In Section 3, we develop a theory for contraction rates. Section 4 consists of the theory on uncertainty quantification. In Section 5 these results are applied to two specific priors, with polynomially and exponentially decaying eigenvalues, respectively. We conclude with a numerical analysis, including various inducing variable methods in Section 6. The proofs are deferred to the Appendix. In Section A we prove the more abstract results, while the proofs for the examples are given in Section B.

Notation For sequences a_n, b_n of non-negative real numbers, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n \leq Cb_n$ for all $n \in \mathbb{N}$. We write $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$ hold. We indicate with an apostrophe the transpose A' of a matrix A .

2. Variational approximations for Gaussian process regression

Throughout the paper we investigate the nonparametric regression model

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n, \tag{2.1}$$

with i.i.d. design points x_i with respect to some common probability measure μ on some $\mathcal{X} \subseteq \mathbb{R}^d$, and i.i.d. mean-zero Gaussian measurement errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, for some $\sigma^2 > 0$. We view the unknown parameter f as an element of the function space $L^2(\mathcal{X}, \mu)$. We endow f with a centered Gaussian process (GP) prior Π determined by its covariance kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.

The GP prior is conjugate for the regression model (2.1), which means that the posterior is also a GP. However, the computational and memory costs of obtaining the posterior are $O(n^3)$ and $O(n^2)$, respectively, which is prohibitive for large data sets. Therefore, in practice various approximation methods are applied for inference, see [18] for a detailed discussion. One of the most commonly used approaches is the sparse variational approximation using inducing variables, proposed by [26]. In the next subsection we give a brief summary of inducing variable variational Bayes methods in general and we will focus on a specific, analytically convenient version of the method, using population spectral features.

2.1. Variational Bayes with inducing variables

In the variational framework, the posterior distribution is approximated by projecting it onto an appropriately selected class \mathcal{Q} of probability measures on $L^2(\mathcal{X}, \mu)$ with respect to the Kullback-Leibler divergence. Letting $\Pi(\cdot | \mathbf{x}, \mathbf{y})$ denote the true posterior, the variational posterior $\Psi(\cdot | \mathbf{x}, \mathbf{y})$ is defined as

$$\Psi(\cdot | \mathbf{x}, \mathbf{y}) = \arg \min_{\Psi \in \mathcal{Q}} D_{\text{KL}}(\Psi || \Pi(\cdot | \mathbf{x}, \mathbf{y})), \tag{2.2}$$

where D_{KL} denotes Kullback-Leibler divergence.

In the inducing variables framework, the variational class \mathcal{Q} is constructed using a collection $\mathbf{u} = (u_1, \dots, u_m)$ of (known, specified) bounded linear functionals evaluated at f . The linearity guarantees that the distribution of \mathbf{u} under Π is m -dimensional multivariate Gaussian. Furthermore, the prior conditional on \mathbf{u} is another GP law. To obtain a low-dimensional optimization problem and preserve aspects of the prior distribution, [26] proposed to fit a variational distribution of \mathbf{u} to its posterior, while keeping the conditional prior distribution for $f|\mathbf{u}$. More concretely, the sparse inducing variable variational class consists of the distributions

$$\Psi = \int \Pi(\cdot | \mathbf{u}) d\Psi_{\mathbf{u}}(\mathbf{u}), \quad (2.3)$$

where $\Psi_{\mathbf{u}}$ is any non-degenerate m -dimensional Gaussian distribution, indexing the variational class \mathcal{Q} . This way the posterior information is compressed into the fitted m -dimensional distribution of \mathbf{u} . Nevertheless, Ψ is still a nonparametric distribution on $L^2(\mathcal{X}, \mu)$ which is equivalent to the prior.

The variational posterior is computed by finding the distribution $\Psi_{\mathbf{u}}$ that minimizes the Kullback-Leibler divergence between Ψ given in (2.3) and the true posterior. As is shown in [26], a unique solution exists and can be computed analytically. The corresponding variational posterior is also GP with respective mean and covariance function

$$\hat{f}_m(x) = K_{xu}(\sigma^2 K_{uu} + K_{uf} K_{fu})^{-1} K_{uf} \mathbf{y}, \quad (2.4)$$

$$\hat{k}_m(x, y) = k(x, y) - K_{xu} K_{uu}^{-1} K_{uy} + K_{xu} (K_{uu} + \sigma^{-2} K_{uf} K_{fu})^{-1} K_{uy}. \quad (2.5)$$

Here $\mathbf{y} = (y_1, \dots, y_n)$ is the vector of response variables, K_{uu} is the covariance matrix of \mathbf{u} under Π (with entries $\text{cov}_{\Pi}(u_i, u_j)$), and similarly $K_{fu} = K'_{uf}$ is the $n \times m$ matrix with entries $\text{cov}_{\Pi}(f(x_i), u_j) = \Pi f(x_i) u_j$, and $K_{ux} = K'_{xu} = \text{cov}_{\Pi}(\mathbf{u}, f(x))$. We note that in the special case $\mathbf{u} = \mathbf{f}$, the true posterior is recovered. The above formulas were derived in [26] for the inducing points method ($u_j = f(z_j)$ for $z_j \in \mathcal{X}$, $j = 1, \dots, m$), but the same computations hold for any choice of the inducing variables. For completeness we provide the details in Appendix C.

The theoretical properties of this approach have been investigated in [6, 17] for various choices of inducing variables. The first paper deals with the accuracy of the variational approximation of the original posterior with respect to the Kullback-Leibler divergence. In the second paper, upper bounds were derived for the posterior contraction rate. We extend the latter result by using a different, kernel ridge regression technique allowing sharper control of the approximation. This analysis allows us to derive lower bounds for the contraction rate and to investigate the frequentist coverage properties of the credible sets resulting from the variational approximation.

2.2. Kernel ridge regression

The posterior mean, which is the maximum a posteriori in case of Gaussian processes, can equivalently be obtained as a kernel ridge regression (KRR) estimator. Let \mathbb{H} be the reproducing kernel Hilbert space (RKHS) associated with the GP kernel k . Then (see e.g. [14]) the mean of the original posterior equals

$$\arg \min_{f \in \mathbb{H}} \sum_{i=1}^n (y_i - f(x_i))^2 + \sigma^2 \|f\|_{\mathbb{H}}^2.$$

It is not difficult to see that the variational posterior mean (2.4) can also be viewed as a KRR estimator, for an appropriate choice of the RKHS. Since the inducing variables u_j are linear functionals of f , the functions

$$h_j : \mathcal{X} \rightarrow \mathbb{R}, \quad x \mapsto \text{cov}_{\Pi}(f(x), u_j) = \Pi f(x) u_j$$

are elements of \mathbb{H} (see [28]). Let \mathbb{H}_m denote the linear subspace of \mathbb{H} spanned by the functions h_1, \dots, h_m . The variational posterior mean (2.4) is an element of \mathbb{H}_m . The following lemma states that the inducing variable variational posterior mean minimizes the same objective function as the mean of the true posterior, but over the subclass $\mathbb{H}_m \subset \mathbb{H}$.

Lemma 1. *The variational posterior mean \hat{f}_m given in (2.4) satisfies*

$$\hat{f}_m = \arg \min_{f \in \mathbb{H}_m} \sum_{i=1}^n (y_i - f(x_i))^2 + \sigma^2 \|f\|_{\mathbb{H}}^2. \tag{2.6}$$

For a similar result in context of the more specific Nyström approximation method we refer to [35]. Although the above lemma holds for arbitrary choices of the inducing variables, in the upcoming sections we focus specifically on the population spectral features approach. We believe that what we infer from our results holds more generally, as illustrated in our numerical analysis.

2.3. Population spectral features

In our theoretical analysis we focus on a choice of inducing variables that gives the variational posterior the interpretation of a spectral approximation to the true posterior. We assume that the prior covariance kernel k is continuous and $k \in L^2(\mathcal{X} \times \mathcal{X}, \mu \times \mu)$, so that it has a Mercer decomposition

$$k(x, y) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(x) \varphi_j(y), \tag{2.7}$$

where λ_j is a decreasing, summable sequence of nonnegative numbers and $(\varphi_j)_{j \in \mathbb{N}}$ is an orthonormal basis of $L^2(\mathcal{X}, \mu)$. Under the current assumptions,

the series (2.7) converges not only in the L^2 sense but also uniformly on compact subsets in the support of $\mu \times \mu$ (see [24], Corollary 3.5). We also associate with k an operator $T : L^2(\mathcal{X}, \mu) \rightarrow L^2(\mathcal{X}, \mu)$ given by

$$Tg(x) = \int_{\mathcal{X}} k(x, y)g(y) d\mu(y), \quad (2.8)$$

which is called the covariance operator. The identity (2.7) is equivalent to the decomposition $T = \sum_{j=1}^{\infty} \lambda_j \langle \cdot, \varphi_j \rangle \varphi_j$. We assume that the set of functions $(\varphi_j)_{j \in \mathbb{N}}$ is uniformly bounded:

Assumption 2. *The functions φ_j satisfy*

$$\sup\{|\varphi_j(x)| : j \in \mathbb{N}, x \in \mathcal{X}\} =: C_\varphi < \infty. \quad (2.9)$$

In the theoretical part of this paper, we consider the inducing variables

$$u_j = \langle f, \varphi_j \rangle = \int f(x)\varphi_j(x) d\mu(x), \quad j = 1, \dots, m,$$

which we refer to as *population spectral features*. We note that this approach requires the explicit knowledge of the basis functions φ_j . Let us write $\varphi_{1:m}(x) = (\varphi_1(x), \dots, \varphi_m(x))$, and take Φ to be the $n \times m$ matrix whose i -th row is $\varphi_{1:m}(x_i)'$. It follows from Fubini's theorem that $\Pi u_i u_j = \lambda_j \delta_{ij}$ and $\Pi f(x)u_j = \lambda_j \varphi_j(x)$, so in this case

$$\begin{aligned} K_{\mathbf{u}\mathbf{u}} &= \text{diag}(\lambda_1, \dots, \lambda_m) =: \Lambda, \\ K_{x\mathbf{u}} &= \varphi_{1:m}(x)' \Lambda, \\ K_{f\mathbf{u}} &= \Phi \Lambda. \end{aligned} \quad (2.10)$$

In view of (2.4) and (2.5), the variational posterior $\Psi(\cdot | \mathbf{x}, \mathbf{y})$ is the law of a Gaussian process with mean and covariance function

$$\hat{f}_m(x) = \varphi_{1:m}(x)' (\Lambda^{-1} + \sigma^{-2} \Phi' \Phi)^{-1} \Phi' \mathbf{y} / \sigma^2, \quad (2.11)$$

$$\begin{aligned} \hat{k}_m(x, y) &= k(x, y) - \varphi_{1:m}(x)' \left(\Lambda - (\Lambda^{-1} + \sigma^{-2} \Phi' \Phi)^{-1} \right) \varphi_{1:m}(y), \\ &= \varphi_{1:m}(x)' (\Lambda^{-1} + \sigma^{-2} \Phi' \Phi)^{-1} \varphi_{1:m}(y) + \sum_{j=m+1}^{\infty} \lambda_j \varphi_j(x) \varphi_j(y), \end{aligned} \quad (2.12)$$

where the last line follows from (2.7). We note that in this setting $\mathbb{H}_m \subset \mathbb{H}$ is the linear span of the first m basis functions. In the next three sections, we develop theory for the population spectral features variational posterior, which is fully characterised by (2.11) and (2.12).

3. Contraction rates

First we investigate the asymptotic recovery property of the population spectral features variational posterior. An upper bound for the contraction rate was

already derived in [17] for general inducing variables methods. However, the implicit approach based on GP concentration functions do not directly result in lower bounds for the contraction and therefore does not imply a lower bound on the number of inducing variables one has to apply to achieve minimax contraction rates. Therefore, in this article we take a more direct approach using kernel ridge regression techniques to understand the limitations of the variational approximation.

In this section we explicitly decompose the contraction rate to bias and variance terms, which in turn illuminates how optimal contraction imposes a condition on the minimal dimension m of the approximation. It also shows why the recovery accuracy does not improve any further by increasing the dimension m beyond this minimum. Moreover, a converse result to the contraction rate statement is also given, which says that the variational posterior does not contract at the optimal rate if m is too low.

3.1. Convergence rate of the mean of the variational posterior

We assume that the data are generated according to some true $f_0 \in L^2(\mathcal{X}, \mu)$. Let us denote by P_0 the measure on $\mathcal{X}^n \times \mathbb{R}^n$ under which (\mathbf{x}, \mathbf{y}) satisfies (2.1) with $f = f_0$. First, we consider the variational posterior mean \hat{f}_m as an estimator of f_0 . The next lemma decomposes the mean squared error of the estimator \hat{f}_m into squared bias and variance terms of the estimator \hat{f}_m under P_0 . The bias term B_n consists of two parts. The first part is accounting for the estimation error in the subspace \mathbb{H}_m , while the second term equals the squared norm of the orthogonal projection of f_0 onto the orthogonal complement \mathbb{H}_m^\perp of \mathbb{H}_m in $L^2(\mathcal{X}, \mu)$. The other term W_n is the variance term. The proof of the lemma is deferred to Section A.6.

Lemma 3. *Define*

$$\nu_j = \frac{n\lambda_j}{\sigma^2 + n\lambda_j}. \tag{3.1}$$

Let $m = m_n$ be such that $m^2 n^{-1} \log n \rightarrow 0$ as $n \rightarrow \infty$. Then for any bounded $f_0 \in L^2(\mathcal{X}, \mu)$,

$$P_0 \|\hat{f}_m - f_0\|^2 \lesssim B_n + W_n, \tag{3.2}$$

where

$$B_n = \sum_{j=1}^m (1 - \nu_j)^2 \langle f_0, \varphi_j \rangle^2 + \sum_{j>m} \langle f_0, \varphi_j \rangle^2 \quad \text{and} \quad W_n = \frac{1}{n} \sum_{j=1}^m \nu_j^2. \tag{3.3}$$

Remark 4. *The mild technical assumption $m^2 n^{-1} \log n$ allows for a clean presentation. It can be weakened, but this requires the introduction of a technical term on the right-hand side of (3.2) as in [12].*

3.2. Contraction rate of the variational posterior

The contraction rate of the variational posterior is determined by the squared bias B_n and variance W_n of the posterior mean introduced above, as well as the term V_n introduced below, which characterises the spread of the variational posterior.

Theorem 5. *Let $f_0 \in L^2(\mathcal{X}, \mu)$ be a bounded function and $m = m_n \rightarrow \infty$ such that $m^2 n^{-1} \log n \rightarrow 0$. Then*

$$P_0 \Psi(\|f - f_0\| \geq M_n \epsilon_n | \mathbf{x}, \mathbf{y}) \rightarrow 0 \tag{3.4}$$

for arbitrary $M_n \rightarrow \infty$, where

$$\epsilon_n^2 = B_n + V_n + W_n,$$

with B_n and W_n as in (3.3), and

$$V_n = \frac{1}{n} \sum_{j=1}^m \nu_j + \sum_{j>m} \lambda_j. \tag{3.5}$$

The variance term W_n is always dominated by the posterior variance term V_n , so it does not increase the rate of contraction. The proof of the theorem is given in Section A.1.

Below we investigate three terms in more detail. We present an alternative formulation which is more convenient to apply in our examples and also sheds light on how the rate depends on the dimension m of the variational approximation.

Remark 6. *By considering separately the cases that $n\lambda_j \geq 1$ and $n\lambda_j < 1$, it follows that*

$$\nu_j \asymp 1 \wedge n\lambda_j, \quad 1 - \nu_j = \frac{\sigma^2}{n\lambda_j} \nu_j \asymp 1 \wedge (n\lambda_j)^{-1}. \tag{3.6}$$

Let us introduce

$$J_n = \max\{j : n\lambda_j \geq 1\}, \tag{3.7}$$

denoting the elbow point of the above quantities. Then the terms in the contraction rate (3.4) can alternatively be written as

$$B_n \asymp \sum_{j=1}^{m \wedge J_n} (n\lambda_j)^{-2} \langle f_0, \varphi_j \rangle^2 + \sum_{j>m \wedge J_n} \langle f_0, \varphi_j \rangle^2, \tag{3.8}$$

$$V_n \asymp \frac{m \wedge J_n}{n} + \sum_{j>m \wedge J_n} \lambda_j, \quad W_n \asymp \frac{m \wedge J_n}{n} + n \sum_{j=J_n+1}^m \lambda_j^2. \tag{3.9}$$

These identities follow immediately from (3.6). They show that the contraction rate does not improve any further if m is increased beyond J_n .

3.3. Lower bounds

Next we derive lower bounds for the variational posterior contraction rate. This in turn implies a lower bound on the number of inducing variables needed to achieve minimax recovery of the truth. Let us assume that f_0 belongs to the β -Sobolev space

$$\mathcal{S}^\beta := \{f \in L^2(\mathcal{X}, \mu) : \|f\|_\beta < \infty\}, \quad \|f\|_\beta = \left(\sum_{j=1}^\infty j^{2\beta/d} \langle f, \varphi_j \rangle^2 \right)^{1/2} \quad (3.10)$$

for some $\beta > 0$. The minimax convergence rate for $f_0 \in \mathcal{S}^\beta$ is $n^{-\beta/(d+2\beta)}$.

The abstract results of Theorem 5 imply that for appropriately chosen eigenvalues λ_j the variational posterior contracts around the truth with the minimax optimal rate; see Section 5 for two specific examples. In both of these examples the number of inducing variables has to be at least of order $n^{d/(d+2\beta)}$, which in fact corresponds to the L^2 -entropy of the β -Sobolev ball, also referred to as the “effective dimension” of the model. The same minimal dimension was obtained for various choices of priors and inducing variable methods in [17]. So far, there was no theoretical underpinning available for the sharpness of this threshold. The following theorem aims to fill this gap for the population spectral features variational approach: it shows optimal posterior contraction can not be achieved when m is below this threshold. Concretely, if m grows as a power of n that is strictly smaller than $d/(d + 2\beta)$, the convergence and contraction rate are strictly slower than the optimal rate $n^{-\beta/(d+2\beta)}$. The proof of the theorem is given in Section A.2.

Theorem 7. *Let $m \asymp n^r$, where $0 < r < \frac{d}{d+2\beta}$. Then there exists $f_0 \in \mathcal{S}^\beta$ and $0 < p < 2\beta/(d + 2\beta)$ such that*

$$\mathbb{P}_0 \|\hat{f}_m - f_0\|^2 \gtrsim n^{-p}, \quad (3.11)$$

irrespective of the choice of prior. Moreover, possibly along a subsequence, we have

$$\mathbb{P}_0 \Psi(\|f - f_0\|^2 \leq Mn^{-p} \mid \mathbf{x}, \mathbf{y}) \rightarrow 0 \quad (3.12)$$

for any $M > 0$.

4. Uncertainty quantification

In this section we present our main results on the frequentist validity of Bayesian uncertainty quantification resulting in from the variational approximation. To this end, let us fix $\gamma \in (0, 1)$ and consider the ball

$$C_\gamma := \{f : \|f - \hat{f}_m\| \leq \rho_n\}, \quad (4.1)$$

where the radius ρ_n is chosen such that $\Psi(f \in C_\gamma \mid \mathbf{x}, \mathbf{y}) = 1 - \gamma$. This set is referred to as the $(1 - \gamma)$ -credible ball of the variational posterior. In the next

theorem, we first study the asymptotic size of the radius ρ_n under the frequentist assumption that some $f_0 \in L^2(\mathcal{X}, \mu)$ generates the data. In short, under P_0 , the asymptotic radius of a credible set is of the order V_n , which was defined in (3.5). This is in line with the remark made earlier that V_n characterises the spread of the variational posterior. The proof is given in Appendix A.3.

Theorem 8. *Suppose that $m = m_n$ is such that $m^2 n^{-1} \log n \rightarrow 0$ as $n \rightarrow \infty$. Then there exists a positive constant C such that the credible ball C_γ defined in (4.1) has radius satisfying*

$$C^{-1}V_n \leq \rho_n^2 \leq CV_n,$$

with P_0 -probability tending to 1 for any $f_0 \in L^2(\mathcal{X}, \mu)$.

We now consider the frequentist coverage of the credible set C_γ , i.e. we are interested in the probability

$$P_0(f_0 \in C_\gamma) = P_0(\|\hat{f}_m - f_0\| \leq \rho_n). \quad (4.2)$$

The next result provides guarantees but also limitations for achieving good coverage. The proof is deferred to Section A.4.

Theorem 9. *For B_n, W_n and V_n defined in (3.3) and (3.5), respectively, consider the ratio*

$$R_n = \frac{B_n + W_n}{V_n}. \quad (4.3)$$

Under the conditions of Theorem 8 and assuming $f_0 \in L^2(\mathcal{X}, \mu)$ is bounded,

1. if $R_n \rightarrow 0$, then $P_0(\|\hat{f}_m - f_0\| \leq \rho_n) \rightarrow 1$;
2. if $R_n \lesssim 1$, then $P_0(\|\hat{f}_m - f_0\| \leq M_n \rho_n) \rightarrow 1$ for any sequence $M_n \rightarrow \infty$;
3. if $R_n \rightarrow \infty$, then $P_0(\|\hat{f}_m - f_0\| \leq M \rho_n) \rightarrow 0$ for any $M > 0$.

In view of (4.2) the above theorem presents the frequentist coverage properties of the variational credible ball. It is determined by the relation of the mean squared error $P_0\|\hat{f}_m - f_0\|^2$, studied in Lemma 3, and the radius ρ_n of the credible set, investigated in Theorem 8. The first two statements are in line with the intuition that good coverage follows from the credible set's radius being larger than the loss. In case the radius and loss are asymptotically comparable, good coverage can be achieved by (slightly) blowing up the credible set with a growing factor M_n . The third statement is the converse, i.e. if the loss exceeds the radius then coverage will be bad.

We note that in statements 2 and 3, the ratio R_n may be replaced by B_n/V_n , since the variance of the posterior mean W_n is always bounded by the denominator V_n . The numerator B_n represents the (order of the) squared bias of the variational posterior mean and the denominator V_n corresponds to the variance of the variational posterior. So Theorem 9 then characterizes coverage by a comparison of bias and variance; the asymptotic coverage is good if variance dominates bias, and bad if the bias strictly dominates the variance.

Below we demonstrate in examples that irrespective of the dimension m of the variational approximation, variational credible sets will cover truths that are at least as smooth as the prior, and coverage may be bad if the prior oversmooths the truth. In this sense the variational posterior has the same behaviour as the true posterior (see for example [15, 12]).

5. Examples

We apply our theoretical results from the previous sections to two different, commonly used choices of the eigenvalues λ_j of the kernel in (2.7). First we investigate polynomially decaying eigenvalues. We note that Matérn covariance kernels (including the Ornstein-Uhlenbeck process) and Riemann-Liouville processes (including integrated Brownian motions) possess such covariance structure. As a second example we consider exponentially decaying eigenvalues. We note that the squared exponential Gaussian processes possess such exponentially decaying eigenvalues. In our examples we choose the eigenfunctions $(\varphi_j)_{j \in \mathbb{N}}$ to meet our (uniform) boundedness assumption in (2.9). This latter condition is not verified in general, but only for specific choices of covariance kernels (e.g. the Ornstein-Uhlenbeck process). Alternatively, it is possible to define a suitable GP prior using (2.7) by specifying $(\lambda_j)_{j \in \mathbb{N}}$ and taking any basis $(\varphi_j)_{j \in \mathbb{N}}$ that satisfies the boundedness assumption.

5.1. Polynomially decaying eigenvalues

In this subsection we consider eigenvalues of the form $\lambda_j \asymp j^{-1-2\alpha/d}$, $j = 1, \dots$, for some given $\alpha > 0$, and eigenfunctions $(\varphi_j)_{j \in \mathbb{N}}$ satisfying (2.9). First, by applying Theorems 5 and 7 to the present setting, we derive contraction rates for the corresponding population spectral features variational Bayes approximations. The proof of the corollary is deferred to Section B.1.

Corollary 10. *Consider the kernel k in (2.7) with $\lambda_j \asymp j^{-1-2\alpha/d}$ for some $\alpha > d/2$ and $(\varphi_j)_{j \in \mathbb{N}}$ satisfying (2.9). Then, for $m \asymp n^r$ with $1/2 > r \geq d/(d + 2\alpha)$, the population spectral features variational posterior contracts around bounded functions $f_0 \in \mathcal{S}^\beta$ at the rate $\epsilon_n = n^{-(\beta \wedge \alpha)/(d+2\alpha)}$, which is minimax optimal for $\alpha = \beta$. Furthermore, for $\alpha = \beta$ and $r < d/(d + 2\alpha)$ there exists a $f_0 \in \mathcal{S}^\beta$ such that the contraction rate $\epsilon_n \gtrsim n^{-p}$, for some $p < \beta/(d + 2\beta)$.*

We note that for $1/2 > r \geq d/(d + 2\alpha)$ the variational approximation inherits the asymptotic recovery property of the true posterior, i.e. the true posterior contracts around $f_0 \in \mathcal{S}^\beta$ with rate $\epsilon_n = n^{-(\beta \wedge \alpha)/(d+2\alpha)}$, which is minimax optimal for $\alpha = \beta$, see e.g. [31, 15]. The characterisation of the contraction rate in Theorem 5 shows that this rate is achieved uniformly over Sobolev balls $\{f \in L^2(\mathcal{X}, \mu) : \|f\|_\beta \leq C\}$. The upper bound $m \lesssim n^r$ with $r < 1/2$ is purely technical, to satisfy the assumption $m^2 n^{-1} \log n$ that has been used throughout this paper and which was justified in Remark 4. In [17] the same contraction rate results were derived using a different proof technique without this upper bound

on m . The lower bound $m \gtrsim n^{d/(d+2\alpha)}$, however, is sharp and of importance. Fewer inducing variables can result in sub-optimal posterior contraction around the truth, not matching the minimax rate.

Next we focus on the reliability of uncertainty quantification resulting in from the variational Bayes approximation. We apply our general Theorem 9 providing guarantees but also limitations for the frequentist coverage for the variational credible sets (4.1). Furthermore, in view of Theorem 8 we derive an upper bound for the size of the credible ball. The proof of the next corollary is given in Section B.2.

Corollary 11. *Let k be the kernel in (2.7) with $\lambda_j \asymp j^{-1-2\alpha/d}$, for some $\alpha > 0$ and eigenfunctions $(\varphi_j)_{j \in \mathbb{N}}$, satisfying (2.9). Consider $m \asymp n^r$ population spectral inducing variables for some $r \in (0, 1/2)$.*

1. *If $\alpha \leq \beta$, then*

$$P_0(\|\hat{f}_m - f_0\| \leq M_n \rho_n) \rightarrow 1$$

for any bounded $f_0 \in \mathcal{S}^\beta$ and any sequence $M_n \rightarrow \infty$.

2. *If $\alpha < \beta$ and $r < d/(d + 2\alpha)$, then*

$$P_0(\|\hat{f}_m - f_0\| \leq \rho_n) \rightarrow 1$$

for any bounded $f_0 \in \mathcal{S}^\beta$.

3. *If $\alpha > \beta > 0$, then there exists $f_0 \in \mathcal{S}^\beta$ such that*

$$P_0(\|\hat{f}_m - f_0\| \leq M \rho_n) \rightarrow 0$$

for all $M > 0$.

Finally, for $\beta = \alpha$ and $m \geq n^{d/(d+2\alpha)}$, there exists $C > 0$ such that for $f_0 \in \mathcal{S}_\beta$, we have $P_0(\rho_n \leq C n^{-\beta/(d+2\beta)}) \rightarrow 1$.

The first statement of the corollary says that by considering a prior which does not oversmooth the true function, the slightly inflated credible sets provide reliable uncertainty quantification from the frequentist perspective. In other words, by blowing up a credible ball (4.1) with a multiplicative factor of $M_n \rightarrow \infty$, its frequentist coverage tends to one. We note that inflating the credible set is in principle equivalent to (slightly) undersmoothing the prior distribution.

In statement 2, we in fact show that such technical post-processing is not necessary by taking $\alpha < \beta$ and considering less than $m \leq n^r$, with $r < d/(d + 2\alpha)$, inducing variables. We note that in view of Corollary 10, such choices of α and m result in an inflated contraction rate. In this case the variational posterior mean is sub-optimally far from the true f_0 (considering squared L^2 -loss). At the same time the size of the credible ball is also of higher magnitude $V_n \asymp \sum_{j>m} \lambda_j$ which is large enough to dominate the expected loss of the variational posterior mean.

When considering smoother priors than the true function f_0 , as in statement 3 of the corollary, the frequentist coverage of the variational credible set will tend to zero, independently from the number of inducing variables applied.

Finally, as a follow up of Corollary 10, the size of the credible set achieves the minimax rate if the regularity of the prior matches the regularity of f_0 . All these results, in fact, are in line with the frequentist coverage results derived for the true posterior distribution [15].

5.2. Exponentially decaying eigenvalues

We also study the series prior with (rescaled) exponentially decaying eigenvalues, motivated by the squared exponential kernel $k(x, y) = \exp(-b\|x - y\|^2)$. In this subsection we consider eigenvalues

$$\lambda_{n,j} \asymp \exp(-\tau_n j^{1/d}) \quad j = 1, \dots, \tag{5.1}$$

in (2.7) for some rescaling sequence $\tau_n \rightarrow 0$. For fixed τ_n , the decay of the eigenvalues is faster than in the previous example, giving the associated prior support consisting only of functions that are substantially smoother than the true function $f_0 \in \mathcal{S}^\beta$. To compensate for the rapidly diminishing eigenvalues a rescaling factor τ_n is introduced, which shrinks the trajectories of the prior draws, making them rougher. Let us consider the rescaling factor τ_n in the form

$$\tau_n := n^{-1/(d+2\alpha)} \log n, \tag{5.2}$$

for some constant $\alpha > 0$. Taking $\alpha = \beta$, in fact, results in minimax contraction rate and reliable uncertainty quantification for the true posterior distribution, see for instance [30, 2, 13]. The following corollary states that the contraction rate of the true posterior is inherited by its variational approximation for sufficiently high m . The proof is deferred to Section B.3.

Corollary 12. *Let k be the kernel in (2.7) with eigenvalues satisfying (5.1) with rescaling sequence (5.2) for some $\alpha > 0$ and eigenfunctions φ_j , $j = 1, \dots$ satisfying (2.9). Then, for $m = m_n \geq (\tau_n^{-1} \log n)^d = n^{d/(d+2\alpha)}$ such that $m^2 n^{-1} \log n \rightarrow 0$, the variational posterior contracts around bounded $f_0 \in \mathcal{H}^\beta$ at the rate $\epsilon_n = n^{-(\beta \wedge \alpha)/(d+2\alpha)}$, which is minimax optimal for $\alpha = \beta$. Furthermore, for $\alpha = \beta$ and $m = n^r$, with $r < d/(d + 2\alpha)$, there exists an $f_0 \in \mathcal{H}^\beta$, for which the posterior contracts at best at the sub-optimal rate n^{-p} , for some $p < \beta/(d + 2\beta)$.*

This corollary, on one hand shows that the variational posterior achieves minimax contraction rates for appropriately chosen rescaling factor and sufficiently many inducing variables $m \gtrsim n^{d/(d+2\alpha)}$. (The technical assumption $m^2 n^{-1} \log n = o(1)$ has been used throughout and was explained in Remark 4.) On the other hand it also implies that we need at least this many inducing variables, otherwise the variational posterior will provide sub-optimal recovery for the truth.

Next we discuss the validity of the variational Bayes uncertainty quantification. The result is slightly different in comparison with the polynomially decaying eigenvalues. The proof of the next corollary is given in Section B.4.

Corollary 13. *Let k be the kernel in (2.7) with eigenvalues satisfying (5.1) with rescaling sequence (5.2) for some $\alpha > 0$ and eigenfunctions φ_j , $j = 1, \dots$, satisfying (2.9). Furthermore, let the approximation have dimension $m = m_n \asymp n^r$ for some $r \in (0, 1/2)$. Then*

1. *if $r \geq d/(d + 2\alpha)$ and $\beta \geq \alpha > 0$, then for any bounded $f_0 \in \mathcal{S}^\beta$ and any $M_n \rightarrow \infty$,*

$$P_0(\|\hat{f}_m - f_0\| \leq M_n \rho_n) \rightarrow 1;$$

2. *if $r < d/(d + 2\alpha)$ then for any bounded $f_0 \in L^2(\mathcal{X}, \mu)$*

$$P_0(\|\hat{f}_m - f_0\| \leq \rho_n) \rightarrow 1;$$

3. *if $\alpha > \beta > 0$ and $m \geq n^{d/(d+2\alpha)}$, then there exists $f_0 \in \mathcal{S}^\beta$ such that for all $M > 0$*

$$P_0(\|\hat{f}_m - f_0\| \leq M \rho_n) \rightarrow 0.$$

Note that the only (real) difference between this result and Corollary 11 is that in the second statement no assumption is made about the (Sobolev) smoothness of f_0 . If the dimension of the approximation is of order $m_n = o(n^{d/(d+2\alpha)})$, then asymptotically the coverage of the variational credible set will always be good. The rest of the results are the same as in Corollary 11, except of the minor difference that in the third statement the lower bound on m holds exactly and not up to a multiplicative constant due to the exponentially decaying form of the eigenvalues, hence the conclusions are the same as well.

6. Numerical experiments

In this section we demonstrate how the developed theory can be applied in practice. Moreover, we show using synthetic data sets that the variational Bayes method proposed by [26] provides reliable uncertainty quantification (from a frequentist perspective) for appropriately tuned Gaussian process priors, independently from the number of inducing variables applied.

6.1. Synthetic data set

In a simulation study, we go beyond the population spectral features inducing variable method considered in our theoretical analysis above, and include other, practically more advantageous and popular approaches as well. We aspire to extrapolate our findings about the principles that govern the method with the investigated particular choice of inducing variables to other choices of inducing variables, for which the theoretical analysis is more complicated. With this goal in mind, in the current section, we aim not only to illustrate our findings but also compare various methods, to study empirically if and how our theoretical results may carry over. Between each of three choices of inducing variables, we shall point out what are common features and what is different. Besides the

population spectral features, we shall also study inducing points and what we call ‘empirical spectral features’.

In inducing points methods, the choice of inducing variables is $u_j = f(z_j)$ for a given set of inducing points $z_1, \dots, z_m \in \mathcal{X}$. In the literature various approaches were proposed to choose the inducing points. In our numerical analysis we consider two specific choices. First we consider the equidistant choice of the inducing points, i.e. we simply take the z_j on an equispaced grid on $[-\pi, \pi]$. This approach is designed to explore all parts of the signal equally. In case the design points x_1, \dots, x_n are not uniformly distributed the grid can be modified accordingly to mimic the underlying distribution of the covariates. We also consider the finite fixed-size determinantal point process (m -DPP), see [9].

The other choice of inducing variables is the sample analogue of the population spectral features $u_j = \langle f, \varphi_j \rangle$. Instead of diagonalizing the covariance operator T in (2.8), we decompose the covariance matrix of the prior at the design points $K_{\mathbf{f}\mathbf{f}} = [k(x_i, x_j)]_{1 \leq i, j \leq n}$. Since $K_{\mathbf{f}\mathbf{f}}$ is positive definite, there exists an orthonormal set of eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ such that $K_{\mathbf{f}\mathbf{f}} = \sum_{j=1}^n \mu_j \mathbf{v}_j \mathbf{v}_j'$, where $\mu_1 \geq \mu_2 \geq \dots \geq \mu_n \geq 0$ are the eigenvalues of $K_{\mathbf{f}\mathbf{f}}$. The *empirical spectral features* are defined $u_j := \mathbf{v}_j' \mathbf{f} = \sum_{i=1}^n v_{j,i} f(x_i)$. In [17] it was shown that the empirical spectral features induce a variational posterior with similar contraction rate results as the population spectral features under similar threshold for the number of inducing variables. Furthermore, the approximation accuracy of the above variational Bayes methods to the true posterior distribution with respect to the expected Kullback-Leibler divergence was studied in [6].

We consider the function space $L^2([-\pi, \pi], \mu)$, where μ is taken as the uniform distribution on $[-\pi, \pi]$. As our underlying true function (plotted in black in the figures) we take

$$f_0 = \sum_{\ell=1}^{\infty} \varphi_{3\ell} (3\ell)^{-1/2-\beta} / \log(3\ell) \in \mathcal{H}^\beta$$

with $\beta = 0.5$, where φ_j denote the standard Fourier basis. We use a synthetic data set with $n = 2500$ realisations $x_i \sim_{\text{i.i.d.}} \mu$, $i = 1, \dots, n$ and independent data points $y_i \sim_{\text{ind}} \mathcal{N}(f_0(x_i), \sigma^2)$ with $\sigma = 0.1$.

We consider both the polynomially and exponentially decaying eigenvalues λ_j from Section 5 and take the GP prior with covariance kernel $k(x, y) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(x) \varphi_j(y)$. Accordingly, in Figure 1 we plot the mean (dark gray) and 95% pointwise credible sets of the associated true posterior when $\lambda_j = j^{-1-2\beta}$ and $\lambda_j = \tau_n \exp(-\tau_n j/4)$, with $\tau_n = n^{-1/(1+2\beta)} \log n$, respectively.

First we investigate the variational Bayes approximations for the posterior corresponding to the polynomial GP priors in Figure 2. We plot the population spectral feature variational approximation in the first line (in blue), the empirical spectral feature approach in the second line (in red), the equidistant inducing points method in the third line (in green) and the m -DPP inducing points method in the fourth line (in yellow). The black curve stands for the underlying true function f_0 , the colored curve for the posterior mean, while the shaded area represents the 95% pointwise credible bands. We consider two choices for the number of inducing variables. For $m = 30$ (left hand side) we are

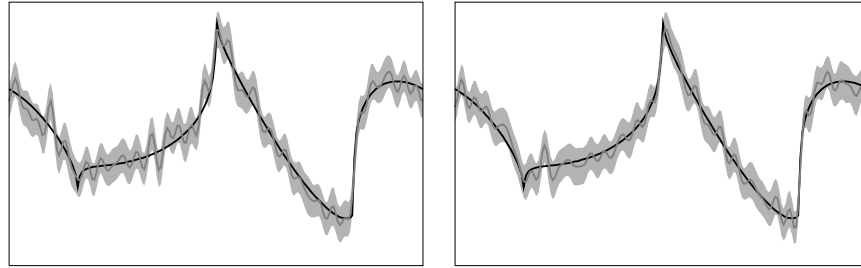


FIG 1. The true posterior distribution corresponding to the GP prior with polynomially (left) and exponentially (right) decaying eigenvalues.

below, while for $m = 60$ (right hand side) we are above the theoretical threshold $m = n^{1/(1+2\beta)} = 50$ obtained in our analysis. One can observe that for all methods the size of the credible bands are overly large for $m = 30$, while for $m = 60$ they closely resemble the true posterior given on the left hand side of Figure 1 (with perhaps the exception of the m -DPP method). Furthermore, for both choices of m the credible bands contain the true functional parameter f_0 , illustrating the good frequentist coverage properties of the variational methods, in line with our theoretical findings. One can also notice that the approximations from the population and empirical spectral features are fairly similar. For the equidistant inducing points method the variational posterior means do not seem to differ significantly of those in the preceding plots, but credible regions look fairly different. At the same time in the m -DPP method the posterior in certain neighbourhoods can be quite different from the true posterior. In case of both inducing points methods the intervals are narrow near those points on the horizontal axis that are close to the inducing points, and widen as the distance to an inducing point increases. Here, too, the credible regions seem to be over-conservative, i.e. their width being about equal to that of the credible set of the true posterior at inducing points and larger elsewhere.

Then in Figure 3 we plot the posterior means and credible sets resulting in from exponentially decaying eigenvalues $(\lambda_j)_{j \in \mathbb{N}}$. We use the same experimental setup as for polynomial eigenvalues, i.e. we take $m = 30$ and $m = 60$ inducing variables on the left and right hand side of the figure, respectively, and consider the above discussed four variational approximations. The plots are rather similar to what was shown for the polynomial eigenvalues in Figure 2, and hence the same conclusions hold for this prior.

6.2. Real world data

To illustrate how the procedure can be applied in practice, we consider a real data set consisting of hourly tin oxide measurements (the PT08.S1 series from [32] used to predict carbon monoxide). Since the series exhibits periodicity, a series prior with Fourier basis is suitable. With the population spectral features

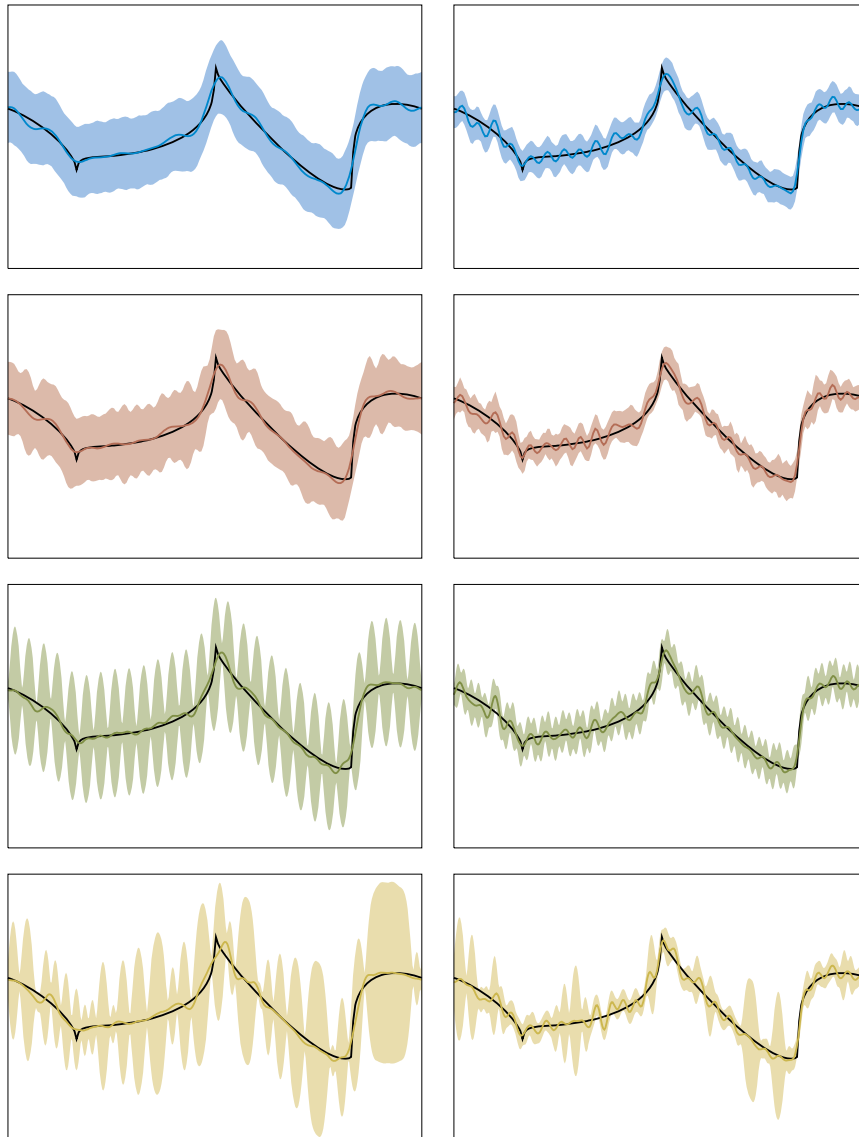


FIG 2. The variational Bayes approximations for GP with polynomially decaying eigenvalues. for $m = 30$ (left) and $m = 60$ (right) inducing variables. The choices of inducing variable methods from top to bottom: population spectral features (blue), empirical spectral features (red), inducing variables with equidistant design (green) and with m -DPP (yellow). In each figure the true function is plotted in black, the posterior mean is drawn with solid colored curves, while the shaded area illustrate the 95% point-wise credible bands.

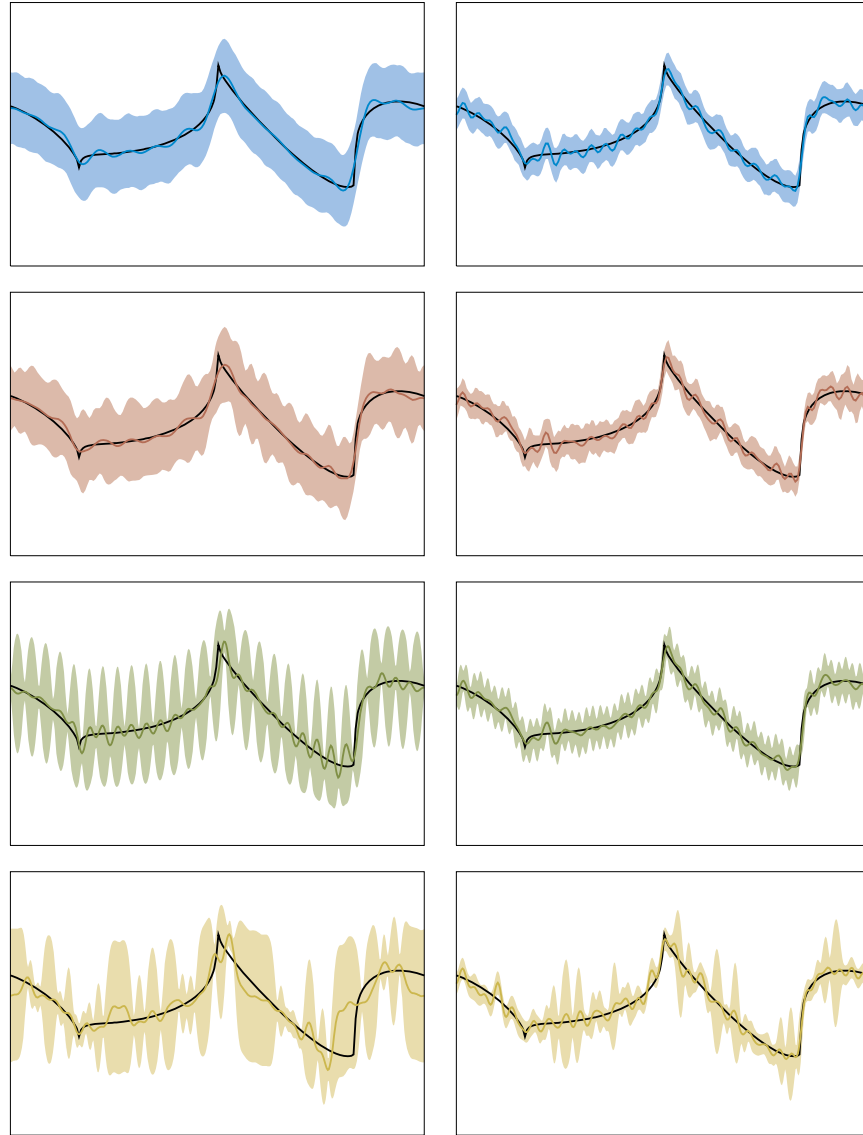


FIG 3. The variational Bayes approximations for GP with exponentially decaying eigenvalues. for $m = 30$ (left) and $m = 60$ (right) inducing variables. The choices of inducing variable methods from top to bottom: population spectral features (blue), empirical spectral features (red), inducing variables with equidistant design (green) and with m -DPP (yellow). In each figure the true function is plotted in black, the posterior mean is drawn with solid colored curves, while the shaded area illustrate the 95% point-wise credible bands.

variational procedure we solve the time series problem of estimating the trend and periodic components.

We start by preprocessing the data. The missing observations are estimated by interpolation of neighbouring days. This introduces bias, which however we do not account for, as the main focus here is on the practical application of the considered variational GP approach. We select the first $n = 9240$ observations, corresponding to 55 weeks of data.

In the preceding synthetic examples, we used the standard ordering of the Fourier basis, meaning the periods are sorted descendingly. The same approach here would result in overly small mass on the basis functions associated with the daily and weekly periodic behaviour. Hence we reorder the basis in a data-driven way, sorting the first n basis functions according to the size of $\varphi_j(\mathbf{x})'\mathbf{y}/n$, which estimates the coefficient $\langle \varphi_j, f_0 \rangle$.

In the Bayesian analysis we consider the series prior with covariance kernel

$$k(x, y) = \sum_{j=1}^{\infty} \tau j^{-1-2\alpha} \varphi_{(j)}(x) \varphi_{(j)}(y),$$

where the subscript between brackets (j) indicates the reordered indexes. We also introduced a rescaling factor $\tau > 0$ for the prior eigenvalues for additional flexibility. We fix the regularity parameter to be $\alpha = 1/2$ (the ‘roughest’ prior allowed in our theory) and the scaling parameter $\tau = 2 \cdot 10^5$ in our experiment. These quantities in practice are typically taken in a data driven, adaptive way. However, in this work we do not address adaptation and leave its theoretical understanding for future work.

In the synthetic examples, we considered the error variance σ^2 fixed. In the real data set it needs to be estimated. A canonical way to do this is by an empirical Bayes procedure, maximising the evidence

$$\hat{\sigma} = \arg \max_{\sigma} p_{\sigma}(\mathbf{x}, \mathbf{y}) = \arg \max_{\sigma} \int p_{f, \sigma}(\mathbf{x}, \mathbf{y}) d\Pi(f).$$

Alternatively, as proposed by [26], in the variational framework, variance estimation can be done by maximising the evidence lower bound (ELBO)

$$\begin{aligned} & \log \int \exp \left(\int \log p_{f, \sigma}(\mathbf{x}, \mathbf{y}) d\Pi(f|\mathbf{u}) \right) d\Pi_{\mathbf{u}}(\mathbf{u}) = \frac{1}{2} \log |\Sigma^{-1} K_{\mathbf{u}\mathbf{u}}| \\ & - \frac{n}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \mathbf{y}' [I - K_{\mathbf{f}\mathbf{u}} \Sigma^{-1} K_{\mathbf{u}\mathbf{f}} / \sigma^2] \mathbf{y} - \frac{1}{2\sigma^2} \text{tr}(K_{\mathbf{f}\mathbf{f}} - K_{\mathbf{f}\mathbf{u}} K_{\mathbf{u}\mathbf{u}}^{-1} K_{\mathbf{u}\mathbf{f}}), \end{aligned}$$

where $\Sigma = K_{\mathbf{u}\mathbf{u}} + \sigma^{-2} K_{\mathbf{u}\mathbf{f}} K_{\mathbf{f}\mathbf{u}}$ (for more details we refer to equation (28) in [26] and our Appendix C). This provides a significant reduction in computation time, since optimisation of the Bayes evidence requires repeated computation of the inverse of an $n \times n$ matrix, whereas the optimisation of the ELBO requires computation of the much smaller $m \times m$ inverse of Σ (note that the matrix $K_{\mathbf{u}\mathbf{u}}$ is $m \times m$ diagonal with entries $\tau j^{-1-2\alpha}$).

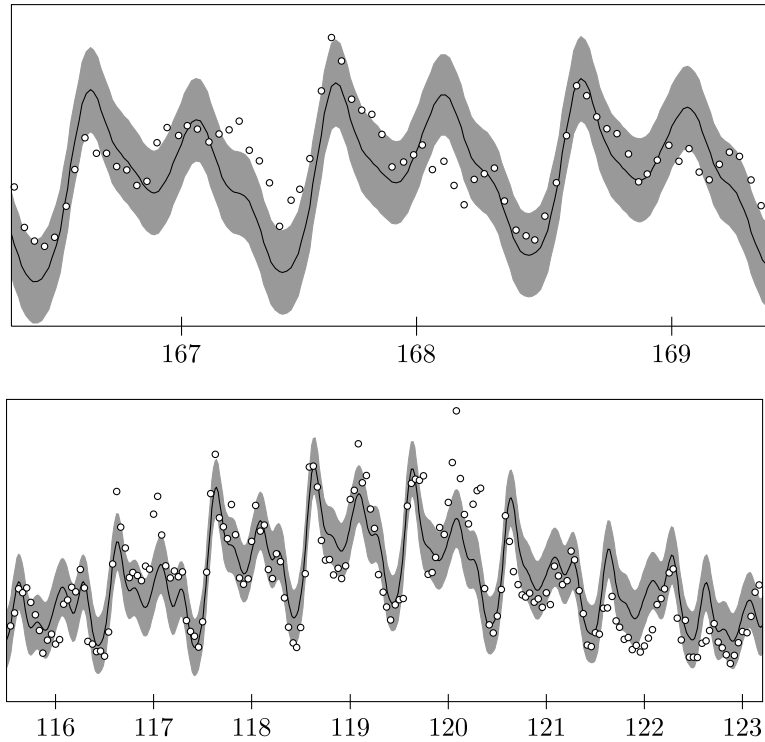


FIG 4. Variational posterior on two intervals of different lengths. On the horizontal axis is the elapsed time in days. We used $m = 100$, $\tau = 2 \cdot 10^5$ and using the ELBO estimated $\hat{\sigma} \approx 124$.

Figure 4 shows the data and variational posterior mean (black line) and 95% pointwise credible sets (gray) for our procedure with $m = 100 \approx n^{1/(1+2\alpha)}$, which is the recommendation that follows from the contraction rate results. We note that our theoretical results give frequentist coverage guarantees for the L^2 -credible sets (which are harder to visualize) under the assumption that $f_0 \in \mathcal{S}^\alpha$, for $\alpha = 1/2$. Nevertheless, together with the simulation study using the synthetic data set, it gives an indication of the reliability of the Bayesian uncertainty quantification from a frequentist perspective.

Appendix A: Proof of the general theorems and lemmas

A.1. Proof of Theorem 5

In view of Lemma 3 and the inequality

$$P_0\Psi(\|f - f_0\|^2 \geq (M_n\epsilon_n)^2 | \mathbf{x}, \mathbf{y}) \lesssim \frac{P_0\Psi[\|f - \hat{f}_m\|^2 | \mathbf{x}, \mathbf{y}] + P_0\|\hat{f}_m - f_0\|^2}{(M_n\epsilon_n)^2}$$

(by Markov’s inequality) it suffices to establish $\mathbb{P}_0\Psi[\|f - \hat{f}_m\|^2 | \mathbf{x}, \mathbf{y}] \lesssim \epsilon_n^2$. Using the formula for the posterior covariance (2.12), it follows that

$$\begin{aligned} \mathbb{P}_0\Psi[\|f - \hat{f}_m\|^2 | \mathbf{x}, \mathbf{y}] &= \mathbb{P}_0 \int \hat{k}_m(x, x) d\mu(x) \\ &= \mathbb{P}_0 \operatorname{tr}((\Lambda^{-1} + \sigma^{-2}\Phi'\Phi)^{-1}) + \sum_{j=m+1}^{\infty} \lambda_j. \end{aligned} \tag{A.1}$$

The proof is completed by showing

$$\mathbb{P}_0 \operatorname{tr}((\Lambda^{-1} + \sigma^{-2}\Phi'\Phi)^{-1}) \lesssim \frac{1}{n} \sum_{j=1}^m \nu_j. \tag{A.2}$$

To do so, the expectation term is distributed over the event

$$\Omega_m := \left\{ \sup_{1 \leq j, k \leq m} \left[\frac{1}{n} \sum_{i=1}^n \varphi_j(x_i)\varphi_k(x_i) - \delta_{jk} \right]^2 > C \frac{\log n}{n} \right\}$$

and its complement. Note that

$$\operatorname{tr}((\Lambda^{-1} + \sigma^{-2}\Phi'\Phi)^{-1}) \leq \operatorname{tr} \Lambda \leq \sum_{j=1}^{\infty} \lambda_j < \infty,$$

hence by the union bound and Hoeffding’s inequality,

$$\begin{aligned} \mathbb{P}_0 1_{\Omega_m} \operatorname{tr}((\Lambda^{-1} + \sigma^{-2}\Phi'\Phi)^{-1}) &\lesssim \mathbb{P}_0(\Omega_m) \leq \sum_{j, k=1}^m 2 \exp\left(\frac{-2n^2 C \log n}{4C_\varphi^2 n^2}\right) \\ &= 2m^2 n^{-C/2C_\varphi^2}, \end{aligned}$$

where C can be chosen arbitrarily large. This term is dominated by the following upper bound that we establish below:

$$\mathbb{P}_0 1_{\Omega_m^c} \operatorname{tr}((\sigma^2\Lambda^{-1} + \Phi'\Phi)^{-1}) \lesssim \operatorname{tr}((\sigma^2\Lambda^{-1} + nI)^{-1}) = \frac{1}{n} \sum_{j=1}^m \nu_j. \tag{A.3}$$

Using the matrix identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ it follows that

$$\begin{aligned} &|\operatorname{tr}((\sigma^2\Lambda^{-1} + \Phi'\Phi)^{-1}) - \operatorname{tr}((\sigma^2\Lambda^{-1} + nI)^{-1})| \\ &= \operatorname{tr}\left((\sigma^2\Lambda^{-1} + \Phi'\Phi)^{-1}(nI - \Phi'\Phi)(\sigma^2\Lambda^{-1} + nI)^{-1}\right) \\ &\leq \|(\sigma^2\Lambda^{-1} + \Phi'\Phi)^{-1}\| \|nI - \Phi'\Phi\| \operatorname{tr}((\sigma^2\Lambda^{-1} + nI)^{-1}), \end{aligned}$$

where $\|A\|$ denotes the operator norm of the matrix $A \in \mathbb{R}^{m \times m}$ with respect to the Euclidean norm on \mathbb{R}^m , which for symmetric A coincides with the largest absolute eigenvalue of A . We now show that on Ω_m^c , the product of the two norms in the preceding display is $o(1)$, implying (A.3).

By Lemma 14 below, there exists j between 1 and m such that the smallest eigenvalue of $\sigma^2\Lambda^{-1} + \Phi'\Phi$ (on Ω_m^c) is bounded from below by

$$\begin{aligned} & \sigma^2\lambda_j^{-1} + [\Phi'\Phi]_{j,j} - \sum_{\substack{1 \leq k \leq m \\ k \neq j}} |[\Phi'\Phi]_{j,k}| \\ &= \sigma^2\lambda_j^{-1} + \left[\sum_{i=1}^n \varphi_j(x_i)^2 - \sum_{\substack{1 \leq k \leq m \\ k \neq j}} \left| \sum_{i=1}^n \varphi_k(x_i)\varphi_j(x_i) \right| \right] \\ &\geq n - m\sqrt{Cn \log n} \geq cn, \end{aligned}$$

where the latter inequality being true for some (deterministic) $c > 0$ provided n is large enough, as follows from the assumption $m^2n^{-1} \log n \rightarrow 0$. Consequently, the largest eigenvalue of the inverse $(\sigma^2\Lambda^{-1} + \Phi'\Phi)^{-1}$ satisfies

$$\|(\sigma^2\Lambda^{-1} + \Phi'\Phi)^{-1}\| \leq \frac{1}{cn}.$$

Another application of Lemma 14 similarly shows that

$$\|nI - \Phi'\Phi\| \leq m\sqrt{Cn \log n}.$$

Since $m\sqrt{n^{-1} \log n} \rightarrow 0$, the product of the above two norms vanishes (deterministically) on Ω_m^c , implying (A.3) and concluding the proof.

Lemma 14 (Gershgorin circle theorem; see [10]). *Let $A \in \mathbb{R}^{n \times n}$ with entries A_{ij} . For any eigenvalue λ of A , there exists $j \in \{1, \dots, n\}$ such that*

$$|\lambda - A_{jj}| \leq \sum_{i \neq j} |A_{ij}|.$$

A.2. Proof of Theorem 7

Take

$$f_0 = \sum_{j=1}^{\infty} j^{-(1+p/r)/2} \varphi_j, \quad \text{for some } \frac{2r\beta}{d} < p < \frac{2\beta}{d+2\beta}$$

and note that $f_0 \in \mathcal{S}^\beta$ and

$$\sum_{j>m} \langle f_0, \varphi_j \rangle^2 \asymp \int_m^\infty x^{-1-p/r} dx \asymp m^{-p/r} \asymp n^{-p} \gg n^{-2\beta/(d+2\beta)}. \tag{A.4}$$

The left-hand side in this display is a lower bound for $\|\hat{f}_m - f_0\|^2$ since $\langle \hat{f}_m, \varphi_j \rangle = 0$ for $j > m$, so (3.11) follows.

For the contraction rate statement we fix $p < p' < \frac{2\beta}{d+2\beta}$ and make a case distinction, first assuming

$$\sum_{j>m} \lambda_j \lesssim n^{-p'}. \tag{A.5}$$

Note that given $M > 0$, by (A.4) there exists $C_0 > 0$ such that for n large enough

$$Mn^{-p'} \leq (1 - C_0) \sum_{j>m} \langle f_0, \varphi_j \rangle^2.$$

Combining this with

$$\|f - f_0\|^2 = \sum_{j=1}^{\infty} \langle f - f_0, \varphi_j \rangle^2 \geq \sum_{j>m} \langle f_0, \varphi_j \rangle^2 - 2 \sum_{j>m} \langle f_0, \varphi_j \rangle \langle f, \varphi_j \rangle$$

yields

$$\begin{aligned} & \Psi(\|f - f_0\|^2 \leq Mn^{-p'} \mid \mathbf{x}, \mathbf{y}) \\ & \leq \Psi\left(\sum_{j>m} \langle f_0, \varphi_j \rangle^2 - Mn^{-p'} \leq 2 \sum_{j>m} \langle f_0, \varphi_j \rangle \langle f, \varphi_j \rangle \mid \mathbf{x}, \mathbf{y}\right) \\ & \leq \Psi\left(\sum_{j>m} \langle f_0, \varphi_j \rangle^2 \leq 2C_0 \sum_{j>m} \langle f_0, \varphi_j \rangle \langle f, \varphi_j \rangle \mid \mathbf{x}, \mathbf{y}\right) \\ & = \frac{1}{2} \Psi\left(\sum_{j>m} \langle f_0, \varphi_j \rangle^2 \leq 2C_0 \left| \sum_{j>m} \langle f_0, \varphi_j \rangle \langle f, \varphi_j \rangle \right| \mid \mathbf{x}, \mathbf{y}\right), \end{aligned}$$

where in the last line we used that under the variational posterior the inner products $\langle f, \varphi_j \rangle, j > m$ have a mean-zero Gaussian distribution. Continuing the preceding display, and using the Markov, triangle and Cauchy-Schwarz inequalities, it follows that

$$\begin{aligned} & \Psi(\|f - f_0\|^2 \leq Mn^{-p'} \mid \mathbf{x}, \mathbf{y}) \\ & \lesssim \frac{\sum_{j>m} \langle f_0, \varphi_j \rangle \Psi(|\langle f, \varphi_j \rangle| \mid \mathbf{x}, \mathbf{y})}{\sum_{j>m} \langle f_0, \varphi_j \rangle^2} \\ & \lesssim n^p \sum_{j>m} \langle f_0, \varphi_j \rangle \sqrt{\lambda_j} \\ & \leq n^p \sqrt{\sum_{j>m} \langle f_0, \varphi_j \rangle^2 \sum_{j>m} \lambda_j} \\ & \lesssim n^{(p-p')/2} \rightarrow 0, \end{aligned}$$

where in the last line we used (A.4) and (A.5).

If the assumption (A.5) does not hold, then along a subsequence,

$$n^{p'} \sum_{j>m} \lambda_j \rightarrow \infty. \tag{A.6}$$

In this case, we note that for $j > m$ the inner products $\langle f, \varphi_j \rangle \stackrel{d}{=} \sqrt{\lambda_j} Z_j$ with $Z_j \sim_{\text{i.i.d.}} N(0, 1)$, hence by Markov's inequality,

$$\Psi(\|f\|^2 \leq Mn^{-p'} \mid \mathbf{x}, \mathbf{y})$$

$$\begin{aligned}
&\leq \Psi\left(\sum_{j>m} \langle f, \varphi_j \rangle^2 \leq Mn^{-p'} \mid \mathbf{x}, \mathbf{y}\right) \\
&= \Psi\left(\exp(-n^{p'} \sum_{j>m} \langle f, \varphi_j \rangle^2) \geq \exp(-M) \mid \mathbf{x}, \mathbf{y}\right) \\
&\lesssim \prod_{j>m} \mathbb{E} \exp\left(-n^{p'} \lambda_j Z_j^2\right) = \prod_{j>m} (1 + 2n^{p'} \lambda_j)^{-1/2} \\
&\leq \left(1 + \sum_{j>m} 2n^{p'} \lambda_j\right)^{-1/2},
\end{aligned}$$

which vanishes along a subsequence by the assumption (A.6). We conclude that in this case the assertion of the theorem holds for $f_0 = 0$.

A.3. Proof of Theorem 8

In view of Markov's inequality and the definition of ρ_n

$$\rho_n^2 = \rho_n^2 \gamma^{-1} \Psi(\|f - \hat{f}_m\| > \rho_n \mid \mathbf{x}, \mathbf{y}) \leq \gamma^{-1} \Psi(\|f - \hat{f}_m\|^2 \mid \mathbf{x}, \mathbf{y}), \quad (\text{A.7})$$

Then the upper bound for ρ_n is implied by the preceding display together with the inequalities (A.1) and (A.2) in the proof of Theorem 5.

Next we establish the lower bound. First note, that in view of Fubini's theorem and the expansion (2.12) of the variational posterior covariance,

$$\begin{aligned}
&\Psi(\langle f - \hat{f}_m, \varphi_i \rangle \langle f - \hat{f}_m, \varphi_j \rangle \mid \mathbf{x}, \mathbf{y}) \\
&= \Psi\left(\int (f - \hat{f}_m) \varphi_i d\mu \int (f - \hat{f}_m) \varphi_j d\mu \mid \mathbf{x}, \mathbf{y}\right) \\
&= \int \int \hat{k}_m(u, v) \varphi_i(u) \varphi_j(v) d\mu(u) d\mu(v) \\
&= \begin{cases} [(\Lambda^{-1} + \sigma^{-2} \Phi' \Phi)^{-1}]_{i,j} & \text{for } i \vee j \leq m, \\ \lambda_j \delta_{ij} & \text{for } i \vee j > m. \end{cases}
\end{aligned}$$

Therefore, under the variational posterior $\Psi(\cdot \mid \mathbf{x}, \mathbf{y})$ the inner products $\langle f - \hat{f}_m, \varphi_j \rangle$ are centered Gaussian random variables, where the vector of the first m variables has covariance matrix $(\Lambda^{-1} + \sigma^{-2} \Phi' \Phi)^{-1}$ and is independent of the remaining $j > m$, which are mutually independent with variance λ_j . Furthermore, note that if Z has an m -dimensional normal distribution with mean zero and covariance matrix A with eigenvalues a_1, \dots, a_m , then

$$\begin{aligned}
\mathbb{E} e^{-v \|Z\|^2} &= \mathbb{E} e^{-v \sum_{j=1}^m Z_j^2} \\
&= \det(I + 2vA)^{-1/2} \\
&= \prod_{j=1}^m (1 + 2va_j)^{-1/2}
\end{aligned}$$

$$\begin{aligned} &\leq \left(1 + 2v \sum_{j=1}^m a_j\right)^{-1/2} \\ &= \left(1 + 2v \operatorname{tr}(A)\right)^{-1/2}. \end{aligned}$$

Therefore, by using Markov’s inequality,

$$\begin{aligned} 1 - \gamma &= \Psi(\|f - \hat{f}_m\| \leq \rho_n \mid \mathbf{x}, \mathbf{y}) \\ &= \Psi(\exp(-\rho_n^{-2}\|f - \hat{f}_m\|^2) \geq \exp(-1) \mid \mathbf{x}, \mathbf{y}) \\ &\leq e\Psi\left(\exp\left(-\rho_n^{-2}\sum_{j=1}^{\infty}\langle f - \hat{f}_m, \varphi_j \rangle^2\right) \mid \mathbf{x}, \mathbf{y}\right) \\ &= e\Psi\left(\exp\left(-\rho_n^{-2}\sum_{j=1}^m\langle f - \hat{f}_m, \varphi_j \rangle^2\right) \mid \mathbf{x}, \mathbf{y}\right) \\ &\quad \cdot \prod_{j>m} \Psi\left(\exp(-\rho_n^{-2}\langle f, \varphi_j \rangle^2) \mid \mathbf{x}, \mathbf{y}\right) \\ &= e \det\left(I + 2\rho_n^{-2}(\Lambda^{-1} + \sigma^{-2}\Phi'\Phi)^{-1}\right)^{-1/2} \cdot \prod_{j>m} (1 + 2\rho_n^{-2}\lambda_j)^{-1/2} \\ &\leq e\left(1 + 2\rho_n^{-2}\operatorname{tr}((\Lambda^{-1} + \sigma^{-2}\Phi'\Phi)^{-1}) + 2\rho_n^{-2}\sum_{j>m}\lambda_j\right)^{-1/2}, \end{aligned}$$

which in turn implies that

$$\rho_n^2 \geq \frac{2}{e^2(1 - \gamma)^{-2} - 1} \left(\operatorname{tr}((\Lambda^{-1} + \sigma^{-2}\Phi'\Phi)^{-1}) + \sum_{j>m}\lambda_j\right).$$

The lower bound in the statement of the theorem now follows by the inequality (12) from [33], which reads

$$\operatorname{tr}((\Lambda^{-1} + \sigma^{-2}\Phi'\Phi)^{-1}) \geq \sum_{j=1}^m \frac{\sigma^2\lambda_j}{\sigma^2 + \lambda_j \sum_{i=1}^n \varphi_j(x_i)^2} \geq (C_\varphi^{-2} \wedge 1)\sigma^2 \frac{1}{n} \sum_{j=1}^m \nu_j.$$

A.4. Proof of Theorem 9

Consider first the statements 1 and 2. Note that, with K as in Theorem 8, by Markov’s inequality,

$$\begin{aligned} &\mathbb{P}_0(\|\hat{f}_m - f_0\| > M_n\rho_n) \\ &\leq \mathbb{P}_0(\rho_n^2 < K^{-1}V_n) + \mathbb{P}_0(\|\hat{f}_m - f_0\|^2 \geq M_n^2K^{-1}V_n) \\ &\lesssim \mathbb{P}_0(\rho_n^2 < K^{-1}V_n) + M_n^{-2} \frac{\mathbb{P}_0\|\hat{f}_m - f_0\|^2}{V_n}. \end{aligned}$$

This can be seen to vanish by applying Theorem 8 to the term on the left in the upper bound, and by combining Lemma 3 with the assumptions on R_n for the term on the right. Statements 1 and 2 follow, the former by taking $M_n = 1$.

Since $\nu_n \leq 1$, the assumption $R_n \rightarrow \infty$ in statement 3 is equivalent to $B_n/V_n \rightarrow \infty$. Then letting $F : L^2(\mathcal{X}, \mu) \rightarrow L^2(\mathcal{X}, \mu)$ be the operator $f \mapsto \sum_{j=1}^m \nu_j \langle f, \varphi_j \rangle \varphi_j$, and id the identity operator, the assumption can be further re-written as

$$V_n / \|(\text{id} - F)f_0\|^2 \rightarrow 0. \tag{A.8}$$

Note that by the triangle inequality,

$$\begin{aligned} & \mathbb{P}_0(\|\hat{f}_m - f_0\| \leq \rho_n) \\ & \leq \mathbb{P}_0(\|(\text{id} - F)f_0\| \leq \|\hat{f}_m - Ff_0\| + M\rho_n) \\ & \leq \mathbb{P}_0(\|(\text{id} - F)f_0\| \leq 2\|\hat{f}_m - Ff_0\|) + \mathbb{P}_0(\|(\text{id} - F)f_0\| \leq 2M\rho_n). \end{aligned}$$

The second probability on the right hand side is seen to vanish as $n \rightarrow \infty$ by combining (A.8) with Theorem 8. Regarding the first probability, (A.10) in the proof of Lemma 3 below, gives

$$\mathbb{P}_0\|\hat{f}_m - Ff_0\|^2 \lesssim n^{-1} \sum_{j=1}^m \nu_j^2$$

so by Markov’s inequality and (A.8) it follows that

$$\begin{aligned} & \mathbb{P}_0(\|(\text{id} - F)f_0\| \leq 2\|\hat{f}_m - Ff_0\|) \\ & \lesssim \frac{\mathbb{P}_0\|\hat{f}_m - Ff_0\|^2}{\|(\text{id} - F)f_0\|^2} \\ & \lesssim \frac{n^{-1} \sum_{j=1}^m \nu_j^2}{V_n} \frac{V_n}{\|(\text{id} - F)f_0\|^2} \rightarrow 0. \end{aligned}$$

A.5. Proof of Lemma 1

Any $f \in \mathbb{H}_m$ is of the form $f(x) = \sum_{j=1}^m a_j h_j(x) = K_{xu} \mathbf{a}$ for some $\mathbf{a} = (a_1, \dots, a_m) \in \mathbb{R}^m$. Moreover, such an f has squared RKHS norm

$$\|f\|_{\mathbb{H}}^2 = \sum_j \sum_k a_j a_k \langle h_j, h_k \rangle_{\mathbb{H}} = \sum_j \sum_k a_j a_k \Pi_{u_j} u_k = \mathbf{a}' K_{uu} \mathbf{a}.$$

It follows that we need to minimise the objective function

$$\mathbf{a} \mapsto (\mathbf{y} - K_{fu} \mathbf{a})' (\mathbf{y} - K_{fu} \mathbf{a}) + \sigma^2 \mathbf{a}' K_{uu} \mathbf{a}.$$

Setting the derivative equal to zero yields

$$-K_{uf} \mathbf{y} + (\sigma^2 K_{uu} + K_{uf} K_{fu}) \mathbf{a} = 0. \tag{A.9}$$

Since the second derivative $2(\sigma^2 K_{uu} + K_{uf} K_{fu})$ is positive definite, the solution $\mathbf{a}^* = (\sigma^2 K_{uu} + K_{uf} K_{fu})^{-1} K_{uf} \mathbf{y}$ minimises the objective function. It follows that

$$\arg \min_{f \in \mathbb{H}_m} \sum_{i=1}^n (y_i - f(x_i))^2 + \sigma^2 \|f\|_{\mathbb{H}}^2 = K_{\cdot u} \mathbf{a}^* = \hat{f}_m.$$

A.6. Proof of Lemma 3

The proof follows similar lines of reasoning as [3] and [12] in context of GP and distributed GP regression, using kernel ridge regression techniques. Here these standard techniques are adapted to the variational approximation.

Letting id denote the identity operator and F denote the operator

$$F = \sum_{j=1}^m \nu_j \langle \cdot, \varphi_j \rangle \varphi_j$$

on $L^2(\mathcal{X}, \mu)$, we obtain an identity for the sum of the bias terms

$$\sum_{j=1}^m (1 - \nu_j)^2 \langle f_0, \varphi_j \rangle^2 + \sum_{j>m} \langle f_0, \varphi_j \rangle^2 = \|(\text{id} - F)f_0\|^2.$$

Define $\Delta f_0 = \hat{f}_m - Ff_0$. Since $\|\hat{f}_m - f_0\|^2 \leq 2\|\Delta f_0\|^2 + 2\|(\text{id} - F)f_0\|^2$, it suffices to show that

$$\mathbb{P}_0 \|\Delta f_0\|^2 \lesssim \frac{1}{n} \sum_{j=1}^m \nu_j^2. \quad (\text{A.10})$$

This is done by characterising the variational posterior mean \hat{f}_m as the root of a ‘‘score’’ function. Let us write $\hat{f}_m = K_{xu} \mathbf{a}^*$ as in the proof of Lemma 1. Combining (A.9) with the identities (2.10), it follows that $\hat{f}_m = \varphi_{1:m}(\cdot)' \Lambda \mathbf{a}^*$ solves the equation

$$\begin{aligned} \mathbf{0} &= \Lambda \Phi' \mathbf{y} - (\sigma^2 + \Lambda \Phi' \Phi) \Lambda \mathbf{a}^* \\ &= \Lambda \Phi' \left(\mathbf{y} - \begin{bmatrix} \hat{f}_m(x_1) \\ \vdots \\ \hat{f}_m(x_n) \end{bmatrix} \right) - \sigma^2 \begin{bmatrix} \langle \hat{f}_m, \varphi_1 \rangle \\ \vdots \\ \langle \hat{f}_m, \varphi_m \rangle \end{bmatrix} \end{aligned} \quad (\text{A.11})$$

where we recall that $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$ and $\Phi_{i,j} = \varphi_j(x_i)$. Multiplying the j -th entry of the vector in (A.11) with φ_j gives that \hat{f}_m is the root of the ‘‘score’’ operator \hat{S}_n

$$\hat{S}_n(g) = \sum_{j=1}^m \left[\frac{1}{n} \sum_{i=1}^n (y_i - g(x_i)) \varphi_j(x_i) \right] \lambda_j \varphi_j - \frac{\sigma^2}{n} g, \quad g \in \mathbb{H}_m,$$

that is, $\hat{S}_n(\hat{f}_m) = 0$. Similarly, Ff_0 is the root of the ‘‘population score’’

$$S_n(g) = \mathbb{P}_0 \hat{S}_n(g) = \sum_{j=1}^m \langle f_0 - g, \varphi_j \rangle \lambda_j \varphi_j - \frac{\sigma^2}{n} g, \quad g \in \mathbb{H}_m.$$

Let $f_{0,m} = \sum_{j=1}^m \langle f_0, \varphi_j \rangle \varphi_j$ denote the orthogonal projection of f_0 onto \mathbb{H}_m . Then

$$S_n(\hat{f}_m) = T f_{0,m} - \sum_{j=1}^m \frac{n \lambda_j + \sigma^2}{n} \langle \hat{f}_m, \varphi_j \rangle \varphi_j = T(f_{0,m} - F^{-1} \hat{f}_m),$$

so $-\Delta f_0 = Ff_0 - \hat{f}_m = FT^{-1}S_n(\hat{f}_m)$ and

$$\begin{aligned} \|\Delta f_0\|^2 &= \|FT^{-1}S_n(\hat{f}_m)\|^2 \\ &\leq 2\|FT^{-1}(\hat{S}_n(Ff_0) + S_n(\hat{f}_m))\|^2 + 2\|FT^{-1}\hat{S}_n(Ff_0)\|^2. \end{aligned}$$

We establish at the end of this proof that

$$P_0\|FT^{-1}(S_n(\hat{f}_m) + \hat{S}_n(Ff_0))\|^2 \lesssim m^4n^{-D} + o(P_0\|\Delta f_0\|^2), \quad (\text{A.12})$$

so the preceding two displays together yield

$$P_0\|\Delta f_0\|^2 \lesssim P_0\|FT^{-1}\hat{S}_n(Ff_0)\|^2 + m^4n^{-D} + o(P_0\|\Delta f_0\|^2). \quad (\text{A.13})$$

Now we use that $y_i = f_0(x_i) + \varepsilon_i$ under P_0 , so

$$\begin{aligned} \hat{S}_n(Ff_0) &= \hat{S}_n(Ff_0) - S_n(Ff_0) \\ &= \sum_{j=1}^m \left[\frac{1}{n} \sum_{i=1}^n (y_i - Ff_0(x_i))\varphi_j(x_i) - \langle (\text{id} - F)f_0, \varphi_j \rangle \right] \lambda_j \varphi_j \\ &= \sum_{j=1}^m \left[\frac{1}{n} \sum_{i=1}^n (f_0 - Ff_0)(x_i)\varphi_j(x_i) - \langle (\text{id} - F)f_0, \varphi_j \rangle \right] \\ &\quad + \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_j(x_i) \lambda_j \varphi_j \end{aligned}$$

so since the ε_i are independent of the x_i and both are i.i.d.,

$$\begin{aligned} P_0\|FT^{-1}\hat{S}_n(Ff_0)\|^2 &= \sum_{j=1}^m \nu_j^2 P_0 \left[\frac{1}{n} \sum_{i=1}^n [(\text{id} - F)f_0](x_i)\varphi_j(x_i) - \langle (\text{id} - F)f_0, \varphi_j \rangle + \frac{1}{n} \sum_{i=1}^n \varepsilon_i \varphi_j(x_i) \right]^2 \\ &= \sum_{j=1}^m \nu_j^2 \frac{1}{n} \text{var}_0 \left([(\text{id} - F)f_0](x_1)\varphi_j(x_1) \right) + \frac{\sigma^2}{n} \sum_{j=1}^m \nu_j^2 \lesssim \frac{1}{n} \sum_{j=1}^m \nu_j^2, \end{aligned}$$

the bottom line following from the boundedness of f_0 . This together with (A.13) for D large enough yields (A.10), which concludes the proof.

Proof of (A.12). Since $S_n(Ff_0) = \hat{S}_n(\hat{f}_m) = 0$,

$$\begin{aligned} \hat{S}_n(Ff_0) + S_n(\hat{f}_m) &= \hat{S}_n(Ff_0) - \hat{S}_n(\hat{f}_m) + S_n(\hat{f}_m) - S_n(Ff_0) \\ &= \sum_{j=1}^m \left[\frac{1}{n} \sum_{i=1}^n (\Delta f_0)(x_i)\varphi_j(x_i) - \langle \Delta f_0, \varphi_j \rangle \right] \lambda_j \varphi_j \end{aligned}$$

and using $\Delta f_0 = \sum_{k=1}^m \langle \Delta f_0, \varphi_k \rangle \varphi_k$ and the Cauchy-Schwarz inequality, it follows that

$$P_0\|FT^{-1}(\hat{S}_n(Ff_0) + S_n(\hat{f}_m))\|^2$$

$$\begin{aligned}
 &= \mathbb{P}_0 \left\| \sum_{j=1}^m \left[\sum_{k=1}^m \langle \Delta f_0, \varphi_k \rangle \left[\frac{1}{n} \sum_{i=1}^n \varphi_k(x_i) \varphi_j(x_i) - \delta_{jk} \right] \right] \nu_j \varphi_j \right\| \\
 &= \sum_{j=1}^m \nu_j^2 \mathbb{P}_0 \left[\sum_{k=1}^m \langle \Delta f_0, \varphi_k \rangle \left[\frac{1}{n} \sum_{i=1}^n \varphi_k(x_i) \varphi_j(x_i) - \delta_{jk} \right] \right]^2 \\
 &\leq \mathbb{P}_0 \|\Delta f_0\|^2 \sum_{j=1}^m \nu_j^2 \sum_{k=1}^m \left[\frac{1}{n} \sum_{i=1}^n \varphi_k(x_i) \varphi_j(x_i) - \delta_{jk} \right]^2. \tag{A.14}
 \end{aligned}$$

Splitting over the events

$$\Omega_{j,m} := \left\{ \sup_{1 \leq k \leq m} \left[\frac{1}{n} \sum_{i=1}^n \varphi_j(x_i) \varphi_k(x_i) - \delta_{jk} \right]^2 > C \frac{\log n}{n} \right\} \tag{A.15}$$

and their complements, and then using Assumption 2, the term in (A.14) is at most of the order

$$\frac{m \log n}{n} \sum_{j=1}^m \nu_j^2 \mathbb{P}_0 \|\Delta f_0\|^2 + C_\varphi^2 m^2 \max_{1 \leq j \leq m} \mathbb{P}_0 1_{\Omega_{j,m}} \|\Delta f_0\|^2.$$

Now (A.12) follows from

$$\frac{m \log n}{n} \sum_{j=1}^m \nu_j^2 \leq \frac{m^2 \log n}{n} \rightarrow 0$$

(by assumption) and, as we prove now

$$\max_{1 \leq j \leq m} \mathbb{P}_0 1_{\Omega_{j,m}} \|\Delta f_0\|^2 \lesssim m^2 n^{-D}. \tag{A.16}$$

To this end, recall that from (2.11), $\hat{f}_m(x) = \varphi_{1:m}(x)' (\sigma^2 \Lambda^{-1} + \Phi' \Phi)^{-1} \Phi' \mathbf{y}$, so

$$\begin{aligned}
 \|\hat{f}_m\|^2 &= \sum_{k=1}^m \langle \hat{f}_m, \varphi_k \rangle^2 = \text{tr} \left[(\sigma^2 \Lambda^{-1} + \Phi' \Phi)^{-1} \Phi' \mathbf{y} \mathbf{y}' \Phi (\sigma^2 \Lambda^{-1} + \Phi' \Phi)^{-1} \right] \\
 &\leq \lambda_1^2 \sigma^{-4} \mathbf{y}' \Phi \Phi' \mathbf{y} \lesssim \sum_{j=1}^m \left[\sum_{i=1}^n y_i \varphi_j(x_i) \right]^2 \lesssim mn \sum_{i=1}^n y_i^2,
 \end{aligned}$$

and hence, since $y_i = f_0(x_i) + \varepsilon_i$ for some bounded f_0 and ε_i independent of x_i ,

$$\mathbb{P}_0 1_{\Omega_{j,m}} \|\hat{f}_m\|^2 \lesssim mn \mathbb{P}_0 1_{\Omega_{j,m}} \left[\sum_{i=1}^n f_0(x_i)^2 + n\sigma^2 \right] \lesssim mn^2 \mathbb{P}_0(\Omega_{j,m}).$$

This together with $\|F f_0\| \leq \|f_0\| < \infty$ yields

$$\mathbb{P}_0 1_{\Omega_{j,m}} \|\Delta f_0\|^2 \leq 2\mathbb{P}_0(\Omega_{j,m}) \|F f_0\|^2 + 2\mathbb{P}_0 1_{\Omega_{j,m}} \|\hat{f}_m\|^2 \lesssim mn^2 \mathbb{P}_0(\Omega_{j,m}).$$

The inequality (A.16) follows from the above by combining a union bound and Hoeffding's inequality:

$$P_0(\Omega_{j,m}) \leq m \cdot 2 \exp\left(\frac{-2n^2 C \log n}{4C_\varphi^2 n^2}\right) = 2mn^{-D-2},$$

with $D = C/2C_\varphi^2 - 2$, where the constant C can be chosen arbitrarily large. \square

Appendix B: Proof of the corollaries

B.1. Proof of Corollary 10

Take $d/(d + 2\alpha) \leq r < 1/2$. Note that $m^2 n^{-1} \log n = o(1)$, hence the conditions of Theorem 5 hold. Next note that

$$J_n = \max\{j : n\lambda_j \geq 1\} \asymp n^{d/(d+2\alpha)} \lesssim m, \tag{B.1}$$

hence in view of (3.9),

$$\begin{aligned} V_n &\asymp J_n/n + \int_{m \wedge J_n}^\infty x^{-1-2\alpha/d} dx \asymp J_n/n + J_n^{-2\alpha/d} \asymp n^{-2\alpha/(d+2\alpha)} \\ W_n &\lesssim J_n/n + \int_{J_n}^\infty x^{-2-4\alpha/d} dx \asymp J_n/n + nJ_n^{-1-4\alpha/d} \asymp n^{-2\alpha/(d+2\alpha)}. \end{aligned} \tag{B.2}$$

To bound the bias term B_n , we consider (3.8) and deal with the two terms on the right hand side separately. For the first term note

$$\begin{aligned} \sum_{j=1}^{m \wedge J_n} (n\lambda_j)^{-2} \langle f_0, \varphi_j \rangle^2 &\lesssim \sum_{j=1}^{J_n} (nj^{-1-2\alpha/d})^{-2} \langle f_0, \varphi_j \rangle^2 \\ &= n^{-2} \sum_{j=1}^{J_n} j^{2+4\alpha/d-2\beta/d} j^{2\beta/d} \langle f_0, \varphi_j \rangle^2 \\ &\lesssim n^{-2} \max_{1 \leq j < J_n} j^{2+4\alpha/d-2\beta/d} \|f_0\|_\beta^2 \\ &\lesssim (n^{-2} \vee n^{-2\beta/(d+2\alpha)}) \|f_0\|_\beta^2, \end{aligned}$$

while for the second term we get

$$\sum_{j > m \wedge J_n}^\infty \langle f_0, \varphi_j \rangle^2 < (J_n \wedge m)^{-2\beta/d} \sum_{j > m \wedge J_n}^\infty j^{2\beta/d} \langle f_0, \varphi_j \rangle^2 \lesssim n^{-2\beta/(d+2\alpha)} \|f_0\|_\beta^2, \tag{B.3}$$

hence the contraction rate in (3.4) is $\epsilon_n = n^{-(\beta \wedge \alpha)/(d+2\alpha)}$.

The sub-optimal contraction rate for insufficient amount of inducing variables is a direct consequence of Theorem 7.

B.2. Proof of Corollary 11

Note that $m^2 n^{-1} \log n = n^{2r-1} \log n \rightarrow 0$, hence we can apply Theorem 9 to the present setting.

By (3.9), similarly to (B.2), we get that

$$V_n \asymp \frac{m \wedge n^{d/(d+2\alpha)}}{n} + m^{-2\alpha/d} \quad \text{and} \quad W_n \lesssim \frac{m \wedge n^{d/(d+2\alpha)}}{n}. \quad (\text{B.4})$$

Then similarly as in the proof of Corollary 10 we get for $f_0 \in \mathcal{S}^\beta$, that

$$\begin{aligned} \sum_{j=1}^m (1 - \nu_j)^2 \langle f_0, \varphi_j \rangle^2 &\lesssim (n^{-2} \vee n^{-2\beta/(d+2\alpha)}) \\ \sum_{j=m+1}^{\infty} \langle f_0, \varphi_j \rangle^2 &\lesssim m^{-2\beta/d} \|f_0\|_\beta^2. \end{aligned}$$

Therefore,

$$R_n = \frac{B_n + W_n}{V_n} \lesssim \frac{n^{-1}(m \wedge n^{d/(d+2\alpha)}) + n^{-2\beta/(d+2\alpha)} + m^{-2\beta/d}}{n^{-1}(m \wedge n^{d/(d+2\alpha)}) + m^{-2\alpha/d}}. \quad (\text{B.5})$$

In the first case, (B.5) is further bounded from above by a multiple of $m^{-2\beta/d}/m^{-2\alpha/d} = O(1)$. In the second case, (B.5) is $o(1)$. Finally, in the third case, let us take $f_0 \in \mathcal{S}^\beta$ given by $f_0 = \sum_{j=1}^{\infty} j^{-1/2-q} \varphi_j$, for some $\beta/d < q < \alpha/d$. Then in view of (3.8),

$$\sum_{j=1}^{m \wedge J_n} (1 - \nu_j)^2 \langle f_0, \varphi_j \rangle^2 + \sum_{j > m \wedge J_n} \langle f_0, \varphi_j \rangle^2 \gtrsim \sum_{j > m \wedge J_n} \langle f_0, \varphi_j \rangle^2 \asymp (m \wedge J_n)^{-2q}$$

hence by recalling (B.4),

$$R_n \gtrsim \frac{m^{-2q} \vee n^{-2qd/(d+2\alpha)}}{m^{-2\alpha/d} + n^{-2\alpha/(d+2\alpha)}} \rightarrow \infty.$$

Therefore, in all three cases the statement follows from Theorem 9.

Finally, the upper bound for the radius follows from Theorem 8 and (B.2) with $\alpha = \beta$.

B.3. Proof of Corollary 12

The proof goes similarly to Corollary 10. First assume that m is lower bounded by

$$J_n := (\tau_n^{-1} \log n)^d = n^{d/(d+2\alpha)} \quad (\text{B.6})$$

and note that $\lambda_{n, J_n} \asymp \exp(-\tau_n J_n^{1/d}) = n^{-1}$. Hence, similarly to (3.8)

$$B_n \asymp \sum_{j=1}^{m \wedge J_n} (n \lambda_{n, j})^{-2} \langle f_0, \varphi_j \rangle^2 + \sum_{j > m \wedge J_n} \langle f_0, \varphi_j \rangle^2.$$

We deal with the two terms on the right hand side of the preceding display separately.

For any $f_0 \in \mathcal{S}^\beta$, we have

$$\sum_{j>m \wedge J_n} \langle f_0, \varphi_j \rangle^2 \leq J_n^{-2\beta/d} \sum_{j>m \wedge J_n} j^{2\beta/d} \langle f_0, \varphi_j \rangle^2 \lesssim n^{-2\beta/(d+2\alpha)} \|f_0\|_\beta^2 \tag{B.7}$$

and

$$\begin{aligned} \sum_{j=1}^{m \wedge J_n} (n\lambda_{n,j})^{-2} \langle f_0, \varphi_j \rangle^2 &\leq \max_{1 \leq i \leq J_n} (n\lambda_{n,i})^{-2} i^{-2\beta/d} \sum_{j=1}^\infty j^{2\beta/d} \langle f_0, \varphi_j \rangle^2 \\ &\lesssim n^{-2} \left(\exp(2\tau_n) \vee \frac{\exp(2\tau_n J_n^{1/d})}{J_n^{2\beta/d}} \right) \|f_0\|_\beta^2 \\ &\lesssim (n^{-2} \vee n^{-2\beta/(d+2\alpha)}) \|f_0\|_\beta^2 \end{aligned} \tag{B.8}$$

where we used that the function $i \mapsto \exp(2\tau_n i) i^{-2\beta/d}$ is convex in i , so the maximum occurs at one of the endpoints.

Next we deal with the variance terms W_n and V_n , defined in (3.3) and (3.5), respectively. Similarly to (3.9) and (B.2), for J_n given in (B.6),

$$\begin{aligned} V_n &\asymp \frac{m \wedge J_n}{n} + \sum_{j>m \wedge J_n} \lambda_{n,j} \asymp \frac{(m \wedge J_n)}{n} + \int_{m \wedge J_n} e^{-\tau_n x^{1/d}} dx \\ &\lesssim n^{-\frac{2\alpha}{d+2\alpha}} + \int_{(m \wedge J_n)^{1/d}} x^{d-1} e^{-\tau_n x} dx. \end{aligned} \tag{B.9}$$

By partial integration and induction we get that the rightmost integral satisfies

$$\int_{(m \wedge J_n)^{1/d}} x^{d-1} e^{-\tau_n x} dx \asymp (\tau_n^{-1} (m \wedge J_n)^{1-1/d} \vee \tau_n^{-d}) \exp(-\tau_n (m \wedge J_n)^{1/d}). \tag{B.10}$$

Since $m \geq J_n = (\tau_n^{-1} \log n)^d$ it is further bounded by

$$n^{d/(d+2\alpha)} (\log n)^{-1} \exp(-\log n) = n^{-2\alpha/(d+2\alpha)} / \log n.$$

Furthermore, by similar computations

$$\begin{aligned} W_n &\asymp \frac{m \wedge J_n}{n} + \sum_{j>m \wedge J_n} \lambda_{n,j} \\ &\lesssim n^{-2\alpha/(d+2\alpha)} + n \int_{(m \wedge J_n)^{1/d}} x^{d-1} e^{-2\tau_n x} dx \\ &\lesssim n^{-2\alpha/(d+2\alpha)} + n(\tau_n^{-1} (m \wedge J_n)^{1-1/d} \vee \tau_n^{-d}) \exp(-2\tau_n (m \wedge J_n)^{1/d}) \\ &\lesssim n^{-2\alpha/(d+2\alpha)}. \end{aligned} \tag{B.11}$$

Our contraction rate result follows from Theorem 5 with $\epsilon_n^2 = n^{-2(\beta \wedge \alpha)/(d+2\alpha)}$.

The sub-optimal contraction rate for insufficient amount of inducing variables is a direct consequence of Theorem 7.

B.4. Proof of Corollary 13

Note that by assumption we have $m^2 n^{-1} \log n \rightarrow 0$, hence we can apply Theorem 9. Therefore it is sufficient to investigate the asymptotic behaviour of the fraction $R_n = (B_n + W_n)/V_n$, with the terms defined in (3.3) and (3.5). We also recall the definition of J_n given in (B.6) for the exponentially decaying eigenvalues.

In the second case, following from the bounds (B.7), (B.8) and (B.11), the numerator of R_n is $o(1)$. At the same time, in view of assertions (B.9), (B.10) and $\tau_n m^{1/d} = n^{r/d-1/(d+2\alpha)} \log n \rightarrow 0$, the denominator is bounded from below as

$$V_n \gtrsim \sum_{j>m} \lambda_{n,j} \gtrsim \tau_n^{-d} \exp(-\tau_n m^{1/d}) \gtrsim \tau_n^{-d} \rightarrow \infty,$$

and therefore $R_n = o(1)$.

Next, in the first case, we have $m \wedge J_n \asymp n^r \wedge J_n \asymp J_n$, hence again in view of (B.7), (B.8), $B_n \lesssim n^{-2} \vee n^{-2\beta/(d+2\alpha)}$. Moreover, following from (B.9) and (B.11), we get that $V_n \gtrsim J_n/n = n^{-2\alpha/(d+2\alpha)}$ and $W_n \lesssim n^{-2\alpha/(d+2\alpha)}$, respectively. Combining this upper bounds results in that $R_n = O(1)$.

Finally, in the third case, as in the proof of Corollary 11, let f_0 be the function $\sum_{j=1}^\infty j^{-1/2-q} \varphi_j \in \mathcal{S}^\beta$ for some $\beta/d < q < \alpha/d$. In view of (3.8)

$$B_n \gtrsim \sum_{j>m \wedge J_n} \langle f_0, \varphi_j \rangle^2 \asymp (m \wedge J_n)^{-2q} \gtrsim n^{-2qd/(d+2\alpha)}.$$

Furthermore, (B.9) and (B.10) together with $m \geq J_n$ imply that,

$$V_n \lesssim J_n/n + \tau_n^{-1} J_n^{1-1/d} \exp(-\tau_n J_n^{1/d}) \lesssim n^{-2\alpha/(d+2\alpha)}.$$

Therefore, $R_n \gtrsim n^{-2qd/(d+2\alpha)} / n^{-2\alpha/(d+2\alpha)} \rightarrow \infty$. We conclude that, in all three cases the statements now follow directly from Theorem 9.

Finally, the upper bound for the radius follows from Theorem 8 and the upper bound for V_n in the proof of Corollary B.3 with $\alpha = \beta$.

Appendix C: Variational posterior for general inducing variables

Recall that the posterior is approximated by the variational distribution

$$\Psi = \int \Pi(\cdot | \mathbf{u}) d\Psi_{\mathbf{u}}(\mathbf{u}) \tag{C.1}$$

on $L^2(\mathcal{X}, \mu)$. The above display is equivalent to

$$\frac{d\Psi}{d\Pi}(f) = \frac{d\Psi_{\mathbf{u}}}{d\Pi_{\mathbf{u}}}(\mathbf{u}(f)), \tag{C.2}$$

where $\Pi_{\mathbf{u}}$ is the distribution of the vector of inducing variables \mathbf{u} under the prior Π (this follows from existence of the Radon-Nikodym derivative on the right,

which in turn is guaranteed by the assumption that $\Psi_{\mathbf{u}}$ is a non-degenerate Gaussian). The identity (C.2) shows that the variational posterior Ψ is parameterised by f through \mathbf{u} only.

Then, in view of Bayes' theorem

$$\frac{d\Pi(\cdot|\mathbf{x},\mathbf{y})}{d\Pi}(f) = \frac{p_f(\mathbf{x},\mathbf{y})}{p(\mathbf{x},\mathbf{y})}$$

and some additional elementary algebraic manipulations

$$\begin{aligned} D_{\text{KL}}(\Psi\|\Pi(\cdot|\mathbf{x},\mathbf{y})) &= \int \log \frac{d\Psi}{d\Pi(\cdot|\mathbf{x},\mathbf{y})} d\Psi \\ &= \int \log \frac{d\Psi}{d\Pi} d\Psi + \int \log \frac{d\Pi}{d\Pi(\cdot|\mathbf{x},\mathbf{y})} d\Psi \\ &= \int \log \frac{d\Psi_{\mathbf{u}}}{d\Pi_{\mathbf{u}}} d\Psi_{\mathbf{u}} - \iint \log p_f(\mathbf{x},\mathbf{y}) d\Psi(f|\mathbf{u}) d\Psi_{\mathbf{u}}(\mathbf{u}) + \log p(\mathbf{x},\mathbf{y}). \end{aligned}$$

In the variational procedure, we aim to minimise this with respect to Ψ . Using that the KL-divergence is non-negative, and $d\Psi(f|\mathbf{u}) = d\Pi(f|\mathbf{u})$ by construction, we obtain

$$\log p(\mathbf{x},\mathbf{y}) \geq \iint \log p_f(\mathbf{x},\mathbf{y}) d\Pi(f|\mathbf{u}) d\Psi_{\mathbf{u}}(\mathbf{u}) - \int \log \frac{d\Psi_{\mathbf{u}}}{d\Pi_{\mathbf{u}}} d\Psi_{\mathbf{u}}.$$

The right-hand side is commonly referred to as evidence lower bound (ELBO). Minimising the KL-divergence is equivalent to maximising the ELBO. Note that by Jensen's inequality,

$$\text{ELBO} = \int \log \left[\frac{d\Pi_{\mathbf{u}}}{d\Psi_{\mathbf{u}}}(\mathbf{u}) \exp \left(\int \log p_f(\mathbf{x},\mathbf{y}) d\Pi(f|\mathbf{u}) \right) \right] d\Psi_{\mathbf{u}}(\mathbf{u}) \quad (\text{C.3})$$

$$\leq \log \int \exp \left(\int \log p_f(\mathbf{x},\mathbf{y}) d\Pi(f|\mathbf{u}) \right) d\Pi_{\mathbf{u}}(\mathbf{u}), \quad (\text{C.4})$$

and the maximum is attained by the distribution $\Psi_{\mathbf{u}}^*$ defined through

$$\frac{d\Psi_{\mathbf{u}}^*}{d\Pi_{\mathbf{u}}}(\mathbf{v}) = \frac{\exp \left(\int \log p_f(\mathbf{x},\mathbf{y}) d\Pi(f|\mathbf{u}=\mathbf{v}) \right)}{\int \exp \left(\int \log p_f(\mathbf{x},\mathbf{y}) d\Pi(f|\mathbf{u}) \right) d\Pi_{\mathbf{u}}(\mathbf{u})}. \quad (\text{C.5})$$

In general there is no guarantee that the maximizer $\Psi_{\mathbf{u}}^*$ of the ELBO is tractable, but in case of the considered Gaussian process regression model and Gaussian variational class it has an explicit form. In view of (C.5),

$$\begin{aligned} \frac{d\Psi_{\mathbf{u}}^*}{d\Pi_{\mathbf{u}}}(\mathbf{u}) &\propto \exp \left(-\frac{1}{2\sigma^2} \int \sum_{i=1}^n (y_i - f(x_i))^2 d\Pi(f|\mathbf{u}) \right) \\ &= \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n [(y_i - \Pi(f(x_i)|\mathbf{u}))^2 + \text{var}_{\Pi(\cdot|\mathbf{u})}(f(x_i))] \right) \end{aligned}$$

$$\begin{aligned} &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i - \Pi(f(x_i)|\mathbf{u}))^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i - K_{x_i\mathbf{u}}K_{\mathbf{u}\mathbf{u}}^{-1}\mathbf{u})^2\right), \end{aligned} \tag{C.6}$$

where $K_{x\mathbf{u}} = \text{cov}_{\Pi}(f(x), \mathbf{u}) = \Pi f(x)\mathbf{u}^T$ and $K_{\mathbf{u}\mathbf{u}} = \text{cov}_{\Pi}(\mathbf{u}, \mathbf{u}) = \Pi\mathbf{u}\mathbf{u}^T$, and we used that

$$\text{var}_{\Pi(\cdot|\mathbf{u})}(f(x)) = k(x, x) - K_{x\mathbf{u}}K_{\mathbf{u}\mathbf{u}}^{-1}K_{\mathbf{u}x},$$

which is constant as a function of \mathbf{u} .

Completing the square in (C.6), it follows that the optimal variational distribution $\Psi_{\mathbf{u}}^*$ of \mathbf{u} is Gaussian with mean $K_{\mathbf{u}\mathbf{u}}(\sigma^2K_{\mathbf{u}\mathbf{u}} + K_{\mathbf{u}f}K_{f\mathbf{u}})^{-1}K_{\mathbf{u}f}\mathbf{y}$ and covariance matrix $K_{\mathbf{u}\mathbf{u}}(K_{\mathbf{u}\mathbf{u}} + \sigma^{-2}K_{\mathbf{u}f}K_{f\mathbf{u}})^{-1}K_{\mathbf{u}\mathbf{u}}$. By (C.1), the variational distribution Ψ^* of f is Gaussian, with mean function

$$\begin{aligned} \hat{f}_m(x) &= \int \Pi(f(x)|\mathbf{u}) d\Psi_{\mathbf{u}}^*(\mathbf{u}) \\ &= K_{x\mathbf{u}}K_{\mathbf{u}\mathbf{u}}^{-1} \int \mathbf{u} d\Psi_{\mathbf{u}}^*(\mathbf{u}) \\ &= K_{x\mathbf{u}}(\sigma^2K_{\mathbf{u}\mathbf{u}} + K_{\mathbf{u}f}K_{f\mathbf{u}})^{-1}K_{\mathbf{u}f}\mathbf{y}. \end{aligned}$$

Furthermore, the covariance function is

$$\begin{aligned} (x, y) &\mapsto \int \Pi((f - \hat{f}_m)(x)(f - \hat{f}_m)(y)|\mathbf{u}) d\Psi_{\mathbf{u}}^*(\mathbf{u}) \\ &= \int \text{cov}_{\Pi(\cdot|\mathbf{u})}(f(x), f(y)) d\Psi_{\mathbf{u}}^*(\mathbf{u}) + \text{cov}_{\Psi_{\mathbf{u}}^*}(\Pi(f(x)|\mathbf{u}), \Pi(f(y)|\mathbf{u})) \\ &= k(x, y) - K_{x\mathbf{u}}K_{\mathbf{u}\mathbf{u}}^{-1}K_{\mathbf{u}y} + K_{x\mathbf{u}}K_{\mathbf{u}\mathbf{u}}^{-1} \text{cov}_{\Psi_{\mathbf{u}}^*}(\mathbf{u}, \mathbf{u})K_{\mathbf{u}\mathbf{u}}^{-1}K_{\mathbf{u}y} \\ &= k(x, y) - K_{x\mathbf{u}}K_{\mathbf{u}\mathbf{u}}^{-1}K_{\mathbf{u}y} + K_{x\mathbf{u}}(K_{\mathbf{u}\mathbf{u}} + \sigma^{-2}K_{\mathbf{u}f}K_{f\mathbf{u}})^{-1}K_{\mathbf{u}y}. \end{aligned}$$

Funding

Co-funded by the European Union (ERC, BigBayesUQ, project number: 101041064). Views and opinions expressed are however those of the author(s) 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.

References

- [1] ALQUIER, P. and RIDGWAY, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics* **48** 1475–1497. [MR4124331](#)

- [2] BHATTACHARYA, A. and PATI, D. (2015). Adaptive Bayesian inference in the Gaussian sequence model using exponential-variance priors. *Statistics & Probability Letters* **103** 100–104. [MR3350868](#)
- [3] BHATTACHARYA, A., PATI, D. and YANG, Y. (2017). Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*.
- [4] BISHOP, C. M. and NASRABADI, N. M. (2006). *Pattern Recognition and Machine Learning*. Springer. [MR2247587](#)
- [5] BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American statistical Association* **112** 859–877. [MR3671776](#)
- [6] BURT, D. R., RASMUSSEN, C. E. and VAN DER WILK, M. (2020). Convergence of sparse variational inference in Gaussian processes regression. *Journal of Machine Learning Research* **21** 1–63. [MR4138115](#)
- [7] CASTILLO, I. and NICKL, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics* **42** 1941–1969. [MR3262473](#)
- [8] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *The Annals of Statistics* **21** 903–923. [MR1232525](#)
- [9] GAUTIER, G., POLITO, G., BARDENET, R. and VALKO, M. (2019). DPPy: DPP sampling with Python. *Journal of Machine Learning Research* **20** 1–7.
- [10] GERSHGORIN, S. A. (1931). Über die Abgrenzung der Eigenwerte einer Matrix. *Bulletin de l'Académie des Sciences de l'URSS* **6** 749–754.
- [11] GIORDANO, R. J., BRODERICK, T. and JORDAN, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. *Advances in Neural Information Processing Systems* **28**.
- [12] HADJI, A., HESSELINK, T. and SZABO, B. (2022). Optimal recovery and uncertainty quantification for distributed Gaussian process regression. *arXiv preprint arXiv:2205.03150*.
- [13] HADJI, A. and SZABÓ, B. (2021). Can we trust Bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel? *SIAM/ASA Journal on Uncertainty Quantification* **9** 185–230. [MR4212884](#)
- [14] KIMELDORF, G. S. and WAHBA, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* **41** 495–502. [MR0254999](#)
- [15] KNAPIK, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *The Annals of Statistics* **39** 2626–2657. [MR2906881](#)
- [16] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag. [MR0856411](#)
- [17] NIEMAN, D., SZABO, B. and VAN ZANTEN, H. (2022). Contraction rates for sparse variational approximations in Gaussian process regression. *Journal of Machine Learning Research* **23** 1–26. [MR4577158](#)
- [18] RASMUSSEN, C. E. and WILLIAMS, C. K. (2006). *Gaussian Processes for Machine Learning*. MIT Press. [MR2514435](#)

- [19] RAY, K. and SZABÓ, B. (2022). Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association* **117** 1270–1281. [MR4480711](#)
- [20] RAY, K., SZABO, B. and CLARA, G. (2020). Spike and slab variational Bayes for high dimensional logistic regression. *Advances in Neural Information Processing Systems* **33** 14423–14434.
- [21] ROUSSEAU, J. and SZABO, B. (2020). Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors. *The Annals of Statistics* **48** 2155–2179. [MR4134790](#)
- [22] SCHÖLKOPF, B., SMOLA, A. J., BACH, F. et al. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.
- [23] SRINIVAS, N., KRAUSE, A., KAKADE, S. and SEEGER, M. (2010). Gaussian process optimization in the bandit setting: no regret and experimental design. In: *International Conference on Machine Learning* 1015–1022. [MR2952544](#)
- [24] STEINWART, I. and SCOVEL, C. (2012). Mercer’s theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation* **35** 363–417. [MR2914365](#)
- [25] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics* **43** 1391–1428. [MR3357861](#)
- [26] TITSIAS, M. (2009). Variational model selection for sparse Gaussian process regression. *Report, University of Manchester, UK*.
- [27] VAKILI, S., SCARLETT, J., SHIU, D.-S. and BERNACCHIA, A. (2022). Improved convergence rates for sparse approximation methods in kernel-based learning. In: *Proceedings of the 39th International Conference on Machine Learning* **162** 21960–21983. PMLR.
- [28] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh* 200–222. [MR2459226](#)
- [29] VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press. [MR1652247](#)
- [30] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electronic Journal of Statistics* **1** 433–448. [MR2357712](#)
- [31] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics* **36** 1435–1463. [MR2418663](#)
- [32] VITO, S. (2016). Air Quality. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C59K5F>.
- [33] VIVARELLI, F. and OPPER, M. (1999). General bounds on Bayes errors for regression with Gaussian processes. *Advances in Neural Information Processing Systems* **11** 302–308. [MR1655599](#)
- [34] WANG, Y. and BLEI, D. M. (2019). Frequentist consistency of variational

- Bayes. *Journal of the American Statistical Association* **114** 1147–1161. [MR4011769](#)
- [35] WILD, V., KANAGAWA, M. and SEJDINOVIC, D. (2021). Connections and equivalences between the Nyström method and sparse variational Gaussian processes. *arXiv preprint [arXiv:2106.01121](#)*.
- [36] YANG, Y., PATI, D. and BHATTACHARYA, A. (2020). α -variational inference with statistical guarantees. *The Annals of Statistics* **48** 886–905. [MR4102680](#)
- [37] ZHANG, F. and GAO, C. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics* **48** 2180–2207. [MR4134791](#)