Journal of Machine Learning Research 23 (2022) 1-26 Submitted 9/21; Revised 3/22; Published 6/22

Contraction rates for sparse variational approximations in Gaussian process regression

Dennis Nieman D.NIEMAN@VU.NL
Department of Mathematics
Vrije Universiteit Amsterdam
De Boelelaan 1111, 1081 HV Amsterdam
The Netherlands

Botond Szabo BOTOND.SZABO@UNIBOCCONI.IT
Department of Decision Sciences,
Bocconi Institute for Data Science and Analytics,
Bocconi University
Via Roentgen 1, Milano, Italy

Harry van Zanten J.H.VAN.ZANTEN@VU.NL
Department of Mathematics
Vrije Universiteit Amsterdam
De Boelelaan 1111, 1081 HV Amsterdam
The Netherlands

Editor: Marc Peter Deisenroth

Abstract

We study the theoretical properties of a variational Bayes method in the Gaussian Process regression model. We consider the inducing variables method introduced by Titsias (2009b) and derive sufficient conditions for obtaining contraction rates for the corresponding variational Bayes (VB) posterior. As examples we show that for three particular covariance kernels (Matérn, squared exponential, random series prior) the VB approach can achieve optimal, minimax contraction rates for a sufficiently large number of appropriately chosen inducing variables. The theoretical findings are demonstrated by numerical experiments.

Keywords: Variational Bayes, Gaussian Process regression, inducing variables, contraction rates

1. Introduction

Suppose we observe n independent pairs $(x_1, y_1), \dots, (x_n, y_n)$, where each x_i has distribution G on a subset $\mathcal{X} \subseteq \mathbb{R}^d$ and

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

with an unknown function $f : \mathcal{X} \to \mathbb{R}$ and $\varepsilon_1, \dots, \varepsilon_n$ independent Gaussian variables with mean zero and variance σ^2. In Gaussian Process (GP) regression we model f a-priori as a centered GP with covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. GP regression has become popular due to the explicit expressions for the posterior (also a GP, see e.g. Rasmussen and

©2022 Dennis Nieman, Botond Szabo and Harry van Zanten.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v23/21-1128.html.

Williams, 2006 and Section 2 ahead) and the marginal likelihood, and the ease with which uncertainty quantification can be obtained. Moreover, there exist mathematical guarantees for consistency, optimal contraction rates, and validity of uncertainty quantification (e.g. van der Vaart and van Zanten, 2008b; Sniekers and van der Vaart, 2015; Rousseau and Szabo, 2017).

A drawback of plain GP regression is the fact that computation of the posterior requires inversion of an $n \times n$ matrix, which becomes computationally demanding for large sample size n. The computational cost typically scales as n^3, which can be prohibitive in practice. To alleviate the computational burden, reduced rank approximations are often employed; see for instance Chapter 8 of Rasmussen and Williams (2006) and the more recent overview in Liu et al. (2020). These approximations somehow summarise the posterior using $m \ll n$ variables instead of n, typically reducing the order of the computational cost from n^3 to nm^2.

In this paper we consider the variational approximation proposed by Titsias (2009b). This approach uses m so-called inducing variables to summarise the posterior (details are given in the next section). It is a true variational Bayes procedure, in the sense that the approximate posterior minimises the Kullback-Leibler (KL) divergence between the true posterior and a parametrised family of approximating distributions.

While the computational aspects of low rank approximations are well understood, little is known about whether the mathematical guarantees for the true posterior carry over to the approximate posterior. Burt et al. (2019) analyse the expected KL-divergence between the posterior and its variational approximation. In particular, they investigate in various cases how large the number of inducing variables m should be chosen in relation to the sample size n in order to ensure that the expected KL-divergence vanishes as n becomes large. However, since they compute the expectation both over the data $(\boldsymbol{x}, \boldsymbol{y})$ and over the prior on f, these results do not translate to (frequentist) guarantees about consistency and contraction rates, which assume that the data is generated from a fixed, “true” regression function f_0.

In this paper we derive contraction rates for the approximate posterior in this frequentist setup. This makes it possible to compare rates with known minimax lower bounds, which explain what the best possible contraction rates are and how these depend on global characteristics of the true regression function f_0, like its degree of smoothness. This in turn gives insight into how the dimension m of the variational approximation should be chosen in order for the variational posterior to have the same contraction rate as the true posterior.

Our findings can be summarised as follows:

(i) In order to have an optimal rate of contraction of the variational posterior around the true regression function f_0, it is not necessary that the KL-divergence between the true posterior and the variational approximation vanishes as $n \to \infty$.

(ii) For appropriately chosen inducing variables, one can recover an α-smooth regression function f_0 at the optimal rate with the VB method using the Matérn kernel or a series kernel with regularity hyper-parameter α if the number of inducing variables m scales at least as $n^{d/(d+2\alpha)}$.

(iii) These inducing variable VB methods also result in minimax contraction rate around α-smooth regression functions f_0 for GP priors with squared exponential covariance kernel (in $d = 1$) if the number of inducing variables m scales at least as $n^{1/(1+2\alpha)} \log n$.

(iv) Choosing fewer inducing points than the optimal number can result in overly smooth posterior means and conservative, sub-optimally large credible sets; see the numerical study in Section 7.

The remainder of the paper is organised as follows. In Section 2 we recall the inducing variable variational Bayes method by Titsias (2009b). Next in Section 3 we briefly discuss contraction rate results for GP posteriors following van der Vaart and van Zanten (2008b). A more detailed description of the frequentist analysis of general (nonparametric) posteriors are given in Appendix A. The main results are presented in Section 4 where sufficient conditions are given on the GP prior and the inducing variables to obtain the contraction rate of the corresponding VB posterior. In Sections 5.1 and 5.2 two specific choices of the inducing variables are described, from the eigendecompositions of respectively the covariance matrix and the covariance operator. We show in Section 6 that these approaches result in rate optimal VB posterior contraction rates for the squared exponential, Matérn and series covariance kernels, matching the optimal behaviour of the appropriately scaled true posterior. Finally we conclude our results with a brief numerical study in Section 7.

1.1 Notation

For two positive sequences a_n, b_n we use the notation $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \leq C b_n$ for all n. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$ are satisfied simultaneously. We denote by tr the trace operator and by $D_{\mathrm{KL}}(\mu, \nu)$ the Kullback-Leibler divergence between the measures μ and ν. The norm $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral/operator norm for matrices. By $L^2(\mathcal{X}, G)$ we denote the space of (almost sure equivalence classes of) Borel measurable real-valued functions f on $\mathcal{X}$ such that $\|f\|_{2,G}^2 := \int_{\mathcal{X}} |f|^2 \, dG$ is finite.

2. Inducing variables variational Bayes

In this section we recall the sparse GP regression approach of Titsias (2009b), introducing the notation that we use throughout the paper.

In the regression model (1), if a centered GP prior Π with covariance kernel k is used, then the true posterior is again a GP, with mean and covariance function given by

$$x \mapsto K_{x\boldsymbol{f}}(\sigma^2 I + K_{\boldsymbol{f}\boldsymbol{f}})^{-1}\boldsymbol{y},$$
$$(x, y) \mapsto k(x, y) - K_{x\boldsymbol{f}}(\sigma^2 I + K_{\boldsymbol{f}\boldsymbol{f}})^{-1} K_{\boldsymbol{f}y},$$

respectively. Here we denote $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))$,

$$\begin{aligned} K_{x\boldsymbol{f}} &= \mathrm{cov}_\Pi(f(x), \boldsymbol{f}) = (k(x, x_1), \ldots, k(x, x_n)) = K_{\boldsymbol{f}x}^T, \\ K_{\boldsymbol{f}\boldsymbol{f}} &= \mathrm{cov}_\Pi(\boldsymbol{f}, \boldsymbol{f}) = [k(x_i, x_j)]_{1 \leq i,j \leq n}, \end{aligned} \tag{2}$$

where we emphasise through the subscript Π that the covariances are computed under the prior Π (and not also the distribution G of the design points). We denote the posterior probability kernel by $\Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})$.

The idea of Titsias (2009b) is to summarise the true posterior through a collection of inducing variables $u_1, \ldots, u_m \in L^2(\Pi)$, which by definition are continuous linear functionals of the prior process on f. By the linearity assumption, the prior process f conditional on $\boldsymbol{u} = (u_1, \dots, u_m)$ is again a GP, with mean and covariance function given by

$$x \mapsto K_{x\boldsymbol{u}} K_{\boldsymbol{uu}}^{-1} \boldsymbol{u}, \tag{3}$$

$$(x, y) \mapsto k(x, y) - K_{x\boldsymbol{u}} K_{\boldsymbol{uu}}^{-1} K_{\boldsymbol{u}y}, \tag{4}$$

where $K_{x\boldsymbol{u}} = \mathrm{cov}_\Pi(f(x), \boldsymbol{u}) = K_{\boldsymbol{u}x}^T$ and $K_{\boldsymbol{uu}} = [\mathrm{cov}_\Pi(u_i, u_j)]_{1 \leq i,j \leq m}$. This motivates the construction of a variational family of measures approximating the posterior by postulating that the vector $\boldsymbol{u}$ has a Gaussian distribution with some mean $\mu \in \mathbb{R}^m$ and $m \times m$ covariance matrix Σ, and that the conditional $f \,|\, \boldsymbol{u}$ is the GP law given by (3)-(4). This results in a variational family of GP laws indexed by variational parameters μ and Σ. Explicitly, for fixed μ and Σ, the variational approximation to the posterior is a GP with mean and covariance function given by

$$x \mapsto K_{x\boldsymbol{u}} K_{\boldsymbol{uu}}^{-1} \mu,$$

$$(x, y) \mapsto k(x, y) - K_{x\boldsymbol{u}} K_{\boldsymbol{uu}}^{-1} (K_{\boldsymbol{uu}} - \Sigma) K_{\boldsymbol{uu}}^{-1} K_{\boldsymbol{u}y},$$

cf. also equation (2) in Burt et al. (2019). We denote this member of the variational family by $\Psi_{\mu,\Sigma}(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})$.

It can be shown that for all μ and Σ, the approximation $\Psi_{\mu,\Sigma}(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})$ and the true posterior $\Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})$ are equivalent measures (the Radon-Nikodym derivative reduces to a finite-dimensional Gaussian derivative, a function of at most $m + n$ variables). Hence their Kullback-Leibler divergence is well defined. Titsias (2009a) proves that there exist optimal μ' and Σ' such that

$$\begin{aligned}\inf_{\mu,\Sigma} D_{\mathrm{KL}}\Big(\Psi_{\mu,\Sigma}(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y}) \,\Big\|\, \Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})\Big) &= D_{\mathrm{KL}}\Big(\Psi_{\mu',\Sigma'}(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y}) \,\Big\|\, \Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})\Big) \\ &= \frac{1}{2}\Big(\boldsymbol{y}^T (Q_n^{-1} - K_n^{-1}) \boldsymbol{y} + \log \frac{|Q_n|}{|K_n|} + \frac{1}{\sigma^2} \operatorname{tr}(K_n - Q_n)\Big). \end{aligned} \tag{5}$$

Here $K_n = \sigma^2 I + K_{\boldsymbol{ff}}$ and $Q_n = \sigma^2 I + Q_{\boldsymbol{ff}}$, where

$$Q_{\boldsymbol{ff}} = K_{\boldsymbol{fu}} K_{\boldsymbol{uu}}^{-1} K_{\boldsymbol{uf}} \tag{6}$$

with $K_{\boldsymbol{uf}} = \mathrm{cov}_\Pi(\boldsymbol{u}, \boldsymbol{f})$. Even though in Titsias (2009a) the considered distributions are jointly over f and $\boldsymbol{u}$, the Kullback-Leibler divergence does not change when we use the f-marginal distributions, as follows from Matthews et al. (2016), noting that the inducing variables are measurable functions of f.

The variational posterior $\Psi_{\mu',\Sigma'}(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})$ can be seen as a particular rank-m approximation of the full posterior $\Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})$. In the next section we present results about the rate at which it contracts around the true regression function f_0 as $n \to \infty$. Since the precise form of the optimal variational parameters is not important here, we simply denote the variational posterior by $\Psi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y}) = \Psi_{\mu',\Sigma'}(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})$.

3. Posterior contraction rates for Gaussian process priors

We give a brief overview of posterior contraction rates for GP priors. In Appendix A we provide further details and discuss general contraction rate results for (nonparametric) Bayesian methods. Here we focus on the results directly used in our main theorem in the upcoming section.

We study the posterior distribution $\Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})$ under the assumption that the data $(\boldsymbol{x}, \boldsymbol{y})$ are generated according to some fixed, "true" regression function $f_0 \in L^2(\mathcal{X}, G)$. In other words, we suppose (1) holds with f_0 instead of f, or equivalently, the pairs (x_i, y_i) are i.i.d. with Gaussian density

$$p_{f_0}(x, y) = (2\pi\sigma^2)^{-1/2} \exp(-(y - f_0(x))^2/(2\sigma^2))$$

relative to the product of the data-generating measure G and the Lebesgue measure. We denote by P_0 the associated joint distribution of the data and by E_0 its according expectation operator. General theory on Bayesian contraction rates gives conditions under which the posterior corresponding to a GP prior in the nonparametric regression model contracts around the true regression function f_0 at a certain rate $\epsilon_n \to 0$ as the sample size n tends to infinity.

The standard approach for establishing contraction rates, as exposed in Ghosal and van der Vaart (2017), relies on the existence of appropriate hypothesis tests. This is guaranteed when the chosen metric is the Hellinger distance, so the contraction rate is naturally measured relative to this metric on the space of joint densities of the pair (x_i, y_i). Given $f_1, f_2 \in L^2(\mathcal{X}, G)$, this Hellinger distance d_{H} between the two associated Gaussian densities p_{f_1}, p_{f_2} is given by

$$\begin{aligned} d_{\mathrm{H}}(p_{f_1}, p_{f_2})^2 &= \frac{1}{2} \iint \Big(\sqrt{p_{f_1}(x,y)} - \sqrt{p_{f_2}(x,y)}\Big)^2 \, dy \, dG(x) \\ &= \int_{\mathcal{X}} 1 - \exp\Big(-\frac{(f_1(x) - f_2(x))^2}{8\sigma^2}\Big) \, dG(x). \end{aligned} \tag{7}$$

Considering this as a function of (f_1, f_2), the distance d_{H} can be viewed as a metric on the function space $L^2(\mathcal{X}, G)$. In the sequel we shall abuse our notation and simply write $d_{\mathrm{H}}(f_1, f_2)$.

The posterior is said to contract around the truth f_0 at the rate ϵ_n with respect to the Hellinger distance d_{H} if for any sequence $M_n \to \infty$,

$$\mathrm{E}_0 \, \Pi\big(f : d_{\mathrm{H}}(f, f_0) \geq M_n \epsilon_n \mid \boldsymbol{x}, \boldsymbol{y}\big) \to 0 \tag{8}$$

as $n \to \infty$. Loosely speaking, (8) entails that if f_0 generated the data, then, asymptotically, all posterior mass lies in Hellinger balls around f_0 with a radius of the order ϵ_n.

In view of van der Vaart and van Zanten (2008b), for GPs the posterior contraction rate is determined by the concentration function $\varphi_{f_0} : (0, \infty) \to \mathbb{R}$ associated to the GP prior Π, which is defined as

$$\varphi_{f_0}(\epsilon) = \inf_{h \in \mathbb{H} : \|h - f_0\|_{2,G} \leq \epsilon} \|h\|_{\mathbb{H}}^2 - \log \Pi(f : \|f\|_{2,G} \leq \epsilon). \tag{9}$$

Here $\mathbb{H}$ is the Reproducing Kernel Hilbert Space (RKHS) associated to the prior, and $\|\cdot\|_{\mathbb{H}}$ is the corresponding RKHS norm (see e.g. van der Vaart and van Zanten, 2008a or Appendix I of Ghosal and van der Vaart, 2017). Specifically, if $\epsilon_n \to 0$ is such that $n\epsilon_n^2 \to \infty$ and

$$\varphi_{f_0}(\epsilon_n) \leq n\epsilon_n^2, \tag{10}$$

then the posterior distribution contracts at the rate ϵ_n. The following is a slightly more refined version of this statement.

Lemma 1 *Suppose that the concentration function inequality* (10) *holds for some sequence of positive numbers $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$. Then, for every constant $C_2 > 0$ there exists an event A_n in the σ-field generated by $(\boldsymbol{x}, \boldsymbol{y})$ such that $\mathrm{P}_0(A_n) \to 1$ and*

$$\mathrm{E}_0\, \Pi(f : d_{\mathrm{H}}(f, f_0) \geq M_n\epsilon_n \mid \boldsymbol{x}, \boldsymbol{y}) 1_{A_n} \lesssim \exp(-C_2 n\epsilon_n^2). \tag{11}$$

Note that the preceding lemma implies the posterior contraction (8). This inequality together with a bound on the Kullback-Leibler divergence in (5) will help establish our main result, a contraction rate statement for the variational posterior; see Theorem 2 ahead and the lemmas below it. The results leading to Lemma 1 are recalled and discussed in Appendix A. We note that in specific examples, verifying the concentration inequality (10) means analysing the so-called small ball behaviour of the prior GP and the approximation properties of its RKHS (see also Section 6 and Appendix B).

4. Main results

In this paper we are interested in contraction rate results like (8), but for the variational posterior $\Psi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})$ instead of the full posterior $\Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})$. It is intuitively clear that in addition to an assumption like (10), this requires control over the approximation properties of the variational family, which depend on the choice of inducing variables $\boldsymbol{u} = (u_1, \dots, u_n)$. In the following theorem, this is measured in terms of the expected "size" of the difference between the matrices $K_{\boldsymbol{ff}}$ and $Q_{\boldsymbol{ff}}$ (defined in (2) and (6), respectively), which is the covariance matrix of the conditional law of the vector $\boldsymbol{f} = (f(x_1), \dots, f(x_n))$ given $\boldsymbol{u}$ (see (4)). The size of $K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}$ measures how well the vector of inducing variables $\boldsymbol{u}$ summarises the full prior distribution. In short, we characterise the contraction rate of the variational posterior by conditions on the inducing variables and the prior.

Below, $\|A\|$ denotes the spectral norm and $\mathrm{tr}(A)$ is the trace of the square matrix A, and $\mathrm{E}_{\boldsymbol{x}}$ is the expectation over the input variables $\boldsymbol{x}$ alone.

Theorem 2 *Suppose that for $f_0 \in L^2(\mathcal{X}, G)$ and $\epsilon_n \to 0$ such that $n\epsilon_n^2 \to \infty$, the concentration function inequality* (10) *holds. If in addition there exists a constant $C > 0$ (independent of n) such that*

$$\mathrm{E}_{\boldsymbol{x}} \|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| \leq C, \tag{12}$$

$$\mathrm{E}_{\boldsymbol{x}}\, \mathrm{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}) \leq Cn\epsilon_n^2, \tag{13}$$

then the variational posterior contracts around f_0 at the rate ϵ_n, that is, for all sequences $M_n \to \infty$,

$$\mathrm{E}_0\, \Psi\big(f : d_{\mathrm{H}}(f, f_0) \geq M_n\epsilon_n \mid \boldsymbol{x}, \boldsymbol{y}\big) \to 0. \tag{14}$$

as $n \to \infty$.

Proof The concentration inequality also holds with $M_n\epsilon_n$ instead of ϵ_n. Hence, by Lemma 1, there exist events A_n and a constant $C_2 > 0$ such that $\mathrm{P}_0(A_n) \to 1$ and

$$\mathrm{E}_0\,\Pi\big(f : d_{\mathrm{H}}(f, f_0) \geq M_n\epsilon_n \mid \boldsymbol{x}, \boldsymbol{y}\big)1_{A_n} \lesssim e^{-C_2 n M_n^2\epsilon_n^2}.$$

Lemma 13 applied with $\delta_n = C_2 n M_n^2 \epsilon_n^2$ yields

$$\mathrm{E}_0\,\Psi(f : d_{\mathrm{H}}(f, f_0) \geq M_n\epsilon_n \mid \boldsymbol{x}, \boldsymbol{y})1_{A_n} \lesssim \frac{\mathrm{E}_0\,D_{\mathrm{KL}}(\Psi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y}) \,\|\, \Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})) + e^{-C_2 n M_n^2\epsilon_n^2}}{n M_n^2 \epsilon_n^2}.$$

The proof is completed by combining this with $\mathrm{P}_0(A_n^c) \to 0$, and, as we prove now,

$$\mathrm{E}_0\,D_{\mathrm{KL}}(\Psi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y}) \,\|\, \Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})) \leq C_1 n\epsilon_n^2 \tag{15}$$

for some positive constant C_1. By the concentration function inequality, there exist an $h \in \mathbb{H}$ such that $\|h\|_{\mathbb{H}}^2 \leq n\epsilon_n^2$ and $\|f_0 - h\|_{2,G} \leq \epsilon_n$. Applying Lemma 3 ahead with that choice for h and using the assumptions on $K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}$ then establishes (15). ■

It can be seen from the proof that the variational posterior contracts at the same rate as the true posterior if the inequality (15) holds. Since $n\epsilon_n^2 \to \infty$, this means that the Kullback-Leibler divergence need not go to zero in P_0-expectation. The inequality, which is an essential step in the above proof, follows from the next lemma. A crucial difference with Lemma 2 of Burt et al. (2019) is that we consider f_0 to be fixed.

Lemma 3 *For every $f_0 \in L^2(\mathcal{X}, G)$ and $h \in \mathbb{H}$ we have*

$$\begin{aligned}\mathrm{E}_0\,D_{\mathrm{KL}}&\Big(\Psi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y}) \Big\| \Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})\Big) \\ &\leq \frac{1}{\sigma^2}\Big(n\|f_0 - h\|_{2,G}^2 + \|h\|_{\mathbb{H}}^2\,\mathrm{E}_{\boldsymbol{x}}\,\|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| + \mathrm{E}_{\boldsymbol{x}}\,\mathrm{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}})\Big).\end{aligned}$$

Proof The matrix $K_n - Q_n = K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}$ is the covariance matrix of the conditional law of the vector $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))$ given $\boldsymbol{u} = (u_1, \ldots, u_m)$. In particular it is positive semidefinite, which implies that $K_n \geq Q_n$, hence $\log(|Q_n|/|K_n|) \leq 0$. Therefore, the KL-divergence between the variational class and the true posterior can be bounded from above by leaving out the logarithmic term on the right hand side of the identity (5), i.e.

$$D_{\mathrm{KL}}\Big(\Psi_{\mu',\Sigma'}(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y}) \Big\| \Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})\Big) \leq \frac{1}{2}\Big(\boldsymbol{y}^T(Q_n^{-1} - K_n^{-1})\boldsymbol{y} + \frac{1}{\sigma^2}\mathrm{tr}(K_n - Q_n)\Big). \tag{16}$$

Now let $\mathrm{E}_{\boldsymbol{y}}$ be the expectation over $\boldsymbol{y}$, assuming the input variables $\boldsymbol{x}$ are fixed and f_0 is the true regression function, so that $\mathrm{E}_0 = \mathrm{E}_{\boldsymbol{x}}\,\mathrm{E}_{\boldsymbol{y}}$. We have

$$\mathrm{E}_{\boldsymbol{y}}\,\boldsymbol{y}^T(Q_n^{-1} - K_n^{-1})\boldsymbol{y} = \boldsymbol{f}_0^T(Q_n^{-1} - K_n^{-1})\boldsymbol{f}_0 + \sigma^2\,\mathrm{tr}(Q_n^{-1} - K_n^{-1}). \tag{17}$$

For the first term on the right-hand side we write, with $\boldsymbol{h} = (h(x_1), \ldots, h(x_n))$,

$$\begin{aligned}\frac{1}{2}\boldsymbol{f}_0^T(Q_n^{-1} - K_n^{-1})\boldsymbol{f}_0 &\leq \boldsymbol{h}^T Q_n^{-1}(K_n - Q_n)K_n^{-1}\boldsymbol{h} + (\boldsymbol{f}_0 - \boldsymbol{h})^T(Q_n^{-1} - K_n^{-1})(\boldsymbol{f}_0 - \boldsymbol{h}) \\ &\leq \|Q_n^{-1}\|\|K_n - Q_n\|\boldsymbol{h}^T K_n^{-1}\boldsymbol{h} + (\boldsymbol{f}_0 - \boldsymbol{h})^T Q_n^{-1}(\boldsymbol{f}_0 - \boldsymbol{h}) \\ &\leq \frac{1}{\sigma^2}\Big(\|K_n - Q_n\|\boldsymbol{h}^T K_{\boldsymbol{ff}}^{-1}\boldsymbol{h} + \sum_{i=1}^n (f_0(x_i) - h(x_i))^2\Big),\end{aligned}$$

where we used that $K_n = \sigma^2 I + K_{\boldsymbol{ff}} \geq K_{\boldsymbol{ff}}$ and $Q_n = \sigma^2 I + Q_{\boldsymbol{ff}} \geq \sigma^2 I$. The quantity $\boldsymbol{h}^T K_{\boldsymbol{ff}}^{-1} \boldsymbol{h}$ is the squared RKHS norm of the orthogonal projection in $\mathbb{H}$ of the function h on the linear span of the functions $k(x_1, \cdot), \dots, k(x_n, \cdot)$. Since orthogonal projections decrease norms, we have $\boldsymbol{h}^T K_{\boldsymbol{ff}}^{-1} \boldsymbol{h} \leq \|h\|_{\mathbb{H}}^2$.

For the second term in (17) we note that

$$\operatorname{tr}(Q_n^{-1} - K_n^{-1}) = \operatorname{tr}(Q_n^{-1}(K_n - Q_n) K_n^{-1}) \leq \|Q_n^{-1}\| \|K_n^{-1}\| \operatorname{tr}(K_n - Q_n),$$

where the matrix norms appearing on the right are both bounded by σ^{-2}.

Together we get

$$\frac{1}{2} \mathrm{E}_{\boldsymbol{y}}\, \boldsymbol{y}^T (Q_n^{-1} - K_n^{-1}) \boldsymbol{y} \leq \frac{1}{\sigma^2} \Big(\|K_n - Q_n\| \|h\|_{\mathbb{H}}^2 + \sum_{i=1}^n (f_0(x_i) - h(x_i))^2 + \frac{1}{2} \operatorname{tr}(K_n - Q_n) \Big).$$

Combining this with (16), taking expectations over $\boldsymbol{x}$ and recalling that $K_n - Q_n = K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}$, we arrive at the statement of the lemma. ■

In the next section, we present two choices of inducing variables, also considered in Burt et al. (2019), to which we apply Theorem 2.

5. Inducing variables from eigendecompositions

The covariance operator T_k on $L^2(\mathcal{X}, G)$ associated with the kernel k is defined as

$$T_k \psi(y) = \int_{\mathcal{X}} k(x, y) \psi(x)\, dG(x). \tag{18}$$

Note that this definition depends on the distribution G of the design points. Since k is a covariance kernel, the operator T_k is positive (meaning $\langle T_k \psi, \psi \rangle \geq 0$ for all $\psi \in L^2(\mathcal{X}, G)$). We assume that $k \in L^\infty(G \times G)$. One of the assertions of Mercer's Theorem (see e.g. König, 1986) is that consequently, T_k is a Hilbert-Schmidt operator, and thus compact. It follows that T_k has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \to 0$.

The covariance kernels used in practice satisfy these mild assumptions. We focus on three such kernels in this paper: the Matérn kernel, the squared exponential kernel, and the kernel of a random series prior. For each kernel we consider one or two choices of inducing variables and discuss the conditions of Theorem 2: we study the concentration function inequality (10) and analyse the expected norm and trace terms (12) and (13). The latter is done with help of the eigenvalues of the operator T_k. We consider kernels whose associated operator T_k has exponentially or polynomially decreasing eigenvalues, that is, for $j = 1, 2, \dots$, we assume one of the conditions

$$\lambda_j \leq C_{\exp} b_n e^{-D_{\exp} b_n j}, \tag{19}$$

$$C_\alpha^{-1} j^{-1-2\alpha/d} \leq \lambda_j \leq C_\alpha j^{-1-2\alpha/d}, \tag{20}$$

for $0 < b_n \leq 1$ and positive constants $C_{\exp}, D_{\exp}, C_\alpha$.

5.1 Using the eigendecomposition of the covariance matrix

In this case we construct inducing variables using the m largest eigenvalues and the corresponding eigenvectors of the matrix $K_{\boldsymbol{ff}} = [k(x_i, x_j)]_{1\leq i,j\leq n}$. We define

$$u_j = \boldsymbol{v}_j^T \boldsymbol{f} = \sum_{i=1}^{n} v_j^i f(x_i), \qquad j = 1, \dots, m, \tag{21}$$

where $\boldsymbol{v}_j = (v_j^1, v_j^2, \dots, v_j^n)$ is the eigenvector corresponding to the jth largest eigenvalue μ_j of the matrix $K_{\boldsymbol{ff}}$. Note that each u_j is a linear functional of f, and more precisely a linear combination of the values of f evaluated at the observations $\boldsymbol{x}$. It is easy to verify (see also Section C.1. of Burt et al., 2019) that in this case we have

$$(K_{\boldsymbol{uu}})_{ij} = \mathrm{cov}_\Pi(u_i, u_j) = \mu_j \delta_{ij},$$
$$(K_{\boldsymbol{fu}})_{ij} = \mathrm{cov}_\Pi(f(x_i), u_j) = \mu_j v_j^i.$$

Hence in view of the identity $K_{\boldsymbol{ff}} = \sum_{j=1}^n \mu_j \boldsymbol{v}_j \boldsymbol{v}_j^T$,

$$Q_{\boldsymbol{ff}} = K_{\boldsymbol{fu}} K_{\boldsymbol{uu}}^{-1} K_{\boldsymbol{uf}} = \sum_{j=1}^{m} \mu_j \boldsymbol{v}_j \boldsymbol{v}_j^T,$$
$$K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}} = \sum_{j=m+1}^{n} \mu_j \boldsymbol{v}_j \boldsymbol{v}_j^T. \tag{22}$$

Note that with this choice of $\boldsymbol{u}$ the matrix $Q_{\boldsymbol{ff}}$ is the optimal rank-m approximation of $K_{\boldsymbol{ff}}$. The computational complexity of obtaining the first m eigenvalues and the corresponding eigenvectors of $K_{\boldsymbol{ff}}$ numerically is $O(mn^2)$, by using for instance the Lanczos iteration (Lanczos, 1950). Analytical expressions for the eigenvalues and eigenvectors of $K_{\boldsymbol{ff}}$ are not available for the majority of commonly used kernels.

Since the eigenvectors $\boldsymbol{v}_j$ are orthogonal,

$$\|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| = \mu_{m+1}, \tag{23}$$
$$\mathrm{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}) = \sum_{j=m+1}^{n} \mu_j. \tag{24}$$

Burt et al. (2020) explain that this choice of $Q_{\boldsymbol{ff}}$ is the minimiser of both these quantities. As such, the right-hand sides of the above identities serve as benchmarks for other choices of inducing variables. To bound these, we will use repeatedly the part of Proposition 2 in Shawe-Taylor and Williams (2002) stating that

$$\mathrm{E}_{\boldsymbol{x}} \sum_{j=j_0}^{n} \mu_j / n \leq \sum_{j=j_0}^{\infty} \lambda_j \tag{25}$$

for all j_0 between 1 and n.

We bound the expected trace and norm terms in Theorem 2. For exponentially decreasing eigenvalues (19) this is straightforward. Indeed, from (24) and (25) we obtain

$$\mathrm{E}_{\boldsymbol{x}} \|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| \leq \mathrm{E}_{\boldsymbol{x}} \operatorname{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}) \leq n \sum_{j=m+1}^{\infty} \lambda_j \lesssim n \sum_{j=m+1}^{\infty} b_n e^{-D_{\exp} b_n j} \lesssim n e^{-D_{\exp} b_n m}, \tag{26}$$

which suffices for our purposes. Polynomially decaying eigenvalues require more work as we need to do better than bounding the operator norm by the trace.

Lemma 4 *If the eigenvalues $\lambda_1, \lambda_2, \ldots$ of the operator* (18) *are polynomially decaying* (20)*, then there is a constant $\bar{C}_\alpha$ such that*

$$\mathrm{E}_{\boldsymbol{x}} \|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| \leq \bar{C}_\alpha n m^{-1-2\alpha/d},$$
$$\mathrm{E}_{\boldsymbol{x}} \operatorname{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}) \leq \bar{C}_\alpha n m^{-2\alpha/d},$$

for any $2 \leq m \leq n$.

Proof We deal with the norm term using (23). We argue by contradiction. Suppose that for all $i \in \{m/2, \ldots, m\}$ we have $\mathrm{E}_{\boldsymbol{x}}\, \mu_i/n > \tilde{C}_\alpha \lambda_i$, where $\tilde{C}_\alpha = 1 + dC_\alpha^2/\alpha$. Since

$$\sum_{i=m+1}^{\infty} \lambda_i \leq C_\alpha \sum_{i=m+1}^{\infty} i^{-1-2\alpha/d} \leq C_\alpha \int_m^\infty t^{-1-2\alpha/d} dt = \frac{C_\alpha d}{2\alpha} m^{-2\alpha/d}, \tag{27}$$

$$\sum_{i=m/2}^{m} \lambda_i \geq C_\alpha^{-1} \sum_{i=m/2}^{m} i^{-1-2\alpha/d} \geq (2C_\alpha)^{-1} m^{-2\alpha/d},$$

we have

$$\mathrm{E}_{\boldsymbol{x}} \sum_{i=m/2}^{n} \mu_i/n \geq \mathrm{E}_{\boldsymbol{x}} \sum_{i=m/2}^{m} \mu_i/n > \tilde{C}_\alpha \sum_{i=m/2}^{m} \lambda_i \geq \sum_{i=m/2}^{\infty} \lambda_i,$$

but this contradicts (25). Therefore there exists $i \in \{m/2, \ldots, m\}$ such that $\mathrm{E}_{\boldsymbol{x}}\, \mu_i/n \leq \tilde{C}_\alpha \lambda_i$. Hence

$$\mathrm{E}_{\boldsymbol{x}}\, \mu_{m+1} \leq \mathrm{E}_{\boldsymbol{x}}\, \mu_i \leq n\tilde{C}_\alpha \lambda_i \leq n\tilde{C}_\alpha \lambda_{m/2} \leq (\tilde{C}_\alpha C_\alpha 2^{1+2\alpha/d}) m^{-1-2\alpha/d} n,$$

hence, recalling (23), we obtain the bound on the expected norm term.

Regarding the trace term, the inequality (25) implies

$$\mathrm{E}_{\boldsymbol{x}} \operatorname{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}) = \mathrm{E}_{\boldsymbol{x}} \sum_{i=m+1}^{n} \mu_i \leq n \sum_{i=m+1}^{\infty} \lambda_i.$$

The inequality regarding the trace in the statement of the lemma then follows immediately using (27). ■

We turn to the second choice of inducing variables before applying these results to the chosen kernels.

5.2 Using the eigendecomposition of the covariance operator

The previous method requires computing the eigenvalues and the eigenvectors of the matrix $K_{\boldsymbol{ff}}$, which for large data sets becomes computationally demanding. Another choice of inducing variables is

$$u_j = \int_{\mathcal{X}} f(x)\varphi_j(x)\, dG(x), \qquad j = 1, \dots, m, \tag{28}$$

where $\varphi_1, \varphi_2, \dots$ are the eigenfunctions of the kernel operator T_k, corresponding to the eigenvalues $\lambda_1, \lambda_2, \dots$, so $\int k(x,y)\varphi_i(x)\, dG(x) = \lambda_i \varphi_i(y)$. In case $\mathcal{X}$ is a compact interval and the functions φ_j form a Fourier series, this choice of inducing variables yields the variational Fourier features described in Hensman et al. (2018).

The relevant covariance matrices for the inducing variables (28) are

$$(K_{\boldsymbol{uu}})_{ij} = \mathrm{cov}_\Pi(u_i, u_j) = \lambda_j \delta_{ij},$$
$$(K_{\boldsymbol{fu}})_{ij} = \mathrm{cov}_\Pi(f(x_i), u_j) = \lambda_j \varphi_j(x_i)$$

(see again Appendix C of Burt et al., 2019 for the proof of these statements). Then in view of Mercer's theorem, $K_{\boldsymbol{ff}} = \sum_{j=1}^\infty \lambda_j \boldsymbol{\varphi}_j \boldsymbol{\varphi}_j^T$ where we denote $\boldsymbol{\varphi}_j = (\varphi_j(x_1), \dots, \varphi_j(x_n))$, so

$$Q_{\boldsymbol{ff}} = \sum_{j=1}^m \lambda_j \boldsymbol{\varphi}_j \boldsymbol{\varphi}_j^T,$$
$$K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}} = \sum_{j=m+1}^\infty \lambda_j \boldsymbol{\varphi}_j \boldsymbol{\varphi}_j^T.$$

(Note that unlike the $\boldsymbol{v}_j$ from the previous section, the vectors $\boldsymbol{\varphi}_j$ do not necessarily form an orthonormal basis of $\mathbb{R}^n$.)

With this choice of inducing variables, we obtain for the expected trace term

$$\mathrm{E}_x \operatorname{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}) = \sum_{j=m+1}^\infty \lambda_j \sum_{i=1}^n \mathrm{E}_{\boldsymbol{x}}\, \varphi_j(x_i)^2 = n \sum_{j=m+1}^\infty \lambda_j. \tag{29}$$

This is exactly the upper bound we obtained for the trace term in the previous section. For the exponentially decaying eigenvalues, we bound the operator norm just as in (26) by

$$\mathrm{E}_{\boldsymbol{x}} \|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| \le \mathrm{E}_{\boldsymbol{x}} \operatorname{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}) \lesssim n e^{-D_{\exp} b_n m}. \tag{30}$$

The results regarding the polynomially decreasing eigenvalues are summarized in the next lemma.

Lemma 5 *Assume that the eigenvalues $\lambda_1, \lambda_2, \dots$ of the operator* (18) *are polynomially decaying* (20)*, with $\alpha > d$. Suppose the corresponding eigenfunctions of the operator T_k are uniformly bounded. Then*

$$\mathrm{E}_{\boldsymbol{x}} \|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| \lesssim 1 + n m^{-1-2\alpha/d} + n^{d/(2\alpha)} m^{-2\alpha/d} \log n,$$
$$\mathrm{E}_{\boldsymbol{x}} \operatorname{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}) \le \frac{C_\alpha d}{2\alpha} n m^{-2\alpha/d}.$$

Proof The expected trace inequality follows upon combining (29) and (27). The expectation of the spectral norm is bounded by distributing over the event

$$A_n(C) = \{\boldsymbol{x} \in \mathcal{X}^n : |\langle \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_k \rangle - n\delta_{jk}| \le C\sqrt{n \log n}, \quad m < j, k \le n^{d/(2\alpha)}\},$$

and its complement.

In view of Lemma 6 below, there exists a large enough $C > 0$ such that $\mathrm{P}_{\boldsymbol{x}}(A_n(C)^c) \le n^{-1}$. Using the crude estimate

$$\|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| \le \mathrm{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}) = \sum_{j=m+1}^{\infty} \sum_{i=1}^{n} \lambda_j \varphi_j(x_i)^2 \le n C_\varphi \sum_{j=1}^{\infty} \lambda_j \lesssim n$$

(the constant C_φ being the uniform bound for the φ_j) we then obtain

$$\mathrm{E}_{\boldsymbol{x}} \mathbf{1}_{A_n(C)^c} \|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| \le n\, \mathrm{P}_{\boldsymbol{x}}(A_n(C)^c) \lesssim 1.$$

On the event $A_n(C)$ we use

$$\begin{aligned}\mathrm{E}_{\boldsymbol{x}} \mathbf{1}_{A_n(C)} \|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| &\le \mathrm{E}_{\boldsymbol{x}} \mathbf{1}_{A_n(C)} \Big\| \sum_{k=m+1}^{n^{d/(2\alpha)}} \lambda_k \boldsymbol{\varphi}_k \boldsymbol{\varphi}_k^T \Big\| + \mathrm{E}_{\boldsymbol{x}} \Big\| \sum_{k>n^{d/(2\alpha)}} \lambda_k \boldsymbol{\varphi}_k \boldsymbol{\varphi}_k^T \Big\| \\ &\le \mathrm{E}_{\boldsymbol{x}} \mathbf{1}_{A_n(C)} \max_{\|v\|_2=1} v^T \Big(\sum_{k=m+1}^{n^{d/(2\alpha)}} \lambda_k \boldsymbol{\varphi}_k \boldsymbol{\varphi}_k^T \Big) v + \mathrm{E}_{\boldsymbol{x}} \,\mathrm{tr} \Big(\sum_{k>n^{d/(2\alpha)}} \lambda_k \boldsymbol{\varphi}_k \boldsymbol{\varphi}_k^T \Big),\end{aligned}$$

where the last inequality follows from the positive semi-definiteness of the matrices $\lambda_k \boldsymbol{\varphi}_k \boldsymbol{\varphi}_k^T$. The second bounding term equals

$$\mathrm{tr} \Big(\sum_{k>n^{d/(2\alpha)}} \lambda_k \, \mathrm{E}_{\boldsymbol{x}} \, \boldsymbol{\varphi}_k \boldsymbol{\varphi}_k^T \Big) = n \sum_{k>n^{d/(2\alpha)}} \lambda_k \lesssim n \sum_{k>n^{d/(2\alpha)}} k^{-1-2\alpha/d} \lesssim 1.$$

Lastly, we deal with the first term by bounding

$$\max_{\|v\|_2=1} v^T \Big(\sum_{k=m+1}^{n^{d/(2\alpha)}} \lambda_k \boldsymbol{\varphi}_k \boldsymbol{\varphi}_k^T \Big) v = \max_{\|v\|_2=1} \sum_{k=m+1}^{n^{d/(2\alpha)}} \lambda_k \langle v, \boldsymbol{\varphi}_k \rangle^2.$$

on the event $A_n(C)$. It is sufficient to consider vectors v of the form $v = \sum_{k=m+1}^{n^{d/(2\alpha)}} \rho_k \boldsymbol{\varphi}_k$. On the event $A_n(C)$, using that $\alpha > d$,

$$\begin{aligned} 1 = \|v\|_2^2 = \sum_{k,j=m+1}^{n^{d/(2\alpha)}} \rho_j \rho_k \langle \boldsymbol{\varphi}_j, \boldsymbol{\varphi}_k \rangle &\ge \sum_{k,j=m+1}^{n^{d/(2\alpha)}} \rho_j \rho_k \Big(n\delta_{jk} - C\sqrt{n \log n} \Big) \\ \ge \sum_{k=m+1}^{n^{d/(2\alpha)}} \rho_k^2 \Big(n - n^{d/(2\alpha)} C \sqrt{n \log n} \Big) &\ge \frac{n}{2} \sum_{k=m+1}^{n^{d/(2\alpha)}} \rho_k^2, \end{aligned}$$

and therefore

$$\begin{aligned}
\max_{\|v\|_2=1}\sum_{k=m+1}^{n^{d/(2\alpha)}}\lambda_k\langle v,\boldsymbol{\varphi}_k\rangle^2 &= \max_{\|v\|_2=1}\sum_{k=m+1}^{n^{d/(2\alpha)}}\lambda_k\Big(\sum_{j=m+1}^{n^{d/(2\alpha)}}\rho_j\langle\boldsymbol{\varphi}_j,\boldsymbol{\varphi}_k\rangle\Big)^2\\
&\leq \max_{\|v\|_2=1}\sum_{k=m+1}^{n^{d/(2\alpha)}}\lambda_k\Big(\sum_{j=m+1}^{n^{d/(2\alpha)}}|\rho_j|(n\delta_{jk}+C\sqrt{n\log n})\Big)^2\\
&\lesssim \max_{\|v\|_2=1}\sum_{k=m+1}^{n^{d/(2\alpha)}}\lambda_k\Big(n^2\rho_k^2+n^{\frac{d+2\alpha}{2\alpha}}\log n\sum_{j=m+1}^{n^{d/(2\alpha)}}\rho_j^2\Big)\\
&\lesssim n\lambda_{m+1}\big(\max_{\|v\|_2=1}n\sum_{k=m+1}^{n^{d/(2\alpha)}}\rho_k^2\big)+n^{d/(2\alpha)}\log n\sum_{k=m+1}^{n^{d/(2\alpha)}}\lambda_k\\
&\lesssim nm^{-1-2\alpha/d}+n^{d/(2\alpha)}m^{-2\alpha/d}\log n.
\end{aligned}$$

The proof is concluded by multiplying with $\mathbf{1}_{A_n(C)}$ and taking expectations $\mathrm{E}_{\boldsymbol{x}}$ in the above display. ■

The following lemma provides the concentration inequality for the empirical inner product of the eigenfunctions, used in the proof of the preceding lemma.

Lemma 6 *For orthonormal functions $\varphi_1,\varphi_2,\ldots,\varphi_{M_n}$ w.r.t. the measure G such that $|\varphi_i|\leq C_\varphi$ on $\mathcal{X}$ and $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)$ i.i.d. with common distribution G, the random vectors $\boldsymbol{\varphi}_\ell=(\varphi_\ell(x_1),\ldots,\varphi_\ell(x_n))$ satisfy*

$$\mathrm{P}_{\boldsymbol{x}}\Big(\sup_{1\leq\ell,k\leq M_n}|\langle\boldsymbol{\varphi}_\ell,\boldsymbol{\varphi}_k\rangle-n\delta_{\ell k}|\geq C\sqrt{n\log n}\Big)\leq M_n^2 n^{-(C/C_\varphi^2)^2/2}$$

for any $C>0$.

Proof By the subadditivity of the probability and using Hoeffding's inequality for bounded random variables we get that

$$\begin{aligned}
&\mathrm{P}_{\boldsymbol{x}}(\sup_{1\leq\ell,k\leq M_n}|\langle\boldsymbol{\varphi}_\ell,\boldsymbol{\varphi}_k\rangle-n\delta_{\ell k}|\geq C\sqrt{n\log n})\\
&\qquad\leq M_n^2\sup_{\ell,k}\mathrm{P}_{\boldsymbol{x}}(|n^{-1}\langle\boldsymbol{\varphi}_\ell,\boldsymbol{\varphi}_k\rangle-\delta_{\ell k}|\geq C\sqrt{n^{-1}\log n})\\
&\qquad\leq M_n^2\exp\{-\frac{2n^2C^2n^{-1}\log n}{n4C_\varphi^4}\}=M_n^2\exp\{-\frac{C^2}{2C_\varphi^4}\log n\},
\end{aligned}$$

finishing the proof of the statement. ■

6. Concrete examples

We consider three explicit examples to demonstrate how the approximation theory from the previous section can be used to apply the main theorem in Section 4. The contraction rates we obtain depend on the smoothness properties of the underlying true regression function f_0. To make this precise, we recall the definition of two smoothness classes.

The Hölder space $C^\alpha(\mathcal{X})$ of smoothness $\alpha > 0$ consists of those functions on $\mathcal{X}$ with Hölder regularity α. This means partial derivatives of order up to $\alpha_0 := \lceil \alpha \rceil - 1$ exist and are uniformly bounded, and derivatives of order equal to α_0 satisfy a Hölder condition with exponent $\alpha - \alpha_0$.

The Sobolev space $H^\alpha(\mathcal{X})$ is the collection of restrictions $f_0|_{\mathcal{X}}$ to $\mathcal{X}$ of functions $f_0 : \mathbb{R}^d \to \mathbb{R}$ with Fourier transform $\hat{f}_0(\lambda) = (2\pi)^{-d} \int_{\mathbb{R}^d} e^{i\langle \lambda, x\rangle} f_0(x)\, dx$ satisfying

$$\int (1 + \|\lambda\|^2)^\alpha |\hat{f}_0(\lambda)|^2\, d\lambda < \infty.$$

For $\alpha \in \mathbb{N}$ the space $H^\alpha(\mathcal{X})$ coincides with the space of functions with square integrable weak α-derivatives over $\mathcal{X}$.

6.1 Matérn kernel

The Matérn prior is the centered GP whose covariance kernel is

$$k(x, y) = c_1 \|x - y\|^\alpha K_\alpha(c_2\|x - y\|), \tag{31}$$

where c_1, c_2, α are positive constants and K_α is the modified Bessel function of the second kind (see Rasmussen and Williams, 2006). If $\mathcal{X} = [0, 1]^d$ and $f_0 \in C^\alpha(\mathcal{X}) \cap H^\alpha(\mathcal{X})$, then it is known that the true posterior contracts around f_0 at the rate $n^{-\alpha/(d+2\alpha)}$; see e.g. van der Vaart and van Zanten (2011). This is the optimal minimax rate of contraction for this problem. The following corollary asserts that if the number of inducing variables is chosen at least of the order $n^{d/(d+2\alpha)}$, then for the first class of inducing variables considered above, the variational posterior attains this optimal rate as well.

Corollary 7 *Let k be the Matérn kernel* (31) *on $\mathcal{X} = [0, 1]^d$ and let G be a distribution with bounded Lebesgue density. Suppose that the inducing variables* (21) *are used and $\alpha > d/2$. Then the variational posterior contracts around $f_0 \in C^\alpha(\mathcal{X}) \cap H^\alpha(\mathcal{X})$ at the rate $\epsilon_n = n^{-\alpha/(d+2\alpha)}$ for $m = m_n \geq n^{d/(d+2\alpha)}$.*

Proof It follows from the assumptions on f_0 and α combined with (results leading to) Theorem 5 in van der Vaart and van Zanten (2011) that $\varphi(\epsilon) \lesssim \epsilon^{-d/\alpha}$, so the concentration function inequality (10) holds for ϵ_n as specified.

The assumptions on G allow for an application of Theorem 1 in Seeger (2007), whose proof yields (20) for the eigenvalues of the kernel operator T_k. Lemma 4 implies that the trace and norm inequalities in Theorem 2 hold for m as given. This yields the contraction statement for the variational posterior. ■

The other choice of inducing variables (28) is not considered here, since for the stationary Matérn process we don't have access to the eigenfunctions of the kernel operator T_k. For

G equal to the uniform distribution on $[0,1]^d$, finding the eigenfunctions and eigenvalues is equivalent to finding the Karhunen-Loève expansion. Explicit expressions appear only to be available for the case $\alpha = 1/2$ of the Ornstein-Uhlenbeck process, see for instance Corlay and Pagès (2015).

6.2 Squared exponential kernel

The squared exponential process on $\mathcal{X} = \mathbb{R}^d$ with length scale $b > 0$ is the centered GP on $\mathbb{R}^d$ with covariance function

$$k(x,y) = \exp(-\|x-y\|^2/b^2). \tag{32}$$

The structure of the RKHS and sharp bounds for the concentration function are known for this process, but in existing results the process is usually viewed on a compact subset of $\mathbb{R}^d$ and the concentration function relative to the uniform norm is considered, see for instance van der Vaart and van Zanten (2009), van der Vaart and van Zanten (2011). In this paper we want to consider the example that G is a normal distribution, in which case the existing results do not directly apply. Therefore we adapt the relevant results, viewing the squared exponential process as a random element in the space $L^2(\mathcal{X}, G)$.

We formulate the following lemma for slightly more general distributions G with sub-Gaussian tails, that is, we assume that there exist constants $C_1, C_2 > 0$ such that

$$G(x : \|x\| > a) \leq C_1 e^{-C_2 a^2} \tag{33}$$

for all $a > 0$ large enough. It is seen from the proof that the statement of the lemma can easily be adapted to cases with different tail behaviours.

Lemma 8 *Let k be the squared exponential kernel* (32) *with length scale $b = b_n = n^{-1/(d+2\alpha)}$. Suppose that $f_0 \in C^\alpha(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, and G satisfies the sub-Gaussian tail bound* (33) *on $\mathcal{X} = \mathbb{R}^d$. Then the concentration function inequality* (10) *is satisfied for ϵ_n a multiple of $n^{-\alpha/(d+2\alpha)} \log^{\kappa/2} n$, where $\kappa = 1 + 3d/2$.*

Proof This follows from combining Lemma 15 and 16 in Appendix B. ■

By the results of van der Vaart and van Zanten (2009), under the assumptions of the above lemma, the true posterior contracts around f_0 at the optimal rate $n^{-\alpha/(d+2\alpha)}$, up to a logarithmic factor. The following corollary asserts that if $d = 1$ and G is a normal distribution, the same is true for the variational posteriors considered above.

Corollary 9 *Let k be the squared exponential kernel* (32) *with $b = b_n = n^{-1/(1+2\alpha)}$, and G a centered Gaussian distribution on $\mathcal{X} = \mathbb{R}$. Then the variational posterior using either choice of inducing variables* (21) *or* (28) *contracts around $f_0 \in C^\alpha(\mathbb{R}) \cap L^2(\mathbb{R})$ at the rate $\epsilon_n = n^{-\alpha/(1+2\alpha)}(\log n)^{5/4}$, provided that $m = m_n \geq D_{\text{exp}}^{-1} n^{1/(1+2\alpha)} \log n$.*

Proof In Lemma 8 we have already established that the concentration function inequality is satisfied for the specified truth f_0, scale b_n, and rate ϵ_n.

We now prove the eigenvalues of the covariance operator are exponentially decaying. For notational convenience suppose that $a > 0$ is such that G has density $p(x) \propto e^{-2ax^2}$. There is an explicit expression for the eigenvalues (see Rasmussen and Williams, 2006)

$$\lambda_j = \sqrt{2a/A_n}\Big(\frac{1}{A_n b_n^2}\Big)^{j-1}, \quad j = 1, 2, \ldots$$

with $A_n = a + b_n^{-2} + \sqrt{a^2 + 2ab_n^{-2}}$. We note that

$$\frac{1}{A_n b_n^2} = 1 - z_n \leq e^{-z_n}$$

for $z_n = \sqrt{a^2 b_n^4 + 2ab_n^2} - ab_n^2$, and $z_n/b_n \to \sqrt{2a}$ as $n \to \infty$, so $z_n > D_{\exp} b_n$ when $0 < D_{\exp} < \sqrt{2a}$ and n is sufficiently large. Then

$$\lambda_j \leq \sqrt{2a/A_n} e^{-z_n j} \lesssim b_n e^{-D_{\exp} b_n j},$$

so we are in the situation of (19). By (26) and (29), the choice of m yields

$$\mathrm{E}_{\boldsymbol{x}} \|K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}\| \leq \mathrm{E}_{\boldsymbol{x}} \operatorname{tr}(K_{\boldsymbol{ff}} - Q_{\boldsymbol{ff}}) \lesssim n e^{-D_{\exp} b_n m} \leq 1,$$

so the conditions of Theorem 2 are satisfied. ■

Remark 10 *A stronger requirement on the smoothness of f_0 is that it belongs to the RKHS $\mathbb{H}$ associated to the prior. In this case the RKHS approximation term in the concentration function is bounded by a constant, so the contraction rate is characterised by the small ball probability which is bounded in Lemma 15. One can take a fixed length scale $b > 0$, so that the concentration function inequality holds for ϵ_n satisfying*

$$\Big(\log \frac{1}{\epsilon_n}\Big)^{\kappa} \lesssim n\epsilon_n^2.$$

This is fulfilled by the rate $\epsilon_n = n^{-1/2}(\log n)^{\kappa/2}$, which is almost the parametric rate $n^{-1/2}$. By the arguments used to establish the above corollary, the variational posterior contracts at this rate when m_n is taken of the order $\log n$. *This is also what Burt et al. (2019) suggest for the exponential kernel. Our Corollary 9 illustrates that this choice may not be optimal if f_0 is not so smooth that it belongs to the RKHS of the squared exponential kernel, which only contains analytic functions. See also the numerical illustration in Section 7.*

6.3 Random series prior

The last choice of kernel is one defined through a series expansion. We take $\mathcal{X} = [0, 1]^d$ and consider a uniform distribution G for the design points. Let (φ_j) be an orthonormal basis of the corresponding function space $L^2[0, 1]^d$. Suppose that the basis functions are continuous and uniformly bounded, that is, $\sup_j \sup_x |\varphi_j(x)| < \infty$. Define, for $\alpha > 0$, the series

$$f(x) = \sum_{j=1}^{\infty} j^{-1/2-\alpha/d} \varphi_j(x) Z_j, \qquad x \in [0, 1]^d, \tag{34}$$

where (Z_j) is a sequence of i.i.d. standard normal random variables. The series converges uniformly and the resulting process $(f(x) : x \in [0,1]^d)$ is a centered GP with covariance function

$$k(x,y) = \sum_{j=1}^{\infty} j^{-1-2\alpha/d} \varphi_j(x)\varphi_j(y). \tag{35}$$

By construction, (φ_j) is the orthonormal eigenbasis of the associated operator T_k with eigenvalues $\lambda_j = j^{-1-2\alpha/d}$. We note that one can generalise these priors to compact Riemannian manifolds $\mathcal{X}$ (in fact, the compactness assumption can also be relaxed for appropriate choice of G) and the coefficients $j^{-1/2-\alpha/d}$ can be replaced by any sequence $\sqrt{\lambda_j}$ such that (34) converges.

We consider contraction of the variational posterior corresponding to this prior. A function $f_0 \in L^2[0,1]^d$ has the expansion $f_0 = \sum_{j=1}^{\infty} f_{0,j}\varphi_j$ where $f_{0,j} = \langle f_0, \varphi_j\rangle$. Here we consider functions in the Sobolev space

$$\tilde{H}^\alpha = \{f \in L^2[0,1]^d : \|f\|_\alpha < \infty\}, \qquad \|f\|_\alpha^2 = \sum_j j^{2\alpha/d} |\langle f, \varphi_j\rangle|^2.$$

In general this space is different from the previously defined $H^\alpha([0,1]^d)$ since it depends on the choice of basis functions φ_j. If (φ_j) is the standard Fourier basis in $d = 1$, however, the spaces coincide.

With either choice of inducing variables discussed earlier, the variational posterior contracts around elements of $\tilde{H}^\alpha$ at the minimax rate.

Corollary 11 *Consider the kernel* (35) *for some uniformly bounded orthonormal basis* (φ_j) *of* $L^2[0,1]$ *consisting of continuous functions. Suppose that either*

- *the inducing variables* (21) *are used and* $\alpha > d/2$, *or*
- *the inducing variables* (28) *are used and* $\alpha > d$.

Then the variational posterior contracts around $f_0 \in \tilde{H}^\alpha$ *at the rate* $\epsilon_n = n^{-\alpha/(d+2\alpha)}$ *for* $m = m_n \geq n^{d/(d+2\alpha)}$.

Proof We start by bounding the concentration function. By Theorem 4.1 in van der Vaart and van Zanten (2008a), the function $h := \sum_{j=1}^J \langle f_0, \varphi_j\rangle\varphi_j = \sum_{j=1}^J f_{0,j}\varphi_j$ is an element of the RKHS $\mathbb{H}$ of the prior with squared norm $\|h\|_{\mathbb{H}}^2 = \sum_{j=1}^J |f_{0,j}|^2/\lambda_j$. If $f_0 \in \tilde{H}^\alpha$ then we have

$$\|h\|_{\mathbb{H}}^2 = \sum_{j=1}^J |f_{0,j}|^2 j^{1+2\alpha/d} \leq J\|f_0\|_\alpha^2.$$

Moreover,

$$\|f_0 - h\|_{2,G}^2 = \sum_{j>J} |f_{0,j}|^2 \leq \|f_0\|_\alpha^2 J^{-2\alpha/d}$$

so by choosing J of the order $\epsilon^{-d/\alpha}$, it follows that

$$\inf_{h\in\mathbb{H}:\|h-f_0\|_{2,G}\leq\epsilon} \|h\|_{\mathbb{H}}^2 \lesssim \epsilon^{-d/\alpha}.$$

By the expansion (34), the centered small ball probability can be written as

$$\Pi(f : \|f\|_{2,G} \leq \epsilon) = \Pr\Big(\sum_{j=1}^{\infty} j^{-1-2\alpha/d} Z_j^2 \leq \epsilon^2\Big).$$

By Corollary 4.3 in Dunker et al. (1998),

$$-\log \Pr\Big(\sum_{j=1}^{\infty} j^{-1-2\alpha/d} Z_j^2 \leq \epsilon^2\Big) \lesssim \epsilon^{-d/\alpha}.$$

It follows that the concentration function inequality (10) holds for ϵ_n as specified (up to a constant), and this is the rate at which the true posterior contracts.

Evidently the eigenvalues satisfy (20). The trace and norm inequalities in Theorem 2 are readily verified for our choice of m with the help of either Lemma 4 or Lemma 5. This yields the contraction statement for the variational posterior. ■

7. Numerical experiments

We illustrate the theoretical results by two numerical experiments, varying both the kernel and the choice of inducing variables.

7.1 Matérn kernel – method 1

We simulate $n = 3000$ samples $x_i \sim$ uniform$[0,1]$ and $y_i \sim \mathcal{N}(f_0(x_i), \sigma^2)$ with $\sigma = 0.2$ and

$$f_0(x) = |x - 0.4|^\alpha - |x - 0.2|^\alpha$$

for $\alpha = 0.6$, which is plotted in Figure 1. We use the Matérn-α kernel for the GP prior and study the variational posterior using the inducing variables obtained from the covariance matrix (Section 5.1).

We compare the behaviour of the true and variational Bayes methods for different choices of the number of inducing points. Figures 2 and 3 show the mean and pointwise 95% credible regions (intervals centered vertically around the posterior mean which have posterior mass 0.95) for both the true and variational posterior. According to Corollary 7, m should be at least $n^{1/(1+2\alpha)} \approx 40$. Figure 2 illustrates this: here $m = 40$, and although the variational posterior is in general a bit smoother, its credible region is hardly larger than that of the true posterior. On the contrary, one can conclude from Figure 3 that it is unwise to take a significantly lower number of inducing variables. The variational posterior mean is far too smooth and credible regions are too wide.

Table 1 shows estimates of the expected Kullback-Leibler divergence (5), computed from 100 repetitions of the above experiment for different n. We used $m = n^{1/(1+2\alpha)}$ inducing variables so that by Corollary 7 the variational posterior contracts at the minimax rate. Note that the KL-divergence increases with n, meaning that it does not vanish. This is in according with our theory, which says that the P_0-expectation of the KL-divergence need only be of the order $n\epsilon_n^2 \to \infty$ (see the proof of Theorem 2).

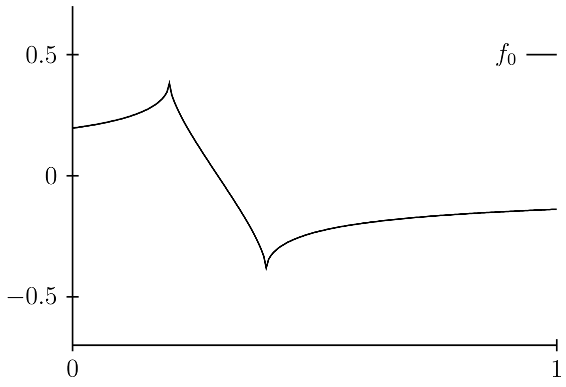

Figure 1: plot of $f_0 = |x+1|^\alpha - |x+3/2|^\alpha$ for $\alpha = 0.8$

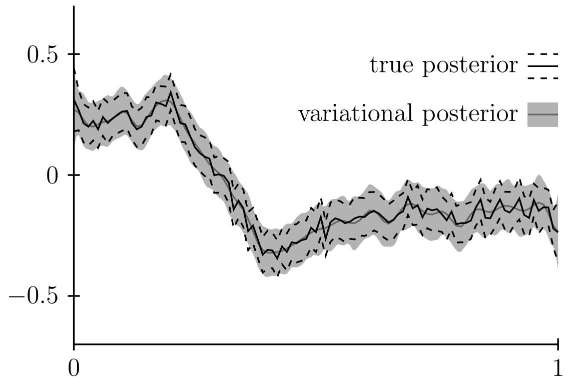

Figure 2: True and variational posterior and credible regions for Matérn prior and $m = 40$ inducing variables from method 1

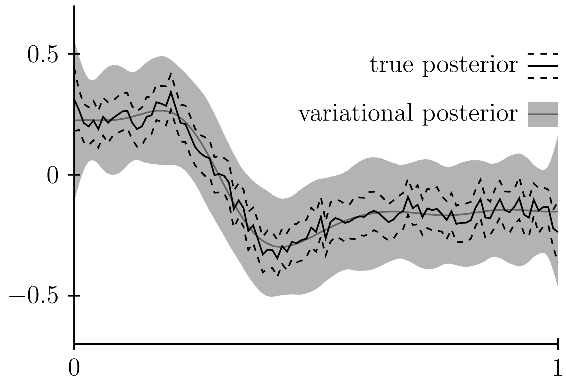

Figure 3: True and variational posterior and credible regions for Matérn prior and $m = 10$ inducing variables from method 1

n	$D_{\mathrm{KL}}(\Psi(\cdot \mid \boldsymbol{x},\boldsymbol{y}) \| \Pi(\cdot \mid \boldsymbol{x},\boldsymbol{y}))$
100	14.71 (1.75)
300	25.20 (2.31)
1000	42.09 (3.23)
3000	68.90 (3.94)

Table 1: Estimates of the KL-divergence between variational and true posterior (average over 100 repeated experiments). Estimated standard deviations are given between brackets.

7.2 Squared exponential kernel – method 2

In a similar fashion, we simulate $n = 5000$ samples $x_i \sim \mathcal{N}(0,1)$ and y_i from the $\mathcal{N}(f_0(x_i), \sigma^2)$ distribution with

$$f_0(x) = |x+1|^\alpha - |x+3/2|^\alpha$$

for $\alpha = 0.8$ and $\sigma = 0.2$. The function f_0 is plotted in Figure 4. Although strictly speaking $f_0 \notin L^2(\mathcal{X}, G)$, one can easily modify its tails maintaining $f_0 \in C^\alpha(\mathbb{R})$ (also note that with high probability all x_i are in a large compact set).

We use the squared exponential kernel as defined in (32) with $b = b_n = 4n^{-1/(1+2\alpha)}$ and the variational Bayes method with operator eigenvectors (Section 5.2) as the inducing variables.

Corollary 9 prescribes that we take m_n at least

$$(D_{\exp} b_n)^{-1} \log n \approx \sqrt{2a}(n^{1/(1+2\alpha)}/4) \log n \approx 80,$$

where $a = 1/4$ to ensure that $G = \mathcal{N}(0,1)$. Figure 5 illustrates that this is indeed a good choice of m. One can observe that the true and variational posterior are virtually indistinguishable, i.e., there is almost no loss of information in the variational Bayes method.

In Figure 6 we take a smaller number $m = 40$ of inducing points than the optimal $m \approx 80$. One can observe that in this case the variational posterior mean is overly smooth, although still gives a reasonable estimate of f_0. The main difference when considering insufficiently many inducing variables, is that the variational posterior overestimates variance. Here, too, the variational Bayes method provides overly conservative, way too large credible sets compared to the true posterior.

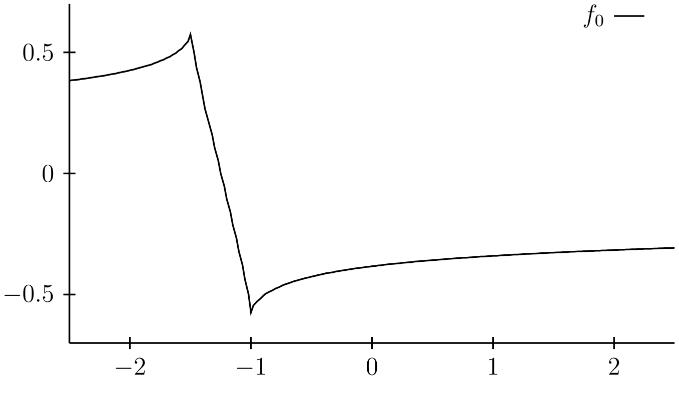

Figure 4: plot of $f_0 = |x+1|^\alpha - |x+3/2|^\alpha$ for $\alpha = 0.8$

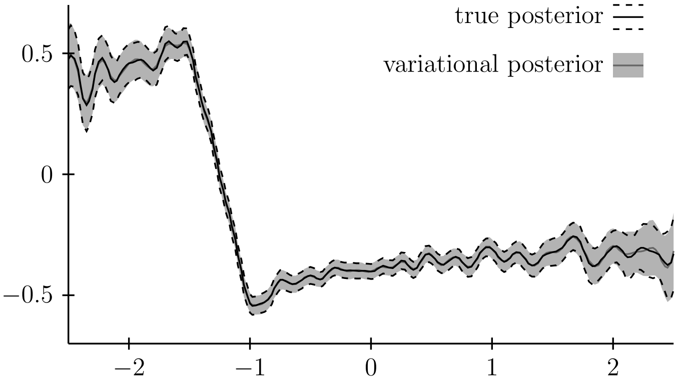

Figure 5: True and variational posterior and credible regions for squared exponential prior and $m = 80$ inducing variables from method 2

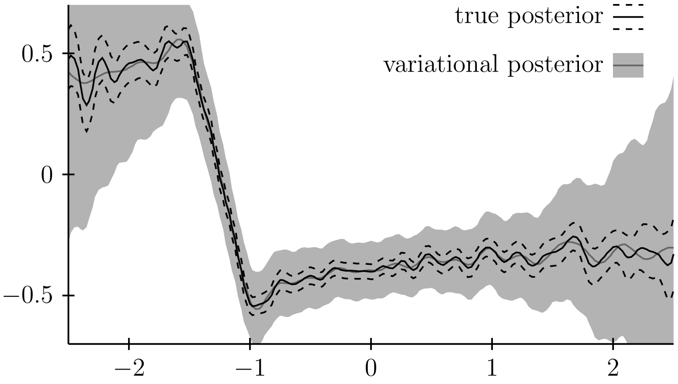

Figure 6: True and variational posterior and credible regions for squared exponential prior and $m = 40$ inducing variables from method 2

8. Conclusion

In this paper we consider the inducing variables variational Bayes method for GP regression and determine sufficient conditions under which the variational approximation achieves the same contraction rate as the original posterior. As examples we consider three commonly used priors and two choices of inducing variables obtained from spectral decompositions and determine a lower bound on the number of inducing variables, which is sufficient for achieving optimal (minimax) contraction rates for the corresponding variational posterior.

The numerical experiments show that credible regions based on the variational posterior are wider than those associated with the true posterior when too few inducing variables are chosen, providing overly conservative uncertainty statements. Nevertheless this suggests that reliable uncertainty quantification should also carry over from the true to the variational posterior, even if the variational approximations are too sparse. If the original credible regions can "capture" the true regression function with P_0-probability tending to one, then so will variational credible regions. A natural next step is to substantiate these experimental results by theory.

Besides the two choices of inducing variables discussed in this paper, inducing point methods that fall within our framework, simply by taking inducing variables of the form $u_j = f(z_j)$ for points $z_j \in \mathcal{X}$. Burt et al. (2020) discuss several methods for selecting the inducing points z_j and obtain bounds on the KL-divergence between the true and variational posterior. It would be interesting to see, by means of an application of Theorem 2, what the minimal number of inducing points has to be in order for these methods to yield optimal contraction rates.

Acknowledgments

We would like to thank the AE and three anonymous reviewers for providing many useful comments that lead to an improved version of the paper.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101041064).

Appendix A. Theory of contraction rates

In this section we provide a brief summary of the frequentist theory of contraction rates for Gaussian Process priors, tailored to our setting. First we start with a general contraction rate result for (nonparametric) posterior distributions. It is a slightly modified version of Theorem 8.9 of Ghosal and van der Vaart (2017) (which also directly follows from their proof), similar to the original statement that appeared in the seminal paper by Ghosal et al. (2000), but simplified and adapted to our setting. It makes use of the so-called covering number (or entropy)

$$N(\epsilon, \mathcal{F}, d_{\mathrm{H}}), \tag{36}$$

which is the minimal number of d_{H}-balls of radius ϵ required to cover the set $\mathcal{F} \subset L^2(\mathcal{X}, G)$.

Lemma 12 *Suppose that there exists a sieve $\mathcal{F} \subset L^2(\mathcal{X}, G)$, a constant $C > 0$, and a sequence of postive numbers ϵ_n with $n\epsilon_n^2 \to \infty$, such that*

$$\Pi(f : \|f - f_0\|_{2,G} < \epsilon_n) \geq \exp(-Cn\epsilon_n^2), \tag{37}$$

$$\log N(\epsilon_n, \mathcal{F}, d_{\mathrm{H}}) \lesssim n\epsilon_n^2, \tag{38}$$

$$\Pi(\mathcal{F}^c) \leq \exp(-(C+4)n\epsilon_n^2). \tag{39}$$

Then there exists an event A_n such that $\mathrm{P}_0(A_n) \to 1$, and

$$\mathrm{E}_0\, \Pi(f : d_{\mathrm{H}}(f, f_0) \geq M_n\epsilon_n \mid \boldsymbol{x}, \boldsymbol{y}) 1_{A_n} \lesssim \exp(-C_2 n\epsilon_n^2)$$

holds for some $C_2 > 0$, and, consequently, the posterior distribution contracts at the rate ϵ_n.

The above lemma can be summarised as follows: the posterior contraction rate at f_0 is ϵ_n if the prior puts sufficient mass on ϵ_n-balls around f_0, and the parameter space can be divided into two sets, of which one has log-entropy of order $n\epsilon_n^2$, and the other attains exponentially small prior mass.

The original statement of this result differs in two ways from ours. Firstly, the original condition (37) uses KL-divergence and KL-variation instead of the L^2-norm $\|\cdot\|_{2,G}$. In our case the statements are equivalent since these are of the same order (see Lemma 2.7 in Ghosal and van der Vaart, 2017). Secondly, the original theorem includes a testing condition, which holds in our case due to the use of the Hellinger distance. For details we refer to Appendix D of Ghosal and van der Vaart (2017).

For GP priors the contraction rate can be characterised by the concentration function inequality (10), since it replaces the conditions of Lemma 12. Indeed, suppose that (10) holds for some $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$. Theorem 2.1 in van der Vaart and van Zanten (2008b) applied to the Banach space $L^2(\mathcal{X}, G)$ with norm $\|\cdot\|_{2,G}$ yields a sieve $\mathcal{F}$ such that (37) and (39) hold, and moreover,

$$\log N(\epsilon_n, \mathcal{F}, \|\cdot\|_{2,G}) \lesssim n\epsilon_n^2. \tag{40}$$

This means there is a bound for a covering number with respect to a different metric. But the elementary inequality $1 - e^{-u} \leq u$ applied to (7) shows that the Hellinger distance d_{H} is bounded by the L^2-norm $\|\cdot\|_{2,G}$ up to a multiplicative constant, so the covering number in condition (38) is bounded by a constant multiplied by the covering number in (40). This means all conditions of Lemma 12 are satisfied and so Lemma 1 is proved.

To connect the contraction rates of the true and variational posterior in the proof of Theorem 2, we use the following result, which is Theorem 5 of Ray and Szabo (2021).

Lemma 13 *Let $\mathcal{F}_n$ be a measurable subset of the parameter space $L^2(\mathcal{X}, G)$, A_n be an event, and Q a distribution for f. If there exist $C, \delta_n > 0$ such that*

$$\mathrm{E}_0\, \Pi(\mathcal{F}_n \mid \boldsymbol{x}, \boldsymbol{y}) 1_{A_n} \leq C e^{-\delta_n},$$

then

$$\mathrm{E}_0\, Q(\mathcal{F}_n) 1_{A_n} \leq \frac{2}{\delta_n}\Big(\mathrm{E}_0\, D_{\mathrm{KL}}(Q \,\|\, \Pi(\,\cdot \mid \boldsymbol{x}, \boldsymbol{y})) 1_{A_n} + C e^{-\delta_n/2} \Big).$$

Although this theorem was applied in context of a high-dimensional parameter space in Ray and Szabo (2021), the result holds for general (possibly infinite-dimensional) parameter spaces, hence can be applied in our setting as well.

Appendix B. The concentration function inequality for the squared exponential prior

We provide the lemmas used in the proof of Lemma 8. The first lemma deals with the $L^2(\mathcal{X}, G)$-entropy of the unit ball $\mathbb{H}_1^b$ of the RKHS of the squared exponential process with length scale b. Recall that N is defined in (36).

Lemma 14 *Let f be the squared exponential process with covariance function* (32) *and suppose that G satisfies the sub-Gaussian tail bound* (33). *There exist a constant $K > 0$ such that for all small enough $\epsilon > 0$, the logarithm of the covering number satisfies*

$$\log N(\epsilon, \mathbb{H}_1^b, \|\cdot\|_{2,G}) \leq K b^{-d} \Big(\log \frac{1}{\epsilon} \Big)^{\kappa},$$

where $\kappa = 1 + 3d/2$.

Proof Let $\mu^b(d\lambda) = (2\pi^{1/2}/b)^{-d} \exp(-\|b\lambda\|^2/4)\, d\lambda$ be the spectral measure of the process f. By Lemma 4.1 of van der Vaart and van Zanten (2009) the RKHS of the process is the collection $\mathbb{H}^b$ of (real parts of) all functions of the form

$$h_\psi(x) = \int e^{i\langle\lambda,x\rangle}\psi(\lambda)\, \mu^b(d\lambda),$$

with $\psi \in L^2(\mu^b)$, and $\|h_\psi\|_{\mathbb{H}^b} = \|\psi\|_{L^2(\mu^b)}$. By Cauchy-Schwarz and the fact that $\mu^b(B) = \mu^1(bB)$ all functions in the RKHS unit ball $\mathbb{H}_1^b$ are uniformly bounded by $C = \sqrt{\mu^1(\mathbb{R}^d)}$. It follows that for $h_1, h_2 \in \mathbb{H}_1^b$ and $a > 0$ we have

$$\|h_1 - h_2\|_{2,G} \le \sup_{\|x\|\le a} |h_1(x) - h_2(x)| + \sqrt{2}C\sqrt{G(x : \|x\| > a)}.$$

The assumption (33) of sub-Gaussian tails implies that for a a large enough multiple of $\sqrt{\log(1/\epsilon)}$ we have $G(x : \|x\| > a) \le \epsilon^2/(2C^2)$, so that

$$\log N\Big(2\epsilon, \mathbb{H}_1^b, L^2(\mathcal{X}, G)\Big) \le \log N(\epsilon, \mathbb{H}_1^b, \ell^\infty[-a, a]).$$

By Lemma 4.5 of van der Vaart and van Zanten (2009) the entropy on the right is bounded by a constant times $(a/b)^d(\log(1/\epsilon))^{1+d}$. ■

Using the well-known connection between the entropy of the RKHS unit ball and the small ball probabilities of a centered Gaussian process as in Lemma 4.6 of van der Vaart and van Zanten (2009), we obtain the following small ball estimate from Lemma 14.

Lemma 15 *Let f be the squared exponential process with covariance function* (32) *and suppose that G satisfies the sub-Gaussian tail bound* (33). *There exist a constant $K > 0$ such that for all small enough $\epsilon > 0$*

$$-\log \Pi(f : \|f\|_{2,G} \le \epsilon) \le K b^{-d}\big(-\log(b\epsilon)\big)^\kappa,$$

where $\kappa = 1 + 3d/2$.

The following lemma deals with the approximation term in the concentration function (9). It follows from the proof of Lemma 4.3 of van der Vaart and van Zanten (2009).

Lemma 16 *Let f be the squared exponential process with covariance function* (32) *and suppose that G satisfies the sub-Gaussian tail bound* (33). *Let $\mathbb{H}^b$ be the RKHS of f. If $f_0 \in L^2(\mathbb{R}^d) \cap C^\alpha(\mathbb{R}^d)$ for $\alpha > 0$, then there exist constants $K_1, K_2 > 0$ such that*

$$\inf_{h\in\mathbb{H}^b : \|h-f_0\|_{2,G}\le K_1 b^\alpha} \|h\|^2_{\mathbb{H}^b} \le K_2 b^{-d}$$

for all $b > 0$ small enough.

References

David R. Burt, Carl Edward Rasmussen, and Mark van der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *International Conference on Machine Learning*, pages 862–871. PMLR, 2019.

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

Sylvain Corlay and Gilles Pagès. Functional quantization-based stratified sampling methods. *Monte Carlo Methods and Applications*, 21(1):1–32, 2015.

Thomas Dunker, Mikhail Lifshits, and Werner Linde. Small deviation probabilities of sums of independent random variables. In *High dimensional probability*, pages 59–74. Springer, 1998.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Subhashis Ghosal, Jayanta K Ghosh, and Aad van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

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

Hermann König. *Eigenvalue distribution of compact operators*. Birkhäuser, 1986.

Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. 1950.

Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When Gaussian process meets big data: A review of scalable GPs. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423, 2020.

Alexander G de G Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Artificial Intelligence and Statistics*, pages 231–239. PMLR, 2016.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT press, 2006.

Kolyan Ray and Botond Szabo. Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, pages 1–12, 2021.

Judith Rousseau and Botond Szabo. Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *The Annals of Statistics*, 45(2): 833–865, 2017.

Matthias Seeger. Addendum to: Information consistency of nonparametric Gaussian process methods. *Max Planck Institute for Biological Cybernetics, Tübingen, Germany, Tech. Rep*, 2007.

John Shawe-Taylor and Christopher KI Williams. The stability of kernel principal components analysis and its relation to the process eigenspectrum. *Advances in neural information processing systems*, pages 383–390, 2002.

Suzanne Sniekers and Aad van der Vaart. Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electronic Journal of Statistics*, 9(2):2475–2527, 2015.

Michalis Titsias. Variational model selection for sparse Gaussian process regression. *Report, University of Manchester, UK*, 2009a.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009b.

Aad van der Vaart and Harry van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, page 200–222, 2008a.

Aad van der Vaart and Harry van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008b.

Aad van der Vaart and Harry van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, pages 2655–2675, 2009.

Aad van der Vaart and Harry van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12(6), 2011.