Journal of Machine Learning Research 12 (2011) 2095-2119 Submitted 12/10; Revised 5/11; Published 6/11

Information Rates of Nonparametric Gaussian Process Methods

Aad van der Vaart AAD@FEW.VU.NL
Department of Mathematics
VU University Amsterdam
De Boelelaan 1081
1081 HV Amsterdam
The Netherlands

Harry van Zanten J.H.V.ZANTEN@TUE.NL
Department of Mathematics
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands

Editor: Manfred Opper

Abstract

We consider the quality of learning a response function by a nonparametric Bayesian approach using a Gaussian process (GP) prior on the response function. We upper bound the quadratic risk of the learning procedure, which in turn is an upper bound on the Kullback-Leibler information between the predictive and true data distribution. The upper bound is expressed in small ball probabilities and concentration measures of the GP prior. We illustrate the computation of the upper bound for the Matérn and squared exponential kernels. For these priors the risk, and hence the information criterion, tends to zero for all continuous response functions. However, the rate at which this happens depends on the combination of true response function and Gaussian prior, and is expressible in a certain concentration function. In particular, the results show that for good performance, the regularity of the GP prior should match the regularity of the unknown response function.

Keywords: Bayesian learning, Gaussian prior, information rate, risk, Matérn kernel, squared exponential kernel

1. Introduction

In this introductory section we first recall some important concepts from Gaussian process regression and then outline our main contributions.

1.1 Gaussian Process Regression

Gaussian processes (GP's) have become popular tools for making inference about unknown functions. They are widely used as prior distributions in nonparametric Bayesian learning to predict a response $Y \in \mathcal{Y}$ from a covariate $X \in \mathcal{X}$. In this approach (cf. Rasmussen and Williams, 2006) a response function $f: \mathcal{X} \to \mathcal{Y}$ is "a-priori" modelled by the sample path of a Gaussian process. This means that for every finite set of points x_j in $\mathcal{X}$, the prior distribution of the vector $(f(x_1), \ldots, f(x_n))$

is multivariate Gaussian. As Gaussian distributions are completely parameterized by their mean and covariance matrix, a GP is completely determined by its *mean function* $m\colon \mathcal{X} \to \mathbb{R}$ and *covariance kernel* $K\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defined as

$$m(x) = \mathrm{E} f(x), \qquad K(x_1, x_2) = \mathrm{cov}\big(f(x_1), f(x_2)\big).$$

The mean function can be any function; the covariance function can be any symmetric, positive semi-definite function. Popular choices are the squared-exponential and Matérn kernels (see Rasmussen and Williams, 2006), or (multiply) integrated Brownian motions (e.g., Wahba, 1978; Van der Vaart and Van Zanten, 2008a). The first two choices are examples of *stationary* GP: the corresponding covariance function has the form $K(x_1, x_2) = K_0(x_1 - x_2)$, for some function K_0 of one argument and hence the distribution of the random function $x \mapsto f(x)$ remains the same under shifting its argument. By Bochner's theorem the stationary covariance functions on $\mathcal{X} = \mathbb{R}^d$ correspond one-to-one to *spectral distributions* (see below for the examples of the squared-exponential and Matérn kernels, or see Rasmussen and Williams, 2006).

In Gaussian process learning the regression function f is modeled as a GP and conditionally on f, observed training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are viewed as independent pairs that satisfy $Y_i = f(X_i) + \varepsilon_i$, for noise variables ε_i. If g denotes the marginal density of the covariates X_i and for $\mu \in \mathbb{R}$, p_μ denotes the density of $\mu + \varepsilon_i$, then conditional on the GP f the pairs (X_i, Y_i) are independently generated according to the probability density $(x, y) \mapsto p_{f(x)}(y) g(x)$. If the errors are normal with mean 0 and variance σ^2 for instance, we have $p_\mu(y) = (2\pi\sigma^2)^{-1/2} \exp(-(y-\mu)^2/(2\sigma^2))$. By Bayes' rule, the posterior distribution for f given the training data is then given by

$$d\Pi_n(f | X_{1:n}, Y_{1:n}) \propto \prod_{i=1}^n p_{f(X_i)}(Y_i)\, d\Pi(f),$$

where $d\Pi(f)$ refers to the prior distribution, and $Z_{1:n}$ is short for the sequence $Z_1, \ldots, Z_n$. After computation (see for instance Rasmussen and Williams, 2006 for methodology), the posterior distribution may be used to predict new responses from covariate values.

1.2 Quantifying Performance

A common approach to assessing the performance of nonparametric Bayes methods is to assume that the data are in actual fact generated according to a fixed, "true" regression function f_0 and to study how well the posterior distribution, which is a distribution over functions, approximates the target f_0 as the number of training data n tends to infinity.

The distance of the posterior to the truth can be measured in various ways. Seeger et al. (2008) discussed the performance of this method in terms of an information criterion due to Barron (1999). They consider the quantity

$$\mathrm{E}_{f_0} \frac{1}{n} \sum_{i=1}^n KL\Big(p_{f_0(X_i)}, \int p_{f(X_i)}\, d\Pi_{i-1}(f | X_{1:i-1}, Y_{1:i-1})\Big). \tag{1}$$

Here $KL(p, q) = \int \log(p/q)\, dP$ denotes the Kullback-Leibler divergence between two probability densities p and q, so that the terms of the sum are the Kullback-Leibler divergences between the density $y \mapsto p_{f_0(X_i)}(y)$ and the *Bayesian predictive density* $y \mapsto \int p_{f(X_i)}(y)\, d\Pi_{i-1}(f | X_{1:(i-1)}, Y_{1:i-1})$ based on the first $(i-1)$ observations, both evaluated for fixed covariate X_i. The expectation E_{f_0}

on the far left is relative to the distribution of $(X_1,Y_1),\ldots,(X_n,Y_n)$. Seeger et al. (2008) obtain a bound on the information criterion (1), which allows them to show for several combinations of true regression functions f_0 and GP priors Π that this tends to zero at a certain rate in the number of observations n.

The information criterion (1) is the Cesàro average of the sequence of *prediction errors*, for $n = 1,2,\ldots$,

$$\mathrm{E}_{f_0} KL\Big(p_{f_0(X_{n+1})}, \int p_{f(X_{n+1})}\, d\Pi_n(f|X_{1:n},Y_{1:n})\Big).$$

By concavity of the logarithm and Jensen's inequality (or the convexity of KL in its second argument), these are bounded above by the *risks*

$$\mathrm{E}_{f_0} \int KL\big(p_{f_0(X_{n+1})}, p_{f(X_{n+1})}\big)\, d\Pi_n(f|X_{1:n},Y_{1:n}). \tag{2}$$

The KL divergence between two normal densities with means μ_1 and μ_2 and common variance σ^2 is equal to $(\mu_1-\mu_2)^2/(2\sigma^2)$. Therefore, in the case of normal errors, with p_f the density of the normal distribution with mean f and variance σ^2, the risks reduce to

$$\frac{1}{2\sigma^2}\mathrm{E}_{f_0} \int \|f_0 - f\|_2^2\, d\Pi_n(f|X_{1:n},Y_{1:n}), \tag{3}$$

where $\|\cdot\|_2$ is the L_2-norm relative to the distribution of the covariate X_{n+1}, that is, $\|f\|_2^2 = \int f^2(x)g(x)\,dx$, and σ^2 is the error variance.

Barron (1999) suggested to use the information criterion (1) as a discrepancy measure, because the risks (2) sometimes behave erratically. However, the risks measure the concentration of the full posterior (both location and spread) near the truth, whereas the prediction errors concern the location of the posterior only. Furthermore, taking Cesàro averages may blur discrepancies in the individual prediction errors. We will show that the present situation is in fact *not* one where the risk (2) behaves badly, and this bigger quantity can be bounded instead of the information criterion (1).

If the risk (3) is bounded by ε_n^2 for some sequence $\varepsilon_n \to 0$, then by another application of Jensen's inequality the posterior mean $\mathrm{E}(f|X_{1:n},Y_{1:n}) = \int f\, d\Pi_n(f|X_{1:n},Y_{1:n})$ satisfies

$$\mathrm{E}_{f_0}\big\|\mathrm{E}(f|X_{1:n},Y_{1:n}) - f_0\big\|_2^2 \le \varepsilon_n^2. \tag{4}$$

Thus the posterior distribution induces a "point estimator" that approximates f_0 at the rate same ε_n. It follows that a bound ε_n^2 on the posterior risk (3) must satisfy the same fundamental lower bound as the (quadratic) risk of general nonparametric estimators for the regression function f_0. Such bounds are usually formulated as *minimax* results: for a given point estimator (for example the posterior mean) one takes the maximum (quadratic) risk over all f_0 in a given "a-priori class" of response functions, and shows that this cannot be smaller than some lower bound (see, e.g., Tsybakov, 2009 for a general introduction to this approach). Typical a-priori classes in nonparametric learning are spaces of "smooth" functions. Several variations exist in the precise definition of such spaces, but they have in common a positive parameter β, which measures the extent of the smoothness or "regularity"; this is roughly the number of times that the functions f_0 are differentiable. It is known that if f_0 is defined on a compact subset of $\mathbb{R}^d$ and has regularity $\beta > 0$, then the optimal, minimax rate ε_n is given by (see, e.g., Tsybakov, 2009)

$$\varepsilon_n = n^{-\beta/(2\beta+d)}. \tag{5}$$

It follows that this is also the best possible bound for the risk (3) if f_0 is a β-regular function of d variables. Recent findings in the statistics literature show that for GP priors, it is typically true that this optimal rate can only be attained if the regularity of the GP that is used matches the regularity of f_0 (see Van der Vaart and Van Zanten, 2008a). Using a GP prior that is too rough or too smooth deteriorates the performance of the procedure. Plain consistency however, that is, the existence of *some* sequence ε_n for which (4) holds, typically obtains for *any* f_0 in the support in the prior.

Seeger et al. (2008) considered the asymptotic performance for the Matérn and squared exponential GP priors, but we will argue in the next subsection that using their approach it is not possible to exhibit the interesting facts that optimal rates are obtained by matching regularities and that consistency holds for any f_0 in the support of the prior. In this paper we will derive these results by following a different approach, along the lines of Ghosal et al. (2000) and Van der Vaart and Van Zanten (2008a).

1.3 Role of the RKHS

A key issue is the fact that Seeger et al. (2008) require the true response function f_0 to be in the reproducing kernel Hilbert space (RKHS) of the GP prior. The RKHS of a GP prior with zero mean function and with covariance kernel K can be constructed by first defining the space $\mathbb{H}_0$ consisting of all functions of the form $x \mapsto \sum_{j=1}^k c_i K(x, y_i)$. Next, the inner product between two functions in $\mathbb{H}_0$ is defined by

$$\big\langle \sum c_i K(\cdot, y_i), \sum c'_j K(\cdot, y'_j) \big\rangle_{\mathbb{H}} = \sum\sum c_i c'_j K(y_i, y'_j),$$

and the associated RKHS-norm by $\|h\|_{\mathbb{H}}^2 = \langle h, h\rangle_{\mathbb{H}}$. Finally, the RKHS $\mathbb{H}$ is defined as the closure of $\mathbb{H}_0$ relative to this norm. Since for all $h \in \mathbb{H}_0$ we have the *reproducing formula*

$$h(x) = \langle h, K(x, \cdot)\rangle_{\mathbb{H}},$$

the RKHS is (or, more precisely, can be identified with) a space of functions on $\mathcal{X}$ and the reproducing formula holds in fact for all $h \in \mathbb{H}$. (For more details, see, e.g., the paper Van der Vaart and Van Zanten, 2008b, which reviews theory on RKHSs that is relevant for Bayesian learning.)

The assumption that $f_0 \in \mathbb{H}$ is very limiting in most cases. The point is that unless the GP prior is a finite-dimensional Gaussian, the RKHS is very small relative to the support of the prior. In the infinite-dimensional case that we are considering here the probability that a draw f from the prior belongs to $\mathbb{H}$ is 0. The reason is that typically, the elements of $\mathbb{H}$ are "smoother" than the draws from the prior. On the other hand, the probability of a draw f falling in a neighbourhood of a given continuous f_0 is typically positive, no matter how small the neighbourhood. (A neighbourhood of f_0 could for instance be defined by all functions with $|f(x) - f_0(x)| < \varepsilon$ for all x, and a given $\varepsilon > 0$.) This means that prior draws can approximate any given continuous function arbitrarily closely, suggesting that the posterior distribution should be able to learn any such function f_0, not just the functions in the RKHS.

Example 1 (Integrated Brownian motion and Matérn kernels) *It is well known that the sample paths $x \mapsto f(x)$ of Brownian motion f have regularity $1/2$. More precisely, for all $\alpha \in (0, 1/2)$ they are almost surely Hölder continuous with exponent α: $\sup_{0\le x<y\le 1} |f(x) - f(y)|/|x-y|^\alpha$ is finite or infinite with probability one depending on whether $\alpha < 1/2$ or $\alpha \ge 1/2$ (see, e.g., Karatzas and Shreve, 1991). Another classical fact is that the RKHS of Brownian motion is the so-called Cameron-Martin space, which consists of functions that have a square integrable derivative (see,*

e.g., Lifshits, 1995). Hence, the functions in the RKHS have regularity 1. *More generally, it can be shown that draws from a k times integrated Brownian motion have regularity* $k+1/2$*, while elements from its RKHS have regularity* $k+1$ *(cf., e.g., Van der Vaart and Van Zanten, 2008b). Analogous statements hold for the Matérn kernel, see Section 3.1 ahead. All these priors can approximate a continuous function* f_0 *arbitrarily closely on any compact domain: the probability that* $\sup_x |f(x)-f_0(x)| < \varepsilon$ *is positive for any* $\varepsilon > 0$.

We show in this paper that if the true response function f_0 on a compact $\mathcal{X} \subset \mathbb{R}^d$ has regularity β, then for the Matérn kernel with smoothness parameter α the (square) risk (3) decays at the rate $n^{-2\min(\alpha,\beta)/(2\alpha+d)}$. This rate is identical to the optimal rate (5) if and only if $\alpha = \beta$. Because the RKHS of the Matérn (α) prior consists of functions of regularity $\alpha+1/2$, it contains functions of regularity β only if $\beta \geq \alpha+1/2$, and this excludes the case $\alpha = \beta$ that the Matérn prior is optimal. Thus if it is assumed a-priori that f_0 is contained in the RKHS, then optimality of Bayesian learning can never be established.

A second drawback of the assumption that $f_0 \in \mathbb{H}$ is that *consistency* (asymptotically correct learning at *some* rate) can be obtained only for a very small class of functions, relative to the support of the GP prior. For instance, Bayesian learning with a Matérn (α) prior is consistent for any continuous true function f_0, not only for f_0 of regularity $\alpha+1/2$ or higher. For the square-exponential process restricting to $f_0 \in \mathbb{H}$ is even more misleading.

Example 2 (Squared exponential kernel) *For the squared exponential GP on a compact subset of* $\mathbb{R}^d$*, every function h in the RKHS has a Fourier transform* $\hat{h}$ *that satisfies*

$$\int |\hat{h}(\lambda)|^2 e^{c\|\lambda\|^2}\, d\lambda < \infty$$

for some $c > 0$ *(see Van der Vaart and Van Zanten, 2009 and Section 3.2 ahead). In particular, every* $h \in \mathbb{H}$ *can be extended to an analytic (i.e., infinitely often differentiable) function on* $\mathbb{C}^d$.

Hence for the squared exponential kernel, restricting to $f_0 \in \mathbb{H}$ only proves consistency for certain analytic regression functions. However, the support of the process is equal to the space of all continuous functions, and consistency pertains for every continuous regression function f_0.

A third drawback of the restriction to $f_0 \in \mathbb{H}$ is that this is the best possible case for the prior, thus giving an inflated idea of its performance. For instance, the squared exponential process gives very fast learning rates for response functions in its RKHS, but as this is a tiny set of analytic functions, this gives a misleading idea of its performance in genuinely nonparametric situations.

1.4 Contributions

In this paper we present a number of contributions to the study of the performance of GP methods for regression.

Firstly, our results give bounds for the risk (2) instead of the information criterion (1). As argued in Section 1.2 the resulting bounds are stronger.

Secondly, our results are not just valid for functions f_0 in the RKHS of the GP prior, but for all functions in the support of the prior. As explained in the preceding section, this is a crucial difference. It shows that in GP regression we typically have plain consistency for all f_0 in the support of the prior and it allows us to study how the performance depends on the relation between

the regularities of the regression function f_0 and typical draws from the prior. We illustrate the general results for the Matérn and squared exponential priors. We present new rate-optimality results for these priors.

A third contribution is that although the concrete GP examples that we consider (Matérn and squared exponential) are stationary, our general results are not limited to stationary processes. The results of Seeger et al. (2008) do concern stationary process and use eigenvalue expansions of the covariance kernels. Underlying our approach are the so-called small deviations behaviour of the Gaussian prior and entropy calculations, following the same basic approach as in our earlier work (Van der Vaart and Van Zanten, 2008a). This allows more flexibility than eigenvalue expansions, which are rarely available and dependent on the covariate distribution. In our approach both stationary and nonstationary prior processes can be considered and it is not necessary to assume a particular relationship between the distribution of the covariates and the prior.

Last but not least, the particular cases of the Matérn and squared exponential kernels that we investigate illustrate that the performance of Bayesian learning methods using GP priors is very sensitive to the fine properties of the priors used. In particular, the relation between the regularity of the response function and the GP used is crucial. Optimal performance is only guaranteed if the regularity of the prior matches the regularity of the unknown function of interest. Serious mismatch leads to (very) slow learning rates. For instance, we show that using the squared-exponential prior, in a situation where a Matérn prior would be appropriate, slows the learning rate from polynomial to logarithmic in n.

1.5 Notations and Definitions

In this section we introduce notation that is used throughout the paper.

1.5.1 SPACES OF SMOOTH FUNCTIONS

As noted in Section 1.2 it is typical to quantify the performance of nonparametric learning procedures relative to a-priori models of smooth functions. The proper definition of "smoothness" or "regularity" depends on the specific situation, but roughly speaking, saying that a function has regularity α means it has α derivatives. In this paper we use two classical notions of finite smoothness: Hölder and Sobolev regularity; and also a scale of infinite smoothness.

For $\alpha > 0$, write $\alpha = m + \eta$, for $\eta \in (0,1]$ and m a nonnegative integer. The *Hölder space* $C^\alpha[0,1]^d$ is the space of all functions whose partial derivatives of orders $(k_1,\ldots,k_d)$ exist for all nonnegative integers $k_1,\ldots,k_d$ such that $k_1+\ldots+k_d \le m$ and for which the highest order partial derivatives are Lipshitz functions of order η. (A function f is *Lipschitz* of order η if $|f(x)-f(y)| \le C|x-y|^\eta$, for every x,y; see for instance Van der Vaart and Wellner (1996), Section 2.7.1, for further details on Hölder classes.)

The *Sobolev space* $H^\alpha[0,1]^d$ is the set of functions $f_0\colon [0,1]^d \to \mathbb{R}$ that are restrictions of a function $f_0\colon \mathbb{R}^d \to \mathbb{R}$ with Fourier transform $\hat f_0(\lambda) = (2\pi)^{-d}\int e^{i\lambda^T t} f(t)\,dt$ such that

$$\|f_0\|_{\alpha|2}^2 := \int \left(1+\|\lambda\|^2\right)^\alpha \left|\hat f_0(\lambda)\right|^2 d\lambda < \infty.$$

Roughly speaking, for integer α, a function belongs to H^α if it has partial derivatives up to order α that are all square integrable. This follows, because the αth derivative of a function f_0 has Fourier transform $\lambda \mapsto (i\lambda)^\alpha \hat f_0(\lambda)$,

Qualitatively both spaces $H^\alpha[0,1]^d$ and $C^\alpha[0,1]^d$ describe "α-regular" functions. Technically their definitions are different, and so are the resulting sets. There are however many functions in the intersection $H^\alpha[0,1]^d \cap C^\alpha[0,1]^d$ and these are α-regular in both senses at the same time.

We also consider functions that are "infinitely smooth". For $r \geq 1$ and $\lambda > 0$, we define the space $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$ of functions $f_0 : \mathbb{R}^d \to \mathbb{R}$ with Fourier transform $\hat{f}_0$ satisfying

$$\|f_0\|_{\mathcal{A}}^2 := \int e^{\gamma\|\lambda\|^r} |\hat{f}_0|^2(\lambda)\, d\lambda < \infty.$$

This requires exponential decrease of the Fourier transform, in contrast to polynomial decrease for Sobolev smooothness. The functions in $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$ are infinitely often differentiable and "increasingly smooth" as γ or r increase. They extend to functions that are analytic on a strip in $\mathbb{C}^d$ containing $\mathbb{R}^d$ if $r = 1$, and to entire functions if $r > 1$ (see, e.g., Bauer, 2001, 8.3.5).

1.5.2 GENERAL FUNCTION SPACES AND NORMS

For a general metric space $\mathcal{X}$ we denote by $C_b(\mathcal{X})$ the space of bounded, continuous functions on $\mathcal{X}$. If the space $\mathcal{X}$ is compact, for example, $\mathcal{X} = [0,1]^d$, we simply write $C(\mathcal{X})$. The supremum norm of a bounded function f on $\mathcal{X}$ is denoted by $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$.

For $x_1, \ldots, x_n \in \mathcal{X}$ and a function $f : \mathcal{X} \to \mathbb{R}$ we define the empirical norm $\|f\|_n$ by

$$\|f\|_n = \Big(\frac{1}{n}\sum_{i=1}^n f^2(x_i)\Big)^{1/2}. \tag{6}$$

For m a (Borel) measure on $A \subset \mathbb{R}^d$ we denote by $L_2(m)$ the associated L^2-space, defined by

$$L_2(m) = \Big\{ f : A \to \mathbb{R} \,\Big|\, \int_A |f(x)|^2\, dm(x) < \infty \Big\}.$$

In a regression setting where the covariates have probability density g on $\mathbb{R}^d$, we denote the corresponding L_2-norm simply by $\|f\|_2$, that is,

$$\|f\|_2 = \int f^2(x) g(x)\, dx.$$

1.5.3 MISCELLANEOUS

The notation $a \lesssim b$ means that $a \leq Cb$ for a universal constant C. We write $a \vee b = \max\{a,b\}$, $a \wedge b = \min\{a,b\}$.

2. General Results

In this section we present general bounds on the posterior risk. The next section treats the special cases of the Matérn and squared exponential kernels. Proofs are deferred to Section 4.

2.1 Fixed Design

In this section we assume that given the function $f : \mathcal{X} \to \mathbb{R}$, the data $Y_1, \ldots, Y_n$ are independently generated according to $Y_j = f(x_j) + \varepsilon_j$, for fixed $x_j \in \mathcal{X}$ and independent $\varepsilon_j \sim N(0, \sigma^2)$. Such a fixed design setting occurs when the covariate values in the training data have been set by an experimenter.

For simplicity we assume that $\mathcal{X}$ is a compact metric space, such as a bounded, closed set in $\mathbb{R}^d$, and assume that the true response function f_0 and the support of the GP prior are included in the space $C_b(\mathcal{X})$ of bounded, continuous functions on the metric space $\mathcal{X}$. This enables to formulate the conditions in terms of the *supremum norm* (also called "uniform" norm). Recall that the supremum norm of $f \in C_b(\mathcal{X})$ is given by $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. (Actually Theorem 1 refers to the functions on the design points only and is in terms of the norm (6). The conditions could be formulated in terms of this norm. This would give a stronger result, but its interpretation is hampered by the fact that the norm (6) changes with n.) The RKHS of the GP prior, as defined in Section 1.3, is denoted by $\mathbb{H}$ and the RKHS-norm by $\|\cdot\|_{\mathbb{H}}$.

The following theorem gives an upper bound for the posterior risk. The bound depends on the "true" response function f_0 and the GP prior Π and its RKHS $\mathbb{H}$ through the so-called *concentration function*

$$\phi_{f_0}(\varepsilon) = \inf_{h\in\mathbb{H}:\|h-f_0\|_\infty<\varepsilon} \|h\|_{\mathbb{H}}^2 - \log\Pi\big(f\colon \|f\|_\infty < \varepsilon\big) \tag{7}$$

and the associated function

$$\psi_{f_0}(\varepsilon) = \frac{\phi_{f_0}(\varepsilon)}{\varepsilon^2}. \tag{8}$$

We denote by $\psi_{f_0}^{-1}$ the (generalized) inverse function of the function ψ_{f_0}, that is, $\psi_{f_0}^{-1}(l) = \sup\{\varepsilon > 0\colon \psi_{f_0}(\varepsilon) \geq l\}$.

The concentration function ϕ_{f_0} for a general response function consists of two parts. The second is the small ball exponent $\phi_0(\varepsilon) = -\log\Pi(f\colon \|f\|_\infty < \varepsilon)$, which measures the amount of prior mass in a ball of radius ε around the zero function. As the interest is in small ε this is (the exponent of) the *small ball probability* of the prior. There is a large literature on small ball probabilities of Gaussian distributions. (See Kuelbs and Li, 1993 and Li and Shao, 2001 and references.) This contains both general methods (probabilistic and analytic) for its computation and many examples, stationary and non-stationary. The first part of the definition of $\phi_{f_0}(\varepsilon)$, the infimum, measures the decrease in prior mass if the (small) ball is shifted from the origin to the true parameter f_0. This is not immediately clear from the definition (7), but it can be shown that up to constants, $\phi_{f_0}(\varepsilon)$ equals $-\log\Pi(f\colon \|f-f_0\|_\infty < \varepsilon)$ (see for instance Van der Vaart and Van Zanten, 2008b, Lemma 5.3). The infimum depends on how well f_0 can be approximated by elements h of the RKHS of the prior, and the quality of this approximation is measured by the size of the approximand h in the RKHS-norm. The infimum is finite for every $\varepsilon > 0$ if and only if f_0 is contained in the closure of $\mathbb{H}$ within $C_b(\mathcal{X})$. The latter closure is the support of the prior (Van der Vaart and Van Zanten, 2008b, Lemma 5.1) and in typical examples it is the full space $C_b(\mathcal{X})$.

Our general upper bound for the posterior risk in the fixed design case takes the following form.

Theorem 1 *For $f_0 \in C_b(\mathcal{X})$ it holds that*

$$\mathrm{E}_{f_0} \int \|f-f_0\|_n^2 \, d\Pi_n\big(f|Y_{1:n}\big) \lesssim \psi_{f_0}^{-1}(n)^2.$$

For $\psi_{f_0}^{-1}(n) \to 0$ as $n\to\infty$, which is the typical situation, the theorem shows that the posterior distribution contracts at the rate $\psi_{f_0}^{-1}(n)$ around the true response function f_0. To connect to Seeger et al. (2008), we have expressed the contraction using the quadratic risk, but the concentration is actually exponential. In particular, the power 2 can be replaced by any finite power.

From the definitions one can show that (see Lemma 17), whenever $f_0 \in \mathbb{H}$,

$$\psi_{f_0}^{-1}(n) \lesssim \frac{\|f_0\|_{\mathbb{H}}}{\sqrt{n}} + \psi_0^{-1}(n). \tag{9}$$

This relates the theorem to formula (3) in Seeger et al., whose $\log\det(I+cK)$ is replaced by $\psi_0^{-1}(n)^2$. However, the left side $\psi_{f_0}^{-1}(n)$ of the preceding display is finite for every f_0 in the support of the prior, which is typically a much large space than the RKHS (see Section 1.3). For instance, functions f_0 in the RKHS of the squared exponential process are analytic, whereas $\psi_{f_0}^{-1}(n)$ is finite for every continuous function f_0 in that case. Thus the theorem as stated is much more refined than if its upper bound would be replaced by the right side of (9). It is true that $\psi_{f_0}^{-1}(n)$ is smallest if f_0 belongs to the RKHS, but typically the posterior also contracts if this is not the case.

In Sections 3.1 and 3.2 we show how to obtain bounds for the concentration function, and hence a risk bound, for two classes of specific priors: the Matérn class and the squared exponential. Other examples, including non-stationary ones like (multiply) integrated Brownian motion, were considered in Van der Vaart and Van Zanten (2008a), Van der Vaart and Van Zanten (2007) and Van der Vaart and Van Zanten (2009).

2.2 Random Design

In this section we assume that given the function $f:[0,1]^d \to \mathbb{R}$ on the d-dimensional unit cube $[0,1]^d$ (or another compact, Lipschitz domain in $\mathbb{R}^d$) the data $(X_1,Y_1),\ldots,(X_n,Y_n)$ are independently generated, X_i having a density g on $[0,1]^d$ that is bounded away from zero and infinity, and $Y_j = f(X_j)+\varepsilon_j$, for errors $\varepsilon_j \sim N(0,\sigma^2)$ that are independent given the X_i's.

We assume that under the GP prior Π the function f is a zero-mean, continuous Gaussian process. The concentration function ϕ_{f_0} and the derived function ψ_{f_0} are defined as before in (7) and (8). Recall that $\|f\|_2$ is the L_2-norm relative to the covariate distribution, that is, $\|f\|_2^2 = \int f^2(x)g(x)\,dx$. The theorem assumes that for some $\alpha > 0$, draws from the prior are α-regular in Hölder sense. This roughly means that α derivatives should exist. See Section 1.5 for the precise definition.

Theorem 2 *Suppose that for some $\alpha > 0$ the prior gives probability one to the Hölder space $C^\alpha[0,1]^d$. For $\psi_{f_0}^{-1}$ the inverse function of ψ_{f_0} and C a constant that depends on the prior and the covariate density, if $\psi_{f_0}^{-1}(n) \le n^{-d/(4\alpha+2d)}$, then*

$$\mathrm{E}_{f_0} \int \|f - f_0\|_2^2 \, d\Pi_n\big(f \,|\, X_{1:n}, Y_{1:n}\big) \le C\psi_{f_0}^{-1}(n)^2.$$

If, on the other hand, $\psi_{f_0}^{-1}(n) \ge n^{-d/(4\alpha+2d)}$, then the assertion is true with the upper bound $Cn\psi_{f_0}^{-1}(n)^{(4\alpha+4d)/d}$.

Unlike in the case of fixed design treated in Theorem 1, this theorem makes assumptions on the regularity of the prior. This seems unavoidable, because the $\|\cdot\|_2$-risk extrapolates from the observed design points to all points in the support of the covariate density.

In the next section we shall see that a typical rate for estimating a β-smooth response function f_0 is given by

$$\psi_{f_0}^{-1}(n) \sim n^{-(\beta\wedge\alpha)/(2\alpha+d)}.$$

(This reduces to the minimax rate $n^{-\alpha/(2\alpha+d)}$ if and only if $\alpha=\beta$.) In this case $\psi_{f_0}^{-1}(n) \leq n^{-d/(4\alpha+2d)}$ if and only if $\alpha \wedge \beta \geq d/2$. In other words, upper bounds for fixed and random design have exactly the same form if prior and true response are not too rough.

For very rough priors and true response functions, the rate given by the preceding theorem is slower than the rate for deterministic design, and for very rough response functions the theorem may not give a rate at all. The latter seems partly due to using the second moment of the posterior, rather than posterior concentration, although perhaps the theorem can be improved.

3. Results for Concrete Priors

In this section we specialize to two concrete classes of Gaussian process priors, the Matérn class and the squared exponential process.

3.1 Matérn Priors

In this section we compute the risk bounds given by Theorems 1 and 2 for the case of the Matérn kernel. In particular, we show that optimal rates are attained if the smoothness of the prior matches the smoothness of the unknown response function.

The *Matérn priors* correspond to the mean-zero Gaussian processes $W = (W_t : t \in [0,1]^d)$ with covariance function

$$\mathrm{E} W_s W_t = \int_{\mathbb{R}^d} e^{i\lambda^T(s-t)} m(\lambda)\, d\lambda,$$

defined through the *spectral densities* $m : \mathbb{R}^d \to \mathbb{R}$ given by, for $\alpha > 0$,

$$m(\lambda) = \frac{1}{\big(1+\|\lambda\|^2\big)^{\alpha+d/2}}. \tag{10}$$

The integral can be expressed in certain special functions (see, e.g., Rasmussen and Williams, 2006). This is important for the numerical implementation of the resulting Bayesian procedure, but not useful for our present purpose.

The sample paths of the Matérn process possess the same smoothness in L_2 as the set of functions $e_t(\lambda) = e^{i\lambda^T t}$ in $L_2(m)$. From this it can be seen that the sample paths are k times differentiable in L_2, for k the biggest integer smaller than α, with kth derivative satisfying

$$\mathrm{E}(W_s^{(k)} - W_t^{(k)})^2 \lesssim \|s-t\|^{2(\alpha-k)}.$$

By Kolmogorov's continuity criterion it follows that the sample paths of the kth derivative can be constructed to be Lipshitz of any order strictly smaller than $\alpha - k$. Thus the Matérn process takes its values in $C^{\underline{\alpha}}[0,1]^d$ for any $\underline{\alpha} < \alpha$. Hence in this specific sense it is α-regular.

By Lemma 4.1 of Van der Vaart and Van Zanten (2009) the RKHS $\mathbb{H}$ of the process W is the space of all (real parts of) functions of the form

$$h_\psi(t) = \int e^{i\lambda^T t} \psi(\lambda) m(\lambda)\, d\lambda, \tag{11}$$

for $\psi \in L_2(m)$, and squared RKHS-norm given by

$$\|h_\psi\|_{\mathbb{H}}^2 = \min_{\phi : h_\phi = h_\psi} \int |\phi|^2(\lambda) m(\lambda)\, d\lambda. \tag{12}$$

This characterization is generic for stationary Gaussian processes. The minimum is unnecessary if the spectral density has exponential tails (as in the next section), but is necessary in the present case.

In the following two lemmas we describe the concentration function (7) of the Matérn prior. The small ball probability can be obtained from the preceding characterization of the RKHS, estimates of metric entropy, and general results on Gaussian processes. See Section 4.3 for proofs.

Lemma 3 *For $\|\cdot\|_\infty$ the uniform norm, and C a constant independent of ε,*

$$-\log \mathrm{P}\big(\|W\|_\infty < \varepsilon\big) \le C\Big(\frac{1}{\varepsilon}\Big)^{d/\alpha}.$$

To estimate the infimum in the definition of the concentration function ϕ_{f_0} for a nonzero response function f_0, we approximate f_0 by elements of the RKHS. The idea is to write f_0 in terms of its Fourier inverse $\hat{f}_0$ as

$$\begin{aligned} f_0(x) &= \int e^{i\lambda^T x} \hat{f}_0(\lambda)\, d\lambda \\ &= \int e^{i\lambda^T x} \frac{\hat{f}_0}{m}(\lambda)\, m(\lambda)\, d\lambda. \end{aligned} \tag{13}$$

If $\hat{f}_0/m$ were contained in $L_2(m)$, then f_0 would be contained in the RKHS, with RKHS-norm bounded by the $L_2(m)$-norm of $\hat{f}_0/m$, that is, the square root of $\int (|\hat{f}_0|^2/m)(\lambda)\, d\lambda$. In general this integral may be infinite, but we can remedy this by truncating the tails of $\hat{f}_0/m$. We then obtain an approximation of f_0 by an element of the RKHS, which is enough to compute the concentration function (8).

A natural a-priori condition on the true response function $f_0\colon [0,1]^d \to \mathbb{R}$ is that this function is contained in a Sobolev space of order β. This space consists roughly of functions that possess β square integrable derivatives. The precise definition is given in Section 1.5.

Lemma 4 *If $f_0 \in C^\beta[0,1]^d \cap H^\beta[0,1]^d$ for $\beta \le \alpha$, then, for $\varepsilon < 1$, and a constant C depending on f_0 and α,*

$$\inf_{h:\|h-f_0\|_\infty<\varepsilon} \|h\|_{\mathbb{H}}^2 \le C\Big(\frac{1}{\varepsilon}\Big)^{(2\alpha+d-2\beta)/\beta}.$$

Combination of the two lemmas yields that for $f_0 \in C^\beta[0,1]^d \cap H^\beta[0,1]^d$ for $\beta \le \alpha$, the concentration function (7) satisfies

$$\phi_{f_0}(\varepsilon) \lesssim \Big(\frac{1}{\varepsilon}\Big)^{(2\alpha+d-2\beta)/\beta} + \Big(\frac{1}{\varepsilon}\Big)^{d/\alpha}.$$

This implies that

$$\psi_{f_0}^{-1}(n) \lesssim \Big(\frac{1}{n}\Big)^{\beta/(2\alpha+d)}.$$

Theorems 1 and 2 imply that the rate of contraction of the posterior distribution is of this order in the case of fixed design, and of this order if $\beta > d/2$ in the case of random design. We summarize these findings in the following theorem.

Theorem 5 *Suppose that we use a Matérn prior with parameter $\alpha > 0$ and $f_0 \in C^\beta[0,1]^d \cap H^\beta[0,1]^d$ for $\beta > 0$. Then in the fixed design case the posterior contracts at the rate $n^{-(\alpha\wedge\beta)/(2\alpha+d)}$. In the random design case this holds as well, provided $\alpha \wedge \beta > d/2$.*

Observe that the optimal rate $n^{-\beta/(2\beta+d)}$ is attained if and only if $\alpha = \beta$. Using a prior that is "rougher" or "smoother" than the truth leads to sub-optimal rates. This is in accordance with the findings for other GP priors in in Van der Vaart and Van Zanten (2008a). It should be remarked here that Theorem 5 only gives an upper bound on the rate of contraction. However, the paper by Castillo (2008) shows that these bounds are typically tight.

3.2 Squared Exponential Kernel

In this section we compute the risk bounds given by Theorems 1 and 2 for the case of the squared exponential kernel.

The *squared exponential process* is the zero-mean Gaussian process with covariance function

$$\mathrm{E}W_s W_t = e^{-\|s-t\|^2}, \qquad s,t \in [0,1]^d.$$

Like the Matérn process the squared exponential process is stationary. Its spectral density is given by

$$m(\lambda) = \frac{1}{2^d \pi^{d/2}} e^{-\|\lambda\|^2/4}. \tag{14}$$

The sample paths of the square exponential process are analytic.

This process was studied already in Van der Vaart and Van Zanten (2007) and Van der Vaart and Van Zanten (2009). The first of the following lemmas is Lemma 4.5 in Van der Vaart and Van Zanten (2009). It deals with the second term in the concentration function (7). As before, let $\|\cdot\|_\infty$ be the uniform norm on the functions $f\colon [0,1]^d \to \mathbb{R}$.

Lemma 6 *There exists a constant C depending only on d such that*

$$-\log \mathrm{P}\Big(\|W\|_\infty \le \varepsilon\Big) \le C\Big(\log \frac{1}{\varepsilon}\Big)^{1+d}.$$

The following lemma concerns the infimum part of the concentration function in the case that the function f_0 belongs to a Sobolev space with regularity β (see Section 1.5).

Lemma 7 *If $f_0 \in H^\beta[0,1]^d$ for $\beta > d/2$, then, for a constant C that depends only on f_0,*

$$\inf_{\|h-f_0\|_\infty \le \varepsilon} \|h\|_{\mathbb{H}}^2 \le \exp\big(C\varepsilon^{-2/(\beta-d/2)}\big).$$

Combination of the preceding two lemmas shows that for a β-regular response function f_0 (in Sobolev sense)

$$\phi_{f_0}(\varepsilon) \lesssim \exp\big(C\varepsilon^{-2/(\beta-d/2)}\big) + \Big(\log \frac{1}{\varepsilon}\Big)^{1+d}.$$

The first term on the right dominates, for any $\beta > 0$. The corresponding rate of contraction satisfies

$$\psi_{f_0}^{-1}(n) \lesssim (1/\log n)^{\beta/2-d/4}.$$

Thus the extreme smoothness of the prior relative to the smoothness of the response function leads to very slow contraction rates for such functions. A remedy for this mismatch is to rescale the sample paths. The length scale of the process can be treated as a hyperparameter and can be endowed with a prior of its own, or can be selected using an empirical Bayes procedure. Van der

Vaart and Van Zanten (2007) and Van der Vaart and Van Zanten (2009) for example show that the prior $x \mapsto f(Ax)$, for f the squared exponential process and A^d an independent Gamma distributed random variable, leads to optimal contraction rates for β-smooth true response functions, for any $\beta > 0$.

Actually, the preceding discussion permits only the derivation of an *upper bound* on the contraction rate. In the next theorem we show that the logarithmic rate is real however. The theorem shows that asymptotically, balls around f_0 of logarithmic radius receive zero posterior mass. The proof, following an idea of Castillo (2008) and given in Section 4.4, is based on the fact that balls of this type also receive very little *prior* mass, essentially because the inequality of the preceding lemma can be reversed.

Theorem 8 *If f_0 is contained in $H^\beta[0,1]^d$ for some $\beta > d/2$, has support within $(0,1)^d$ and possesses a Fourier transform satisfying $|\hat{f}_0(\lambda)| \gtrsim \|\lambda\|^{-k}$ for some $k > 0$ and every $\|\lambda\| \geq 1$, then there exists a constant l such that $\mathrm{E}_{f_0}\Pi\big(f : \|f - f_0\|_2 \leq (\log n)^{-l} | X_{1:n}, Y_{1:n}\big) \to 0$.*

As the prior puts all of its mass on analytic functions, perhaps it is not fair to study its performance only for β-regular functions, and it makes sense to study the concentration function also for "supersmooth", analytic response functions as well. The functions in the RKHS of the squared exponential process are examples of supersmooth functions, and for those functions we obtain the rate $\psi_0^{-1}(n)$ determined by the (centered) small ball probability only. In view of Lemma 6 this is a $1/\sqrt{n}$-rate up to a logarithmic factor.

The following lemma deals with the infimum part of the concentration function in the case that that the function f_0 is supersmooth. Recall the definition of the space $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$ of analytic functions given in Section 1.5.

Lemma 9

- *If f_0 is the restriction to $[0,1]^d$ of an element of $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$, for $r > 2$, or for $r \geq 2$ with $\gamma \geq 4$, then $f_0 \in \mathbb{H}$.*
- *If f_0 is the restriction to $[0,1]^d$ of an element of $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$ for $r < 2$, then there exist a constant C depending on f_0 such that*

$$\inf_{\|h-w\|_\infty \leq \varepsilon} \|h\|_{\mathbb{H}}^2 \leq Ce^{\left(\log(1/\varepsilon)\right)^{2/r}/(4\gamma^{2/r})}.$$

Combination of Lemmas 6 and 9 with the general theorems yields the following result.

Theorem 10 *Suppose that we use a squared exponential prior and f_0 is the restriction to $[0,1]^d$ of an element of $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$, for $r \geq 1$ and $\gamma > 0$. Then both in the fixed and the random design cases the posterior contracts at the rate $(\log n)^{1/r}/\sqrt{n}$.*

Observe that the rate that we get in the last theorem is up to a logarithmic factor equal to the rate $1/\sqrt{n}$ at which the posterior typically contracts for parametric models (cf., the Bernstein-von Mises theorem, for example, Van der Vaart, 1998). This "almost parametric rate" is explainable from the fact that spaces of analytic functions are only slightly bigger than finite-dimensional spaces in terms of their metric entropy (see Kolmogorov and Tihomirov, 1961).

Together, Theorems 8 and 10 give the same general message for the squared exponential kernel as Theorem 5 does for the Matérn kernel: fast convergence rates are only attained if the smoothness of the prior matches the smoothness of the response function f_0. However, generally the

assumption of existence of infinitely many derivatives of a true response function ($f_0 \in \mathcal{A}^{g,r}(\mathbb{R}^d)$) is considered too strong to define a test case for nonparametric learning. If this assumption holds, then the response function f_0 can be recovered at a very fast rate, but this is poor evidence of good performance, as only few functions satisfy the assumption. Under the more truly "nonparametric assumption" that f_0 is β-regular, the performance of the squared-exponential prior is disastrous (unless the length scale is changed appropriately in a data-dependent way).

4. Proofs

This section contains the proofs of the presented results.

4.1 Proof of Theorem 1

The proof of Theorem 1 is based on estimates of the prior mass near the true parameter f_0 and on the metric entropy of the support of the prior. This is expressed in the following proposition.

We use the notation $D(\varepsilon, \mathcal{A}, d)$ for the ε-packing number of the metric space $(\mathcal{A}, d)$: the maximal number of points in $\mathcal{A}$ such that every pair has distance at least ε relative to d.

Proposition 11 *Suppose that for some $\varepsilon > 0$ with $\sqrt{n}\varepsilon \geq 1$ and for every $r > 1$ there exists a set $\mathcal{F}_r$ such that*

$$D\big(\varepsilon, \mathcal{F}_r, \|\cdot\|_n\big) \leq e^{n\varepsilon^2 r^2}, \qquad \Pi(\mathcal{F}_r) \geq 1 - e^{-2n\varepsilon^2 r^2}. \tag{15}$$

Furthermore, suppose that

$$\Pi\big(f : \|f - f_0\|_n \leq \varepsilon\big) \geq e^{-n\varepsilon^2}. \tag{16}$$

Then

$$P_{n,f_0} \int \|f - f_0\|_n^l \, d\Pi_n\big(f \,|\, Y_{1:n}\big) \lesssim \varepsilon^l.$$

For $\theta \in \mathbb{R}^n$ let $P_{n,\theta}$ be the normal distribution $N_n(\theta, I)$. In the following three lemmas let $\|\cdot\|$ be the Euclidean norm on $\mathbb{R}^n$.

Lemma 12 *For any $\theta_0, \theta_1 \in \mathbb{R}^n$, there exists a test ϕ based on $Y \sim N_n(\theta, I)$ such that, for every $\theta \in \mathbb{R}^n$ with $\|\theta - \theta_1\| \leq \|\theta_0 - \theta_1\|/2$,*

$$P_{n,\theta_0}\phi \vee P_{n,\theta}(1-\phi) \leq e^{-\|\theta_0 - \theta_1\|^2/8}.$$

Proof For simplicity of notation we can choose $\theta_0 = 0$. If $\|\theta - \theta_1\| \leq \|\theta_1\|/2$, then $\|\theta\| \geq \|\theta_1\|/2$ and hence $\langle \theta, \theta_1 \rangle = \big(\|\theta\|^2 + \|\theta_1\|^2 - \|\theta - \theta_1\|^2\big)/2 \geq \|\theta_1\|^2/2$. Therefore, the test $\phi = 1_{\theta_1^T Y > D\|\theta_1\|}$ satisfies, with Φ the standard normal cdf,

$$\begin{aligned} P_{n,\theta_0}\phi &= 1 - \Phi(D), \\ P_{n,\theta}(1-\phi) &= \Phi\big((D\|\theta_1\| - \langle \theta, \theta_1 \rangle)/\|\theta_1\|\big) \leq \Phi(D - \rho), \end{aligned}$$

for $\rho = \|\theta_1\|/2$. The infimum over D of $\big(1 - \Phi(D)\big) + \Phi(D - \rho)$ is attained for $D = \rho/2$, for which $D - \rho = -\rho/2$. We substitute this in the preceding display and use the bound $1 - \Phi(x) \leq e^{-x^2/2}$,

valid for $x \geq 0$. ■

Let $D(\varepsilon, \Theta)$ be the maximal number of points that can be placed inside the set $\Theta \subset \mathbb{R}^n$ such that any pair has Euclidean distance at least ε.

Lemma 13 *For any $\Theta \subset \mathbb{R}^n$ there exists a test ϕ based on $Y \sim N_n(\theta, I)$ with, for any $r > 1$ and every integer $j \geq 1$,*

$$P_{n,\theta_0}\phi \leq 9D(r/2, \Theta)\exp(-r^2/8),$$
$$\sup_{\theta \in \Theta: \|\theta - \theta_0\| \geq jr} P_{n,\theta}(1-\phi) \leq \exp(-j^2r^2/8).$$

Proof The set Θ can be partitioned into the shells

$$C_{j,r} = \big\{\theta \in \Theta: jr \leq \|\theta - \theta_0\| < (j+1)r\big\}.$$

We place in each of these shells a maximal collection Θ_j of points that are $jr/2$-separated, and next construct a test ϕ_j as the maximum of all the tests as in the preceding lemma attached to one of these points. The number of points is equal to $D(jr/2, C_{j,r})$. Every $\theta \in C_{j,r}$ is in a ball of radius $jr/2$ of some point $\theta_1 \in \Theta_j$ and satisfies $\|\theta - \theta_1\| \leq jr/2 \leq \|\theta_0 - \theta_1\|/2$, since $\theta_1 \in C_{j,r}$. Hence each test satisfies the inequalities of the preceding lemma. It follows that

$$P_{n,\theta_0}\phi_j \leq D(jr/2, C_{j,r})e^{-j^2r^2/8},$$
$$\sup_{\theta \in C_{j,r}} P_{n,\theta}(1-\phi_j) \leq e^{-j^2r^2/8}.$$

Finally, we construct ϕ as the supremum over all tests ϕ_j, for $j \geq 1$. We note that $\sum_{j \geq 1} D(jr/2, C_{j,r})e^{-j^2r^2/8} \leq D(r/2, \Theta)e^{-r^2/8}/(1 - e^{-r^2/8})$, and $1/(1 - e^{-1/8}) \approx 8.510$. ■

Lemma 14 *For any probability distribution Π on $\mathbb{R}^n$ and $x > 0$,*

$$P_{n,\theta_0}\Big(\int \frac{p_{n,\theta}}{p_{n,\theta_0}}\, d\Pi(\theta) \leq e^{-\sigma_0^2/2 - \|\mu_0\|x}\Big) \leq e^{-x^2/2},$$

for $\mu_0 = \int (\theta - \theta_0)\, d\Pi(\theta)$ and $\sigma_0^2 = \int \|\theta - \theta_0\|^2\, d\Pi(\theta)$. Consequently, for any probability distribution Π on $\mathbb{R}^n$ and any $r > 0$,

$$P_{n,\theta_0}\Big(\int \frac{p_{n,\theta}}{p_{n,\theta_0}}\, d\Pi(\theta) \geq e^{-r^2}\Pi\big(\theta: \|\theta - \theta_0\| < r\big)\Big) \geq 1 - e^{-r^2/8}.$$

Proof Under θ_0 the variable $\int \log(p_{n,\theta}/p_{n,\theta_0})\, d\Pi(\theta) = \mu_0^T(Y - \theta_0) - \sigma_0^2/2$ is normally distributed with mean $-\sigma_0^2/2$ and variance $\|\mu_0\|^2$. Therefore, the event B_n that this variable is smaller than $-\sigma_0^2/2 - \|\mu_0\|x$ has probability bounded above by $\Phi(-x) \leq e^{-x^2/2}$. By Jensen's inequality applied to the logarithm, the event in the left side of the lemma is contained in B_n.

To prove the second assertion we first restrict the integral $\int p_{n,\theta}/p_{n,\theta_0}\, d\Pi(\theta)$ to the ball $\{\theta: \|\theta - \theta_0\| \leq r\}$, which makes it smaller. Next we divide by $\Pi\big(\theta: \|\theta - \theta_0\| < r\big)$ to renormalize Π to a

probability measure on this ball, and apply the first assertion with this renormalized measure Π. The relevant characteristics of the renormalized measure satisfy $\|\mu_0\| \le r$ and $\sigma_0^2 \le r^2$. Therefore the assertion follows upon choosing $x = r/2$. ■

Proof [Proof of Proposition 11] For any event $\mathcal{A}$, any test ϕ and any $r > 1$, the expected value $P_{n,f_0}\Pi\big(f\colon \|f - f_0\|_n > 4\varepsilon r | Y_{1:n}\big)$ is bounded by $A+B+C+D$, for

$$\begin{aligned} A &= P_{n,f_0}\phi, \\ B &= P_{n,f_0}(\mathcal{A}^c) \\ C &= P_{n,f_0}\Pi_n\big(f \notin \mathcal{F}_r | Y_{1:n}\big)1_{\mathcal{A}}, \\ D &= P_{n,f_0}\Pi_n\big(f \in \mathcal{F}_r\colon \|f - f_0\|_n > 4\varepsilon r | Y_{1:n}\big)(1-\phi)1_{\mathcal{A}}. \end{aligned}$$

For the test ϕ given by Lemma 13 with Θ the set of all vectors $\big(f(x_1), \ldots, f(x_n)\big)$ as f ranges over $\mathcal{F}_r$, with θ_0 this vector at $f = f_0$, and with r taken equal to $4\sqrt{n}\varepsilon r$, we obtain, for $4\sqrt{n}\varepsilon r > 1$,

$$A \le 9D(2\sqrt{n}\varepsilon r, \Theta)e^{-2n\varepsilon^2 r^2} \le 9e^{-n\varepsilon^2 r^2}.$$

In view of Lemma 14 applied with r equal to $\sqrt{n}\varepsilon r$, there exists an event $\mathcal{A}$ such that

$$B \le e^{-n\varepsilon^2 r^2/8},$$

while on the event $\mathcal{A}$,

$$\int \frac{p_{n,f}}{p_{n,f_0}}\, d\Pi(f) \ge e^{-n\varepsilon^2 r^2}\Pi\big(f\colon \|f - f_0\|_n < \varepsilon r\big) \ge e^{-n\varepsilon^2(r^2+1)}.$$

It follows that on the event $\mathcal{A}$, for any set $\mathcal{B}$,

$$\Pi_n(\mathcal{B}|Y_{1:n}) \le e^{n\varepsilon^2(r^2+1)} \int_{\mathcal{B}} p_{n,f}/p_{n,f_0}\, d\Pi(f).$$

Therefore, in view of the fact that $P_{n,f_0}(p_{n,f}/p_{n.f_0}) \le 1$, we obtain,

$$\begin{aligned} C &\le e^{n\varepsilon^2(r^2+1)} P_{n,f_0} \int_{\mathcal{F}_r^c} p_{n,f}/p_{n,f_0}\, d\Pi(f) \\ &\le e^{n\varepsilon^2(r^2+1)}\Pi(\mathcal{F}_r^c) \le e^{-n\varepsilon^2(r^2-1)}. \end{aligned} \tag{17}$$

Finally, in view of the fact that $P_{n,f_0}(p_{n,f}/p_{n.f_0})(1-\phi) \le P_{n,f}(1-\phi)$, which is bounded above by $e^{-2j^2 n\varepsilon^2 r^2}$ for f contained in $C_{j,r} := \{f \in \mathcal{F}_{n,r}\colon 4j\varepsilon r \le \|f - f_0\|_n < 4(j+1)\varepsilon r\}$ by the second inequality in Lemma 13, we obtain, again using Fubini's theorem,

$$\begin{aligned} D &\le e^{n\varepsilon^2(r^2+1)} \sum_{j\ge 1} P_{n,f_0}(1-\phi) \int_{C_{j,r}} p_{n,f}/p_{n,f_0}\, d\Pi(f) \\ &\le e^{n\varepsilon^2(r^2+1)} \sum_{j\ge 1} e^{-2j^2 n\varepsilon^2 r^2} \le 9e^{-n\varepsilon^2(r^2-1)}, \end{aligned}$$

for $n\varepsilon^2 r^2 \ge 1/16$, as $1/(1-e^{-1/8}) \approx 8.5$.

Finally we write

$$\begin{aligned} P_{n,f_0} \int \|f-f_0\|_n^l \, d\Pi_n\big(f|Y_{1:n}\big) \\ &= P_{n,f_0} \int_0^\infty l r^{l-1} \Pi_n\big(\|f-f_0\|_n > 4\varepsilon r | Y_{1:n}\big)\, dr\, (4\varepsilon)^l \\ &\le (8\varepsilon)^l + (4\varepsilon)^l P_{n,f_0} \int_2^\infty l r^{l-1} (A+B+C+D)(r)\, dr. \end{aligned}$$

Inserting the bound on $A+B+C+D$ obtained previously we see that the integral is bounded by $10\int_2^\infty (e^{-r^2/8} + e^{-(r^2-1)})\, dr < \infty$. ■

Proof [Proof of Theorem 1] Theorem 1 is a specialization of Proposition 11 to Gaussian priors, where the conditions of the proposition are reexpressed in terms of the concentration function ϕ_{f_0} of the prior. The details are the same as in Van der Vaart and Van Zanten (2008a).

First we note that $\varepsilon := 2\psi_{f_0}^{-1}(n)$ satisfies $\phi_{f_0}(\varepsilon/2) \le n\varepsilon^2/4 \le n\varepsilon^2$. It is shown in Kuelbs et al. (1994) (or see Lemma 5.3 in Van der Vaart and Van Zanten, 2008b) that the concentration function ϕ_{f_0} determines the small ball probabilities around f_0, in the sense that, for the given ε,

$$\Pi\big(f: \|f-f_0\|_\infty < \varepsilon\big) \ge e^{-n\varepsilon^2}. \tag{18}$$

Because $\|\cdot\|_n \le \|\cdot\|_\infty$, it follows that (16) is satisfied.

For $\mathbb{H}_1$ and $\mathbb{B}_1$ the unit balls of the RKHS and $\mathbb{B}$ and $M_r = -2\Phi^{-1}(e^{-n\varepsilon^2 r^2})$, we define sets $\mathcal{F}_r = \varepsilon\mathbb{B}_1 + M_r\mathbb{H}_1$. By Borell's inequality (see Borell, 2008, or Theorem 5.1 in Van der Vaart and Van Zanten, 2008b) these sets have prior probability $\Pi(\mathcal{F}_r)$ bounded below by $1-\Phi(\alpha+M_r)$, for Φ the standard normal distribution function and α the solution to the equation $\Phi(\alpha) = \Pi\big(f: \|f\|_\infty < \varepsilon\big) = e^{-\phi_0(\varepsilon)}$. Because $\Phi(\alpha) \ge e^{-n\varepsilon^2} \ge e^{-n\varepsilon^2 r^2}$, we have $\alpha + M_r \ge -\Phi^{-1}(e^{-n\varepsilon^2 r^2})$. We conclude that $\Pi(\mathcal{F}_r) \ge 1 - e^{-n\varepsilon^2 r^2}$.

It is shown in the proof of Theorem 2.1 of Van der Vaart and Van Zanten (2008a) that the sets $\mathcal{F}_r$ also satisfy the entropy bound (15), for the norm $\|\cdot\|_\infty$, and hence certainly for $\|\cdot\|_n$. ■

4.2 Proof of Theorem 2

For a function $f: [0,1]^d \to \mathbb{R}$ and $\alpha > 0$ let $\|f\|_{\alpha|\infty}$ be the Besov norm of regularity α measured using the $L_\infty - L_\infty$-norms (see (19) below). This is bounded by the Hölder norm of order α (see for instance Cohen et al., 2001 for details).

Lemma 15 *Let $\mathcal{X} = [0,1]^d$ and suppose that the density of the covariates is bounded below by a constant c. Then $\|f\|_\infty \lesssim c^{-2\alpha/(2\alpha+d)} \|f\|_{\alpha|\infty}^{d/(2\alpha+d)} \|f\|_2^{2\alpha/(2\alpha+d)}$, for any function $f: [0,1]^d \to \mathbb{R}$.*

Proof We can assume without loss of generality that the covariate distribution is the uniform distribution. We can write the function as the Fourier series $f = \sum_{j=0}^\infty \sum_k \sum_v \beta_{j,k,v} e_{j,k,v}$ relative to a basis $(e_{j,k,v})$ of orthonormal wavelets in $L_2(\mathbb{R}^d)$. (Here k runs for each fixed j through an index set for of the order $O(2^{jd})$ translates, and v runs through $\{0,1\}^d$ when $j=0$ and $\{0,1\}^d \setminus \{0\}$ when $j \ge 1$.)

For wavelets constructed from suitable scaling functions, the various norms of f can be expressed in the coefficients through (up to constants, see for instance Cohen et al., 2001, Section 2)

$$\begin{aligned}\|f\|_2 &= \Big(\sum_j \sum_k \sum_v \beta_{j,k,v}^2\Big)^{1/2},\\ \|f\|_\infty &\le \sum_j \max_k \max_v |\beta_{j,k,v}| 2^{jd/2},\\ \|f\|_{\alpha|\infty} &= \sup_j \max_k \max_v |\beta_{j,k,v}| 2^{j(\alpha+d/2)}. \end{aligned} \tag{19}$$

For given J let $f_J = \sum_{j\le J}\sum_k \sum_v \beta_{j,k,v} e_{j,k,v}$ be the projection of f on the base elements of resolution level bounded by J. Then

$$\begin{aligned}\|f - f_J\|_\infty &\le \sum_{j>J} \max_k \max_v |\beta_{j,k,v}| 2^{jd/2}\\ &\le \sum_{j>J} 2^{-j(\alpha+d/2)} \|f\|_{\alpha|\infty} 2^{jd/2} \le 2^{-J\alpha}\|f\|_{\alpha|\infty}.\end{aligned}$$

Furthermore, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\|f_J\|_\infty &\le \sum_{j\le J} \max_k \max_v |\beta_{j,k,v}| 2^{jd/2}\\ &\le \Big(\sum_{j\le J} \max_k \max_v \beta_{j,k,v}^2\Big)^{1/2} \Big(\sum_{j\le J} 2^{jd}\Big)^{1/2}\\ &\le \|f\|_2 2^{Jd/2},\end{aligned}$$

where in the last inequality we have bounded the maximum over (k,v) by the sum.

Combining the two preceding displays we see that $\|f\|_\infty \le 2^{-J\alpha}\|f\|_{\alpha|\infty} + \|f\|_2 2^{Jd/2}$. We finish the proof by choosing J to balance the two terms on the right. ■

Proof [Proof of Theorem 2] Let $\varepsilon = 2\psi_{f_0}^{-1}(n)$ so that $\phi_{f_0}(\varepsilon/2) \le n\varepsilon^2$ and (18) holds. By the definition of ϕ_{f_0} there exists an element f_ε of the RKHS of the prior with $\|f_\varepsilon - f_0\|_\infty \le \varepsilon/2$ and $\|f_\varepsilon\|_{\mathbb{H}}^2 \le \phi_{f_0}(\varepsilon/2) \le n\varepsilon^2$. Because $\|f_\varepsilon - f_0\|_2 \le \|f_\varepsilon - f_0\|_\infty \le \varepsilon$, the posterior second moments of $\|f - f_\varepsilon\|_2$ and $\|f - f_0\|_2$ are within a multiple of ε^2, and hence it suffices to bound the former of the two.

For any positive constants γ, τ, any $\eta \ge \varepsilon$, and any events $\mathcal{A}_r$ we can bound

$$\begin{aligned}&\frac{1}{\eta^2} \mathrm{E}_{f_0} \int \|f - f_\varepsilon\|_2^2 \, d\Pi(f|X_{1:n}, Y_{1:n})\\ &\qquad = \mathrm{E}_{f_0} \int_0^\infty r\Pi\big(f: \|f - f_\varepsilon\|_2 > \eta r | X_{1:n}, Y_{1:n}\big)\, dr\end{aligned}$$

by $I+II+III+IV$, for

$$\begin{aligned}
I &= \mathrm{E}_{f_0} \int_0^\infty r\Pi\big(f\colon 2\|f-f_\varepsilon\|_n > \eta r | X_{1:n}, Y_{1:n}\big)\, dr,\\
II &= \mathrm{E}_{f_0} \int_0^\infty r 1_{\mathcal{A}_r^c}\, dr,\\
III &= \mathrm{E}_{f_0} \int_0^\infty r 1_{\mathcal{A}_r} \Pi\big(\|f\|_{\alpha|\infty} > \tau\sqrt{n}\eta r^\gamma | X_{1:n}, Y_{1:n}\big)\, dr,\\
IV &= \mathrm{E}_{f_0} \int_0^\infty r 1_{\mathcal{A}_r} \Pi\big(f\colon \|f-f_\varepsilon\|_2 > \eta r \geq 2\|f-f_\varepsilon\|_n,\\
&\qquad\qquad \|f\|_{\alpha|\infty} \leq \tau\sqrt{n}\eta r^\gamma | X_{1:n}, Y_{1:n}\big)\, dr.
\end{aligned}$$

The term I is the quadratic risk in terms of the empirical norm, centered at f_ε. Conditioned on the design points and centered at f_0 this was seen to be bounded in the previous section (as $\eta \geq \varepsilon$), uniformly in the design points. Because $\|f_0 - f_\varepsilon\|_\infty \leq \varepsilon$, the term I is bounded by a constant.

In view of Lemma 14, with r of the lemma equal to $\sqrt{n}\varepsilon r^\gamma$, there exist events $\mathcal{A}_r$ such that

$$II \leq \int_0^\infty r e^{-n\varepsilon^2 r^{2\gamma}/8}\, dr \lesssim 1,$$

while on the event $\mathcal{A}_r$,

$$\begin{aligned}
\int \frac{p_{n,f}}{p_{n,f_0}}\, d\Pi(f) &\geq e^{-n\varepsilon^2 r^{2\gamma}} \Pi\big(f\colon \|f-f_0\|_n < \varepsilon r^\gamma\big)\\
&\geq e^{-n\varepsilon^2(r^{2\gamma}+1)},
\end{aligned} \tag{20}$$

by (18) and because $\|\cdot\|_n \leq \|\cdot\|_\infty$.

Because the prior Π is concentrated on the functions with $\|f\|_{\alpha|\infty} < \infty$ by assumption, it can be viewed as the distribution of a Gaussian random element with values in the Hölder space $C^\alpha[0,1]^d$. It follows that $\tau^2 := 16\int \|f\|_{\alpha|\infty}^2\, d\Pi(f)$ is finite, and $\Pi\big(f\colon \|f\|_{\alpha|\infty} > \tau x\big) \leq e^{-2x^2}$, for every $x > 0$, by Borell's inequality (e.g., Van der Vaart and Wellner, 1996, A.2.1.). By the same argument as used to obtain (17) in the proof of Proposition 11, we see that

$$\begin{aligned}
III &\leq 1 + \int_1^\infty r e^{n\varepsilon^2(r^{2\gamma}+1)} \Pi\big(f\colon \|f\|_{\alpha|\infty} > \tau\sqrt{n}\eta r^\gamma\big)\, dr\\
&\leq 1 + \int_1^\infty r e^{n\varepsilon^2(r^{2\gamma}+1)} e^{-2n\eta^2 r^{2\gamma}}\, dr \lesssim 2.
\end{aligned}$$

It remains to prove that IV is bounded as well.

The squared empirical norm $\|f-f_\varepsilon\|_n^2$ is the average of the independent random variables $(f-f_\varepsilon)^2(X_i)$, which have expectation $\|f-f_\varepsilon\|_2^2$, and variance bounded by $P(f-f_\varepsilon)^4 \leq \|f-f_\varepsilon\|_2^2\|f-f_\varepsilon\|_\infty^2$. Therefore, we can apply Bernstein's inequality (see, e.g., Lemma 2.2.9 in Van der Vaart and Wellner, 1996) to see that

$$\mathrm{P}\big(\|f-f_\varepsilon\|_2 \geq 2\|f-f_\varepsilon\|_n\big) \leq e^{-(n/5)\|f-f_\varepsilon\|_2^2/\|f-f_\varepsilon\|_\infty^2}.$$

The unit ball of the RKHS of a GP f is always contained in c times the unit ball of the Banach space on which it is supported, for $c^2 = \mathrm{E}\|f\|^2$, where $\|\cdot\|$ is the norm of the Banach space (see, e.g.,

Van der Vaart and Van Zanten, 2008b), formula (2.5)). An equivalent statement is that the Banach norm $\|f\|$ of an element of the RKHS is bounded above by c times its RKHS-norm. Because Π is concentrated on $C^\alpha[0,1]^d$, we can apply this general fact with $\|\cdot\|$ the α-Hölder norm, and conclude that the α-Hölder norm of an element of the RKHS is bounded above by $\tau/4$ times its RKHS-norm, for $\tau/4$ the second moment of the prior norm defined previously. In particular $\|f_\varepsilon\|_{\alpha|\infty} \le \tau\|f_\varepsilon\|_{\mathbb{H}} \le \tau\sqrt{n}\varepsilon$. Therefore, for f in the set $\mathcal{F}$ of functions with $\|f\|_{\alpha|\infty} \le \tau\sqrt{n}\varepsilon r^\gamma$, we have $\|f - f_\varepsilon\|_{\alpha|\infty} \le 2\tau\sqrt{n}\varepsilon r^\gamma$, whence by Lemma 15 for $f \in \mathcal{F}$ we can replace $\|f - f_\varepsilon\|_\infty$ in the preceding display by $c(2\tau\sqrt{n}\varepsilon r^\gamma)^{d/(2\alpha+d)}\|f - f_\varepsilon\|_2^{2\alpha/(2\alpha+d)}$, for a constant c depending on the covariate density. We then have

$$\begin{aligned}
&\mathrm{E}_{f_0}\Pi\big(f \in \mathcal{F}: \|f - f_\varepsilon\|_2 > \eta r \ge 2\|f - f_\varepsilon\|_n\big) \\
&\quad\le \int_{f\in\mathcal{F}:\|f-f_\varepsilon\|_2>\eta r} \mathrm{P}\big(\|f - f_\varepsilon\|_2 \ge 2\|f - f_\varepsilon\|_n\big)\, d\Pi(f) \\
&\quad\le \int_{\|f-f_\varepsilon\|_2>\eta r} \exp\Big(-\frac{n}{5c^2}\Big(\frac{\|f - f_\varepsilon\|_2}{2\tau\sqrt{n}\varepsilon r^\gamma}\Big)^{2d/(2\alpha+d)}\Big)\, d\Pi(f) \\
&\quad\le \exp\Big(-Cn^{2\alpha/(2\alpha+d)}(\eta r^{1-\gamma}/\varepsilon)^{2d/(2\alpha+d)}\Big),
\end{aligned}$$

for $1/C = 5c^2(2\tau)^{2d/(2\alpha+d)}$. Substitution of this bound and the lower bound (20) in IV yields

$$IV \le 1 + \int_1^\infty r e^{n\varepsilon^2(r^{2\gamma}+1)} e^{-Cn^{2\alpha/(2\alpha+d)}(\eta r^{1-\gamma}/\varepsilon)^{2d/(2\alpha+d)}}\, dr.$$

For $Cn^{2\alpha/(2\alpha+d)}(\eta/\varepsilon)^{2d/(2\alpha+d)} \ge n\varepsilon^2$ this is finite if $\gamma > 0$ is chosen sufficiently small. Equivalently, IV is bounded if $\eta \gtrsim \sqrt{n}\varepsilon^{(2\alpha+2d)/d}$.

We must combine this with the requirement made at the beginning of the proof that $\eta \ge \varepsilon \ge 2\psi_{f_0}^{-1}(n)$. If $\varepsilon \le n^{-d/(4\alpha+2d)}$, then $\sqrt{n}\varepsilon^{(2\alpha+2d)/d} \le \varepsilon$ and hence the requirement $\eta \gtrsim \sqrt{n}\varepsilon^{(2\alpha+2d)/d}$ is satisfied for $\eta = \varepsilon$. Otherwise, we choose $\eta \sim \sqrt{n}\varepsilon^{(2\alpha+2d)/d} \gg \varepsilon$. In both cases we have proved that the posterior second moment has mean bounded by a multiple of η^2. ■

4.3 Proofs for Section 3

Proof [Proof of Lemma 3] The Fourier transform of h_ψ given in (11) is, up to constants, the function $\phi = \psi m$, and for ψ the minimal choice as in (12) this function satisfies (cf., (10))

$$\int |\phi(\lambda)|^2 \big(1 + \|\lambda\|^2\big)^{\alpha+d/2}\, d\lambda = \|h_\psi\|_{\mathbb{H}}^2.$$

In other words, the unit ball $\mathbb{H}_1$ of the RKHS is contained in a Sobolev ball of order $\alpha + d/2$. (See Section 1.5 for the definition of Sobolev spaces.) The metric entropy relative to the uniform norm of such a Sobolev ball is bounded by a constant times $(1/\varepsilon)^{d/(\alpha+d/2)}$ (see Theorem 3.3.2 on p. 105 in Edmunds and Triebel, 1996). The lemma next follows from the results of Kuelbs and Li (1993) and Li and Linde (1998) that characterize the small ball probability in terms of the entropy of the RKHS-unit ball. ■

Proof [Proof of Lemma 4] Let $\kappa:\mathbb{R}\to\mathbb{R}$ be a function with a real, symmetric Fourier transform $\hat{\kappa}$, which equals $1/(2\pi)$ in a neighborhood of 0 and which has compact support. From $\hat{\kappa}(\lambda)=(2\pi)^{-1}\int e^{i\lambda t}\kappa(t)\,dt$ it then follows that $\int\kappa(t)\,dt=1$ and $\int(it)^k\kappa(t)\,dt=0$ for $k\geq 1$. For $t=(t_1,\ldots,t_d)$, define $\phi(t)=\kappa(t_1)\cdots\kappa(t_d)$. Then ϕ integrates to 1, has finite absolute moments of all orders, and vanishing moments of all orders bigger than 0.

For $\sigma>0$ set $\phi_\sigma(x)=\sigma^{-d}\phi(x/\sigma)$ and $h=\phi_\sigma * f_0$. Because ϕ is a higher order kernel, standard arguments from the theory of kernel estimation shows that $\|f_0-\phi_\sigma * f_0\|_\infty\lesssim\sigma^\beta$.

The Fourier transform of h is the function $\lambda\mapsto\hat{h}(\lambda)=\hat{\phi}(\sigma\lambda)\,\hat{f}_0(\lambda)$, and therefore (12) and (13) show that

$$\begin{aligned}\|h\|_{\mathbb{H}}^2&\lesssim\int\bigl|\hat{\phi}(\sigma\lambda)\,\hat{f}_0(\lambda)\bigr|^2\frac{1}{m(\lambda)}\,d\lambda\\&\lesssim\sup_\lambda\Bigl[\bigl(1+\|\lambda\|^2\bigr)^{\alpha+d/2-\beta}\bigl|\hat{\phi}(\sigma\lambda)\bigr|^2\Bigr]\|f_0\|_{\beta|2}^2\\&\lesssim C(\sigma)\sup_\lambda\Bigl[\bigl(1+\|\lambda\|^2\bigr)^{\alpha+d/2-\beta}\bigl|\hat{\phi}(\lambda)\bigr|^2\Bigr]\|f_0\|_{\beta|2}^2.\end{aligned}$$

for

$$C(\sigma)=\sup_\lambda\Bigl(\frac{1+\|\lambda\|^2}{1+\|\sigma\lambda\|^2}\Bigr)^{\alpha+d/2-\beta}\lesssim\Bigl(\frac{1}{\sigma}\Bigr)^{2\alpha+d-2\beta},$$

if $\sigma\leq 1$. The assertion of the lemma follows upon choosing $\sigma\sim\varepsilon^{1/\beta}$. ■

Proof [Proof of Lemma 7] For given $K>0$ let $\psi(\lambda)=(\hat{f}_0/m)(\lambda)1_{\|\lambda\|\leq K}$. The function h_ψ defined by (11) with m given in (14) satisfies

$$\begin{aligned}\|h_\psi-f_0\|_\infty&\leq\int_{\|\lambda\|>K}|\hat{f}_0(\lambda)|\,d\lambda\\&\leq\|f_0\|_{\beta|2}\Bigl(\int_{\|\lambda\|>K}\bigl(1+\|\lambda\|^2\bigr)^{-\beta}d\lambda\Bigr)^{1/2}\\&\lesssim\|f_0\|_{\beta|2}\frac{1}{K^{\beta-d/2}}.\end{aligned}$$

Furthermore, the squared RKHS-norm of h_ψ is given by

$$\begin{aligned}\|h_\psi\|_{\mathbb{H}}^2&=\int_{\|\lambda\|\leq K}\frac{|\hat{f}_0|^2}{m}(\lambda)\,d\lambda\\&\leq\sup_{\|\lambda\|\leq K}m(\lambda)^{-1}\bigl(1+\|\lambda\|^2\bigr)^{-\beta}\|f_0\|_{\beta|2}^2\\&\lesssim e^{K^2/4}\|f_0\|_{\beta|2}^2.\end{aligned}$$

We conclude the proof by choosing $K\sim\varepsilon^{-1/(\beta-d/2)}$. ■

Proof [Proof of 9] The first assertion is proved in Van der Vaart and Van Zanten (2009), Lemma 4.4. The second assertion is proved in the same way as Lemma 7, where this time, with $\|f_0\|_{\mathcal{A}}$ the norm

of f_0 in $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$,

$$\begin{aligned}\|h_\psi - f_0\|_\infty^2 &\le \int_{\|\lambda\|>K} e^{-\gamma\|\lambda\|^r}\, d\lambda \|f_0\|_{\mathcal{A}}^2 \\ &\le e^{-\gamma K^r} K^{-r+1} \|f_0\|_{\mathcal{A}}^2, \\ \|h_\psi\|_{\mathbb{H}}^2 &\le \sup_{\|\lambda\|\le K} e^{\|\lambda\|^2/4 - \gamma\|\lambda\|^r} \|f_0\|_{\mathcal{A}}^2 \le e^{K^2/4} \|f_0\|_{\mathcal{A}}^2.\end{aligned}$$

We finish by choosing $K \sim \big(\gamma^{-1}\log(1/\varepsilon)\big)^{1/r}$. ■

4.4 Miscellaneous Results

Proof [Proof of Theorem 8] We start by proving the following lower bound on the concentration function: there exists $b, v > 0$ such that for $\varepsilon \downarrow 0$,

$$\begin{aligned}\phi_{f_0}(\varepsilon) &\ge \inf_{\psi:\|h_\psi - f_0\|_2 < \varepsilon} \|h_\psi\|_{\mathbb{H}}^2 \\ &\gtrsim \exp\big(b\varepsilon^{-v}\big).\end{aligned} \tag{21}$$

For given $\varepsilon > 0$ let h_ψ be a function in the RKHS of the form (11) such that $\|h_\psi - f_0\|_2 < \varepsilon$. Let r be a function which is equal to 1 on the support of f_0, has itself support within $[0,1]$ and Fourier transform with exponentially small tails: $|\hat{r}(\lambda)\exp(|\lambda|^u) \to 0$ as $|\lambda| \to \infty$, for some $u > 0$. (Such a function exists for $u < 1$.) Then $h_\psi r$ has support inside $[0,1]$ and $f_0 r = f_0$, so that $\|h_\psi r - f_0\|_{2,\mathbb{R}} \le \|h_\psi - f_0\|_2$, where $\|\cdot\|_{2,\mathbb{R}}$ is the norm of $L_2(\mathbb{R}^d)$ and $\|\cdot\|_2$ the norm of $L_2[0,1]^d$. The function $h_\psi r$ has Fourier transform $(\psi m) * \hat{r}$, and hence by Parseval's identity $\big\|(\psi m) * \hat{r} - \hat{f}_0\big\|_{2,\mathbb{R}} < \varepsilon$. Therefore, for $K > 0$ and χ_K the indicator of the set $\{\lambda \in \mathbb{R}^d : \|\lambda\| > K\}$,

$$\big\|(\psi m) * \hat{r}\chi_{2K}\big\|_{2,\mathbb{R}} \ge \|\hat{f}_0\chi_{2K}\|_{2,\mathbb{R}} - \varepsilon \ge c(1/K)^{k-d/2} - \varepsilon,$$

by the assumption on $\hat{f}_0$, for some constant c. By Lemma 16 with $A = K/2$ and $2K$ instead of K, it follows that

$$\|\psi m \chi_K\|_{2,\mathbb{R}} \|\hat{r}(1-\chi_K)\|_{1,\mathbb{R}} \ge c(1/K)^{k-d/2} - \varepsilon - \|\psi m\|_{2,\mathbb{R}} \|\hat{r}\chi_K\|_1.$$

In view of (12) we have that $\|h_\psi\|_{\mathbb{H}} = \|\psi\sqrt{m}\|_{2,\mathbb{R}}$ and hence $\|\psi m \chi_K\|_{2,\mathbb{R}} \le \sqrt{m(K)}\|h_\psi\|_{\mathbb{H}}$, and $\|\psi m\|_{2,\mathbb{R}} \le \|h_\psi\|_{\mathbb{H}}$. Combining this with the preceding display we see that

$$\big(\|\hat{r}(1-\chi_K)\|_{1,\mathbb{R}}\sqrt{m(K)} + \|\hat{r}\chi_K\|_1\big)\|h_\psi\|_{\mathbb{H}} \ge c(1/K)^{k-d/2} - \varepsilon = \varepsilon,$$

for $K = (c/2\varepsilon)^{1/(k-d/2)}$. Here $\|\hat{r}(1-\chi_K)\|_{1,\mathbb{R}}\sqrt{m(K)}$ is of the order $\exp(-K^2/4)$, in view of the definition (14) of m and the fact that $\hat{r}$ is integrable, and $\|\hat{r}\chi_K\|_1$ is of the order $\exp(-dK^u)$, by construction. The proof of (21) is complete upon substituting $K = (c/2\varepsilon)^{1/(k-d/2)}$ and rearranging the preceding display.

The prior mass of a ball of radius ε around f_0 is bounded below by $e^{-\phi_{f_0}(\varepsilon/2)}$ and bounded above by $e^{-\phi_{f_0}(\varepsilon)}$, where we can use any norm. In view of (21) and Lemmas 6 and 7 we conclude that

there exist constants such

$$\exp\bigl(-e^{a\varepsilon^{-u}}\bigr) \le \Pi(f\colon \|f-f_0\|_\infty < \varepsilon),$$
$$\Pi(f\colon \|f-f_0\|_2 < \varepsilon) \le \exp\bigl(-e^{b\varepsilon^{-v}}\bigr).$$

By choosing η_n, ε_n such that $a\varepsilon_n^{-u} = \log n^s$ and $b\eta_n^{-v} = \log n^t$, we obtain that

$$\frac{\Pi\bigl(f\colon \|f-f_0\|_2 < \eta_n\bigr)}{\Pi\bigl(f\colon \|f-f_0\|_\infty < \varepsilon_n\bigr)} \le \exp(-n^t + n^s) \ll e^{-2n\varepsilon_n^2},$$

if $t > 1 \vee s$. It then follows that $\mathrm{E}_{f_0}\Pi\bigl(f\colon \|f-f_0\|_2 < \eta_n | X_{1:n}, Y_{1:n}\bigr) \to 0$, by the same argument as given to prove (17). ■

If the convolution of a function f with a light-tailed function g has heavy tails, then f itself must have heavy tails. The following quantitative version of this principle underlies the preceding proof.

Lemma 16 *For arbitrary functions $f, g\colon \mathbb{R} \to \mathbb{R}$, χ_K the indicator function of $\{\lambda \in \mathbb{R}^d\colon \|\lambda\| > K\}$, and $0 < A < K$,*

$$\|f\chi_{K-A}\|_2 \|g(1-\chi_A)\|_1 \ge \|(f*g)\chi_K\|_2 - \|f\|_2 \|g\chi_A\|_1.$$

Proof For f_t the function $\lambda \mapsto f(\lambda - t)$, we have $\|f_t\chi_K\|_2 \le \|f\chi_{K-A}\|$ if $\|t\| \le A$, and $\|f_t\chi_K\|_2 \le \|f\|_2$ for every t. Therefore

$$\begin{aligned}\Bigl\|\int f_t\chi_K\, g(t)\, dt\Bigr\|_2 &\le \int \|f_t\chi_K\|_2\, |g(t)|\, dt \\ &\le \|f\chi_{K-A}\|_2 \int_{\|t\|\le A} |g(t)|\, dt + \|f\|_2 \int_{\|t\|>A} |g(t)|\, dt.\end{aligned}$$

It suffices to arrange this inequality. ■

Lemma 17 *For ψ_{f_0} defined by (8) and $f_0 \in \mathbb{H}$ we have (9).*

Proof Because the function ψ_{f_0} is decreasing, the relation $\psi_{f_0}(\varepsilon) \le n$ for some ε implies that $\psi_{f_0}^{-1}(n) \le \varepsilon$. Consequently, if $\tilde{\psi}_{f_0}$ is an upper bound on ψ_{f_0}, then $\tilde{\psi}_{f_0}(\varepsilon) \le n$ for some ε implies that $\psi_{f_0}^{-1}(n) \le \varepsilon$. If $f_0 \in \mathbb{H}$, then we can choose $h = f_0$ in the infimum in the definition of ϕ_{f_0}, and hence we obtain

$$\phi_{f_0}(\varepsilon) \le \|f_0\|_{\mathbb{H}}^2 + \phi_0(\varepsilon).$$

If both $\|f_0\|_{\mathbb{H}}^2 \le n\varepsilon^2/2$ and $\psi_0(\varepsilon) \le n\varepsilon^2/2$, then $\tilde{\psi}_{f_0}(\varepsilon) \le n$ and hence $\psi_{f_0}^{-1}(n) \le \varepsilon$. ■

References

A. R. Barron. Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In *Bayesian Statistics, 6 (Alcoceber, 1998)*, pages 27–52. Oxford Univ. Press, New York, 1999.

H. Bauer. *Measure and integration theory*, volume 26 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 2001.

C. Borell. Inequalities of the Brunn-Minkowski type for Gaussian measures. *Probab. Theory Related Fields*, 140(1-2):195–205, 2008.

I. Castillo. Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.*, 2: 1281–1299, 2008.

A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore. Tree approximation and optimal encoding. *Appl. Comput. Harmon. Anal.*, 11(2):192–226, 2001.

D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*, volume 120 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1996.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*, 2nd edition. Springer-Verlag, New York, 1991.

J. Kuelbs and W. V. Li. Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.*, 116(1):133–157, 1993.

J. Kuelbs, W. V. Li, and W. Linde. The Gaussian measure of shifted balls. *Probab. Theory Related Fields*, 98(2):143–162, 1994.

A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17: 277–364, 1961

W. V. Li and Q.-M. Shao. Gaussian processes: inequalities, small ball probabilities and applications. In *Stochastic Processes: Theory and Methods*, volume 19 of *Handbook of Statist.*, pages 533–597. North-Holland, Amsterdam, 2001.

W. V. Li and W. Linde. Existence of small ball constants for fractional Brownian motions. *C. R. Acad. Sci. Paris Sér. I Math.*, 326(11):1329–1334, 1998.

M. A. Lifshits. *Gaussian Random Functions*. Kluwer Academic Publishers, Dordrecht, 1995.

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

M. W. Seeger, S. M. Kakade, and D. P. Foster. Information consistency of nonparametric Gaussian process methods. *IEEE Trans. Inform. Theory*, 54(5):2376–2382, 2008.

A. B. Tsybakov., *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

A. W. van der Vaart and J. H. van Zanten. Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.*, 1:433–448 (electronic), 2007.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 2008a.

A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3 of *Inst. Math. Stat. Collect.*, pages 200–222. Inst. Math. Statist., Beachwood, OH, 2008b.

A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

G. Wahba. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B*, 40(3): 364–372, 1978.