Asymptotic Bounds for Smoothness Parameter Estimates in Gaussian Process Interpolation[*]

Toni Karvonen

Department of Mathematics and Statistics
University of Helsinki, Finland

Abstract. It is common to model a deterministic response function, such as the output of a computer experiment, as a Gaussian process with a Matérn covariance kernel. The smoothness parameter of a Matérn kernel determines many important properties of the model in the large data limit, including the rate of convergence of the conditional mean to the response function. We prove that the maximum likelihood estimate of the smoothness parameter cannot asymptotically undersmooth the truth when the data are obtained on a fixed bounded subset of $\mathbb{R}^d$. That is, if the data-generating response function has Sobolev smoothness $\nu_0 > d/2$, then the smoothness parameter estimate cannot be asymptotically less than ν_0. The lower bound is sharp. Additionally, we show that maximum likelihood estimation recovers the true smoothness for a class of compactly supported self-similar functions. For cross-validation we prove an asymptotic lower bound $\nu_0 - d/2$, which however is unlikely to be sharp. The results are based on approximation theory in Sobolev spaces and some general theorems that restrict the set of values that the parameter estimators can take.

Keywords: Gaussian processes, parameter estimation, Matérn kernels, self-similarity

MSC2020: 60G15, 62G20, 46E22, 65D12

1   Introduction

Gaussian process interpolation is commonly used to approximate a deterministic response or data-generating function which may, for example, represent the output of a computer experiment (Sacks et al., 1989). A zero-mean Gaussian process is defined by a positive-definite covariance kernel K_θ with parameters $\theta \in \Theta$. To ensure that Gaussian process interpolation yields a good approximation and reasonable quantification of uncertainty for the response function at unseen data locations, it is necessary to estimate the kernel parameters from the data. Due to its flexibility and interpretability, the Matérn class of stationary covariance kernels is often preferred in applications (Stein, 1999). Let ν, σ, and λ be positive smoothness, magnitude, and scale parameters, respectively. A Matérn kernel on $\mathbb{R}^d$ is defined as

$$K_\nu(x,y) = \sigma^2 c(\nu) \left(\frac{\sqrt{2\nu}\,\|x-y\|}{\lambda} \right)^\nu \mathcal{K}_\nu \left(\frac{\sqrt{2\nu}\,\|x-y\|}{\lambda} \right) \quad \text{for} \quad x,y \in \mathbb{R}^d, \qquad (1.1)$$

where $\mathcal{K}_\nu$ is the modified Bessel function of the second kind of order ν and $c(\nu)$ a positive ν-dependent scaling factor. Much is known about fixed-domain asymptotics of various estimators for the parameters σ and λ (as well as $\sigma^2 \lambda^{2\nu}$ for a fixed ν) for Matérns and related kernels in both the *Bayesian* setting where the response function is assumed to be a Gaussian process (e.g., Ying, 1991; Loh, 2005; Anderes, 2010; Bachoc et al., 2017) and the *frequentist* setting where the

response function is a fixed deterministic function (Xu and Stein, 2017; Karvonen et al., 2020). As it defines assumed degree of differentiability of the response function and is microergdic (Stein, 1999, Section 6.2), the smoothness parameter ν is arguably the most important parameter of a Matérn kernel. While it is common to fix ν beforehand, doing so is problematic when the smoothness of the response function is unknown (though data-driven estimation of σ and λ may overcome these issues):

- If the model *undersmooths* the truth (i.e., the response function is smoother than assumed), uncertainty quantification is reliable, in the sense that the response function is contained in a credible set centered at the conditional mean for some fixed credible level [see (4.1) and (4.2)]. However, this comes at the cost of (likely) sub-optimal approximation accuracy.

- If the model *oversmooths* the truth, the approximation quality is best possible, in that the Narcowich–Ward–Wendland escape theorem [see (4.3)] guarantees convergence of the conditional mean to the response function with a rate that is worst-case optimal in any Sobolev space which contains the response function. However, uncertainty quantification may be unreliable.

The effects of under- and oversmoothing in the frequentist setting, as well as connections to the literature on construction of adaptive confidence and credible sets, are discussed in more detail in Section 4.1.

Maximum likelihood estimation is perhaps the most popular data-driven approach to select the parameters of a Matérn kernel. It seems that the only theoretical results concerning maximum likelihood estimation of the smoothness parameter of a Matérn-type kernel have been obtained by Chen et al. (2021) and Petit (2023), who consider the periodic version of the Matérn kernel on $[0,1]^d$ (see Stein, 1999, Section 6.7) and show that maximum likelihood estimators are consistent. Szabó et al. (2015) and Knapik et al. (2016) have derived results for maximum likelihood smoothness estimation in a related white noise model. Loh (2015); Loh et al. (2021); and Loh and Sun (2023) construct other smoothness estimators for the Matérn model whose consistency they prove under certain sampling schemes on $[0,1]^d$. Other work on maximum likelihood estimation, as well as cross-validation, of parameters in Gaussian process and related models can be found in Bachoc (2013); Szabó et al. (2013); Szabo and Rousseau (2017); Bachoc et al. (2017); Xu and Stein (2017); and Hadji and Szábo (2021). This article contains what appear to be the first theoretical results on maximum likelihood estimation (as well as cross-validation) of the smoothness parameter of the Matérn class on subsets of $\mathbb{R}^d$. These results are described next for maximum likelihood estimation. In short, we prove that (a) asymptotic undersmoothing is not possible and (b) smoothness is estimated consistently for a class of compactly supported self-similar functions.

Let f_0 be a real-valued response function that is defined on a sufficiently regular bounded connected open subset Ω of $\mathbb{R}^d$, such as $\Omega = (0,1)^d$, and suppose that $\{x_i\}_{i=1}^\infty$ is any quasi-uniform sequence (see Definition 3.4) of pairwise distinct points in Ω. Let

$$\hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) = \arg\min_{\nu \in \Theta} \big\{ f_0(X_n)^\mathsf{T} K_\nu(X_n)^{-1} f_0(X_n) + \log \det K_\nu(X_n) \big\},$$

where $\Theta \subset (0,\infty)$ is some interval, denote any maximum likelihood estimate of the smoothness parameter ν given the data vector $f_0(X_n) = (f_0(x_1), \ldots, f_0(x_n)) \in \mathbb{R}^n$ consisting of noiseless evaluations of f_0 at the points $X_n = \{x_i\}_{i=1}^n$. Here $K_\nu(X_n) \in \mathbb{R}^{n \times n}$ is the kernel matrix for the Matérn kernel (1.1) with elements $(K_\nu(X_n))_{ij} = K_\nu(x_i, x_j)$. Let $\Theta = [\nu_{\min}, \nu_{\max}]$ for $0 < \nu_{\min} < \nu_{\max} < \infty$ be an interval large enough that the bounds and limits below can hold and suppose that the scaling factor $c(\nu)$ is bounded away from zero and infinity on Θ (e.g., being a continuous function of ν). The function f_0 is an element of $H^\alpha(\Omega)$, the Sobolev space of order $\alpha > d/2$, if it admits an extension $f_e \colon \mathbb{R}^d \to \mathbb{R}$ (i.e., $f_e|_\Omega = f_0$) whose Fourier transform $\widehat{f_e}$ satisfies

$$\int_{\mathbb{R}^d} (1 + \|\xi\|^2)^\alpha |\widehat{f_e}(\xi)|^2 \, \mathrm{d}\xi < \infty.$$

2

See Section 3.1 for more details on Sobolev spaces. We prove that inclusion in a Sobolev space implies an asymptotic lower bound on the maximum likelihood estimate.

No undersmoothing — Theorem 3.11. *Let $\nu_0 > d/2$. If f_0 is an element of $H^{\nu_0}(\Omega)$, then*

$$\liminf_{n \to \infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \geq \nu_0. \tag{1.2}$$

This bound is sharp in the sense that for every $\varepsilon > 0$ there is $f_0 \in H^{\nu_0}(\Omega)$ such that

$$\limsup_{n \to \infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \leq \nu_0 + \varepsilon.$$

Let $\nu(f_0) = \sup\{\nu > 0 \,:\, f_0 \in H^\nu(\Omega)\}$ be the smoothness of f_0. From (1.2) we get

$$\liminf_{n \to \infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \geq \nu(f_0).$$

As satisfying as it would be, it does *not* follow that $\hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \to \nu(f_0)$. In the context of density estimation and the Gaussian white noise model, it is well known that consistent estimation of smoothness and construction of adaptive confidence sets over Sobolev classes is impossible (Picard and Tribouley, 2000; Giné and Nickl, 2010; Bull, 2012; Szabó et al., 2015; Nickl and Szabó, 2016; Giné and Nickl, 2016, Chapter 8). Additional self-similarity assumptions are needed to exclude "inconvenient" or "deceptive" functions whose smoothness cannot be estimated (see in particular Szabó et al., 2015, Section 3). In this vein, we say that f_0 is β-self-similar if it admits an extension f_e such that

$$\sup_{\xi \in \mathbb{R}^d} \|\xi\|^{2\beta+d} |\widehat{f_e}(\xi)|^2 < \infty \quad \text{and} \quad \int_{\|\xi\| \geq R} |\widehat{f_e}(\xi)|^2 \, \mathrm{d}\xi \geq CR^{-2\beta}$$

for some positive C and R_0 and all $R \geq R_0$. The Fourier transform of a prototypical f_e that satisfies these conditions is of order $\|\xi\|^{-(\beta+d/2)}$ as $\|\xi\| \to \infty$. See Section 3.2 for more details on self-similar functions. We prove that maximum likelihood estimation of smoothness is consistent if f_0 is self-similar and supported on Ω. Because $\nu(f_0) = \beta$ if f_0 is β-self-similar (see Lemma 3.2), in our context "consistency" simply means that the true smoothness of f_0 is recovered.

Consistent estimation for self-similar functions — Theorem 3.12. *If f_0 is ν_0-self-similar and has its support contained in Ω, then*

$$\lim_{n \to \infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) = \nu_0.$$

Except for the requirement that f_0 be supported in Ω in the latter theorem, the assumptions of our results are not particularly restrictive. More detailed discussion on the assumptions is deferred to Section 3.3.3.

Because the samples of a Gaussian process with a Matérn covariance kernel of smoothness ν_0 have Sobolev smoothness ν_0 but the reproducing kernel Hilbert space (RKHS) of the kernel is norm-equivalent to the Sobolev space of smoothness $\nu_0 + d/2$ (e.g., Steinwart, 2019; see Sections 3.1 and 3.5 for more details), the results indicate that maximum likelihood estimation recovers the smoothness for which f_0 "resembles" a sample from the corresponding Gaussian process rather than the smoothness for which f_0 is an element of the RKHS of the kernel. When f_0 *is* a zero-mean Gaussian process whose covariance kernel is a Matérn of smoothness ν_0, then Corollary 3.14, a straightforward consequence of Theorem 3.11, states that

$$\liminf_{n \to \infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \geq \nu_0 \quad \text{almost surely.}$$

However, we emphasise that when f_0 is assumed a deterministic function, it does not have to be (nor should it be thought of as) a fixed sample path from a Gaussian process with some

3

Matérn kernel—or in any other way related to some other stochastic process. We also consider leave-one-out cross-validation estimation, for which we however can supply no upper bounds or results pertaining to self-similar functions. Even our lower bounds for cross-validation are likely to be off by $d/2$.

Our proofs make use of RKHSs and techniques from approximation theory in Sobolev spaces. This particular approach has begun to gain popularity in various corners of the Gaussian process literature roughly within the past decade (e.g., Bull, 2011; Stuart and Teckentrup, 2018; Briol et al., 2019; Wang et al., 2020; Wynne et al., 2021). In Section 2, we begin by proving a number of general results (Theorems 2.3, 2.5 and 2.7) on parameter sets which cannot contain the parameter estimates. The essence of these results is that maximum likelihood estimation and cross-validation attempt to find the simplest possible model, as quantified by the rate of decay of the conditional variance, that adequately explains the data. Section 3 is then devoted to applying the general results to estimation of the Matérn smoothness parameter. In Section 4.2, we discuss the application of Theorem 2.7 to estimation of the scale parameter of infinitely smooth stationary kernels, such as the Gaussian kernel, though are unable to furnish any rigorous proofs.

2 General Results

This section reviews basic facts about Gaussian process interpolation and RKHSs and proves some general results on maximum likelihood estimation and cross-validation of covariance kernel parameters.

2.1 Gaussian Process Interpolation

Let Ω be an arbitrary infinite set which we call a *domain* throughout this article. By *kernel* we mean a function $K_\theta \colon \Omega \times \Omega \to \mathbb{R}$ which is *symmetric* and *positive-definite*, in that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K_\theta(x_i, x_j) > 0 \tag{2.1}$$

for any $n \in \mathbb{N}$, any pairwise distinct $x_1, \ldots, x_n \in \Omega$, and any non-zero vector $(a_1, \ldots, a_n) \in \mathbb{R}^n$. All kernels in this article are parametrised by some collection of parameters θ in a feasible parameter set Θ. Equation (2.1) implies that for any set $X = \{x_i\}_{i=1}^{n} \subset \Omega$ of n pairwise distinct points the *kernel matrix* $K_\theta(X) \in \mathbb{R}^{n \times n}$ with elements $(K_\theta(X))_{ij} = K_\theta(x_i, x_j)$ is positive-definite and thus invertible. A Matérn kernel (1.1) with any positive parameters is an example of a kernel on $\Omega = \mathbb{R}^d$.

That a stochastic process f_{GP} is a zero-mean Gaussian process with covariance K_θ implies that for any points X the vector $(f_{\mathrm{GP}}(x_1), \ldots, f_{\mathrm{GP}}(x_n))$ is an n-dimensional normal random vector with mean zero and covariance $K_\theta(X)$. Suppose that a *deterministic* response function $f_0 \colon \Omega \to \mathbb{R}$ is modelled as a Gaussian process f_{GP}. Conditioning this process on the exact evaluations (i.e., data) $f_0(X) = (f_0(x_1), \ldots, f_0(x_n)) \in \mathbb{R}^n$ of f_0 at some distinct points X yields a conditional Gaussian process with the mean

$$\mu_{\theta, f_0}(x \mid X) = \mathbb{E}[f_{\mathrm{GP}}(x) \mid X, f_0(X)] = K_\theta(x, X)^\mathsf{T} K_\theta(X)^{-1} f_0(X) \tag{2.2}$$

and variance

$$\mathbb{V}_\theta(x \mid X) = \mathrm{Var}[f_{\mathrm{GP}}(x) \mid X, f_0(X)] = K_\theta(x, x) - K_\theta(x, X)^\mathsf{T} K_\theta(X)^{-1} K_\theta(x, X), \tag{2.3}$$

where $K_\theta(x, X)$ is an n-vector with elements $(K_\theta(x, X))_i = K_\theta(x, x_i)$. Note that the variance can depend on f_0 only if θ is estimated from the data $f_0(X)$.

2.2 Reproducing Kernel Hilbert Spaces

Every symmetric positive-definite kernel $K_\theta \colon \Omega \times \Omega \to \mathbb{R}$ induces a unique *reproducing kernel Hilbert space* (RKHS), $H(K_\theta)$. This space consists of functions $f \colon \Omega \to \mathbb{R}$ and is equipped with an inner product $\langle \cdot, \cdot \rangle_\theta$ and the associated norm $\|\cdot\|_\theta$. The terminology comes from the kernel K_θ having the *reproducing property*

$$\langle f, K_\theta(\cdot, x)\rangle_\theta = f(x) \quad \text{for all} \quad f \in H(K_\theta) \text{ and } x \in \Omega. \tag{2.4}$$

It may be difficult to determine if a given function is contained in the RKHS based merely on the algebraic form of the kernel and the function. However, many general properties of the kernel, such as its continuity or degree of differentiability, are inherited by the functions in $H(K_\theta)$ (e.g., Steinwart and Christmann, 2008, Section 4.3). Results on the relationship between RKHSs of stationary kernels whose Fourier transforms decay polynomially on $\mathbb{R}^d$ and Sobolev spaces are reviewed in Section 3. See the textbooks Berlinet and Thomas-Agnan (2004) and Paulsen and Raghupathi (2016) for a wealth of additional information on RKHSs.

Most of our proofs rely on the connection between Gaussian process interpolation and optimal interpolation in an RKHS. The history of this rather well known connection goes back at least to the work of Kimeldorf and Wahba (1970). We refer to Berlinet and Thomas-Agnan (2004, Section 2.4); Scheuerer et al. (2013); and Kanagawa et al. (2018, Section 3) for recent reviews on the topic. In short, the Gaussian process conditional mean equals the unique *minimum-norm interpolant* in the RKHS and the conditional variance is the squared *worst-case approximation error*. That is,

$$\mu_{\theta, f_0}(\cdot \mid X) = \operatorname*{arg\,min}_{s \in H(K_\theta)} \big\{ \|s\|_\theta \, : \, s(x_i) = f_0(x_i) \text{ for every } i = 1, \ldots, n \big\} \tag{2.5}$$

and

$$\mathbb{V}_\theta(x \mid X) = \sup_{\|f\|_\theta \le 1} |f(x) - \mu_{\theta, f}(x \mid X)|^2 \tag{2.6}$$

for every $x \in \Omega$. Note that the correspondence (2.5) does not require that f_0 be an element of $H(K_\theta)$. From (2.6) it is straightforward to derive the fundamental error estimate

$$|f(x) - \mu_{\theta, f}(x \mid X)| \le \|f\|_\theta \, \mathbb{V}_\theta(x \mid X)^{1/2}, \tag{2.7}$$

which holds for every $f \in H(K_\theta)$ and $x \in \Omega$.

2.3 Parameter Estimation in a General Setting

Let $\{x_i\}_{i=1}^\infty$ be a set of pairwise distinct points in Ω and denote $X_n = \{x_i\}_{i=1}^n$. Given evaluations of f_0 at points X_n, a *maximum likelihood estimate*, $\hat\theta_{\mathrm{ML}}^{f_0}(X_n)$, of θ is any minimiser of the function

$$\ell_{\mathrm{ML}}^{f_0}(\theta \mid X_n) = f_0(X_n)^\mathsf{T} K_\theta(X_n)^{-1} f_0(X_n) + \log \det K_\theta(X_n), \tag{2.8}$$

while a *leave-one-out cross-validated estimate*, $\hat\theta_{\mathrm{CV}}^{f_0}(X_n)$, is any minimiser of the function

$$\ell_{\mathrm{CV}}^{f_0}(\theta \mid X_n) = \sum_{i=1}^n \left[\frac{(f_0(x_i) - \mu_{\theta, f_0}(x_i \mid X_n^i))^2}{\mathbb{V}_\theta(x_i \mid X_n^i)} + \log \mathbb{V}_\theta(x_i \mid X_n^i) \right],$$

where we use superscripted i to denote that the ith point has been removed (i.e., $X_n^i = X_n \setminus \{x_i\}$); see, for example, Section 5.4 in Rasmussen and Williams (2006). The following lemmas are useful. Here we use the convention $\mathbb{V}_\theta(x \mid X_0) = \mathbb{V}_\theta(x \mid \emptyset) = K_\theta(x, x)$.

Lemma 2.1. *It holds that* $\log \det K_\theta(X_n) = \sum_{i=1}^n \log \mathbb{V}_\theta(x_i \mid X_{i-1})$.

Proof. The claim follows from straightforward iteration of the variance formula (2.3) and the block determinant identity

$$\det \begin{pmatrix} a & b \\ b^{\mathsf{T}} & C \end{pmatrix} = \det(C)(a - b^{\mathsf{T}} C^{-1} b)$$

for any $a \in \mathbb{R}$, $b \in \mathbb{R}^{n-1}$ and any invertible $C \in \mathbb{R}^{(n-1)\times(n-1)}$. $\qquad\square$

Lemma 2.2. *For any $f_0 \colon \Omega \to \mathbb{R}$ we have*

$$f_0(X_n)^{\mathsf{T}} K_\theta(X_n)^{-1} f_0(X_n) = \|\mu_{\theta,f_0}(\cdot \mid X_n)\|_\theta^2. \tag{2.9}$$

Moreover, if $f_0 \in H(K_\theta)$, then

$$f_0(X_n)^{\mathsf{T}} K_\theta(X_n)^{-1} f_0(X_n) = \|\mu_{\theta,f_0}(\cdot \mid X_n)\|_\theta^2 \leq \|f_0\|_\theta^2.$$

Proof. Equation (2.9) follows from the expression for the conditional mean in (2.2) and the fact, which is a consequence of the reproducing property in (2.4), that

$$\left\| \sum_{i=1}^{n} a_i K_\theta(\cdot, x_i) \right\|_\theta^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \langle K_\theta(\cdot, x_i), K_\theta(\cdot, x_j) \rangle_\theta = a^{\mathsf{T}} K_\theta(X_n) a$$

for any $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. The inequality is a consequence of the minimum-norm interpolation property in (2.5) and the fact that f_0 trivially interpolates itself. $\qquad\square$

From the block matrix inversion formula one easily obtains the relatively well known (e.g., Xu and Stein, 2017, Section 4.2.2) expansion

$$f_0(X_n)^{\mathsf{T}} K_\theta(X_n)^{-1} f_0(X_n) = \sum_{i=1}^{n} \frac{(f_0(x_i) - \mu_{\theta,f_0}(x_i \mid X_{i-1}))^2}{\mathbb{V}_\theta(x_i \mid X_{i-1})}. \tag{2.10}$$

Applying (2.10) and Lemma 2.1 to (2.8) shows that the objective functions for maximum likelihood estimation and cross-validation are of similar form. It should therefore be no surprise that the two parameter estimation methods share many properties (an additional interesting connection can be found in Fong and Holmes, 2020).

2.3.1 Lower Bounds

The following theorem yields lower bounds on smoothness parameter estimates.

Theorem 2.3. *Let $\Delta \subset \Theta$ and $\theta_0 \in \Theta$.*

1. *If B is a set of real-valued functions on Ω such that*

$$\limsup_{n\to\infty} \sup_{\theta \in \Delta} \sup_{f_0 \in B} \left[\|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}^2 + \sum_{i=1}^{n} \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_{i-1})}{\mathbb{V}_\theta(x_i \mid X_{i-1})} \right] < 0, \tag{2.11}$$

   then $\hat{\theta}_{\mathrm{ML}}^{f_0}(X_n) \notin \Delta$ for every $f_0 \in B$ when n is sufficiently large.

2. *If B is a set of real-valued functions on Ω such that*

$$\limsup_{n\to\infty} \sup_{\theta \in \Delta} \sup_{f_0 \in B} \sum_{i=1}^{n} \left[\frac{(f_0(x_i) - \mu_{\theta_0,f_0}(x_i \mid X_n^i))^2}{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)} + \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)}{\mathbb{V}_\theta(x_i \mid X_n^i)} \right] < 0, \tag{2.12}$$

   then $\hat{\theta}_{\mathrm{CV}}^{f_0}(X_n) \notin \Delta$ for every $f_0 \in B$ when n is sufficiently large.

Proof. Let us consider maximum likelihood estimation first. Let $\theta \in \Theta$. Lemmas 2.1 and 2.2 yield

$$
\begin{aligned}
\ell_{\mathrm{ML}}^{f_0}(\theta_0 \mid X_n) &= \|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}^2 + \log \det K_{\theta_0}(X_n) \\
&\leq \|\mu_{\theta,f_0}(\cdot \mid X_n)\|_{\theta}^2 + \|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}^2 + \log \det K_{\theta_0}(X_n) \\
&= \|\mu_{\theta,f_0}(\cdot \mid X_n)\|_{\theta}^2 + \|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}^2 + \sum_{i=1}^n \log \mathbb{V}_{\theta_0}(x_i \mid X_{i-1}) \\
&= \|\mu_{\theta,f_0}(\cdot \mid X_n)\|_{\theta}^2 + \|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}^2 \\
&\quad + \sum_{i=1}^n \left[\log \mathbb{V}_{\theta}(x_i \mid X_{i-1}) + \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_{i-1})}{\mathbb{V}_{\theta}(x_i \mid X_{i-1})} \right] \\
&= \ell_{\mathrm{ML}}^{f_0}(\theta \mid X_n) + \|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}^2 + \sum_{i=1}^n \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_{i-1})}{\mathbb{V}_{\theta}(x_i \mid X_{i-1})}.
\end{aligned}
$$

It now follows from (2.11) that $\ell_{\mathrm{ML}}^{f_0}(\theta_0 \mid X_n) < \inf_{\theta \in \Delta} \ell_{\mathrm{ML}}^{f_0}(\theta \mid X_n)$ for every $f_0 \in B$ when n is sufficiently large, which means that the maximum likelihood estimate, being a minimiser of $\ell_{\mathrm{ML}}^{f_0}(\cdot \mid X_n)$, must be outside of Δ for all $f_0 \in B$ when n is sufficiently large. The proof for cross-validation is analogous. $\square$

The role of θ_0 in Theorem 2.3 is somewhat subtle. For $\theta = \theta_0$ the logarithmic terms in (2.11) and (2.12) are non-negative. Because the non-logarithmic terms are always non-negative, this means that θ_0 cannot be an element of Δ. Therefore one should select θ_0 such that the set Δ can be made as large as possible. When we consider smoothness estimation for Matérns in Section 3, assumptions (2.11) and (2.12) are verified by taking θ_0 and Δ such that [specifically, see (3.25) and (3.26)]

$$
\lim_{n \to \infty} \sup_{\theta \in \Delta} \frac{\mathbb{V}_{\theta_0}(x_n \mid X_{n-1})}{\mathbb{V}_{\theta}(x_n \mid X_{n-1})} = 0, \tag{2.13}
$$

which states that the conditional variance should decay faster for the parameter θ_0 than for any parameter $\theta \in \Delta$. Because the conditional variance is the supremum over the unit ball of the RKHS by the equivalence in (2.6), the limit in (2.13) implies that $H(K_{\theta_0})$ is *essentially* smaller as a set than $H(K_{\theta})$ for any $\theta \in \Delta$. Note that (2.13) is not necessary for (2.11) and (2.12) to hold. For example, RKHSs of two Matérn kernels in (1.1) with any positive scale parameters λ_1 and λ_2 are norm-equivalent; see (3.5) and (3.6). If one is estimating the scale parameter and Δ is an interval bounded away from zero and infinity, the ratio in (2.13) cannot tend to zero.

For maximum likelihood estimation it is not necessary to use Lemma 2.1 to decompose the determinant as a sum of variances. By writing

$$
\log \det K_{\theta_0}(X_n) = \log \det K_{\theta}(X_n) + \log \det \left[K_{\theta_0}(X_n) K_{\theta}(X_n)^{-1} \right],
$$

we could have replaced (2.11) with the equivalent condition

$$
\limsup_{n \to \infty} \sup_{\theta \in \Delta} \sup_{f_0 \in B} \left[\|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}^2 + \log \frac{\det K_{\theta_0}(X_n)}{\det K_{\theta}(X_n)} \right] < 0. \tag{2.14}
$$

Because the kernel matrix determinant is known as the *model complexity* and the *data-fit term* $\|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}$ quantifies how well the model fits the data, the condition (2.14) gives the following interpretation of Theorem 2.3 for maximum likelihood estimation: If there is a parameter θ_0 such that the model for this parameter fits the data sufficiently well (i.e., the data-fit is bounded or grows slowly) and each parameter in Δ corresponds to a model more complex than that for θ_0 (i.e., the log-ratio of model complexities is sufficiently small or tends to negative

infinity sufficiently fast), then the parameter estimate cannot be contained in Δ. That is, maximum likelihood estimation prefers simple models that fit the data well.

The following corollary is a specialisation of Theorem 2.3 to a setting where Θ is an interval, which we take to be any connected subset of $\mathbb{R}$, and θ can be thought of as a smoothness parameter, so that $H(K_{\theta_1}) \subsetneq H(K_{\theta_2})$ whenever $\theta_1 > \theta_2$.[1] Under suitable conditions, the implication $\theta_0 \notin \Delta$ may then be expressed as an inequality that provides an asymptotic lower bound on the smoothness estimates.

Corollary 2.4. *Let $\Theta \subset \mathbb{R}$ be an interval and $\theta_0 \in \Theta$.*

1. If B is a set of real-valued functions on Ω such that

$$\limsup_{n \to \infty} \sup_{\theta \leq \theta_1} \sup_{f_0 \in B} \left[\|\mu_{\theta_0, f_0}(\cdot \mid X_n)\|_{\theta_0}^2 + \sum_{i=1}^{n} \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_{i-1})}{\mathbb{V}_{\theta}(x_i \mid X_{i-1})} \right] < 0$$

for every $\theta_1 < \theta_0$, then

$$\liminf_{n \to \infty} \inf_{f_0 \in B} \hat{\theta}_{\mathrm{ML}}^{f_0}(X_n) \geq \theta_0.$$

2. If B is a set of real-valued functions on Ω such that

$$\limsup_{n \to \infty} \sup_{\theta \leq \theta_1} \sup_{f_0 \in B} \sum_{i=1}^{n} \left[\frac{(f_0(x_i) - \mu_{\theta_0, f_0}(x_i \mid X_n^i))^2}{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)} + \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)}{\mathbb{V}_{\theta}(x_i \mid X_n^i)} \right] < 0$$

for every $\theta_1 < \theta_0$, then

$$\liminf_{n \to \infty} \inf_{f_0 \in B} \hat{\theta}_{\mathrm{CV}}^{f_0}(X_n) \geq \theta_0.$$

Proof. Let Θ be an interval with endpoints $a_\Theta \leq b_\Theta$ that are possibly infinite. The claim follows by applying Theorem 2.3 to $\Delta = \Theta \cap [a_\Theta, \theta_1]$ for each $\theta_1 < \theta_0$ and using the definition of the lower limit. □

In Section 2.4 we shall assume that B is a subset of $H(K_{\theta_0})$, which simplifies the role of θ_0 but renders the results somewhat sub-optimal.

2.3.2 Upper Bounds

The following theorem yields upper bounds on smoothness parameter estimates.

Theorem 2.5. *Let $\Sigma \subset \Theta$ and suppose that B is a bounded subset of $H(K_{\theta_B})$ for some $\theta_B \in \Theta$.*

1. If

$$\limsup_{n \to \infty} \mathbb{V}_{\theta_B}(x_n \mid X_{n-1}) < 1 \tag{2.15}$$

and

$$\liminf_{n \to \infty} \inf_{\theta \in \Sigma} \inf_{f_0 \in B} \ell_{\mathrm{ML}}^{f_0}(\theta \mid X_n) > -\infty, \tag{2.16}$$

then $\hat{\theta}_{\mathrm{ML}}^{f_0}(X_n) \notin \Sigma$ for every $f_0 \in B$ when n is sufficiently large.

2. Let $b = \sup_{f_0 \in B} \|f_0\|_{\theta_B}$. If

$$\limsup_{n \to \infty} \max_{1 \leq i \leq n} \mathbb{V}_{\theta_B}(x_i \mid X_n^i) < \exp(-b^2) \tag{2.17}$$

and

$$\liminf_{n \to \infty} \inf_{\theta \in \Sigma} \inf_{f_0 \in B} \ell_{\mathrm{CV}}^{f_0}(\theta \mid X_n) > -\infty, \tag{2.18}$$

then $\hat{\theta}_{\mathrm{CV}}^{f_0}(X_n) \notin \Sigma$ for every $f_0 \in B$ when n is sufficiently large.

---

[1] See Gualtierotti (2015, Section 3.2) for a general treatment of such *contractive inclusions* of RKHSs.

Proof. Let $b = \sup_{f_0 \in B} \|f_0\|_{\theta_B} < \infty$. Let us consider maximum likelihood estimation first. By Lemma 2.2 and (2.15),

$$\ell_{\text{ML}}^{f_0}(\theta_B \mid X_n) = \|\mu_{\theta_B, f_0}(\cdot \mid X_n)\|_{\theta_B}^2 + \sum_{i=1}^n \log \mathbb{V}_{\theta_B}(x_i \mid X_{i-1}) \le b^2 + \sum_{i=1}^n \log \mathbb{V}_{\theta_B}(x_i \mid X_{i-1})$$

tends to negative infinity as $n \to \infty$ uniformly over $f_0 \in B$. It thus follows from (2.16) that

$$\sup_{f_0 \in B} \ell_{\text{ML}}^{f_0}(\theta_B \mid X_n) < \inf_{\theta \in \Sigma} \inf_{f_0 \in B} \ell_{\text{ML}}^{f_0}(\theta \mid X_n)$$

for all sufficiently large n, which gives the claim for maximum likelihood estimation. The proof for cross-validation is analogous, except that now we use (2.7) to get

$$\ell_{\text{CV}}^{f_0}(\theta_B \mid X_n) = \sum_{i=1}^n \left[\frac{(f_0(x_i) - \mu_{\theta_B, f_0}(x_i \mid X_n^i))^2}{\mathbb{V}_{\theta_B}(x_i \mid X_n^i)} + \log \mathbb{V}_{\theta_B}(x_i \mid X_n^i) \right]$$

$$\le \sum_{i=1}^n \left[b^2 + \log \mathbb{V}_{\theta_B}(x_i \mid X_n^i) \right],$$

which tends to negative infinity as $n \to \infty$ by (2.17). $\square$

The assumptions (2.15) and (2.17) usually hold. For instance, if Ω is a compact metric space, the sequence $\{x_i\}_{i=1}^\infty$ is dense in Ω, and K_θ is continuous, then $\sup_{x \in \Omega} \mathbb{V}_\theta(x \mid X_n) \to 0$ as $n \to \infty$. That (2.17) can be likely improved somewhat is discussed later in Remark 2.8. The gist of Theorem 2.5 is in the interplay between θ_B and Σ: By (2.15) and (2.17), both $\ell_{\text{ML}}^{f_0}(\theta_B \mid X_n)$ and $\ell_{\text{CV}}^{f_0}(\theta_B \mid X_n)$ tend to negative infinity, which may be interpreted as θ_B being a plausible parameter estimate. The assumptions (2.16) and (2.18) then state that no parameter in Σ is plausible, which limits the size of Δ by prohibiting $\theta \in \Delta$ such that $H(K_{\theta_B}) \subset H(K_\theta)$. For if this inclusion were true and $\mathbb{V}_\theta(x_n \mid X_{n-1})$ tended to zero, both

$$\ell_{\text{ML}}^{f_0}(\theta \mid X_n) = \|\mu_{\theta, f_0}(\cdot \mid X_n)\|_\theta^2 + \sum_{i=1}^n \log \mathbb{V}_\theta(x_i \mid X_{i-1}) \le \|f_0\|_\theta^2 + \sum_{i=1}^n \log \mathbb{V}_\theta(x_i \mid X_{i-1})$$

and $\ell_{\text{CV}}^{f_0}(\theta \mid X_n)$ would tend to negative infinity, thus violating (2.16) and (2.18). The following corollary provides a version of Theorem 2.5 adapted to intervals and complements Corollary 2.4.

Corollary 2.6. *Let $\Theta \subset \mathbb{R}$ be an interval and $\theta_0 \in \Theta$. Suppose that B is a bounded subset of $H(K_{\theta_B})$ for some $\theta_B \in \Theta$.*

1. *If*

$$\limsup_{n \to \infty} \mathbb{V}_{\theta_B}(x_n \mid X_{n-1}) < 1 \quad \text{and} \quad \liminf_{n \to \infty} \inf_{\theta \ge \theta_1} \inf_{f_0 \in B} \ell_{\text{ML}}^{f_0}(\theta \mid X_n) > -\infty,$$

   for every $\theta_1 > \theta_0$, then
$$\limsup_{n \to \infty} \sup_{f_0 \in B} \hat{\theta}_{\text{ML}}^{f_0}(X_n) \le \theta_0.$$

2. *Let $b = \sup_{f_0 \in B} \|f_0\|_{\theta_B}$. If*

$$\limsup_{n \to \infty} \max_{1 \le i \le n} \mathbb{V}_{\theta_B}(x_i \mid X_n^i) < \exp(-b^2) \quad \text{and} \quad \liminf_{n \to \infty} \inf_{\theta \ge \theta_1} \inf_{f_0 \in B} \ell_{\text{CV}}^{f_0}(\theta \mid X_n) > -\infty,$$

   for every $\theta_1 > \theta_0$, then
$$\limsup_{n \to \infty} \sup_{f_0 \in B} \hat{\theta}_{\text{CV}}^{f_0}(X_n) \le \theta_0.$$

Unlike the assumption that $B \subset H(K_{\theta_0})$ in the next section, the purpose of the assumption in Theorem 2.5 and Corollary 2.6 that B be a subset of *some* RKHS is only to guarantee that the objective functions do not tend to negative infinity for all possible parameters.

9

2.4 Parameter Estimation in an RKHS Setting

By assuming that $B \subset H(K_{\theta_0})$ we obtain a weaker version of Theorem 2.3.

Theorem 2.7. *Let $\Delta \subset \Theta$ and $\theta_0 \in \Theta$. Suppose that B is a bounded subset of $H(K_{\theta_0})$.*

1. If

$$\limsup_{n \to \infty} \sup_{\theta \in \Delta} \frac{\mathbb{V}_{\theta_0}(x_n \mid X_{n-1})}{\mathbb{V}_\theta(x_n \mid X_{n-1})} < 1, \tag{2.19}$$

then $\hat{\theta}_{\mathrm{ML}}(X_n) \notin \Delta$ for every $f_0 \in B$ when n is sufficiently large.

2. Let $b = \sup_{f_0 \in B} \|f_0\|_{\theta_0}$. If

$$\limsup_{n \to \infty} \sup_{\theta \in \Delta} \max_{1 \le i \le n} \frac{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)}{\mathbb{V}_\theta(x_i \mid X_n^i)} < \exp(-b^2), \tag{2.20}$$

then $\hat{\theta}_{\mathrm{CV}}(X_n) \notin \Delta$ for every $f_0 \in B$ when n is sufficiently large.

Proof. Let $b = \sup_{f_0 \in B} \|f_0\|_{\theta_0} < \infty$. Let us consider maximum likelihood estimation first. Because $B \subset H(K_{\theta_0})$, we get from Lemma 2.2 that

$$\sup_{\theta \in \Delta} \sup_{f_0 \in B} \left[\|\mu_{\theta_0, f_0}(\cdot \mid X_n)\|_{\theta_0}^2 + \sum_{i=1}^n \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_{i-1})}{\mathbb{V}_\theta(x_i \mid X_{i-1})} \right]$$
$$\le b^2 + \sum_{i=1}^n \sup_{\theta \in \Delta} \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_{i-1})}{\mathbb{V}_\theta(x_i \mid X_{i-1})},$$

which tends to negative infinity as $n \to \infty$ by (2.19). Therefore (2.11) holds, so that the claim follows from Theorem 2.3. Let us then consider cross-validation. Because $B \subset H(K_{\theta_0})$, we may use (2.7) to get

$$\sup_{\theta \in \Delta} \sum_{i=1}^n \left[\frac{(f_0(x_i) - \mu_{\theta_0, f_0}(x_i \mid X_n^i))^2}{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)} + \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)}{\mathbb{V}_\theta(x_i \mid X_n^i)} \right]$$
$$\le \sum_{i=1}^n \left[b^2 + \sup_{\theta \in \Delta} \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)}{\mathbb{V}_\theta(x_i \mid X_n^i)} \right] \le n \left[b^2 + \sup_{\theta \in \Delta} \max_{1 \le i \le n} \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)}{\mathbb{V}_\theta(x_i \mid X_n^i)} \right],$$

which tends to negative infinity as $n \to \infty$ by (2.20). Therefore (2.12) holds, so that the claim follows from Theorem 2.3. $\square$

Remark 2.8. Suppose for simplicity that $B = \{f_0\}$. It is likely that (2.20), and similarly (2.17), can be improved to requiring simply that the upper limit be less than one. For had we used (2.10) and (2.7) and proceeded as we did in the case of cross-validation, we would have arrived at the similar assumption

$$\limsup_{n \to \infty} \sup_{\theta \in \Delta} \frac{\mathbb{V}_{\theta_0}(x_n \mid X_{n-1})}{\mathbb{V}_\theta(x_n \mid X_{n-1})} < \exp(-\|f_0\|_{\theta_0}^2)$$

for maximum likelihood estimation from

$$\sup_{\theta \in \Delta} \left[\|\mu_{\theta_0, f_0}(\cdot \mid X_n)\|_{\theta_0}^2 + \sum_{i=1}^n \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_{i-1})}{\mathbb{V}_\theta(x_i \mid X_{i-1})} \right]$$
$$= \sup_{\theta \in \Delta} \sum_{i=1}^n \left[\frac{(f_0(x_i) - \mu_{\theta, f_0}(x_i \mid X_{i-1}))^2}{\mathbb{V}_\theta(x_i \mid X_{i-1})} + \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_{i-1})}{\mathbb{V}_\theta(x_i \mid X_{i-1})} \right]$$
$$\le n \|f_0\|_{\theta_0}^2 + \sum_{i=1}^n \sup_{\theta \in \Delta} \log \frac{\mathbb{V}_{\theta_0}(x_i \mid X_{i-1})}{\mathbb{V}_\theta(x_i \mid X_{i-1})}.$$

This indicates that using (2.7) should be avoided. It is straightforward to improve (2.7) to $|f(x) - \mu_{\theta,f}(x \mid X)| \leq \|f - \mu_{\theta,f}(\cdot \mid X)\|_\theta \, \mathbb{V}_\theta(x \mid X)^{1/2}$. However, controlling the RKHS norm $\|f - \mu_{\theta,f}(\cdot \mid X)\|_\theta$ is challenging.

To see that Theorem 2.7 is weaker than Theorem 2.3, observe that assumptions (2.11) and (2.12) can hold even when

$$\|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}^2 \to \infty \quad \text{or} \quad \frac{(f_0(x_i) - \mu_{\theta_0,f_0}(x_i \mid X_n^i))^2}{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)} \to \infty,$$

which, as we saw in the proof of Theorem 2.7, cannot happen if $B \subset H(K_{\theta_0})$. This weakness of Theorem 2.7 is explained by the fact that $B \subset H(K_{\theta_0})$ is a "wrong" assumption to make because the samples of a Gaussian process with covariance kernel K_θ are *not* elements of $H(K_\theta)$ but of a somewhat larger RKHS (Driscoll 1973; we discuss this more in Sections 2.5 and 3.5). That is, maximum likelihood estimation and cross-validation do not attempt to find θ_0 such that $B \subset H(K_{\theta_0})$ but θ_0 for which the elements of B resemble, in some sense, the samples of a Gaussian process with covariance kernel K_{θ_0}. We shall see this phenomenon in action in Section 3 because for Matérn kernels on $\mathbb{R}^d$ the samples have $d/2$ orders of smoothness less than the RKHS. By applying Theorem 2.7 to the setting where Θ is an interval we obtain the following corollary.

Corollary 2.9. *Let $\Theta \subset \mathbb{R}$ be an interval and $\theta_0 \in \Theta$. Suppose that B is a bounded subset of $H(K_{\theta_0})$.*

1. If

$$\limsup_{n \to \infty} \sup_{\theta \leq \theta_1} \frac{\mathbb{V}_{\theta_0}(x_n \mid X_{n-1})}{\mathbb{V}_\theta(x_n \mid X_{n-1})} < 1 \quad \text{for every} \quad \theta_1 < \theta_0, \tag{2.21}$$

then

$$\liminf_{n \to \infty} \inf_{f_0 \in B} \hat{\theta}_{\mathrm{ML}}^{f_0}(X_n) \geq \theta_0.$$

2. Let $b = \sup_{f_0 \in B} \|f_0\|_{\theta_0}$. If

$$\limsup_{n \to \infty} \sup_{\theta \leq \theta_1} \max_{1 \leq i \leq n} \frac{\mathbb{V}_{\theta_0}(x_i \mid X_n^i)}{\mathbb{V}_\theta(x_i \mid X_n^i)} < \exp(-b^2) \quad \text{for every} \quad \theta_1 < \theta_0, \tag{2.22}$$

then

$$\liminf_{n \to \infty} \inf_{f_0 \in B} \hat{\theta}_{\mathrm{CV}}^{f_0}(X_n) \geq \theta_0.$$

2.5 On Driscoll's Theorem

The determinantal condition (2.14) has a connection to sample path properties of Gaussian processes that is worth elucidating. Because the product of two positive-definite matrices has positive eigenvalues, the inequality of arithmetic and geometric means yields

$$\det\left[K_{\theta_0}(X_n)K_\theta(X_n)^{-1}\right] \leq \left(\frac{1}{n}\mathrm{tr}\left[K_{\theta_0}(X_n)K_\theta(X_n)^{-1}\right]\right)^n. \tag{2.23}$$

From (2.14) and (2.23) we obtain the following variant of Theorem 2.7 for maximum likelihood estimation.

Theorem 2.10. *Let $\Delta \subset \Theta$ and $\theta_0 \in \Theta$. If B is a set of real-valued functions on Ω such that*

$$\limsup_{n \to \infty} \sup_{\theta \in \Delta} \sup_{f_0 \in B} \left[\|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}^2 + n \log\left(\frac{1}{n}\mathrm{tr}\left[K_{\theta_0}(X_n)K_\theta(X_n)^{-1}\right]\right)\right] < 0, \tag{2.24}$$

then $\hat{\theta}_{\mathrm{ML}}^{f_0}(X_n) \notin \Delta$ for every $f_0 \in B$ when n is sufficiently large.

11

Suppose that Ω is a separable metric space, that $\{x_i\}_{i=1}^\infty$ is dense in Ω, and that the kernels K_{θ_0} and K_θ are continuous. Driscoll (1973, Theorem 3) has proved that, under certain additional assumptions,

$$\lim_{n\to\infty} \text{tr}\big[K_{\theta_0}(X_n)K_\theta(X_n)^{-1}\big] < \infty \tag{2.25}$$

if and only if the samples from a Gaussian process with covariance kernel K_{θ_0} are contained in $H(K_\theta)$ with probability one. In particular, setting $\theta = \theta_0$ shows that the samples are not contained in the RKHS of K_{θ_0}. See Lukić and Beder (2001) for the equivalence of (2.25) to a *nuclear dominance* condition between K_θ and K_{θ_0}. If the set B is such that $\|\mu_{\theta_0,f_0}(\cdot \mid X_n)\|_{\theta_0}^2$ does not grow faster than linearly in n, which is the case if B is a bounded subset of $H(K_{\theta_0})$, then (2.24) is implied by the following version of Driscoll's condition (2.25) that is uniform over $\theta \in \Delta$:

$$\limsup_{n\to\infty} \sup_{\theta\in\Delta} \text{tr}\big[K_{\theta_0}(X_n)K_\theta(X_n)^{-1}\big] < C < \infty \tag{2.26}$$

for some $C > 0$. If (2.26) holds, then

$$\limsup_{n\to\infty} \sup_{\theta\in\Delta} \left[n\log\left(\frac{1}{n}\text{tr}\big[K_{\theta_0}(X_n)K_\theta(X_n)^{-1}\big]\right)\right] \le n\log\frac{C}{n} = n\log C - n\log n,$$

which implies (2.24) if the RKHS norm is assumed to grow at most linearly.

3   Smoothness Estimation for Matérns

In this section we apply the results of Section 2 to estimation of the smoothness parameter ν of the Matérn class in (1.1). We fix the positive magnitude and scale parameters σ and λ and write the Matérn kernel of smoothness $\nu > 0$ as

$$K_\nu(x,y) = \Phi_\nu(x-y),$$

where the function

$$\Phi_\nu(z) = \sigma^2 c(\nu)\left(\frac{\sqrt{2\nu}\,\|z\|}{\lambda}\right)^\nu \mathcal{K}_\nu\left(\frac{\sqrt{2\nu}\,\|z\|}{\lambda}\right) \tag{3.1}$$

is defined on $\mathbb{R}^d$. Recall that $\mathcal{K}_\nu$ is the modified Bessel function of the second kind of order ν. In most results of this section we shall consider smoothness estimation over the bounded interval $\Theta = [\nu_{\min}, \nu_{\max}]$ for $0 < \nu_{\min} \le \nu_{\max} < \infty$ and employ the following assumptions on the positive scaling factor $c(\nu)$:

$$\inf_{\nu\in[\nu_{\min},\nu_{\max}]} c(\nu) > 0, \tag{3.2a}$$

and

$$\sup_{\nu\in[\nu_{\min},\nu_{\max}]} c(\nu) < \infty. \tag{3.2b}$$

These assumptions hold if the scaling factor is a continuous function of ν. Let Γ denote the Gamma function. The scaling factor $c(\nu) = 2^{1-\nu}/\Gamma(\nu)$, which is typically used because it ensures that K_ν tends pointwise to the Gaussian kernel

$$K(x,y) = \sigma^2 \exp\left(-\frac{\|x-y\|^2}{2\lambda^2}\right)$$

as $\nu \to \infty$ (Stein, 1999, pp. 49–50), is obviously continuous.

We use $\lesssim_n$ (with $\gtrsim_n$ defined analogously) to denote an inequality that holds up to a constant factor for all $n \in \mathbb{N}$. That is, $a_n \lesssim_n b_n$ means that there is a non-negative constant C such that $a_n \le Cb_n$ for all n. We write $a_n \asymp_n b_n$ if $a_n \lesssim_n b_n$ and $a_n \gtrsim_n b_n$.

3.1 Sobolev Spaces

The Fourier transform of $f \in L^2(\mathbb{R}^d)$ is defined as $\widehat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-ix^\mathsf{T}\xi} \, \mathrm{d}x$. For $\alpha > 0$, the *Sobolev space* $H^\alpha(\mathbb{R}^d)$ consists of functions $f \in L^2(\mathbb{R}^d)$ such that

$$\|f\|^2_{H^\alpha(\mathbb{R}^d)} = \int_{\mathbb{R}^d} (1 + \|\xi\|^2)^\alpha |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi < \infty. \tag{3.3}$$

By the Sobolev embedding theorem, assuming that $\alpha > d/2$ ensures that every element of $H^\alpha(\mathbb{R}^d)$ can be uniquely identified with a continuous function, in which case $H^\alpha(\mathbb{R}^d)$ can be interpreted as a space of functions rather than of their equivalence classes. If α is an integer, the Sobolev space consists of functions whose weak derivatives up to order α exist and are in $L^2(\mathbb{R}^d)$. Moreover, every function in $H^\alpha(\mathbb{R}^d)$ is $\lfloor \alpha - d/2 \rfloor$ times differentiable in the classical sense. On a subset Ω of $\mathbb{R}^d$ the Sobolev space $H^\alpha(\Omega)$ is defined as the set of functions $f \colon \Omega \to \mathbb{R}$ for which there exists an *extension* $f_e \in H^\alpha(\mathbb{R}^d)$ such that $f_e|_\Omega = f$. The norm of $H^\alpha(\Omega)$ is

$$\|f\|_{H^\alpha(\Omega)} = \min\{\|f_e\|_{H^\alpha(\mathbb{R}^d)} \,:\, f_e \in H^\alpha(\mathbb{R}^d) \text{ and } f_e|_\Omega = f\}. \tag{3.4}$$

We shall formulate all our auxiliary results in $H^\alpha(\mathbb{R}^d)$ and use $H^\alpha(\Omega)$ only in what we consider to be main results of this section.

The RKHS of any stationary kernel of the form $K_\theta(x, y) = \Phi_\theta(x - y)$ for an integrable and continuous $\Phi_\theta \colon \mathbb{R}^d \to \mathbb{R}$ can be expressed in terms of the Fourier transform of Φ_θ (Wendland, 2005, Theorem 10.12). Namely, $H(K_\theta)$ contains those square-integrable and continuous functions $f \colon \mathbb{R}^d \to \mathbb{R}$ for which

$$\|f\|^2_\theta = \int_{\mathbb{R}^d} \frac{|\widehat{f}(\xi)|^2}{\widehat{\Phi}_\theta(\xi)} \, \mathrm{d}\xi < \infty. \tag{3.5}$$

The function Φ_ν in (3.1), which defines the Matérn class, has the Fourier transform

$$\widehat{\Phi}_\nu(\xi) = \frac{1}{C_\nu} \left(\frac{2\nu}{\lambda^2} + \|\xi\|^2 \right)^{-(\nu + d/2)} \qquad \text{with} \qquad C_\nu = \frac{\pi^{d/2}}{\sigma^2 c(\nu) 2^{\nu - 1} \Gamma(\nu + d/2)} \left(\frac{\lambda^2}{2\nu} \right)^\nu. \tag{3.6}$$

Therefore the norm $\|\cdot\|_\nu$ of the Matérn RKHS $H(K_\nu)$ is

$$\|f\|^2_\nu = C_\nu \int_{\mathbb{R}^d} \left(\frac{2\nu}{\lambda^2} + \|\xi\|^2 \right)^{\nu + d/2} |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi. \tag{3.7}$$

It is straightforward to compute that

$$C_\nu \|f\|^2_{H^{\nu + d/2}(\mathbb{R}^d)} \min\{1, b_\nu\} \leq \|f\|^2_\nu \leq C_\nu \max\{1, b_\nu\} \|f\|^2_{H^{\nu + d/2}(\mathbb{R}^d)}, \tag{3.8}$$

where $b_\nu = (2\nu/\lambda^2)^{\nu + d/2}$. This shows (as is well known) that $H(K_\nu)$ is *norm-equivalent* to the Sobolev space $H^{\nu + d/2}(\mathbb{R}^d)$, which is to say that $H(K_\nu) = H^{\nu + d/2}(\mathbb{R}^d)$ as sets and the norms $\|\cdot\|_\nu$ and $\|\cdot\|_{H^{\nu + d/2}(\mathbb{R}^d)}$ are equivalent.

3.2 Self-Similar Functions

Lower bounds on smoothness parameter estimates that we prove apply to any Sobolev function. But to obtain upper bounds we need to work with a class of self-similar functions that we define as follows.

Definition 3.1. Let $\beta > 0$. We say that a function $f \in L^2(\mathbb{R}^d)$ is β-*self-similar* if

$$\sup_{\xi \in \mathbb{R}^d} \|\xi\|^{2\beta + d} |\widehat{f}(\xi)|^2 < \infty \tag{3.9}$$

13

and there are positive constants C and R_0 such that

$$\int_{\|\xi\| \geq R} |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi \geq C R^{-2\beta} \tag{3.10}$$

for all $R \geq R_0$. A function on Ω is β-self-similar if it has a β-self-similar extension.

As discussed in the introduction, self-similarity assumptions are routinely used in the literature on non-parametric statistics to exclude "inconvenient" or "deceptive" functions whose smoothness cannot be estimated (Bull, 2012; Szabó et al., 2015; Nickl and Szabó, 2016). See in particular Section 3 in Szabó et al. (2015). Functions such that

$$C_1 \|\xi\|^{-(\beta+d/2)} \leq |\widehat{f}(\xi)| \leq C_2 \|\xi\|^{-(\beta+d/2)}$$

for some positive C_1 and C_2 and all ξ outside some ball centered at the origin are prototypical examples of β-self-similar functions because

$$\sup_{\xi \in \mathbb{R}^d} \|\xi\|^{2\beta+d} |\widehat{f}(\xi)|^2 \leq C_2^2$$

and

$$\int_{\|\xi\| \geq R} |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi \geq C_1^2 \int_{\|\xi\| \geq R} \|\xi\|^{-2\beta-d} \, \mathrm{d}\xi = C_1^2 C_d R^{-2\beta} \tag{3.11}$$

for a certain positive constant C_d.

Lemma 3.2. *If $f \in L^2(\mathbb{R}^d)$ is β-self-similar, then $f \in H^\alpha(\mathbb{R}^d)$ if $\alpha < \beta$ and $f \notin H^\alpha(\mathbb{R}^d)$ if $\alpha > \beta$.*

Proof. By (3.9), there is a non-negative constant c such that

$$\|f\|_{H^\alpha(\mathbb{R}^d)}^2 = \int_{\mathbb{R}^d} (1 + \|\xi\|^2)^\alpha |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi \leq c \int_{\mathbb{R}^d} (1 + \|\xi\|^2)^\alpha \|\xi\|^{-2\beta-d} \, \mathrm{d}\xi,$$

which is finite if $\alpha < \beta$. Therefore a β-self-similar function is in $H^\alpha(\mathbb{R}^d)$ for every $\alpha < \beta$. On the other hand, from (3.10) we get

$$\begin{aligned}
\|f\|_{H^\alpha(\mathbb{R}^d)}^2 = \int_{\mathbb{R}^d} (1 + \|\xi\|^2)^\alpha |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi &\geq \int_{\|\xi\| \geq R} (1 + \|\xi\|^2)^\alpha |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi \\
&\geq 2^\alpha R^{2\alpha} \int_{\|\xi\| \geq R} |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi \\
&\geq C 2^\alpha R^{2(\alpha-\beta)}
\end{aligned}$$

for every $R \geq \max\{1, R_0\}$. Therefore $f \notin H^\alpha(\mathbb{R}^d)$ if $\alpha > \beta$. $\qquad\square$

We shall work with self-similar functions that are supported in a given open set. Let $\Omega \subset \mathbb{R}^d$ be an open set and define

$$H_{\mathrm{ss}}^\beta(\Omega) = \{f \in L^2(\mathbb{R}^d) \, : \, f \text{ is } \beta\text{-self-similar and the support of } f \text{ is contained in } \Omega\}.$$

It seems likely that the requirement that the functions be supported in Ω is not necessary in our results; see the discussion in Section 3.3.3. The following basic construction shows that $H_{\mathrm{ss}}^\beta(\Omega)$ is non-empty.

Lemma 3.3. *Let $\Omega \subset \mathbb{R}^d$ be an open set. For every $\beta > d/2$ there is $f \in L^2(\mathbb{R}^d)$ such that the support of f is contained in Ω and*

$$C_1(1 + \|\xi\|^2)^{-(\beta+d/2)} \leq |\widehat{f}(\xi)|^2 \leq C_2(1 + \|\xi\|^2)^{-(\beta+d/2)} \tag{3.12}$$

for some positive C_1 and C_2 and all $\xi \in \mathbb{R}^d$. This function is in $H_{\mathrm{ss}}^\beta(\Omega)$ and an element of $H^\alpha(\mathbb{R}^d)$ if and only if $\alpha < \beta$.

Proof. We may assume without loss of generality that Ω is any open ball. By first setting $\nu = \beta/2 - d/4 > 0$ and then selecting σ and λ properly in (3.1) and (3.6), we obtain a function g with the Fourier transform

$$\widehat{g}(\xi) = (1 + \|\xi\|^2)^{-(\beta+d/2)/2}.$$

There exists a non-negative *bump function* $\phi \colon \mathbb{R}^d \to \mathbb{R}$ that (i) is supported on the unit ball, (ii) satisfies $\sup_{x\in\mathbb{R}^d} \phi(x) = \phi(0) = 1$ and (iii) is infinitely differentiable and hence has a Fourier transform which decays faster than any polynomial. The standard example of such a bump function is given by

$$\phi(x) = \exp\left(-\frac{1}{1 - \|x\|^2}\right) \text{ if } \|x\| < 1 \quad\text{and}\quad \phi(x) = 0 \text{ if } \|x\| \geq 1. \tag{3.13}$$

Because ϕ is radial, its Fourier transform is real-valued and therefore the function $\phi_2 = \phi * \phi$ has non-negative Fourier transform by the convolution theorem. Because ϕ is supported on the unit ball, so are ϕ_2 and $f := \phi_2 g$. It remains to show that f satisfies other requirements in the lemma.

First, because the Fourier transform of ϕ decays faster than any polynomial, there is $C_\phi > 0$ such that

$$\widehat{\phi_2}(\xi) = \widehat{\phi}(\xi)^2 \leq C_\phi (1 + \|\xi\|^2)^{-(\beta+d/2)/2}$$

for all $\xi \in \mathbb{R}^d$. From $\|\xi\| \leq \|\xi - \omega\| + \|\omega\|$ it follows that

$$
\begin{aligned}
|\widehat{f}(\xi)| &= \int_{\mathbb{R}^d} \widehat{\phi_2}(\omega) \widehat{g}(\xi - \omega)\, d\omega \\
&\leq C_\phi \int_{\mathbb{R}^d} (1 + \|\omega\|^2)^{-(\beta+d/2)/2}(1 + \|\xi - \omega\|^2)^{-(\beta+d/2)/2}\, d\omega \\
&\leq C_\phi \int_{\|\xi-\omega\| \leq \frac{1}{2}\|\xi\|} \left(1 + \frac{\|\xi\|^2}{4}\right)^{-(\beta+d/2)/2}(1 + \|\xi - \omega\|^2)^{-(\beta+d/2)/2}\, d\omega \\
&\quad + C_\phi \int_{\|\xi-\omega\| \geq \frac{1}{2}\|\xi\|} (1 + \|\omega\|^2)^{-(\beta+d/2)/2}\left(1 + \frac{\|\xi\|^2}{4}\right)^{-(\beta+d/2)/2}\, d\omega \\
&\leq 2^{\beta+d/2} C_\phi \left[\int_{\mathbb{R}^d} (1 + \|\omega\|^2)^{-(\beta+d/2)/2}\, d\omega\right](1 + \|\xi\|^2)^{-(\beta+d/2)/2},
\end{aligned}
$$

which establishes the upper bound in (3.12).

Because $\widehat{\phi_2}$ is continuous and $\widehat{\phi_2}(0) = \int_{\mathbb{R}^d} \phi_2(x)\, dx > 0$, there is a positive constant δ such that $c_\phi := \min_{\|\omega\| \leq \delta} \widehat{\phi_2}(\omega) > 0$. Therefore

$$
\begin{aligned}
|\widehat{f}(\xi)| &= \int_{\mathbb{R}^d} \widehat{\phi_2}(\xi - \omega) \widehat{g}(\omega)\, d\omega \geq \int_{\|\xi-\omega\| \leq \delta} \widehat{\phi_2}(\xi - \omega)(1 + \|\omega\|^2)^{-(\beta+d/2)/2}\, d\omega \\
&\geq c_\phi \int_{\|\xi-\omega\| \leq \delta} (1 + \|\omega\|^2)^{-(\beta+d/2)/2}\, d\omega.
\end{aligned}
$$

Let $C_{d,\delta}$ be the volume of a d-dimensional δ-ball. For $\|\xi\| \geq \delta$ we get

$$
\begin{aligned}
|\widehat{f}(\xi)| &\geq c_\phi \int_{\|\xi-\omega\| \leq \delta} (1 + \|\omega\|^2)^{-(\beta+d/2)/2}\, d\omega \\
&\geq c_\phi \int_{\|\xi-\omega\| \leq \delta} (1 + (\|\xi\| + \delta)^2)^{-(\beta+d/2)/2}\, d\omega \\
&\geq C_{d,\delta}\, c_\phi (1 + 4\|\xi\|^2)^{-(\beta+d/2)/2} \\
&\geq 2^{-(\beta+d/2)} C_{d,\delta}\, c_\phi (1 + \|\xi\|^2)^{-(\beta+d/2)/2}.
\end{aligned}
$$

The lower bound in (3.12) follows from this estimate and $\inf_{\|\xi\| \leq \delta} |\widehat{f}(\xi)| > 0$.

It is clear that f satisfies (3.9), while (3.10) follows from a computation similar to (3.11). Finally, that $f \in H^\alpha(\mathbb{R}^d)$ if and only if $\alpha < \beta$ is a consequence of (3.12) and the fact that

$$\int_{\mathbb{R}^d} (1 + \|\xi\|^2)^\alpha (1 + \|\xi\|^2)^{-(\beta + d/2)} \, \mathrm{d}\xi < \infty$$

if and only if $\alpha < \beta$. $\qquad\square$

3.3 Smoothness Estimation

Let $\Omega \subset \mathbb{R}^d$ be a bounded set. The *fill-distance* $h_{n,\Omega}$ of the points $X_n = \{x_i\}_{i=1}^n$ is

$$h_{n,\Omega} = \sup_{x \in \Omega} \operatorname{dist}(x, X_n) = \sup_{x \in \Omega} \min_{i=1,\ldots,n} \|x - x_i\|$$

and their *separation radius* q_n is

$$q_n = \frac{1}{2} \min_{i \neq j} \|x_i - x_j\|.$$

The fill-distance equals the radius of the largest ball in Ω which does not contain any of the points in X_n, while an open ball with radius q_n can contain at most one point in X_n.

Definition 3.4 (Quasi-uniformity). Suppose that $\Omega \subset \mathbb{R}^d$ is bounded. A point sequence $\{x_i\}_{i=1}^\infty \subset \Omega$ is *quasi-uniform* on Ω if there is a positive constant c_{qu} such that

$$q_n \leq h_{n,\Omega} \leq c_{\mathrm{qu}} q_n \quad \text{for every} \quad n \in \mathbb{N}. \tag{3.14}$$

Quasi-uniform points cover the domain Ω in a sufficiently uniform manner, in that the ratio between the distance of the two nearest points in X_n and the diameter of the largest empty region in Ω which does not contain any of the points in X_n remains bounded from above and below. Quasi-uniformity implies that (e.g., Wendland, 2005, Proposition 14.1)

$$q_n \asymp_n h_{n,\Omega} \asymp_n n^{-1/d}. \tag{3.15}$$

Although the quasi-uniformity assumption is not satisfied by random points, extensions of our results for random points may be possible by following Krieg and Sonnleitner (2023) and considering $L^p(\Omega)$-norms of the distance function $\operatorname{dist}(\cdot, X_n)$ for $p < \infty$. The following assumption on regularity of the domain Ω is needed in some of our results.

Assumption 3.5. The domain $\Omega \subset \mathbb{R}^d$ is a bounded connected open set which satisfies an interior cone condition and has a Lipschitz boundary.

See Definition 3.6 in Wendland (2005) for the interior cone condition and p. 189 in Stein (1970) for the definition of a Lipschitz boundary. The former of the assumptions prohibits the existence of pinch points on the boundary of Ω by requiring that each $x \in \Omega$ be a vertex of a cone contained in Ω while the latter prescribes that the boundary of Ω is sufficiently regular. Standard domains, such as $(0,1)^d$ and open bounded convex sets, satisfy Assumption 3.5.

3.3.1 Variance Bounds

Propositions 3.6 and 3.7 provide upper and lower bounds on the conditional variance for Matérn kernels. Although both propositions are well known in the literature (e.g., Novak, 1988; Novak and Triebel, 2006), we include a full (and fairly thorough) proof of the latter proposition because we need to keep track of the constants that appear in the bounds.

Proposition 3.6. *Suppose that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 3.5 and that the points $\{x_i\}_{i=1}^{\infty}$ are such that $h_{n,\Omega} \lesssim_n n^{-1/d}$. Then*

$$\sup_{x \in \Omega} \mathbb{V}_\nu(x \mid X_n) \lesssim_n n^{-2\nu/d}$$

for every $\nu > 0$.

Proof. This result is a direct consequence of Corollary 4.1 in Arcangéli et al. (2007) and the norm-equivalence of $H(K_\nu)$ and $H^{\nu+d/2}(\mathbb{R}^d)$. A slightly less general version would follow from Proposition 3.6 in Wendland and Rieger (2005). $\square$

Proposition 3.7. *Suppose that $\Omega \subset \mathbb{R}^d$ is bounded, that $c(\nu)$ satisfies (3.2a), and that the points $\{x_i\}_{i=1}^{\infty}$ are quasi-uniform. Then*

$$\inf_{\nu \in [\nu_{\min}, \tau]} \min_{1 \le i \le n} \mathbb{V}_\nu(x_i \mid X_n^i) \gtrsim_n n^{-2\tau/d}$$

for every $\tau \in [\nu_{\min}, \nu_{\max}]$.

Proof. Let $\tau \in [\nu_{\min}, \nu_{\max}]$ and $\nu \in [\nu_{\min}, \tau]$. Recall from (2.6) that

$$\mathbb{V}_\nu(x_i \mid X_n^i) = \sup_{\|f\|_\nu \le 1} |f(x) - \mu_{\nu,f}(x_i \mid X_n^i)|^2.$$

From this expression it follows that $\mathbb{V}_\nu(x_i \mid X_n^i) \ge |f(x_i)|^2$ if $f \in H(K_\nu) = H^{\nu+d/2}(\mathbb{R}^d)$ is a function such that $\|f\|_\nu \le 1$ and $f(x) = 0$ for every $x \in X_n^i$ since, as can be seen from (2.2), in this case $\mu_{\nu,f}(\cdot \mid X_n^i) \equiv 0$. We now construct such a function.

Let ϕ be the bump function in (3.13). For any $q > 0$, define the function $g \colon \mathbb{R}^d \to \mathbb{R}$ via

$$g(x) = \phi\left(\frac{x - x_i}{q}\right). \tag{3.16}$$

This function is an element of $H^{\nu+d/2}(\mathbb{R}^d)$ for every $\nu > 0$ because the Fourier transform of ϕ decays faster than any polynomial, which implies that the Sobolev norm in (3.3) is finite. Suppose that $q \le 1$. Equation (3.7) and the scaling properties of the Fourier transform give

$$\|g\|_\nu^2 = C_\nu \int_{\mathbb{R}^d} \left(\frac{2\nu}{\lambda^2} + \|\xi\|^2\right)^{\nu+d/2} |\widehat{g}(\xi)|^2 \, \mathrm{d}\xi$$

$$\le C_\nu q^{2d} \int_{\mathbb{R}^d} \left(\frac{2\nu}{\lambda^2} + \|\xi\|^2\right)^{\nu+d/2} |\widehat{\phi}(q\xi)|^2 \, \mathrm{d}\xi.$$

A change of variables and some basic estimates based on $q \le 1$ and $\nu \le \tau$ then yield

$$\|g\|_\nu^2 \le C_\nu q^d \int_{\mathbb{R}^d} \left(\frac{2\nu}{\lambda^2} + \frac{\|\xi\|^2}{q^2}\right)^{\nu+d/2} |\widehat{\phi}(\xi)|^2 \, \mathrm{d}\xi$$

$$= C_\nu \lambda^{-(2\nu+d)} q^{-2\nu} \int_{\mathbb{R}^d} (2\nu q^2 + \lambda^2 \|\xi\|^2)^{\nu+d/2} |\widehat{\phi}(\xi)|^2 \, \mathrm{d}\xi$$

$$\le C_\nu \lambda^{-(2\nu+d)} q^{-2\nu} \int_{\mathbb{R}^d} (2\nu + \lambda^2 \|\xi\|^2)^{\nu+d/2} |\widehat{\phi}(\xi)|^2 \, \mathrm{d}\xi$$

$$\le C_\nu B_\lambda q^{-2\nu} \int_{\mathbb{R}^d} (2\tau + \lambda^2 \|\xi\|^2)^{\tau+d/2} |\widehat{\phi}(\xi)|^2 \, \mathrm{d}\xi,$$

17

where $B_\lambda = \max\{1, \lambda^{-(2\tau+d)}\}$. By assumption (3.2a), there is $c > 0$ such that $\Gamma(\nu + d/2) \geq c$ and $c(\nu) \geq c$ for all $\nu \in [\nu_{\min}, \tau]$. Moreover, $\max_{\nu>0}(2\nu)^{-\nu} = e^{1/(2e)}$. Thus

$$C_\nu = \frac{\pi^{d/2}}{\sigma^2 c(\nu) 2^{\nu-1} \Gamma(\nu + d/2)} \left(\frac{\lambda^2}{2\nu}\right)^\nu \leq \frac{2\pi^{d/2}}{\sigma^2 c^2} \max\{1, \lambda^{2\tau}\}(2\nu)^{-\nu}$$
$$\leq \frac{\pi^{d/2}}{\sigma^2 c^2} \max\{1, \lambda^{2\tau}\} e^{1/(2e)}$$

for all $\nu \in [\nu_{\min}, \tau]$. It follows that there is a constant $C > 0$ such that $\|g\|_\nu \leq Cq^{-\nu}$ for every $\nu \in [\nu_{\min}, \tau]$. Therefore the $H(K_\nu)$-norm of the function $f = C^{-1}q^\nu g$ does not exceed one.

Set $q = q_n$ and assume that n is sufficiently large that $q_n \leq 1$ holds. It follows from ϕ being supported on the unit ball and (3.16), as well as the definition of the separation radius, that $f(x) = 0$ for every $x \in X_n^i$. By the argument given in the beginning of the proof and $q_n \leq 1$,

$$\mathbb{V}_\nu(x_i \mid X_n^i) \geq |f(x_i)|^2 = C^{-2} q_n^{2\nu} g(x_i) = C^{-2} q_n^{2\nu} \phi(0) = C^{-2} q_n^{2\nu} \geq C^{-2} q_n^{2\tau}$$

for every $\nu \in [\nu_{\min}, \tau]$. The claim then follows from the assumption that $\{x_i\}_{i=1}^\infty$ are quasi-uniform and (3.15). $\qquad\square$

3.3.2 Some Norm Bounds

We need upper and lower bounds on the RKHS norm of the conditional mean. These are given in Propositions 3.8 and 3.10, respectively.

Proposition 3.8. *Suppose that $\Omega \subset \mathbb{R}^d$ is bounded and that the points $\{x_i\}_{i=1}^\infty$ are quasi-uniform on Ω. Let B be a bounded subset of $H^{\tau+d/2}(\mathbb{R}^d)$ for $\tau > 0$. Then*

$$\sup_{f \in B} \|\mu_{\nu,f}(\cdot \mid X_n)\|_\nu^2 \lesssim_n n^{2(\nu-\tau)/d}$$

for every $\nu \geq \tau$.

Proof. The claim follows from Lemma A.1 in Karvonen et al. (2020), quasi-uniformity, and the norm-equivalence of $H(K_\nu)$ and $H^{\nu+d/2}(\mathbb{R}^d)$. $\qquad\square$

The following lemma states that the Matérn norm for $\tau > 0$ is weaker, in a uniform sense, than that for all $\nu \geq \tau$ if the standard scaling $c(\nu) = 2^{1-\nu}/\Gamma(\nu)$ is used. It seems likely that this lemma exists in some form in the literature. Note that for Sobolev norms we obtain from (3.3) the simpler result that $\|f\|_{H^\beta(\mathbb{R}^d)} \leq \|f\|_{H^\alpha(\mathbb{R}^d)}$ if $\alpha \geq \beta$.

Lemma 3.9. *Let $\tau > 0$ and suppose that there is a positive constant C_τ such that*

$$c(\nu) \leq \frac{C_\tau}{2^\nu \Gamma(\nu)} \tag{3.17}$$

for all $\nu \geq \tau$. Then there is a positive constant C such that

$$\|f\|_\tau^2 \leq C \|f\|_\nu^2 \tag{3.18}$$

for all $\nu \geq \tau$ and $f \in H(K_\tau)$, where we set $\|f\|_\nu = \infty$ if $f \notin H(K_\nu)$. In particular, the estimate (3.18) holds for all $\nu \in [\tau, \nu_{\max}]$ if $\tau \in [\nu_{\min}, \nu_{\max}]$ and $c(\nu)$ satisfies (3.2b).

Proof. It is straightforward to compute that the function

$$p_{\tau,\nu}(x) = \frac{(2\tau/\lambda^2 + x)^{\tau+d/2}}{(2\nu/\lambda^2 + x)^{\nu+d/2}}$$

18

attains its maximum on $[0, \infty)$ at $x = d/\lambda^2$. Therefore

$$
\max_{x \geq 0} p_{\tau,\nu}(x) = \frac{(2\tau/\lambda^2 + d/\lambda^2)^{\tau+d/2}}{(2\nu/\lambda^2 + d/\lambda^2)^{\nu+d/2}} = \left(\frac{\lambda^2}{2}\right)^{\nu-\tau} \frac{\tau^{\tau+d/2}}{\nu^{\nu+d/2}} \cdot \frac{(1 + d/(2\tau))^{\tau+d/2}}{(1 + d/(2\nu))^{\nu+d/2}}
$$
$$
\leq \left(\frac{\lambda^2}{2}\right)^{\nu-\tau} \frac{\tau^{\tau+d/2}}{\nu^{\nu+d/2}} (1 + d/(2\tau))^{\tau+d/2}. \tag{3.19}
$$

Let $\kappa_\tau = c(\tau)2^{\tau-1}\Gamma(\tau + d/2)$ and $\lambda_\tau = (1 + d/(2\tau))^{\tau+d/2}$. From (3.19) we get

$$
\begin{aligned}
\|f\|_\tau^2 &= \frac{\pi^{d/2}}{\sigma^2 \kappa_\tau}\left(\frac{\lambda^2}{2\tau}\right)^\tau \int_{\mathbb{R}^d} \left(\frac{2\tau}{\lambda^2} + \|\xi\|^2\right)^{\tau+d/2} |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi \\
&= \frac{\pi^{d/2}}{\sigma^2 \kappa_\tau}\left(\frac{\lambda^2}{2\tau}\right)^\tau \int_{\mathbb{R}^d} p_{\tau,\nu}(\|\xi\|^2)\left(\frac{2\nu}{\lambda^2} + \|\xi\|^2\right)^{\nu+d/2} |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi \\
&\leq \frac{\pi^{d/2}}{\sigma^2 \kappa_\tau}\left(\frac{\lambda^2}{2\tau}\right)^\tau \left(\frac{\lambda^2}{2}\right)^{\nu-\tau} \frac{\tau^{\tau+d/2}}{\nu^{\nu+d/2}} \lambda_\tau \int_{\mathbb{R}^d} \left(\frac{2\nu}{\lambda^2} + \|\xi\|^2\right)^{\nu+d/2} |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi \\
&= \frac{\lambda_\tau \tau^{d/2}}{\kappa_\tau} \cdot \frac{\pi^{d/2}}{\sigma^2 \nu^{d/2}}\left(\frac{\lambda^2}{2\nu}\right)^\nu \int_{\mathbb{R}^d} \left(\frac{2\nu}{\lambda^2} + \|\xi\|^2\right)^{\nu+d/2} |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi \\
&= \frac{\lambda_\tau \tau^{d/2}}{\kappa_\tau} \cdot \frac{c(\nu)2^{\nu-1}\Gamma(\nu + d/2)}{\nu^{d/2}} \|f\|_\nu^2.
\end{aligned}
$$

The claim then follows from (3.17) and the Gamma function asymptotics $\Gamma(\nu + d/2) \sim \nu^{d/2}\Gamma(\nu)$ as $\nu \to \infty$. $\qquad\square$

For self-similar functions whose support is contained in the domain the RKHS norm of the conditional mean can be bounded from below by the following proposition, which originally appeared as Theorem 8 in van der Vaart and van Zanten (2011).

Proposition 3.10. *Suppose that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 3.5, that $c(\nu)$ satisfies* (3.2b), *and that the points $\{x_i\}_{i=1}^\infty$ are quasi-uniform on Ω. Let $\tau \in (0, \nu_{\max}]$. Then for every $f \in H_{\mathrm{ss}}^{\tau+d/2}(\Omega)$ we have*

$$
\inf_{\nu \in [\nu_1, \nu_2]} \|\mu_{\nu,f}(\cdot \mid X_n)\|_\nu^2 \gtrsim_n n^{2(\nu_1-\tau)/d - 4(\nu_1-\tau)\delta} \tag{3.20}
$$

for any $\delta \in (0, \tau)$, $\nu_1 \in [\tau, \nu_{\max}]$, and $\nu_2 \in [\nu_1, \nu_{\max}]$.

Proof. By the minimum-norm property (2.5) and Lemma 3.9, there is a positive constant C such that

$$
\|\mu_{\nu_1,f}(\cdot \mid X_n)\|_{\nu_1}^2 \leq \|\mu_{\nu,f}(\cdot \mid X_n)\|_{\nu_1}^2 \leq C \|\mu_{\nu,f}(\cdot \mid X_n)\|_\nu
$$

for all $\nu \in [\tau, \nu_{\max}]$. It therefore suffices to prove (3.20) for the fixed smoothness $\nu = \nu_1$ (i.e., when $\nu_2 = \nu_1$). Proposition 4.7 in Karvonen et al. (2020) with $\alpha = \nu_1 + d/2$, $\beta = \tau + d/2 \leq \alpha$ and $\gamma = \tau - \delta + d/2 < \beta$ states that

$$
\|\mu_{\nu_1,f}(\cdot \mid X_n)\|_{\nu_1}^2 \gtrsim_n n^{2\gamma(\alpha/\beta-1)/d} = n^{2(\nu_1-\tau)/d - 2(\nu_1-\tau)\delta/(d(\tau+d/2))} \tag{3.21}
$$

if $f \in H^\gamma(\mathbb{R}^d)$ is such that $|\widehat{f}(\xi)|^2 \geq C(1 + \|\xi\|^2)^{-(\beta+d/2)}$ for some $C > 0$ and all $\xi \in \mathbb{R}^d$ and the support of f is contained in Ω. This assumption on the decay of the Fourier transform implies (3.10). However, by inspection of the proofs of Proposition 4.7 in Karvonen et al. (2020) [specifically, the equation after (A.10)] and Theorem 8 in van der Vaart and van Zanten (2011) we see that (3.10) is sufficient to establish (3.21). Moreover, that $\lfloor \gamma \rfloor > d/2$ may be relaxed to $\gamma > d/2$ by using Corollary 4.1 of Arcangéli et al. (2007) to derive Equation (A.9) in Karvonen et al. (2020). For any $f \in H_{\mathrm{ss}}^{\tau+d/2}(\Omega)$ we therefore have

$$
\|\mu_{\nu_1,f}(\cdot \mid X_n)\|_{\nu_1}^2 \gtrsim_n n^{2(\nu_1-\tau)/d - 2(\nu_1-\tau)\delta/(d(\tau+d/2))} \geq n^{2(\nu_1-\tau)/d - 4(\nu_1-\tau)\delta},
$$

which completes the proof. $\qquad\square$

3.3.3 Main Result for Matérns

Propositions 3.6 to 3.8 and 3.10 yield our main result on smoothness estimation in the Matérn model. We shall use a parameter space of the following form:

$$\Theta = [\nu_{\min}, \nu_{\max}] \quad \text{for} \quad \nu_{\min} \in (0, \nu_{\max}] \text{ and } \nu_{\max} \in (d/2, \infty). \tag{3.22}$$

Theorem 3.11. *Let Θ be given in (3.22). Suppose that $c(\nu)$ satisfies (3.2a), that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 3.5, and that the points $\{x_i\}_{i=1}^{\infty}$ are quasi-uniform on Ω. If B is a bounded subset of $H^{\nu_0}(\Omega)$ for $\nu_0 \in (d/2, \nu_{\max}]$, then*

$$\liminf_{n \to \infty} \inf_{f_0 \in B} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \geq \nu_0 \quad \text{and} \quad \liminf_{n \to \infty} \inf_{f_0 \in B} \hat{\nu}_{\mathrm{CV}}^{f_0}(X_n) \geq \nu_0 - d/2. \tag{3.23}$$

Moreover, the bound for maximum likelihood estimation is sharp if $c(\nu)$ satisfies also (3.2b), in the sense that for every $\varepsilon > 0$ there is $f_0 \in H^{\nu_0}(\Omega)$ such that

$$\limsup_{n \to \infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \leq \nu_0 + \varepsilon. \tag{3.24}$$

Proof. By (3.4), each $f_0 \in B \subset H^{\nu_0}(\Omega)$ may be identified with $f_e \in H^{\nu_0}(\mathbb{R}^d)$ such that $f_e|_{\Omega} = f_0$ and the Sobolev norms of f_0 and f_e are equal. The set of these extensions f_e is bounded in $H^{\nu_0}(\mathbb{R}^d)$. We may therefore proceed as if B were a bounded subset of $H^{\nu_0}(\mathbb{R}^d)$. Let $\tilde{\nu} > 0$ and $\nu_1 \in \Theta$. Propositions 3.6 and 3.7 yield

$$\sup_{\nu \in [\nu_{\min}, \nu_1]} \max_{1 \leq i \leq n} \frac{\mathbb{V}_{\tilde{\nu}}(x_i \mid X_n^i)}{\mathbb{V}_{\nu}(x_i \mid X_n^i)} \lesssim_n \frac{n^{-2\tilde{\nu}/d}}{n^{-2\nu_1/d}} = n^{-2(\tilde{\nu} - \nu_1)/d}. \tag{3.25}$$

Because $H(K_{\nu_0 - d/2})$ is norm-equivalent to $H^{\nu_0}(\mathbb{R}^d)$, the bound on the lower limit for $\hat{\nu}_{\mathrm{CV}}^{f_0}(X_n)$ follows by setting $\theta_0 = \tilde{\nu} = \nu_0 - d/2$ and $\theta_1 = \nu_1 < \nu_0 - d/2 < \tilde{\nu}$ in Corollary 2.9, in which case $n^{-2(\tilde{\nu} - \nu_1)/d} \to 0$. Note that the lower limit holds trivially whenever $\nu_0 - d/2 < \nu_{\min}$.

To prove the lower bound for the maximum likelihood estimator we shall apply Corollary 2.4 with $\theta_0 = \nu_0$. If $\nu_0 < \nu_{\min}$, the lower bound holds trivially, so we may assume that $\nu_0 \in \Theta$. Setting $\tilde{\nu} = \nu_0$ and considering the case $i = n$ (so that $X_n^i = X_{n-1}$) in (3.25) yields

$$\sup_{\nu \in [\nu_{\min}, \nu_1]} \frac{\mathbb{V}_{\nu_0}(x_n \mid X_{n-1})}{\mathbb{V}_{\nu}(x_n \mid X_{n-1})} \lesssim_n n^{-2(\nu_0 - \nu_1)/d}. \tag{3.26}$$

Proposition 3.8 with $\nu = \nu_0$ and $\tau = \nu_0 - d/2$ yields

$$\sup_{f_0 \in B} \|\mu_{\nu_0, f_0}(\cdot \mid X_n)\|_{\nu_0}^2 \lesssim_n n^{2(\nu - \tau)/d} = n.$$

Therefore,

$$\begin{aligned}
Q_n(\nu_1) &:= \sup_{\nu \in [\nu_{\min}, \nu_1]} \sup_{f_0 \in B} \left[\|\mu_{\nu_0, f_0}(\cdot \mid X_n)\|_{\nu_0}^2 + \sum_{i=1}^{n} \log \frac{\mathbb{V}_{\nu_0}(x_i \mid X_{i-1})}{\mathbb{V}_{\nu}(x_i \mid X_{i-1})} \right] \\
&\lesssim_n n - \frac{2(\nu_0 - \nu_1)}{d} \sum_{i=1}^{n} \log i \\
&= n - \frac{2(\nu_0 - \nu_1)}{d} \log n!.
\end{aligned}$$

From Stirling's formula $n! \sim \sqrt{2\pi} n^{n+1/2} e^{-n}$ as $n \to \infty$ we get

$$\frac{2(\nu_0 - \nu_1)}{d} \log n! \sim \frac{2(\nu_0 - \nu_1)}{d} n \log n.$$

Hence

$$Q_n(\nu_1) \lesssim_n n - \frac{2(\nu_0 - \nu_1)}{d} \log n! \to -\infty$$

as $n \to \infty$ for every ν_1 such that $2(\nu_0 - \nu_1)/d > 0$, which is equivalent to $\nu_1 < \nu_0$. We may thus use Corollary 2.4 with $\theta_0 = \nu_0$.

We are left to prove that the lower bound for the maximum likelihood estimator is sharp in the sense that for every $\varepsilon > 0$ there is $f_0 \in H^{\nu_0}(\Omega)$ for which (3.24) holds. We may assume that $\nu_0 + \varepsilon < \nu_{\max}$, for otherwise (3.24) holds trivially. We shall use Corollary 2.6. By Lemmas 3.2 and 3.3 and Proposition 3.10 with $\tau = \nu_0 + \varepsilon/2 - d/2$, $\nu_1 = \nu_0 + \varepsilon$, and δ sufficiently small, there exists $f_0 \in H^{\nu_0}(\mathbb{R}^d)$ such that

$$\inf_{\nu \in [\nu_1, \nu_2]} \|\mu_{\nu, f_0}(\cdot \mid X_n)\|_\nu^2 \gtrsim_n n^{2(\nu_1 - \tau)/d - 4(\nu_1 - \tau)\delta} = n^{1 + \varepsilon/d - 4(d/2 + \varepsilon/2)\delta} \tag{3.27}$$

$$\geq n^{1 + \varepsilon/(2d)}$$

for every $\nu_2 \in [\nu_1, \nu_{\max}]$. The lower bound (3.27) and Proposition 3.7 yield

$$
\begin{aligned}
P_n(\nu_1, \nu_2) &:= \inf_{\nu \in [\nu_1, \nu_2]} \ell_{\mathrm{ML}}^{f_0}(\nu \mid X_n) \\
&= \inf_{\nu \in [\nu_1, \nu_2]} \left[\|\mu_{\nu, f_0}(\cdot \mid X_n)\|_\nu^2 + \sum_{i=1}^n \log \mathbb{V}_\nu(x_i \mid X_{i-1}) \right] \tag{3.28} \\
&\gtrsim_n n^{1 + \varepsilon/(2d)} - \frac{2\nu_2}{d} \log n!,
\end{aligned}
$$

so that $P_n(\nu_1, \nu_2) \to \infty$ as $n \to \infty$ by the same arguments that we used to control $Q_n(\nu_1)$ above. The claim follows by applying Corollary 2.6 with $\theta_0 = \nu_1 = \nu_0 + \varepsilon$ and $B = \{f_0\}$ for every $\theta_1 = \nu_2 > \nu_1$. $\qquad\square$

As discussed in Section 2.4, Corollary 2.9 that we applied to cross-validation is weaker than Corollary 2.4 that we applied to maximum likelihood estimation. Corollary 2.9 is capable of establishing a lower bound ν_0 only if B is a bounded subset of $H^{\nu_0 + d/2}(\Omega)$ while Corollary 2.4 is more flexible. The use of Corollary 2.4 to improve the lower bound for cross-validation would require that one proved a variant of Proposition 3.8 to bound

$$\sum_{i=1}^n \frac{(f_0(x_i) - \mu_{\nu, f_0}(x_i \mid X_n^i))^2}{\mathbb{V}_\nu(x_i \mid X_n^i)}$$

from above when $\nu \geq \nu_0$. We do not presently know how to do this. Similarly, proving a version of the upper bound (3.24) for cross-validation would require controlling the above quantity from below.

The following theorem optimises Theorem 3.11 for the maximum likelihood estimator. For a function f that is defined and square-integrable at least on Ω, let

$$\nu(f) = \sup\{\nu > 0 \,:\, f|_\Omega \in H^\nu(\Omega)\},$$

so that $f \in H^{\nu(f) - \varepsilon}(\Omega)$ for every $\varepsilon \in (0, \nu(f))$. Wang and Jing (2022, Section 5.1.1) call $\nu(f)$ the *smoothness* of f.

Theorem 3.12. *Let Θ be given in (3.22). Suppose that $c(\nu)$ satisfies (3.2a), that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 3.5, and that the points $\{x_i\}_{i=1}^\infty$ are quasi-uniform on Ω.*

1. If $f_0 \colon \Omega \to \mathbb{R}$ is such that $\nu(f_0) \in (d/2, \nu_{\max}]$, then

$$\liminf_{n \to \infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \geq \nu(f_0). \tag{3.29}$$

21

2. *If $f_0 \in H_{\mathrm{ss}}^{\nu_0}(\Omega)$ for $\nu_0 \in (d/2, \nu_{\max}]$ and $c(\nu)$ satisfies also (3.2b), then*

$$\lim_{n \to \infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) = \nu_0. \tag{3.30}$$

Proof. The bound (3.29) follows from (3.23) because $B = \{f_0\} \subset H^{\nu(f_0)-\varepsilon}(\Omega)$ for every sufficiently small $\varepsilon > 0$. Suppose that $f_0 \in H_{\mathrm{ss}}^{\nu_0}(\Omega)$ and $\nu_0 < \nu_{\max}$. Let $\varepsilon > 0$ be small enough that $\nu_0 + \varepsilon < \nu_{\max}$. With $\tau = \nu_0 - d/2$, $\nu_1 = \nu_0 + \varepsilon$, and δ sufficiently small, Proposition 3.10 yields

$$\inf_{\nu \in [\nu_1, \nu_2]} \|\mu_{\nu,f_0}(\cdot \mid X_n)\|_{\nu}^2 \gtrsim_n n^{2(\nu_1 - \tau)/d - 4(\nu_1 - \tau)\delta} = n^{1 + 2\varepsilon/d - 4(d/2 + \varepsilon)\delta} \geq n^{1 + \varepsilon/d}$$

for every $\nu_2 \in [\nu_1, \nu_{\max}]$. By bounding $P_n(\nu_1, \nu_2) = \inf_{\nu \in [\nu_1, \nu_2]} \ell_{\mathrm{ML}}^{f_0}(\nu \mid X_n)$ as in the proof of Theorem 3.11 and applying Corollary 2.6 with $\theta_0 = \nu_0$ and $B = \{f_0\}$ to the parameter space $\Theta = [\nu_{\min}, \nu_{\max}]$, we then get $\limsup_{n \to \infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \leq \nu_0$. If $\nu_0 = \nu_{\max}$, this upper bound holds trivially. To establish a matching lower bound, observe that, by Lemma 3.2, $f_0|_\Omega \in H^\nu(\Omega)$ for every $\nu < \nu_0$. Therefore $\nu(f_0) \geq \nu_0$, and it thus follows from (3.29) that $\liminf_{n \to \infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \geq \nu_0$. This concludes the proof of (3.30). $\qquad\square$

Let us say a few words about the assumptions in Theorems 3.11 and 3.12. As discussed earlier, Assumption 3.5 on the domain is non-restrictive. In comparison to other spatial sampling assumptions employed in the literature, the quasi-uniformity assumption is rather non-restrictive and generic. However, random points are not quasi-uniform. Significant relaxations of the quasi-uniformity assumption may require a new assumption on the spatial homogeneity of f_0. The assumption that the parameter space is compact and bounded away from zero is not practically restrictive. Generalisations to the case $\Theta = (0, \infty)$ would require much more careful handling of smoothness-dependent constant coefficients. For example, in the proof of (3.24) the two terms that make up $P_n(\nu_1, \nu_2)$ in (3.28) are bounded from below individually, which allows us to ignore potential dependencies on ν_2 in their constant coefficients.

That the self-similar functions for which the limit (3.30) is obtained are required to have their supports in Ω is the most unsatisfactory part of Theorems 3.11 and 3.12. Because the samples from a Gaussian process with a Matérn covariance kernel are not compactly supported, this assumption is likely superfluous. The assumption propagates from Proposition 3.10 and Theorem 8 in van der Vaart and van Zanten (2011) and to remove it a new technique to obtain lower bounds on RKHS norms is needed.

3.4 Infinitely Smooth Functions

The following corollary shows that the parameter estimators detect infinite smoothness of the response function. Here we obviously need to consider an infinite parameter space.

Corollary 3.13. *Let $\Theta = [\nu_{\min}, \infty)$ for $\nu_{\min} > 0$. Suppose that $c(\nu)$ satisfies (3.2a) for every $\nu_{\max} > \nu_{\min}$, that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 3.5, and that the points $\{x_i\}_{i=1}^\infty$ are quasi-uniform on Ω. If B is a bounded subset of $H^\nu(\Omega)$ for every $\nu > d/2$, then*

$$\lim_{n \to \infty} \inf_{f_0 \in B} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) = \lim_{n \to \infty} \inf_{f_0 \in B} \hat{\nu}_{\mathrm{CV}}^{f_0}(X_n) = \infty.$$

Proof. The claims follow because the bounds (3.23) hold for every $\nu_0 > d/2$. $\qquad\square$

Even though the Matérn kernel, with proper parametrisation, tends to the infinitely smooth Gaussian kernel as $\nu \to \infty$, it is not required in Corollary 3.13 that f_0 be an element of the RKHS of the Gaussian kernel or of an RKHS that contains the samples of a Gaussian process with the Gaussian kernel. However, as we now argue, it is likely that membership in the Gaussian RKHS causes the smoothness parameter estimates to diverge faster as n increases than if this

were not the case. For simplicity, consider Theorem 2.7 and maximum likelihood estimation. From the proof of Theorem 2.7 we see that $\hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \geq \nu_1$ if $f \in H(K_{\nu_0})$ and

$$\|f_0\|_{\nu_0}^2 + \sup_{\nu \in [\nu_{\min}, \nu_1]} \sum_{i=1}^n \log \frac{\mathbb{V}_{\nu_0}(x_i \mid X_{i-1})}{\mathbb{V}_\nu(x_i \mid X_{i-1})} < 0$$

for $\nu_1 < \nu_0$. Notably, the second term above does not depend on the response function. This suggests (though does not rigorously prove) that the larger $\|f_0\|_{\nu_0}^2$ is, the larger an n is needed for the maximum likelihood estimator to exceed ν_1. Or, in other words, a fast rate of growth of $\|f_0\|_\nu$ as ν increases ought to imply a slow rate of growth of $\hat{\nu}_{\mathrm{ML}}^{f_0}(X_n)$. Next we estimate $\|f_0\|_\nu$ for two functions, of which one is not an element of the RKHS of the Gaussian kernel and the other is.

Consider the Matérn kernel

$$K_\nu(x, y) = \frac{2^{1-\nu}}{\Gamma(\nu)} \big(\sqrt{\nu}\, |x - y| \big)^\nu \mathcal{K}_\nu \big(\sqrt{\nu}\, |x - y| \big)$$

for $d = 1$, where we have set $\sigma = 1$ and $\lambda = \sqrt{2}$. As $\nu \to \infty$, the Matérn kernel above tends to the Gaussian kernel $K(x, y) = \exp(-(x - y)^2/4)$. By the characterisation (3.5), the RKHS of the Gaussian kernel consists of functions whose Fourier transforms satisfy

$$\int_{\mathbb{R}} \exp(\xi^2) |\widehat{f}(\xi)|^2 \, \mathrm{d}\xi < \infty. \tag{3.31}$$

Consider the function $f_0(x) = 1/(1/4 + x^2)$ with the Fourier transform $\widehat{f}_0(\xi) = \exp(-\,|\xi|\,/2)$. It is clear from (3.7) and (3.31) that this function is an element of every Matérn RKHS but not an element of the RKHS of the Gaussian kernel. Now, from (3.7) we have

$$\begin{aligned}
\|f_0\|_\nu^2 = C_\nu \int_{\mathbb{R}} (\nu + \xi^2)^{\nu+1/2} \exp(-\,|\xi|)\, \mathrm{d}\xi &= 2C_\nu \int_0^\infty (\nu + \xi^2)^{\nu+1/2} \exp(-\xi)\, \mathrm{d}\xi \\
&\geq 2C_\nu \int_0^\infty \xi^{2\nu+1} \exp(-\xi)\, \mathrm{d}\xi \\
&= 2\sqrt{\pi}\, \frac{\Gamma(\nu)\Gamma(2\nu + 2)}{\Gamma(\nu + 1/2)} \nu^{-\nu}.
\end{aligned}$$

From $\Gamma(\nu + 1/2) \sim \Gamma(\nu)\sqrt{\nu}$ and Stirling's formula we get

$$\frac{\Gamma(\nu)\Gamma(2\nu + 2)}{\Gamma(\nu + 1/2)} \nu^{-\nu} \sim \Gamma(2\nu + 2)\nu^{-\nu-1/2} \sim \sqrt{2\pi}\, \frac{2^{\nu+1/2}}{e^{2\nu+1}} (2\nu + 1)^{\nu+1},$$

which shows that $\|f_0\|_\nu \to \infty$ very fast as $\nu \to \infty$. We therefore expect that for this response function $\hat{\nu}_{\mathrm{ML}}^{f_0}(X_n)$ grows rather slowly.

As a second example, consider the function $f_0(x) = (2\sqrt{\pi}\,)^{-1} \exp(-x^2/4)$ with the Fourier transform $\widehat{f}_0(\xi) = \exp(-\xi^2)$. Unlike the previous function, this function is an element of the Gaussian RKHS because

$$\int_{\mathbb{R}} \exp(\xi^2) |\widehat{f}_0(\xi)|^2 \, \mathrm{d}\xi = \int_{\mathbb{R}} \exp(-\xi^2)\, \mathrm{d}\xi = \sqrt{\pi} < \infty.$$

Furthermore,

$$\|f_0\|_\nu^2 = C_\nu \int_\mathbb{R} (\nu + \xi^2)^{\nu+1/2} \exp(-2\xi^2) \, d\xi = C_\nu \nu^{\nu+1/2} \int_\mathbb{R} \left(1 + \frac{\xi^2}{\nu}\right)^{\nu+1/2} \exp(-2\xi^2) \, d\xi$$

$$\leq C_\nu \nu^{\nu+1/2} \int_\mathbb{R} \exp(\xi^2) \sqrt{1 + \frac{\xi^2}{\nu}} \exp(-2\xi^2) \, d\xi$$

$$= \sqrt{\pi} \frac{\Gamma(\nu)\sqrt{\nu}}{\Gamma(\nu+1/2)} \int_\mathbb{R} \sqrt{1 + \frac{\xi^2}{\nu}} \exp(-\xi^2) \, d\xi$$

$$\sim \sqrt{\pi} \int_\mathbb{R} \exp(-\xi^2) \, d\xi$$

$$= \pi.$$

Therefore $\|f_0\|_\nu$ is bounded as $\nu \to \infty$ and we can thus expect that $\hat{\nu}_{\mathrm{ML}}^{f_0}(X_n)$ grows fast as n increases.

3.5 Sample Paths

Let us then assume that f_0 is a version of a zero-mean Gaussian process with a Matérn covariance kernel K_{ν_0} that has continuous sample paths (such a version always exists). It is well known that under these assumptions almost all samples of f_0 are in the Sobolev space of order $\nu_0 - \varepsilon$ for every $\varepsilon > 0$. Specifically, if $\Omega \subset \mathbb{R}^d$ satisfies Assumption 3.5, then $f_0 \in H^{\nu_0-\varepsilon}(\Omega)$ almost surely for every $\varepsilon > 0$. For results of this type, see Scheuerer (2010); Corollary 4.15 in Kanagawa et al. (2018); Corollaries 4.5 and 5.7 in Steinwart (2019); Karvonen (2023); and Henderson (2023). From this result we obtain a Bayesian version of Theorem 3.11.

Corollary 3.14. *Let Θ be given in* (3.22)*. Suppose that $c(\nu)$ satisfies* (3.2a)*, that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 3.5, and that the points $\{x_i\}_{i=1}^\infty$ are quasi-uniform on Ω. If f_0 is a version of a zero-mean Gaussian process with a Matérn covariance kernel K_{ν_0} for some $\nu_0 \in (d/2, \nu_{\max}]$ such that almost all of its samples are continuous, then*

$$\liminf_{n\to\infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \geq \nu_0 \quad and \quad \liminf_{n\to\infty} \hat{\nu}_{\mathrm{CV}}^{f_0}(X_n) \geq \nu_0 - \frac{d}{2} \quad almost\ surely.$$

Proof. Because $f_0 \in H^{\nu_0-1/k}(\Omega)$ almost surely for every $k \in \mathbb{N}$, we obtain from Theorem 3.11 that

$$\liminf_{n\to\infty} \hat{\nu}_{\mathrm{ML}}^{f_0}(X_n) \geq \nu_0 - \frac{1}{k} \quad \text{almost surely}$$

for every $k \in \mathbb{N}$. The claims follow by taking intersections over $k \in \mathbb{N}$ of sets of measure one. The proof for cross-validation is analogous. $\square$

As discussed in Section 3.3.3, the result in Corollary 3.14 for cross-validation is likely sub-optimal. Because the samples are not supported on Ω, we are unable to exploit results that require f_0 to be an element of $H_{\mathrm{ss}}^{\nu_0}(\Omega)$. Nevertheless, Corollary 3.14 is a step towards showing that the maximum likelihood and cross-validation estimators are consistent or strongly consistent. Chen et al. (2021, Theorem 2.7) and Petit (2023) have proved consistency results for smoothness estimators of periodic Matérn-type kernels.

4 Discussion

This section contains some discussion on the implications of the smoothness estimation results in Section 3. We also discuss the use of Theorem 2.7 to estimate the scale parameter of an infinitely smooth stationary kernel.

4.1 Approximation and Uncertainty Quantification

Suppose that $f_0 \in H^{\nu_0 + d/2}(\Omega)$ and that the assumptions of Theorem 3.11 are satisfied. We now discuss what happens in the frequentist setting when a Matérn model of smoothness ν either over- or undersmooths the truth. For a succinct review on the closely related topic of coverage properties of Bayesian credible intervals we refer to Szabó et al. (2015, pp. 1391–2).

Undersmoothing: $\nu_0 > \nu$. If the response function is smoother than the Matérn prior, the current theory does not guarantee that the conditional mean tends to the response function with a rate that adapts to the smoothness, except if (essentially) $\nu_0 \geq 2\nu$, in which case one can expect the rate to be approximately $n^{-2\nu/d}$. This is known as *superconvergence* (Schaback, 2018) or the *improved rate* (Wendland, 2005, Section 11.5) of kernel-based approximation. See Karvonen et al. (2020, Sections 3.4 and 4.5) and Tuo et al. (2020, Section 2.3) for discussion on superconvergence in the context of Gaussian process interpolation. By (2.7) and the fact that $f_0 \in H^{\nu_0 + d/2}(\Omega) \subset H^{\nu + d/2}(\Omega)$,

$$|f_0(x) - \mu_{\nu, f_0}(x \mid X_n)| \leq \|f_0\|_\nu \, \mathbb{V}_\nu(x \mid X_n)^{1/2} \tag{4.1}$$

for every $x \in \Omega$. Therefore f_0 is contained in the credible set

$$\mathcal{C}_\rho^n(f_0) = \left\{ f \colon \Omega \to \mathbb{R} \, : \, |f(x) - \mu_{\nu, f_0}(x \mid X_n)| \leq \rho \mathbb{V}_\nu(x \mid X_n)^{1/2} \text{ for every } x \in \Omega \right\} \tag{4.2}$$

for any $\rho \geq \|f_0\|_\nu$ and $n \in \mathbb{N}$. Reliable uncertainty quantification is therefore possible when the model undersmooths, in that there is a fixed credible level for which the credible set centered at the conditional mean contains the truth for all n. If f_0 is smooth enough to benefit from superconvergence (or other such phenomenon), then $f_0 \in \mathcal{C}_{\rho_n}^n(f_0)$ can hold even for a sequence $(\rho_n)_{n=1}^\infty$ that decays fast. Since $\|\mu_{\nu, f_0}(\cdot \mid X_n)\|_\nu^2 \to \|f_0\|_\nu^2$ as $n \to \infty$ if $h_{n, \Omega} \to 0$ (e.g., Iske, 2018, Theorem 8.37), one may use (2.9) to construct a sequence $(\bar\rho_n)_{n=1}^\infty$ with limit $\bar\rho = \|f_0\|_\nu = \lim_{n \to \infty} \bar\rho_n$ such that $f_0 \in \mathcal{C}_{\bar\rho}^n(f_0)$ for every n.

Oversmoothing: $\nu_0 < \nu$. By the Narcowich–Ward–Wendland escape theorem (Narcowich et al. 2006; see Wynne et al. 2021, Theorem 1 for a Gaussian process formulation), the Matérn conditional mean tends to $f_0 \in H^{\nu_0 + d/2}(\Omega)$ with a rate which depends on ν_0:

$$\sup_{x \in \Omega} |f_0(x) - \mu_{\nu, f_0}(x \mid X_n)| \lesssim_n n^{-\nu_0/d}. \tag{4.3}$$

The rate in (4.3) is worst-case optimal because from Propositions 3.6 and 3.7 and (2.6) it follows that the worst-case error in $H^{\nu_0 + d/2}(\Omega)$ decays with the same $n^{-\nu_0/d}$ rate. This means that oversmoothing does not have an adverse effect on the rate of convergence of the conditional mean. However, the model may provide overconfident uncertainty quantification because, by a combination of Propositions 3.6 and 3.7, the conditional standard deviation decays with a rate $n^{-\nu/d}$, which is (potentially) faster than the rate on the right-hand side in (4.3). Therefore it is possible that

$$\limsup_{n \to \infty} \frac{|f_0(x) - \mu_{\nu, f_0}(x \mid X_n)|}{\mathbb{V}_\nu(x \mid X_n)^{1/2}} = \infty$$

for some $x \in \Omega$, which implies there is no fixed ρ for which f_0 is contained in the credible set in (4.2) for all $n \in \mathbb{N}$.

Theorem 3.11 ensures that maximum likelihood estimation cannot can yield a Gaussian process model which is asymptotically undersmoothing. The conditional mean therefore tends to the response function with a worst-case optimal rate. However, there is much statistical literature on the impossibility of constructing, in the presence of noise, non-parametric confidence or credible bands (or sets) that are *adaptive* over natural smoothness classes, such as Sobolev or Hölder spaces, in the sense that (i) the band contains the truth with high probability and (ii) the radius of the band decays with the worst-case optimal rate for any truth in one of the smoothness classes (Low, 1997; Cai and Low, 2004; Robins and van der Vaart, 2006). While these results do

not directly apply to our setting, they do suggest that it is likely that more restrictive assumptions on f_0 than membership in Sobolev spaces are necessary if it is to be guaranteed that uncertainty quantification is not overconfident. In the context of the white noise and the Gaussian sequence space model, it is known that adaptivity can be guaranteed over self-similar and related classes that exclude "inconvenient" or "deceptive" response functions whose smoothness cannot be estimated (Bull, 2012; Szabó et al., 2015; Nickl and Szabó, 2016). Theorem 3.12 and these connections suggest that it is the class of self-similar Sobolev functions (or a closely related class) for which one should attempt to establish non-overconfidence of uncertainty quantification.

4.2 Scale Estimation for Infinitely Smooth Kernels

Let $\lambda > 0$ be a scale parameter and consider the infinitely smooth Gaussian kernel

$$K_\lambda(x, y) = \Phi_\lambda(x - y) = \Phi\left(\frac{x - y}{\lambda}\right), \quad \text{where} \quad \Phi(z) = \exp\left(-\frac{1}{2}\|z\|^2\right).$$

Rescaled Gaussian processes with Gaussian covariance kernel are studied in van der Vaart and van Zanten (2009), while results on scale parameter estimation for a kernel related to the Gaussian can be found in Hadji and Szábo (2021). The Fourier transform of Φ is

$$\widehat{\Phi}(\xi) = (2\pi)^{d/2}\exp\left(-\frac{1}{2}\|\xi\|^2\right),$$

and from the scaling properties of the Fourier transform we get

$$\widehat{\Phi}_\lambda(\xi) = \lambda^d\,\widehat{\Phi}(\lambda\xi) = (2\pi\lambda^2)^{d/2}\exp\left(-\frac{1}{2}\lambda^2\|\xi\|^2\right).$$

Recall from Section 3.1 that $H(K_\lambda)$ contains those square-integrable and continuous functions $f\colon \mathbb{R}^d \to \mathbb{R}$ for which

$$\|f\|_\lambda^2 = \int_{\mathbb{R}^d}\frac{|\widehat{f}(\xi)|^2}{\widehat{\Phi}_\lambda(\xi)}\,\mathrm{d}\xi = \frac{1}{(2\pi\lambda^2)^{d/2}}\int_{\mathbb{R}^d}\exp\left(\frac{1}{2}\lambda^2\|\xi\|^2\right)|\widehat{f}(\xi)|^2\,\mathrm{d}\xi < \infty. \tag{4.4}$$

From (4.4) we see that $H(K_{\lambda_0})$ is a proper subset of $H(K_{\lambda_1})$ whenever $\lambda_0 > \lambda_1$, which is a well-known result. The reasoning in Section 2.4 therefore suggests that (2.21) and (2.22) may hold if $f_0 \in H(K_{\lambda_0})$ for some $\lambda_0 > 0$ and, consequently,

$$\liminf_{n\to\infty}\hat{\lambda}_{\mathrm{ML}}^{f_0}(X_n) \geq \lambda_0 \quad \text{and} \quad \liminf_{n\to\infty}\hat{\lambda}_{\mathrm{CV}}^{f_0}(X_n) \geq \lambda_0.$$

Similar reasoning applies to many other infinitely smooth stationary kernels, such as the Cauchy kernel defined by $\Phi(z) = (1 + \|z\|^2)^{-1}$, whose Fourier transforms decay with super-algebraic rates.

However, to prove that either of the assumptions (2.21) or (2.22) holds for the Gaussian kernel does not appear to be possible at the moment. Upper bounds on the conditional variance of the form

$$\sup_{x\in\Omega}\mathbb{V}_\lambda(x \mid X_n) \lesssim_n \exp(-c_\lambda n^{1/d}\log n),$$

where and c_λ is a positive constant which depends on λ, are available in the scattered data approximation literature (e.g., Rieger and Zwicknagl, 2010, Theorem 6.1) if $\Omega \subset \mathbb{R}^d$ is bounded and sufficiently regular and the points $\{x_i\}_{i=1}^\infty$ are such that $h_{n,\Omega} \lesssim_n n^{-1/d}$, but no corresponding lower bounds exist. Moreover, the constant c_λ in these results is unlikely to be optimal. The result that is closest to being useful is a theorem in Karvonen (2022), which states that for the univariate Gaussian kernel we have

$$C_{1,\lambda}\frac{1}{(4\lambda^2)^n n!} \leq \sup_{x\in[-1,1]}\mathbb{V}_\lambda(x \mid X_n) \leq C_{2,\lambda}\,n^{-1/4}e^{2\sqrt{n}/\lambda}\frac{1}{(4\lambda^2)^n n!}$$

for sufficiently large n and positive constants $C_{1,\lambda}$ and $C_{2,\lambda}$ that depend on λ if $X_n = \{x_{n,i}\}_{i=1}^n$ are the non-nested Chebyshev nodes $x_{n,i} = \cos(\pi(i - 1/2)/n)$. See also Theorem 2 in Yarotsky (2013). In this setting the rate of convergence is controlled by λ because

$$n^{1/4} e^{-2\sqrt{n}/\lambda} \left(\frac{\lambda^2}{\lambda_0^2} \right)^n \lesssim_n \frac{\sup_{x \in [-1,1]} \mathbb{V}_{\lambda_0}(x \mid X_n)}{\sup_{x \in [-1,1]} \mathbb{V}_\lambda(x \mid X_n)} \lesssim_n n^{-1/4} e^{2\sqrt{n}/\lambda_0} \left(\frac{\lambda^2}{\lambda_0^2} \right)^n \qquad (4.5)$$

tends to zero if $\lambda_0 > \lambda$ and explodes if $\lambda_0 < \lambda$. However, Equation (4.5) is clearly not sufficient to establish either (2.21) or (2.22).

Acknowledgements

References

Anderes, E. (2010). On the consistent separation of scale and variance for Gaussian random fields. *The Annals of Statistics*, 38(2):870–893.

Arcangéli, R., de Silanes, M. C. L., and Torrens, J. J. (2007). An extension of a bound for functions in Sobolev spaces, with applications to (m, s)-spline interpolation and smoothing. *Numerische Mathematik*, 107(2):181–211.

Bachoc, F. (2013). Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69.

Bachoc, F., Lagnoux, A., and Nguyen, T. M. N. (2017). Cross-validation estimation of covariance parameters under fixed-domain asymptotics. *Journal of Multivariate Analysis*, 160:42–67.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer.

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. (2019). Probabilistic integration: A role in statistical computation? (with discussion and rejoinder). *Statistical Science*, 34(1):1–22.

Bull, A. D. (2011). Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12:2879–2904.

Bull, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics*, 6:1490–1516.

Cai, T. T. and Low, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32(5):1805–1840.

Chen, Y., Owhadi, H., and Stuart, A. M. (2021). Consistency of empirical Bayes and kernel flow for hierarchical parameter estimation. *Mathematics of Computation*, 90(332):2527–2578.

Driscoll, M. F. (1973). The reproducing kernel Hilbert space structure of the sample paths of a Gaussian process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 26(4):309–316.

Fong, E. and Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2):489–496.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *The Annals of Statistics*, 38(2):1122–1170.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Number 40 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Gualtierotti, A. F. (2015). *Detection of Random Signals in Dependent Gaussian Noise*. Springer.

Hadji, A. and Szábo, B. (2021). Can we trust Bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel? *SIAM/ASA Journal on Uncertainty Quantification*, 9(1):185–230.

Henderson, I. (2023). Sobolev regularity of Gaussian random fields. *Journal of Functional Analysis*. To appear.

Iske, A. (2018). *Approximation Theory and Algorithms for Data Analysis*. Springer.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1807.02582v1*.

Karvonen, T. (2022). Approximation in Hilbert spaces of the Gaussian and other weighted power series kernels. *arXiv:2209.12473v2*.

Karvonen, T. (2023). Small sample spaces for Gaussian processes. *Bernoulli*, 29(2):875–900.

Karvonen, T., Wynne, G., Tronarp, F., Oates, C. J., and Särkkä, S. (2020). Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):926–958.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Knapik, B. T., Szabó, B. T., van der Vaart, A. W., and van Zanten, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probability Theory and Related Fields*, 164:771–813.

Krieg, D. and Sonnleitner, M. (2023). Random points are optimal for the approximation of Sobolev functions. *IMA Journal of Numerical Analysis*. To appear.

Loh, W.-L. (2005). Fixed-domain asymptotics for a subclass of Matérn-type Gaussian random fields. *The Annals of Statistics*, 33(5):2344–2394.

Loh, W.-L. (2015). Estimating the smoothness of a Gaussian random field from irregularly spaced data via higher-order quadratic variations. *The Annals of Statistics*, 43(6):2766–2794.

Loh, W.-L. and Sun, S. (2023). Estimating the parameters of some common Gaussian random fields with nugget under fixed-domain asymptotics. *Bernoulli*, 29(3):2519–2534.

Loh, W.-L., Sun, S., and Wen, J. (2021). On fixed-domain asymptotics, parameter estimation and isotropic Gaussian random fields with Matérn covariance functions. *The Annals of Statistics*, 49(6):3127–3152.

Low, M. G. (1997). On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554.

Lukić, M. N. and Beder, J. H. (2001). Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969.

Narcowich, F. J., Ward, J. D., and Wendland, H. (2006). Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation*, 24(2):175–186.

Nickl, R. and Szabó, B. (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Processes and their Applications*, 126(12):3913–3934.

Novak, E. (1988). *Deterministic and Stochastic Error Bounds in Numerical Analysis*. Number 1349 in Lecture Notes in Mathematics. Springer-Verlag.

Novak, E. and Triebel, H. (2006). Function spaces in Lipschitz domains and optimal rates of convergence for sampling. *Constructive Approximation*, 23:325–350.

Paulsen, V. I. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Number 152 in Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Petit, S. (2023). An asymptotic study of the joint maximum likelihood estimation of the regularity and the amplitude parameters of a Matérn model on the circle. *arXiv:2209.07791v6*.

Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *The Annals of Statistics*, 28(1):298–335.

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

Rieger, C. and Zwicknagl, B. (2010). Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. *Advances in Computational Mathematics*, 32:103–129.

Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1):229–253.

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

Schaback, R. (2018). Superconvergence of kernel-based interpolation. *Journal of Approximation Theory*, 235:1–19.

Scheuerer, M. (2010). Regularity of the sample paths of a general second order random field. *Stochastic Processes and their Applications*, 120(10):1879–1897.

Scheuerer, M., Schaback, R., and Schlather, M. (2013). Interpolation of spatial data – A stochastic or a deterministic problem? *European Journal of Applied Mathematics*, 24(4):601–629.

Stein, E. M. (1970). *Singular Integrals and Differentiability Properties of Functions*. Number 30 in Princeton Mathematical Series. Princeton University Press.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer.

Steinwart, I. (2019). Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties. *Potential Analysis*, 51:361–395.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Information Science and Statistics. Springer.

Stuart, A. M. and Teckentrup, A. L. (2018). Posterior consistency for Gaussian process approximations of Bayesian posterior distributions. *Mathematics of Computation*, 87:721–753.

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

Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electronic Journal of Statistics*, 7:991–1018.

Szabó, B., van der Vaart, A. W., and van Zanten, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428.

Tuo, R., Wang, Y., and Wu, C. F. J. (2020). On the improved rates of convergence for Matérn-type kernel ridge regression with application to calibration of computer models. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4):1522–1547.

van der Vaart, A. and van Zanten, H. (2011). Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12(6):2095–2119.

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

Wang, W. and Jing, B.-Y. (2022). Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression. *Journal of Machine Learning Research*, 23(193):1–67.

Wang, W., Tuo, R., and Wu, C. F. J. (2020). On prediction properties of kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 115(530):920–930.

Wendland, H. (2005). *Scattered Data Approximation*. Number 17 in Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Wendland, H. and Rieger, C. (2005). Approximate interpolation with applications to selecting smoothing parameters. *Numerische Mathematik*, 101(4):729–748.

Wynne, G., Briol, F.-X., and Girolami, M. (2021). Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness. *Journal of Machine Learning Research*, 22(123):1–40.

Xu, W. and Stein, M. L. (2017). Maximum likelihood estimation for a smooth Gaussian random field model. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):138–175.

Yarotsky, D. (2013). Examples of inconsistency in optimization by expected improvement. *Journal of Global Optimization*, 56:1773–1790.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, 36(2):280–296.