Maximum Likelihood Estimation and Uncertainty Quantification for Gaussian Process Approximation of Deterministic Functions[*]

Toni Karvonen[†], George Wynne[‡], Filip Tronarp[§], Chris J. Oates[¶], and Simo Särkkä[‖]

Abstract. Despite the ubiquity of the Gaussian process regression model, few theoretical results are available that account for the fact that parameters of the covariance kernel typically need to be estimated from the dataset. This article provides one of the first theoretical analyses in the context of Gaussian process regression with a noiseless dataset. Specifically, we consider the scenario where the scale parameter of a Sobolev kernel (such as a Matérn kernel) is estimated by maximum likelihood. We show that the maximum likelihood estimation of the scale parameter alone provides significant adaptation against misspecification of the Gaussian process model in the sense that the model can become "slowly" overconfident at worst, regardless of the difference between the smoothness of the data-generating function and that expected by the model. The analysis is based on a combination of techniques from nonparametric regression and scattered data interpolation. Empirical results are provided in support of the theoretical findings.

Key words. nonparametric regression, scattered data approximation, credible sets, Bayesian cubature, model misspecification

AMS subject classifications. 60G15, 62G20, 68T37, 65D05, 46E22

1. Introduction. This article considers the related tasks of approximation and integration of a *deterministic* function $f \colon \Omega \to \mathbb{R}$, defined on $\Omega \subset \mathbb{R}^d$, using Gaussian process (GP) regression based on a noiseless dataset $\mathcal{D} \coloneqq \{(x_n, f(x_n))\}_{n=1}^N$. In GP regression the true function f is formally considered unknown and is modelled *a priori* with a GP $f_{\mathrm{GP}} \sim \mathrm{GP}(m, K)$, which is characterised by a *mean* function $m \colon \Omega \to \mathbb{R}$ and a symmetric positive (semi-)definite covariance function $K \colon \Omega \times \Omega \to \mathbb{R}$, called a *kernel*. The GP is conditioned on the dataset $\mathcal{D}$ and the conditional GP is used to produce credible sets for quantities of interest, such as the function f itself or its integral. The popularity of the GP model can be attributed, at least in part, to its elegance, flexibility and computational tractability, and as such GPs underpin much of the modern statistical toolkit for both regression and classification (Rasmussen and Williams, 2006). In the last decade GPs have been adopted in a wide variety of applications, a selection of which includes time series analysis (Wang et al., 2006), astrophysical data analysis (Rajpaul et al., 2015), spatial statistics (Lindgren et al., 2011), bioinformatics (Gao et al., 2008), robotics (Yang

et al., 2013), functional data analysis (Shi and Wang, 2008), computer science (Manogaran and Lopez, 2018), emulation of computer models (Sacks et al., 1989; Kennedy and O'Hagan, 2001), and probabilistic numerical computation (Larkin, 1972; Hennig et al., 2015; Cockayne et al., 2019).

The GP model is typically *misspecified*: the deterministic data-generating function f is not, or does not "resemble", a sample path of f_{GP}. Accordingly, the critical importance of selecting an appropriate kernel K in GP regression is well-understood (MacKay, 1992). Different approaches include selecting a single kernel from a continuously parametrised family $\{K^{\theta}\}_{\theta \in \Theta}$ (Rasmussen and Williams, 2006, Chapter 5), selecting a kernel from an arbitrarily rich dictionary of possibilities (Duvenaud, 2014; Sun et al., 2018), or even learning a kernel in a nonparametric manner from the data itself (Băzăvan et al., 2012; Oliva et al., 2016). In the parametric case, maximisation of (marginal) likelihood is the most common way to select the kernel parameter θ, for example being the default in well-documented software packages (e.g., Rasmussen and Williams, 2006). Despite their ubiquity in the applied context, little is known about the circumstances in which these approaches to kernel parameter selection work well and, by extension, when the credible sets arising from the GP regression model can be trusted. The increasing use of GP regression models and their associated credible sets in strategic and safety-critical systems, such as monitoring mine gas emissions (Dong, 2012), assessing the health of lithium-ion batteries (Liu et al., 2013), and detecting anomalous or malicious maritime activity (Kowalska and Peel, 2012), as well as in more general adaptive numerical computation routines (e.g., Rathinavel and Hickernell, 2019), has led to an urgent need to better understand approaches to kernel parameter selection and model misspecification at a theoretical level.

This article shows that one of the simplest and most commonly used techniques, maximum likelihood estimation of a single *scale parameter* of the kernel, provides a certain amount of protection against model misspecification. We consider a kernel $K^{\sigma}(x, y) := \sigma^2 K(x, y)$ that depends on a scale parameter $\sigma > 0$ and analyse the asymptotic (as $N \to \infty$) behaviour of $\sigma_{\mathrm{ML}}(f, X_N)$, the *maximum likelihood estimate* of σ given noiseless evaluations of f at a set $X_N \subset \Omega$ consisting of N points, and implications on the coverage of the credible sets derived from the fitted GP model. For finitely smooth kernels (e.g., Matérn) we show that the maximum likelihood estimate detects the smoothness of the data-generating function: if K induces a Sobolev space of smoothness α, f is in a certain sense *exactly* of smoothness $\beta \leq \alpha$, and the points X_N cover Ω in a sufficiently uniform manner, then $\sigma_{\mathrm{ML}}(f, X_N)$ is of order $N^{(\alpha-\beta)/d-1/2}$, up to logarithmic factors. Because f being akin to a sample of f_{GP} roughly speaking corresponds to $\beta = \alpha - d/2$ (see Section 4.2), the maximum likelihood estimate inflates the conditional variance if f is rougher than the samples and deflates it f is smoother than the samples. If f is in the Sobolev space of smoothness $\beta \geq \alpha$, then $\sigma_{\mathrm{ML}}(f, X_N)$ is of order $N^{-1/2}$. We then use these result to prove that, no matter the degree of over or undersmoothing of f by the kernel, the model can become at most "slowly" overconfident in that the GP conditional standard deviation can decay at most with rate $N^{-1/2}$ faster than the true estimation error. If the scale parameter is held fixed (4.10) demonstrates that the model may become significantly more overconfident than this.

The results are reviewed in more detail in Section 2.7. Section 3 considers the case where f is an element of the reproducing kernel Hilbert space of the kernel K and therefore smoother than expected by the GP. Section 4 extends the results for kernels that induce Sobolev spaces by

allowing the function to live in a rougher Sobolev space than the one induced by the kernel, in which case the results are dependent on the degree of oversmoothing by the kernel. Numerical examples are used to validate the theoretical results in Section 5. In most applications of GP regression there will be several kernel parameters in addition of the scale parameter that must be jointly estimated; our analysis does not extend to that more general setting. The opportunities and challenges associated with estimation of other kernel parameters are discussed in Section 6.

2. Background. In this section we introduce the GP regression model and recall how the kernel scale parameter can be estimated using maximum likelihood. Then we discuss how credible sets can be obtained based on the fitted GP model and what it means to say that the model is asymptotically underconfident or overconfident. For the latter, we focus on credible sets both for function values and integrals of the function of interest.

2.1. Gaussian Process Regression. Let Ω be an arbitrary subset of $\mathbb{R}^d$ and $f : \Omega \to \mathbb{R}$ a deterministic function of interest. In GP regression the function f is modelled using a GP f_{GP}, for which $(f_{\mathrm{GP}}(x_1), \ldots, f_{\mathrm{GP}}(x_N))$ is Gaussian-distributed for any finite collection $\{x_1, \ldots, x_N\} \subset \Omega$ of points. Let $\mathbb{P}$ denote the law of the GP and let $\mathbb{E}$, $\mathbb{V}$, and $\mathbb{C}$, respectively, denote the expectation, variance, and covariance with respect to $\mathbb{P}$. The law $\mathbb{P}$ of a GP is characterised by a mean function $m : \Omega \to \mathbb{R}$, such that $m(x) = \mathbb{E}[f_{\mathrm{GP}}(x)]$ for all $x \in \Omega$, and a symmetric positive definite covariance function $K : \Omega \times \Omega \to \mathbb{R}$, called a kernel, such that $K(x, y) = \mathbb{C}[f_{\mathrm{GP}}(x), f_{\mathrm{GP}}(y)]$ for all $x, y \in \Omega$. Although the kernel can be allowed to be positive semi-definite, in this article we only consider positive definite kernels. It is common to denote the GP via the shorthand $f_{\mathrm{GP}} \sim \mathrm{GP}(m, K)$. Throughout the article and without a loss of generality[1] we assume that f_{GP} is centred (i.e., $m(x) = 0$ for all $x \in \Omega$). Further details on GP regression can be found in Bogachev (1998); Stein (1999); and Rasmussen and Williams (2006).

Given a set $\mathcal{D} = \{(x_i, f(x_i))\}_{i=1}^N$ consisting of *exact* evaluations of f at distinct points $X = \{x_1, \ldots, x_N\} \subset \Omega$, the conditional process is again Gaussian: $f_{\mathrm{GP}} \mid \mathcal{D} \sim \mathrm{GP}(s_{f,X}, P_X)$, with the conditional mean and covariance functions

$$(2.1) \qquad s_{f,X}(x) := k_X(x)^\mathsf{T} K_X^{-1} f_X \quad \text{and} \quad P_X(x, y) := K(x, y) - k_X(x)^\mathsf{T} K_X^{-1} k_X(y),$$

where $(k_X(x))_i = K(x, x_i)$, $(K_X)_{i,j} = K(x_i, x_j)$, and $(f_X)_i = f(x_i)$. The conditional process quantifies the uncertainty associated with f after the data $\mathcal{D}$ have been observed and can be summarised in terms of a *credible set* for a quantity of interest. Let F stand for the cumulative distribution function of the standard normal distribution and denote $\psi_a := F^{-1}(1 - a/2)$. For any $0 < a < 1$, the Gaussian model implies that

$$(2.2) \qquad \mathbb{P}\left[\left|f_{\mathrm{GP}}(x) - s_{f,X}(x)\right| \le \psi_a P_X(x, x)^{1/2} \mid \mathcal{D}\right] = 1 - a \quad \text{for any} \quad x \in \Omega.$$

Thus (if $P_X(x, x) \ne 0$) the interval bounded by $s_{f,X}(x) \pm \psi_a P_X(x, x)^{1/2}$ is a $(1 - a)$ credible set for the unknown quantity $f(x)$ at fixed $x \in \Omega \setminus X$ under the GP model. However, as is evident from its algebraic expression in (2.1), the conditional covariance P_X does *not* depend

---

[1] If m is non-zero then the true function f can just be replaced by $f - m$, since m is considered to be known and can thus $f - m$ can also be pointwise evaluated.

on the function evaluations f_X, which is clearly undesirable as this implies that the size of the credible set is identical for two wildly different functions evaluated at the same inputs X. It is well-understood that, for sensible uncertainty quantification to be performed, the kernel should be adapted to the dataset (MacKay, 1992). When the kernel is parametrised by a collection of parameters θ (i.e., $K = K^\theta$), this means that θ should be estimated based on the dataset. Standard approaches to estimation of θ are reviewed in Section 2.3.

2.2. Bayesian Cubature. It is convenient to consider and easier to visualise credible sets for scalar quantities derived from f, rather than f itself.[2] Moreover, approximation of integrals (i.e., numerical integration) is among the most prevalent applications where noiseless data are provided. For these reasons we also focus on integrals of f as scalar quantities of interest. The use of the GP regression model as a means to perform numerical integration is called *Bayesian cubature* (*quadrature* if $d = 1$) and is due to Larkin (1972). See also O'Hagan (1991) and Briol et al. (2019) for background. For a Lebesgue measurable[3] $\Omega \subset \mathbb{R}^d$ and a positive, bounded, and measurable weight function $w \colon \Omega \to \mathbb{R}$ we consider the integral

$$(2.3) \qquad I(f) := \int_\Omega f(x)w(x)\,\mathrm{d}x$$

as a scalar quantity of interest. Because the integration operator is a linear functional, the random variable $I(f_{\mathrm{GP}}) \mid \mathcal{D}$ is Gaussian if $\int_\Omega K(x,x)w(x)\,\mathrm{d}x < \infty$. Its mean and variance are

$$(2.4a) \qquad Q_X(f) := \mathbb{E}\big[I(f_{\mathrm{GP}}) \mid \mathcal{D}\big] = \int_\Omega s_{f,X}(x)w(x)\,\mathrm{d}x,$$

$$(2.4b) \qquad V_X := \mathbb{V}\big[I(f_{\mathrm{GP}}) \mid \mathcal{D}\big] = \int_\Omega \int_\Omega P_X(x,y)w(x)w(y)\,\mathrm{d}x\,\mathrm{d}y.$$

The Gaussian model for the integral then implies that

$$(2.5) \qquad \mathbb{P}\Big[\big|I(f_{\mathrm{GP}}) - Q_X(f)\big| \le \psi_a V_X^{1/2} \,\Big|\, \mathcal{D}\Big] = 1 - a$$

and thus (if $V_X \neq 0$) the interval bounded by $Q_X(f) \pm \psi_a V_X^{1/2}$ is a $(1 - a)$ credible set, or *credible interval* for $I(f)$ under the GP model.

2.3. Scale Parameter Estimation. In this section we consider the case where the kernel $K^\sigma(x,y) = \sigma^2 K(x,y)$ depends on a single fixed scale parameter $\sigma > 0$. Under the law $f_{\mathrm{GP}} \sim \mathrm{GP}(0, K^\sigma)$ the conditional distribution is $f_{\mathrm{GP}} \mid \mathcal{D} \sim \mathrm{GP}(s_{f,X}, P_X^\sigma)$, where the conditional mean remains unchanged from (2.1) and the covariance is

$$P_X^\sigma(x,y) := \sigma^2 P_X(x,y) = \sigma^2 \big[K(x,y) - k_X(x)^\mathsf{T} K_X^{-1} k_X(y)\big].$$

---

[2]Indeed, unlike the scalar case there is no general consensus on how one should aim to construct a credible set in a function space; see for example Liebl and Reimherr (2019).

[3]Whenever Bayesian cubature is discussed or results for it are provided, it is implicitly assumed in this article that Ω is Lebesgue measurable.

The purpose of this article is to analyse the maximum likelihood estimate, $\sigma_{\mathrm{ML}}(f, X)$, of σ and its effect on the credible sets (2.2) and (2.5). The MLE is defined as the maximiser

$$(2.6) \qquad \sigma_{\mathrm{ML}}(f, X) := \arg\max_{\sigma > 0} L(\sigma \mid \mathcal{D}) = \sqrt{\frac{f_X^{\mathsf{T}} K_X^{-1} f_X}{N}}.$$

of the log marginal likelihood,

$$\log L(\sigma \mid \mathcal{D}) := -\frac{1}{2} \left(\frac{f_X^{\mathsf{T}} K_X^{-1} f_X}{\sigma^2} + N \log \sigma^2 + \log \det K_X + N \log(2\pi) \right).$$

Equation (2.6) is easy to verify by finding the root of the derivative of $L(\sigma \mid \mathcal{D})$. The estimator $\sigma_{\mathrm{ML}}(f, X)$ is sometimes called a *maximum marginal likelihood* or *empirical Bayes* estimator. In applications where additional parameters are present in the kernel, these could be simultaneously estimated based on the dataset. However, our focus on the scale parameter is due to the closed-form expression in (2.6); such expressions are not be available in general.

2.4. Credible Sets and Maximum Likelihood. Adopting the maximum likelihood approach to parameter selection means that σ is replaced by $\sigma_{\mathrm{ML}}(f, X)$ in (2.2) and (2.5) to produce

$$\mathbb{P}\left[\left| f_{\mathrm{GP}}(x) - s_{f,X}(x) \right| \le \psi_a \sigma_{\mathrm{ML}}(f, X) P_X(x, x)^{1/2} \,\middle|\, \mathcal{D} \right] = 1 - a,$$

$$\mathbb{P}\left[\left| I(f_{\mathrm{GP}}) - Q_X(f) \right| \le \psi_a \sigma_{\mathrm{ML}}(f, X) V_X^{1/2} \,\middle|\, \mathcal{D} \right] = 1 - a.$$

We use the compact notation

$$(2.7) \qquad R_{\mathrm{GP}}(x, f, X) := \sigma_{\mathrm{ML}}(f, X) P_X(x, x)^{1/2} \quad \text{and} \quad R_{\mathrm{BC}}(f, X) := \sigma_{\mathrm{ML}}(f, X) V_X^{1/2}$$

for the unscaled widths of the credible sets and denote the credible sets as

$$(2.8a) \qquad \mathcal{C}_{\mathrm{GP}}^a(x, f, X) := \left\{ y \in \mathbb{R} \,:\, \frac{|y - s_{f,X}(x)|}{R_{\mathrm{GP}}(x, f, X)} \le \psi_a \right\},,$$

$$(2.8b) \qquad \mathcal{C}_{\mathrm{BC}}^a(f, X) := \left\{ \mu \in \mathbb{R} \,:\, \frac{|\mu - Q_X(f)|}{R_{\mathrm{BC}}(f, X)} \le \psi_a \right\}.$$

These credible sets underpin inferences and decisions based on the fitted GP regression model, with applications in diverse fields, including strategic and safety-critical systems, several of which were mentioned in Section 1. It is therefore important to understand when these sets can and cannot be trusted to accurately reflect the function f or its integral.

It is immediately clear from (2.6) that credible sets are invariant to scaling of f, in the sense that the transformation $f \mapsto \lambda f$ for some constant λ leads to $\sigma_{\mathrm{ML}}(f, X) \mapsto |\lambda| \sigma_{\mathrm{ML}}(f, X)$. However, it is far from clear how these credible sets behave as a function of the point set X. In particular, we consider the limit of a large number of points next.

2.5. Asymptotics of Credible Sets. Consider a sequence $(X_N)_{N=1}^\infty \subset \Omega$ of point sets such that X_N contains N distinct points. The function f is fixed and our focus is on the behaviour of credible sets when $N \to \infty$, a setting called *fixed domain asymptotics* by Stein (1999). Specifically, we are interested in whether or not $f(x)$ or $I(f)$ can be expected to fall within the relevant credible set, $\mathcal{C}_{\mathrm{GP}}^a(x, f, X_N)$ or $\mathcal{C}_{\mathrm{BC}}^a(f, X_N)$, for large N. To avoid confusion, it is important to note that our focus is distinct from the assessment of *frequentist coverage* that is more commonplace in the statistical literature. There, it is most common for N and f to be fixed and for observations of f to be contaminated with noise; one can then ask for credible sets to have correct coverage with respect to realisations of the noise generating process. Equally, our analysis is distinct from an assessment of frequentist coverage in which f is considered to be drawn at random from $\mathbb{P}$ and observed (without noise) at N locations. To emphasise, in this article the dataset $\mathcal{D}$ and function f are deterministic and the only source of uncertainty is the *epistemic uncertainty* from the GP regression model.

We say that a GP model with a covariance kernel K is *asymptotically overconfident* for approximation at $x \in \Omega$ (respectively, integration) of a function $f\colon \Omega \to \mathbb{R}$ if

$$(2.9) \qquad \liminf_{N\to\infty} \frac{|f(x) - s_{f,X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} = \infty \qquad \left(\liminf_{N\to\infty} \frac{|I(f) - Q_{X_N}(f)|}{R_{\mathrm{BC}}(f, X_N)} = \infty \right)$$

and *asymptotically underconfident* if

$$(2.10) \qquad \lim_{N\to\infty} \frac{|f(x) - s_{f,X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} = 0 \qquad \left(\lim_{N\to\infty} \frac{|I(f) - Q_{X_N}(f)|}{R_{\mathrm{BC}}(f, X_N)} = 0 \right).$$

Conforming to conventional statistical terminology we call the ratios in (2.9) and (2.10) *standard scores*. Note that $R_{\mathrm{GP}}(x, f, X_N) = 0$ only if $f(x) = s_{f,X_N}(x)$; in this case we set $0/0 = 1$. Asymptotic overconfidence means that the width of the credible set decays faster than the true approximation or integration error: for any fixed $a \in (0,1)$ we have $f(x) \notin \mathcal{C}_{\mathrm{GP}}^a(x, f, X_N)$ or $I(f) \notin \mathcal{C}_{\mathrm{BC}}^a(f, X_N)$ for all sufficiently large N. Conversely, asymptotic underconfidence implies that for any $a \in (0,1)$ we have $f(x) \in \mathcal{C}_{\mathrm{GP}}^a(x, f, X_N)$ or $I(f) \in \mathcal{C}_{\mathrm{BC}}^a(f, X_N)$ for all N large enough.

Overconfidence can have disastrous effect in particular in safety-critical applications while underconfidence results in inefficiency as more data than is necessary is needed to attain the same level of assurance. The ideal state of affairs is thus for the model to be neither asymptotically overconfident nor underconfident, a situation which we call *asymptotic honest* as this implies that the size of the credible sets decay at rate that is commensurate with the true approximation error. See Szabó et al. (2015) for a similar concept. In practice asymptotic honesty is a weak requirement and does not guarantee credible sets can be trusted at finite values of N. Our tools are not powerful enough to identify or prove the existence of meaningful collections of functions for which the model is asymptotically honest, and our results concern only asymptotic overconfidence and underconfidence.

2.6. Prior Work on Maximum Likelihood Estimation. The only prior work in an identical setting, to the best of our knowledge, is by Xu and Stein (2017) and Karvonen et al. (2019). Xu and Stein (2017) considered the Gaussian kernel $K(x, y) = \exp(-(x-y)^2/(2\ell^2))$ with $\ell > 0$

fixed and monomials $f(x) = x^p$ on $[0, 1]$, evaluated at successive sets of N equispaced points, $X_N = \{1/N, 2/N, \ldots, 1\}$. They conjectured an asymptotic equivalence

$$\sigma_{\mathrm{ML}}(f, X_N) \sim \frac{\ell^{2p}}{\sqrt{2\pi}(p + 1/2)} N^{p - 1/2} \quad \text{for any} \quad p \geq 0$$

and proved this for $p = 0$ and partially for $p = 1$ using an explicit Cholesky decomposition of the kernel matrix. Karvonen et al. (2019) worked with the Ornstein–Uhlenbeck kernel $K(x, y) = \exp(-\lambda |x - y|) - \exp(-\lambda(x + y))$, with $\lambda > 0$ fixed, and equispaced evaluation points on $[0, 1]$. They proved that $\lim_{N \to \infty} \sigma_{\mathrm{ML}}(f, X_N)$ is proportional to the quadratic variation $V^2(f)$ of f. Consequently, the MLE converges to zero if the Hölder exponent of f exceeds $1/2$ (e.g., the function is differentiable) and to a positive constant if $V^2(f) \in (0, \infty)$. As almost all sample paths of the Ornstein–Uhlenbeck process have a finite non-zero quadratic variation, this is in agreement with the intuition that the maximum likelihood estimate should behave reasonably if the function is plausible as a sample from the GP.

In addition, frequentist coverage of Bayesian credible sets when various hyperparameters of a GP are selected with maximum likelihood has been extensively studied by Szabó et al. (2013, 2015) and Hadji and Szabó (2019). In these articles the model of interest differs from ours, being the Gaussian *white noise model* for an unknown function $f(x) = \sum_{i=1}^{\infty} \vartheta_i \varphi_i(x)$ expressed in a basis $\{\varphi_i\}_{i=1}^{\infty}$. A sequence $Y = (Y_i)_{i=1}^{\infty}$ of noisy observations are made directly on the square-summable parameter $\vartheta = (\vartheta_i)_{i=1}^{\infty}$ via

$$Y_i = \vartheta_i + \frac{1}{\sqrt{\eta}} Z_i, \quad \text{where} \quad Z_i \sim \mathrm{N}(0, 1) \text{ are i.i.d.}$$

In Szabó et al. (2013, 2015) the parameter ϑ was assigned a Gaussian prior distribution that is analogous to GPs with Sobolev kernels that we analyse. Behaviour as $\eta \to \infty$ (i.e., the noise level decreases) of the MLE of the scaling parameter of this prior and the coverage properties of the resulting credible sets were analysed in Szabó et al. (2013) for the true parameter satisfying $\vartheta_i^2 \leq C_2^2 i^{-1-2\beta}$ or $C_1^2 i^{-1-2\beta} \leq \vartheta_i^2 \leq C_2^2 i^{-1-2\beta}$ for some $C_1, C_2 > 0$ and a smoothness parameter $\beta > 0$. These sets are analogous to our $S_-^{\beta}(\mathbb{R}^d)$ and $S^{\beta}(\mathbb{R}^d)$ defined in Section 4.1. The white noise model is widely studied as a theoretically tractable analogue of regression with noisy data. As such the results are not directly applicable in our context where the function f is exactly evaluated.

For other work related to GP misspecification and kernel parameter estimation in a variety of settings, see Stein (1993); Bachoc (2013); Bachoc et al. (2017); Bachoc (2018); Bachoc et al. (2019); and Teckentrup (2019).

2.7. Our Contributions. Let $(X_N)_{N=1}^{\infty} \subset \Omega$ be a sequence of point sets, each containing N distinct points, and let the function $f \colon \Omega \to \mathbb{R}$ be fixed. Our results concern (i) the behaviour, as $N \to \infty$, of the maximum likelihood estimate, $\sigma_{\mathrm{ML}}(f, X_N)$ in (2.6), of the GP scale parameter based on exact evaluation of f on X_N and (ii) whether or not this induces asymptotic overconfidence or underconfidence in the GP model, as defined in (2.9) and (2.10).

Reproducing kernel Hilbert spaces. In Section 3 we do not place any restrictions on the covariance kernel K. We first prove the surprising result that if f is an element of $\mathcal{H}_K(\Omega)$, the

reproducing kernel Hilbert space of K, then

$$\sigma_{\mathrm{ML}}(f, X_N) \asymp N^{-1/2} \tag{2.11}$$

regardless of the point sets X_N used, provided the X_N share a common element x^* such that $f(x^*) \neq 0$ (Proposition 3.1). Theorem 3.2, an implication of this, states that for such functions and point sets the model cannot become overconfident "too fast", meaning that

$$\sup_{x \in \Omega} \frac{|f(x) - s_{f,X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} = O(N^{1/2}) \quad \text{and} \quad \frac{|I(f) - Q_{X_N}(f)|}{R_{\mathrm{BC}}(f, X_N)} = O(N^{1/2}). \tag{2.12}$$

Note that this does *not* imply that the model is asymptotically overconfident. Indeed, in Theorem 3.4 we show that *underconfidence* occurs if f belongs to a certain subspace of $\mathcal{H}_K(\Omega)$.

Sobolev spaces. Section 4 focusses on *Sobolev kernels*, which induce Sobolev spaces and include the popular Matérn kernels. The restrictive assumption $f \in \mathcal{H}_K(\Omega)$ is relaxed and it is proven in Proposition 4.5 that if K induces a Sobolev space of smoothness α and f is in the Sobolev space of smoothness $\beta < \alpha$, then

$$\sigma_{\mathrm{ML}}(f, X_N) = O\big(N^{(\alpha-\beta)/d - 1/2}\big) \tag{2.13}$$

assuming that X_N cover the domain Ω in a uniform fashion; (2.11) is applicable if $\beta \geq \alpha$. Moreover, a similar lower bound is available when a lower bound on the smoothness of f is known (Proposition 4.7). If it is known that f is of *exact smoothness* $\beta \leq \alpha$, in that it belongs to the set $S^\beta(\Omega)$ in (4.4), then the rate (2.13) is sharp up to logarithmic factors (Theorem 4.9). In particular, if f is of exact smoothness $\beta = \alpha - d/2$, which roughly speaking corresponds to f having the same regularity as samples from the GP (see Section 4.2), then $\sigma_{\mathrm{ML}}(f, X_N)$ is constant up to logarithmic factors. If the exact smoothness of f is known, bounds similar to (2.12) on the standard scores then hold by Theorem 4.10. These results thus show that maximum likelihood estimation of the scale parameter is a useful tool in adapting the GP model to misspecified smoothness of the data-generating function. Finally, according to Theorem 4.11, f being much smoother than the kernel implies underconfidence of the GP model.

Empirical results in Section 5 verify the MLE asymptotics (2.11) and (2.13) but suggest that the standard score bounds (2.12) and their extensions in the Sobolev setting are not tight. Although sufficient conditions for asymptotic honesty of a GP model are not provided here, the collection of results that we establish represents a substantial expansion of what is currently known in the context of maximum likelihood estimation with a noiseless dataset.

2.8. Notation. For $x \in \mathbb{R}^d$ we let $\|x\| := (x_1^2 + \cdots + x_d^2)^{1/2}$ be the Euclidean norm. The space $L^p(\Omega)$ stands for the space of p-integrable functions on a Lebesgue measurable set $\Omega \subset \mathbb{R}^d$. For non-negative sequences $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$ we denote $a_n \lesssim b_n$ ($a_n \gtrsim b_n$) if there is $C > 0$ such that $a_n \leq C b_n$ ($a_n \geq C b_n$) for every sufficiently large n. When $a_n \lesssim b_n \lesssim a_n$, we write $a_n \asymp b_n$. Analogous notation is used for non-negative functions. For example, $h(x) \lesssim g(x)$ means that there is $C > 0$ such that $h(x) \leq C g(x)$ for all sufficiently large $\|x\|$.

The *restriction* of a function $g \colon A \to \mathbb{R}$ on a subset $B \subset A$ is the function $g|_B \colon B \to \mathbb{R}$ such that $g|_B(b) = f(b)$ for every $b \in B$. In particular, the statement that $h|_X = g|_X$ for a

set $X \subset \mathbb{R}^d$ means that the function g *interpolates* h on X. Conversely, if $g\colon A \to \mathbb{R}$ and $h\colon B \to \mathbb{R}$ are such that $g|_B = h$, then g is said to be an *extension* of h (onto A).

In what follows the set $X = \{x_1, \dots, x_N\}$ always denotes a collection of $N \in \mathbb{N}$ distinct points contained in the domain $\Omega \subset \mathbb{R}^d$ of the function f of interest. If it is necessary to emphasise the number of points in the set, we write X_N for a set of N points.

3. Approximation of Functions in the RKHS. In this section we introduce reproducing kernel Hilbert spaces (RKHSs) and study the maximum likelihood estimate and implications for the standard scores when f is regular enough to be contained in the RKHS of the covariance kernel. Results for less regular functions are deferred until Section 4.

3.1. Positive-Definite Kernels and RKHSs. The monograph of Berlinet and Thomas-Agnan (2004) is a standard introduction to the theory of RKHS. Let Ω be an arbitrary subset of $\mathbb{R}^d$. We say that a function $K\colon \Omega \times \Omega \to \mathbb{R}$ is a kernel (on Ω) if it is *positive-definite*. Positive-definiteness entails that, for any $N \in \mathbb{N}$, the $N \times N$ *kernel matrix* $(K_X)_{i,j} = K(x_i, x_j)$ is positive-definite for any set $X = \{x_1, \dots, x_N\} \subset \mathbb{R}^d$ of N distinct points. Every kernel induces a unique *reproducing kernel Hilbert space* $\mathcal{H}_K(\Omega)$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K(\Omega)}$ and the induced norm $\|\cdot\|_{\mathcal{H}_K(\Omega)}$. This space consists of certain sufficiently regular functions $g\colon \Omega \to \mathbb{R}$ and is characterised by

(i) $K(\cdot, x) \in \mathcal{H}_K(\Omega)$ for every $x \in \Omega$ and

(ii) $\langle g, K(\cdot, x) \rangle_{\mathcal{H}_K(\Omega)} = g(x)$ for every $g \in \mathcal{H}_K(\Omega)$ and $x \in \Omega$ (the *reproducing property*).

Note the RKHS $\mathcal{H}_K(\Omega)$ and its norm are always those of the "unscaled" kernel K. That is, they do not depend on the scale parameter σ.

Throughout the article we assume that K is a kernel. In this section the kernel K is arbitrary, meaning that it is not necessarily straightforward to verify if a given function is contained in its RKHS. However, in Section 4 the kernel is selected such that the RKHS is a Sobolev space so that the differentiability of a function determines if it is a member of the RKHS. We occasionally define the kernel on the whole of $\mathbb{R}^d$ and then consider the restriction of $\mathcal{H}_K(\mathbb{R}^d)$ to $\Omega \subset \mathbb{R}^d$. The restriction consists of functions $g\colon \Omega \to \mathbb{R}$ that admit an extension $g_0 \in \mathcal{H}_K(\mathbb{R}^d)$ and its norm is

$$\|g\|_{\mathcal{H}_K(\Omega)} := \inf\big\{\|g_0\|_{\mathcal{H}_K(\mathbb{R}^d)} \,:\, g_0 \in \mathcal{H}_K(\mathbb{R}^d) \text{ such that } g_0|_\Omega = g\big\}.$$

3.2. Kernel Interpolation and Error Estimates. It is necessary to recognise the equivalence of GP regression and *kernel* or *radial basis function interpolation* (Wendland, 2005; Fasshauer and McCourt, 2015): the GP conditional mean (2.1) is the kernel interpolant to f at X, which is to say that it is the unique function g in $\mathrm{span}\{K(\cdot, x_i)\}_{i=1}^N$ such that $g|_X = f|_X$. Equivalently, $s_{f,X}$ is the interpolant to f of minimal norm among the functions in the RKHS of the kernel:

$$(3.1) \qquad s_{f,X} = \arg\min\big\{\|g\|_{\mathcal{H}_K(\Omega)} \,:\, g \in \mathcal{H}_K(\Omega) \text{ such that } g|_X = f|_X\big\}.$$

This property implies in particular that $\|s_{f,X}\|_{\mathcal{H}_K(\Omega)} \leq \|f\|_{\mathcal{H}_K(\Omega)}$ if $f \in \mathcal{H}_K(\Omega)$. If $f \notin \mathcal{H}_K(\Omega)$, the conditional mean is still an element of the RKHS but its norm diverges to infinity as X becomes denser. Further discussion on the relationship between GP regression and kernel-based minimum-norm interpolation can be found in Scheuerer et al. (2013); Kanagawa et al.

(2018); Karvonen (2019); and Fasshauer and McCourt (2015, Chapter 17). Oettershagen (2017, Chapter 3) contains a compact collection of basic results on approximation in RKHS.

The RKHS framework is useful in deriving generic estimates for GP approximation or integration error. The conditional variances (2.1) and (2.4b) are equal to squared *worst-case errors* in function and integral approximation in the RKHS of the covariance kernel:

$$(3.2) \quad P_X(x,x)^{1/2} = \sup_{\|g\|_{\mathcal{H}_K(\Omega)} \leq 1} |g(x) - s_{g,X}(x)| \quad \text{and} \quad V_X^{1/2} = \sup_{\|g\|_{\mathcal{H}_K(\Omega)} \leq 1} |I(g) - Q_X(g)|.$$

Furthermore, the reproducing property of the kernel can be used in bounding the approximation or integration error for a specific function $f \in \mathcal{H}(K)$ using the standard deviations:

$$(3.3) \quad |f(x) - s_{f,X}(x)| \leq \|f\|_{\mathcal{H}_K(\Omega)} P_X(x,x)^{1/2} \quad \text{and} \quad |I(f) - Q_X(f)| \leq \|f\|_{\mathcal{H}_K(\Omega)} V_X^{1/2}.$$

3.3. Maximum Likelihood Estimation in the RKHS. In this section we study the maximum likelihood estimator $\sigma_{\mathrm{ML}}(f, X_N)$ and asymptotic underconfidence and overconfidence of the GP model when the function f is sufficiently regular to belong to $\mathcal{H}_K(\Omega)$.[4]

All results in this article are based on the following expression for the MLE that, simple as it is, appears to have been seldom exploited in the GP literature:

$$(3.4) \qquad\qquad \sigma_{\mathrm{ML}}(f, X_N) = \frac{1}{\sqrt{N}} \|s_{f,X_N}\|_{\mathcal{H}_K(\Omega)}.$$

Note that this equation does not require that $f \in \mathcal{H}_K(\Omega)$. This connection between the maximum likelihood estimate of the scale parameter and the RKHS norm of the conditional mean is made explicit in Fasshauer and McCourt (2015, Remark 9.2), and the straightforward proof, based on the reproducing property and equations (2.1) and (2.6), can also be found in, for example, Fasshauer (2011, Section 5.1). Bull (2011, Section 3.3) uses (3.4) in the context of Bayesian optimisation. Equation (3.4) leads immediately to our first result for $f \in \mathcal{H}_K(\Omega)$.

Proposition 3.1 (MLE in the RKHS). *If $f \in \mathcal{H}_K(\Omega)$, then $\sigma_{\mathrm{ML}}(f, X_N) \leq N^{-1/2} \|f\|_{\mathcal{H}_K(\Omega)}$. Furthermore, if there exists a point $x^* \in \Omega$ such that $f(x^*) \neq 0$ and $x^* \in X_N$ for all sufficiently large N, then $\sigma_{\mathrm{ML}}(f, X_N) \asymp N^{-1/2}$.*

Proof. If $f \in \mathcal{H}_K(\Omega)$, it follows from (3.4) and the minimum-norm characterisation (3.1) of the conditional mean that

$$\sigma_{\mathrm{ML}}(f, X_N) = \frac{\|s_{f,X_N}\|_{\mathcal{H}_K(\Omega)}}{N^{1/2}} \leq \frac{\|f\|_{\mathcal{H}_K(\Omega)}}{N^{1/2}}.$$

The minimum-norm characterisation also implies that $0 < \|s_{f,\{x^*\}}\|_{\mathcal{H}_K(\Omega)} \leq \|s_{f,X_N}\|_{\mathcal{H}_K(\Omega)}$ if $x^* \in X_N$ and $f(x^*) \neq 0$, which proves the second claim and completes the proof. ∎

The reasonableness or otherwise of this behaviour for the maximum likelihood estimate is best assessed in the context of its implied conditional GP, and in particular the behaviour of its credible sets.

---

[4]Note that, as discussed in detail in Section 4.2, samples from the GP do not lie in this RKHS with probability 1 if the RKHS is infinite-dimensional.

   Theorem 3.2 (Slow overconfidence at worst in the RKHS). *If $f \in \mathcal{H}_K(\Omega)$ and there is $x^* \in \Omega$ such that $f(x^*) \neq 0$ and $x^* \in X_N$ for all sufficiently large N, then*

$$(3.5) \qquad \sup_{x \in \Omega} \frac{|f(x) - s_{f,X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} \lesssim N^{1/2} \quad and \quad \frac{|I(f) - Q_{X_N}(f)|}{R_{\mathrm{BC}}(f, X_N)} \lesssim N^{1/2}.$$

   Proof. From (3.3) we know that $|f(x) - s_{f,X_N}(x)| \leq \|f\|_{\mathcal{H}_K(\Omega)} P_{X_N}(x,x)^{1/2}$ for every $x \in \Omega$ if $f \in \mathcal{H}_K(\Omega)$. By this estimate and Proposition 3.1,

$$\frac{|f(x) - s_{f,X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} = \frac{|f(x) - s_{f,X_N}(x)|}{\sigma_{\mathrm{ML}}(f, X_N) P_{X_N}(x,x)^{1/2}} \leq \frac{\|f\|_{\mathcal{H}_K(\Omega)}}{\sigma_{\mathrm{ML}}(f, X_N)} \asymp N^{1/2}.$$

The argument for integration is identical.   ■

   The interpretation of Theorem 3.2 is that a GP model can become at worst slowly overconfident, in the sense that the credible sets are asymptotically $O(N^{1/2})$ times narrower than they would be if the model was asymptotically honest. After the present work was completed, a closely related result appeared as Proposition 3.1 in Wang (2020).

   Remark 3.3. Szabó et al. (2015) included a blow-up factor $L_N > 0$ in the studied credible sets, which in our setting is equivalent to using the scale parameter $\sigma = L_N \sigma_{\mathrm{ML}}(f, X_N)$. If L_N is set to grow sufficiently fast, our results guarantee that the model is not asymptotically overconfident. For example, if $L_N \gtrsim N^{1/2}$ a modification of Theorem 3.2 would state that the standard scores are $O(1)$. It is not clear to us that such artificial inflation of σ can be statistically justified.

   Theorem 3.2 establishes only upper bounds on standard scores and it does not follow that there is a function for which the model is asymptotically overconfident—let alone that this is the case for all functions in the RKHS. In fact, the upper bounds (3.5) can be improved to $\|f - s_{f,X_N}\|_{\mathcal{H}_K(\Omega)} N^{1/2}$ by the use of improved versions (e.g., Wendland, 2005, p. 192) of the generic error estimates (3.3):

$$(3.6a) \qquad\qquad |f(x) - s_{f,X}(x)| \leq \|f - s_{f,X}\|_{\mathcal{H}_K(\Omega)} P_X(x,x)^{1/2},$$

$$(3.6b) \qquad\qquad |I(f) - Q_X(f)| \leq \|f - s_{f,X}\|_{\mathcal{H}_K(\Omega)} V_X^{1/2}.$$

If the RKHS error $\|f - s_{f,X_N}\|_{\mathcal{H}_K(\Omega)}$ decays sufficiently fast it can be established that the model is not asymptotically overconfident. Although it is known that $\|f - s_{f,X_N}\|_{\mathcal{H}_K(\Omega)} \to 0$ as $N \to \infty$ if the kernel is continuous and the point-set sequence $(X_N)_{N=1}^{\infty}$ is space-filling (in the sense that the fill-distance, to defined in Section 4.3, decays to zero), this convergence in the RKHS norm can be arbitrarily slow (Iske, 2018, Theorem 8.37 and Exercise 8.64). It is therefore interesting to ask whether there is a well-characterised subset of the RKHS for which the GP model is not asymptotically overconfident. Such a subset is identified next.

   3.4. Asymptotic Underconfidence for a Subset of the RKHS. In this section we characterise a subset of the RKHS, related to an $L^2(\Omega)$ integral operator, where the true approximation error can be shown to decay faster than the width of the credible set. If $\Omega \subset \mathbb{R}^d$ is compact and

the kernel K continuous, it follows that the integral operator $T \colon L^2(\Omega) \to L^2(\Omega)$ defined as

$$(3.7) \qquad Tg(x) \coloneqq \int_\Omega g(y) K(x,y) \, \mathrm{d}y \quad \text{for} \quad g \in L^2(\Omega)$$

is self-adjoint and compact. By the spectral theorem there exists a sequence of positive and decreasing eigenvalues $(\lambda_n)_{n=1}^\infty$ and corresponding eigenfunctions $(\varphi_n)_{n=1}^\infty \subset L^2(\Omega)$ such that $T\varphi_n = \lambda_n \varphi_n$. Since K is assumed continuous, Mercer's theorem implies that $(\varphi_n)_{n=1}^\infty$ form an orthonormal basis of $L^2(\Omega)$ and $(\lambda_n^{1/2} \varphi_n)_{n=1}^\infty$ form an orthonormal basis of $\mathcal{H}_K(\Omega)$ when each φ_n is uniquely identified with a continuous element of the RKHS. Therefore the kernel has the uniformly convergent expansion $K(x,y) = \sum_{n=1}^\infty \lambda_n \varphi_n(x) \varphi_n(y)$ on $\Omega \times \Omega$ and the RKHS is

$$\mathcal{H}_K(\Omega) = \left\{ g \in L^2(\Omega) \, : \, \sum_{n=1}^\infty \frac{\langle g, \varphi_n \rangle_{L^2(\Omega)}^2}{\lambda_n} < \infty \right\}.$$

It can be then shown that the range of T is

$$(3.8) \qquad T(L^2(\Omega)) = \left\{ g \in L^2(\Omega) \, : \, \sum_{n=1}^\infty \frac{\langle g, \varphi_n \rangle_{L^2(\Omega)}^2}{\lambda_n^2} < \infty \right\} \subset \mathcal{H}_K(\Omega).$$

It is easy to prove that the error estimates (3.3) can be improved if f is in $T(L^2(\Omega))$ (Wendland, 2005, Section 11.5). Namely, if there is $v \in L^2(\Omega)$ such that $f = Tv$, then

$$\|f - s_{f,X}\|_{\mathcal{H}_K(\Omega)} \le \left(\int_\Omega P_X(x,x) \, \mathrm{d}x \right)^{1/2} \|v\|_{L^2(\Omega)} =: \|P_X^{1/2}\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}$$

and therefore by (3.6) the error estimates become

$$(3.9\mathrm{a}) \qquad |f(x) - s_{f,X}(x)| \le P_X(x,x)^{1/2} \|P_X^{1/2}\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \, ,$$

$$(3.9\mathrm{b}) \qquad |I(f) - Q_X(f)| \le V_X^{1/2} \|P_X^{1/2}\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \, .$$

The standard convergence rates are thus effectively squared, this being occasionally referred to as *superconvergence* (Schaback, 2018). See Schaback (1999, 2000); Fasshauer and McCourt (2015, Section 9.4.3); and Bach (2017, Section 5) for additional results and discussion and Kanagawa et al. (2020, Section 6.2) for numerical examples. Also note the connection of the space (3.8) to powers of RKHSs (Steinwart and Scovel, 2012) and Hilbert scales (Dashti and Stuart, 2017, Appendix A.1.3). Unfortunately, the argument that yields the improved rates (3.9) does not appear amenable to handling more general subspaces of $\mathcal{H}_K(\Omega)$.

By replacing (3.3) with (3.9) in the proof of Theorem 3.2 we establish that the GP model is asymptotically underconfident for $f \in T(L^2(\Omega))$.

Theorem 3.4 (Asymptotic underconfidence for sufficiently regular functions). *Suppose that $\Omega \subset \mathbb{R}^d$ is compact, K is continuous, $f \in T(L^2(\Omega)) \subset \mathcal{H}_K(\Omega)$, and there is $x^* \in \Omega$ such that $f(x^*) \neq 0$ and $x^* \in X_N$ for all sufficiently large N. Then*

$$\sup_{x \in \Omega \setminus X_N} \frac{|f(x) - s_{f,X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} \lesssim N^{1/2} \|P_{X_N}^{1/2}\|_{L^2(\Omega)} \quad \text{and} \quad \frac{|I(f) - Q_{X_N}(f)|}{R_{\mathrm{BC}}(f, X_N)} \lesssim N^{1/2} \|P_{X_N}^{1/2}\|_{L^2(\Omega)} \, .$$

That is, the model is asympotically underconfident if the sequence $(X_N)_{N=1}^\infty \subset \Omega$ is such that $N^{1/2} \| P_{X_N}^{1/2} \|_{L^2(\Omega)} \to 0$ as $N \to \infty$.

Proof. Let $f = Tv$ for $v \in L^2(\Omega)$. By using (3.9) instead of (3.3) in the proof of Theorem 3.2 we get

$$\sup_{x \in \Omega \setminus X_N} \frac{|f(x) - s_{f,X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} \leq \frac{\| P_{X_N}^{1/2} \|_{L^2(\Omega)} \, \| v \|_{L^2(\Omega)}}{\sigma_{\mathrm{ML}}(f, X_N)} \lesssim N^{1/2} \| P_{X_N}^{1/2} \|_{L^2(\Omega)}.$$

The supremum is over $x \notin X_N$ because for $x \in X_N$ we have defined the standard score to be one. The argument for integration is identical. ∎

If $(X_N)_{N=1}^\infty$ are quasi-uniform (see Section 4.3 for details), then $N^{1/2} \| P_{X_N}^{1/2} \|_{L^2(\Omega)} \to 0$ is true, for example, when K is one of the popular infinitely smooth kernels associated with super-algebraic rates of convergence such as a Gaussian or an inverse multiquadric (Rieger and Zwicknagl, 2010). A specialisation to Sobolev kernels will be given in Section 4.5.

4. Sobolev Kernels and Functions Outside the RKHS.
This section extends the results of Section 3 for functions outside the RKHS when the kernel K is a Sobolev kernel.

4.1. Sobolev Spaces and Kernels.
Let $\widehat{g}(\xi) := \int_{\mathbb{R}^d} g(x) e^{-i\xi^\mathsf{T} x} \, \mathrm{d}x$ denote the Fourier transform of $g \in L^1(\mathbb{R}^d)$. The Sobolev space $W_2^\alpha(\mathbb{R}^d)$ of order $\alpha \geq 0$ is the Hilbert space

$$W_2^\alpha(\mathbb{R}^d) := \left\{ g \in L^2(\mathbb{R}^d) \, : \, \int_{\mathbb{R}^d} \left(1 + \|\xi\|^2\right)^\alpha |\widehat{g}(\xi)|^2 \, \mathrm{d}\xi < \infty \right\}$$

equipped with the inner product

$$\langle h, g \rangle_{W_2^\alpha(\mathbb{R}^d)} := \int_{\mathbb{R}^d} \left(1 + \|\xi\|^2\right)^\alpha \widehat{h}(\xi) \overline{\widehat{g}(\xi)} \, \mathrm{d}\xi,$$

where $\bar{z}$ is the complex conjugate of $z \in \mathbb{C}$. When $\alpha \in \mathbb{N}$, the space $W_2^\alpha(\mathbb{R}^d)$ can be equivalently defined as consisting of those functions whose weak derivatives up to order α exist and are in $L^2(\mathbb{R}^d)$. For $\alpha \notin \mathbb{N}$, $W_2^\alpha(\mathbb{R}^d)$ can also be defined as an interpolation or Besov space, up to equivalent norms (e.g., Triebel, 2006). If $\alpha > d/2$, then every element of $W_2^\alpha(\mathbb{R}^d)$ can be uniquely identified with a continuous function from its $L^2(\mathbb{R}^d)$ equivalence class and $W_2^\alpha(\mathbb{R}^d)$ can be viewed as an RKHS of continuous functions on $\mathbb{R}^d$. This identification will be implicitly assumed throughout the article.

Let $\Omega \subset \mathbb{R}^d$ be Lebesgue measurable and let $W_2^\alpha(\Omega)$ be the restriction of $W_2^\alpha(\mathbb{R}^d)$ to Ω, as defined in Section 3.1. We say that a kernel $K \colon \Omega \times \Omega \to \mathbb{R}$ is a *Sobolev kernel of order* $\alpha > d/2$ (on Ω) if its RKHS $\mathcal{H}_K(\Omega)$ is *norm-equivalent* to $W_2^\alpha(\Omega)$. That is, $\mathcal{H}_K(\Omega)$ equals $W_2^\alpha(\Omega)$ as a set of functions and there exist positive constants C_K and C_K' such that

(4.1) $$C_K \| g \|_{W_2^\alpha(\Omega)} \leq \| g \|_{\mathcal{H}_K(\Omega)} \leq C_K' \| g \|_{W_2^\alpha(\Omega)}$$

for all $g \in \mathcal{H}_K(\Omega)$. Stationary kernels with prescribed Fourier decay form an important subclass of Sobolev kernels: if there is $\Phi \colon \mathbb{R}^d \to \mathbb{R}$ such that $K(x, y) = \Phi(x - y)$ and

$$C_1 \left(1 + \|\xi\|^2\right)^{-\alpha} \leq \widehat{\Phi}(\xi) \leq C_2 \left(1 + \|\xi\|^2\right)^{-\alpha} \quad \text{for some } C_1, C_2 > 0 \text{ and all } \xi \in \mathbb{R}^d,$$

then K is a Sobolev kernel of order α and $\mathcal{H}_K(\mathbb{R}^d)$ is norm-equivalent to $W_2^\alpha(\mathbb{R}^d)$. Perhaps the most ubiquitous Sobolev kernels are the Matérn kernels

$$(4.2) \qquad K_{\nu,\ell}(x,y) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\,\|x-y\|}{\ell} \right)^\nu \mathrm{K}_\nu \left(\frac{\sqrt{2\nu}\,\|x-y\|}{\ell} \right),$$

where $\nu > 0$ is a smoothness parameter, $\ell > 0$ a length-scale parameter, Γ the Gamma function, and K_ν the modified Bessel function of the second kind of order ν. The Fourier transform of a Matérn kernel is (Stein, 1999, p. 49)

$$(4.3) \qquad K_{\nu,\ell}(x,y) = \Phi_{\nu,\ell}(x-y), \qquad \widehat{\Phi}_{\nu,\ell}(\xi) = \frac{\Gamma(\nu+d/2)}{\pi^{d/2}\Gamma(\nu)} \left(\frac{2\nu}{\ell^2} \right)^\nu \left(\frac{2\nu}{\ell^2} + \|\xi\|^2 \right)^{-(\nu+d/2)},$$

and its RKHS is thus norm-equivalent to the Sobolev space $W_2^\alpha(\mathbb{R}^d)$ with $\alpha = \nu + d/2$. See Wendland (2005, Chapter 10) for proofs and further detail.

Functions that lie on the "boundary" of a Sobolev space play an important role in our analysis. For this purpose, define the sets

$$S_-^\alpha(\mathbb{R}^d) := \left\{ g \in L^2(\mathbb{R}^d) \,:\, |\widehat{g}(\xi)|^2 \lesssim \left(1 + \|\xi\|^2 \right)^{-(\alpha+d/2)} \right\},$$
$$S_+^\alpha(\mathbb{R}^d) := \left\{ g \in L^2(\mathbb{R}^d) \,:\, |\widehat{g}(\xi)|^2 \gtrsim \left(1 + \|\xi\|^2 \right)^{-(\alpha+d/2)} \right\},$$

and

$$(4.4) \qquad S^\alpha(\mathbb{R}^d) := S_-^\alpha(\mathbb{R}^d) \cap S_+^\alpha(\mathbb{R}^d).$$

From the fact that $\int_{\mathbb{R}^d} (1+\|\xi\|^2)^\alpha (1+\|\xi\|^2)^{-(\beta+d/2)} \, \mathrm{d}\xi$ is finite if and only if $\beta > \alpha$ it follows that $S^\beta(\mathbb{R}^d)$ and $S_-^\beta(\mathbb{R}^d)$ are subsets of $W_2^\alpha(\mathbb{R}^d)$ if and only if $\beta > \alpha$. Similarly, $S_+^\beta(\mathbb{R}^d) \cap W_2^\alpha(\mathbb{R}^d)$ is non-empty if and only if $\beta > \alpha$, and it may therefore be helpful to think of $S_+^\alpha(\mathbb{R}^d)$ as approximately the collection of square-integrable functions that are not in $W_2^\alpha(\mathbb{R}^d)$. A function $g \colon \Omega \to \mathbb{R}$ is said to be in $S_-^\alpha(\Omega)$ ($S_+^\alpha(\Omega)$) if it has an extension $g_0 \in S_-^\alpha(\mathbb{R}^d)$ ($g_0 \in S_+^\alpha(\mathbb{R}^d)$). As an aside, we note the similarity of these sets to the sequence hyperrectangles analysed in Szabó et al. (2013, 2015) and Hadji and Szabó (2019).

4.2. Motivation: Sample Path Properties of Gaussian Processes. In this article the function f is fixed, but nevertheless it seems reasonable that a statistical estimation method based on a GP model ought to perform well when the assumptions of the GP model are satisfied. This motivates us to consider the regularity of samples from the GP model, which will later form the basis of regularity assumptions on f. The most important results relating the samples and the RKHS are the following (for a recent review, see Kanagawa et al., 2018):

- If $\mathcal{H}_K(\Omega)$ is infinite-dimensional, then the sample paths of the GP belong to $\mathcal{H}_K(\Omega)$ with probability 0. In general, the samples being contained in the RKHS of a different kernel R with probability 0 or 1 depends on whether or not a certain nuclear dominance condition between the kernels K and R holds (Driscoll, 1973; Lukić and Beder, 2001).
- If K is a Sobolev kernel of order $\alpha > d/2$, then the GP sample paths are in $W_2^\beta(\Omega)$ with probability 1 if $\beta < \alpha - d/2$ and with probability 0 if $\beta \geq \alpha - d/2$ (Scheuerer, 2010; Steinwart, 2019).

The latter result essentially says that for Sobolev kernels the sample paths are rougher than elements in $\mathcal{H}_K(\Omega)$ by order $d/2$. Furthermore, it follows that sample paths are in the set

$$(4.5) \qquad\qquad\qquad W_2^{\alpha - d/2 - \varepsilon}(\Omega) \setminus W_2^{\alpha - d/2}(\Omega)$$

with probability 1 for any $\varepsilon > 0$. We are not aware of more advanced developments than this but, encouraged by (4.9), conjecture that the set $S^{\alpha - d/2}(\Omega)$, which is a subset of (4.5) for any $\varepsilon > 0$, is in some sense the smallest set (or closely related to such a set) that contains almost all sample paths of a GP with a Sobolev covariance kernel of order α.

4.3. Error Estimates for Sobolev Kernels. In this section we present bounds on the GP approximation and integration errors and sharp rates (i.e., the upper and lower bounds are of matching order) of decay of $\sup_{x \in \Omega} P_X(x, x)^{1/2}$ and $V_X^{1/2}$ when K is a Sobolev kernel; these will be used to study the maximum likelihood estimator in Section 4.4. Define the *fill-distance* h_X and the *separation radius* q_X of a set of distinct points $X = \{x_1, \ldots, x_N\} \subset \Omega$ as

$$h_X := \sup_{x \in \Omega} \min_{i = 1, \ldots, N} \|x - x_i\| \quad \text{and} \quad q_X := \frac{1}{2} \min_{i \neq j} \|x_i - x_j\|.$$

Also define the *mesh ratio* $\rho_X := h_X / q_X \geq 1$. A sequence $(X_N)_{N=1}^{\infty} \subset \Omega$ is *quasi-uniform* if $\rho_{X_N} \lesssim 1$, which implies that $q_{X_N} \asymp h_{X_N} \asymp N^{-1/d}$ (Wendland, 2005, Proposition 14.1).

The domain $\Omega \subset \mathbb{R}^d$ will often be assumed to satisfy the following requirement, which will be made explicit when required.

Assumption 4.1. *The set $\Omega \subset \mathbb{R}^d$ is bounded and connected, has a non-empty interior and a Lipschitz boundary, and satisfies an interior cone condition.*

The Lipschitz boundary condition says that the boundary is sufficiently regular in that it is locally the graph of a Lipschitz function, while the interior cone condition prohibits the existence of pinch points; for technical definitions see for example Kanagawa et al. (2020, Section 3). These conditions are standard in the theory of Sobolev spaces and error analysis of kernel-based approximation methods.[5] In particular, they guarantee that various different notions of integer and fractional order Sobolev spaces defined on Ω result in identical function spaces up to equivalent norms. Assumption 4.1 is satisfied by all typical domains and in particular by $\Omega = [0,1]^d$, which is used in the numerical examples in Section 5.

The following theorem provides bounds on the approximation and integration error by a GP conditional mean when the kernel is Sobolev and f does not necessarily lie in the RKHS. The theorem as we state it is a consequence of results in the scattered data approximation literature (Wendland and Rieger, 2005; Narcowich et al., 2006). For completeness and to simplify later developments the proof is provided in Appendix A.

Theorem 4.2. *Let $\alpha \geq \beta$ and $\lfloor \beta \rfloor > d/2$. Suppose that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 4.1 and K is a Sobolev kernel of order α. If $f \in W_2^{\beta}(\Omega)$, then there are $C_1, C_2, h_0 > 0$, which do not*

---

[5] In the results we cite it is often assumed that Ω is open. Because these results provide bounds on $L^p(\Omega)$ norms and a Lipschitz boundary is of measure zero, they remain valid whenever Ω has a non-empty interior.

depend on f or X, such that

$$\sup_{x \in \Omega} |f(x) - s_{f,X}(x)| \leq C_1 h_X^{\beta-d/2} \rho_X^{\alpha-\beta} \|f\|_{W_2^\beta(\Omega)} \quad and \quad |I(f) - Q_X(f)| \leq C_2 h_X^\beta \rho_X^{\alpha-\beta} \|f\|_{W_2^\beta(\Omega)}$$

whenever $h_X \leq h_0$. For a quasi-uniform sequence $(X_N)_{N=1}^\infty \subset \Omega$ these bounds become

$$\sup_{x \in \Omega} |f(x) - s_{f,X_N}(x)| \lesssim N^{-\beta/d+1/2} \|f\|_{W_2^\beta(\Omega)} \quad and \quad |I(f) - Q_{X_N}(f)| \lesssim N^{-\beta/d} \|f\|_{W_2^\beta(\Omega)}.$$

See Arcangéli et al. (2007, 2012) and Wynne et al. (2020) for a collection of marginally more general versions of Theorem 4.2. These generalisations are not used here because proofs of some of the results in Section 4.4 require understanding of the dependency, which is much less transparent in the generalisations, on the Sobolev smoothness parameters of the constants C_1 and C_2. The following extension for $f \in S_-^\beta(\Omega)$, that we have not found in the literature, will be useful. Its proof is given in Appendix A.

Theorem 4.3. *Suppose that the other assumptions of Theorem 4.2 are satisfied but $f \in S_-^\beta(\Omega)$. Then for a quasi-uniform sequence $(X_N)_{N=1}^\infty \subset \Omega$,*

$$\sup_{x \in \Omega} |f(x) - s_{f,X_N}(x)| \lesssim N^{-\beta/d+1/2}(\log N)^{1/2} \quad and \quad |I(f) - Q_{X_N}(f)| \lesssim N^{-\beta/d}(\log N)^{1/2}.$$

The implicit constants in the estimates of Theorem 4.3 depend on the smallest $C > 0$ such that $|\widehat{f_0}(\xi)|^2 \leq C(1 + \|\xi\|^2)^{-(\beta+d/2)}$ for a Sobolev exntesion f_0 of f and all sufficiently large $\xi \in \mathbb{R}^d$. Similar dependencies are implicit in the bounds of Propositions 4.6 and 4.8 and Theorems 4.9 and 4.10.

Due to (3.2) the error estimates of Theorem 4.2 for $\beta = \alpha$ are also upper bounds on the conditional standard deviations. It is possible to establish matching lower bounds, which leads to the following standard result, the proof of which is given in Appendix A.

Theorem 4.4. *Suppose that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 4.1. If K is a Sobolev kernel of order $\alpha > \lfloor d/2 \rfloor$ and the sequence $(X_N)_{N=1}^\infty \subset \Omega$ is quasi-uniform, then*

$$\sup_{x \in \Omega} P_{X_N}(x,x)^{1/2} \asymp N^{-\alpha/d+1/2} \quad and \quad V_{X_N}^{1/2} \asymp N^{-\alpha/d}.$$

Furthermore, $P_{X_N}(x,x)^{1/2} \asymp N^{-\alpha/d+1/2}$ for any $x \notin \bigcup_{N=1}^\infty X_N$.

4.4. Maximum Likelihood Estimation. This section contains upper and lower bounds on $\sigma_{\mathrm{ML}}(f, X_N)$ when K is a Sobolev kernel of order α and f is not necessarily in $W_2^\alpha(\Omega)$. If $f \in W_2^\alpha(\Omega)$, which below corresponds to either $\beta \geq \alpha$ or $\beta < \alpha$, then the results in Section 3.3 can be used instead. The main result on maximum likelihood estimation is Theorem 4.9 which provides sharp (up to logarithmic factors) asymptotics for the maximum likelihood estimate under certain conditions on f. Propositions 4.5 to 4.8 contain individual upper and lower bounds. The bounds are used to discuss credible sets and asymptotic overconfidence and underconfidence in Section 4.5. The proofs of this section are provided in Appendix A.

Proposition 4.5. *Let $\alpha \geq \beta$ and $\lfloor \beta \rfloor > d/2$. Suppose that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 4.1 and K is a Sobolev kernel of order α. If $f \in W_2^\beta(\Omega)$, then there are $C, h_0 > 0$, which do not depend on f or X_N, such that*

$$(4.6) \qquad \sigma_{\mathrm{ML}}(f, X_N) \leq C N^{-1/2} q_{X_N}^{\beta - \alpha} \|f\|_{W_2^\beta(\Omega)}$$

whenever $h_{X_N} \leq h_0$. For a quasi-uniform sequence $(X_N)_{N=1}^\infty \subset \Omega$ this bound becomes

$$\sigma_{\mathrm{ML}}(f, X_N) \lesssim N^{(\alpha - \beta)/d - 1/2} \|f\|_{W_2^\beta(\Omega)}.$$

Proposition 4.5 holds in a slightly modified form if $f \in S_-^\beta(\Omega)$ (recall $S_-^\beta(\Omega) \cap W_2^\beta(\Omega) = \emptyset$).

Proposition 4.6. *Suppose that the other assumptions of Theorem 4.5 are satisfied but $f \in S_-^\beta(\mathbb{R}^d)$. Then for a quasi-uniform sequence $(X_N)_{N=1}^\infty \subset \Omega$,*

$$\sigma_{\mathrm{ML}}(f, X_N) \lesssim N^{(\alpha - \beta)/d - 1/2} (\log N)^{1/2}.$$

Lower bounds require some additional assumptions and take a more cumbersome form. Recall that the support of a function is the closed set $\mathrm{supp}(f) := \overline{\{x \in \Omega : f(x) \neq 0\}}$ and the interior $\mathrm{int}(\Omega)$ of $\Omega \subset \mathbb{R}^d$ is the largest open set contained in Ω.

Proposition 4.7. *Let $\alpha \geq \beta > \gamma$ and $\lfloor \gamma \rfloor > d/2$. Suppose that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 4.1 and K is a Sobolev kernel of order α. If $\mathrm{supp}(f) \subset \mathrm{int}(\Omega)$ and f has an extension $f_0 \in W_2^\gamma(\mathbb{R}^d) \cap S_+^\beta(\mathbb{R}^d)$ such that $\mathrm{supp}(f_0) \subset \mathrm{int}(\Omega)$, then there are $C, h_0 > 0$, which do not depend on X_N, such that*

$$(4.7) \qquad \sigma_{\mathrm{ML}}(f, X_N) \geq C N^{-1/2} h_{X_N}^{\gamma(1 - \alpha/\beta)} \rho_{X_N}^{-(\alpha - \gamma)(\alpha - \beta)/\beta} \|f\|_{W_2^\gamma(\Omega)}^{1 - \alpha/\beta}$$

whenever $h_{X_N} \leq h_0$. For a quasi-uniform sequence $(X_N)_{N=1}^\infty \subset \Omega$ this bound becomes

$$(4.8) \qquad \sigma_{\mathrm{ML}}(f, X_N) \gtrsim N^{\gamma(\alpha/\beta - 1)/d - 1/2} \|f\|_{W_2^\gamma(\Omega)}^{1 - \alpha/\beta}.$$

Proposition 4.8. *Suppose that the other assumptions of Theorem 4.7 are satisfied but $f_0 \in S^\beta(\mathbb{R}^d)$. Then for a quasi-uniform sequence $(X_N)_{N=1}^\infty \subset \Omega$,*

$$\sigma_{\mathrm{ML}}(f, X_N) \gtrsim N^{(\alpha - \beta)/d - 1/2} (\log N)^{(1 - \alpha/\beta)/2}.$$

By combining Propositions 4.6 and 4.8 we obtain a rate for $f \in S^\beta(\Omega)$ that is sharp up to logarithmic factors. The empirical results in Section 5.1 suggest that elimination of the logarithmic factors and the support conditions may be possible with more careful analysis.

Theorem 4.9 (Asymptotics of the MLE). *Let $\alpha \geq \beta$ and $\lfloor \beta \rfloor > d/2$. Suppose that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 4.1 and K is a Sobolev kernel of order α. If $\mathrm{supp}(f) \subset \mathrm{int}(\Omega)$ and f has an extension $f_0 \in S^\beta(\mathbb{R}^d)$ such that $\mathrm{supp}(f_0) \subset \mathrm{int}(\Omega)$, then for a quasi-uniform sequence $(X_N)_{N=1}^\infty \subset \Omega$,*

$$N^{(\alpha - \beta)/d - 1/2} (\log N)^{(1 - \alpha/\beta)/2} \lesssim \sigma_{\mathrm{ML}}(f, X_N) \lesssim N^{(\alpha - \beta)/d - 1/2} (\log N)^{1/2}.$$

In particular, for $\beta = \alpha - d/2$ we have $(\alpha - \beta)/d - 1/2 = 0$ so that the maximum likelihood estimates are asymptotically constant, up to logarithmic factors:

$$(4.9) \qquad (\log N)^{-d/(4\alpha - 2d)} \lesssim \sigma_{\mathrm{ML}}(f, X_N) \lesssim (\log N)^{1/2}$$

if $f \in S^{\alpha - d/2}(\Omega)$. As discussed in Section 4.2, this corresponds to the case where f has essentially the same regularity as samples from a GP whose covariance kernel is a Sobolev kernel of order α.

4.5. Credible Sets. We now use the bounds on the maximum likelihood estimates to prove an overconfidence result similar to Theorem 3.2, but this time for functions outside the RKHS. First, it is instructive to study what can happen if the scale parameter is held fixed. If K is a Sobolev kernel of order α, $f \in W_2^\beta(\Omega)$ for $\beta \leq \alpha$, and $(X_N)_{N=1}^\infty$ are quasi-uniform, then Theorems 4.2 and 4.4 yield

$$(4.10) \quad \frac{|I(f) - Q_{X_N}(f)|}{R_{\mathrm{BC}}(f, X_N)} = \frac{|I(f) - Q_{X_N}(f)|}{\sigma V_{X_N}^{1/2}} \lesssim \frac{N^{-\beta/d} \|f\|_{W_2^\beta(\Omega)}}{\sigma N^{-\alpha/d}} = \sigma^{-1} N^{(\alpha - \beta)/d} \|f\|_{W_2^\beta(\Omega)}.$$

That is, there is potential for significant overconfidence if K is smoother than f. The following theorem shows that maximum likelihood estimation provides protection against such model misspecification.

Theorem 4.10 (Slow overconfidence at worst outside the RKHS). *Let $\alpha \geq \beta$ and $\lfloor \beta \rfloor > d/2$. Suppose that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 4.1 and K is a Sobolev kernel of order α. If $\mathrm{supp}(f) \subset \mathrm{int}(\Omega)$ and f has an extension $f_0 \in S^\beta(\mathbb{R}^d)$ such that $\mathrm{supp}(f_0) \subset \mathrm{int}(\Omega)$, then for a quasi-uniform sequence $(X_N)_{N=1}^\infty \subset \Omega$,*

$$\frac{|f(x) - s_{f, X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} \lesssim N^{1/2} (\log N)^{\alpha/(2\beta)} \quad \textit{for any} \quad x \in \Omega$$

and

$$\frac{|I(f) - Q_{X_N}(f)|}{R_{\mathrm{BC}}(f, X_N)} \lesssim N^{1/2} (\log N)^{\alpha/(2\beta)}.$$

Proof. Consider first approximation with GPs. For N such that $x \in X_N$ the standard score in (2.9) and (2.10) is by definition equal to one. We can thus assume that $x \notin \bigcup_{N=1}^\infty X_N$. Then the estimates in Theorems 4.3 and 4.4 and Proposition 4.8 yield

$$\frac{|f(x) - s_{f, X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} = \frac{|f(x) - s_{f, X_N}(x)|}{\sigma_{\mathrm{ML}}(f, X_N) P_{X_N}(x, x)^{1/2}} \lesssim \frac{N^{-\beta/d + 1/2} (\log N)^{1/2}}{N^{(\alpha - \beta)/d - 1/2} (\log N)^{(1 - \alpha/\beta)/2} N^{-\alpha/d + 1/2}}$$

$$= N^{1/2} (\log N)^{\alpha/(2\beta)}.$$

The proof for integration is essentially identical. ∎

Interestingly, the case $\beta = \alpha - d/2$, which essentially corresponds to f having the same regularity as samples from the GP, plays no special role in Theorem 4.10. We are uncertain

if this is due to an inadequacy in the analysis or if there in fact exist GP samples for which the model is overconfident. In practice one rarely knows the exact smoothness of f (or the function is not an element of $S^\beta(\Omega)$ for any β) and can only guess, for example, that f has weak derivatives at least up to some order β. If $\beta < \alpha$, then nothing can be inferred about the credible sets based on our results; if $\beta \geq \alpha$, then Theorem 3.2 can be used.

As our final result we present a specialisation of Theorem 3.4 to Sobolev kernels. The proof is a straightforward application of the estimates in (3.9), Proposition 3.1, and Theorem 4.4:

$$\sup_{x \in \Omega \setminus X_N} \frac{|f(x) - s_{f,X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} \lesssim N^{1/2} \left(\int_\Omega P_{X_N}(x, x) \, \mathrm{d}x \right)^{1/2} \leq N^{1/2} \sup_{x \in \Omega} P_{X_N}(x, x)^{1/2} \asymp N^{-\alpha/d+1}$$

if the point sequence is quasi-uniform. Recall that $T(L^2(\Omega))$ is the range of the integral operator in (3.7).

Theorem 4.11 (Asymptotic underconfidence for sufficiently regular functions). *Suppose that $\Omega \subset \mathbb{R}^d$ is compact, K is a Sobolev kernel of order $\lfloor \alpha \rfloor > d/2$, and $f \in T(L^2(\Omega))$. Then for a quasi-uniform sequence $(X_N)_{N=1}^\infty \subset \Omega$ such that there is $x^* \in X_N$ for which $f(x^*) \neq 0$ for all sufficiently large N,*

$$\sup_{x \in \Omega \setminus X_N} \frac{|f(x) - s_{f,X_N}(x)|}{R_{\mathrm{GP}}(x, f, X_N)} \lesssim N^{-\alpha/d+1} \quad \text{and} \quad \frac{|I(f) - Q_{X_N}(f)|}{R_{\mathrm{BC}}(f, X_N)} \lesssim N^{-\alpha/d+1}.$$

We thus have asymptotic underconfidence for approximation and integration of $f \in T(L^2(\Omega))$ at least when $\alpha > d$. Note that for Sobolev kernels of order α the range $T(L^2(\Omega))$ is related to the Sobolev space of smoothness 2α (see, e.g., Tuo et al., 2020, Section 2.3).

Theorem 4.11 can be illustrated in detail using the Brownian motion kernel $K(x, y) = \min\{x, y\}$ on $\Omega = [0, 1]$. Its RKHS consists of functions $f \in W_2^1([0, 1])$ such that $f(0) = 0$. It is well-known that the GP conditional mean for this kernel is the piecewise linear spline interpolant. Furthermore, for the weight $w \equiv 1$ in (2.3) the Bayesian quadrature estimator is the *trapezoidal rule* if $x_N = 1$ and $f(0) = 0$ (e.g., Karvonen, 2019, Section 5.5):

$$Q_{X_N}(f) = \sum_{n=1}^N \frac{f(x_{n-1}) + f(x_n)}{2}(x_n - x_{n-1}),$$

where the convention $x_0 = 0$ is used. If the equispaced points $x_n = n/N$ are used, the integral conditional variance (2.4b) has the simple form (Ritter, 2000, p. 26)

$$(4.11) \qquad\qquad V_{X_N} = \frac{1}{12} \sum_{n=1}^N (x_n - x_{n-1})^3 = \frac{1}{12N^2}.$$

If $f : [0, 1] \to \mathbb{R}$ is twice-differentiable with f'' bounded and $f(0) = 0$, which means that $f \in W_2^2([0, 1])$, the standard error formula for the trapezoidal rule with equispaced points is (Atkinson, 1989, Section 5.1)

$$(4.12) \qquad\qquad I(f) - Q_{X_N}(f) = -\frac{1}{12N^2} f''(\xi_N) \quad \text{for some} \quad \xi_N \in [0, 1].$$

Because such a function is in the RKHS, Proposition 3.1 gives $\sigma_{\mathrm{ML}}(f, X_N) \asymp N^{-1/2}$. This and the estimates (4.11) and (4.12) thus yield

$$(4.13) \qquad \frac{|I(f) - Q_{X_N}(f)|}{R_{\mathrm{BC}}(f, X_N)} = \frac{|I(f) - Q_{X_N}(f)|}{\sigma_{\mathrm{ML}}(f, X_N) V_{X_N}^{1/2}} \leq \frac{\sup_{x \in [0,1]} |f''(x)|}{\sqrt{12} N \sigma_{\mathrm{ML}}(f, X_N)} \lesssim N^{-1/2},$$

which is the statement of Theorem 4.11 with $\alpha = 1$ and $d = 1$. If we further assume that there is $C > 0$ such that $f''(x) > C$ for all $x \in [0, 1]$ (i.e., f is strictly convex), then (4.12) implies that $|I(f) - Q_{X_N}(f)| \asymp N^{-2}$, and the standard score (4.13) hence has a lower bound of matching order $N^{-1/2}$.

5. Numerical Illustration. This section numerically investigates the sharpness of the results in Section 4. Examples in Section 5.1 verify that the bounds on $\sigma_{\mathrm{ML}}(f, X_N)$ in Theorem 4.9 are valid. Section 5.2 contains limited evidence that the bounds in Section 4.5 are not tight: the credible sets do *not* appear to contract with a rate $O(N^{-1/2})$ faster than the true error.

5.1. Maximum Likelihood Estimation. In these examples we illustrate the behaviour of the maximum likelihood estimate $\sigma_{\mathrm{ML}}(f, X_N)$ using the Matérn kernel $K_{\nu,\ell}$ in (4.2). Recall that the RKHS of $K_{\nu,\ell}$ is norm-equivalent to the Sobolev space $W_2^\alpha(\mathbb{R}^d)$ with $\alpha = \nu + d/2$. We select $\Omega = [0, 1]^d$ and use test functions constructed out of Matérn kernels of smoothness η:

$$(5.1) \qquad f(x) = \sum_{i=1}^m a_i K_{\eta,\ell}(x, z_i), \quad \text{with some} \quad a_i \in \mathbb{R} \quad \text{and} \quad z_i \in [0, 1]^d.$$

By (4.3), the Fourier transform of such a function satisfies $|\widehat{f}(\xi)|^2 \propto (2\eta/\ell^2 + \|\xi\|^2)^{-(2\eta+d)}$. The function is thus an element $S^{2\eta+d/2}(\mathbb{R}^d)$ and, except for the support condition $\mathrm{supp}(f) \subset (0, 1)^d$, satisfies the assumptions of Theorem 4.9 with $\beta = 2\eta + d/2$. For a quasi-uniform point sequence we therefore expect that (possibly up to logarithmic factors)

$$(5.2) \qquad \sigma_{\mathrm{ML}}(f, X_N) \asymp N^{(\nu-2\eta)_+/d-1/2} \text{ if } \nu \geq 2\eta \quad \text{and} \quad \sigma_{\mathrm{ML}}(f, X_N) \asymp N^{-1/2} \text{ if } \nu < 2\eta.$$

MLE when $d = 1$. In the first example we set $d = 1$, $\ell = 0.2$, $\eta = 0.5$, $m = 3$, $(a_1, a_2, a_3) = (1, 0.5, 0.2)$, and $(z_1, z_2, z_3) = (0.2, 0.55, 0.78)$. Figure 1 displays the behaviour of the maximum likelihood estimates and the predicted theoretical rates (5.2) for different values of the smoothness parameter ν of the Matérn kernel. On the first and second row of Figure 1 the point sets are the non-nested uniform grids

$$(5.3) \qquad X_N = \left\{ 0, \frac{1}{N-1}, \dots, \frac{N-2}{N-1}, 1 \right\}$$

with fill-distances $h_{X_N} = 1/(N-1)$. The maximum likelihood estimates exhibit wild oscillations which seem to be related to placement of the evaluation points in relation to the points z_i defining f. Nevertheless, it is clear that the rates predicted by (5.2) are realised in all cases. On the third and fourth row of Figure 1 the point sets are nested: X_N consists of the first N elements of the low-discrepancy van der Corput sequence

$$(5.4) \qquad 0, \, 0.5, \, 0.75, \, 0.125, \, 0.625, \, 0.375, \, 0.875, \, \dots$$
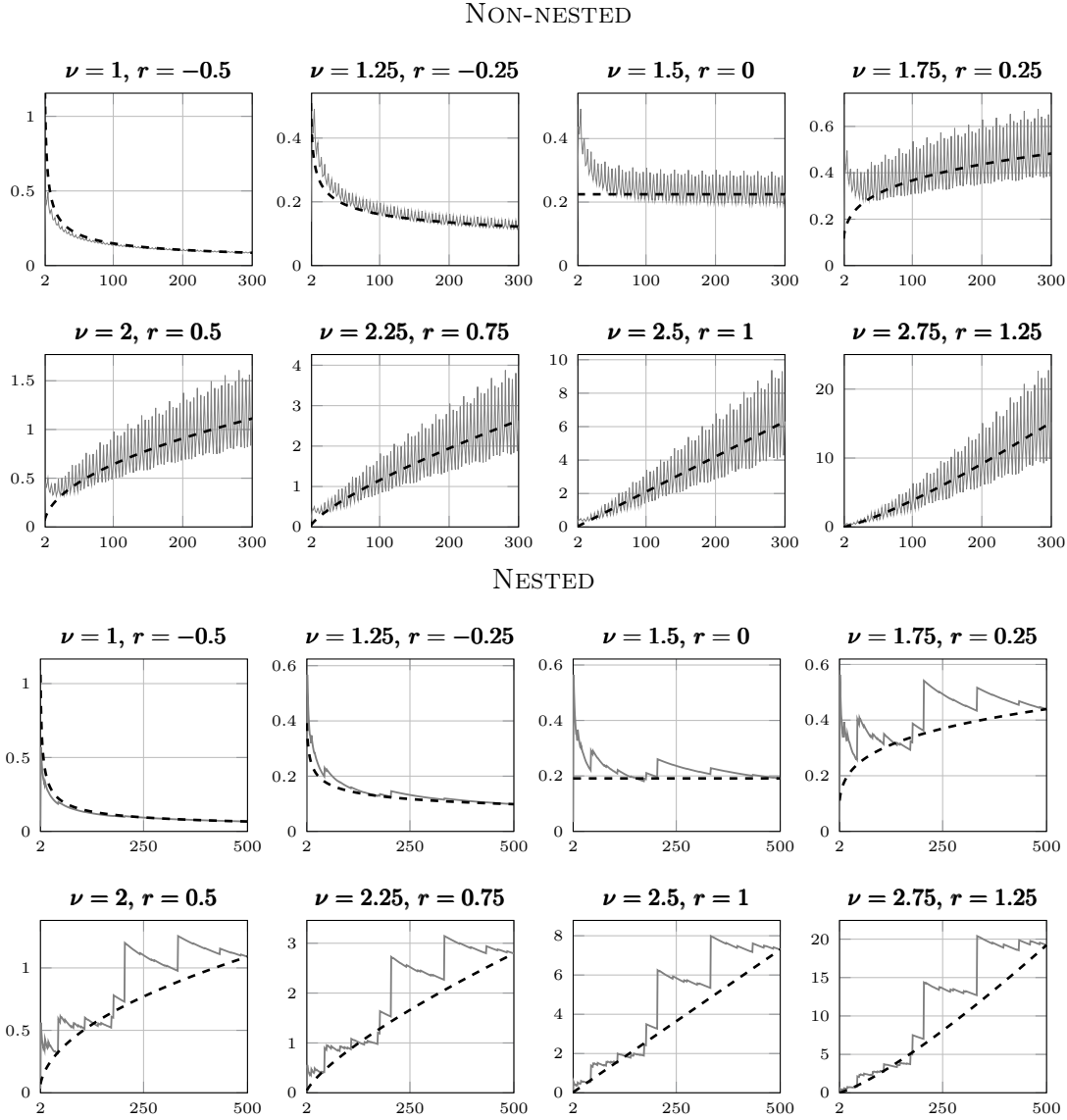
Figure 1. *Maximum likelihood estimates $\sigma_{\mathrm{ML}}(f, X_N)$ (grey) and theoretically predicted rates N^r with $r = \nu - 3/2$ (dashed black) when $d = 1$, as a function of the size N of the point set. Above: non-nested uniform point sets (5.3) for $N = 2, \ldots, 300$. Below: nested van der Corput points (5.4) for $N = 2, \ldots, 500$. The function f is of the form (5.1) with $\eta = 0.5$. The GP covariance kernel is a Matérn (4.2) with smoothness ν.*

Because the fill-distances and separation radii of these sets are not equal, the maximum likelihood estimates exhibit sudden increases intercepted by periods of decay contributed by the $N^{-1/2}$ term in (4.6) and (4.7). However, overall behaviour of $\sigma_{\mathrm{ML}}(f, X_N)$ appears to be compatible with the rate (5.2). Even though the plots are seemingly similar, $\sigma_{\mathrm{ML}}(f, X_N)$ grows much faster for larger ν as attested by changing y-scaling of the figures.
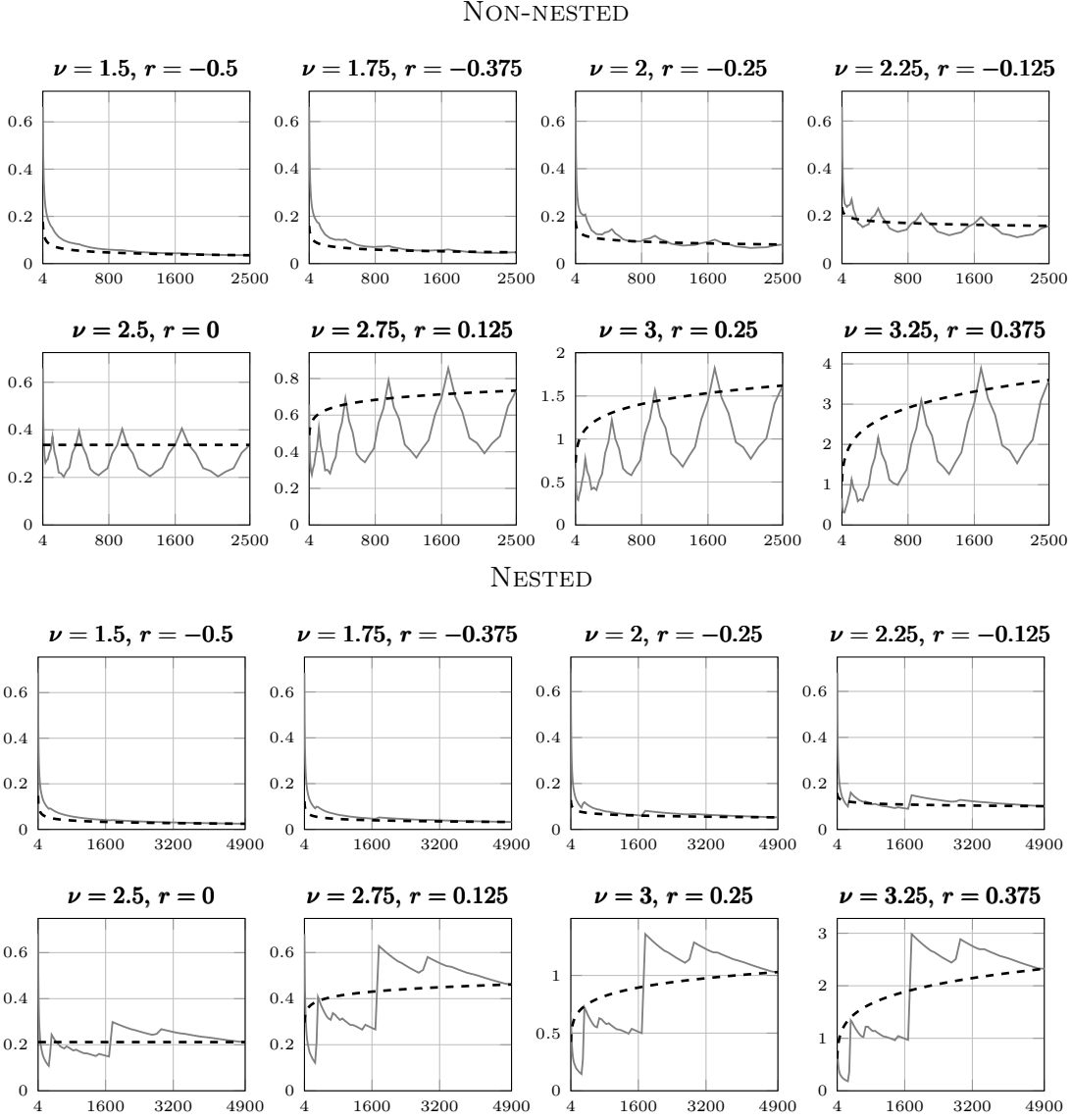
Figure 2. *Maximum likelihood estimates $\sigma_{\mathrm{ML}}(f, X_N)$ (grey) and theoretically predicted rates N^r with $r = \nu/2 - 5/4$ (dashed black) when $d = 2$, as a function of the size N of the point set.* Above: *Cartesian products of non-nested uniform point sets (5.3) for $N = 2^2, \ldots, 50^2$.* Below: *nested van der Corput points (5.4) for $N = 2^2, \ldots, 70^2$. The function f is of the form (5.1) with $\eta = 0.75$. The GP covariance kernel is a Matérn (4.2) with smoothness ν.*

MLE when $d = 2$. In the second example we set $d = 2$, $\ell = 0.8$, $\eta = 0.75$, $m = 3$, $(a_1, a_2, a_3) = (1, 0.5, 0.2)$, and $(z_1, z_2, z_3) = ((0.1, 0.1), (0.5, 0.1), (0.725, 0.565))$. The results are displayed in Figure 2. The point sets are now Cartesian products of the point sets used in the previous one-dimensional example. The maximum likelihood estimates again appear to behave as predicted by (5.2).

5.2. Credible Sets. The uncertainty quantification provided by GPs, as measured by the true function or its integral being contained in the credible sets (2.8), has been empirically assessed by various authors in a number of problems of varying character (Karvonen et al., 2018; Briol et al., 2019; Rathinavel and Hickernell, 2019). The theoretical results that we report may help to explain such empirical results previously observed.

In this example we study asymptotic overconfidence and underconfidence on $\Omega = [0, 1] \subset \mathbb{R}$ using the released once integrated Brownian motion kernel

$$(5.5) \qquad K(x, y) = 1 + xy + \frac{1}{3} \min\{x, y\}^3 + \frac{1}{2} |x - y| \min\{x, y\}^2$$

whose RKHS is $W_2^2([0, 1])$ (in parlance of Section 4, $\alpha = 2$). The term $1 + xy$ "releases" the standard integrated Brownian motion by removing the requirement that $f(0) = f'(0) = 0$. See van der Vaart and van Zanten (2008, Section 10) and Karvonen (2019, Section 2.2.3) for details about integrated Brownian motion kernels. For simplicity we only consider Bayesian quadrature for the computation of unweighted Lebesgue integrals on $[0, 1]$. Our integrands are of the form (5.1) with fixed $m = 3$, $\ell = 0.7$, $(a_1, a_2, a_3) = (1, 2, 0.5)$, and $(z_1, z_2, z_3) = (0.125, 0.5, 0.75)$. Six different smoothness parameters are used: $\eta = n/4$ for $n = 1, \ldots, 6$. As described in Section 5.1, this implies that the integrands are elements of $S^\beta([0, 1])$ with $\beta = 2\eta + 1/2 = (n + 1)/2$ for $n = 1, \ldots, 6$. For each $N \geq 1$ the point set X_N consists of the N first elements in the van der Corput sequence (5.4).

The results are depicted in Figure 3. In the right hand panel we see that asymptotic overconfidence appears to occur when f is less smooth than the RKHS ($\eta = 0.25$ and $\eta = 0.50$), though it is not clear with which rate this happens. When $\eta = 0.75$, which corresponds to f being on the "boundary" of the RKHS, credible sets appear to be either slowly asymptotically overconfident or asymptotically honest. When $f \in W_2^2([0, 1])$ ($\eta = 1.00$ and $\eta = 1.25$) the GP model appears to be asymptotically honest. Note that in all cases the relevant theoretical result, Theorem 4.10, only guarantees overconfidence, if it happens, cannot happen too fast in that $|I(f) - Q_{X_N}(f)| / R_{\mathrm{BC}}(f, X_N) = O(N^{1/2}(\log N)^{1/\beta})$.

6. Conclusion. In this article we analysed the asymptotic behaviour, as the number of data points grows, of maximum likelihood estimates of the scale parameter of a GP in the context of approximation and integration of a function that is exactly observed. The results on maximum likelihood estimation were then used to show that in some settings the GP model can become at worst "slowly" overconfident.

Similar analysis of other common kernel parameters, in particular the length-scale parameter of a stationary kernel, and their effect on uncertainty quantification would be a logical next step. Some work exists in the setting of the white noise model (Szabó et al., 2015; Hadji and Szabó, 2019). However, such analysis is greatly complicated by the lack of a closed-form expression like (2.6) for more general maximum likelihood estimators. For the length-scale parameter there is some evidence that its MLE can converge to a constant $\ell_\infty \in (0, \infty)$ even when the GP model is misspecified (Karvonen et al., 2019). Although proper selection of this parameter is often a prerequisite for an accurate and meaningful uncertainty quantification when N is small, it would follow that the parameter has no effect on asymptotic overconfidence or underconfidence of the model. Teckentrup (2019) has recently proved approximation error bounds for GPs over
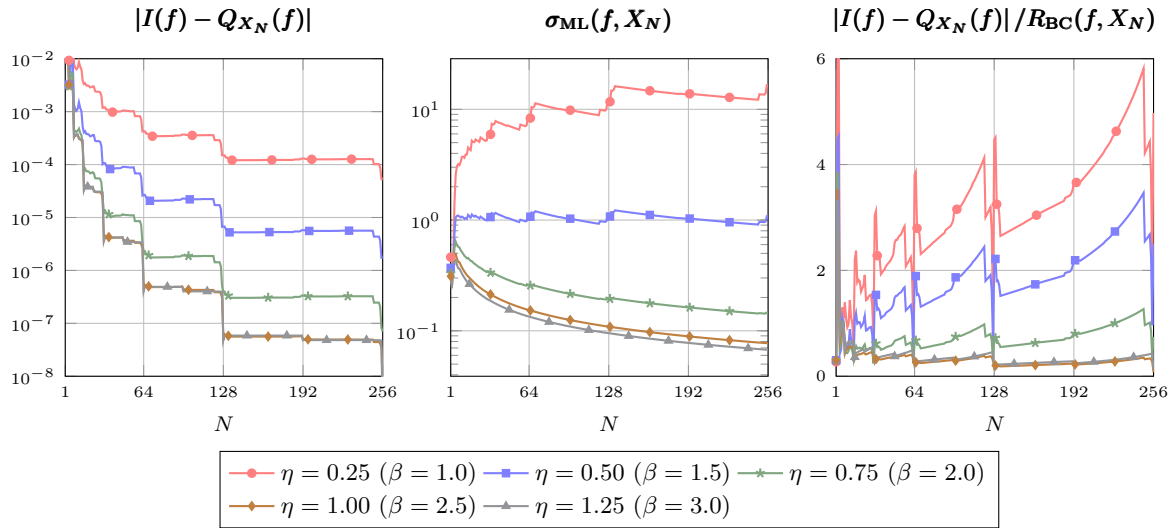
Figure 3. *Behaviour of the integration error (*left*), maximum likelihood estimate of the kernel scale parameter (*centre*), and the standard score in* (2.9) *and* (2.10) *(*right*) for a function f of the form* (5.1) *with $\eta = n/4$, $n = 1, \ldots, 6$, corresponding to $f \in S^\beta([0,1])$ for $\beta = (n+1)/2$, $n = 1, \ldots, 6$. The functions are evaluated at the nested van der Corput points* (5.4) *for $N = 1, \ldots, 256$. The GP covariance kernel is the released integrated Brownian motion kernel* (5.5) *whose RKHS is $W_2^2([0,1])$.*

compact sets of kernel parameters by assuming that the associated RKHS norm-equivalence constants can be uniformly bounded. If the existence of a limit ℓ_∞ could be established, it is likely that results in Teckentrup (2019) could be leveraged to extend the results in Section 4.5 to simultaneous maximum likelihood estimation of the scale and length-scale parameters.

There are also other popular approaches to kernel parameter selection. In *marginalisation*, or *full Bayes*, the scale parameter is treated as random and assigned an improper prior with density $p(\sigma^2) \propto 1/\sigma^2$ before being marginalised out (see MacKay, 1996). If $N \geq 3$, the conditional process becomes a Student's t process with N degrees of freedom, whose mean function is still $s_{f,X}$ but whose covariance function is now $(f_X^\mathsf{T} K_X^{-1} f_X/(N-2))P_X(x,y)$. The Student's t distribution converges to a Gaussian when its degrees of freedom increases, which implies that the resulting posterior is indistinguishable from the one obtained using maximum likelihood in the large N limit. As a consequence, the asymptotic results of this article apply equally to the case where σ is marginalised. *Cross-validation* offers more possibilities, both rooted (Fong and Holmes, 2019) and not rooted (e.g., Rathinavel and Hickernell, 2019, Section 2.2.3) in the GP model, to some of which our results on maximum likelihood may be relevant. An empirical investigation on cross-validation has been performed in Bachoc (2013).

Acknowledgements. We thank Gabriele Santin for pointing out some useful references.

Appendix A. Proofs for Sections 4.3 and 4.4.

This appendix contains proofs for the results in Sections 4.3 and 4.4. Unlike in the statements of the results, here various constants are tracked carefully because these constants

need to be controlled in the proofs of Theorem 4.3 and Propositions 4.6 and 4.8.

Lemma A.1. *Suppose that $\alpha \geq \beta > d/2$ and $\Omega \subset \mathbb{R}^d$ satisfies Assumption 4.1. If $f \in W_2^\beta(\Omega)$ and $X \subset \Omega$ is a finite set of points, then there is $f_\beta \in W_2^\alpha(\Omega)$ such that $f_\beta|_X = f|_X$,*

$$
(A.1) \qquad \|f - f_\beta\|_{W_2^\beta(\Omega)} \leq 5 \|f\|_{W_2^\beta(\Omega)}, \quad \text{and} \quad \|f_\beta\|_{W_2^\alpha(\Omega)} \leq C_\beta q_X^{\beta-\alpha} \|f\|_{W_2^\beta(\Omega)},
$$

where the constant $C_\beta > 0$ does not depend on f or X and varies continuously with β.

Proof. For any band-limited f_σ with band-limit $\sigma \geq 1$ we have

$$
(A.2) \quad
\begin{aligned}
\|f_\sigma\|_{W_2^\alpha(\mathbb{R}^d)}^2 &= \int_{\|\xi\| \leq \sigma} \left(1 + \|\xi\|^2\right)^\alpha |\widehat{f_\sigma}(\xi)|^2 \, d\xi \\
&\leq (1 + \sigma^2)^{\alpha-\beta} \int_{\|\xi\| \leq \sigma} \left(1 + \|\xi\|^2\right)^\beta |\widehat{f_\sigma}(\xi)|^2 \, d\xi \\
&= (1 + \sigma^2)^{\alpha-\beta} \|f_\sigma\|_{W_2^\beta(\mathbb{R}^d)}^2 \\
&\leq 2^{\alpha-\beta} \sigma^{2(\alpha-\beta)} \|f_\sigma\|_{W_2^\beta(\mathbb{R}^d)}^2 .
\end{aligned}
$$

Let $f_0 \in W_2^\beta(\mathbb{R}^d)$ be any extension of f. By Theorem 3.4 in Narcowich et al. (2006) there exists $f_\beta \in W_2^\alpha(\mathbb{R}^d)$ with band-width $\sigma = \kappa_\beta q_X^{-1}$ such that $f_\beta|_X = f|_X$ and

$$
(A.3) \qquad\qquad \|f_0 - f_\beta\|_{W_2^\beta(\mathbb{R}^d)} \leq 5 \|f_0\|_{W_2^\beta(\mathbb{R}^d)} .
$$

The constant $\kappa_\beta > 0$ depends only on d and β and can be selected such that $\sigma = \kappa_\beta q_X^{-1} \geq 1$ for any points $X \subset \Omega$. That it varies continuously with β is ascertained by observing that, according to the proof of Lemma 3.3 in Narcowich et al. (2006), it is a continuous combination of the constants $C_{\beta,d} = \Phi_\beta(0) + \sum_{n=1}^\infty 3d(n+2)^{d-1}\Phi_\beta(n)$, where $\Phi_\beta(x) = (1+x^2)^{-\beta}$ for $x \in \mathbb{R}$, and $c_{\beta,d}$, given in Wendland (2005, Theorem 12.3), which are continuous functions of β. The second claim follows from (A.2) and (A.3):

$$
\begin{aligned}
\|f_\beta\|_{W_2^\alpha(\mathbb{R}^d)} &\leq 2^{(\alpha-\beta)/2}\sigma^{\alpha-\beta}\|f_\beta\|_{W_2^\beta(\mathbb{R}^d)} \leq 2^{(\alpha-\beta)/2}\sigma^{\alpha-\beta}\left(\|f_0\|_{W_2^\beta(\mathbb{R}^d)} + \|f_0 - f_\beta\|_{W_2^\beta(\mathbb{R}^d)}\right) \\
&\leq 6 \times 2^{(\alpha-\beta)/2}\kappa_\beta^{\alpha-\beta} q_X^{\beta-\alpha}\|f_0\|_{W_2^\beta(\mathbb{R}^d)} .
\end{aligned}
$$

That is, $C_\beta = 6 \times 2^{(\alpha-\beta)/2}\kappa_\beta^{\alpha-\beta}$. The Sobolev norms over $\mathbb{R}^d$ in the inequalities can be replaced with norms over Ω because the inequalities are valid for any extension of f. ∎

Theorem A.2 (Wendland and Rieger 2005, Theorem 2.6). *Let $\alpha \geq \beta$, $\lfloor\beta\rfloor > d/2$, and $p \in [1, \infty]$. Suppose that $\Omega \subset \mathbb{R}^d$ satisfies Assumption 4.1 and K is a Sobolev kernel of order α. If $f \in W_2^\beta(\Omega)$, then there are constants $C_\beta, h_{0,\beta} > 0$ such that*

$$
\|f - s_{f,X}\|_{L^p(\Omega)} \leq C_\beta h_X^{\beta - d(1/2-1/p)_+} \|f - s_{f,X}\|_{W_2^\beta(\Omega)}
$$

whenever $h_X \leq h_{0,\beta}$, where $(x)_+ := \max\{x, 0\}$. The constant C_β depends on d, Ω, p, and $\lfloor\beta\rfloor$ and $h_{0,\beta}$ on d, Ω, and $\lfloor\beta\rfloor$.

Proof. The result as stated here follows by setting $k = \lfloor \beta \rfloor$, $s = \beta - \lfloor \beta \rfloor$, $p = 2$, $q = p$, $m = 0$, and $u = f - s_{f,X} \in W_2^\beta(\Omega)$ in Theorem 2.6 of Wendland and Rieger (2005). ∎

Proof of Theorem 4.2. Theorem A.2 with $\beta = \alpha$ and $p = 2$, the norm-equivalence (4.1), and $\|g - s_{g,X}\|_{\mathcal{H}_K(\Omega)} \le \|g\|_{\mathcal{H}_K(\Omega)}$ for any $g \in \mathcal{H}_K(\Omega)$ give $\|g - s_{g,X}\|_{W_2^\alpha(\Omega)} \le C_K^{-1} C_K' \|g\|_{W_2^\alpha(\Omega)}$ and, if $h_X \le h_{0,\alpha}$,

$$\|g - s_{g,X}\|_{L^2(\Omega)} \le C_\alpha h_X^\alpha \|g - s_{g,X}\|_{W_2^\alpha(\Omega)} \le C_K^{-1} C_K' C_\alpha h_X^\alpha \|g\|_{W_2^\alpha(\Omega)}.$$

Lemma 2.1 in Narcowich et al. (2006) therefore holds with the mapping $Tg = g - s_{g,X}$ and constants $\tau = \alpha$, $C_1 = C_K^{-1} C_K' C_\alpha h_X^\alpha$, and $C_2 = C_K^{-1} C_K'$. It follows that

$$\|Tg\|_{W_2^\beta(\Omega)} = \|g - s_{g,X}\|_{W_2^\beta(\Omega)} \le C_1^{1-\beta/\alpha} C_2^{\beta/\alpha} \|g\|_{W_2^\alpha(\Omega)} = C_K^{-1} C_K' C_\alpha^{1-\beta/\alpha} h_X^{\alpha-\beta} \|g\|_{W_2^\alpha(\Omega)}$$

for $g \in W_2^\beta(\Omega)$ and $h_X \le h_{0,\alpha}$. Select now g as the function $f_\beta \in W_2^\alpha(\Omega)$ in Lemma A.1 and let C_β' be the constant in (A.1). Then, exploiting the fact that $f_\beta|_X = f|_X$ and thus $s_{f_\beta,X} = s_{f,X}$,

$$\|f - s_{f,X}\|_{W_2^\beta(\Omega)} \le \|f - f_\beta\|_{W_2^\beta(\Omega)} + \|f_\beta - s_{f_\beta,X}\|_{W_2^\beta(\Omega)} + \|s_{f_\beta,X} - s_{f,X}\|_{W_2^\beta(\Omega)}$$

$$\le 5 \|f\|_{W_2^\beta(\Omega)} + C_K^{-1} C_K' C_\alpha^{1-\beta/\alpha} h_X^{\alpha-\beta} \|f_\beta\|_{W_2^\alpha(\Omega)}$$

$$\le \left(5 + C_K^{-1} C_K' C_\alpha^{1-\beta/\alpha} C_\beta' \rho_X^{\alpha-\beta} \right) \|f\|_{W_2^\beta(\Omega)}.$$

Since the mesh ratio satisfies $\rho_X \ge 1$, we can write this as $\|f - s_{f,X}\|_{W_2^\beta(\Omega)} \le C_\beta^* \rho_X^{\alpha-\beta} \|f\|_{W_2^\beta(\Omega)}$ for a constant $C_\beta^* > 0$ varying continuously with β. Finally, Theorem A.2 yields

(A.4)
$$\|f - s_{f,X}\|_{L^p(\Omega)} \le C_\beta h_X^{\beta - d(1/2 - 1/p)_+} \|f - s_{f,X}\|_{W_2^\beta(\Omega)}$$
$$\le C_\beta C_\beta^* h_X^{\beta - d(1/2-1/p)_+} \rho_X^{\alpha-\beta} \|f\|_{W_2^\beta(\Omega)}$$

if $h_X \le \min\{h_{0,\alpha}, h_{0,\beta}\}$. The claims of Theorem 4.2 follow from the above inequality with $p = 1$ and $p = \infty$, the inequality $|I(f) - Q_X(f)| \le \sup_{x \in \Omega} w(x) \|f - s_{f,X}\|_{L^1(\Omega)}$ (w is the weight function from Section 2.2), and that ρ_{X_N} is bounded for quasi-uniform $(X_N)_{N=1}^\infty \subset \Omega$. ∎

Note that for any bounded set $B \subset [\lceil d/2 \rceil, \infty)$ the constants related to (A.4) satisfy

(A.5)
$$\sup_{\beta \in B} C_\beta C_\beta^* < \infty \quad \text{and} \quad \inf_{\beta \in B} \min\{h_{0,\alpha}, h_{0,\beta}\} > 0$$

because, as remarked in the proof, C_β^* is a continuous function of β and C_β and $h_{0,\beta}$, which are the constants in Theorem A.2, depend on $\lfloor \beta \rfloor$ and can thus take only a finite number of distinct values for $\beta \in B$. This observation is used in the next proof.

Proof of Theorem 4.3. The proof is based on the fact that $S_-^\beta(\mathbb{R}^d) \subset W_2^\gamma(\mathbb{R}^d)$ for every $\beta > \gamma$ and given here only for approximation. Let $f_0 \in S_-^\beta(\mathbb{R}^d) \cap W_2^\gamma(\mathbb{R}^d)$ be an extension of $f \in S_-^\beta(\Omega)$. For a quasi-uniform sequence $(X_N)_{N=1}^\infty \subset \Omega$ it follows from (A.4) and (A.5) that

(A.6)
$$\sup_{x \in \Omega} |f(x) - s_{f,X_N}(x)| \le C N^{-\gamma/d + 1/2} \|f\|_{W_2^\gamma(\Omega)} \le C N^{-\gamma/d + 1/2} \|f_0\|_{W_2^\gamma(\mathbb{R}^d)}$$

for every $\gamma \in B := [\lceil d/2 \rceil, \beta)$ and all $N \geq N_0$, where

$$C := \sup_{\gamma \in B} C_\gamma C_\gamma^* < \infty \quad \text{and} \quad N_0 := \left(C_{\mathrm{qu}} \inf_{\gamma \in B} \min\{h_{0,\alpha}, h_{0,\gamma}\} \right)^{-d}$$

are independent of γ and $C_{\mathrm{qu}} > 0$ is a constant such that $C_{\mathrm{qu}}^{-1} N^{-1/d} \leq h_{X_N} \leq C_{\mathrm{qu}} N^{-1/d}$ for all $N \geq 1$ (the existence of which follows from quasi-uniformity). Set $\gamma = \gamma_N := \beta - 1/\log N \to \beta$. Because $f_0 \in S_-^\beta(\mathbb{R}^d)$, a spherical coordinate transform gives, with constants $C_1, C_2 > 0$ that depend on γ and β, d, and f_0 and remain bounded away from zero and infinity as $\gamma \to \beta$,

$$\|f_0\|_{W_2^\gamma(\mathbb{R}^d)}^2 = \int_{\mathbb{R}^d} \left(1 + \|\xi\|^2\right)^\gamma |\widehat{f_0}(\xi)|^2 \, d\xi \leq C_1 \int_{\|\xi\| \geq 1} \|\xi\|^{2(\gamma - \beta) - d} \, d\xi = C_1 C_2 \int_1^\infty r^{2(\gamma - \beta) - 1} \, dr$$
$$= \frac{C_1 C_2}{2(\beta - \gamma)}.$$

By inserting $\gamma = \gamma_N = \beta - 1/\log N$ into (A.6) and exploiting the estimate above we thus get

$$\sup_{x \in \Omega} |f(x) - s_{f, X_N}(x)| \lesssim N^{-\gamma_N/d + 1/2} \|f\|_{W_2^{\gamma_N}(\mathbb{R}^d)} \lesssim N^{-\beta/d + 1/2} N^{1/(d \log N)} (\log N)^{1/2}$$
$$= \mathrm{e}^{1/d} N^{-\beta/d - 1/2} (\log N)^{1/2},$$

as claimed. ∎

Proof of Theorem 4.4. The rates $\sup_{x \in \Omega} P_{X_N}(x, x)^{1/2} \asymp N^{-\alpha/d + 1/2}$ and $V_{X_N}^{1/2} \asymp N^{-\alpha/d}$ follow the worst-case interpretation (3.2) of the standard deviations, Theorem 4.2 with $\beta = \alpha$, and standard results on fundamental lower bounds for the rate of convergence of approximation and integration algorithms in Sobolev spaces, which can be found in Novak (1988, Sections 1.3.11 and 1.3.12); Ritter (2000, Section 1.2, Chapter VI); and Novak and Woźniakowski (2008, Section 4.2.4). We are left to prove the lower bound $P_{X_N}(x, x)^{1/2} \gtrsim N^{-\alpha/d + 1/2}$ for fixed x. Although this lower bound is more or less standard (e.g., Schaback, 1995), we have not found the exact version given here in the literature.

By (3.2) the conditional standard deviation has the worst-case interpretation

$$P_{X_N}(x, x)^{1/2} = \sup_{\|g\|_{\mathcal{H}_K(\Omega)} \leq 1} |g(x) - s_{g, X_N}(x)|.$$

If there is $g \in \mathcal{H}_K(\Omega)$ such that $g|_{X_N} \equiv 0$ it follows that $P_{X_N}(x, x)^{1/2} \geq |g(x)| \|g\|_{\mathcal{H}_K(\Omega)}^{-1}$ because in this case $s_{g, X_N} \equiv 0$. We follow the proof of Theorem 1 in De Marchi and Schaback (2010), a standard bump function argument, to construct this function. Let $\phi \colon \mathbb{R}^d \to \mathbb{R}$ be an infinitely smooth bump function that is supported on the unit ball and satisfies $\sup_{x \in \mathbb{R}^d} \phi(x) = \phi(0) = 1$. Let $\delta_{x, X_N} := \min_{i=1,\ldots,N} \|x - x_i\|$ be the distance between $x \in \Omega$ and $X_N \subset \Omega$. Define $\phi_x := \phi(\cdot - x)$ and $g_x := \phi_x(\cdot / \delta_{x, X_N}) \in W_2^\alpha(\Omega)$, which satisfies $g_x|_{X_N} \equiv 0$. Using the properties of the Fourier transform, a change of variables, and the fact that $\delta_{x, X_N} \leq h_{X_N} \leq 1$ for all

sufficiently large N due to quasi-uniformity we get

$$\|g_x\|_{W_2^\alpha(\Omega)}^2 \leq \delta_{x,X_N}^{2d} \int_{\mathbb{R}^d} \left(1 + \|\xi\|^2\right)^\alpha |\widehat{\phi}_x(\delta_{x,X_N}\xi)|^2 \, \mathrm{d}\xi = \delta_{x,X_N}^d \int_{\mathbb{R}^d} \left(1 + \frac{\|\xi\|^2}{\delta_{x,X_N}^2}\right)^\alpha |\widehat{\phi}_x(\xi)|^2 \, \mathrm{d}\xi$$

$$\leq \delta_{x,X_N}^{d-2\alpha} \int_{\mathbb{R}^d} \left(1 + \|\xi\|^2\right)^\alpha |\widehat{\phi}(\xi)|^2 \, \mathrm{d}\xi$$

$$= \delta_{x,X_N}^{d-2\alpha} \|\phi\|_{W_2^\alpha(\mathbb{R}^d)}^2.$$

By norm-equivalence we thus have $\|g_x\|_{\mathcal{H}_K(\Omega)} \leq C\delta_{x,X_N}^{-\alpha+d/2}$ for a constant $C > 0$ which is independent of x and X_N. Therefore

$$(A.7) \qquad\qquad P_{X_N}(x,x)^{1/2} \geq C^{-1}\delta_{x,X_N}^{\alpha-d/2}.$$

Because the point set sequence is quasi-uniform, there is a constant $C_{\mathrm{qu}} > 0$ such that $C_{\mathrm{qu}}^{-1}N^{-1/d} \leq q_{X_N}$ for all $N \geq 1$. If $\delta_{x,X_N} \geq q_{X_N}$, then $\delta_{x,X_N} \geq c_1 N^{-1/d}$ and (A.7) gives

$$(A.8) \qquad\qquad P_{X_N}(x,x)^{1/2} \geq C^{-1}\delta_{x,X_N}^{\alpha-d/2} \geq C^{-1}c_1^{-\alpha+d/2}N^{-\alpha/d+1/2},$$

the claimed lower bound. Assume thus that $\delta_{x,X_N} \leq q_{X_N}$. Let $x^* \in X_N$ be a point closest to $x \notin X_N$. Then $2q_{X_N} \leq \|x^* - x\| + \|x - x'\| = \delta_{x,X_N} + \|x - x'\|$ for any $x' \in X_N \setminus \{x^*\}$. If N is such that $\delta_{x,X_N} < \delta_{x,X_{N-1}}$ there is x' such that $\|x - x'\| = \delta_{x,X_{N-1}}$. For such N we thus have $2q_{X_N} \leq \delta_{x,X_N} + \delta_{x,X_{N-1}}$, which together with $\delta_{x,X_N} \geq q_{X_N}$ and quasi-uniformity yields

$$\delta_{x,X_{N-1}} \geq C_{\mathrm{qu}}^{-1}N^{-1/d} \geq 2^{-1/d}C_{\mathrm{qu}}^{-1}(N-1)^{-1/d}$$

for $N \geq 2$. Again, a lower bound of the form (A.8) thus holds. This completes the proof. ∎

Proof of Proposition 4.5. For any distinct points $X \subset \Omega$ Lemma A.1, $s_{f,X} = s_{f_\beta,X}$, the minimum-norm property of the GP conditional mean, and the norm-equivalence (4.1) yield

$$\|s_{f,X}\|_{\mathcal{H}_K(\Omega)} = \|s_{f_\beta,X}\|_{\mathcal{H}_K(\Omega)} \leq \|f_\beta\|_{\mathcal{H}_K(\Omega)} \leq C'_K \|f_\beta\|_{W_2^\alpha(\Omega)} \leq C'_K C_\beta q_X^{\beta-\alpha} \|f\|_{W_2^\beta(\Omega)}.$$

The claims follow from (3.4) and the fact that $q_{X_N} \gtrsim N^{-1/d}$ for quasi-uniform points. ∎

Proof of Proposition 4.6. The proof is similar to that of Theorem 4.3. The use of Theorem 4.2 is replaced with Proposition 4.5 which implies that

$$\sigma_{\mathrm{ML}}(f,X_N) \leq C_\gamma N^{(\alpha-\gamma)/d-1/2} \|f_0\|_{W_2^\gamma(\mathbb{R}^d)},$$

where $f_0 \in S_-^\beta(\mathbb{R}^d) \cap W_2^\gamma(\mathbb{R}^d)$ is an extension of f and C_γ again satisfies $\sup_{\gamma \in [\lceil d/2\rceil,\beta)} C_\gamma < \infty$. ∎

Proof of Proposition 4.7. This proof is adapted from the proof of Theorem 8 in van der Vaart and van Zanten (2011). Let $X \subset \Omega$ be any distinct points. By Theorem 4.2 there are $C_\gamma, h_{0,\gamma} > 0$ such that for $h_X \leq h_{0,\gamma}$,

$$(A.9) \qquad\qquad \|f - s_{f,X}\|_{L^2(\Omega)} \leq C_\gamma h_X^\gamma \rho_X^{\alpha-\gamma} \|f_0\|_{W_2^\gamma(\mathbb{R}^d)} =: \varepsilon_X.$$

In the proof of Theorem 4.3 it is shown that C_γ and $h_{0,\gamma}$ are bounded away from zero and infinity if γ remains in a bounded interval. Because the support of f_0 is compact and contained in the interior of Ω, there is a non-negative bump function $\phi\colon \mathbb{R}^d \to \mathbb{R}$ such that $\phi|_{\mathrm{supp}(f_0)} \equiv 1$, $\phi|_{\mathbb{R}^d \setminus \mathrm{int}(\Omega)} \equiv 0$, $\sup_{x \in \Omega} \phi(x) = 1$, and $|\widehat{\phi}(\xi)\mathrm{e}^{\|\xi\|^u}| \to 0$ as $\|\xi\| \to \infty$ for some $u > 0$. Let $s_{f,X,0} \in W_2^\alpha(\mathbb{R}^d)$ be an extension of $s_{f,X} \in W_2^\alpha(\Omega)$. By Parseval's identity and $f_0 = f_0\phi$,

$$(A.10) \qquad \|\widehat{f_0} - \widehat{s}_{f,X,0} * \widehat{\phi}\|_{L^2(\mathbb{R}^d)} = \|f_0 - s_{f,X,0}\phi\|_{L^2(\mathbb{R}^d)} \leq \|f - s_{f,X}\|_{L^2(\Omega)} \leq \varepsilon_X.$$

For $R > 0$ let $\mathbb{1}_R^{\mathsf{c}}$ be the indicator function of the set $\{x \in \mathbb{R}^d : \|x\| > R\}$. Because $f_0 \in S_+^\beta(\mathbb{R}^d)$, for sufficiently large R we have

$$\|\widehat{f_0}\mathbb{1}_{2R}^{\mathsf{c}}\|_{L^2(\mathbb{R}^d)}^2 = \int_{\|\xi\|>2R} |\widehat{f_0}(\xi)|^2 \,\mathrm{d}\xi \geq C_f \int_{\|\xi\|>2R} \|\xi\|^{-2\beta-d} \,\mathrm{d}\xi \geq C_f \widetilde{C} R^{-2\beta} =: C_1 R^{-2\beta},$$

where $C_f > 0$ depends on f_0 and $\widetilde{C} > 0$ on β and d. The reverse triangle inequality and (A.10) thus yield

$$\|(\widehat{s}_{f,X,0} * \widehat{\phi})\mathbb{1}_{2R}^{\mathsf{c}}\|_{L^2(\mathbb{R}^d)} \geq \|\widehat{f_0}\mathbb{1}_{2R}^{\mathsf{c}}\|_{L^2(\mathbb{R}^d)} - \|(\widehat{f_0} - \widehat{s}_{f,X,0} * \widehat{\phi})\mathbb{1}_{2R}^{\mathsf{c}}\|_{L^2(\mathbb{R}^d)} \geq C_1 R^{-\beta} - \varepsilon_X.$$

By Lemma 16 in van der Vaart and van Zanten (2011),

$$(A.11) \qquad \|\widehat{s}_{f,X,0}\mathbb{1}_R^{\mathsf{c}}\|_{L^2(\mathbb{R}^d)} \|\widehat{\phi}(1 - \mathbb{1}_R^{\mathsf{c}})\|_{L^1(\mathbb{R}^d)} \geq C_1 R^{-\beta} - \varepsilon_X - \|\widehat{s}_{f,X,0}\|_{L^2(\mathbb{R}^d)} \|\widehat{\phi}\mathbb{1}_R^{\mathsf{c}}\|_{L^1(\mathbb{R}^d)}.$$

Let $m(\xi) := (1 + \|\xi\|^2)^{\alpha/2}$ so that $\|s_{f,X,0}\|_{W_2^\alpha(\mathbb{R}^d)} = \|\widehat{s}_{f,X,0}m\|_{L^2(\mathbb{R}^d)}$. We engage in slight abuse of notation by also writing $m(r) = (1 + r^2)^{\alpha/2}$ for $r \in \mathbb{R}$. By the definition of Sobolev norm,

$$\|\widehat{s}_{f,X,0}\mathbb{1}_R^{\mathsf{c}}\|_{L^2(\mathbb{R}^d)} = \|\widehat{s}_{f,X,0}m\mathbb{1}_R^{\mathsf{c}}m^{-1}\|_{L^2(\mathbb{R}^d)} \leq m(R)^{-1} \|s_{f,X,0}\|_{W_2^\alpha(\mathbb{R}^d)}$$

and $\|\widehat{s}_{f,X,0}\|_{L^2(\mathbb{R}^d)} \leq \|s_{f,X,0}\|_{W_2^\alpha(\mathbb{R}^d)}$. Set $R = C_1^{1/\beta}(2\varepsilon_X)^{-1/\beta}$ and use these estimates to rearrange (A.11) as

$$\left[m(R)^{-1} \|\widehat{\phi}(1 - \mathbb{1}_R^{\mathsf{c}})\|_{L^1(\mathbb{R}^d)} + \|\widehat{\phi}\mathbb{1}_R^{\mathsf{c}}\|_{L^1(\mathbb{R}^d)} \right] \|s_{f,X,0}\|_{W_2^\alpha(\mathbb{R}^d)} \geq C_1 R^{-\beta} - \varepsilon_X = \varepsilon_X.$$

By construction, $\widehat{\phi} \in L^1(\mathbb{R}^d)$ and $\|\widehat{\phi}\mathbb{1}_R^{\mathsf{c}}\|_{L^1(\mathbb{R}^d)} \leq C_2 \mathrm{e}^{-dR^u}$ for some constant $C_2 > 0$. Let $C_3 > 0$ be a constant such that $C_2 \mathrm{e}^{-dr^u} \leq C_3 m(r)^{-1}$ for all $r \geq 0$. Then

$$\begin{aligned}
\|s_{f,X,0}\|_{W_2^\alpha(\mathbb{R}^d)} &\geq \varepsilon_X \left[m(R)^{-1} \|\widehat{\phi}(1 - \mathbb{1}_R^{\mathsf{c}})\|_{L^1(\mathbb{R}^d)} + \|\widehat{\phi}\mathbb{1}_R^{\mathsf{c}}\|_{L^1(\mathbb{R}^d)} \right]^{-1} \\
&\geq \varepsilon_X \left[m(R)^{-1} \|\widehat{\phi}\|_{L^1(\mathbb{R}^d)} + C_2 \mathrm{e}^{-dR^u} \right]^{-1} \\
&\geq \varepsilon_X \left[m(R)^{-1} \big(\|\widehat{\phi}\|_{L^1(\mathbb{R}^d)} + C_3 \big) \right]^{-1} \\
&\geq C_4 \varepsilon_X^{1-\alpha/\beta},
\end{aligned}$$

where $C_4 = 2^{-1/\beta} C_1^{1/\beta} (\|\widehat{\phi}\|_{L^1(\mathbb{R}^d)} + C_3)^{-1}$ does not depend on γ. The definition of ε_X in (A.9) therefore gives

$$
\begin{aligned}
\|s_{f,X,0}\|_{W_2^\alpha(\mathbb{R}^d)} &\geq C_4 \big(C_\gamma h_X^\gamma \rho_X^{\alpha-\gamma} \|f_0\|_{W_2^\gamma(\mathbb{R}^d)} \big)^{1-\alpha/\beta} \\
&= C_4 C_\gamma^{1-\alpha/\beta} h_X^{\gamma(1-\alpha/\beta)} \rho_X^{-(\alpha-\gamma)(\alpha-\beta)/\beta} \|f_0\|_{W_2^\gamma(\mathbb{R}^d)}^{1-\alpha/\beta}.
\end{aligned}
$$

If $(X_N)_{N=1}^\infty$ is a quasi-uniform sequence, ρ_{X_N} remains bounded and $h_{X_N} \gtrsim N^{-1/d}$. Thus

$$
\|s_{f,X_N,0}\|_{W_2^\alpha(\mathbb{R}^d)} \gtrsim N^{\gamma(\alpha/\beta-1)/d} \|f_0\|_{W_2^\gamma(\mathbb{R}^d)}^{1-\alpha/\beta}.
$$

The claims now follow from the norm-equivalence of $\mathcal{H}_K(\Omega)$ and $W_2^\alpha(\Omega)$, the Sobolev extension theorem (Grisvald, 1985, Theorem 1.4.3.1), and (3.4). ∎

Proof of Proposition 4.8. Let $\gamma_N = \beta - 1/\log N$. The arguments in the proof of Theorem 4.6, the Sobolev extension theorem, and (4.8) yield

$$
\begin{aligned}
\sigma_{\mathrm{ML}}(f, X_N) &\gtrsim N^{\gamma_N(\alpha/\beta-1)/d-1/2} \|f_0\|_{W_2^{\gamma_N}(\mathbb{R}^d)}^{1-\alpha/\beta} \\
&\gtrsim N^{(\alpha-\beta)/d-1/2} N^{(1-\alpha/\beta)/(d\log N)} (\log N)^{(1-\alpha/\beta)/2} \\
&= \mathrm{e}^{(1-\alpha/\beta)/d} N^{(\alpha-\beta)/d-1/2} (\log N)^{(1-\alpha/\beta)/2},
\end{aligned}
$$

which is the claimed bound. ∎

References.

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.

Arcangéli, R., de Silanes, M. C. L., and Torrens, J. J. (2012). Extension of sampling inequalities to Sobolev semi-norms of fractional order and derivative data. *Numerische Mathematik*, 121(3):587–608.

Atkinson, K. E. (1989). *An Introduction to Numerical Analysis*. John Wiley & Sons, 2nd edition.

Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(19):1–38.

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. (2018). Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case. *Bernoulli*, 24(2):1531–1575.

Bachoc, F., Lagnoux, A., and Lopera-López, A. (2019). Maximum likelihood estimation for Gaussian processes under inequality constraints. *Electronic Journal of Statistics*, 13(2):2921–2969.

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.

Băzăvan, E. G., Li, F., and Sminchisescu, C. (2012). Fourier kernel learning. In *European Conference on Computer Vision*, pages 459–473. Springer.

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

Bogachev, V. I. (1998). *Gaussian Measures*. American Mathematical Society.

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.

Cockayne, J., Oates, C., Sullivan, T., and Girolami, M. (2019). Bayesian probabilistic numerical methods. *SIAM Review*, 61(4):756–789.

Dashti, M. and Stuart, A. M. (2017). The Bayesian approach to inverse problems. In *Handbook of Uncertainty Quantification*, pages 311–428. Springer.

De Marchi, S. and Schaback, R. (2010). Stability of kernel-based interpolation. *Advances in Computational Mathematics*, 32(2):155–161.

Dong, D. (2012). Mine gas emission prediction based on Gaussian process model. *Procedia Engineering*, 45:334–338.

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.

Duvenaud, D. (2014). *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge.

Fasshauer, G. and McCourt, M. (2015). *Kernel-based Approximation Methods Using MATLAB*. World Scientific Publishing.

Fasshauer, G. E. (2011). Positive definite kernels: past, present and future. *Dolomite Research Notes on Approximation*, 4:21–63.

Fong, E. and Holmes, C. C. (2019). On the marginal likelihood and cross-validation. *arXiv:1905.08737v2*.

Gao, P., Honkela, A., Rattray, M., and Lawrence, N. (2008). Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, 24(16):i70–i75.

Grisvald, P. (1985). *Elliptic Problems in Nonsmooth Domains*. Pitman Publishing.

Hadji, A. and Szabó, B. (2019). Can we trust Bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel? *arXiv:1904.01383v1*.

Hennig, P., Osborne, M. A., and Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 471(2179).

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*.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. (2020). Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *Foundations of Computational Mathematics*, 20:155–194.

Karvonen, T. (2019). *Kernel-Based and Bayesian Methods for Numerical Integration*. PhD thesis, Department of Electrical Engineering and Automation, Aalto University.

Karvonen, T., Oates, C. J., and Särkkä, S. (2018). A Bayes–Sard cubature method. In *Advancges in Neural Information Processing Systems 31*, pages 5882–5893.

Karvonen, T., Tronarp, F., and Särkkä, S. (2019). Asymptotics of maximum likelihood parameter estimates for Gaussian processes: The Ornstein–Uhlenbeck prior. In *Proceedings of the 29th IEEE International Workshop on Machine Learning for Signal Processing*.

Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464.

Kowalska, K. and Peel, L. (2012). Maritime anomaly detection using Gaussian process active learning. In *Proceedings of the 15th International Conference on Information Fusion*, pages 1164–1171.

Larkin, F. M. (1972). Gaussian measure in Hilbert space and applications in numerical analysis. *Rocky Mountain Journal of Mathematics*, 2(3):379–422.

Liebl, D. and Reimherr, M. (2019). Fast and fair simultaneous confidence bands for functional parameters. *arXiv:1910.00131v2*.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Liu, D., Pang, J., Zhou, J., Peng, Y., and Pecht, M. (2013). Prognostics for state of health estimation of lithium-ion batteries based on combination Gaussian process functional regression. *Microelectronics Reliability*, 53(6):832–839.

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.

MacKay, D. J. (1992). Bayesian interpolation. *Neural Computation*, 4(3):415–447.

MacKay, D. J. (1996). Hyperparameters: Optimize, or integrate out? In *Maximum Entropy and Bayesian Methods*, pages 43–59. Springer.

Manogaran, G. and Lopez, D. (2018). A Gaussian process based big data processing framework in cluster computing environment. *Cluster Computing*, 21(1):189–204.

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.

Novak, E. (1988). *Deterministic and Stochastic Error Bounds in Numerical Analysis*. Springer-Verlag.

Novak, E. and Woźniakowski, H. (2008). *Tractability of Multivariate Problems. Volume I: Linear Information*. European Mathematical Society.

Oettershagen, J. (2017). *Construction of Optimal Cubature Algorithms with Applications to Econometrics and Uncertainty Quantification*. PhD thesis, University of Bonn.

O'Hagan, A. (1991). Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260.

Oliva, J. B., Dubey, A., Wilson, A. G., Póczos, B., Schneider, J., and Xing, E. P. (2016). Bayesian nonparametric kernel-learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1078–1086.

Rajpaul, V., Aigrain, S., Osborne, M., Reece, S., and Roberts, S. (2015). A Gaussian process

framework for modelling stellar activity signals in radial velocity data. *Monthly Notices of the Royal Astronomical Society*, 452(3):2269–2291.

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

Rathinavel, J. and Hickernell, F. (2019). Fast automatic Bayesian cubature using lattice sampling. *Statistics and Computing*, 29(6):1215–1229.

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.

Ritter, K. (2000). *Average-Case Analysis of Numerical Problems*. Springer.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Schaback, R. (1995). Error estimates and condition numbers for radial basis function interpolation. *Advances in Computational Mathematics*, 3(3):251–264.

Schaback, R. (1999). Improved error bounds for scattered data interpolation by radial basis functions. *Mathematics of Computation*, 68(225):201–216.

Schaback, R. (2000). A unified theory of radial basis functions: Native Hilbert spaces for radial basis functions II. *Journal of Computational and Applied Mathematics*, 121(1–2):165–177.

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.

Shi, J. and Wang, B. (2008). Curve prediction and clustering with mixtures of Gaussian process functional regression models. *Statistics and Computing*, 18(3):267–283.

Stein, M. L. (1993). Spline smoothing with an estimated order parameter. *The Annals of Statistics*, 21(3):1522–1544.

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

Steinwart, I. (2019). Convergence types and rates in generic Karhunen-Loève expansions with applications to sample path properties. *Potential Analysis*, 51(3):361–395.

Steinwart, I. and Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417.

Sun, S., Zhang, G., Wang, C., Zeng, W., Li, J., and Grosse, R. (2018). Differentiable compositional kernel learning for Gaussian processes. In *35th International Conference on Machine Learning*, volume 31, pages 4828–4837.

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.

Teckentrup, A. L. (2019). Convergence of Gaussian process regression with estimated hyperparameters and applications in Bayesian inverse problems. *arXiv:1909.00232v2*.

Triebel, H. (2006). *Theory of Function Spaces III*. Birkhäuser Basel.

Tuo, R., Wang, W., 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. *arXiv:2001.00152v1*.

van der Vaart, A. W. and van Zanten, J. H. (2008). *Reproducing Kernel Hilbert spaces of Gaussian Priors*, volume 3 of *IMS Collections*, pages 200–222. Institute of Mathematical Statistics.

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

Wang, J., Hertzmann, A., and Fleet, D. J. (2006). Gaussian process dynamical models. In *Advances in Neural Information Processing Systems 19*, pages 1441–1448.

Wang, W. (2020). On the inference of applying Gaussian process modeling to a deterministic function. *arXiv:2002.01381v1*.

Wendland, H. (2005). *Scattered Data Approximation*. 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. (2020). Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness. *arXiv:2001.10818v2*.

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.

Yang, K., Keat Gan, S., and Sukkarieh, S. (2013). A Gaussian process-based RRT planner for the exploration of an unknown and cluttered environment with a UAV. *Advanced Robotics*, 27(6):431–443.