Journal of Machine Learning Research 22 (2021) 1-40 Submitted 6/20; Revised 5/21; Published 5/21

Convergence Guarantees for Gaussian Process Means With Misspecified Likelihoods and Smoothness

George Wynne G.WYNNE18@IMPERIAL.AC.UK
Department of Mathematics
Imperial College London
London, SW7 2AZ, UK

François-Xavier Briol F.BRIOL@UCL.AC.UK
Department of Statistical Science
University College London
London, WC1E 7HB, UK

Mark Girolami MAG92@ENG.CAM.AC.UK
Department of Engineering
University of Cambridge
Cambridge, CB2 1PZ, UK

Editor: Jean-Philippe Vert

Abstract

Gaussian processes are ubiquitous in machine learning, statistics, and applied mathematics. They provide a flexible modelling framework for approximating functions, whilst simultaneously quantifying uncertainty. However, this is only true when the model is well-specified, which is often not the case in practice. In this paper, we study the properties of Gaussian process means when the smoothness of the model and the likelihood function are misspecified. In this setting, an important theoretical question of practical relevance is how accurate the Gaussian process approximations will be given the chosen model and the extent of the misspecification. The answer to this problem is particularly useful since it can inform our choice of model and experimental design. In particular, we describe how the experimental design and choice of kernel and kernel hyperparameters can be adapted to alleviate model misspecification.

Keywords: Gaussian Processes, Kriging, Nonparametric Regression, Reproducing Kernel Hilbert Space, Sampling Inequality

1. Introduction

Gaussian processes (GPs) have found widespread use in machine learning (Rasmussen and Williams, 2006) as they offer flexible and interpretable models with uncertainty quantification. Applications include reinforcement learning (Kuss and Rasmussen, 2004), time-series modelling (Roberts et al., 2013), robotics and control (Deisenroth et al., 2015), as well as Bayesian numerical methods including Bayesian quadrature (Briol et al., 2019; Kanagawa et al., 2020), Bayesian optimization (Mockus, 1989; Snoek et al., 2012; Bull, 2011) and

Bayesian differential equations solvers (Cockayne et al., 2016). Outside of machine learning, Gaussian process regression was first used in geostatistics (Krige, 1951; Cressie, 1990; Matheron, 1963), where the procedure was originally known as kriging and is a current active field (Wang et al., 2019; Lederer et al., 2019). Gaussian processes are used for tackling problems ranging from computer models (Kennedy and O'Hagan, 2001) to inverse problems (Stuart, 2010; Stuart and Teckentrup, 2018), health monitoring (Stegle et al., 2008), engineering design (Forrester et al., 2008) and tsunami modelling (Sarri et al., 2012), to name but a few.

In most of the applications above, the central task is to approximate a function of interest given pointwise evaluations of this function which may be corrupted by some unknown noise. To do so, practitioners carefully design their algorithms such that the approximation error decreases at a fast rate in the number of data points. Several modelling choices need to be made, including the selection of a GP model and hyperparameters, of a likelihood, and of the locations at which to obtain data. Making appropriate choices for a given application is an extremely difficult task, and poor choices can lead to poor empirical performance. One way to tackle this problem in a unified manner is to turn to theoretical convergence guarantees which explicitly account for these modelling choices, and to select specific algorithms which minimise upper bounds on the approximation error.

Of course, this approach is only sensible if the bounds apply to the problem at hand, but most existing bounds are rather restrictive and require assumptions which users might not be able to verify. The novel contributions of this paper include convergence guarantees in the presence of two common modelling errors, and suggestions as to how to construct algorithms which can mitigate these.

The first is *likelihood misspecification*, meaning that the observations follow a distribution which is different from the one assumed by the model. This often occurs because conditioning of Gaussian process means on data is only possible in closed-form if assuming the data is noiseless, or contains independently and identically distributed Gaussian noise with known variance. For more complex observations, such as input-dependent noise (Goldberg et al., 1998; Le et al., 2005) or distributions with heavy tails (Vanhatalo et al., 2009), a closed-form expression for the mean is not available. In order to maintain a closed-form expression, practitioners often use simplistic models which may not be a faithful representation of the data-generating process, leading to a lack of robustness and poor approximations (Goldberg et al., 1998; Jylänki et al., 2011).

The second is *smoothness misspecification*, meaning that the Gaussian process mean is either too rough or too smooth relative to the target function. Here, the smoothness of a function is measured in terms of number of derivatives in the sense of Sobolev spaces. This is known to guide the rate of convergence of Gaussian process approximations, with faster rates attainable for smoother functions if the mean and covariance functions are chosen appropriately. However, for many of the aforementioned applications, it is difficult to identify the smoothness of the target function. This commonly leads to sub-optimal choices of GPs, and as a result potentially slower convergence rates.

Our novel convergence guarantees highlight the impact that both types of misspecification can have on rates of convergence, and can provide guidance on model choice for practitioners at risk of misspecification. In particular, the impact of the experimental design and covariance function is made clear in the bounds. The bounds employ results from the

scattered data approximation (SDA) literature (Wendland, 2005), which has been applied to GP related methods in numerous works (Bull, 2011; Stuart and Teckentrup, 2018; Xi et al., 2018; Briol et al., 2019; Teckentrup, 2020; Tuo and Wang, 2020). Smoothness misspecification has previously been considered in this context (Narcowich et al., 2006; Teckentrup, 2020; Kanagawa et al., 2020) as has corrupted observations (Rieger and Zwicknagl, 2009; Arcangéli et al., 2007). However, the interplay of smoothness and likelihood misspecification has not been investigated to date. Our paper therefore unifies and extends existing work in this area.

The main results in this paper are Theorem 1, Theorem 2, Theorem 4 and Theorem 7 which, respectively, concern the cases when a likelihood reflecting no noise is correctly assumed, a Gaussian likelihood is correctly assumed, a Gaussian likelihood is incorrectly assumed and a likelihood of no noise is assumed but there is arbitrary corruption. In each case the results also facilitate the smoothness of the target function being different from the smoothness of the approximating function. To highlight the relevance of these novel bounds, in Section 5 we derive implications for the convergence of Bayesian numerical methods based on GPs, specifically Bayesian quadrature and Bayesian optimization.

The paper is structured as follows. Section 2 reviews background material on GPs and reproducing kernel Hilbert spaces. Section 3 introduces and discusses assumptions on the design region, design points and GP model required for our theory to hold. Existing convergence results are also covered. Section 4 contains the error bounds. Section 5 demonstrates implications of these bounds for Bayesian quadrature and Bayesian optimization. Section 6 provides concluding remarks.

2. Background on Gaussian Processes and Kernel Methods

In this section, we start by introducing notation for GPs conditioned on data and recall some of their properties, then we highlight how the smoothness of GPs can be measured using Sobolev spaces.

2.1 Interpolation and Regression

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{X} \subseteq \mathbb{R}^d$. A Gaussian process (Stein, 1999; Rasmussen and Williams, 2006) is a stochastic process $g : \mathcal{X} \times \Omega \to \mathbb{R}$ whose properties are captured by its *mean* $m : \mathcal{X} \to \mathbb{R}$, $m(x) = \mathbb{E}[g(x, \cdot)]$, and *covariance function* $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $k(x, x') = \mathbb{E}[(g(x, \cdot) - m(x))(g(x', \cdot) - m(x'))]$. The defining property of a GP with mean m and covariance k, denoted $g \sim \mathcal{GP}(m, k)$, is that for any finite set of points $X = \{x_i\}_{i=1}^n$, the random vector $(g(x_1, \cdot), \ldots, g(x_n, \cdot))^\top \in \mathbb{R}^n$ follows the multivariate normal distribution $\mathcal{N}(m_X, k_{XX})$ with mean vector given by $m_X = (m(x_1), \ldots, m(x_n))^\top \in \mathbb{R}^n$ and covariance matrix $k_{XX} = (k(x_i, x_j))_{1 \leq i,j \leq n} \in \mathbb{R}^{n \times n}$.

The covariance function is symmetric ($k(x, x') = k(x', x) \forall x, x' \in \mathcal{X}$) and positive definite ($\forall n \in \mathbb{N}$, $a_1, \ldots, a_n \in \mathbb{R}$, $\{x_i\}_{i=1}^n \subset \mathcal{X}$, $\sum_{i,j=1}^n a_i a_j k(x_i, x_j) \geq 0$) and we shall call any function satisfying these two properties a *kernel*. A GP induces a probability measure over functions which we denote Π_k. A significant advantage of GPs over other stochastic processes is our ability to condition on data in closed form in some settings. Let $f_{\text{GP}} \sim \mathcal{GP}(m, k)$, $X = (x_1, \ldots, x_n)^\top$ be a finite collection of design points and for some deterministic function f denote by $f_X = (f(x_1), \ldots, f(x_n))^\top$ the corresponding function values. Conditioning the

stochastic process f_{GP} on noisy function evaluations, often called the *regression setting*, observed with independent, identically distributed Gaussian noise ε_i with mean zero, variance σ^2, gives another GP, denoted $f_{\text{GP}} \mid X, y \sim \mathcal{GP}(\bar{m}_{\sigma^2}, \bar{k}_{\sigma^2})$, where $y_i = f(x_i) + \varepsilon_i$, $\bar{m}_{\sigma^2}(x) = m(x) + k_{xX}(k_{XX} + \sigma^2 I_{n\times n})^{-1}(y - m_X)$, $\bar{k}_{\sigma^2}(x, x') = k(x, x') - k_{xX}(k_{XX} + \sigma^2 I_{n\times n})^{-1}k_{Xx'}$, with $k_{xX} = (k(x, x_1), \ldots, k(x, x_n))$ and $I_{n\times n}$ is an identity matrix of size n. This will also be the case if f_X is observed without noise, also called the *interpolation setting*, in which case the conditioned GP is denoted $f_{\text{GP}} \mid X, f_X \sim \mathcal{GP}(\bar{m}, \bar{k})$ where $\bar{m}(x) = k_{xX}k_{XX}^{-1}(f_X - m_X)$ and $\bar{k} = \bar{k}_0$.

Although the expressions for $\bar{m}$ and $\bar{m}_{\sigma^2}$ were obtained through conditioning of a GP, they can also arise through non-probabilistic function approximation schemes. The function spaces used are the reproducing kernel Hilbert spaces (RKHS) (Berlinet and Thomas-Agnan, 2004) associated with the kernel k of the GP. A Hilbert space of functions on $\mathcal{X}$, denoted $\mathcal{H}(\mathcal{X})$, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}(\mathcal{X})}$ and norm $\| \cdot \|_{\mathcal{H}(\mathcal{X})}$ is called a reproducing kernel Hilbert space if there exists a kernel k, such that the following two conditions are satisfied (i) $\forall x \in \mathcal{X}$ we have $k(\cdot, x) \in \mathcal{H}(\mathcal{X})$, and (ii) $\forall x \in \mathcal{X}$ and $\forall f \in \mathcal{H}(\mathcal{X})$, we have $\langle f, k(\cdot, x) \rangle_{\mathcal{H}(\mathcal{X})} = f(x)$ which is called the reproducing property. By the Moore-Aronszajn theorem, the relationship between kernels and RKHS is one-to-one, so we denote the RKHS by $\mathcal{H}_k(\mathcal{X})$ instead of $\mathcal{H}(\mathcal{X})$.

The optimisation problem for the interpolation setting is the following constrained problem

$$\underset{g \in \mathcal{H}_k(\mathcal{X})}{\arg\min} \|g\|^2_{\mathcal{H}_k(\mathcal{X})} \text{ such that } g(x_i) = f(x_i)\ \forall i \in \{1, \ldots, n\}.$$

The optimisation problem corresponding to regression is similar but does not require the approximating function to be exactly equal to observed data at the observation points

$$\underset{g \in \mathcal{H}_k(\mathcal{X})}{\arg\min}\, S(g, \lambda_n, \mathcal{X}) = \underset{g \in \mathcal{H}_k(\mathcal{X})}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (g(x_i) - y_i)^2 + \lambda_n \|g\|^2_{\mathcal{H}_k(\mathcal{X})}.$$

The fit at X and the complexity of the approximating function are traded off using a regularisation parameter $\lambda_n > 0$. When $\varepsilon_i = 0\ \forall i \in \{1, \ldots, n\}$, kernel regression is sometimes referred to as approximate kernel interpolation (Wendland and Rieger, 2005) due to the fact that it differs from kernel interpolation as $\lambda_n > 0$. For further discussion regarding the relationship between kernel methods of approximating functions and GP methods, see e.g. (Berlinet and Thomas-Agnan, 2004; Scheuerer et al., 2013; Kanagawa et al., 2018).

To unify notation, given a function m, a vector $\varepsilon \in \mathbb{R}^n$ and $\lambda > 0$, define the function

$$R^m_{f,\lambda,\varepsilon}(x) \coloneqq m(x) + k_{xX}(k_{XX} + \lambda I_{n\times n})^{-1}(f_X + \varepsilon - m_X), \tag{1}$$

then $\bar{m} = R^m_{f,0,0}$ and $\bar{m}_{\sigma^2} = R^m_{f,\sigma^2,\varepsilon}$ and the functions solving the kernel interpolation and regression problems are $R^0_{f,0,0}, R^0_{f,n\lambda_n,\varepsilon}$ respectively and for ease of notation we will drop the variables which are zero throughout the rest of the paper.

2.2 The Smoothness of Reproducing kernel Hilbert Spaces

As previously mentioned, we measure the smoothness of functions using Sobolev spaces, and this smoothness will control approximation rates. For $\tau \in \mathbb{N}, q \in [1, \infty]$ and a domain

$\mathcal{X} \subseteq \mathbb{R}^d$, meaning a non-empty, open, connected set, define the integer order Sobolev space $W_q^\tau(\mathcal{X})$

$$W_q^\tau(\mathcal{X}) = \left\{ f \in L^q(\mathcal{X}) \colon \forall \alpha \in \mathbb{N}^d \, |\alpha| \leq \tau, D^\alpha f \in L^q(\mathcal{X}) \right\},$$

where $\mathbb{N}^d$ is the set of multi-indices of size d, $|\alpha| = \sum_{i=1}^d \alpha_i$ and D^α is the weak derivative operator corresponding to α, see e.g. Arcangéli et al. (2012). Sobolev spaces can also be defined for $\tau \notin \mathbb{N}$ through a standard interpolation space argument (Arcangéli et al., 2012). In particular for $\tau > d/2$, the Sobolev space $W_2^\tau(\mathbb{R}^d)$ may be written as

$$W_2^\tau(\mathbb{R}^d) \coloneqq \left\{ f \in L^2(\mathbb{R}^d) \colon \|f\|^2_{W_2^\tau(\mathbb{R}^d)} \coloneqq \int_{\mathbb{R}^d} \left(1 + \|x\|_2^2\right)^\tau |\hat{f}(x)|^2 dx < \infty \right\},$$

where $\hat{f}$ is the Fourier transform of f and $\|\cdot\|_2$ denotes the Euclidean norm. Our theoretical results shall apply to functions defined over $\mathcal{X} \subseteq \mathbb{R}^d$, recall the definition of $W_2^\tau(\mathcal{X})$ via restriction

$$W_2^\tau(\mathcal{X}) \coloneqq \left\{ f \colon \mathcal{X} \to \mathbb{R}^d \colon \exists f^\circ \in W_2^\tau(\mathbb{R}^d) \text{ such that } f^\circ(x) = f(x) \; \forall x \in \mathcal{X} \right\},$$

with norm

$$\|f\|_{W_2^\tau(\mathcal{X})} = \inf \left\{ \|f^\circ\|_{W_2^\tau(\mathbb{R}^d)} \colon f^\circ \in W_2^\tau(\mathbb{R}^d) \text{ and } f^\circ(x) = f(x) \; \forall x \in \mathcal{X} \right\}.$$

Similarly, starting from $\mathcal{H}_k(\mathbb{R}^d)$, we may define $\mathcal{H}_k(\mathcal{X})$ via restriction. This function space is still an RKHS with the kernel being the restriction of k to $\mathcal{X} \times \mathcal{X}$ (Berlinet and Thomas-Agnan, 2004, Theorem 6). If $\mathcal{H}_k(\mathbb{R}^d)$ is norm equivalent to $W_2^\tau(\mathbb{R}^d)$ and $\mathcal{X}$ is regular in some sense to be outlined in Section 3, then $\mathcal{H}_k(\mathcal{X})$ is norm equivalent to $W_2^\tau(\mathcal{X})$. We call a kernel τ-*smooth* if $\mathcal{H}_k(\mathcal{X})$ is norm equivalent to $W_2^\tau(\mathcal{X})$.

A frequently used example of $\tau-$smooth kernel is the Matérn kernel. For $\tau > d/2$, it is given by

$$k_{\text{Mat}}(x, x') = \frac{2^{1-(\tau-\frac{d}{2})} A}{\Gamma(\tau - \frac{d}{2})} \left(\sqrt{2\left(\tau - \frac{d}{2}\right)} \frac{\|x - x'\|_2}{l} \right)^{\tau - \frac{d}{2}} K_{\tau - \frac{d}{2}} \left(\sqrt{2\left(\tau - \frac{d}{2}\right)} \frac{\|x - x'\|_2}{l} \right), \quad (2)$$

where $l > 0, A > 0$. Here, Γ is the Gamma function and $K_{\tau - d/2}$ is the modified Bessel function of second kind of order $\tau - d/2$. The parameter l is called the lengthscale, A is the amplitude. If $\tau = m + 1/2 + d/2$ for some $m \in \mathbb{N}$ then the expression drastically simplifies thanks to properties of Bessel functions (Kanagawa et al., 2018). Another kernel which has RKHS norm equivalent to a Sobolev space is the Wendland kernel (Wendland, 2005, Chapter 9). This kernel is popular in the SDA literature due to the fact that it is compactly supported and thus offers favourable computational advantages. Both these kernels are translation invariant meaning there exists a function ϕ such that $k(x, y) = \phi(x - y)$.

3. Experimental Setting

We now highlight assumptions on the experimental setting for which our theoretical results hold. Section 3.1 outlines properties of the domain over which the approximation occurs and of the points at which the target function is evaluated, Section 3.2 outlines properties of the GP model and Section 3.3 compares our assumptions to those in related literature.

3.1 The Experimental Design

Throughout this paper, we will follow Arcangéli et al. (2012) and assume the domain $\mathcal{X}$ is bounded and satisfies an (R, δ) interior cone condition with a Lipschitz boundary. Such domains will be called $\mathcal{L}(R, \delta)$-domains, see Section A in the Appendix for full details. This is a standard assumption to make when applying scattered data approximation type results (Kanagawa et al., 2020; Arcangéli et al., 2012; Teckentrup, 2020; Narcowich et al., 2006). As discussed by Stein (1970), any open bounded convex set in $\mathbb{R}^d$ has Lipschitz boundary. This includes for example any open hypercube $(0, 1)^d$ and indeed any hyper cuboid. An example of a non-Lipschitz boundary is a domain of two polygons with boundaries touching at only one point.

The experimental design problem is well studied for GP surrogate models (Sacks et al., 1989; Santner et al., 2018) and an intuitive requirement is that the point set X somehow covers the whole domain $\mathcal{X}$. Designs based on this rule-of-thumb are usually referred to as space-filling designs, see the review by Pronzato and Müller (2012).

Given a bounded set $\mathcal{X} \subseteq \mathbb{R}^d$ and a collection of points $X \subseteq \mathcal{X}$, the *fill distance* h_X, *separation radius* q_X and *mesh ratio* ρ_X are defined as

$$h_X \coloneqq \sup_{x \in \mathcal{X}} \inf_{y \in X} \|x - y\|_2, \qquad q_X \coloneqq \min_{\substack{x,y \in X \\ x \neq y}} \frac{1}{2}\|x - y\|_2, \qquad \rho_X = \frac{h_X}{q_X}.$$

A small fill distance guarantees that no point in the domain $\mathcal{X}$ is too far away from a point in the design X, while a large separation radius guarantees that points in the design X are not too close to one another and the mesh-ratio measures the uniformity of the points. All of our bounds will be expressed in terms of these quantities. A sequence of points sets $\{X_n\}_{n \in \mathbb{N}}$ is said to be *quasi-uniform*, if $\exists\, C > 0$ such that $C q_{X_n} \geq h_{X_n}\ \forall n \in \mathbb{N}$. Note that quasi-uniformity is equivalent to a bounded mesh-ratio ρ_{Xn}. Quasi-uniform points achieve optimal rates for the fill distance on $\mathcal{L}(R, \delta)$-domains, namely Müller (2008, Satz 2.1.7) showed that $\exists C_1, C_2 > 0$ such that $C_1 n^{-1/d} \leq h_{X_n} \leq C_2 n^{-1/d}\ \forall n \in \mathbb{N}$. We now provide several examples of point sets for which results on the fill distance or separation radius are available

- Regular grid points in a hypercube $\mathcal{X} = (0, 1)^d$ form a quasi-uniform point set (Johnson et al., 1990).

- Random points sampled according to some probability measure on $\mathcal{X}$ can be shown to decrease the fill distance at a near-optimal rate in expectation. Indeed, Oates et al. (2019) showed that on a $\mathcal{L}(R, \delta)$-domain, for any $\epsilon > 0$, $\mathbb{E}[h_{X_n}] = \mathcal{O}(n^{-1/d+\epsilon})$ whenever the density $p > 0$ on all of $\mathcal{X}$.

- Points chosen in a restricted greedy fashion to minimise the GP posterior variance for a τ-smooth kernel with $\tau > d/2 + 1$ result in quasi-uniform points (Wenzel et al., 2019).

- Another possible choice are quasi-Monte Carlo (QMC) point sets. Since quasi-uniformity as defined above is not studied in QMC, it is unclear when common QMC point sets are quasi-uniform. However, several special cases are known, see e.g. (Breger et al., 2018) for quasi-uniform QMC point sets on compact Riemmanian manifolds.

- Some design schemes aim to minimise energy functionals. For the case of the Riesz energy, Hardin et al. (2012) showed that minimum energy point sets on compact metric spaces can be quasi-uniform.

- The seminal work of Johnson et al. (1990) termed points globally minimising h_X "minimax-distance designs", and points globally maximising q_X "maximin-distance designs".

There are several popular choices of point sets for which exact rates for h_{X_n} or q_{X_n} are unknown, but which minimise these quantities numerically. The bounds in our paper clearly motivate these designs. We now present several examples:

- Smolyak sparse grids, which originate from the partial differential equations literature, are also popular in the GP literature. It was shown by Teckentrup (2020, Theorem 3.9) that these points are marginally quasi-uniform when projected onto the coordinate axis, but these will not be quasi-uniform in general.

- Latin hypercube designs (LHDs) (McKay et al., 1979). Unfortunately, these are not necessarily quasi-uniform point sets. However, several authors have proposed what they call maximin and minimax LHDs (Morris and Mitchell, 1995; Joseph and Hung, 2008; Wang et al., 2018), which search the space of LHDs for a design optimising the fill distance or separation radius.

- Many designs are model-based, the point sets depend on properties of the GPs. Two popular examples include D-optimal designs, which aim to minimise the differential Shannon information, and G-optimal designs which are selected to minimise the maximum variance of the predicted values. It was shown by Johnson et al. (1990) that these choices are asymptotically equivalent to minimax or maximin design when taking a radial kernel with lengthscale going towards zero.

3.2 The Gaussian Process Model and Hyperparameter Selection

Let $m(\theta)$ and $k(\theta)$ denote the mean function and covariance kernel in the GP model parameterised by some $\theta \in \Theta \subset \mathbb{R}^{d_\Theta}$. In practice, it is common to learn hyperparameters as more data points are observed, and our convergence results will allow for such adaptivity. There exists a vast literature on parameter estimation for GPs; for an overview, see e.g. (Stein, 1999, Chapter 6), which includes a detailed discussion of Matérn kernels, or (Rasmussen and Williams, 2006, Chapter 5).

For the mean function $m(\theta)$, it is common to use a parametric model whose parameters are estimated using least-squares. Of course, other methods, such as empirical-risk minimisation and gradient-based optimisation could also be used. For the covariance function $k(\theta)$, parameters controlling lengthscales, amplitudes and smoothness need to be estimated. Common approaches include maximum marginal likelihood estimation, sometimes referred to as empirical Bayes, and cross-validation. In Bayesian settings, it is also common to provide a full prior on these hyperparameters and consider a predictive distribution taking into account uncertainty in the parameters.

Our bounds will be independent of the method used for parameter estimation, following the approach of Teckentrup (2020). The convergence rates will only depend on how the smallest and largest smoothness of the approximation function $R^m_{f,\lambda,\varepsilon}$ for $\theta \in \Theta$ and the corresponding norm-equivalence constants. For this reason, we will use the notation $R^m_{f,\lambda,\varepsilon}(\theta)$ to emphasise the dependence on the parameter values. If $k(\theta)$ is $\tau(\theta)$-smooth, then we denote the norm equivalence constants by

$$C_l(\theta)\|\cdot\|_{\mathcal{H}_{k(\theta)}(\mathcal{X})} \leq \|\cdot\|_{W_2^{\tau(\theta)}(\mathcal{X})} \leq \ C_u(\theta)\|\cdot\|_{\mathcal{H}_{k(\theta)}(\mathcal{X})}. \tag{3}$$

Assume that the parameter estimation method gives a sequence of hyperparameters $\{\theta_n\}_{n=1}^{\infty}$ so that once the n-th data point has been observed, the parameters θ_n are used. Following Teckentrup (2020) given $N \in \mathbb{N}$ define $\tau_k^- := \inf_{n\geq N} \tau(\theta_n)$, $\tau_k^+ := \sup_{n\geq N} \tau(\theta_n)$ and $C_N = \sup_{n\geq N} C_u(\theta_n)C_l(\theta_n)^{-1}$. This set of extreme values is denoted by $\Theta_N^* = \{\tau_k^+, \tau_k^-, C_N\}$. These quantities represent the extremes of the smoothness of the kernel and the ratio of norm equivalence constants after the N-th data point has been observed. Of course, these parameters are often selected as data is observed. As a result, to bound expressions for $n \geq N$, we need to ensure the parameter selection methods used does not result in extreme values regardless of the data observed. We note that in the context of Gaussian regression, the observation noise parameter σ could also be estimated from data, leading to a sequence of parameters $\{\sigma_n\}_{n=1}^{\infty}$. A common approach is to maximise the marginal likelihood.

3.3 Comparison to Related Literature

Now that we have discussed our experimental setting, we briefly remark on connections with related literature using kernel approximations. In our work, the target function is modelled as an unknown deterministic function, possibly corrupted by noise, with no assumption on the distribution of the design points. The error bounds shall be expressed in terms of the smoothness of the approximating function, the smoothness of the true function, and geometric properties of the design points.

The closest approach to the work in this paper can be found in the scattered data approximation (Wendland, 2005) literature. Indeed, our proofs harness multiple results from the field. The main difference is that we tackle the combination of corrupted data, misspecified smoothness and misspecified likelihood, whereas existing works have only covered these cases individually. Examples include approximate interpolation (Wendland and Rieger, 2005), deterministic corruption (Rieger and Zwicknagl, 2009) and random error satisfying a regularity condition (Arcangéli et al., 2012; Utreras, 1988). A framework for managing smoothness misspecification was presented by Narcowich et al. (2006) which uses

quasi-uniform point placement. Adapting hyperparameters with no observation corruption was studied by Teckentrup (2020).

Statistical learning theory (Steinwart and Christmann, 2008; Cucker and Zhou, 2007) takes the view of approximation as an optimisation problem in an RKHS, outlined in Section 2, with the target function specified by some joint probability distribution on the input and output spaces. A sampling distribution for the location of the data points is assumed, which is employed as the weight measure for the norm used to measure error of the approximation. This statistical assumption is the main difference with the SDA view, which we use, since in SDA the error bounds are expressed in terms of the experimental design. Additionally the remedy for smoothness misspecification in SLT is altering the parameters of the approximating function (Steinwart et al., 2009) as opposed to quasi-uniform points.

Nonparametric regression (Györfi et al., 2002; Wahba, 1990) is an approach to regression which assumes no parametric underlying form for the target function. Such techniques bare a lot of resemblance to SDA and statistical learning theory, and indeed have similar methods for obtaining approximating functions. Sometimes the unknown function is assumed to be a draw from a distribution over a space of functions, for example in Kriging (Matheron, 1963; Stein, 1999). This is clearly different from the SDA paradigm which assumes the quantity of interest is a fixed deterministic function. As done in statistical learning theory a sampling distribution of data locations is also often assumed. Within this nonparametric paradigm, an important subclass is Bayesian nonparametric regression (Ghosal and van der Vaart, 2017; Giné and Nickl, 2016). These take the Bayesian view of modelling by placing a prior measure on the unknown target function, and using a likelihood and Bayes' rule to obtain a posterior measure on the unknown quantity given observed data. Contraction of the entire posterior measure is studied which is stronger than contraction of the posterior mean function, the focus of Section 4. Again the assumption of a sampling distribution of the points and the method of dealing with smoothness misspecification makes this modelling paradigm distinct from the one considered in this paper.

4. Convergence Guarantees for Gaussian Process Means

We are now ready to present the main results of the paper. All of the proofs are provided in the appendix. We will use the following notation $x \wedge y = \min(x, y)$, $x \vee y = \max(x, y)$, $(x)_+ = \max(x, 0)$. $\lfloor x \rfloor$ denotes the integer part of x and $\lceil x \rceil$ the ceiling of x. The integrability parameter in the Sobolev norms will be $q \in [1, \infty]$. Following Arcangéli et al. (2012) define $\tau_0 \coloneqq \tau - d(1/2 - 1/q)_+$ and $\tau^* \coloneqq \tau_0$ if $\tau \in \mathbb{N}$ and either $2 < q < \infty$ and $\tau_0 \in \mathbb{N}$, or $q = 2$, else we will have $\tau^* \coloneqq \lceil \tau_0 \rceil - 1$. Finally, for $a, b > 0$, let $\tilde{a} = a - \lfloor a \rfloor$ and define: $\Lambda_{a,b} \coloneqq (b\tilde{a}(1 - \tilde{a}))^{1/b}$, if $\tilde{a} \in (0, 1)$ and $\Lambda_{a,b} \coloneqq 1$ if $\tilde{a} = 0$.

4.1 Convergence Guarantees for Interpolation

This section considers approximations with noiseless function evaluations observed at a finite collection of n points $X_n \subset \mathcal{X}$. We will assume that the likelihood is well specified in that the data is indeed noiseless. The interpolation setting is of particular interest since it leads to a closed-form approximation, and corresponds to the use of GPs for range of applications including to computer models (Kennedy and O'Hagan, 2001), Bayesian inverse problems (Teckentrup, 2020) and Bayesian numerical methods (Bull, 2011; Xi et al., 2018;

Briol et al., 2019; Chen et al., 2019). From a practical point of view, the result provide insights into point-picking strategies and hyperparameter selection for these applications. Before stating the first bound, we summarise all of the necessary assumptions which were mentioned in the previous section

Assumption 1 (Assumptions on the Domain) *$\mathcal{X}$ is an $\mathcal{L}(R, \delta)$-domain for some $R > 0$ and $\delta \in (0, \pi/2)$.*

Assumption 2 (Assumptions on the Kernel Parameters) *Given $N \in \mathbb{N}$, for $n \geq N$, $k(\theta_n)$ is $\tau(\theta_n)$-smooth and the elements of Θ_N^* are finite with $\tau_k^- > d/2$.*

Assumption 3 (Assumptions on the Kernel Smoothness Range) *Given $N \in \mathbb{N}$, the set $\{\tau(\theta_n)\}_{n \geq N}$ has finitely many values.*

Assumption 4 (Assumptions on the Target Function and Mean Function) *The target function satisfies $f \in W_2^{\tau_f}(\mathcal{X})$ for some $\tau_f > d/2$ and given $N \in \mathbb{N}$ the mean function satisfies $\sup_{n \geq N} \|m(\theta_n)\|_{W_2^{\tau_f}(\mathcal{X})} < \infty$.*

Assumption 1 ensures that the domain is sufficiently regular to use extension and embedding theorems. For a discussion about examples of domains satisfying the assumptions see Section A. Assumption 2 ensures that the RKHS of $k(\theta_n)$ is norm equivalent to a Sobolev space with smoothness $\tau(\theta_n)$ and that the parameters for the model are not so extreme as to result in arbitrarily smooth or arbitrarily rough functions. This assumption also concerns the ratio of the norm equivalence constants to ensure that their ratio is finite. For the case of k being a Matérn kernel, a sufficient condition was given by Teckentrup (2020, Lemma 3.4) which shows that $C_N \leq \sup_{n \geq N} \max(l_n, l_n^{-1})$ where l_n is the lengthscale when using parameter setting θ_n.

The N term facilitates a "burn-in" period for narrowing down the desired range of hyperparameters. Assumption 3 is required in order to provide a uniform bound over parameter values. The assumption is satisfied in the common scenario where cross validation is used for smoothness parameter selection where there is a finite candidate set of smoothness parameters. For example, the widely used Matérn kernel has a convenient closed form for $\tau = m + 1/2 + d/2$ for $m \in \mathbb{N}$, whereas other smoothness level require evaluations of Bessel functions which is computationally challenging. In practice, it is therefore very common to focus on $\{\tau(\theta_n)\}_{n \geq N} = \{m + 1/2 + d/2\}_{m \in M}$ for some finite set $M \subset \mathbb{N}$. We note that Assumption 3 is not required by Teckentrup (2020) since weaker sampling inequalities depending only on the integer part of the smoothness parameter were used in that paper. Assumption 4 ensures that the target function has a minimal level of regularity and that the parameterised mean function used in the prior GP is at least as smooth as the target function.

We are now ready to state our main result for GP interpolation. This will be split into two parts covering the well-specified ($\tau_f \geq \tau_k^+$) and misspecified ($\tau_f < \tau_k^+$) smoothness settings.

Theorem 1 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-4 hold. Let $q \in [1, \infty]$ and $s \in [0, (\tau_f \wedge \tau_k^-)^*]$. Then, $\exists C_0, h_0 > 0$ such that $\forall n \geq N$, $\forall X_n \subseteq \mathcal{X}$ with $h_{X_n} \leq h_0$, when $\tau_f \geq \tau_k^+$*

$$\left\|f - R_f^m(\theta_n)\right\|_{W_q^s(\mathcal{X})} \leq C h_{X_n}^{\tau_k^- - s - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n \geq N} \left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})} \right),$$

and when $\tau_f < \tau_k^+$

$$\left\|f - R_f^m(\theta_n)\right\|_{W_q^s(\mathcal{X})} \leq C h_{X_n}^{(\tau_f \wedge \tau_k^-) - s - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \rho_{X_n}^{(\tau_k^+ - \tau_f)} \left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n \geq N} \left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})} \right),$$

where $C = C_0 \Lambda_{s,q}$ with $C_0 = C_0(\mathcal{X}, d, \tau_f, q, \Theta_N^*)$ and $h_0 = h_0(R, \delta, d, \tau_f, \Theta_N^*)$.

This theorem is an extension of the result by Teckentrup (2020, Theorem 3.5) since it holds for a wider range of target functions f. In particular, it only requires $\tau_f > d/2$ rather than $\lfloor \tau_f \rfloor > d/2$, as such, alleviates the issues mentioned by Teckentrup (2020, Remark 3.7). The range of the smoothness parameter s in the norm is dictated by τ_f, the smoothness of the target function, and τ_k^-, the minimum smoothness of the approximating function. There is a large freedom in the norm choice, for example a bound for L^2 approximation can be recovered by setting $s = 0, q = 2$, and L^∞ is recovered with $s = 0, q = \infty$. We will see in Section 5 that this flexibility can be useful for applications.

The upper bound holds only when the data points provide a sufficient initial covering of $\mathcal{X}$, as measured via the h_0 term, see e.g (Arcangéli et al., 2012, Remark 3.2) for a discussion of h_0. The behaviour of the constant $\Lambda_{s,q}$ is discussed further by Arcangéli et al. (2012, Section 4.2). Aside from the exponent of h_{X_n}, $\Lambda_{s,q}$ is the only term on the right hand side that depends on s, therefore the same C_0 value can be used for different s values. We now highlight how the bound depends on model-specific choices.

- *Experimental design:* The terms h_{X_n} and ρ_{X_n} quantify the impact of the experimental design. A detailed discussion of these quantities was provided in Section 3.1. In general, the approximation error bound is always minimised by making h_{X_n} and ρ_{X_n} as small as possible. We recall that the optimal decay of h_{X_n} is $n^{-\frac{1}{d}}$ and the optimal case for ρ_{X_n} is when it is bounded by a constant independent of n. Both of these properties occur when quasi-uniform points are used, and this is therefore a reasonable criterion for point selection. When quasi-uniform points are used, the optimal error rate is obtained in terms of worst case complexity (Novak and Woźniakowski, 2008, Theorem 4.17).

- *Kernel smoothness:* The rate of convergence, as a function of $h_{X_n}, q_{X_n}, \rho_{X_n}$, is controlled by τ_f, τ_k^+ and τ_k^-. In general, the larger the value of τ_f, the faster the convergence rate can be. Two regimes are highlighted. When $\tau_f < \tau_k^+$, meaning smoothness is misspecified, then $(\tau_k^+ - \tau_f)$ penalises overestimation of τ_f by increasing the exponent of ρ_{X_n}. Therefore, if one believes they are in danger of over estimating the smoothness of the true function, then quasi-uniform points should be used. When $\tau_f \geq \tau_k^+$, we see τ_k^- penalises underestimation of τ_f by limiting the exponent of h_{X_n}.

- *Other kernel parameters:* The bound can also be helpful when it comes to understanding the impact of adapting hyperparameters which do not change the smoothness of the RKHS. For those, adaptively choosing the hyperparameters does not impact the rate of convergence in n, but only constants of the bound. Indeed, It can be seen in the proof that C_0 depends on the extremes of the norm equivalence constants.

4.2 Convergence Guarantees for Regression with Gaussian Likelihood

This section considers observations that are corrupted with independently and identically distributed Gaussian noise so the data is $y_i = f(x_i) + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Once again, the mean of the GP conditioned on this data is available in closed-form, and a well-specified likelihood is used. Three further assumptions are required.

Assumption 5 (Additional Assumptions on Kernel Parameters) *Given $N \in \mathbb{N}$ and $\tau_f > d/2$ for all $n \geq N$ we have $\tau(\theta_n) \in (d/2, \tau_f] \cup [\lceil \tau_f \rceil, \infty)$.*

Assumption 6 (Assumption on Small Ball Probabilities) *Given $N \in \mathbb{N}$, $\exists c, \alpha_N > 0$ such that $\Pi_{k(\theta_n)}\left(\|f\|_{L^\infty(\overline{\mathcal{X}})} \leq c\right) \leq \exp(-\alpha_N)$ $\forall n \geq N$.*

Assumption 7 (Additional Assumption on the Target Function) *Given $\tau > d/2$, f has an extension $f^\circ \in C^{\tau_f}(\mathbb{R}^d) \cap W_2^{\tau_f}(\mathbb{R}^d)$ where $C^{\tau_f}(\mathbb{R}^d)$ is the space of τ_f Hölder continuous functions.*

Assumption 5 restricts slightly the smoothness values that f can take. It is required due to the double use of a sampling inequality in our proof, see the proof of Theorem 4 for further explanation. This is not a very restrictive assumptions since the length of interval containing disallowed values is less than one. Assumption 6 involves the measure on functions induced by the GP with parameters θ and ensures the size of the GP samples cannot be uniformly small with arbitrarily high probability, since this would result in a somehow degenerate GP. This assumption is implicitly used by Li and Linde (1999, Theorem 1.2) which is a key auxiliary result for van der Vaart and van Zanten (2011, Theorem 1), which our proof follows closely. In the case where an amplitude parameter is used for the kernel (e.g. A for the Matérn kernel in Section 2), the assumption is satisfied if this parameter is bounded away from zero. This can be seen by using concentration inequalities for the supremum of a Gaussian process, see e.g. (Giné and Nickl, 2016, Chapter 2.4). It should be noted that the commonly used maximum likelihood procedure can result in A decaying to zero (Karvonen et al., 2020). Assumption 7 concerns the regularity of the target function. This is a requirement for Lemma 4 of van der Vaart and van Zanten (2011) which is an auxiliary result that is employed in our bound, see Appendix F for details.

We can now present our main theorem for GP regression, which is stated in expectation over the distribution of the noise. Again, we separate the well-specified and misspecified smoothness settings.

Theorem 2 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-7 hold. Let $q \in [1, \infty]$ and $s \in [0, (\tau_f \wedge \tau_k^-)^*]$. Then, $\exists C, h_0 > 0$ such that $\forall n \geq N$, $\forall X_n \subseteq \mathcal{X}$ with $h_{X_n} \leq h_0$, when*

$\tau_f \geq \tau_k^+$

$$\mathbb{E}\left[\left\|f - R^m_{f,\sigma^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}\right] \leq C h_{X_n}^{\frac{d}{\gamma}-s}\left[h_{X_n}^{\tau_k^- - \frac{d}{2}}\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n\geq N}\left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})}\right) + n^{\frac{1}{2}} h_{X_n}^{\tau_k^- - \frac{d}{2}} + n^{\frac{d}{4\tau_k^-}}\right],$$

and when $\tau_f < \tau_k^+$

$$\mathbb{E}\left[\left\|f - R^m_{f,\sigma^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}\right] \leq C h_{X_n}^{\frac{d}{\gamma}-s}\left[h_{X_n}^{(\tau_f\wedge\tau_k^-)-\frac{d}{2}}\rho_{X_n}^{(\tau_k^+-\tau_f)_+}\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n\geq N}\left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})}\right) + n^{\frac{1}{2}} h_{X_n}^{\tau_k^- - \frac{d}{2}} + n^{\left(\frac{1}{2}-\frac{\tau_f}{2\tau_k^+}\right)_+ \vee \left(\frac{d}{4\tau_k^-}\right)}\right],$$

where $C = C_0\Lambda_{s,q}$ *with* $C_0 = C_0\left(\mathcal{X}, d, \tau_f, q, \|f\|_{W_2^{\tau_f}(\mathcal{X})}, \sup_{n\geq N}\|m(\theta_n)\|_{W_2^{\tau_f}(\mathcal{X})}, \Theta_N^*\right)$, $h_0 = h_0\left(R, \delta, d, \tau_f, \Theta_N^*\right)$ *and* $\gamma = 2 \vee q$.

As far as we are aware this is the first combination of SDA and Bayesian nonparametrics techniques. The closest result that we know of is by Arcangéli et al. (2007, Theorem 8.1) but does not cover the present scenario due to that result having requirements on the noise not satisfied by Gaussian noise.

The bounds contain a sum of three terms. The first term gives a rate identical to the interpolation case, and the later two describe the impact of the Gaussian noise. These last two terms will usually decrease to zero at a slower rate in n, and again, one can notice a clear advantage of using quasi-uniform points. The dependence on the norms of f and m in C_0 arises from the use of the result by van der Vaart and van Zanten (2011, Lemma 4). This dependence is made explicit in the proof and occurs in a small-ball probability bound. Due to Assumption 5, there is a limitation in our theory for $d = 1$. Specifically, $\tau_f + d/2$ could be smaller than $\lceil\tau_f\rceil$ when $d = 1$ so Assumption 4 might not be satisfied. But in two dimensions and higher, Assumption 5 does not impose extra restrictions since then $\tau_f + d/2 \geq \lceil\tau_f\rceil$ so $\tau(\theta_n) = \tau_f + d/2$ is a permissible value.

We once again comment on the impact of model choice. The advice in terms of experimental design is once again to use quasi-uniform points. The main difference with the previous theorem is for the smoothness parameters of the kernel.

- *Kernel smoothness:* The equality $\tau_f + d/2 = \tau_k^+ = \tau_k^-$ optimises the bound when using quasi-uniform points. This corresponds to the sample paths of the GP matching the smoothness of the target function (Kanagawa et al., 2018; Lukić and Beder, 2001). This is a phenomenon that occurs in this setting due to the true and assumed likelihood both being Gaussian, which is in distinct contrast to the interpolation case where the bound is optimised when $\tau_f = \tau_k^+ = \tau_k^-$. This choice of smoothness parameter might seem unintuitive from the perspective of kernel ridge regression. However this can be rationalised by observing that the connection to kernel ridge regression can only be

made if the regularisation being kept constant and not altering with added data. An in depth discussion is provided by Kanagawa et al. (2018, Section 5.1).

Corollary 3 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-7 hold with $\tau_k^+ = \tau_k^- = \tau_f + d/2$. Let $q \in [1,2]$ and $s \in [0, \tau_f^*]$. Then, $\exists C, h_0 > 0$ such that $\forall n \geq N$, $\forall X_n \subseteq \mathcal{X}$ quasi-uniform with $h_{X_n} \leq h_0$*

$$\mathbb{E}\left[\left\|f - R^m_{f,\sigma^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}\right] \leq C n^{-\frac{\tau_f}{2\tau_f+d}+\frac{s}{d}},$$

where $C = C_0 \Lambda_{s,q}$ with $C_0 = C_0\left(\mathcal{X}, d, \tau_f, q, \|f\|_{W_2^{\tau_f}(\mathcal{X})}, \sup_{n\geq N}\|m(\theta_n)\|_{W_2^{\tau_f}(\mathcal{X})}, \Theta_N^*\right)$ and $h_0 = h_0\left(R, \delta, d, \tau_f, \Theta_N^*\right)$.

When $q = 2, s = 0$ this is the mini-max optimal rate for non-parametric regression within the Bayesian nonparametric paradigm, see e.g. (Tsybakov, 2009, Chapter 2) and references therein. A comparison can be made to a recent result in statistical learning theory (Fischer and Steinwart, 2020, Corollary 6) which has $s/(2\tau + d)$ instead of s/d meaning it is stronger than our result. However, as discussed in Section 3, the assumptions in the statistical learning paradigm is somewhat different to our setting as we do not consider a norm weighted by the point sampling distribution.

4.3 Convergence Guarantees with Misspecified Likelihoods

Now that we have presented our results for well-specified likelihoods, we extend these to the misspecified case. We recall that GP approximations based on interpolation or Gaussian likelihoods are often used due to their closed form expressions, but that these are often idealisations of the problem.

This section illustrates the impact of this idealisation on convergence. In each case, the bound allows for arbitrary corruption $y_i = f(x_i) + \varepsilon_i$ where $\{\varepsilon_i\}_{i=1}^n$ do not have to be i.i.d. nor Gaussian, and could even be deterministic. The corruption is manifested in the bounds only in a $\mathbb{E}[\|\varepsilon\|_2]$ term. The main point of this section is that quasi-uniform points are not only essential for smoothness misspecification, but can also be of significant help to counter likelihood misspecification. Due to the lack of assumptions on the type of corruption, our bounds should be interpreted as worst-case type bounds.

They are particularly suitable for misspecification models studied in the robust regression literature (Rousseeuw and Leroy, 1987; Huber and Ronchetti, 2009; Christmann and Steinwart, 2007) in particular Christmann and Steinwart (2007) studies the case of kernel ridge regression. For example, bias robustness corresponds to the setting where the i.i.d Gaussian assumption is satisfied up to a small number of corruptions, usually called outliers. This is common when the data are collected from physical or medical sensors as these tend to have faults after a certain period of time, see e.g. (Armstrong and Boufassa, 1988). It also occurs in many applications of Bayesian optimisation (Martinez-Cantin et al., 2018). If a fixed number of data points is contaminated by outliers, then $\|\varepsilon\|_2 = O(1)$. Alternatively, it could be that some proportion of the total number of data points is corrupted. For example, if $n^\alpha, \alpha \in (0,1)$ data points are corrupted, then $\|\varepsilon\|_2 = O(n^{\alpha/2})$, whereas if $\beta n, \beta \in (0,1]$ data points are corrupted, then $\|\varepsilon\|_2 = O(\sqrt{\beta n})$.

Another possible scenario is the (pessimistic) case of arbitrary random and unbounded noise, see e.g. (Stegle et al., 2008). In this case, assuming that $\mathbb{E}[\varepsilon_i] < \infty$ and $\mathbb{E}[\varepsilon_i^2] < \infty$ $\forall i$, we get that: $\mathbb{E}[\|\varepsilon\|_2] = O(\sqrt{n})$ regardless of the distribution of these corruptions or of any correlation. We note that the exact distribution of these norms have been derived for a range of distributions (Mathai and Provost, 1992). Of course, it should be possible to improve on these worst-case bounds by using stronger assumptions on the distributions of the noise terms. For example, the Laplace to Gaussian misspecification was previously studied by Kleijn and van der Vaart (2006). It would be interesting, but beyond the scope of this paper, to combine such results with scattered data approximation bounds to produce bounds in the same fashion as Theorem 2.

Finally, one setting where our bounds will not be of help is when the noise distribution has infinite first or second moment. In this case, $\mathbb{E}[\|\varepsilon\|_2] = \infty$ and the bounds will be vacuous. This will be the case for Cauchy noise, or for certain instances of student-t or Pareto noise.

4.3.1 Misspecified Gaussian Regression Likelihood

For the first result, the Gaussian likelihood $\mathcal{N}(0, \sigma_n^2)$ is implicitly assumed. This corresponds to considering $R^m_{f,\sigma_n^2,\varepsilon}$ with $\sigma_n > 0$ as the approximating function. The subscript in σ_n is kept since we might want to vary the parameter with n in order to improve the convergence rate. This is to be interpreted as a worst-case type result since no assumption is placed upon the corruption.

Theorem 4 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-5 hold. Let $q \in [1, \infty]$, $s \in [0, (\tau_f \wedge \tau_k^-)^*]$ and $\sigma_n > 0 \, \forall n \in \mathbb{N}$. Then, $\exists C, h_0 > 0$ such that $\forall n \geq N$, $\forall X_n \subseteq \mathcal{X}$ with $h_{X_n} \leq h_0$, when $\tau_f \geq \tau_k^+$*

$$\mathbb{E}\left[\left\|f - R^m_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}\right] \leq C h_{X_n}^{\frac{d}{\gamma}-s}\Bigg[\left(h_{X_n}^{\tau_k^- - \frac{d}{2}} + \sigma_n\right)\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n\geq N}\left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})}\right) + \left(h_{X_n}^{\tau_k^- - \frac{d}{2}}\sigma_n^{-1} + 1\right)\mathbb{E}[\|\varepsilon\|_2]\Bigg],$$

and when $\tau_f < \tau_k^+$

$$\begin{aligned}&\mathbb{E}\left[\left\|f - R^m_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}\right] \\ &\leq C h_{X_n}^{\frac{d}{\gamma}-s}\Bigg[\left(h_{X_n}^{(\tau_f\wedge\tau_k^-)-\frac{d}{2}}\rho_{X_n}^{(\tau_k^+-\tau_f)} + \sigma_n q_{X_n}^{-(\tau_k^+-\tau_f)}\right)\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n\geq N}\left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})}\right) \\ &\qquad + \left(h_{X_n}^{\tau_k^- - \frac{d}{2}}\sigma_n^{-1} + 1\right)\mathbb{E}[\|\varepsilon\|_2]\Bigg],\end{aligned}$$

where $C = C_0\Lambda_{s,q}$ with $C_0 = C_0(\mathcal{X}, d, q, \tau_f, \Theta_N^*)$, $h_0 = h_0(R, \delta, d, \tau_f, \Theta_N^*)$ and $\gamma = 2 \vee q$.

This generalizes the results by Wendland and Rieger (2005, Proposition 3.6) and Arcangéli et al. (2007, Theorem 7.1) to misspecified likelihood and smoothness. Assumptions 6 and 7, used in Theorem 2, are not required since the corruption is not assumed Gaussian.

The effect of the corruption is manifested solely in the $\mathbb{E}[\|\varepsilon\|_2]$ term and to conclude the right hand side converges to zero, the growth of $\mathbb{E}[\|\varepsilon\|_2]$ needs to be sufficiently bounded. The theorem leads us to a useful recommendation for σ_n in settings where the data is noiseless.

- *Adaptive Likelihood/Nugget:* As noted in Section 3, it is common to add a "nugget" term to kernel matrices in order to improve numerical stability. This corresponds to taking $\sigma_n > 0$, and larger values of σ_n are known to provide greater stability at the cost of a slower convergence rate. When there is no corruption this is referred to as approximate interpolation (Wendland and Rieger, 2005). Theorem 4 provides a way of choosing σ_n for this scenario. Setting $q = \infty$ and $\tau_k^- = \tau_k^+ = \tau_f$, meaning we are in the well specified smoothness case, the choice $\sigma_n \propto h_{X_n}^{(\tau_f \wedge \tau_k^-)+(\tau_k^+ - \tau_f)-d/2} = h_{X_n}^{\tau_f - d/2}$ optimises the bound. This coincides with the choice proven to be optimal for matrix conditioning by Wendland and Rieger (2005, Corollary 3.7).

Thinking of this suggestion from the point of view of adaptive likelihood may seem unnatural at first since the likelihood is normally a fixed object which is independent of data. However, this suggestion can also be viewed from a regularisation perspective as altering the penalisation in the optimisation problem S, see Section 2.

We now give two corollaries of Theorem 4 under different assumptions on τ_k^- and τ_k^+. In each case, these provide insights into model choices which optimise the bounds. First, consider $\tau_k^- = \tau_k^+ = \tau_f$, in which case the smoothness is correctly specified. The next result gives a bound when h_{X_n} has the optimal rate $n^{-1/d}$ and σ_n is kept constant.

Corollary 5 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-5 hold. Let $q \in [1, \infty]$, $s \in [0, \tau_f^*]$, $\tau_k^- = \tau_k^+ = \tau_f$ and $\sigma_n = \sigma$. Assume $h_{X_n} \leq C_1 n^{-\frac{1}{d}}$ for some $C_1 > 0$. Then, $\exists C, h_0 > 0$ such that $\forall n \geq N$ with $h_{X_n} \leq h_0$*

$$\mathbb{E}\left[\left\|f - R^m_{f,\sigma^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}\right] \leq Cn^{-\frac{1}{\gamma}+\frac{s}{d}}\left(\mathbb{E}[\|\varepsilon\|_2] + n^{-\frac{\tau_f}{d}+\frac{1}{2}}\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n\geq N}\left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})}\right)\right),$$

where $C = C_0\Lambda_{s,q}$ with $C_0 = C_0\left(\mathcal{X}, d, q, \tau_f, \Theta_N^*\right)$, $h_0 = h_0\left(R, \delta, d, \tau_f, \Theta_N^*\right)$ and $\gamma = 2 \vee q$.

We note that when the smoothness is well specified, the value of τ_f does not impact the decay rate of the right hand side since the rate will be slowed down by the $h_{X_n}^{d/\gamma - s}$ term which does not depend on τ_f. This differs significantly from the noiseless case in Section 4.1 where a large value of τ_f led to faster convergence rates, and demonstrates how a small amount of noise can significantly impact the convergence rate.

Well-specified smoothness is a strong requirement. For the second corollary, we instead consider $\tau_k^+ = \tau_k^- = \tau$ for some $\tau \in \mathbb{R}$, but not necessarily $\tau_f = \tau$, when using quasi-uniform points and varying σ_n according to the fill distance. Surprisingly, this is enough to obtain the same bound as Corollary 5.

Corollary 6 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-5 hold. Let $q \in [1, \infty]$, $s \in [0, (\tau_f \wedge \tau)^*]$, and $\tau_k^+ = \tau_k^- = \tau$ and $\sigma_n = O(h_{X_n}^{\tau - d/2})$. Then, $\exists C, h_0 > 0$ such that $\forall n \geq N$, $\forall X_n \subseteq \mathcal{X}$ quasi-uniform with $h_{X_n} \leq h_0$*

$$\mathbb{E}\left[\left\|f - R^m_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}\right] \leq Cn^{-\frac{1}{\gamma}+\frac{s}{d}}\left(\mathbb{E}[\|\varepsilon\|_2] + n^{-\frac{(\tau_f\wedge\tau)}{d}+\frac{1}{2}}\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n\geq N}\left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})}\right)\right),$$

where $C = C_0\Lambda_{s,q}$ with $C_0 = C_0\left(\mathcal{X}, d, q, \tau_f, \Theta_N^*\right)$, $h_0 = h_0\left(R, \delta, d, \tau_f, \Theta_N^*\right)$ and $\gamma = 2 \vee q$.

On top of using quasi-uniform points, this corollary suggests the following practical approach.

- *Adaptive likelihood/Nugget*: When practitioners suspect that their likelihood might be misspecified, a sensible choice of nugget is $\sigma_n \propto h_{X_n}^{\tau - d/2}$. Interestingly, this is the same suggestion as for the case of Gaussian regression for noiseless data, which suggests that this choice may be sensible more broadly.

4.3.2 Misspecified Interpolation Likelihood

Our last main result considers arbitrary corruption when an interpolant is used, which is equivalent to taking $\sigma_n = 0$.

Theorem 7 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-5 hold. Let $q \in [1, \infty]$ and $s \in [0, (\tau_f \wedge \tau_k^-)^*]$. Then, $\exists C, h_0 > 0$ such that $\forall n \geq N$, $\forall X_n \subseteq \mathcal{X}$ with $h_{X_n} \leq h_0$, when $\tau_f \geq \tau_k^+$*

$$\mathbb{E}\left[\left\|f - R^m_{f,0,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}\right] \leq Ch_{X_n}^{\frac{d}{\gamma}-s}\left[h_{X_n}^{\tau_k^- - \frac{d}{2}}\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n\geq N}\left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})}\right) + \rho_{X_n}^{(\tau_k^+ - \frac{d}{2})}\mathbb{E}[\|\varepsilon\|_2]\right],$$

and when $\tau_f < \tau_k^+$

$$\mathbb{E}\left[\left\|f - R^m_{f,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}\right] \leq Ch_{X_n}^{\frac{d}{\gamma}-s}\rho_{X_n}^{(\tau_k^+ - \tau_f)}\left[h_{X_n}^{(\tau_f\wedge\tau_k^-) - \frac{d}{2}}\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n\geq N}\left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})}\right) + \rho_{X_n}^{(\tau_f - \frac{d}{2})}\mathbb{E}[\|\varepsilon\|_2]\right],$$

where $C = C_0\Lambda_{s,q}$ with $C_0 = C_0\left(\mathcal{X}, d, q, \tau_f, \Theta_N^*\right)$, $h_0 = h_0\left(R, \delta, d, \tau_f, \Theta_N^*\right)$ and $\gamma = 2 \vee q$.

If $\varepsilon = 0$, then there is no noise and we recover the well-specified likelihood result for interpolation from Theorem 1. We now study the impact of model choice.

- *Experimental design:* In this bound, there is a ρ_{X_n} term multiplied by $\mathbb{E}[\|\varepsilon\|_2]$ whose exponent does not vanish when the smoothness is well specified. This is in contrast

to Theorem 4 where the exponent of the ρ_{X_n} term vanished when the smoothness was well specified, and ρ_{X_n} did not interact with $\mathbb{E}[\|\varepsilon\|_2]$. This can be interpreted as $R^m_{f,\varepsilon}$ being less stable than $R^m_{f,\sigma_n^2,\varepsilon}$ with respect to noise and point placement, and suggests that the use of quasi-uniform point is strongly recommended, even when the smoothness is well-specified.

The following corollary shows the same bound as Corollary 5 and Corollary 6 can be obtained without the assumption of fixed kernel smoothness as long as quasi-uniform points are used.

Corollary 8 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-5 hold. Let $q \in [1, \infty]$ and $s \in [0, (\tau_f \wedge \tau_k^-)^*]$. Then, $\exists C, h_0 > 0$ such that $\forall n \geq N$, $\forall X_n \subseteq \mathcal{X}$ quasi-uniform with $h_{X_n} \leq h_0$*

$$\mathbb{E}\left[\left\|f - R^m_{f,0,\varepsilon}(\theta_n)\right\|_{W^s_q(\mathcal{X})}\right] \leq Cn^{-\frac{1}{\gamma}+\frac{s}{d}}\left(\mathbb{E}[\|\varepsilon\|_2] + n^{-\frac{(\tau_f\wedge\tau_k^-)}{d}+\frac{1}{2}}\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n\geq N}\left\|m(\theta_n)\right\|_{W_2^{\tau}(\mathcal{X})}\right)\right),$$

where $C = C_0\Lambda_{s,q}$ with $C_0 = C_0\left(\mathcal{X}, d, q, \tau_f, \Theta_N^*\right)$, $h_0 = h_0\left(R, \delta, d, \tau_f, \Theta_N^*\right)$ and $\gamma = 2 \vee q$.

Compared to Corollary 5 the requirement of quasi-uniform points is stronger than just $h_{X_n} \leq Cn^{-\frac{1}{d}}$, but this allows us to weaken the assumptions on the smoothness of the kernel. Indeed, as opposed to Corollary 6, the kernel smoothness is allowed to alter with n. However, $\sigma_n = 0$ means the approximation is harder to compute due to the matrix inversion being less stable.

5. Implications for Bayesian Numerical Methods

We demonstrate the applicability of our theorems to Bayesian probabilistic numerical methods, specifically Bayesian quadrature and Bayesian optimisation. These methods use GP approximations to solve numerical tasks, and can therefore inherit some of the convergence guarantees presented in the previous section.

5.1 Bayesian Quadrature

In Bayesian quadrature (BQ), the goal is to approximate some integral $\int_{\mathcal{X}} f(x)p(x)dx$. To do so, a GP prior is placed on f. This is conditioned on function evaluations to obtain a posterior on f, which itself implies a Gaussian posterior on the value of the integral. The posterior mean on this integral is used as an estimate of the integral, see e.g. (Briol et al., 2019) and the accompanying discussion for an in-depth overview. The most up-to-date convergence guarantees are available from Kanagawa et al. (2020). These consider the problem of smoothness misspecification in the interpolation setting.

We now highlight how the results of this paper can refine theory for BQ, but also lead to results in settings with likelihoods which have not yet been considered. First we consider interpolation, the proof is a combination of Theorem 1 with $q = 1, s = 0$ and Hölder's inequality.

Theorem 9 *Fix $N \in \mathbb{N}$ suppose Assumptions 1-4 hold. Then $\exists C_0, h_0 > 0$ such that $\forall n \geq N$, $\forall X_n \subseteq \mathcal{X}$ with $h_{X_n} \leq h_0$ and $\forall p \in L^2(\mathcal{X})$*

$$\left| \int_{\mathcal{X}} f(x)p(x)dx - \int_{\mathcal{X}} R_f^m(\theta_n)(x)p(x)dx \right| \\ \leq C\|p\|_{L^2(\mathcal{X})} h_{X_n}^{(\tau_f \wedge \tau_k^-)} \rho_{X_n}^{(\tau_k^+ - \tau_f)_+} \left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sup_{n \geq N} \left\| m(\theta_n) \right\|_{W_2^{\tau_f}(\mathcal{X})} \right),$$

where $C = C(\mathcal{X}, d, \tau_f, \Theta_N^*)$, $h_0 = h_0(R, \delta, d, \tau_f, \Theta_N^*)$.

This result generalizes (Kanagawa et al., 2020, Theorem 3) by allowing a greater range of values for τ_k^-, τ_k^+ and τ_f. It also takes into account the adaptation of hyperparameters with n, which has not been considered in the literature. Next, we consider a correctly specified Gaussian likelihood. The proof is a combination of Corollary 3 with $q = 1, s = 0$ and Hölder's inequality.

Theorem 10 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-7 hold. Let $\tau_k^+ = \tau_k^- = \tau_f + d/2$. Then, $\exists C, h_0 > 0$ such that $\forall n \geq N$, $\forall X_n \subseteq \mathcal{X}$ quasi-uniform with $h_{X_n} \leq h_0$ and $\forall p \in L^2(\mathcal{X})$*

$$\mathbb{E}\left[\left| \int_{\mathcal{X}} f(x)p(x)dx - \int_{\mathcal{X}} R_{f,\sigma^2,\epsilon}^m(\theta_n)(x)p(x)dx \right| \right] \leq C\|p\|_{L^2(\mathcal{X})} n^{-\frac{\tau_f}{2\tau_f + d}},$$

where $C = C\left(\mathcal{X}, d, \tau_f, \|f\|_{W_2^{\tau_f}(\mathcal{X})}, \sup_{n \geq N} \|m(\theta_n)\|_{W_2^{\tau_f}(\mathcal{X})}, \Theta_N^*\right)$, $h_0 = h_0\left(R, \delta, d, \tau_f, \Theta_N^*\right)$.

This result provides the very first bound for BQ with a correctly specified Gaussian likelihood. This may be particularly useful for applications of BQ in inverse problems and computer models, where the integrand cannot be evaluated exactly.

The two results above are illustrations of bounds that can be obtained using the theory in our paper. However, it would be straightforward to obtain results in other settings, including misspecified smoothness or misspecified likelihoods, using the same proof technique with some of the other bounds in Section 4. All of the previous recommendation on model choice are also appropriate for BQ, with the exception of the experimental design, for which it is recommended to use quasi-uniform points which concentrate in areas where p is large.

5.2 Bayesian Optimization

In Bayesian optimization (BO), the goal is to maximise some unknown function. This is done using a GP surrogate, and points are usually chosen using an acquisition function which balances exploration and exploitation of the GP model given the observed data up to that iteration. Common examples include the Upper Confidence Bound and Expected Improvement acquisition functions (Shahriari et al., 2016). In the noiseless case, SDA results were employed by Bull (2011) and a modification was proposed to the standard expected improvement acquisition function, to ensure greater coverage of the domain[1].

1. It is important to note that the definition of "quasi-uniformity" by Bull (2011) is strictly weaker than the standard definition in SDA. It only requires $h_{X_n} \leq Cn^{-1/d}$, which is implied by standard definition used in this paper.

Existing theoretical work on Bayesian optimization that establishes convergence under various acquisition functions do not accommodate for misspecification of functions smoothness (Bull, 2011; Srinivas et al., 2010; Vazquez and Bect, 2010). This problem is addressed by Berkenkamp et al. (2019) using a hyperparameter alteration regime which enlarges the RKHS until the target function is contained in it. Motivated by the content of Theorem 1, we investigate a different approach to tackle smoothness misspecification, relying on a modification of existing acquisition functions to promote quasi-uniform points and then employing the proof technique by Bull (2011).

The γ-stabilized algorithm framework (Wenzel et al., 2019) facilitates such a modification. For any acquisition function $F\colon \mathcal{X} \to \mathbb{R}$, kernel k and $\gamma \in (0, 1]$, the $(n+1)$-th step consists of picking $x_{n+1} = \sup_{x \in \mathcal{X}_{n,\gamma}} F(x)$ where $\mathcal{X}_{n,\gamma} = \{x \in \mathcal{X}\colon P_n(x) \geq \gamma \|P_n\|_{L^\infty}\}$, $P_n(x) = \bar{k}(x,x)^{\frac{1}{2}}$ and $\bar{k}$ is the posterior variance after observing the first n points, see Section 2. Such point selection encourages exploration since it only allows points to be picked from areas of non-trivial variance. If k is translation invariant and τ-smooth with $\tau > d/2 + 1$ then the resulting point set is quasi-uniform (Wenzel et al., 2019, Theorem 14, Theorem 18). This is a modification to the standard BO procedure of picking x_{n+1} as the maximum of F over all of $\mathcal{X}$.

For $n \in \mathbb{N}$ and any acquisition function F define the (γ, F, n) strategy as picking x_1 arbitrarily, then points $\{x_i\}_{i=2}^{n-1}$ according to the γ-stabilized F, and x_n as the maximum of R_f, the kernel interpolant of f based on $\{x_i\}_{i=1}^{n-1}$. The next result gives a bound for the performance of this strategy.

Theorem 11 *Suppose Assumptions 1 & 4 hold, k is a τ-smooth translation invariant kernel with $\tau > d/2 + 1$. Then, $\exists n_0 \in \mathbb{N}$ such that if $n \geq n_0$, the (γ, F, n)-strategy satisfies:*

$$|\arg\max_{x \in \mathcal{X}} f(x) - f(x_n)| = C n^{-\frac{(\tau \wedge \tau_f)}{d} + \frac{1}{2}},$$

where $C = C(\mathcal{X}, d, \tau_f)$ and $n_0 = n_0(R, \delta, d, \tau_f)$.

In terms of worst-case error, in which the slowest rate is considered over the unit ball of the RKHS, this is the best possible rate given the smoothness of the target function and kernel, as shown by Bull (2011, Theorem 1). The 1/2 appears in our bound due to a different parameterisation of kernel smoothness than Bull (2011). However, Theorem 11 is more general than the result by Bull (2011, Theorem 1) since it applies to functions outside the RKHS of k. This is the first BO strategy which achieves the optimal rate in the case of smoothness misspecification.

Once again, we conclude by noting that this theorem is only an illustration of the implications of our results on convergence guarantees for GPs to the BO setting, and many other cases could be considered including likelihood misspecification.

6. Conclusions

In this paper, we have presented novel error bounds for GP means under misspecified likelihoods and smoothness, expressed in terms of observation error, point placement and choice of GP model. Our results apply under four different observation models. Where

the assumption of no noise is correct, where the assumption of no noise is incorrect, where the assumption of a Gaussian likelihood is correct and when the assumption of a Gaussian likelihood is incorrect. In each setting, our results demonstrate the impact of the choice of hyperparameters and the experimental design. As such, our results can guide practitioners who need to select a specific GP algorithm, by allowing them to tailor this choice to the application at hand.

The bounds offer improvements over existing results which we have highlighted. Applications to Bayesian numerical methods were presented such as the first error bounds for BQ with deterministic point selection and Gaussian observation noise and BO with misspecified smoothness. In both instances the use of point picking strategies which produce quasi-uniform points, as opposed to specific hyperparameter selection methods, are of critical importance.

We believe there are many more opportunities to combine GP and SDA methods. For example dealing with smoothness and likelihood misspecification when the approximating function is infinitely smooth, such as when a Gaussian kernel is used for approximation a common choice in practice. Additionally, analogies of the results in this paper for functions with vector valued output or structured output, such as additive functions, would be an important avenue of research and would follow naturally from the insights that SDA offers for such scenarios.

Acknowledgments

We would like to thank Toni Karvonen, Andrew Duncan, three anonymous reviewers and the editor for helpful comments when writing this paper. GW was supported by an EPSRC Industrial CASE award [EP/S513635/1] in partnership with Shell UK Ltd. FXB was supported by an Amazon Research Award on "Transfer Learning for Numerical Integration in Expensive Machine Learning Systems". MG was supported by the EPSRC grants [EP/T000414/1, EP/R018413/2, EP/P020720/2, EP/R034710/1, EP/R004889/1]. FXB and MG were also supported by the Lloyds Register Foundation Programme on Data-Centric Engineering and The Alan Turing Institute under the EPSRC grant [EP/N510129/1].

Appendix A. The Design Region

In this first appendix, we briefly recall common terminology from the literature on scattered data approximation which is used throughout the paper.

A domain shall mean an open connected set in $\mathbb{R}^d$. A domain satisfies the (R, δ) interior cone condition if for $R > 0$ and angle $\delta \in (0, \pi/2)$ we have that $\forall x \in \mathcal{X}$, $\exists\, \xi(x)$ such that the cone

$$C\left(x, \xi(x), \delta, R\right) = \left\{x + \lambda y \,:\, y \in \mathbb{R}^d, \|y\|_2 = 1, y^\top \xi(x) \geq \cos(\delta), \lambda \in [0, R]\right\},$$

is contained in $\mathcal{X}$. An open set $\mathcal{X}_i \subseteq \mathbb{R}^d$ is called a special Lipschitz domain (Stein, 1970, Page 181) if there exists a rotation of $\mathcal{X}_i$, denoted by $\tilde{\mathcal{X}}_i$, and a function $\psi : \mathbb{R}^{d-1} \to \mathbb{R}$ which satisfies the following

1. $\tilde{\mathcal{X}}_i = \{(x, y) \in \mathbb{R}^d \;\; y > \psi(x)\}$,

2. ψ is a Lipschitz function such that $|\psi(x) - \psi(x')| \leq M\|x - x'\|_2 \ \forall x, x' \in \mathbb{R}^{d-1}$ where $M > 0$.

Consider a domain $\mathcal{X} \subseteq \mathbb{R}^d$ and denote its boundary by $\partial\mathcal{X}$. We say $\partial\mathcal{X}$ is a Lipschitz boundary (Stein, 1970, Page 189) if $\exists\, \varepsilon > 0$, $N \in \mathbb{N}$, $M > 0$, and open sets $U_1, U_2, \ldots, U_L \subset \mathbb{R}^d$, where $L \in \mathbb{N} \cup \{\infty\}$, such that the following conditions are satisfied

1. For any $x \in \partial\mathcal{X}$, there exists an index i such that $B(x, \varepsilon) \subset U_i$,

2. $U_{i_1} \cap \cdots \cap U_{i_{N+1}} = \emptyset$ for any distinct indices $\{i_1, \ldots i_{N+1}\}$,

3. For each index i there exists a special Lipschitz domain $\mathcal{X}_i \subset \mathbb{R}^d$ with Lipschitz bound b such that $b \leq M$ and $U_i \cap \mathcal{X} = U_i \cap \mathcal{X}_i$.

and we call any bounded domain satisfying the (R, δ) interior cone condition with a Lipschitz boundary a $\mathcal{L}(R, \delta)$-domain.

Appendix B. Preliminary Results

This section covers results to be used throughout the rest of the proofs, namely a sampling inequality, restriction and extension of functions in RKHS and the Pythagorian property.

Sampling inequalities (Narcowich et al., 2006; Rieger et al., 2010; Arcangéli et al., 2012; Arcangéli and Torrens, 2014) are powerful inequalities for functions in Sobolev spaces which facilitate the systemisation of approximation error bounds. The result below is a special case of the result by Arcangéli et al. (2012, Theorem 3.2) where the integrability parameter in the right hand side Sobolev norm is set to two and so is the parameter p of the l_p norm.

Theorem 12 *Let $\mathcal{X}$ be a $\mathcal{L}(R, \delta)$-domain, $\tau > d/2$ and $q \in [1, \infty]$. Then, $\exists C, h_0 > 0$ such that $\forall X \subseteq \mathcal{X}$ with $h_X \leq h_0$, any $f \in W_2^\tau(\mathcal{X})$ and any $s \in [0, \tau^*]$*

$$\|f\|_{W_q^s(\mathcal{X})} \leq C\Lambda_{s,q} \left(h_X^{\tau - s - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \|f\|_{W_2^\tau(\mathcal{X})} + h_X^{\frac{d}{\gamma} - s} \|f_X\|_2 \right),$$

where $C = C(\mathcal{X}, d, \tau, q)$, $h_0 = h_0(\delta, R, d, \tau)$ and $\gamma = 2 \vee q$.

Discussion of how the domain, smoothness of the function and point set affect the constants is provided by Arcangéli et al. (2012). It is important to note that the dependence on τ in h_0 is only through $\lfloor \tau \rfloor$, this can be seen from inspection of the proof. The sampling inequality above is defined for norms over $\mathcal{X}$, but our proofs will be based on Fourier transforms which will be defined for functions over $\mathbb{R}^d$ therefore results facilitating the restriction and extension of functions between $\mathcal{X}$ and $\mathbb{R}^d$ are required. To this end the Sobolev extension theorem is required, stated below.

Theorem 13 *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a bounded Lipschitz domain, $\tau \geq 0$ and $p \in [1, \infty)$. There exists an extension map $\mathcal{E} \colon W_p^\tau(\mathcal{X}) \to W_p^\tau(\mathbb{R}^d)$ such that $\forall f \in W_p^\tau(\mathcal{X})$ we have $\mathcal{E}f|_{\mathcal{X}} = f|_{\mathcal{X}}$ and $\|\mathcal{E}f\|_{W_p^\tau(\mathbb{R}^d)} \leq C\|f\|_{W_p^\tau(\mathcal{X})}$ where $C = C(\mathcal{X}, d, \tau, p) > 0$ is a constant independent of f.*

The Sobolev extension theorem is used by Wendland (2005, Corollary 10.48) to ensure that, along with some assumptions on $\mathcal{X}$ satisfied by Assumption 1, if $\mathcal{H}_k(\mathbb{R}^d)$ is norm equivalent to $W_2^\tau(\mathbb{R}^d)$ then $\mathcal{H}_k(\mathcal{X})$ is norm equivalent to $W_2^\tau(\mathcal{X})$. Finally, the next two lemmas assure us that the minimal norm properties of the kernel interpolant and kernel regression function still hold along with the Pythagorean property for kernel interpolant. For a proof, see e.g. (Wendland, 2005, Corollary 10.25).

Lemma 14 *Let $\mathcal{X} \subseteq \mathbb{R}^d$, $X \subseteq \mathcal{X}$ a finite subset, k a kernel over $\mathbb{R}^d \times \mathbb{R}^d$ and $f \in \mathcal{H}_k(\mathcal{X})$ then*

$$R_f|_{\mathcal{X}} = \operatorname*{arg\,min}_{\substack{g \in \mathcal{H}_k(\mathcal{X}) \\ g_X = f_X}} \|g\|_{\mathcal{H}_k(\mathcal{X})} \qquad R_{f,n\lambda,\varepsilon}|_{\mathcal{X}} = \operatorname*{arg\,min}_{g \in \mathcal{H}_k(\mathcal{X})} S(g, \lambda, \mathcal{X}),$$

where $S(g, \lambda, \mathcal{X})$ is the regularized least squares problem defined in Section 2.

Proof The case of $\mathcal{X} = \mathbb{R}^d$ is obtained by standard arguments (Kanagawa et al., 2018, Theorem 3.4, Theorem 3.5) so we restrict to the case when $\mathcal{X}$ is a strict subset of $\mathbb{R}^d$. We shall only prove the first statement since the second proof is analogous. The interpolant restricted to $\mathcal{X}$ equals f on X since $X \subseteq \mathcal{X}$ and by definition $R_f|_{\mathcal{X}} \in \mathcal{H}_k(\mathcal{X})$, therefore

$$\left\|R_f|_{\mathcal{X}}\right\|_{\mathcal{H}_k(\mathcal{X})} \geq \min_{\substack{g \in \mathcal{H}_k(\mathcal{X}) \\ g|_X = f|_X}} \|g\|_{\mathcal{H}_k(\mathcal{X})}.$$

The rest of the proof will be done by contradiction. Suppose $\exists g \in \mathcal{H}_k(\mathcal{X})$ such that $g|_X = f|_X$ and $\|g\|_{\mathcal{H}_k(\mathcal{X})} < \left\|R_f|_{\mathcal{X}}\right\|_{\mathcal{H}_k(\mathcal{X})}$. Then, by definition of the norm on $\mathcal{H}_k(\mathcal{X})$

$$\|g\|_{\mathcal{H}_k(\mathcal{X})} = \inf_{\substack{h \in \mathcal{H}_k(\mathbb{R}^d) \\ h|_{\mathcal{X}} = g|_{\mathcal{X}}}} \|h\|_{\mathcal{H}_k(\mathbb{R}^d)} < \left\|R_f|_{\mathcal{X}}\right\|_{\mathcal{H}_k(\mathcal{X})} \leq \left\|R_f\right\|_{\mathcal{H}_k(\mathbb{R}^d)}.$$

By definition of the infimum, $\exists h \in \mathcal{H}_k$ such that $h|_{\mathcal{X}} = g|_{\mathcal{X}}$ and $\|h\|_{\mathcal{H}_k(\mathbb{R}^d)} < \|R_f\|_{\mathcal{H}_k(\mathbb{R}^d)}$. But $X \subseteq \mathcal{X}$ hence $h_X = g_X = f_X$ which contradicts norm minimality of R_f over $\mathbb{R}^d$. This completes the proof. ■

Lemma 15 *Let $\mathcal{X} \subseteq \mathbb{R}^d$, k a kernel over $\mathcal{X} \times \mathcal{X}$ and $f \in \mathcal{H}_k(\mathcal{X})$ then we have the Pythagorean property for the interpolant: $\left\|f - R_f\right\|^2_{\mathcal{H}_k(\mathcal{X})} + \left\|R_f\right\|^2_{\mathcal{H}_k(\mathcal{X})} = \|f\|^2_{\mathcal{H}_k(\mathcal{X})}$.*

Appendix C. Proof of Theorem 1

A key intermediate result is a slight generalisation of the sampling inequality by Narcowich et al. (2006, Theorem 4.2) which facilitates bounds for the misspecified smoothness scenario.

Theorem 16 *Suppose $\mathcal{X}$ is a $\mathcal{L}(R, \delta)$-domain and k is γ-smooth for $\gamma > d/2$. Then, $\exists C, h_0 > 0$ such that $\forall X \subseteq \mathcal{X}$ with $h_X \leq h_0$, we have $\forall f \in W_2^\tau(\mathcal{X})$ $\forall \mu \in [0, \tau]$*

$$\left\|f - R_f\right\|_{W_2^\mu(\mathcal{X})} \leq C h_X^{\tau-\mu} \rho_X^{\gamma-\tau} \|f\|_{W_2^\tau(\mathcal{X})},$$

where $C = C(\mathcal{X}, d, \tau)$, $h_0 = h_0(\delta, R, d, \tau)$.

Proof The proof is identical to the proof by Narcowich et al. (2006, Theorem 4.2), but with different assumptions on γ and $\mathcal{X}$. Specifically the proof by Narcowich et al. (2006, Theorem 4.2) uses the result by Narcowich et al. (2006, Lemma 4.1) for which a strictly smaller range of γ is permitted. However Theorem 12 generalises the older bound by Narcowich et al. (2006, Lemma 4.1) and simply requires $\gamma > d/2$. Additionally compactness of $\mathcal{X}$ was assumed by Narcowich et al. (2006, Theorem 4.2) to use a version of the Sobolev extension theorem but Theorem 13 can instead be used to obtain the same conclusion for $\mathcal{L}(R, \delta)$-domains. ■

We begin by expressing the error for the interpolant $R_f^m(\theta_n)$ in terms of two zero-mean GP interpolation problems

$$\begin{aligned}\left\|f - R_f^m(\theta_n)\right\|_{W_q^s(\mathcal{X})} &= \left\|f - R_f(\theta_n) - m(\theta_n) + R_{m(\theta_n)}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \\ &\leq \left\|f - R_f(\theta_n)\right\|_{W_q^s(\mathcal{X})} + \left\|m(\theta_n) - R_{m(\theta_n)}(\theta_n)\right\|_{W_q^s(\mathcal{X})}. \end{aligned} \tag{4}$$

The equality follows by the definition in (1) and the inequality is the triangle inequality. Therefore zero-mean GP interpolation problems only needs to be dealt with. An upper-bound on the first term naturally leads to an upper bound on the second since Assumption 4 imposes that $m(\theta_n)$ is at least as smooth as the target function. For $n \geq N$ and $s \in \left[0, (\tau_f \wedge \tau_N^-)^*\right]$, applying Theorem 12 to the function $f - R_f(\theta_n)$ over all smoothness levels $\{\tau_f \wedge \tau(\theta_n)\}_{n \geq N}$ yields

$$\left\|f - R_f(\theta_n)\right\|_{W_q^s(\mathcal{X})} \leq C_1 \Lambda_{s,q} h_{X_n}^{(\tau_f \wedge \tau(\theta_n)) - s - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \left\|f - R_f(\theta_n)\right\|_{W_2^{\tau_f \wedge \tau(\theta_n)}(\mathcal{X})}, \tag{5}$$

for $h_{X_n} \leq h_1$ where $C_1 = C_1(\mathcal{X}, d, \tau_f, q, \Theta_N^*)$ and $h_1 = h_1(R, \delta, d, \tau_f, \Theta_N^*)$ are respectively the supremum and infimum over $n \geq N$ of the constants obtained from applying Theorem 12 with smoothness parameter $\tau_f \wedge \tau(\theta_n)$. Due to Assumption 3, $\tau_f \wedge \tau(\theta)$ takes finitely many values so the infimum and supremum are over a finite number of values. This immediately gives $C_1 < \infty$ and $h_1 > 0$ and the same logic will be employed whenever Theorem 12 is used again. The residual terms are zero since $R_f(\theta_n)$ interpolates f at the observation points.

For the case $\tau_f \geq \tau(\theta_n)$, the target function f is in the RKHS of $k(\theta_n)$ so we can derive the following inequality

$$\begin{aligned}\left\|f - R_f(\theta_n)\right\|_{W_2^{\tau_f \wedge \tau(\theta_n)}(\mathcal{X})} &= \left\|f - R_f(\theta_n)\right\|_{W_2^{\tau(\theta_n)}(\mathcal{X})} \\ &\leq C_u(\theta_n) \|f - R_f(\theta_n)\|_{\mathcal{H}_{k(\theta_n)}(\mathcal{X})} &(6)\\ &\leq C_u(\theta_n) \|f\|_{\mathcal{H}_{k(\theta_n)}(\mathcal{X})} &(7)\\ &\leq C_u(\theta_n) C_l(\theta_n)^{-1} \|f\|_{W_2^{\tau(\theta_n)}(\mathcal{X})} &(8)\\ &\leq C_N \|f\|_{W_2^{\tau_f}(\mathcal{X})}. &(9)\end{aligned}$$

The inequalities in (6) and (8) follow from the norm equivalence between the RKHSs and Sobolev spaces with constants given in (3). The inequality in (7) is due to the Pythagorean property in Lemma 15, (9) is obtained by upper bounding by the largest constants over all

values of $\{\theta_n\}_{n\geq N}$, which can be done by Assumption 2, and the fact that the $\|\cdot\|_{W_2^{\tau_f}(\mathcal{X})}$ norm which is larger than the $\|\cdot\|_{W_2^{\tau(\theta_n)}(\mathcal{X})}$ norm since we are currently dealing with the case $\tau_f \geq \tau(\theta_n)$.

For the case $\tau(\theta_n) > \tau_f$, setting $\gamma = \tau(\theta_n)$ and $\mu = \tau_f$ in Theorem 16 gives

$$\left\|f - R_f(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})} \leq C_2 \rho_{X_n}^{\tau(\theta_n)-\tau_f} \|f\|_{W_2^{\tau_f}(\mathcal{X})}, \tag{10}$$

for $h_X \leq h_2$ where $C_2 = C_2(\mathcal{X}, d, \tau_f, \Theta_N^*)$ and $h_2 = h_2(R, \delta, d, \tau_f, \Theta_N^*)$. By the same reasoning as the discussion after (5) $h_2 > 0$ and $C_2 < \infty$. Now combine (5), (9) and (10)

$$\begin{aligned}
&\left\|f - R_f(\theta_n)\right\|_{W_q^s(\mathcal{X})} \\
&\quad\leq \begin{cases} C_1 C_N \Lambda_{s,q} h_X^{(\tau_f \wedge \tau(\theta_n))-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+} \|f\|_{W_2^{\tau_f}(\mathcal{X})} & \text{if } \tau_f < \tau(\theta_n) \\ C_1 C_2 \Lambda_{s,q} h_X^{(\tau_f \wedge \tau(\theta_n))-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+} \rho_X^{\tau(\theta_n)-\tau_f} \|f\|_{W_2^{\tau_f}(\mathcal{X})} & \text{if } \tau_f \geq \tau(\theta_n) \end{cases} \\
&\quad\leq C_3 \Lambda_{s,q} h_X^{(\tau_f \wedge \tau_k^-)-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+} \rho_X^{(\tau_k^+-\tau_f)_+} \|f\|_{W_2^{\tau_f}(\mathcal{X})},
\end{aligned} \tag{11}$$

where the inequality in (11) is obtained by taking the largest bound over parameter values $\{\theta_n\}_{n\geq N}$ and $C_3 = \max(C_1 C_N, C_1 C_2)$. To conclude the proof apply the upper bound in (11) to each term of (4) then set C_0 to be two times the maximum of the constants for each term and h_0 the minimum of the fill distance constants related to each term.

Appendix D. Proof of Theorem 4

To obtain a bound in the scenario of corrupted data, we cannot use Theorem 16 or Lemma 15 since they only apply to interpolants. Instead, Theorem 4 will follow from Theorem 18 and Lemma 17 along with the band-limited function techniques pioneered by Narcowich et al. (2006).

Lemma 17 *Let k be a kernel on $\mathcal{X} \times \mathcal{X}$, $f \in \mathcal{H}_k(\mathcal{X})$, $\sigma > 0$ and assume observations $y_i = f(x_i) + \varepsilon_i$ at $X = \{x_i\}_{i=1}^n$ for some $\varepsilon \in \mathbb{R}^n$ then*

$$\left\|R_{f,\sigma^2,\varepsilon}\right\|_{\mathcal{H}_k(\mathcal{X})} \leq \left(\sigma^{-2}\|\varepsilon\|_2^2 + \|f\|_{\mathcal{H}_k(\mathcal{X})}^2\right)^{\frac{1}{2}}$$

$$\left\|(f - R_{f,\sigma^2,\varepsilon})_X\right\|_2 \leq \|\varepsilon\|_2 + \left(\|\varepsilon\|_2^2 + \sigma^2\|f\|_{\mathcal{H}_k(\mathcal{X})}^2\right)^{\frac{1}{2}}.$$

Proof By triangle inequality

$$\left\|(f - R_{f,\sigma^2,\varepsilon})_X\right\|_2 = \left\|(y - \varepsilon - R_{f,\sigma^2,\varepsilon})_X\right\|_2 \leq \left\|(y - R_{f,\sigma^2,\varepsilon})_X\right\|_2 + \|\varepsilon\|_2.$$

Combining this with the inequality below completes the proof

$$\begin{aligned}
&\max\left(\left\|\left(y - R_{f,\sigma^2,\varepsilon}\right)_X\right\|_2^2,\ \sigma^2\left\|R_{f,\sigma^2,\varepsilon}\right\|_{\mathcal{H}_k(\mathcal{X})}^2\right) \\
&\qquad\leq nS(R_{f,\sigma^2,\varepsilon}, \sigma^2 n^{-1}, \mathcal{X}) \qquad (12) \\
&\qquad\leq nS(f, \sigma^2 n^{-1}, \mathcal{X}) = \|\varepsilon\|_2^2 + \sigma^2\|f\|_{\mathcal{H}_k(\mathcal{X})}^2. \qquad (13)
\end{aligned}$$

Where (12) uses the definition of the optimisation problem S, see Section 2, and (13) follows since $R_{f,\sigma^2,\varepsilon}$ solves the optimisation problem S. ■

Theorem 18 *Fix $N \in \mathbb{N}$ suppose Assumptions 1-5 hold and each observation is corrupted by some ε_i and let $q \in [1, \infty]$. Then, $\exists C, h_0 > 0$ such that $\forall n \geq N$, $\forall X_n \subseteq \mathcal{X}$ with $h_{X_n} \leq h_0$ and $\forall s \in [0, (\tau_f \wedge \tau_k^-)^*]$ the approximation error is bounded as*

$$\begin{aligned}
&\mathbb{E}\left[\left\|f - R^m_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}\right] \\
&\quad\leq\ C\Lambda_{s,q}\Bigg[h_{X_n}^{(\tau_f\wedge\tau_k^-)-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+}\rho_{X_n}^{(\tau_k^+-\tau_f)_+}\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \left\|m(\theta_n)\right\|_{W_2^{\tau_f}(\mathcal{X})}\right) \\
&\qquad\qquad + h_{X_n}^{\tau_k^- - s - d\left(\frac{1}{2}-\frac{1}{q}\right)_+}\sigma_n^{-1}\mathbb{E}[\|\varepsilon\|_2] \\
&\qquad\qquad + h_{X_n}^{\frac{d}{\gamma}-s}\mathbb{E}\left[\left\|(f - R_{f,\sigma_n^2,\varepsilon}(\theta_n))_{X_n}\right\|_2\right] \\
&\qquad\qquad + h_{X_n}^{\frac{d}{\gamma}-s}\mathbb{E}\left[\left\|(m(\theta_n) - R_{m(\theta_n),\sigma_n^2,\varepsilon}(\theta_n))_{X_n}\right\|_2\right]\Bigg],
\end{aligned}$$

where $C = C\left(\mathcal{X}, d, q, \tau_f, \Theta_N^*\right)$, $h_0 = h_0\left(R, \delta, d, \tau_f, \Theta_N^*\right)$, $\gamma = 2 \vee q$.

Proof Expectation with respect to ε shall be taken at the final step. By the definition of $R^m_{f,\sigma_n^2,\varepsilon}(\theta_n)$

$$R^m_{f,\sigma_n^2,\varepsilon}(\theta_n) = m(\theta_n) + R_{f,\sigma_n^2,\varepsilon}(\theta_n) - R_{m,\sigma_n^2,\varepsilon}(\theta_n) + R_{0,\sigma_n^2,\varepsilon}(\theta_n),$$

therefore

$$\begin{aligned}
\left\|f - R^m_{f,\sigma_n^2,\epsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} &\leq \left\|f - R_{f,\sigma_n^2,\epsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \\
&\quad + \left\|m(\theta_n) - R_{m(\theta_n),\sigma_n^2,\epsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \\
&\quad + \left\|R_{0,\sigma_n^2,\varepsilon}\right\|_{W_q^s(\mathcal{X})}, \qquad (14)
\end{aligned}$$

so as in the proof of Theorem 1, see (4), without loss of generality it suffices to only consider the case $m = 0$. Use Theorem 12 on $f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)$ to see $\exists h_1 > 0$ such that for $h_{X_n} \leq h_1$ and any $s \in [0, \left(\tau_f \wedge \tau_k^-\right)^*]$

$$\begin{aligned}
\left\|f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \leq C_1\Lambda_{s,q}\Bigg(&h_{X_n}^{(\tau_f\wedge\tau(\theta_n))-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+}\left\|f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_2^{\tau_f\wedge\tau(\theta_n)}(\mathcal{X})} \\
&+ h_{X_n}^{\frac{d}{\gamma}-s}\left\|\left(f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right)_{X_n}\right\|_2\Bigg), \qquad (15)
\end{aligned}$$

where $C_1 = C_1(\mathcal{X}, d, q, \tau_f, \Theta_N^*)$, $h_1 = h_1(R, \delta, d, \tau_f, \Theta_N^*)$ and $\gamma = \max(2, q)$. The rest of the proof is spent bounding the $W_2^{\tau_f\wedge\tau(\theta_n)}(\mathcal{X})$ norm term.

For the case $\tau(\theta_n) \leq \tau_f$, the triangle inequality and Lemma 17 can be employed

$$\begin{aligned}\left\|f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_2^{\tau_f \wedge \tau(\theta_n)}(\mathcal{X})} &= \left\|f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_2^{\tau(\theta_n)}(\mathcal{X})} \\ &\leq \|f\|_{W_2^{\tau(\theta_n)}(\mathcal{X})} + \left\|R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_2^{\tau(\theta_n)}(\mathcal{X})} \\ &\leq C_2\big(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + \sigma_n^{-1}\|\varepsilon\|_2\big),\end{aligned}$$

with C_2 bounding the ratio of norm equivalence constants which facilitates the use of RKHS norms in Lemma 17, this is analogous to the use of ratio of norm equivalence constants in (9). Then, combined with (15)

$$\begin{aligned}\left\|f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \leq\ C'\Lambda_{s,q}\Bigg[&h_{X_n}^{\tau(\theta_n)-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+}\|f\|_{W_2^{\tau_f}(\mathcal{X})} \\ &+ h_{X_n}^{\tau(\theta_n)-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+}\sigma_n^{-1}\|\varepsilon\|_2 \\ &+ h_{X_n}^{\frac{d}{\gamma}-s}\left\|(f - R_{f,\sigma_n^2,\varepsilon}(\theta_n))_{X_n}\right\|_2\Bigg],\end{aligned} \tag{16}$$

which recovers the desired result for this case.

For the case when $\tau(\theta_n) > \tau_f$ first apply the triangle inequality

$$\left\|f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \leq \left\|f - R_f(\theta_n)\right\|_{W_q^s(\mathcal{X})} + \left\|R_f(\theta_n) - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}, \tag{17}$$

where $R_f(\theta_n)$ is an interpolant of f at the data points. The first term on the right hand side of (17) can be bounded by a direct application of Theorem 1 (the bound for GPs in the interpolation setting with well-specified likelihood). The second term can be bounded using (16) by replacing f with $R_f(\theta_n)$. This can be done since f and $R_f(\theta_n)$ agree at the data points and $R_f(\theta_n)$ and $R_{f,\sigma_n^2,\varepsilon}(\theta_n)$ have the same smoothness. Combining these two bounds yields

$$\begin{aligned}&\left\|f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \\ &\leq C'\Lambda_{s,q}\Bigg[h_{X_n}^{\tau(\theta_n)-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+}\left\|R_f\right\|_{W_2^{\tau(\theta_n)}(\mathcal{X})} + h_{X_n}^{\tau(\theta_n)-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+}\sigma_n^{-1}\|\varepsilon\|_2 \\ &+ h_{X_n}^{\frac{d}{\gamma}-s}\left\|\big(R_f(\theta_n) - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\big)_{X_n}\right\|_2\Bigg], \\ &\leq C''\Lambda_{s,q}\Bigg[h_{X_n}^{\tau_f-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+}\rho_{X_n}^{\left(\tau(\theta_n)-\tau_f\right)}\left\|f\right\|_{W_2^{\tau_f}(\mathcal{X})} + h_{X_n}^{\tau(\theta_n)-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+}\sigma_n^{-1}\|\varepsilon\|_2 \\ &+ h_{X_n}^{\frac{d}{\gamma}-s}\left\|\big(f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\big)_{X_n}\right\|_2\Bigg],\end{aligned} \tag{18}$$

where (18) uses the fact that $R_f(\theta_n)$ interpolates f to obtain the final term and employs the proof technique involving band limited functions by (Narcowich et al., 2006, Theorem 4.2) to express the $W_2^{\tau(\theta_n)}(\mathcal{X})$ norm in terms of the $W_2^{\tau_f}(\mathcal{X})$ norm.

Combining (18) with the bound obtained when applying Theorem 1 to the first term on the right hand side of (17) then taking appropriate upper and lower bounds of $\tau(\theta_n)$

completes the proof. ■

Combining the following lemma with Theorem 18 completes the proof of Theorem 4.

Lemma 19 *Suppose the assumptions of Theorem 4 hold then for any $f \in W_2^{\tau_f}(\mathcal{X})$*

$$\left\| \left(f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right)_{X_n} \right\|_2 \leq C \left(\|\varepsilon\|_2 + \sigma_n q_{X_n}^{-(\tau_k^+ - \tau_f)_+} \|f\|_{W_2^{\tau_f}(\mathcal{X})} \right),$$

where $C = C\left(\mathcal{X}, d, \tau_f, \Theta_N^*\right)$.

Proof As discussed in the proof by Narcowich et al. (2006, Theorem 4.2) for each n there exists a band-limited function f_{α_n}, where α_n is the bandwidth and depends on q_{X_n}, such that f_α equals f at the points X_n and $\|f_\alpha\|_{W_2^{\tau(\theta_n)}(\mathcal{X})} \leq C_1 q_{X_n}^{-(\tau(\theta_n)-\tau_f)_+}$ for some $C_1 = C_1(\mathcal{X}, d, \tau_f, \tau(\theta_n))$. Using this,

$$\begin{aligned}
\left\| \left(f - R_{f,\sigma_n^2,\varepsilon}(\theta_n)\right)_{X_n} \right\|_2 &= \left\| \left(f_\alpha - R_{f_\alpha,\sigma_n^2,\varepsilon}(\theta_n)\right)_{X_n} \right\|_2 && (19)\\
&\leq \|\varepsilon\|_2 + \left(\|\varepsilon\|_2^2 + \sigma_n^2 \|f_{\alpha_n}\|^2_{\mathcal{H}_{k(\theta_n)}(\mathcal{X})} \right)^{\frac{1}{2}} && (20)\\
&\leq C \left(\|\varepsilon\|_2 + \sigma_n q_{X_n}^{-(\tau_k^+ - \tau_f)_+} \|f\|_{W_2^{\tau_f}(\mathcal{X})} \right) && (21)
\end{aligned}$$

where (20) used Lemma 17 and (21) used the norm equivalence of the RKHS to a Sobolev space and the aforementioned property of f_{α_n} to obtain the q_{X_n} term. ■

Appendix E. Proof of Theorem 7

We will denote by R_ε the kernel interpolant of the the noise, meaning $R_\varepsilon(x) = k_{xX} k_{XX}^{-1} \varepsilon$.

Lemma 20 *Let k be a τ-smooth kernel for $\tau > d/2$, $\varepsilon \in \mathbb{R}^n$ and $X_n \subset \mathcal{X}$. Then, $\exists C > 0$ such that $\|R_\varepsilon\|_{\mathcal{H}_k(\mathcal{X})} \leq C \|\varepsilon\|_2 q_{X_n}^{-(\tau - d/2)}$ for some $C = C(d,k)$ with the dependence on k entering through the RKHS norm equivalence constants.*

Proof Denote by $\lambda_{\min}(A), \lambda_{\max}(A)$ the minimum and maximum eigenvalues of some matrix A. Then:

$$\begin{aligned}
\|R_\varepsilon(\theta_n)\|^2_{H_k(\mathcal{X})} = \varepsilon^\top k_{XX}^{-1} \varepsilon \leq \|\varepsilon\|_2^2 \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{x^\top k_{XX}^{-1} x}{\|x\|_2^2} &= \|\varepsilon\|_2^2 \lambda_{\max}(k_{XX}^{-1}) && (22)\\
&= \|\varepsilon\|_2^2 \lambda_{\min}(k_{XX})^{-1} \leq C \|\varepsilon\|_2^2 q_{X_n}^{-(2\tau - d)}, && (23)
\end{aligned}$$

where the first inequality in (22) is by the reproducing property and the last inequality is by the Rayleigh-Ritz theorem (Horn and Johnson, 2013), (23) is by using the bounds on minimum eigenvalues of kernel matrices discussed by Wendland (2005, Theorem 12.3) which are applicable since k is τ-smooth. See e.g. (Narcowich et al., 2006, Section 3) for

further discussion. ■

To prove Theorem 7 proceed as in the proof of Theorem 18 up to (15) and note the residual term is simply $\|\varepsilon\|_2$ since $R_{f,0,\varepsilon}(\theta_n)$ interpolates the corrupted data, rather than f_X. So $\exists\, C_1, h_1 > 0$ with $C_1 = C_1(\mathcal{X}, d, q, \tau_f, \Theta_N^*)$, $h_1 = h_1(R, \delta, d, \tau_f, \Theta_N^*)$ such that for any X_n with $h_{X_n} \leq h_1$

$$\begin{aligned}
&\left\|f - R_{f,0,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \\
&\quad \leq C_1 \Lambda_{s,q} \left(h_{X_n}^{(\tau_f \wedge \tau(\theta_n)) - s - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \left\|f - R_{f,0,\varepsilon}(\theta_n)\right\|_{W_2^{\tau_f \wedge \tau(\theta_n)}(\mathcal{X})} + h_{X_n}^{\frac{d}{\gamma} - s} \|\varepsilon\|_2 \right). \qquad (24)
\end{aligned}$$

First consider the case when $\tau_f \geq \tau(\theta_n)$

$$\begin{aligned}
\|f - R_{f,0,\varepsilon}(\theta_n)\|_{W_2^{\tau_f \wedge \tau(\theta_n)}(\mathcal{X})} &\leq \|f - R_f(\theta_n)\|_{W_2^{\tau_f}(\mathcal{X})} + \|R_\varepsilon(\theta_n)\|_{W_2^{\tau(\theta_n)}(\mathcal{X})} \qquad (25) \\
&\leq \|f\|_{W_2^{\tau_f}(\mathcal{X})} + C_2 \|\varepsilon\|_2 q_{X_n}^{-(\tau(\theta_n) - \frac{d}{2})}, \qquad (26)
\end{aligned}$$

where (25) is the triangle inequality and (26) is by Lemma 15 and Lemma 20 with $C_2 = C(d, \Theta_N^*)$, the dependency on Θ_N^* manifested by using the ratio of norm equivalence constants to move from Sobolev to RKHS norm. Combining this with (24)

$$\begin{aligned}
&\left\|f - R_{f,0,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \\
&\quad \leq C_1 \Lambda_{s,q} \left(h_{X_n}^{\tau(\theta_n) - s - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \left(\|f\|_{W_2^{\tau_f}(\mathcal{X})} + C_2 \|\varepsilon\|_2 q_{X_n}^{-(\tau(\theta_n) - \frac{d}{2})} \right) + h_{X_n}^{\frac{d}{\gamma} - s} \|\varepsilon\|_2 \right) \\
&\quad \leq C_3 \Lambda_{s,q} \left(h_{X_n}^{\tau(\theta_n) - s - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \|f\|_{W_2^{\tau_f}(\mathcal{X})} + \left(h_{X_n}^{\frac{d}{\gamma} - s} + \rho_{X_n}^{\tau(\theta_n) - \frac{d}{2}} h_{X_n}^{\frac{d}{2} - s - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \right) \|\varepsilon\|_2 \right) \qquad (27) \\
&\quad \leq C_4 \Lambda_{s,q} \left(h_{X_n}^{\tau(\theta_n) - s - d\left(\frac{1}{2} - \frac{1}{q}\right)_+} \|f\|_{W_2^{\tau_f}(\mathcal{X})} + h_{X_n}^{\frac{d}{\gamma} - s} \rho_{X_n}^{\tau(\theta_n) - \frac{d}{2}} \|\varepsilon\|_2 \right), \qquad (28)
\end{aligned}$$

where (27) is absorbing the constants to the front and (28) is because the exponents of the h_{X_n} terms that are multiplied by $\|\varepsilon\|_2$ are the same and $\rho_{X_n} \geq 1$. Taking upper and lower bounds of $\tau(\theta_n)$ completes the proof for this case.

Now consider $\tau_f < \tau(\theta_n)$. In a similar fashion to the second part of the proof of Theorem 18, we use the triangle inequality

$$\left\|f - R_{f,0,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \leq \left\|f - R_f(\theta_n)\right\|_{W_q^s(\mathcal{X})} + \left\|R_f(\theta_n) - R_{f,0,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})}, \qquad (29)$$

where the first term on the right hand side can be bounded by Theorem 1 since it does not involve observation corruption, and the second term on the right hand side can be bounded by (28) by replacing f with $R_f(\theta_n)$ since $R_f(\theta_n)$ and $R_{f,0,\varepsilon}(\theta_n)$ have the same smoothness.

Therefore

$$\begin{aligned}
&\left\|f - R_{f,0,\varepsilon}(\theta_n)\right\|_{W_q^s(\mathcal{X})} \\
&\quad \leq C_5 \Lambda_{s,q} \left(h_{X_n}^{\tau(\theta_n)-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+} \|R_f(\theta_n)\|_{W_2^{\tau(\theta_n)}(\mathcal{X})} + h_{X_n}^{\frac{d}{\gamma}-s} \rho_{X_n}^{\tau(\theta_n)-\frac{d}{2}} \|\varepsilon\|_2 \right) && (30) \\
&\quad \leq C_6 \Lambda_{s,q} \left(h_{X_n}^{\tau(\theta_n)-s-d\left(\frac{1}{2}-\frac{1}{q}\right)_+} q_{X_n}^{-(\tau(\theta_n)-\tau_f)} \|f\|_{W_2^{\tau_f}(\mathcal{X})} + h_{X_n}^{\frac{d}{\gamma}-s} \rho_{X_n}^{\tau(\theta_n)-\frac{d}{2}} \|\varepsilon\|_2 \right) && (31) \\
&\quad \leq C_6 \Lambda_{s,q} h_{X_n}^{\frac{d}{\gamma}-s} \rho_{X_n}^{\tau(\theta_n)-\tau_f} \left(h_{X_n}^{\tau_f-\frac{d}{2}} \|f\|_{W_2^{\tau_f}(\mathcal{X})} + \rho_{X_n}^{\tau_f-\frac{d}{2}} \|\varepsilon\|_2 \right), && (32)
\end{aligned}$$

where (30) is applying Theorem 1 and (28) to the terms on the right hand side of (29). Then (31) uses, as was done in the proof of Theorem 18, the proof technique involving band-limited functions by Narcowich et al. (2006, Theorem 4.2). Finally (32) is collecting the mesh ratio and fill distance terms to the front. Taking appropriate upper and lower bounds for $\tau(\theta_n)$ in terms of τ_k^+, τ_k^- completes the proof of Theorem 7.

Appendix F. Proof of Theorem 2

The proof of Theorem 2 is simply combination of Theorem 18 with a bound on the residual terms and substituting in a bound for $\mathbb{E}[\|\varepsilon\|_2]$. The bound on the residual terms is obtained from an adaptation of the result by van der Vaart and van Zanten (2011, Theorem 1, Theorem 5) to the case of altering hyperparamters. Proving this adaptation is a tedious matter of checking that the constants involved in the bound by van der Vaart and van Zanten (2011, Theorem 1, Theorem 5), which are different for each parameter value, may be controlled given our assumptions on the hyperparameters. The adaptation is stated next along with an explanation of how it is used to prove Theorem 2 and then a proof of the adaptation is given.

Proposition 21 *Suppose Assumptions 1-7 then* $\exists C = C\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})}, \Theta_N^*\right)$ *such that for* $n \geq N$

$$\mathbb{E}\left[\left\|\left(f - R_{f,\sigma^2,\varepsilon}(\theta_n)\right)_{X_n}\right\|_2\right] \leq C \left(n^{\left(\frac{1}{2}-\frac{\tau_f}{2\tau_k^+}\right)_+ \vee \frac{d}{4\tau_k^-}} \right).$$

Direct substitution of this bound into the residuals in Theorem 18 and noting that $\sigma^{-1}\mathbb{E}[\|\varepsilon\|_2] \leq n^{\frac{1}{2}}$ completes the proof of Theorem 2. Proposition 21 is proved by using Jensen's inequality on Theorem 22 in combination with Corollary 26 to obtain the desired bound. The rest of this section shall prove these intermediate results.

Before starting the details of the proof of Proposition 21 we recall the definition of *Hölder spaces of functions* $C^\tau(\mathcal{X})$. For $\tau > 0$ and $\mathcal{X} \subseteq \mathbb{R}^d$ an open set, $C^\tau(\mathcal{X})$ is the space of functions $f\colon \mathcal{X} \to \mathbb{R}$ with $\|f\|_{C^\tau(\mathcal{X})} < \infty$ where

$$\|f\|_{C^\tau(\mathcal{X})} = \max_{m\colon |m| \leq \lfloor \tau - 1 \rfloor} \sup_{x \in \mathcal{X}} |D^m f(x)| + \max_{m\colon |m| \leq \lfloor \tau - 1 \rfloor} \sup_{\substack{x,y \in \mathcal{X} \\ x \neq y}} \frac{|D^m f(x) - D^m f(y)|}{\|x - y\|_2^{\tau - \lfloor \tau - 1 \rfloor}},$$

where $m = (m_1, \ldots, m_d)$ is a multi-index, $|m| = \sum_{i=1}^d m_i$ and D^m is the partial differential operator corresponding to m. Now the framework by van der Vaart and van Zanten (2011) is presented which views the Gaussian process as a measure on function space. The techniques discussed are detached from the results in the present paper and are discussed only to prove Proposition 21, for further details of their origin and use in Bayesian nonparametrics see e.g. (Ghosal and van der Vaart, 2017).

Let $\Pi_{k(\theta_n)}$ denote the probability measure associated with a GP with zero mean and kernel $k(\theta_n)$ over $\mathcal{X}$. Set we $\Pi_{\theta_n} = \Pi_{k(\theta_n)}$ for ease of notation. Given a target function f and a set of points $X_n = \{x_i\}_{i=1}^n$ and observations $y_i = f(x_i) + \varepsilon_i$ with ε_i i.i.d $\mathcal{N}(0, \sigma^2)$ denote the posterior distribution of Π_{θ_n} given $\{y_i\}_{i=1}^n$ as $\Pi_{\theta_n}(\cdot|y_{1:n})$. For $\varepsilon > 0$ and f a continuous function over the closure of $\mathcal{X}$ define the concentration function

$$\phi_{\theta_n, f}(\varepsilon) = \inf_{\substack{h \in \mathcal{H}_{k(\theta_n)}(\mathcal{X}) \\ \|h - f\|_{L^\infty(\mathcal{X})} < \varepsilon}} \frac{1}{2}\|h\|^2_{\mathcal{H}_{k(\theta_n)}(\mathcal{X})} - \log \Pi_{\theta_n}(g \,:\, \|g\|_{L^\infty(\mathcal{X})} < \varepsilon).$$

The first term is called the decentering function and the second the small ball probability. This is finite if and only if f is contained in the closure of $\mathcal{H}_{k(\theta_n)}(\mathcal{X})$ with respect to the supremum norm, which will be true under the assumptions of the theorems in Section 4.2. The next result by van der Vaart and van Zanten (2011, Theorem 1) shows the residuals may be controlled by the concentration function.

Theorem 22 *Let $\mathcal{X}$ be a compact set then $\exists C > 0$ such that for every $f \in C(\mathcal{X})$*

$$\frac{1}{n}\mathbb{E}\left[\int \|(g - f)_{X_n}\|_2^2 d\Pi_{\theta_n}(g|y_{1:n})\right] \leq C\psi^{-1}_{\theta_n, f}(n)^2,$$

where the expectation is being taken with respect to the noise, $\psi_{\theta_n, f}(\varepsilon) = \phi_{\theta_n, f}(\varepsilon)/\varepsilon^2$ and $\psi^{-1}_{\theta_n, f}$ is the generalised inverse of $\psi_{\theta_n, f}$.

Compactness of $\mathcal{X}$ is assumed whereas in Theorem 2 we assumed $\mathcal{X}$ is open. This is not an issue since it is assumed the target function can be extended to all of $\mathbb{R}^d$ so Theorem 22 may be applied to the restriction of the extension to the closure of $\mathcal{X}$, which is compact and contains all the observation points. The decentering function and the small ball probability needs to be bounded. The decentering function is bounded in Lemma 23 and the small ball probability in Lemma 25 which requires more technical work.

Specifically the decentering term is dealt with by upper bounding norm equivalence constants that occur in the proof by van der Vaart and van Zanten (2011, Lemma 4) when performing the kernel convolution approximation argument in that proof, this is summarised in the next lemma.

Lemma 23 *Let f be the restriction to the closure of $\mathcal{X}$ of some $f^\circ \in C^{\tau_f}(\mathbb{R}^d) \cap W_2^{\tau_f}(\mathbb{R}^d)$ with $\tau(\theta_n) > \tau_f > d/2$. Then, $\exists C = C\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})}, \Theta_N^*\right)$ such that for $\varepsilon < 1$*

$$\inf_{\substack{h \in \mathcal{H}_{k(\theta_n)}(\mathcal{X}) \\ \|h - f\|_{L^\infty(\mathcal{X})} < \varepsilon}} \frac{1}{2}\|h\|^2_{\mathcal{H}_{k(\theta_n)}(\mathcal{X})} \;\leq\; C\varepsilon^{-\frac{2(\tau(\theta_n) - \tau_f)}{\tau_f}} \;\leq\; C\varepsilon^{-2\frac{(\tau_k^+ - \tau_f)}{\tau_f}}.$$

If $\tau_f \geq \tau(\theta_n)$ then $f \in \mathcal{H}_{k(\theta_n)}(\mathcal{X})$ therefore we could take $h = f$ and the bound would be $\frac{1}{2}\|f\|_{W_2^{\tau(\theta_n)}(\mathcal{X})}$ which has no dependence on ε so in this case the growth of the concentration function is dictated entirely by the small ball probability term.

The small ball probability bound requires the result by Li and Linde (1999, Theorem 1.2) which relates small ball probabilities to the metric entropy of the unit ball of the RKHS corresponding to the kernel. Metric entropy is a method of measuring the size of a given function space denoted $H(M, \varepsilon)$ and is defined as the logarithm of the ε-covering number of M, for more discussion see e.g. (Giné and Nickl, 2016, Chapter 2.3). A bound on metric entropy is given by Giné and Nickl (2016, Theorem 4.3.36) which illuminates the way the hyperparameters effect the bound. The constants in the proof can easily be bounded by replacing $\tau(\theta_n)$ by τ_k^+ and τ_k^- where appropriate and doing so yields the next lemma.

Lemma 24 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-7 hold. Let $\mathcal{H}^{(1)}_{k(\theta_n)}(\overline{\mathcal{X}})$ denote the unit ball of the RKHS of $k(\theta_n)$ over the closure of $\mathcal{X}$. Then $\exists C_{met} = C_{met}(\Theta_N^*)$ such that $\forall n \geq N$ and $\forall \varepsilon < 1$*

$$H\left(\mathcal{H}^{(1)}_{k(\theta_n)}\left(\overline{\mathcal{X}}\right), \varepsilon\right) \leq C_{met}\, \varepsilon^{-\frac{d}{\tau(\theta_n)}}.$$

The proof by van der Vaart and van Zanten (2011, Theorem 5) is now followed to link metric entropy to small ball probability. This will involve going through auxiliary results by Li and Linde (1999) to make sure the possible altering hyperparameters result in constants that are controlled, this is a tedious process and the referenced paper should be consulted for greater context. Before this is started note that the two auxiliary results by Li and Linde (1999, Lemma 2.1, Lemma 2.2) used to link entropy numbers to GPs do not depend on the hyperparameter choices since they hold with constants not depending on the smoothness of the RKHS, see e.g. (Pisier, 1989, Theorem 9.1) and (Artstein et al., 2004, Page 1315). The first step is to use (Li and Linde, 1999, Proposition 2.4) in combination with the bound we have derived for metric entropy to get that $\forall \gamma > 2d/(2\tau(\theta_n) - d)$, $\exists\, C(\theta_n, \gamma) > 0$ such that

$$\phi_{\theta_n}(\varepsilon) \coloneqq -\log \Pi_{\theta_n}\left(\|f\|_\infty \leq \varepsilon\right) \leq C(\theta_n, \gamma)\varepsilon^{-\gamma}. \tag{33}$$

Next we explain the result by Li and Linde (1999, Proposition 3.1). First, using the result by Li and Linde (1999, Equation 3.4) and Lemma 24

$$\begin{aligned}\phi_{\theta_n}(\varepsilon) &\leq \log 2 + H\left(\mathcal{H}^{(1)}_{k(\theta_n)}, \varepsilon(8\phi_{\theta_n}(\varepsilon/2))^{-\frac{1}{2}}\right) \\ &\leq \log 2 + C_{met}\varepsilon^{-\frac{d}{\tau(\theta_n)}} 8^{\frac{d}{2\tau(\theta)}} \phi_{\theta_n}\left(\varepsilon/2\right)^{\frac{d}{2\tau(\theta_n)}}, \end{aligned} \tag{34}$$

then once ε gets smaller than some ε^* the second term on the right hand side of (34) becomes greater than $L \log 2$ for some constant $L > 0$ (Li and Linde, 1999, Equation 3.4). This argument does not consider changing hyperparameters. Indeed, for a fixed constant $L > 0$ for different hyperparameters θ_n we might need different ε_n^* to conclude that the second term is greater than $L \log 2$. Assumption 5 introduces the required uniformity by allowing us to say that once ε is small enough (34) can be bounded by a constant times the second term in (34) for all hyperparameter choices. Specifically, by Assumption 6 we know

that if $\varepsilon < c$ then $\phi_{\theta_n}(\varepsilon/2) \geq \alpha_N$, therefore if we set

$$\varepsilon^* := \min\left(c, \left(\alpha_N^{\frac{d}{2\tau_k^+}} (\log 2)^{-1}\right)^{\frac{\tau_k^-}{d}}\right), \tag{35}$$

then for $\varepsilon < \varepsilon^*$, we have

$$\begin{aligned}\phi_{\theta_n}(\varepsilon) &\leq \log 2 + C_{met}\, \varepsilon^{-\frac{d}{\tau(\theta_n)}} 8^{\frac{d}{2\tau(\theta_n)}} \phi_{\theta_n}\left(\varepsilon/2\right)^{\frac{d}{2\tau(\theta_n)}} \\ &\leq (C_{met}+1)\, \varepsilon^{-\frac{d}{\tau(\theta_n)}} 8^{\frac{d}{2\tau(\theta_n)}} \phi_{\theta_n}\left(\varepsilon/2\right)^{\frac{d}{2\tau(\theta_n)}}.\end{aligned}$$

Now take logarithms and employ the iterative procedure from the proof by Li and Linde (1999, Proposition 3.1). Taking logarithms gives

$$\log \phi_{\theta_n}(\varepsilon) \leq \frac{d}{2\tau(\theta_n)} \log \phi_\theta\left(\varepsilon/2\right) + \log \chi_n(\varepsilon),$$

where $\chi_n(\varepsilon) = (C_{met}+1)\varepsilon^{-\frac{d}{\tau(\theta_n)}} 8^{\frac{d}{2\tau(\theta_n)}}$. Now iterate this inequality so that for any $m \in \mathbb{N}$

$$\log \phi_{\theta_n}(\varepsilon) \leq \left(\frac{d}{2\tau(\theta_n)}\right)^m \log \phi_{\theta_n}\left(2^{-m}\varepsilon\right) + \sum_{j=0}^{m-1} \left(\frac{d}{2\tau(\theta_n)}\right)^j \log \chi\left(2^{-j}\varepsilon\right), \tag{36}$$

and note that the left hand side does not depend on m and substituting the bound in (33) reveals the first term on the right hand side of (36) converges to zero as $m \to \infty$

$$\begin{aligned}\left(\frac{d}{2\tau(\theta_n)}\right)^m \log \phi_{\theta_n}\left(2^{-m}\varepsilon\right) &\leq \left(\frac{d}{2\tau(\theta_n)}\right)^m \log\left(C(\theta_n,\gamma) 2^{m\gamma}\varepsilon^{-\gamma}\right) \\ &= \left(\frac{d}{2\tau(\theta_n)}\right)^m \left(m\gamma \log 2 + \log(C(\theta_n,\gamma)\varepsilon^{-\gamma})\right) \xrightarrow{m\to\infty} 0.\end{aligned}$$

So taking the limit of m in (36) gives

$$\begin{aligned}\log \phi_{\theta_n}(\varepsilon) &\leq \sum_{j=0}^{\infty} \left(\frac{d}{2\tau(\theta_n)}\right)^j \log \chi_n\left(2^{-j}\varepsilon\right) \\ &= \frac{2\tau(\theta_n)}{(2\tau(\theta_n)-d)} \log \chi_n(\varepsilon) + \sum_{j=0}^{\infty} \left(\frac{d}{2\tau(\theta_n)}\right)^j \log\left(\frac{\chi_n(2^{-j}\varepsilon)}{\chi_n(\varepsilon)}\right) \\ &\leq \frac{2\tau(\theta_n)}{(2\tau(\theta_n)-d)} \log \chi_n(\varepsilon) + \log(2)\frac{d}{\tau_k^-} \sum_{j=0}^{\infty} \left(\frac{d}{2\tau_k^-}\right)^j j,\end{aligned}$$

and sum has a closed form which we can upper bound

$$\log \phi_{\theta_n}(\varepsilon) \leq \frac{2\tau(\theta_n)}{(2\tau(\theta_n)-d)} \log \chi_n(\varepsilon) + \log(2)\left(\frac{d}{\tau_k^-}\right)\left(\frac{d}{2\tau_k^-}\right)\left(\frac{d}{2\tau_k^-}-1\right)^{-2}.$$

Finally, exponentiating tells us that $\forall \varepsilon < \varepsilon^*$:

$$\phi_{\theta_n}(\varepsilon) \leq C^* \varepsilon^{-2d/(2\tau(\theta_n)-d)} \leq C^* \varepsilon^{-2d/(2\tau_k^- - d)},$$

where we have collected the dependencies on Θ_N^* into C^*. In summary the following lemma which is analogous to the result by van der Vaart and van Zanten (2011, Lemma 3), but with possibly changing hyperparameters, has been proved.

Lemma 25 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-7 hold. Then, for $\varepsilon < \varepsilon^*$, where ε^* is from (35), and $n \geq N$: $-\log \Pi_{\theta_n}\left(\|f\|_{L^\infty(\overline{\mathcal{X}})} \leq \varepsilon\right) \leq C^* \varepsilon^{-2d/(2\tau_k^- - d)}$.*

Corollary 26 *Fix $N \in \mathbb{N}$ and suppose Assumptions 1-7 hold. Then, $\exists C = C\left(\|f\|_{W_2^{\tau_f}(\mathcal{X})}, \Theta_N^*\right)$ such that for $n \geq N$, $\psi^{-1}_{\theta_n, f}(n) \leq C \max\left(n^{-\tau_f/2\tau_k^+}, n^{d/4\tau_k^- - 1/2}\right)$.*

Proof By Lemma 23 and Lemma 25, using the restriction of f° to the closure of $\mathcal{X}$, $\exists C_1 > 0$ such that $\forall\, n \geq N$ and $\varepsilon < \varepsilon^*$, where ε^* is from (35)

$$\begin{aligned}\phi_{\theta_n, f}(\varepsilon)\varepsilon^{-2} &\leq C_1\left(\varepsilon^{-(2d/(2\tau(\theta_n)-d))-2} + \varepsilon^{-2((\tau(\theta_n)-\tau)/\tau)-2}\right) \\ &\leq C_1 \varepsilon^{-\frac{2\tau(\theta_n)}{\min(\tau, \tau(\theta_n)-(d/2))}} \leq C_1\left(\varepsilon^{-2\tau_N^+/\tau} \vee \varepsilon^{\left(d/(4\tau_N^-)-1/2\right)^{-1}}\right).\end{aligned}$$

Set $\varepsilon_n = n^{-\left(\tau/2\tau_k^+\right) \wedge (d/4\tau_k^- - 1/2)}$ then we know once n is large enough that we have $\varepsilon_n < \varepsilon^*$ therefore $\exists\, C_2$ such that $\forall n \geq N$ $\phi_{\theta_n, f}(\varepsilon_n)\varepsilon_n^{-2} \leq C_2 n$. Multiplying ε_n by a constant to remove the factor of C_2 in the previous expression completes the proof. ■

Appendix G. Proof of Theorem 11

The proof follows the proof by (Bull, 2011, Theorem 1). The point x_n satisfies

$$f(x^*) - f(x_n) \leq f(x^*) - R_f(x^*) - f(x_n) + R_f(x_n) \leq 2\left\|f - R_f\right\|_{L^\infty(\mathcal{X})},$$

since $R_f(x_n) \geq R_f(x^*)$ since x_n was chosen as the maximizer of R_f. The points picked by the (γ, F, n) strategy are quasi-uniform by Wenzel et al. (2019, Theorem 14, Theorem 18) therefore taking n_0 to be large enough to ensure that the fill distance obtained from the strategy is small enough to employ Theorem 1 completes the proof.

References

Rémi Arcangéli and Juan José Torrens. Sampling inequalities in Sobolev spaces. *Journal of Approximation Theory*, 182:18–28, 2014.

Rémi Arcangéli, María Cruz López de Silanes, and Juan José Torrens. 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, 2007.

Rémi Arcangéli, María Cruz López de Silanes, and Juan José Torrens. Extension of sampling inequalities to Sobolev semi-norms of fractional order and derivative data. *Numerische Mathematik*, 121(3):587–608, 2012.

Margaret Armstrong and Azeddine Boufassa. Comparing the robustness of ordinary kriging and lognormal kriging: Outlier resistance. *Mathematical Geology*, 20(4):447–457, 1988.

Shirit Artstein, Vitali Milman, and Stanisław J. Szarek. Duality of metric entropy. *Annals of Mathematics*, 159(3):1313–1328, 2004.

Felix Berkenkamp, Angela P. Schoellig, and Andreas Krause. No-regret bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 20(50): 1–24, 2019.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, MA, 2004.

Anna Breger, Martin Ehler, and Manuel Gräf. Points on manifolds with asymptotically optimal covering radius. *Journal of Complexity*, 48:1–14, 2018.

François-Xavier Briol, Chris J. Oates, Mark Girolami, Michael A. Osborne, and Dino Sejdinovic. Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1), 2019.

Adam D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(88):2879–2904, 2011.

Zhehui Chen, Simon Mak, and C. F. Jeff Wu. A hierarchical expected improvement method for Bayesian optimization. *arXiv:1911.07285*, 2019.

Andreas Christmann and Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

Jon Cockayne, Chris J. Oates, Tim Sullivan, and Mark Girolami. Probabilistic meshless methods for partial differential equations and Bayesian inverse problems. *arXiv:1605.07811*, 2016.

Noel Cressie. The origins of kriging. *Mathematical Geology*, 22(3):239–252, 1990.

Felipe Cucker and Ding-Xuan Zhou. *Learning theory: An Approximation Theory Viewpoint*, volume 24 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007.

Marc Peter Deisenroth, Dieter Fox, and Carl Edward Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):408–423, 2015.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

Alexander I. J. Forrester, András Sóbester, and Andy J. Keane. *Engineering Design via Surrogate Modelling*. Wiley, 2008.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2017.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*, volume 40 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, New York, 2016.

Paul W. Goldberg, Christopher K. I. Williams, and Christopher M. Bishop. Regression with input-dependent noise: a Gaussian process treatment. In *Advances in Neural Information Processing Systems*, pages 493–499, 1998.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer New York, 2002.

Doug P. Hardin, Edward B. Saff, and J. Tyler Whitehouse. Quasi-uniformity of minimal weighted energy points on compact metric spaces. *Journal of Complexity*, 28(2):177–191, 2012.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, second edition, 2013.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, 2009.

Mark E. Johnson, Leslie M. Moore, and Donald Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2):131–148, 1990.

V. Roshan Joseph and Ying Hung. Orthogonal-maximin Latin hypercube designs. *Statistica Sinica*, 18(1):171–186, 2008.

Pasi Jylänki, Jarno Vanhatalo, and Aki Vehtari. Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research*, 12(99):3227–3257, 2011.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv:1807.02582*, 2018.

Motonobu Kanagawa, Bharath K. Sriperumbudur, and Kenji Fukumizu. Convergence analysis of deterministic kernel-based quadrature rules in misspecified settings. *Foundations of Computational Mathematics*, 20(1):155–194, 2020.

Toni Karvonen, George Wynne, Filip Tronarp, Chris Oates, and Simo Särkkä. Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):926–958, 2020.

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

Bas Kleijn and Aad van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877, 2006.

Danie G. Krige. *A Statistical Approach to Some Basic Mine Valuation Problems on the Witwatersrand*. PhD thesis, University of Witwatersrand, 1951.

Malte Kuss and Carl Edward Rasmussen. Gaussian processes in reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 751–758, 2004.

Quoc V. Le, Alex J. Smola, and Stéphane Canu. Heteroscedastic Gaussian process regression. In *International Conference on Machine Learning*, pages 489–496, 2005.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Uniform error bounds for Gaussian process regression with application to safe control. In *Advances in Neural Information Processing Systems*, pages 657–667, 2019.

Wenbo V. Li and Werner Linde. Approximation, metric entropy and small ball estimates for Gaussian measures. *The Annals of Probability*, 27(3):1556–1578, 1999.

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

Rubin Martinez-Cantin, Kevin Tee, and Michael McCourt. Practical Bayesian optimization in the presence of outliers. *International Conference on Artificial Intelligence and Statistics*, pages 1722–1731, 2018.

Arak M. Mathai and Serge B. Provost. *Quadratic Forms in Random Variables*, volume 126 of *Statistics: Textbooks and Monographs*. Marcel Dekker, Inc., New York, 1992.

Georges Matheron. Principles of Geostatistics. *Economic Geology*, 58(8):1246–1266, 1963.

Michael D. McKay, Richard J. Beckman, and William J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

Jonas Mockus. *Bayesian Approach to Global Optimization*, volume 37 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1989.

Max D. Morris and Toby J. Mitchell. Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43(3):381–402, 1995.

Stefan Müller. *Komplexität und Stabilität von kernbasierten Rekonstruktionsmethoden*. PhD thesis, Universität Göttingen, 2008.

Francis J. Narcowich, Joseph D. Ward, and Holger Wendland. Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation. An International Journal for Approximations and Expansions*, 24(2):175–186, 2006.

Erich Novak and Henryk Woźniakowski. *Tractability of Multivariate Problems*. European Mathematical Society Publishing House, 2008.

Chris J. Oates, Jon Cockayne, François-Xavier Briol, and Mark Girolami. Convergence rates for a class of estimators based on Stein's method. *Bernoulli*, 25(2):1141–1159, 2019.

Gilles Pisier. *The Volume of Convex Bodies and Banach space Geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

Luc Pronzato and Werner G. Müller. Design of computer experiments: space filling and beyond. *Statistics and Computing*, 22(3):681–701, 2012.

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2006.

Christian Rieger and Barbara Zwicknagl. Deterministic error analysis of support vector regression and related regularized kernel methods. *Journal of Machine Learning Research*, 10(73):2115–2132, 2009.

Christian Rieger, Robert Schaback, and Barbara Zwicknagl. Sampling and stability. In *Mathematical Methods for Curves and Surfaces*, volume 5862 of *Lecture Notes in Computer Science*, pages 347–369. Springer, Berlin, 2010.

Stephen Roberts, Michael A. Osborne, Mark Ebden, Steven Reece, Neale Gibson, and Suzanne Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371 (1984), 2013.

Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.

Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435, 1989.

Thomas J. Santner, Brian J. Williams, and William I. Notz. *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. Springer, New York, 2018.

Andria Sarri, Serge Guillas, and Frederic Dias. Statistical emulation of a tsunami model for sensitivity analysis and uncertainty quantification. *Natural Hazards and Earth System Science*, 12(6):2003–2018, 2012.

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

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances In Neural Information Processing Systems*, pages 2951–2959, 2012.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, page 1015–1022, 2010.

Oliver Stegle, Sebastian V. Fallert, David J. C. MacKay, and Soren Brage. Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering*, 55(9):2143–2151, 2008.

Elias M. Stein. *Singular Integrals and Differentiability Properties of Functions*. Princeton Mathematical Series, No. 30. Princeton University Press, Princeton, N.J., 1970.

Michael L. Stein. *Interpolation of Spatial Data*. Springer New York, 1999.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008.

Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Conference on Learning Theory*, 2009.

Andrew M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

Andrew M. Stuart and Aretha L. Teckentrup. Posterior consistency for Gaussian process approximations of Bayesian posterior distributions. *Mathematics of Computation*, 87 (310):721–753, 2018.

Aretha L. Teckentrup. Convergence of Gaussian process regression with estimated hyperparameters and applications in Bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4):1310–1337, 2020.

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

Rui Tuo and Wenjia Wang. Kriging prediction with isotropic Matérn correlations: Robustness and experimental designs. *Journal of Machine Learning Research*, 21(187):1–38, 2020.

Florencio I. Utreras. Convergence rates for multivariate smoothing spline functions. *Journal of Approximation Theory*, 52(1):1–27, 1988.

Aad van der Vaart and Harry van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 2011.

Jarno Vanhatalo, Pasi Jylänki, and Aki Vehtari. Gaussian process regression with Student-t likelihood. In *Advances in Neural Information Processing Systems*, pages 1910–1918, 2009.

Emmanuel Vazquez and Julien Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and Inference*, 140(11):3088–3095, 2010.

Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

Wenjia Wang, Rui Tuo, and C. F. Jeff Wu. On prediction properties of kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 115(530): 920–930, 2019.

Yaping Wang, Jianfeng Yang, and Hongquan Xu. On the connection between maximin distance designs and orthogonal designs. *Biometrika*, 105(2):471–477, 2018.

Holger Wendland. *Scattered Data Approximation*, volume 17 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2005.

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

Tizian Wenzel, Gabriele Santin, and Bernard Haasdonk. A novel class of stabilized greedy kernel approximation algorithms: Convergence, stability & uniform point distribution. *arXiv:1911.04352*, 2019.

Xiaoyue Xi, François-Xavier Briol, and Mark Girolami. Bayesian quadrature for multiple related integrals. In *International Conference on Machine Learning*, pages 5369–5378, 2018.