Journal of Machine Learning Research 27 (2026) 1-36 Submitted 6/24; Revised 1/26; Published 3/26

Vector-Valued Gaussian Processes for Approximating Divergence- or Rotation-free Vector Fields

Quoc Thong Le Gia QLEGIA@UNSW.EDU.AU
School of Mathematics and Statistics
University of New South Wales
Sydney, NSW 2052, Australia

Ian Hugh Sloan I.SLOAN@UNSW.EDU.AU
School of Mathematics and Statistics
University of New South Wales
Sydney, NSW 2052, Australia

Holger Wendland HOLGER.WENDLAND@UNI-BAYREUTH.DE
Department of Mathematics
University of Bayreuth
95440 Bayreuth, Germany

Editor: Massimiliano Pontil

Abstract

In this paper, we discuss vector-valued Gaussian processes for the approximation of divergence- or rotation-free functions. We establish the theory for such Gaussian processes, then link the theory to multivariate approximation theory, and finally give error estimates for the predictive mean in various situations.

Keywords: Gaussian Process, Kernel-Based Learning, Divergence-Free, Rotation-Free

1. Introduction

Gaussian processes are established tools for approximating unknown functions under the influence of uncertainties; see Rasmussen and Williams (2006). The approximation properties of Gaussian processes of scalar-valued functions, the *so-called regression problem*, have extensively been studied, see for example Nieman et al. (2022); Pati et al. (2015); Teckentrup (2020); van der Vaart and van Zanten (2011); Wynne et al. (2021); Zhou et al. (2023). However, many applications, particularly from physics and engineering, require the reconstruction of vector-valued or, more generally, Hilbert-space valued functions. For example, when modeling fluid flow problems with incompressible flows, the velocity field at a fixed point in time is given by a differentiable function $\boldsymbol{v} : \mathcal{D} \to \mathbb{R}^d$ with $\mathcal{D} \subseteq \mathbb{R}^d$ and usually $d = 2, 3$ satisfying $\nabla \cdot \boldsymbol{v} = \sum_j \partial_j v_j = 0$ if $\boldsymbol{v} = (v_1, \dots, v_d)$.

Another application is the modeling of magnetic fields using Maxwell's equations. Here two fields $\boldsymbol{H} : \mathcal{D} \to \mathbb{R}^3$ and $\boldsymbol{B} : \mathcal{D} \to \mathbb{R}^3$ are usually used, which have to satisfy, among other things, $\nabla \cdot \boldsymbol{B} = 0$ and $\nabla \times \boldsymbol{H} = \boldsymbol{0}$, respectively.

In this paper, we do not consider the underlying partial differential equations, but rather assume that we have some *training data* in the form of *data sites* $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ and vector *data values* $\boldsymbol{v}(\boldsymbol{x}_j)$, $1 \le j \le N$, with $\boldsymbol{v} : \mathcal{D} \to \mathbb{R}^3$.

It is the goal of this paper to study generalizations of scalar-valued Gaussian processes for vector-valued or even Hilbert-space-valued functions. We are particularly interested in modeling vector-valued functions which are either divergence or rotation free. Vector-valued and even Hilbert-space-valued Gaussian processes are well-known in the statistics literature (see for example Adler (1990)). Such Gaussian processes and other kernel-based learning methods have extensively been used in the context of learning vector-valued functions or solving multitask problems, see for example Bonilla et al. (2007); Álvarez et al. (2012); Caponnetto et al. (2008); Carmeli et al. (2006, 2010); Micchelli and Pontil (2004, 2005); Parra and Tobar (2017). Typical methods described in these papers, comprise penalized least-squares regression, interpolation, Kriging and Gaussian processes.

However, none of these papers discusses divergence-free or rotation-free kernels with the exception of Álvarez et al. (2012), where such kernels are at least mentioned but not further investigated. Moreover, none of these papers provide an error analysis at all.

The need for divergence-free or rotation-free Gaussian processes became apparent for example in Bilionis et al. (2013); Gunes and Rist (2008); Owhadi (2023); Wahlström et al. (2013). Again, these papers do not contain any error analysis and, if at all, only test divergence-free or rotation-free kernels numerically.

There are some papers discussing error estimates in the context of learning rates. Most notably Li et al. (2013); Caponnetto and Vito (2007). The analysis and results there differ significantly from ours. Notably, in Li et al. (2013), only essentially diagonal kernels are considered, which immediately excludes the kernels necessary for divergence- or rotation-free approximations. Moreover, both papers are interested only in regression and not interpolation and give error estimates only in probability under the assumption that the data sites follow a given distribution. We, however, are interested in both interpolation and regression and want to derive deterministic, worst case error estimates, yielding also strategies for the selection of the smoothing parameter in the regression case. As pointed out in Wynne et al. (2021), such estimates for "functions with vector valued output [...] would be an important avenue of research".

The new contributions of this paper are the mathematically precise introduction and discussion of divergence- and rotation-free Gaussian processes together with the analysis of their connection to matrix-valued positive definite kernels and vector-valued reproducing kernel Hilbert spaces, and finally error analysis in Sobolev spaces. For the latter, we can rely on results from approximation theory, particularly Fuselier (2008a), but require new results on the approximation operator.

2. Gaussian Processes and Approximation

We shortly review the necessary material on (scalar-valued) Gaussian processes and their connection to multivariate, kernel-based approximation theory, laying out the path we will also follow for vector-valued functions. Basic definitions and details can be found in Rasmussen and Williams (2006).

Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ consisting of the abstract domain Ω, a σ-algebra $\mathcal{A}$ on Ω and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{A})$ and an *index domain* $\mathcal{D} \subseteq \mathbb{R}^d$, a Gaussian process is a family of random variables $g : \Omega \times \mathcal{D} \to \mathbb{R}$, such that for any finite set of pairwise distinct points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$, the random variable $\boldsymbol{g} = (g(\cdot, \boldsymbol{x}_1), \ldots, g(\cdot, \boldsymbol{x}_N))^{\mathrm{T}}$

is jointly normally distributed with mean $\boldsymbol{m} = (m(\boldsymbol{x}_1), \dots, m(\boldsymbol{x}_N))^{\mathrm{T}}$ and covariance matrix $\Sigma = (\mathbb{E}[(g(\cdot, \boldsymbol{x}_i) - m(\boldsymbol{x}_i))(g(\cdot, \boldsymbol{x}_j) - m(\boldsymbol{x}_j))])_{1 \le i,j \le N} \in \mathbb{R}^{N \times N}$. Obviously, this means that the Gaussian process is uniquely determined by a function $m : \mathcal{D} \to \mathbb{R}$ giving the mean and a function $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ describing the covariance via $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Sigma_{ij}$. As usual in this context, we will omit the dependence on the unknown variable from Ω and simply write $g(\boldsymbol{x})$ instead of $g(\cdot, \boldsymbol{x})$. Throughout we shall use the notation

$$g \sim \mathrm{GP}(m, K).$$

The Gaussian process is called *centered* if it has mean zero.

Gaussian processes are often used to model the approximation of an unknown function $f : \mathcal{D} \to \mathbb{R}$. If nothing at all is known about this function, then we may try to approximate f by an initial Gaussian process $f_0 : \Omega \times \mathcal{D} \to \mathbb{R}$ with mean function $m(\boldsymbol{x}) = 0$, $\boldsymbol{x} \in \mathcal{D}$, though other mean functions are possible, and a preselected kernel function $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$. Often, this kernel depends on several parameters, but we will suppress this for simplicity. Obviously, f_0 may not be a particularly good approximation. This will depend significantly on the choice of the kernel.

Importantly, if we have N observations of the unknown function f at data sites $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ (also called *design or training points*), with data values either in the form $y_j = f(\boldsymbol{x}_j)$ $1 \le j \le N$ without noise, or in the form $y_j = f(\boldsymbol{x}_j) + \epsilon_j$, $1 \le j \le N$ with noise, then we can condition the initial Gaussian process using these data to obtain an *a posteriori Gaussian process.* In the case of no noise, $f(\boldsymbol{x}_j) = y_j$, this process is given by $f_{N,0} \sim \mathrm{GP}(m_N, K_N)$ with new mean and kernel functions

$$\begin{aligned} m_N(\boldsymbol{x}) &:= m(\boldsymbol{x}) + K(\boldsymbol{x}, X)K(X, X)^{-1}\left(\boldsymbol{y} - m(X)\right), &(1)\\ K_N(\boldsymbol{x}, \boldsymbol{x}') &:= K(\boldsymbol{x}, \boldsymbol{x}') - K(\boldsymbol{x}, X)K(X, X)^{-1}K(\boldsymbol{x}', X)^{\mathrm{T}}, &(2)\end{aligned}$$

where $\boldsymbol{y} = (y_1, \dots, y_N)^{\mathrm{T}} \in \mathbb{R}^N$ are the observations at X, $m(X) := (m(\boldsymbol{x}_1), \dots, m(\boldsymbol{x}_N))^{\mathrm{T}} \in \mathbb{R}^N$ is the vector of means and $K(\boldsymbol{x}, X) := (K(\boldsymbol{x}, \boldsymbol{x}_1), \dots, K(\boldsymbol{x}, \boldsymbol{x}_N)) \in \mathbb{R}^{1 \times N}$. Finally, $K(X, X) \in \mathbb{R}^{N \times N}$ is the matrix with elements $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

The functions m_N and K_N are also called *predictive mean* and *predictive covariance*, respectively.

In the context of noisy observations $y_j = f(\boldsymbol{x}_j) + \epsilon_j$, $1 \le j \le N$, the assumption that the noise ϵ_j, $1 \le j \le N$, follows an independent, identical Gaussian distribution with mean zero and variance $\sigma^2 > 0$ leads to the posterior Gaussian process $f_{N,\sigma} \sim \mathrm{GP}(m_{N,\sigma}, K_{N,\sigma})$ with

$$\begin{aligned} m_{N,\sigma}(\boldsymbol{x}) &:= m(\boldsymbol{x}) + K(\boldsymbol{x}, X)(K(X, X) + \sigma^2 I)^{-1}\left(\boldsymbol{y} - m(X)\right), &(3)\\ K_{N,\sigma}(\boldsymbol{x}, \boldsymbol{x}') &:= K(\boldsymbol{x}, \boldsymbol{x}') - K(\boldsymbol{x}, X)(K(X, X) + \sigma^2 I)^{-1}K(\boldsymbol{x}', X)^{\mathrm{T}}. &(4)\end{aligned}$$

Here, $I \in \mathbb{R}^{N \times N}$ denotes the identity matrix. Otherwise, the notation is as in (1) and (2). From the above equations, we see that $f_{N,\sigma}$ reduces to $f_{N,0}$ if $\sigma = 0$, justifying our notation.

There is a strong connection to multivariate interpolation theory, particularly approximation by kernels or *radial basis functions.* To specify this connection, we will require the following terminology.

Definition 1 *1. A continuous, symmetric function $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is called* positive definite *if for all $N \in \mathbb{N}$ and all sets of points, the kernel matrix $K(X,X) \in \mathbb{R}^{N\times N}$ is positive definite (that is, if $\boldsymbol{\alpha}^{\mathrm{T}} K(X,X)\boldsymbol{\alpha} > 0$ for all $\boldsymbol{\alpha} \in \mathbb{R}^N \setminus \{\mathbf{0}\}$). Such a function is also called a* positive definite kernel*.*

2. A Hilbert space $\mathcal{H} = \{f : \mathcal{D} \to \mathbb{R}\} \subseteq C(\mathcal{D})$ of functions is a reproducing kernel Hilbert space (RKHS) *if there is a function $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ satisfying $K(\cdot, \boldsymbol{x}) \in \mathcal{H}$ for all $\boldsymbol{x} \in \mathcal{D}$ and $f(\boldsymbol{x}) = \langle f, K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and all $\boldsymbol{x} \in \mathcal{D}$. The function K is called the* reproducing kernel *of $\mathcal{H}$.*

It is well known, see for example Aronszajn (1950); Wendland (2005), that the reproducing kernel of a RKHS is always positive semi-definite (i.e. $K(X,X)$ is symmetric and positive semi-definite for all X) and that the kernel is uniquely determined by the Hilbert space. Moreover, for every positive definite kernel there is a unique RKHS having this kernel as its reproducing kernel.

For us, the following connections between positive definite kernels, RKHS and Gaussian processes are important. We start with the no-noise case. The proofs of the different statements of the next theorem can be found in or follow easily from Wendland (2005) and Scheuerer et al. (2013); Stuart and Teckentrup (2017).

Theorem 2 *Let $K : \mathcal{D}\times\mathcal{D} \to \mathbb{R}$ be a positive definite kernel and let $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ consist of pairwise distinct points. Define the N-dimensional approximation space $V_X = \operatorname{span}\{K(\cdot, \boldsymbol{x}_j) : 1 \le j \le N\}$ and let $\mathcal{H}$ be the RKHS of the kernel K.*

1. *For every $\boldsymbol{y} \in \mathbb{R}^N$ there is a unique function $s_{0,\boldsymbol{y}} \in V_X$ interpolating the data, i.e. satisfying $s_{0,\boldsymbol{y}}(\boldsymbol{x}_i) = y_i$, $1 \le i \le N$. This defines an interpolation operator $I_X : \mathbb{R}^N \to V_X$, $I_X\boldsymbol{y} := s_{0,\boldsymbol{y}}$ and also an operator $I_X : C(\mathcal{D}) \to V_X$ by setting $I_X f = I_X(f|X)$.*

2. *The function $s_{0,\boldsymbol{y}} \in V_X$ is the solution of*

$$\min\left\{\|s\|_{\mathcal{H}} : s \in \mathcal{H} \text{ with } s(\boldsymbol{x}_i) = y_i, 1 \le i \le N\right\}.$$

3. *If m_N is the predictive mean from (1) with data given by $y_i = f(\boldsymbol{x}_i)$, $1 \le i \le N$, then $m_N = m + I_X f - I_X m$.*

4. *If K_N is the predictive covariance kernel from (2) then*

$$K_N(\boldsymbol{x}, \boldsymbol{x}') = K(\boldsymbol{x}, \boldsymbol{x}') - I_X K(\cdot, \boldsymbol{x}')(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x}') - I_X K(\cdot, \boldsymbol{x})(\boldsymbol{x}'),$$

i.e. the predictive covariance kernel is the error between the given kernel and its interpolant with respect to one of its arguments.

For the predictive variance $\sigma_N^2(\boldsymbol{x}) := K_N(\boldsymbol{x}, \boldsymbol{x})$ this means

$$\begin{aligned}\sigma_N^2(\boldsymbol{x}) &= \sup_{\|f\|_{\mathcal{H}}=1} |f(\boldsymbol{x}) - I_X f(\boldsymbol{x})|^2 \\ &= K(\boldsymbol{x}, \boldsymbol{x}) - 2K(\boldsymbol{x}, X)\boldsymbol{u}(\boldsymbol{x}) + \boldsymbol{u}(\boldsymbol{x})^{\mathrm{T}} K(X,X)\boldsymbol{u}(\boldsymbol{x}),\end{aligned}$$

where the vector $\boldsymbol{u}(\boldsymbol{x}) = (u_1(\boldsymbol{x}), \dots, u_N(\boldsymbol{x}))^{\mathrm{T}} \in \mathbb{R}^N$ contains the cardinal or Lagrange functions $u_j \in V_X$ evaluated at $\boldsymbol{x}$. They satisfy $u_j(\boldsymbol{x}_i) = \delta_{ij}$.

The connection between Gaussian processes and RKHS and positive definite kernels in the situation of noisy data is summarized in the next theorem. It is a specific version of the *representer theorem* from machine learning, see Schölkopf and Smola (2002); Steinwart and Christmann (2008).

Theorem 3 *Let $K : \mathcal{D}\times\mathcal{D} \to \mathbb{R}$ be a positive definite kernel and let $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ consist of pairwise distinct points. Define the N-dimensional approximation space $V_X = \operatorname{span}\{K(\cdot, \boldsymbol{x}_j) : 1 \le j \le N\}$ and let $\mathcal{H}$ be the RKHS of the kernel K.*

1. *Given $\lambda > 0$ and $\boldsymbol{y} \in \mathbb{R}^N$, there is a unique function $s_{\lambda,\boldsymbol{y}} \in V_X$ solving*

$$\min \left\{ \sum_{j=1}^{N} |s(\boldsymbol{x}_j) - y_j|^2 + \lambda \|s\|_{\mathcal{H}}^2 : s \in \mathcal{H} \right\}.$$

   This defines an approximation operator $Q_{X,\lambda} : \mathbb{R}^N \to V_X$, $Q_{X,\lambda}\boldsymbol{y} := s_{\lambda,\boldsymbol{y}}$ and also an operator $Q_{X,\lambda} : C(\mathcal{D}) \to V_X$ by setting $Q_{X,\lambda} f := Q_{X,\lambda}(f|X)$.

2. *If $m_{N,\sigma}$ is the predictive mean from (3) with data $\boldsymbol{y} \in \mathbb{R}^N$ and prior mean m and covariance kernel K, then $m_{N,\sigma} = m - Q_{X,\sigma^2} m + Q_{X,\sigma^2}\boldsymbol{y}$.*

Again, we see the connection between the no-noise and the noisy case, as we have $Q_{X,0} = I_X$. Moreover, we can use the approximation operator also for no-noise data, and the interpolation operator for noisy data with either known or unknown distribution.

The connection between the mean of the predictive or posterior Gaussian process and kernel based interpolation and approximation, leads to error estimates typically of the form

$$\|f(\boldsymbol{x}) - m_N(\boldsymbol{x})\|_{L_\infty(\mathcal{D})} \le C h_{X,\mathcal{D}}^{\tau - d/2} \|f\|_{\mathcal{H}},$$

where $h_{X,\mathcal{D}}$ is the so-called fill distance of X in $\mathcal{D}$, defined as

$$h_{X,\mathcal{D}} = \sup_{\boldsymbol{x}\in\mathcal{D}} \min_{1\le j\le N} \|\boldsymbol{x} - \boldsymbol{x}_j\|_2. \tag{5}$$

Here the data-giving function f is assumed to be in the Sobolev space $\mathcal{H} = H^\tau(\mathcal{D})$, $\mathcal{D}$ is supposed to have a Lipschitz boundary and K is supposed to be the reproducing kernel of $H^\tau(\mathcal{D})$. More general estimates that involve the error and relax the condition on the smoothness of f can be found in Wynne et al. (2021). Moreover, as pointed out in Scheuerer et al. (2013), the variance σ_N is the *power function* $P_{X,K}$ in approximation theory. Thus, a bound of the form $P_{X,K}(x) \le C h_{X,\mathcal{D}}^{\tau-d/2}$ and the Chebyshev inequality immediately leads to the following result, which measures the deviation of a realisation from its mean.

Theorem 4 *Assume that the covariance kernel K is the reproducing kernel of a Sobolev space $H^\tau(\mathcal{D})$ with $\tau > d/2$. Then, there are constants $C > 0$ and $h_0 > 0$ such that for all $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ with $h_{X,\mathcal{D}} \le h_0$, all $f \in H^\tau(\mathcal{D})$ and every $\epsilon \in (0, 1)$, the bound*

$$\mathbb{P}\left(|f_{N,0}(\boldsymbol{x}) - m_N(\boldsymbol{x})| \ge \epsilon\right) \le C \frac{h_{X,\mathcal{D}}^{2\tau-d}}{\epsilon^2}$$

holds.

3. Vector-Valued Function Approximation

In this section, we want to generalize the concept of Gaussian processes to Hilbert-space valued functions. We are particularly interested in vector-valued Gaussian processes which model physical vector fields from fluid dynamics and electromagnetics. Thus, we will introduce divergence-free or rotation-free Gaussian processes.

After that we will discuss generalizations of positive definite kernels and reproducing kernel Hilbert spaces to the situation of Hilbert-space valued functions and derive the connection to Gaussian processes.

3.1 Generalized Gaussian Processes

The generalization of Gaussian processes to Hilbert-space valued functions is, more or less, straightforward, see also (Adler, 1990, Ch. 1, Sec. 7, p. 35) for separable Banach-space-valued Gaussian processes.

We will use the following notation. Let W be a Hilbert space and let $\mathcal{L}(W)$ be the space of bounded and linear operators $A : W \to W$. On W we can, as usual, define the Borel σ-algebra $\mathcal{B}(W)$ which is the smallest σ-algebra on W which contains the open sets of W. Measurability of functions mapping into W will always be with respect to $(W, \mathcal{B}(W))$. Next, for $N \in \mathbb{N}$ we let $W^N = W \times \cdots \times W$ be the N-fold Cartesian product of W, which is equipped with the Borel σ-algebra $\mathcal{B}(W^N) = \mathcal{B}(W) \times \cdots \times \mathcal{B}(W)$.

Definition 5 *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and let W be a separable Hilbert space with orthonormal basis $\{\boldsymbol{e}_j\}_{j\in\mathbb{N}}$. A measurable function $\boldsymbol{g} : \Omega \to W$ has a* Gaussian or normal distribution *if the family $(\langle \boldsymbol{g}, \boldsymbol{e}_k\rangle_W)_{k\in\mathbb{N}}$ is Gaussian in the sense that finitely many of the Fourier coefficients have a joint normal distribution. For a given $\mathcal{D} \subseteq \mathbb{R}^d$, a function $\boldsymbol{g} : \Omega \times \mathcal{D} \to W$ describes a* generalized or W-valued *Gaussian process if for any finite set of pairwise distinct points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$, the random variable $\boldsymbol{G} := (\boldsymbol{g}(\cdot, \boldsymbol{x}_1), \ldots, \boldsymbol{g}(\cdot, \boldsymbol{x}_N)) : \Omega \to W^N$ is jointly normally distributed, i.e. if for any finite index set $J \subseteq \mathbb{N}$, the family $(\langle \boldsymbol{g}(\cdot, \boldsymbol{x}_i), \boldsymbol{e}_k\rangle_W)_{1\le i\le N, k\in J}$ has a joint normal distribution.*

In the case $W = \mathbb{R}$ the above definition reduces the above concept of a normal distribution and a generalized Gaussian process to the classical concept.

The distribution of a W-valued Gaussian random variable $\boldsymbol{g} : \Omega \to W$ is uniquely determined by its mean $\boldsymbol{m}(\boldsymbol{g})$ and covariance matrix $\Sigma(\boldsymbol{g})$, which take the form

$$\begin{aligned} m(\boldsymbol{g})_k &:= \mathbb{E}[\langle \boldsymbol{g}, \boldsymbol{e}_k\rangle_W], \qquad k \in \mathbb{N}, \\ \Sigma(\boldsymbol{g})_{k\ell} &:= \mathbb{E}[(\langle \boldsymbol{g}, \boldsymbol{e}_k\rangle_W - m(\boldsymbol{g})_k)(\langle \boldsymbol{g}, \boldsymbol{e}_\ell\rangle_W - m(\boldsymbol{g})_\ell)], \qquad k, \ell \in \mathbb{N}. \end{aligned}$$

The same is true for a W-valued Gaussian process g, where mean $\boldsymbol{m} : \mathcal{D} \to W$ and covariance $\boldsymbol{K} : \mathcal{D} \times \mathcal{D} \to \mathcal{L}(W)$ are now determined by

$$\begin{aligned} \boldsymbol{m}(\boldsymbol{x}) &:= \boldsymbol{m}(\boldsymbol{g}(\cdot, \boldsymbol{x})), \\ \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')_{k\ell} &:= \mathbb{E}[(\langle \boldsymbol{g}(\cdot, \boldsymbol{x}), \boldsymbol{e}_k\rangle_W - m(\boldsymbol{x})_k)(\langle \boldsymbol{g}(\cdot, \boldsymbol{x}'), \boldsymbol{e}_\ell\rangle_W - m(\boldsymbol{x}')_\ell)]. \end{aligned}$$

We can interpret the kernel indeed as a map $K : \mathcal{D} \times \mathcal{D} \to \mathcal{L}(W)$. Any element $\boldsymbol{v} \in W$ can uniquely be written as $\boldsymbol{v} = \sum \langle \boldsymbol{v}, \boldsymbol{e}_k\rangle_W \boldsymbol{e}_k$ with Fourier coefficients $(\langle \boldsymbol{v}, \boldsymbol{e}_k\rangle_W)_{k\in\mathbb{N}}$. Hence, for such an element we can define $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')\boldsymbol{v}$ by

$$\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')\boldsymbol{v} := \sum_{k\in\mathbb{N}}\left(\sum_{\ell\in\mathbb{N}}\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')_{k\ell}\langle \boldsymbol{v},\boldsymbol{e}_\ell\rangle_W\right)\boldsymbol{e}_k.$$

We are particularly interested in the case of vector-valued functions $\boldsymbol{g}:\mathcal{D}\to\mathbb{R}^n$, i.e. where $W=\mathbb{R}^n$, and here our main interest will be in the case $d=n$ such that the dimension of domain and range agree. But for the time being, we keep the possibility of having different dimensions. In any case, we can use the standard basis vectors $\boldsymbol{e}_1,\dots,\boldsymbol{e}_n\in\mathbb{R}^n$ such that the above general theory reduces to the following definition.

Definition 6 *Given a probability space $(\Omega,\mathcal{A},\mathbb{P})$, a finite number of vector-valued random variables $\boldsymbol{g}_i=(g_1^{(i)},\dots,g_n^{(i)})^{\mathrm{T}}:\Omega\to\mathbb{R}^n$, $1\le i\le N$, have a* joint Gaussian distribution *if the random variable $\boldsymbol{G}:=(\boldsymbol{g}_1,\dots,\boldsymbol{g}_N):\Omega\to\mathbb{R}^{n\times N}$ is uniquely determined by its matrix-valued mean*

$$m_k^{(i)}=\mathbb{E}(g_k^{(i)}),\qquad 1\le i\le N, 1\le k\le n,$$

and its tensor-valued covariance kernel

$$\Sigma_{k\ell}^{(ij)}=\mathbb{E}[(g_k^{(i)}-m_k^{(i)})(g_\ell^{(j)}-m_\ell^{(j)})],\qquad 1\le i,j\le N,\quad 1\le k,\ell\le n.$$

For $\mathcal{D}\subseteq\mathbb{R}^d$, we will call a vector-valued random process $\boldsymbol{g}:\Omega\times\mathcal{D}\to\mathbb{R}^n$ a Gaussian process *if for every $N\in\mathbb{N}$ and every $X=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N\}\subseteq\mathcal{D}$, the vector-valued variables $\boldsymbol{g}_j=\boldsymbol{g}(\cdot,\boldsymbol{x}_j)$, $1\le j\le N$ have a joint Gaussian distribution. This is written as $\boldsymbol{g}\sim\mathrm{GP}(\boldsymbol{m},\boldsymbol{K})$, where the vector-valued function $\boldsymbol{m}:\mathcal{D}\to\mathbb{R}^n$ gives the mean function and $K:\mathcal{D}\times\mathcal{D}\to\mathbb{R}^{n\times n}$ is the matrix-valued covariance kernel, satisfying*

$$\boldsymbol{m}(\boldsymbol{x})=\mathbb{E}[\boldsymbol{g}(\cdot,\boldsymbol{x})],\qquad \boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')=\mathbb{E}[(\boldsymbol{g}(\cdot,\boldsymbol{x})-\boldsymbol{m}(\boldsymbol{x}))(\boldsymbol{g}(\cdot,\boldsymbol{x}')-\boldsymbol{m}(\boldsymbol{x}'))^{\mathrm{T}}].$$

Obviously, we have for fixed i,j the relation $\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j)=[\Sigma_{k\ell}^{(ij)}:1\le k,\ell\le n]$.

We will also need a vector-valued version of the Karhunen-Loève expansion. To this end, let $L_2(\mathcal{D})^n=\{\boldsymbol{w}=(w_1,\dots,w_n):w_k\in L_2(\mathcal{D})\}$ be the space of all vector-valued square-integrable functions with inner product $\langle\boldsymbol{u},\boldsymbol{w}\rangle_{L_2(\mathcal{D})^n}:=\sum_{k=1}^n\langle u_k,w_k\rangle_{L_2(\mathcal{D})}$.

Noting that our covariance kernel $\boldsymbol{K}:\mathcal{D}\to\mathcal{D}\to\mathbb{R}^{n\times n}$ of the Gaussian process $\boldsymbol{g}:\Omega\times\mathcal{D}\to\mathbb{R}^n$ is not only continuous but also symmetric in the sense that $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')=\boldsymbol{K}(\boldsymbol{x}',\boldsymbol{x})^{\mathrm{T}}$, we see that the integral operator $A:L_2(\mathcal{D})^n\to L_2(\mathcal{D})^n$, given by

$$A\boldsymbol{w}(\boldsymbol{x}):=\int_{\mathcal{D}}\boldsymbol{K}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{w}(\boldsymbol{y})d\boldsymbol{y} \tag{6}$$

is a self-adjoint and compact operator. The operator is also positive, as we have, assuming for simplicity a centered Gaussian process,

$$\begin{aligned}
\langle A\boldsymbol{w},\boldsymbol{w}\rangle_{L_2(\mathcal{D})^n} &= \sum_{k,\ell=1}^n\int_{\mathcal{D}}\int_{\mathcal{D}}\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')_{k\ell}w_k(\boldsymbol{x})w_\ell(\boldsymbol{x}')d\boldsymbol{x}d\boldsymbol{x}'\\
&= \sum_{k,\ell=1}^n\int_{\mathcal{D}}\int_{\mathcal{D}}\int_{\Omega}g_k(\omega,\boldsymbol{x})g_\ell(\omega,\boldsymbol{x}')d\mathbb{P}(\omega)w_k(\boldsymbol{x})w_\ell(\boldsymbol{x}')d\boldsymbol{x}d\boldsymbol{x}'\\
&= \int_{\Omega}\left|\int_{\mathcal{D}}\sum_{k=1}^n g_k(\omega,\boldsymbol{x})w_k(\boldsymbol{x})d\boldsymbol{x}\right|^2d\mathbb{P}(\omega)\ge 0.
\end{aligned}$$

Thus, the spectral theorem for compact, self-adjoint and positive operators gives us an orthonormal basis of eigenfunctions of A with corresponding nonnegative eigenvalues. As we are only interested in positive eigenvalues, we can proceed as in the scalar-valued case to derive Mercer's representation of the kernel and the Karhunen-Loève expansion for vector-valued Gaussian processes.

Proposition 7 *Let $\mathcal{D} \subseteq \mathbb{R}^d$ be compact, and let $\boldsymbol{g} : \Omega \times \mathcal{D} \to \mathbb{R}^n$ be an integrable, centered Gaussian process with continuous covariance kernel $\boldsymbol{K} : \mathcal{D}\times\mathcal{D} \to \mathbb{R}^{n\times n}$. Let $\{\boldsymbol{\phi}_k\} \subseteq L_2(\mathcal{D})^d$ be an orthonormal subset of $L_2(\mathcal{D})^d$ consisting of eigenfunctions of the integral operator associated to the kernel with corresponding positive eigenvalues $\{\lambda_k\}$. Then, the kernel has the* Mercer representation

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{k=1}^{\infty} \lambda_k \boldsymbol{\phi}_k(\boldsymbol{x}) \boldsymbol{\phi}_k(\boldsymbol{x}')^{\mathrm{T}}, \qquad \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}.$$

The series converges uniformly and absolutely. Moreover, the Gaussian process has the Karhunen-Loève *representation*

$$\boldsymbol{g}(\omega, \boldsymbol{x}) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} \xi_k(\omega) \boldsymbol{\phi}_k(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{D}, \omega \in \Omega, \tag{7}$$

where the convergence with respect to ω is in $L_2(\Omega, \mathbb{P})$ and with respect to $\boldsymbol{x}$ is uniform. The random variables $\xi_k : \Omega \to \mathbb{R}$ are given in terms of g by

$$\xi_k(\omega) = \frac{1}{\sqrt{\lambda_k}} \int_{\mathcal{D}} \langle \boldsymbol{g}(\omega, \boldsymbol{x}), \boldsymbol{\phi}_k(\boldsymbol{x}) \rangle_2 d\boldsymbol{x}.$$

With the notation above, we can now use vector-valued Gaussian processes to form approximations to vector-valued functions $\boldsymbol{v} : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}^n$. As in the scalar-valued case we start with a prior model $\boldsymbol{v}_0 : \Omega \times \mathcal{D} \to \mathbb{R}^n$ with mean function $\boldsymbol{m} : \mathcal{D} \to \mathbb{R}^n$, usually assumed to be zero, and a matrix-valued covariance kernel $\boldsymbol{K} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^{n\times n}$. Next, assuming again that we have N observations at design points $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ in the form $\boldsymbol{y}_j = \boldsymbol{v}(\boldsymbol{x}_j) + \boldsymbol{\epsilon}_j$, $1 \le j \le N$, where the $\boldsymbol{\epsilon}_j$ are either zero (exact data) or follow a normal distribution with mean zero and variance $\sigma^2 I_n$ with $I_n \in \mathbb{R}^{n\times n}$ being the identity matrix and $\sigma^2 > 0$. In both cases, we can write the conditional distribution as a posterior Gaussian process $\boldsymbol{v}_{N,\sigma} \sim \mathrm{GP}(\boldsymbol{m}_{N,\sigma}, \boldsymbol{K}_{N,\sigma})$, $\sigma \ge 0$. To describe the mean $\boldsymbol{m}_{N,\sigma}$ and the covariance $\boldsymbol{K}_{N,\sigma}$ we will interpret $\boldsymbol{m}(X) := (\boldsymbol{m}(\boldsymbol{x}_1), \dots, \boldsymbol{m}(\boldsymbol{x}_N)) \in \mathbb{R}^{n\times N}$as a vector in $\mathbb{R}^{nN}$ rather than a matrix in $\mathbb{R}^{n\times N}$. In the same fashion, we will interpret the N $(n \times n)$-matrices $\boldsymbol{K}(\boldsymbol{x}, X) = (\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}_1), \dots, \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}_N))$ as a matrix from $\mathbb{R}^{n\times nN}$ and the fourth order tensor $\boldsymbol{K}(X, X) = (\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j))_{1\le i,j\le N}$ consisting of N^2 small $(n \times n)$ matrices as a matrix from $\mathbb{R}^{nN\times nN}$. Finally, we set $\boldsymbol{Y} = (\boldsymbol{y}_1^{\mathrm{T}}, \dots, \boldsymbol{y}_N^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{nN}$ and use $I \in \mathbb{R}^{nN\times nN}$ as the identity matrix.

Lemma 8 *With the notation above, the predictive mean and covariance of the posterior Gaussian process $\boldsymbol{v}_{N,\sigma}$ conditioned by the data $\boldsymbol{Y}$ and X are*

$$\begin{aligned} \boldsymbol{m}_{N,\sigma}(\boldsymbol{x}) &:= \boldsymbol{m}(\boldsymbol{x}) + \boldsymbol{K}(\boldsymbol{x}, X)(\boldsymbol{K}(X, X) + \sigma^2 I)^{-1} (\boldsymbol{Y} - \boldsymbol{m}(X)), \\ \boldsymbol{K}_{N,\sigma}(\boldsymbol{x}, \boldsymbol{x}') &:= \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{K}(\boldsymbol{x}, X)(\boldsymbol{K}(X, X) + \sigma^2 I)^{-1} \boldsymbol{K}(\boldsymbol{x}', X)^{\mathrm{T}}. \end{aligned}$$

In the case $\sigma = 0$, we will again write $\boldsymbol{m}_N$ and $\boldsymbol{K}_N$, respectively.

This is in accordance with scalar-valued Gaussian processes, and follows immediately from the standard formulas of conditional Gaussian processes.

3.2 Divergence- and Rotation-free Gaussian Processes

Having introduced the concept of vector-valued Gaussian processes, we can now proceed to discuss divergence- and rotation-free Gaussian processes. To this end, we recall that divergence and rotation of a continuously differentiable vector field $\boldsymbol{v} : \mathcal{D} \subseteq \mathbb{R}^d \to \mathbb{R}^d$ are defined by

$$\operatorname{div} \boldsymbol{v} = \nabla \cdot \boldsymbol{v} = \sum_{k=1}^{d} \frac{\partial v_k}{\partial x_k},$$

for any $d \in \mathbb{N}$ and, for $d = 2, 3$,

$$\operatorname{curl} \boldsymbol{v} = \nabla \times \boldsymbol{v} = \begin{cases} \left(\frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3}, \frac{\partial v_1}{\partial x_3} - \frac{\partial v_3}{\partial x_1}, \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \right)^{\mathrm{T}} & \text{for } d = 3 \\ \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} & \text{for } d = 2. \end{cases}$$

The vector field is called divergence-free or rotation-free if $\operatorname{div} \boldsymbol{v}(\boldsymbol{x}) = 0$ or $\operatorname{curl} \boldsymbol{v}(\boldsymbol{x}) = \boldsymbol{0}$, respectively.

We can now carry this concept over to vector-valued Gaussian processes.

Definition 9 *Let $\boldsymbol{g} : \Omega \times \mathcal{D} \to \mathbb{R}^d$ be a Gaussian process defined on a compact subset $\mathcal{D} \subseteq \mathbb{R}^d$ which is continuously differentiable with respect to the variable $\boldsymbol{x} \in \mathbb{R}^d$.*

- *The process is called* divergence-free *if it satisfies* $\operatorname{div} \boldsymbol{g}(\cdot, \boldsymbol{x}) = 0$ *for all* $\boldsymbol{x} \in \mathcal{D}$ *and* $\mathbb{P}$*-almost everywhere on* Ω*.*
- *The process is called* rotation-free or curl-free *if it satisfies* $\operatorname{curl} \boldsymbol{g}(\cdot, \boldsymbol{x}) = 0$ *for all* $\boldsymbol{x} \in \mathcal{D}$ *and* $\mathbb{P}$*-almost everywhere on* Ω*.*

In both cases, the differential operators act only with respect to the spatial variable $\boldsymbol{x} \in \mathcal{D}$.

The property of being divergence-free or rotation-free carries over from a Gaussian process to its mean and covariance kernel.

Lemma 10 *Let $\boldsymbol{g} : \Omega \times \mathcal{D} \to \mathbb{R}^d$ be a divergence-free Gaussian process. Then the mean $\boldsymbol{m}$ is also divergence-free, and the matrix-valued covariance kernel $\boldsymbol{K} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^{d \times d}$ has the following property. For a fixed $\boldsymbol{x}' \in \mathcal{D}$, each column of the map $\boldsymbol{K}(\cdot, \boldsymbol{x}')$ is in $C^1(\mathcal{D})^d$ and divergence-free.*

In the case of $d = 2, 3$ this statement holds in a similar way for a rotation-free Gaussian process.

Proof We will prove the result only for a divergence-free Gaussian process. We denote the components of $\boldsymbol{g}$ by g_k, $1 \leq k \leq d$. For the mean

$$\boldsymbol{m}(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{g}(\boldsymbol{x})] = \int_{\Omega} \boldsymbol{g}(\omega, \boldsymbol{x}) d\mathbb{P}(\omega),$$

we simply take the divergence under the integral, using the Leibniz theorem to justify differentiating under the integral sign.

For the covariance kernel, we first fix the column index $1 \le \ell \le d$. Then, we have for the k-th component of the ℓ-th column

$$\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')_{k\ell} = \int_\Omega (g_k(\omega,\boldsymbol{x}) - m_k(\boldsymbol{x}))\left(g_\ell(\omega,\boldsymbol{x}') - m_\ell(\boldsymbol{x}')\right) d\mathbb{P}(\omega).$$

That the right-hand side as a function of $\boldsymbol{x}$ is in $C^1(\mathcal{D})$ for fixed $\boldsymbol{x}'$ follows again from the Leibniz theorem. We can then conclude that

$$\begin{aligned}
\operatorname{div}_{\boldsymbol{x}} \boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')\boldsymbol{e}_\ell &= \sum_{k=1}^d \frac{\partial}{\partial x_k} \boldsymbol{K}(\boldsymbol{x},\boldsymbol{x}')_{k\ell} \\
&= \sum_{k=1}^d \frac{\partial}{\partial x_k} \int_\Omega (g_k(\omega,\boldsymbol{x}) - m_k(\boldsymbol{x}))\left(g_\ell(\omega,\boldsymbol{x}') - m_\ell(\boldsymbol{x}')\right) d\mathbb{P}(\omega) \\
&= \int_\Omega [\operatorname{div}_{\boldsymbol{x}} (\boldsymbol{g}(\omega,\boldsymbol{x}) - \boldsymbol{m}(\boldsymbol{x}))]\left(g_\ell(\omega,\boldsymbol{x}') - m_\ell(\boldsymbol{x}')\right) d\mathbb{P}(\omega) \\
&= 0,
\end{aligned}$$

where $\boldsymbol{e}_\ell$ is the ℓth unit vector in $\mathbb{R}^d$, using the fact that both $\boldsymbol{g}$ and $\boldsymbol{m}$ are divergence-free. ■

A Gaussian process is uniquely determined by its mean and covariance kernel. This remains obviously true for vector-valued Gaussian processes. As we have seen, a divergence-free Gaussian process has a divergence-free mean and covariance kernel. Hence, it is natural to ask the question whether it is enough for a Gaussian process to have a divergence-free mean and covariance kernel to be divergence-free, as well. The next theorem shows that this is indeed the case.

Theorem 11 *Let $\mathcal{D} \subseteq \mathbb{R}^d$ be compact. Let $\boldsymbol{g} : \Omega \times \mathcal{D} \to \mathbb{R}^d$ be a continuously differentiable vector-valued Gaussian process whose mean is divergence-free and whose covariance kernel has divergence-free columns. Then $\boldsymbol{g}$ is divergence-free.*

An analogous statement holds in the case $d = 2, 3$ for Gaussian processes with rotation-free mean and covariance.

Proof Again, we give the proof only for divergence-free case. Here, we will use the vector-valued Karhunen-Loève expansion from Proposition 7 with $n = d$. We first have a look at the eigenfunctions $\boldsymbol{\phi} \in L_2(\mathcal{D})^d$ of the integral operator A. From (6) we have

$$\lambda_k \boldsymbol{\phi}_k(\boldsymbol{x}) = \int_{\mathcal{D}} \boldsymbol{K}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{\phi}_k(\boldsymbol{y}) d\boldsymbol{y}.$$

On taking the divergence of each side and using the fact that the divergence of each column of $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{y})$ with respect to $\boldsymbol{x}$ is zero we see that the eigenfunctions ϕ_k are divergence-free. Thus, the partial sums of the expansion (7),

$$\sum_{k=1}^m \sqrt{\lambda_k}\xi_k(\omega)\boldsymbol{\phi}_k(\boldsymbol{x}),$$

have divergence zero, implying that the limit as $m \to \infty$ is also divergence-free. ■

.

3.3 Vector-Valued Reproducing Kernel Hilbert Spaces

As in the scalar-valued case, there is a strong connection to kernel-based approximation theory and reproducing kernel Hilbert spaces. We start this discussion by introducing both concepts for functions that are Hilbert-space-valued. After that, we look at the specific cases of vector-valued functions and then at divergence- and rotation-free functions. The concept of Hilbert-space-valued reproducing kernel Hilbert spaces was introduced in Pedrick (1957). It became popular in the approximation theory community, particularly in the context of approximating divergence- and rotation-free vector fields after the initial work in Narcowich and Ward (1994). Moreover, in Micchelli and Pontil (2005) they were most notably studied in the context of learning theory.

Let W again be a Hilbert space and $\mathcal{L}(W)$ the space of all bounded operators from W to W. As usual, for $A \in \mathcal{L}(W)$ the *adjoint operator* $A^* \in \mathcal{L}(W)$ is the unique operator satisfying $\langle Av, w\rangle_W = \langle v, A^*w\rangle_W$ for all $v, w \in W$.

Definition 12 1. *A continuous map* $\boldsymbol{K} : \mathcal{D} \times \mathcal{D} \to \mathcal{L}(W)$ *is called* positive definite *if it satisfies* $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')^* = \boldsymbol{K}(\boldsymbol{x}', \boldsymbol{x})$ *for all* $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$ *and*

$$\sum_{i,j=1}^{N} \langle w_i, \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) w_j\rangle_W > 0.$$

for all $N \in \mathbb{N}$*, all data sets* $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ *of pairwise distinct points, and all* $\boldsymbol{w} \in W^N \setminus \{\boldsymbol{0}\}$*.*

2. *A Hilbert space* $\mathcal{H} := \mathcal{H}(\mathcal{D}; W) = \{w : \mathcal{D} \to W\}$ *of* W*-valued functions is called a* reproducing kernel Hilbert space *(RKHS) if there is a function* $\boldsymbol{K} : \mathcal{D} \times \mathcal{D} \to \mathcal{L}(W)$ *with*

   (a) $\boldsymbol{K}(\cdot, \boldsymbol{x})w \in \mathcal{H}(\mathcal{D}; W)$ *for all* $\boldsymbol{x} \in \mathcal{D}$ *and all* $w \in W$*.*

   (b) $\langle v(\boldsymbol{x}), w\rangle_W = \langle v, \boldsymbol{K}(\cdot, \boldsymbol{x})w\rangle_{\mathcal{H}}$ *for all* $v \in \mathcal{H}(\mathcal{D}; W)$*, all* $\boldsymbol{x} \in \mathcal{D}$ *and all* $w \in W$*.*

   The function $\boldsymbol{K}$ *is called the* reproducing kernel *of* $\mathcal{H}(\mathcal{D}; W)$*.*

As in the case of Gaussian processes, this definition reduces to the classical definition of positive definite kernels and reproducing kernel Hilbert spaces for $W = \mathbb{R}$. While it is possible to draw connections between the generalized concepts of Gaussian processes, positive definite kernels and reproducing kernel Hilbert spaces, we restrict ourselves here again to the case of vector-valued processes.

The restriction of the above definition to this case yields the following concept of vector-valued RKHS and matrix-valued kernels.

Definition 13 1. *A continuous, matrix-valued map* $\boldsymbol{K} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^{n \times n}$ *is called* positive definite *if it satisfies* $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')^{\mathrm{T}} = \boldsymbol{K}(\boldsymbol{x}', \boldsymbol{x})$ *for all* $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$ *and*

$$\sum_{i,j=1}^{N} \boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) \boldsymbol{w}_j > 0. \tag{8}$$

for all $N \in \mathbb{N}$ and all data sets $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ of pairwise distinct points and all $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N \in \mathbb{R}^n$, not all being the zero vector.

2. *A Hilbert space $\mathcal{H} := \mathcal{H}(\mathcal{D}) = \{\boldsymbol{w} : \mathcal{D} \to \mathbb{R}^n\}$ of vector-valued functions is called a* reproducing kernel Hilbert space *(RKHS) if there is a function $\boldsymbol{K} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^{n\times n}$ with*

   (a) *$\boldsymbol{K}(\cdot, \boldsymbol{x})\boldsymbol{w} \in \mathcal{H}(\mathcal{D})$ for all $\boldsymbol{x} \in \mathcal{D}$ and all $\boldsymbol{w} \in \mathbb{R}^n$.*

   (b) *$\boldsymbol{v}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{w} = \langle \boldsymbol{v}, \boldsymbol{K}(\cdot, \boldsymbol{x})\boldsymbol{w}\rangle_{\mathcal{H}}$ for all $\boldsymbol{v} \in \mathcal{H}(\mathcal{D})$, all $\boldsymbol{x} \in \mathcal{D}$ and all $\boldsymbol{w} \in \mathbb{R}^n$.*

   The function $\boldsymbol{K}$ is called the reproducing kernel *of $\mathcal{H}(\mathcal{D})$.*

The following properties hold and are proven as in the scalar-valued case. The reproducing kernel of a vector-valued RKHS is unique and positive semi-definite, i.e. it satisfies (8) with ≥ 0 rather than > 0. The kernel is positive definite if point-evaluations are linearly independent. Moreover, for every positive definite kernel $\boldsymbol{K} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}^{n\times n}$ there is a unique vector-valued RKHS $\mathcal{H} = \mathcal{H}(\mathcal{D}) \subseteq C(\mathcal{D})^n$, in which $\boldsymbol{K}$ is the reproducing kernel. It is, as in the scalar-valued case, constructed by forming the closure of the linear space

$$F_{\boldsymbol{K}}(\mathcal{D}) = \left\{ \sum_{j=1}^{M} \boldsymbol{K}(\cdot, \boldsymbol{z}_j)\boldsymbol{\beta}_j : \boldsymbol{z}_j \in \mathcal{D}, \boldsymbol{\beta}_j \in \mathbb{R}^n, M \in \mathbb{N} \right\} \tag{9}$$

with respect to the norm which is induced by the inner product

$$\langle \boldsymbol{K}(\cdot, \boldsymbol{x})\boldsymbol{\alpha}, \boldsymbol{K}(\cdot, \boldsymbol{x}')\boldsymbol{\beta}\rangle_{\boldsymbol{K}} := \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}')\boldsymbol{\beta}. \tag{10}$$

Details can, for example, be found in Fuselier (2008b).

We will now describe the connection between vector-valued Gaussian processes and vector-valued RKHS and matrix-valued positive definite kernels. As in the scalar-valued case, we start with the no-noise situation, meaning that our vector-valued observations at $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ are given by $\boldsymbol{y}_j = \boldsymbol{v}(\boldsymbol{x}_j)$, $1 \leq j \leq N$ and thus we have $\sigma = 0$ in Lemma 8. The first three statements of the following theorem are essentially proven as in the scalar-valued case but only a proof for the first one can be found in Micchelli and Pontil (2005) in the even more general situation of Hilbert-space-valued functions. However, we have included its proof as it introduces the notation we need for the rest of the proof. The fourth statement is a new, somewhat unexpected generalization of a similar result in the scalar-valued case.

Theorem 14 *Let $\boldsymbol{K} : \mathcal{D}\times\mathcal{D} \to \mathbb{R}^{n\times n}$ be a positive definite kernel and $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ consist of pairwise distinct points. Define the finite dimensional approximation space $V_X = \{\sum_{j=1}^N \boldsymbol{K}(\cdot, \boldsymbol{x}_j)\boldsymbol{\beta}_j : \boldsymbol{\beta}_j \in \mathbb{R}^n, 1 \leq j \leq N\}$ and let $\mathcal{H} = \mathcal{H}(\mathcal{D})$ be the RKHS of the kernel $\boldsymbol{K}$.*

1. *For every set of training observations $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\} \subseteq \mathbb{R}^n$ there is a unique function $\boldsymbol{s}_{0,\boldsymbol{Y}} \in V_X$ interpolating the data, i.e. satisfying $\boldsymbol{s}_{0,\boldsymbol{Y}}(\boldsymbol{x}_i) = \boldsymbol{y}_i$, $1 \leq i \leq N$. This defines an interpolation operator $\boldsymbol{I}_X : \mathbb{R}^{N\times n} \to V_X$, $\boldsymbol{I}_X\boldsymbol{v} = \boldsymbol{s}_{0,\boldsymbol{Y}}$ and also an operator $\boldsymbol{I}_X : C(\mathcal{D})^n \to V_X$ by setting $\boldsymbol{I}_X\boldsymbol{v} = \boldsymbol{I}_X(\boldsymbol{v}|X)$.*

2. *The function* $\boldsymbol{s}_{0,\boldsymbol{Y}} \in V_X$ *is the solution of*

$$\min\left\{\|\boldsymbol{s}\|_{\mathcal{H}} : \boldsymbol{s} \in \mathcal{H} \text{ with } \boldsymbol{s}(\boldsymbol{x}_i) = \boldsymbol{y}_i, 1 \le i \le N\right\}.$$

3. *If* $\boldsymbol{m}_N$ *is the predictive mean from Lemma 8 in the no-noise case* $\boldsymbol{y}_j = \boldsymbol{v}(\boldsymbol{x}_j)$, $1 \le j \le N$, *i.e. with* $\sigma = 0$, *then*

$$\boldsymbol{m}_N = \boldsymbol{m} + \boldsymbol{I}_X \boldsymbol{v} - \boldsymbol{I}_X \boldsymbol{m}. \tag{11}$$

4. *If* $\boldsymbol{K}_N$ *is the predictive covariance kernel from Lemma 8 in the no-noise case, i.e. with* $\sigma = 0$, *then*

$$\boldsymbol{K}_N(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{I}_X \boldsymbol{K}(\cdot, \boldsymbol{x}')(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{I}_X \boldsymbol{K}(\cdot, \boldsymbol{x})(\boldsymbol{x}'),$$

where the interpolation operator $\boldsymbol{I}_X$ *is now applied to each column of the matrix-valued kernel. This again means that the predictive covariance kernel is the error between the given kernel and its interpolant with respect to one of its arguments.*

For the predictive variance $\boldsymbol{\sigma}_N^2(\boldsymbol{x}) := \boldsymbol{K}_N(\boldsymbol{x}, \boldsymbol{x}) \in \mathbb{R}^{n \times n}$ *this means*

$$\|\boldsymbol{\sigma}_N^2(\boldsymbol{x})\|_2 = \lambda_{\max}(\boldsymbol{\sigma}_N^2(\boldsymbol{x})) = \sup_{\|\boldsymbol{v}\|_{\mathcal{H}}=1} \|\boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{I}_X \boldsymbol{v}(\boldsymbol{x})\|_2^2,$$

where $\lambda_{\max}(\boldsymbol{\sigma}_N^2(\boldsymbol{x}))$ *denotes the maximum eigenvalue of the positive and symmetric matrix* $\boldsymbol{K}_N(\boldsymbol{x}, \boldsymbol{x}) \in \mathbb{R}^{n \times n}$.

Proof If we write the interpolant in the form

$$\boldsymbol{s}_{0,\boldsymbol{Y}} = \sum_{j=1}^{N} \boldsymbol{K}(\cdot, \boldsymbol{x}_j)\boldsymbol{\alpha}_j$$

with $\boldsymbol{\alpha}_j \in \mathbb{R}^n$, then the interpolation conditions and the notational convention immediately before Lemma 8 show $\boldsymbol{\alpha} = \boldsymbol{K}(X, X)^{-1}\boldsymbol{Y}$ and thus

$$\boldsymbol{s}_{0,\boldsymbol{Y}}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x}, X)\boldsymbol{\alpha} = \boldsymbol{K}(\boldsymbol{x}, X)\boldsymbol{K}(X, X)^{-1}\boldsymbol{Y},$$

proving the first point, from which also the third point follows immediately. For the second point, we proceed similarly to the scalar-valued case, see for example Wendland (2005). We first note that if $\boldsymbol{v} \in \mathcal{H}$ is any function with $\boldsymbol{v}(\boldsymbol{x}_j) = \boldsymbol{0}$, $1 \le j \le N$, then

$$\langle \boldsymbol{v}, \boldsymbol{s}_{0,\boldsymbol{Y}} \rangle_{\mathcal{H}} = \sum_{j=1}^{N} \langle \boldsymbol{v}, \boldsymbol{K}(\cdot, \boldsymbol{x}_j)\boldsymbol{\alpha}_j \rangle_{\mathcal{H}} = \sum_{j=1}^{N} \boldsymbol{v}(\boldsymbol{x}_j)^{\mathrm{T}} \boldsymbol{\alpha}_j = 0,$$

using the reproducing kernel property. Thus, for any other interpolant $\boldsymbol{s} \in \mathcal{H}$, $\boldsymbol{s}(\boldsymbol{x}_j) = \boldsymbol{y}_j$, we can conclude

$$\|\boldsymbol{s}_{0,\boldsymbol{Y}}\|_{\mathcal{H}}^2 = \langle \boldsymbol{s}_{0,\boldsymbol{Y}}, \boldsymbol{s}_{0,\boldsymbol{Y}} - \boldsymbol{s} + \boldsymbol{s} \rangle_{\mathcal{H}} = \langle \boldsymbol{s}_{0,\boldsymbol{Y}}, \boldsymbol{s} \rangle_{\mathcal{H}} \le \|\boldsymbol{s}_{0,\boldsymbol{Y}}\|_{\mathcal{H}} \|\boldsymbol{s}\|_{\mathcal{H}}.$$

Dividing by $\|\boldsymbol{s}_{0,\boldsymbol{Y}}\|_{\mathcal{H}}$ shows the second statement. For the fourth statement, we first note that for $1 \le j \le N$ and $1 \le \ell \le n$, there is a unique function $\boldsymbol{u}_j^{(\ell)} \in V_X$ satisfying $\boldsymbol{u}_j^{(\ell)}(\boldsymbol{x}_i) = \delta_{ij}\boldsymbol{e}_\ell$ for $1 \le i, j \le N$ and $1 \le \ell \le n$, where $\boldsymbol{e}_\ell$ is the ℓ-th unit vector in $\mathbb{R}^n$. Thus, the matrix-valued function $\boldsymbol{u}_j := (\boldsymbol{u}_j^{(1)}, \ldots, \boldsymbol{u}_j^{(n)}) : \mathcal{D} \to \mathbb{R}^{n\times n}$, satisfies $\boldsymbol{u}_j(\boldsymbol{x}_i) = \delta_{ij} I_n$, where I_n is the identity matrix in $\mathbb{R}^{n\times n}$. The functions $\boldsymbol{u}_j$ belong to the finite dimensional space $\operatorname{span}\{\boldsymbol{K}(\cdot, \boldsymbol{x}_j) : 1 \le j \le N\}$. Moreover, the functions $\boldsymbol{u}_j^{\mathrm{T}}$ also satisfy $\boldsymbol{u}_j(\boldsymbol{x}_i)^{\mathrm{T}} = \delta_{ij} I_n$ and belong to the space $\operatorname{span}\{\boldsymbol{K}(\boldsymbol{x}_j, \cdot) : 1 \le j \le N\}$. Hence, they are, for a fixed $\boldsymbol{x} \in \mathcal{D}$, the unique solution of the linear matrix system

$$\sum_{j=1}^{N} \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)\boldsymbol{u}_j(\boldsymbol{x})^{\mathrm{T}} = \boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}), \qquad 1 \le i \le N.$$

With this, we can, again for a fixed $\boldsymbol{x} \in \mathcal{D}$ and $\boldsymbol{\alpha} \in \mathbb{R}^n$, express the error as

$$\begin{aligned}
\boldsymbol{\alpha}^{\mathrm{T}}[\boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{I}_X \boldsymbol{v}(\boldsymbol{x})] &= \boldsymbol{\alpha}^{\mathrm{T}}\left[\boldsymbol{v}(\boldsymbol{x}) - \sum_{j=1}^{N} \boldsymbol{u}_j(\boldsymbol{x})\boldsymbol{v}(\boldsymbol{x}_j)\right] \\
&= \left\langle \boldsymbol{v}, \left[\boldsymbol{K}(\cdot, \boldsymbol{x})\boldsymbol{\alpha} - \sum_{j=1}^{N} \boldsymbol{K}(\cdot, \boldsymbol{x}_j)\boldsymbol{u}_j(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\alpha}\right]\right\rangle_{\mathcal{H}} \\
&\le \|\boldsymbol{v}\|_{\mathcal{H}} P_{\boldsymbol{K},X}(\boldsymbol{x}, \boldsymbol{\alpha}) \qquad (12)
\end{aligned}$$

with

$$\begin{aligned}
[P_{\boldsymbol{K},X}(\boldsymbol{x}, \boldsymbol{\alpha})]^2 &= \left\| K(\cdot, \boldsymbol{x})\boldsymbol{\alpha} - \sum_{j=1}^{N} \boldsymbol{K}(\cdot, \boldsymbol{x}_j)\boldsymbol{u}_j(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\alpha} \right\|_{\mathcal{H}}^2 \\
&= \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^{\mathrm{T}} \sum_{j=1}^{N} \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}_j)\boldsymbol{u}_j(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\alpha} \\
&\quad + \boldsymbol{\alpha}^{\mathrm{T}} \sum_{j,k=1}^{N} \boldsymbol{u}_k(\boldsymbol{x})\boldsymbol{K}(\boldsymbol{x}_k, \boldsymbol{x}_j)\boldsymbol{u}_j(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^{\mathrm{T}}\left[\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) - \sum_{j=1}^{N} \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}_j)\boldsymbol{u}_j(\boldsymbol{x})^{\mathrm{T}}\right]\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^{\mathrm{T}}\left[\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{K}(\boldsymbol{x}, X)\boldsymbol{K}(X, X)^{-1}\boldsymbol{K}(X, \boldsymbol{x})\right]\boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^{T}\boldsymbol{K}_N(\boldsymbol{x}, \boldsymbol{x})\boldsymbol{\alpha}.
\end{aligned}$$

Obviously, for a fixed $\boldsymbol{x} \in \mathcal{D}$ we have equality in (12) for

$$\boldsymbol{v} = \boldsymbol{K}(\cdot, \boldsymbol{x})\boldsymbol{\alpha} - \sum_{j=1}^{N} \boldsymbol{K}(\cdot, \boldsymbol{x}_j)\boldsymbol{u}_j(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\alpha},$$

which implies

$$\sup_{\substack{\boldsymbol{v}\in\mathcal{H}\\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{|\boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{I}_X\boldsymbol{v}(\boldsymbol{x}))|^2}{\|\boldsymbol{v}\|_{\mathcal{H}}^2} = P_{\boldsymbol{K},X}(\boldsymbol{x},\boldsymbol{\alpha})^2 = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x})\boldsymbol{\alpha}.$$

This now leads to

$$\begin{aligned}
\lambda_{\max}(\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x})) &= \sup_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n\\ \boldsymbol{\alpha}\neq\boldsymbol{0}}} \frac{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_N(\boldsymbol{x},\boldsymbol{x})\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|_2^2}\\
&= \sup_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n\\ \boldsymbol{\alpha}\neq\boldsymbol{0}}} \sup_{\substack{\boldsymbol{v}\in\mathcal{H}\\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{|\langle \boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{I}_X\boldsymbol{v}(\boldsymbol{x}), \boldsymbol{\alpha}\rangle_2|^2}{\|\boldsymbol{v}\|_{\mathcal{H}}^2\|\boldsymbol{\alpha}\|_2^2}\\
&= \sup_{\substack{\boldsymbol{v}\in\mathcal{H}\\ \boldsymbol{v}\neq\boldsymbol{0}}} \sup_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n\\ \boldsymbol{\alpha}\neq\boldsymbol{0}}} \frac{|\langle \boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{I}_X\boldsymbol{v}(\boldsymbol{x}), \boldsymbol{\alpha}\rangle_2|^2}{\|\boldsymbol{v}\|_{\mathcal{H}}^2\|\boldsymbol{\alpha}\|_2^2}\\
&= \sup_{\substack{\boldsymbol{v}\in\mathcal{H}\\ \boldsymbol{v}\neq\boldsymbol{0}}} \frac{\|\boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{I}_X\boldsymbol{v}(\boldsymbol{x})\|_2^2}{\|\boldsymbol{v}\|_{\mathcal{H}}^2}
\end{aligned}$$

■

The connection between Gaussian processes and positive definite kernels in the situation of noisy data is the following straight-forward generalisation of the representer theorem from machine learning, Theorem 3, to the vector-valued situation.

Theorem 15 *Let $\boldsymbol{K}:\mathcal{D}\times\mathcal{D}\to\mathbb{R}^{n\times n}$ be a positive definite kernel and $X=\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N\}\subseteq\mathcal{D}$ consist of pairwise distinct points. Define the finite dimensional approximation space $V_X=\{\sum_{j=1}^N \boldsymbol{K}(\cdot,\boldsymbol{x}_j)\boldsymbol{\beta}_j : \boldsymbol{\beta}_j\in\mathbb{R}^n, 1\le j\le N\}$ and let $\mathcal{H}=\mathcal{H}(\mathcal{D};\mathbb{R}^n)$ be the RKHS of the kernel $\boldsymbol{K}$.*

1. *Given $\lambda>0$ and $Y=\{\boldsymbol{y}_1,\dots,\boldsymbol{y}_N\}\subseteq\mathbb{R}^n$, there is a unique function $\boldsymbol{s}_{\lambda,\boldsymbol{Y}}\in V_X$ solving*

$$\min\left\{\sum_{j=1}^N \|\boldsymbol{s}(\boldsymbol{x}_j)-\boldsymbol{y}_j\|_2^2 + \lambda\|\boldsymbol{s}\|_{\mathcal{H}}^2 : \boldsymbol{s}\in\mathcal{H}\right\}.$$

   This defines an approximation operator $\boldsymbol{Q}_{X,\lambda}:\mathbb{R}^{N\times n}\to V_X$, $\boldsymbol{Q}_{X,\lambda}\boldsymbol{Y}:=\boldsymbol{s}_{\lambda,\boldsymbol{Y}}$ and also an operator $\boldsymbol{Q}_{X,\lambda}:C(\mathcal{D})^n\to V_X$ by setting $\boldsymbol{Q}_{X,\lambda}\boldsymbol{v}:=\boldsymbol{Q}_{X,\lambda}(\boldsymbol{v}|X)$.

2. *If $\boldsymbol{m}_{N,\sigma}$ is the predictive mean from (3) with data $\boldsymbol{Y}$ and prior mean $\boldsymbol{m}$ and covariance kernel $\boldsymbol{K}$, then $\boldsymbol{m}_{N,\sigma}=\boldsymbol{m}-\boldsymbol{Q}_{X,\sigma^2}\boldsymbol{m}+\boldsymbol{Q}_{X,\sigma^2}(\boldsymbol{Y})$.*

Proof The proof of the first statement is with mild modification the same proof as the one for the scalar-valued case. It is divided into two steps. First, we need to show that the minimum is attained for an element $\boldsymbol{s}\in V_X$. To see this, we note that V_X is a closed subspace of $\mathcal{H}$ and hence we can decompose $\mathcal{H}$ into a direct sum of V_X and its orthogonal complement. Hence, any $\boldsymbol{s}\in\mathcal{H}$ can be written as $\boldsymbol{s}=\boldsymbol{s}_1+\boldsymbol{s}_2$ with $\boldsymbol{s}_1\in V_X$ and

$\boldsymbol{s}_2 \in V_X^\perp$. As we have $\|\boldsymbol{s}\|_{\mathcal{H}}^2 = \|\boldsymbol{s}_1\|_{\mathcal{H}}^2 + \|\boldsymbol{s}_2\|_{\mathcal{H}}^2$ and $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{s}_2(\boldsymbol{x}_j) = \langle \boldsymbol{s}_2, \boldsymbol{K}(\cdot,\boldsymbol{x}_j)\boldsymbol{\alpha}\rangle_{\mathcal{H}} = 0$ for all $\boldsymbol{\alpha} \in \mathbb{R}^n$ and hence $\boldsymbol{s}_2(\boldsymbol{x}_j) = \boldsymbol{0}$ for $1 \le j \le N$, we see that the minimum indeed must be attained for an element from V_X. In the second step, we write an arbitrary element of V_X as $\boldsymbol{s} = \sum_{j=1}^N \boldsymbol{K}(\cdot,\boldsymbol{x}_j)\boldsymbol{\beta}_j$. Minimizing the above expression then means determining the optimal $\boldsymbol{\beta}_j$'s, which eventually gives the stated result. ■

For us the existence of extension and restriction operators will be important. If the kernel $\boldsymbol{K}$ is actually defined on a larger set $\widetilde{\mathcal{D}} \supseteq \mathcal{D}$, we can study the connection between the reproducing kernel Hilbert space $\mathcal{H}(\widetilde{\mathcal{D}})$ and $\mathcal{H}(\mathcal{D})$ of $\boldsymbol{K}$ and $\boldsymbol{K}|(\mathcal{D}\times\mathcal{D})$. The following result is a generalization of the scalar-valued result, see Wendland (2005).

Theorem 16 *Let $\mathcal{D} \subseteq \widetilde{\mathcal{D}} \subseteq \mathbb{R}^d$ be given. Let $\boldsymbol{K} : \widetilde{\mathcal{D}} \times \widetilde{\mathcal{D}} \to \mathbb{R}^{n\times n}$ be positive definite and let $\mathcal{H}(\widetilde{\mathcal{D}})$ and $\mathcal{H}(\mathcal{D})$ be the RKHS in which $\boldsymbol{K}$ and $\boldsymbol{K}|(\mathcal{D}\times\mathcal{D})$ are reproducing kernels, respectively.*

- *There is a linear isometric* extension operator *$E : \mathcal{H}(\mathcal{D}) \to \mathcal{H}(\widetilde{\mathcal{D}})$ with $(E\boldsymbol{v})|\mathcal{D} = \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathcal{H}(\mathcal{D})$.*
- *The linear* restriction operator *$R : \mathcal{H}(\widetilde{\mathcal{D}}) \to \mathcal{H}(\mathcal{D})$, defined by $R\boldsymbol{v} = \boldsymbol{v}|\mathcal{D}$ for all $\boldsymbol{v} \in \mathcal{H}(\widetilde{\mathcal{D}})$ satisfies $\|R\| = 1$.*

Proof Because $\mathcal{D} \subseteq \widetilde{\mathcal{D}}$, we have a natural extension operator $E : F_{\boldsymbol{K}}(\mathcal{D}) \to F_{\boldsymbol{K}}(\widetilde{\mathcal{D}})$ by simply evaluating any $\boldsymbol{v} \in F_{\boldsymbol{K}}(\mathcal{D})$ also at points from $\widetilde{\mathcal{D}}$. (For the definition of $F_{\boldsymbol{K}}$ see (9).) This operator is obviously linear and an isometry, since the norm of such a $\boldsymbol{v}$ depends only on the centers and coefficients of $\boldsymbol{v}$. As $F_{\boldsymbol{K}}(\mathcal{D})$ is dense in $\mathcal{H}(\mathcal{D})$, this extension operator extends to $\mathcal{H}(\mathcal{D})$, keeping the isometry property.

We now turn to the restriction operator. Here, we need a vector-valued version of the results of Theorems 10.22, 10.26 and 10.46 of Wendland (2005). We will adapt these results and condense the arguments to the specific situation. Hence, let $\boldsymbol{v} \in \mathcal{H}(\widetilde{\mathcal{D}})$ be given. We first need to show that $\boldsymbol{v}|\mathcal{D} \in \mathcal{H}(\mathcal{D})$. To this end, we define a linear functional $\psi_{\boldsymbol{v}} : F_{\boldsymbol{K}}(\mathcal{D}) \to \mathbb{R}$ by setting $\psi_{\boldsymbol{v}}(\boldsymbol{K}(\cdot,\boldsymbol{x})\boldsymbol{\alpha}) = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{D}$, $\boldsymbol{\alpha} \in \mathbb{R}^n$, and linearization. This is obviously linear. Moreover, for an arbitrary $\boldsymbol{w} = \sum_{j=1}^N \boldsymbol{K}(\cdot,\boldsymbol{x}_j)\boldsymbol{\alpha}_j \in F_{\boldsymbol{K}}(\mathcal{D})$ we have

$$\begin{aligned}
\psi_{\boldsymbol{v}}(\boldsymbol{w}) &= \sum_{j=1}^N \boldsymbol{\alpha}_j^{\mathrm{T}}\boldsymbol{v}(\boldsymbol{x}_j) = \sum_{j=1}^N \boldsymbol{\alpha}_j^{\mathrm{T}}\boldsymbol{I}_X\boldsymbol{v}(\boldsymbol{x}_j) = \left\langle \boldsymbol{I}_X\boldsymbol{v}, \sum_{j=1}^N \boldsymbol{K}(\cdot,\boldsymbol{x}_j)\boldsymbol{\alpha}_j \right\rangle_{\mathcal{H}(\mathcal{D})} \\
&= \langle \boldsymbol{I}_X\boldsymbol{v}, \boldsymbol{w}\rangle_{\mathcal{H}(\mathcal{D})} \le \|\boldsymbol{I}_X\boldsymbol{v}\|_{\mathcal{H}(\mathcal{D})}\|\boldsymbol{w}\|_{\mathcal{H}(\mathcal{D})} \le \|\boldsymbol{v}\|_{\mathcal{H}(\widetilde{\mathcal{D}})}\|\boldsymbol{w}\|_{\mathcal{H}(\mathcal{D})},
\end{aligned}$$

where in the last step, we have used $\boldsymbol{I}_X\boldsymbol{v} \in F_{\boldsymbol{K}}(\mathcal{D}) \subseteq F_{\boldsymbol{K}}(\widetilde{\mathcal{D}}) \subseteq \mathcal{H}(\widetilde{\mathcal{D}})$ and the norm-minimality of the interpolant. This shows boundedness of the functional $\psi_{\boldsymbol{v}}$ with $\|\psi_{\boldsymbol{v}}\| \le \|\boldsymbol{v}\|_{\mathcal{H}(\widetilde{\mathcal{D}})}$. As $F_{\boldsymbol{K}}(\mathcal{D})$ is dense in $\mathcal{H}(\mathcal{D})$, we can extend $\psi_{\boldsymbol{v}}$ to a functional $\psi_{\boldsymbol{v}} : \mathcal{H}(\mathcal{D}) \to \mathbb{R}$ having the same norm. For simplicity, we will use the same notation for the original functional and its extension. By the Riesz' representation theorem, there is a $\boldsymbol{w}_{\boldsymbol{v}} \in \mathcal{H}(\mathcal{D})$ with $\|\boldsymbol{w}_{\boldsymbol{v}}\|_{\mathcal{H}(\mathcal{D})} = \|\psi_{\boldsymbol{v}}\|$ and $\psi_{\boldsymbol{v}}(\boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{w}_{\boldsymbol{v}}\rangle_{\mathcal{H}(\mathcal{D})}$ for all $\boldsymbol{w} \in \mathcal{H}(\mathcal{D})$. It now suffices to show that $\boldsymbol{v}|\mathcal{D} = \boldsymbol{w}_{\boldsymbol{v}}$. To see this, let $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\boldsymbol{x} \in \mathcal{D}$ be fixed but arbitrary and conclude

$$\begin{aligned}
\boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{w}_{\boldsymbol{v}}(\boldsymbol{x})) &= \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}(\boldsymbol{x}) - \langle \boldsymbol{w}_{\boldsymbol{v}}, K(\cdot,\boldsymbol{x})\boldsymbol{\alpha}\rangle_{\mathcal{H}(\mathcal{D})} \\
&= \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}(\boldsymbol{x}) - \psi_{\boldsymbol{v}}(K(\cdot,\boldsymbol{x})\boldsymbol{\alpha}) \\
&= \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}(\boldsymbol{x}) = 0.
\end{aligned}$$

Now that we know that R maps $\mathcal{H}(\widetilde{\mathcal{D}})$ indeed to $\mathcal{H}(\mathcal{D})$, we can look at the norm and see that for every $\boldsymbol{v} \in \mathcal{H}(\widetilde{\mathcal{D}})$ we have

$$\|R\boldsymbol{v}\|_{\mathcal{H}(\mathcal{D})} = \|\boldsymbol{w}_{\boldsymbol{v}}\|_{\mathcal{H}(\mathcal{D})} = \|\psi_{\boldsymbol{v}}\| \le \|\boldsymbol{v}\|_{\mathcal{H}(\widetilde{\mathcal{D}})},$$

showing $\|R\| \le 1$. However, any $\boldsymbol{v} = \boldsymbol{K}(\cdot,\boldsymbol{x})\boldsymbol{\alpha}$ with a fixed $\boldsymbol{x} \in \mathcal{D}$ and $\boldsymbol{\alpha} \in \mathbb{R}^n$ belongs to $F_{\boldsymbol{K}}(\mathcal{D})$ and $F_{\boldsymbol{K}}(\widetilde{\mathcal{D}}) \subseteq \mathcal{H}(\widetilde{\mathcal{D}})$. Obviously, its restriction $R\boldsymbol{v}$ is $\boldsymbol{v}$ itself, showing $\|R\| = 1$. ■

3.4 Sobolev Spaces

In this paper, we are mainly interested in Sobolev spaces as RKHSs. However, we will use some leeway in defining the inner product and norm on these spaces, as is customary in the scalar-valued case. The advantage of this procedure is that we can have different reproducing kernels for the same algebraic space, using different inner products. We start with recalling the definitions of vector-valued Sobolev spaces and vector-valued mixed-regularity Sobolev spaces. We do this not only for Hilbert spaces but also for general L_p spaces. To this end, we first recall the definition of the L_p-norm for vector-valued functions. For $1 \le p < \infty$ and $\boldsymbol{v} \in L_p(\mathcal{D})^n$, this is defined as

$$\|\boldsymbol{v}\|_{L_p(\mathcal{D})^n} := \left(\sum_{\ell=1}^n \|v_\ell\|_{L_p(\mathcal{D})}^p\right)^{1/p} = \left(\sum_{\ell=1}^n \int_{\mathcal{D}} |v_\ell(\boldsymbol{x})|^p d\boldsymbol{x}\right)^{1/p},$$

i.e. by taking the ℓ_p norm of the L_p norms. For $p = \infty$ the definition of the norm uses the usual modifications.

For $k \in \mathbb{N}_0$, $1 \le p < \infty$ the Sobolev space of order k is defined as expected:

$$\begin{aligned}
W_p^k(\mathcal{D})^n &:= \{\boldsymbol{v} \in L_p(\mathcal{D})^n : \|\boldsymbol{v}\|_{W_p^k(\mathcal{D})^n} < \infty\}, \\
\|\boldsymbol{v}\|_{W_p^k(\mathcal{D})^n} &:= \left(\sum_{\ell=0}^k |\boldsymbol{v}|_{W_p^\ell(\mathcal{D})^n}^p\right)^{1/p}, \\
|\boldsymbol{v}|_{W_p^\ell(\mathcal{D})^n} &:= \left(\sum_{|\boldsymbol{\alpha}|=\ell} \|D^{\boldsymbol{\alpha}}\boldsymbol{v}\|_{L_p(\mathcal{D})^n}^2\right)^{1/2}.
\end{aligned}$$

As usual, the derivatives are meant in the weak sense. For $p = 2$ we will also use the notation $H^k(\mathcal{D})^n := W_2^k(\mathcal{D})^n$. Moreover, for $1 \le p < \infty$, using for example interpolation theory, the definition can be extended to define also fractional order Sobolev spaces $W_p^\tau(\Omega)^n$.

Before we specify the connection between Sobolev spaces and reproducing kernel Hilbert spaces, we want to generalize a *sampling inequality* to the vector valued case. Such sampling inequalities were introduced in the context of scattered data interpolation and approximation in Narcowich et al. (2005); Wendland and Rieger (2005) and then refined, for example, in Arcangéli et al. (2007, 2012); Narcowich et al. (2006). They are a universal tool to provide immediate and optimal error estimates for a variety of approximation processes based on discrete data, as long as these processes have small or vanishing error on the data and are stable in some Sobolev norm. More details can be found in the papers mentioned above.

While an extension of such sampling inequalities for functions vanishing on the data to vector-valued functions is straight-forward and has been used, for example, in Fuselier (2008a); Schräder and Wendland (2011); Wendland (2009), we require the following generalization for functions not vanishing on the data. We formulate it only in the required situation and not in the most general case. We will also need a discrete ℓ_p norm, which we define for $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and $1 \le p < \infty$ by

$$\|\boldsymbol{v}\|_{\ell_p(X)^n} := \left(\sum_{\ell=1}^{n} \|v_\ell\|_{\ell_p(X)}^p \right)^{1/p} = \left(\sum_{\ell=1}^{n} \sum_{j=1}^{N} |v_\ell(\boldsymbol{x}_j)|^p \right)^{1/p}$$

with the obvious modification for $p = \infty$. This, of course is a generalization of the ℓ_p norm for vectors of vectors. For $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$ with $\boldsymbol{y}_j \in \mathbb{R}^n$, we write

$$\|\boldsymbol{Y}\|_{\ell_p} := \left(\sum_{j=1}^{N} \|\boldsymbol{y}_j\|_{\ell_p}^p \right)^{1/p}.$$

Recall that any bounded domain with a Lipschitz boundary also satisfies an interior cone condition, see for example Grisvard (1985). In the sampling theorem that follows we use the fill distance from (5) and $(x)_+ := \max\{x, 0\}$.

Theorem 17 *Let $\mathcal{D} \subseteq \mathbb{R}^d$ be a bounded domain with a Lipschitz boundary. Let $p, q \in [1, \infty]$ and $n \in \mathbb{N}$. Let $\tau > d/2$ and $\gamma := \max\{2, p, q\}$. Then, there exist constants $h_0 > 0$ (depending on $\mathcal{D}$ and τ) and $C > 0$ (depending on $\mathcal{D}$, n, τ and q) such that for all $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ with $h = h_{X,\mathcal{D}} \le h_0$ and all $\boldsymbol{v} \in H^\tau(\mathcal{D})^n$, we have*

$$|\boldsymbol{v}|_{W_q^s(\mathcal{D})^n} \le C \left(h^{\tau - s - d(1/2 - 1/q)_+} |\boldsymbol{v}|_{H^\tau(\mathcal{D})^n} + h^{d/\gamma - s} \|\boldsymbol{v}\|_{\ell_p(X)^n} \right)$$

for all $0 \le s \le \ell(\tau, q, d)$, where $\ell(\tau, q, d) = \ell_0(\tau, q, d) := \tau - d(1/2 - 1/q)_+$ in the case $\tau \in \mathbb{N}$ and either $q > 2$ and $\ell_0(\tau, q, d) \in \mathbb{N}$ or $q = 2$. Otherwise $\ell(\tau, q, d) = \lceil \ell_0(\tau, q, d) \rceil - 1$. Moreover, $s \in \mathbb{N}_0$ in the case $q = \infty$.

Proof In the case $n = 1$, this is a special case of Theorems 3.1 and 3.2 from Arcangéli et al. (2012). For $n \ge 2$ we set $e_1 := \tau - s - d(1/2 - 1/q)_+$ and $e_2 := d/\gamma - s$ to simplify the notation. Then, using the scalar-valued case, we have for $1 \le q < \infty$,

$$\begin{aligned}
|\boldsymbol{v}|_{W_q^s(\mathcal{D})^n} &= \left(\sum_{\ell=1}^n |v_\ell|_{W_q^s(\mathcal{D})}^q \right)^{1/q} \\
&\le \left(\sum_{\ell=1}^n \left[C \left(h^{e_1} |v_\ell|_{H^\tau(\mathcal{D})} + h^{e_2} \|v_\ell\|_{\ell_p(X)} \right) \right]^q \right)^{1/q} \\
&\le C \left[h^{e_1} \left(\sum_{\ell=1}^n |v_\ell|_{H^\tau(\mathcal{D})}^q \right)^{1/q} + |h^{e_2} \left(\sum_{\ell=1}^n \|v_\ell\|_{\ell_p(X)}^q \right)^{1/q} \right] \\
&\le C \left[c_{2q} h^{e_1} |\boldsymbol{v}|_{H^\tau(\mathcal{D})^n} + c_{pq} h^{e_2} \|\boldsymbol{v}\|_{\ell_p(X)^n} \right],
\end{aligned}$$

where we have used in the first step the scalar-valued bound, in the second step the triangle inequality for ℓ_q-norms and in the final step the equivalence of ℓ_q and ℓ_p and ℓ_2 norms. The equivalence constants are given by $c_{pq} = n^{1/q-1/p}$ if $p \ge q$ and $c_{pq} = 1$ otherwise. The case $q = \infty$ is treated similarly. ■

Typically, the above sampling inequality is not used with the target function $\boldsymbol{v}$ directly but with the error $\boldsymbol{v} - \boldsymbol{A}_X \boldsymbol{v}$, where $\boldsymbol{A}_X \boldsymbol{v}$ is some linear approximation process using only the data $(X, \boldsymbol{v}(X))$ or $(X, \boldsymbol{v}(X) + \boldsymbol{\epsilon})$. In this situation the sampling inequality immediately yields

$$|\boldsymbol{v} - \boldsymbol{A}_X \boldsymbol{v}|_{W_q^s(\mathcal{D})^n} \le C \left(h^{\tau - s - d(1/2-1/q)_+} |\boldsymbol{v} - \boldsymbol{A}_X \boldsymbol{v}|_{H^\tau(\mathcal{D})^n} + h^{d/\gamma - s} \|\boldsymbol{v} - \boldsymbol{A}_X \boldsymbol{v}\|_{\ell_p(X)^n} \right),$$

which becomes a proper error bound if the error $\boldsymbol{v} - \boldsymbol{A}_X \boldsymbol{v}$ is small on the data sites X and the process is stable in the sense of $|\boldsymbol{A}_X \boldsymbol{v}|_{H^\tau(\mathcal{D})^n} \le C \|\boldsymbol{v}\|_{H^\tau(\mathcal{D})^n}$. We will employ this procedure for interpolation, $\boldsymbol{A}_X = \boldsymbol{I}_X$, and regression, $\boldsymbol{A}_X = \boldsymbol{Q}_{X,\lambda}$.

The Sobolev embedding theorem guarantees that for $\tau > d/2$, the space $H^\tau(\mathcal{D})$ is a reproducing kernel Hilbert space. However, for general $\mathcal{D}$ the reproducing kernel is usually not known in explicit form. A typical remedy to this problem, which we employ here, is to use a reproducing kernel of $H^\tau(\mathbb{R}^d)$, which can, for example, be given as $K_\tau(\boldsymbol{x}, \boldsymbol{y}) = \Phi_\tau(\boldsymbol{x} - \boldsymbol{y})$ with a continuous and integrable function $\Phi_\tau : \mathbb{R}^d \to \mathbb{R}$, which has a Fourier transform

$$\widehat{\Phi}_\tau(\boldsymbol{\omega}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \Phi_\tau(\boldsymbol{\omega}) e^{-i\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\omega}} d\boldsymbol{x}$$

having the following decay property: there exist constants $c_1, c_2 > 0$ such that

$$c_1 (1 + \|\boldsymbol{\omega}\|_2^2)^{-\tau} \le \widehat{\Phi}_\tau(\boldsymbol{\omega}) \le c_2 (1 + \|\boldsymbol{\omega}\|_2^2)^{-\tau}, \qquad \boldsymbol{\omega} \in \mathbb{R}^d. \tag{13}$$

To be more precise, we have the following well-known result see Wendland (2005).

Lemma 18 *Suppose $\Phi_\tau \in L_1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ possesses a Fourier transform with decay (13). Then, $K_\tau : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $K_\tau(\boldsymbol{x}, \boldsymbol{y}) = \Phi_\tau(\boldsymbol{x} - \boldsymbol{y})$ is the reproducing kernel of $H^\tau(\mathbb{R}^d)$ with respect to the inner product*

$$\langle f, g \rangle_{K_\tau} := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\widehat{f}(\boldsymbol{\omega}) \overline{\widehat{g}(\boldsymbol{\omega})}}{\widehat{\Phi}_\tau(\boldsymbol{\omega})} d\boldsymbol{\omega}.$$

To use a kernel which is defined on all of $\mathbb{R}^d$ and then restricted to $\mathcal{D} \subseteq \mathbb{R}^d$ in the context of kernel-based learning can be done in two equivalent ways. Both involve the existence of a bounded, linear extension operator $E_{\mathcal{D}} : H^\tau(\mathcal{D}) \to H^\tau(\mathbb{R}^d)$, which exists, for example, if $\mathcal{D}$ is a bounded domain with a Lipschitz boundary, see for example Brenner and Scott (1994). The first approach is based on the fact that the interpolation/approximation only uses data of the unknown function f on data sites $X \subseteq \mathcal{D}$. Thus, the interpolation/approximation to f is the same as the interpolation/approximation to $E_{\mathcal{D}} f$, which allows us to work in the RKHS over all of $\mathbb{R}^d$. Here, we will use the second approach, which uses the fact that there is also a canonical extension operator for $E : \mathcal{H}_K(\mathcal{D}) \to \mathcal{H}_K(\mathbb{R}^d)$ for any positive definite kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, which is the identity on $F_K(\mathcal{D}) = \operatorname{span}\{K(\cdot, \boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}\}$. This approach leads to the following result; a proof can be found in (Wendland, 2005, Corollary 10.48) for the integer case and follows by interpolation theory for the fractional-order case.

Lemma 19 *Assume $\Phi \in L_1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ has a Fourier transform $\widehat{\Phi}$ satisfying (13) with $\tau > d/2$. Let $\mathcal{D} \subseteq \mathbb{R}^d$ be a bounded domain with a Lipschitz boundary. Let $K_\tau : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ be defined by $K_\tau(\boldsymbol{x}, \boldsymbol{y}) = \Phi(\boldsymbol{x} - \boldsymbol{y})$, $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$. Then, there exists an inner product $\langle \cdot, \cdot \rangle_{K_\tau} : H^\tau(\mathcal{D}) \times H^\tau(\mathcal{D}) \to \mathbb{R}$ on $H^\tau(\mathcal{D})$ such that K_τ is the reproducing kernel of $H^\tau(\mathcal{D})$ with respect to this inner product. The norm $\|\cdot\|_{K_\tau}$ induced by this inner product is equivalent to the standard norm on $H^\tau(\mathcal{D})$, i.e. there are constants $C_1(\tau), C_2(\tau) > 0$ such that*

$$C_1(\tau)\|u\|_{K_\tau} \le \|u\|_{H^\tau(\mathcal{D})} \le C_2(\tau)\|u\|_{K_\tau}, \qquad u \in H^\tau(\mathcal{D}). \tag{14}$$

The inner product is first defined on $F_K(\mathcal{D}) := \operatorname{span}\{K(\cdot, \boldsymbol{x}) : \boldsymbol{x} \in \mathcal{D}\}$ by

$$\langle K_\tau(\cdot, \boldsymbol{x}), K_\tau(\cdot, \boldsymbol{y}) \rangle_{K_\tau} := K_\tau(\boldsymbol{x}, \boldsymbol{y})$$

and linearisation and then extended to the closure of $F_K(\mathcal{D})$ with respect to the norm induced by this inner product. For details, see Wendland (2005). The above lemma can immediately be carried over to vector-valued Sobolev spaces, using Theorem 16. However, we will do this for certain subspaces in the next section.

3.5 Divergence- and Rotation-Free Kernels

In the context of kernel-based approximation theory, divergence- and rotation-free kernels were introduced in Narcowich and Ward (1994). The construction of divergence-free and rotation-free matrix-valued kernels is quite simple. Starting with a positive definite function $\Phi : \mathbb{R}^d \to \mathbb{R}$, which is at least twice differentiable, we can define functions $\boldsymbol{\Phi}_{\mathrm{div}}, \boldsymbol{\Phi}_{\mathrm{curl}} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ by

$$\boldsymbol{\Phi}_{\mathrm{div}} := (-\Delta I + \nabla\nabla^{\mathrm{T}})\Phi, \tag{15}$$

$$\boldsymbol{\Phi}_{\mathrm{curl}} := -\nabla\nabla^{\mathrm{T}}\Phi. \tag{16}$$

It is easy to see that for three times differentiable Φ the associated kernels

$$\boldsymbol{K}_{\mathrm{div}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Phi}_{\mathrm{div}}(\boldsymbol{x} - \boldsymbol{y}) \text{ and } \boldsymbol{K}_{\mathrm{curl}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Phi}_{\mathrm{curl}}(\boldsymbol{x} - \boldsymbol{y}) \tag{17}$$

have for a fixed argument $\boldsymbol{y}$ divergence-free and rotation-free columns respectively, with the latter of course only for $d = 2, 3$.

These kernels are the reproducing kernels of function spaces consisting of divergence- and rotation-free functions. To be more precise, the following result comes from Fuselier (2008a,b); Schräder and Wendland (2011); Wendland (2009). To formulate it, we first introduce the spaces of divergence- and rotation-free Sobolev functions. For $\tau > 1$ and $\mathcal{D} \subseteq \mathbb{R}^d$, where $d = 2, 3$ in the rotation-free case, let

$$\begin{aligned} H^\tau(\mathcal{D}; \operatorname{div}) &:= \{\boldsymbol{v} \in H^\tau(\mathcal{D})^d : \operatorname{div} \boldsymbol{v} = 0\}, \\ H^\tau(\mathcal{D}; \operatorname{curl}) &:= \{\boldsymbol{v} \in H^\tau(\mathcal{D})^d : \operatorname{curl} \boldsymbol{v} = \mathbf{0}\}. \end{aligned}$$

Of course, as long as we do not have $\tau > d/2 + 1$, the derivatives are meant to be in the weak sense and hence zero divergence and rotation are meant only up to a set of measure zero. We will also need the modified Sobolev spaces

$$\begin{aligned} \widetilde{H}^\tau(\mathbb{R}^d; \operatorname{div}) &= \{\boldsymbol{v} \in H^\tau(\mathbb{R}^d; \operatorname{div}) : \|\boldsymbol{v}\|_{\widetilde{H}^\tau(\mathbb{R}^d)^d} < \infty\}, \\ \widetilde{H}^\tau(\mathbb{R}^d; \operatorname{curl}) &= \{\boldsymbol{v} \in H^\tau(\mathbb{R}^d; \operatorname{curl}) : \|\boldsymbol{v}\|_{\widetilde{H}^\tau(\mathbb{R}^d)^d} < \infty\}, \end{aligned}$$

which are both equipped with the inner product

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle_{\widetilde{H}^\tau(\mathbb{R}^d)^d} := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\overline{\widehat{\boldsymbol{w}}(\boldsymbol{\omega})}^{\mathrm{T}} \widehat{\boldsymbol{v}}(\boldsymbol{\omega})}{\|\boldsymbol{\omega}\|_2^2} (1 + \|\boldsymbol{\omega}\|_2^2)^{\tau+1} d\boldsymbol{\omega}. \tag{18}$$

Here, the Fourier transform of a vector-valued function is defined component-wise. Note that for any $\boldsymbol{v} \in \widetilde{H}^\tau(\mathbb{R}^d; \operatorname{div})$ or $\boldsymbol{v} \in \widetilde{H}^\tau(\mathbb{R}^d; \operatorname{curl})$ we have the obvious but also important bound $\|\boldsymbol{v}\|_{H^\tau(\mathbb{R}^d)^d} \leq \|\boldsymbol{v}\|_{\widetilde{H}^\tau(\mathbb{R}^d)^d}$.

In this and the next section, we will use (13) with τ replaced by $\tau + 1$,

$$c_3(1 + \|\boldsymbol{\omega}\|_2^2)^{-\tau-1} \leq \widehat{\Phi}_{\tau+1}(\boldsymbol{\omega}) \leq c_4(1 + \|\boldsymbol{\omega}\|_2^2)^{-\tau-1}, \qquad \boldsymbol{\omega} \in \mathbb{R}^d. \tag{19}$$

The connection between these modified Sobolev spaces and the above-defined kernels is as follows, see Wendland (2009); Fuselier (2008a).

Proposition 20 *Let $\tau > d/2$. Let $\Phi = \Phi_{\tau+1}$ define a reproducing kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ of $H^{\tau+1}(\mathbb{R}^d)$ in the sense that it has a Fourier transform $\widehat{\Phi}$ satisfying (19). Assume that $\Phi \in W_1^2(\mathbb{R}^d)$. Then, $H^{\tau+1}(\mathbb{R}^d) \subseteq C^1(\mathbb{R}^d) \cap W_\infty^1(\mathbb{R}^d)$ and $\Phi \in C^2(\mathbb{R}^d)$. Moreover, the kernels $\boldsymbol{K}_{\operatorname{div}}$ and $\boldsymbol{K}_{\operatorname{curl}}$ defined from $\boldsymbol{\Phi}_{\operatorname{div}}$ and $\boldsymbol{\Phi}_{\operatorname{curl}}$ as in (17) are the reproducing kernels of $\widetilde{H}^\tau(\mathbb{R}^d; \operatorname{div})$ and $\widetilde{H}^\tau(\mathbb{R}^d; \operatorname{curl})$, respectively, where in both cases the spaces are equipped with the inner product*

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle_{\widetilde{H}} = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\overline{\widehat{\boldsymbol{w}}(\boldsymbol{\omega})}^{\mathrm{T}} \widehat{\boldsymbol{v}}(\boldsymbol{\omega})}{\widehat{\Phi}(\boldsymbol{\omega}) \|\boldsymbol{\omega}\|_2^2} d\boldsymbol{\omega}.$$

Next, we need extension operators to define kernel-induced inner products on $H^\tau(\mathcal{D}; \operatorname{div})$ and $H^\tau(\mathcal{D}, \operatorname{curl})$. The following result is from Wendland (2009); Fuselier (2008a).

Proposition 21 *Let $\tau \geq 0$ and let $\mathcal{D} \subseteq \mathbb{R}^d$ be a simply-connected, bounded domain with $d = 2, 3$.*

1. *If $\mathcal{D}$ has a $C^{\lceil\tau\rceil,1}$ boundary then there is a bounded, linear operator $E_{\mathrm{div}} : H^\tau(\mathcal{D};\mathrm{div}) \to \widetilde{H}^\tau(\mathbb{R}^d;\mathrm{div})$ with $(E_{\mathrm{div}}\boldsymbol{v})|\mathcal{D} = \boldsymbol{v}$ for all $\boldsymbol{v} \in H^\tau(\mathcal{D};\mathrm{div})$.*

2. *If $\mathcal{D}$ has a Lipschitz boundary then there exists a bounded and linear operator E_{curl} : $H^\tau(\mathcal{D};\mathrm{curl}) \to \widetilde{H}^\tau(\mathbb{R}^d;\mathrm{curl})$ with $(E_{\mathrm{curl}}\boldsymbol{v})|\mathcal{D} = \boldsymbol{v}$ for all $\boldsymbol{v} \in H^\tau(\mathcal{D};\mathrm{curl})$.*

Note that we also need $d = 2, 3$ in the divergence-free case for the existence of the extension operator. For that reason, we will, from now on restrict ourselves to this situation.

With this, we can now prove a result corresponding to Lemma 19 for our divergence-free and rotation-free kernels.

Theorem 22 *Let $d = 2, 3$. Assume $\Phi = \Phi_{\tau+1} \in W_1^2(\mathbb{R}^d) \cap C^2(\mathbb{R}^d)$ has a Fourier transform satisfying (19) and $\tau > d/2$. Let $\boldsymbol{K}_{\mathrm{div}}$ be defined by $\boldsymbol{\Phi}_{\mathrm{div}}$ from (15) and let $\boldsymbol{K}_{\mathrm{curl}}$ be defined by $\boldsymbol{\Phi}_{\mathrm{curl}}$ from (16). Let $\mathcal{D} \subseteq \mathbb{R}^d$ be a bounded, simply-connected domain.*

1. *If $\mathcal{D}$ has a $C^{\lceil\tau\rceil,1}$ boundary then there exists an inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{K}_{\mathrm{div}}} : H^\tau(\mathcal{D};\mathrm{div}) \times H^\tau(\mathcal{D};\mathrm{div}) \to \mathbb{R}$ such that $\boldsymbol{K}_{\mathrm{div}}$ is the reproducing kernel with respect to this inner product. The norm defined by this inner product is equivalent to the standard norm on $H^\tau(\mathcal{D};\mathrm{div})$.*

2. *If $\mathcal{D}$ has a Lipschitz boundary then there is an inner product $\langle\cdot,\cdot\rangle_{\boldsymbol{K}_{\mathrm{curl}}} : H^\tau(\mathcal{D};\mathrm{curl}) \times H^\tau(\mathcal{D};\mathrm{curl}) \to \mathbb{R}$ such that $\boldsymbol{K}_{\mathrm{curl}}$ is the reproducing kernel with respect to this inner product. The norm defined by this inner product is equivalent to the standard norm on $H^\tau(\mathcal{D};\mathrm{curl})$.*

Proof As the proof of both cases is similar, we will only prove the first one, i.e. for $\boldsymbol{K}_{\mathrm{div}}$. The proof itself is similar to the proof of the scalar-valued case. We first recall, that $\boldsymbol{K}_{\mathrm{div}}$ can be seen as the reproducing kernel of a unique Hilbert space $\mathcal{H}(\mathcal{D})$ but also as the reproducing kernel of a Hilbert space $\mathcal{H}(\mathbb{R}^d)$, both of which are constructed by completing the pre-Hilbert spaces $F_{\boldsymbol{K}_{\mathrm{div}}}(\mathcal{D})$ and $F_{\boldsymbol{K}_{\mathrm{div}}}(\mathbb{R}^d)$, respectively. By Theorem 16, there are natural extension and restriction operators $E : \mathcal{H}(\mathcal{D}) \to \mathcal{H}(\mathbb{R}^d)$ and $R : \mathcal{H}(\mathbb{R}^d) \to \mathcal{H}(\mathcal{D})$, both having norm one.

Next, on $\mathbb{R}^d$, we know by Proposition 20 that the spaces $\mathcal{H}(\mathbb{R}^d)$ and $\widetilde{H}^\tau(\mathbb{R}^d;\mathrm{div})$ are algebraically the same with equivalent norms. To be more precise, the identity maps ι : $\mathcal{H}(\mathbb{R}^d) \to \widetilde{H}^\tau(\mathbb{R}^d;\mathrm{div})$ and $\widetilde{\iota} : \widetilde{H}^\tau(\mathbb{R}^d;\mathrm{div}) \to \mathcal{H}(\mathbb{R}^d)$ satisfy $\|\iota\| \le 1/C_1(\tau+1)^{1/2}$ and $\|\widetilde{\iota}\| \le C_2(\tau+1)^{1/2}$, where the constants come from (14).

Finally, we have the extension operator $E_{\mathrm{div}} : H^\tau(\mathcal{D};\mathrm{div}) \to \widetilde{H}^\tau(\mathbb{R}^d;\mathrm{div})$ from Proposition 21 and note that the restriction of a function $\boldsymbol{v} \in \widetilde{H}^\tau(\mathbb{R}^d;\mathrm{div})$ to $\mathcal{D}$ gives a function $\boldsymbol{v}|\mathcal{D} \in H^\tau(\mathcal{D};\mathrm{div})$ and thus defines a restriction operator $R_{\mathrm{div}} : \widetilde{H}^\tau(\mathbb{R}^d;\mathrm{div}) \to H(\mathcal{D};\mathrm{div})$ with $\|R_{\mathrm{div}}\| = 1$.

Having all these operators at hand, we are now able to show that the spaces $\mathcal{H}(\mathcal{D})$ and $H^\tau(\mathcal{D};\mathrm{div})$ are algebraically the same spaces with equivalent norms. On the one hand, for $\boldsymbol{v} \in \mathcal{H}(\mathcal{D})$ we have

$$\boldsymbol{v} = E\boldsymbol{v}|\mathcal{D} = \iota \circ E\boldsymbol{v}|\mathcal{D} = R_{\mathrm{div}} \circ \iota \circ E\boldsymbol{v} \in H^\tau(\mathcal{D};\mathrm{div}).$$

On the other hand, for $\boldsymbol{v} \in H^\tau(\mathcal{D};\mathrm{div})$ we have

$$\boldsymbol{v} = E_{\mathrm{div}}\boldsymbol{v}|\mathcal{D} = \widetilde{\iota} \circ E_{\mathrm{div}}\boldsymbol{v}|\mathcal{D} = R \circ \widetilde{\iota} \circ E_{\mathrm{div}}\boldsymbol{v} \in \mathcal{H}(\mathcal{D}),$$

showing that both spaces algebraically coincide. Moreover, these relations also show

$$\begin{aligned}\|\boldsymbol{v}\|_{H^\tau(\mathcal{D};\mathrm{div})} &\leq \frac{1}{C_1(\tau+1)^{1/2}}\|\boldsymbol{v}\|_{\mathcal{H}(\mathcal{D})},\\ \|\boldsymbol{v}\|_{\mathcal{H}(\mathcal{D})} &\leq C_2(\tau+1)^{1/2}C_E\|\boldsymbol{v}\|_{H^\tau(\mathcal{D};\mathrm{div})},\end{aligned}$$

where C_E is the norm of E_{div}. This is the stated norm equivalence. ■

4. Error Analysis

In this section we discuss the quality of the predictive mean as an estimate of the data-generating field $\boldsymbol{v} : \mathcal{D} \to \mathbb{R}^d$, under various assumptions on the noise.

Error estimates for vector-valued function approximation can often be deduced from the scalar-valued case. However, in our situation this is not possible, as the components cannot be approximated independently due to the divergence-free or rotation-free condition. Hence, in this section we use and extend the existing error analysis for such kernel methods. For exact data, the existing results for example from Fuselier (2008a) have to be extended to deal with less smooth target functions and in the case of the regression operator. Moreover, new results are needed in the case of noisy data.

Throughout we will make the following general assumptions on the geometry and the field $\boldsymbol{v}$.

Assumption 23 *For $d = 2$ or 3, $\mathcal{D} \subseteq \mathbb{R}^d$ is a bounded domain with boundary satisfying the conditions in Theorem 22, Part 1 or Part 2, as appropriate to the divergence-free or curl-free situations.*

Assumption 24 *The field $\boldsymbol{v} : \mathcal{D} \to \mathbb{R}^d$ is a continuously differentiable function modelled by the Gaussian process $\boldsymbol{v}_0 \sim GP(\boldsymbol{m}, \boldsymbol{K})$, where $\boldsymbol{K}$ is one of $\boldsymbol{K}_{\mathrm{div}}$ or $\boldsymbol{K}_{\mathrm{curl}}$ as defined in Theorem 22.*

Remark 25 *In practical situations it will often be desirable to modify the kernel by the inclusion of "hyperparameters", for example replacing $\boldsymbol{K}_{\mathrm{div}}(\boldsymbol{x}, \boldsymbol{x}')$ by $\alpha^2\boldsymbol{K}_{\mathrm{div}}(\kappa\boldsymbol{x}, \kappa\boldsymbol{x}')$, where α and κ are positive parameters with α^2 controlling the variance and κ the length scale. For simplicity we ignore hyperparameters in the present discussion, noting that the predictive mean is independent of α^2, while changing the inverse length scale parameter κ would change only the constants in the error bounds in subsequent theorems.*

Assumption 26 *The kernel $\boldsymbol{K}$ is the reproducing kernel of $H^\tau(\mathcal{D};\mathrm{div})$ or $H^\tau(\mathcal{D};\mathrm{curl})$, respectively, and the target function $\boldsymbol{v}$ and the mean $\boldsymbol{m}$ of $\boldsymbol{v}_0$ belong to $H^\beta(\mathcal{D};\mathrm{div})$ or $H^\beta(\mathcal{D};\mathrm{curl})$, respectively, where $\beta, \tau > d/2$ in order to ensure that functions in H^β and H^τ are continuous.*

Within this setting we will consider both the *matching smoothness* case $\beta \geq \tau$ and the *non-matching smoothness* case $\tau > \beta$.

We will study the predictive mean computed under both the assumption of no noise in the data, in which case we shoud look at $\boldsymbol{m}_N$, and of noisy data, in which case we look at

$\boldsymbol{m}_{N,\sigma}$. In the assumed no-noise case we will naturally derive error estimates when there is no noise in the data, but we will also look at the misrepresented case in which we assume no noise but in fact there is noise in the data. In the case of assumed noisy data we will study first the misrepresented case that there is no noise, then the misrepresented case that there is noise but we do not know the kind of noise, and finally the case that there is Gaussian noise.

For all of these cases, we will assume that the prior mean $\boldsymbol{m}$ is divergence- or rotation-free, respectively, and has the same regularity as $\boldsymbol{v}$. While the first assumption is natural, a more natural smoothness assumption might be that $\boldsymbol{m}$ belongs to the RKHS of the kernel used for the prior approximation, i.e. $\boldsymbol{m} \in H^\tau(\mathcal{D}; \mathrm{div})$, for example. However, this choice would not change the approximation results below significantly, as the approximation order is always determined by $\min\{\beta, \tau\}$. It would only cause the statements of results to be lengthier.

4.1 Error Analysis for the Predictive Mean using Interpolation

In this section, we discuss the error $\boldsymbol{v} - \boldsymbol{m}_N$, i.e. the error of the predictive mean under the assumption that there is no noise in the data, even though there might be. Thus we will use the interpolation operator $\boldsymbol{I}_X$ to study the error.

We start with the case that the observations $\boldsymbol{y}_i = \boldsymbol{v}(\boldsymbol{x}_i)$, $1 \le i \le N$, indeed do not contain noise. We will state and prove error estimates for the *matching case* that the smoothness of $\boldsymbol{v}$ is aligned with the smoothness of the kernel, i.e. that $\boldsymbol{v} \in H^\tau(\mathcal{D}; \mathrm{div})$ and $\boldsymbol{K}_{\mathrm{div}}$ is the reproducing kernel of the latter space in the sense of Theorem 22. We will also have a result in which the smoothness of the target function is less than required. In approximation theory, this situation has been named *escaping the native space*. As the only difference between the divergence-free and the rotation-free case is the assumption on the smoothness of the boundary of the domain, we will state both results together.

We recall the fill distance from (5) and introduce the separation radius and the mesh ratio:

$$\begin{aligned} h_{X,\mathcal{D}} &= \sup_{\boldsymbol{x}\in\mathcal{D}} \min_{\boldsymbol{x}_j \in X} \|\boldsymbol{x} - \boldsymbol{x}_j\|_2, \\ q_{X,\mathcal{D}} &= \frac{1}{2} \min_{\boldsymbol{x}_j \neq \boldsymbol{x}_k} \|\boldsymbol{x}_j - \boldsymbol{x}_k\|_2, \\ \rho_{X,\mathcal{D}} &= h_{X,\mathcal{D}} / q_{X,\mathcal{D}}. \end{aligned}$$

To simplify the formulation of the approximation results, we will employ the notation

$$\mathcal{N}_\gamma(\boldsymbol{v}, \boldsymbol{m}) := \|\boldsymbol{v}\|_{H^\gamma(\mathcal{D})^d} + \|\boldsymbol{m}\|_{H^\gamma(\mathcal{D})^d}. \tag{20}$$

Theorem 27 (Interpolation without Noise) *Let Assumption 23 on the domain $\mathcal{D}$, Assumption 24 on the kernel $\boldsymbol{K}$ and Assumption 26 on the kernel $\boldsymbol{K}$, the target function $\boldsymbol{v}$ and mean $\boldsymbol{m}$ hold. Let $1 \le q \le \infty$. Let $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ and let $\boldsymbol{y}_j = \boldsymbol{v}(\boldsymbol{x}_j)$, $1 \le j \le N$. Finally, let $\boldsymbol{m}_N$ be the predictive, vector-valued mean based on the information $(\boldsymbol{x}_j, \boldsymbol{y}_j)$, $1 \le j \le N$. Then, there are constants $h_0, C > 0$ such that for all X with $h_{X,\mathcal{D}} \le h_0$ the following bounds hold.*

1. *If $\beta \geq \tau$ then*
$$\|\boldsymbol{v} - \boldsymbol{m}_N\|_{W_q^s(\mathcal{D})^d} \leq C h_{X,\mathcal{D}}^{\tau - s - d(1/2-1/q)_+} \mathcal{N}_\tau(\boldsymbol{v}, \boldsymbol{m})$$
for all $0 \leq s \leq \ell(\tau, q, d)$ with $\ell(\tau, q, d)$ given in Theorem 17 and $s \in \mathbb{N}_0$ if $q = \infty$.

2. *If $\beta < \tau$ then*
$$\|\boldsymbol{v} - \boldsymbol{m}_N\|_{W_q^s(\mathcal{D})^d} \leq C h_{X,\mathcal{D}}^{\beta - s - d(1/2-1/q)_+} \rho_{X,\mathcal{D}}^{\tau-\beta} \mathcal{N}_\beta(\boldsymbol{v}, \boldsymbol{m})$$
for all $0 \leq s \leq \ell(\beta, q, d)$ and $s \in \mathbb{N}_0$ if $q = \infty$.

Proof For $\beta \geq \tau$ the statement is essentially proved as Theorem 5 in Fuselier (2008a). However, we can shorten the proof by employing Theorem 22. We first note that with the interpolation operator $\boldsymbol{I}_X : C(\mathcal{D})^d \to \mathcal{H}$ with $\mathcal{H} = H^\tau(\mathcal{D}; \mathrm{div})$ or $\mathcal{H} = H^\tau(\mathcal{D}; \mathrm{curl})$, respectively, from (11) we have

$$\|\boldsymbol{v} - \boldsymbol{m}_N\|_{W_q^s(\mathcal{D})^d} \leq \|\boldsymbol{v} - \boldsymbol{I}_X \boldsymbol{v}\|_{W_q^s(\mathcal{D})^d} + \|\boldsymbol{m} - \boldsymbol{I}_X \boldsymbol{m}\|_{W_q^s(\mathcal{D})^d}. \tag{21}$$

As we assume that both $\boldsymbol{v}$ and $\boldsymbol{m}$ have Sobolev regularity $\beta \geq \tau > d/2$, both terms above are bounded in the same way and it suffices to concentrate on one of them, say the first one. As we have a sufficiently smooth boundary and as $\boldsymbol{v} - \boldsymbol{I}_X \boldsymbol{v}$ vanishes on X, we can use the sampling theorem, Theorem 17, to derive

$$\|\boldsymbol{v} - \boldsymbol{I}_X \boldsymbol{v}\|_{W_q^s(\mathcal{D})^d} \leq C h^{\tau - s - d(1/2-1/q)_+} \|\boldsymbol{v} - \boldsymbol{I}_X \boldsymbol{v}\|_{H^\tau(\mathcal{D})^d}.$$

By Theorem 22 we know there is an equivalent norm on $\mathcal{H}$ such that $\boldsymbol{K}$ is the reproducing kernel of $\mathcal{H}$ with respect to the corresponding inner product. Hence, by the second part of Theorem 14, we have $\|\boldsymbol{I}_X v\|_{H^\tau(\mathcal{D})^d} \leq C\|v\|_{H^\tau(\mathcal{D})^d}$ and hence $\|v - \boldsymbol{I}_X v\|_{H^\tau(\mathcal{D})^d} \leq C\|\boldsymbol{v}\|_{H^\tau(\mathcal{D})^d}$.

For $\beta < \tau$, we still have (21) and the fact that $\boldsymbol{v}$ and $\boldsymbol{m}$ belong to the same Sobolev space. Hence, it again is enough to study one of the errors. Theorem 17 yields this time

$$\|\boldsymbol{v} - \boldsymbol{I}_X \boldsymbol{v}\|_{W_q^s(\mathcal{D})^d} \leq C h^{\beta - s - d(1/2-1/q)_+} \|\boldsymbol{v} - \boldsymbol{I}_X \boldsymbol{v}\|_{H^\beta(\mathcal{D})^d}.$$

We can now proceed as in the proof of Theorem 6 in Fuselier (2008a) to use a specific auxiliary band-limited approximation to $\boldsymbol{v}$ together with the sampling theorem and a Bernstein inequality to finally derive

$$\|\boldsymbol{v} - \boldsymbol{I}_X \boldsymbol{v}\|_{H^\beta(\mathcal{D})^d} \leq C {\rho_{X,\mathcal{D}}}^{\tau-\beta} \|\boldsymbol{v}\|_{H^\beta(\mathcal{D})^d}. \tag{22}$$

■

Remark 28 *Note that in the matching case, we do not need* quasi-uniform *point sets, i.e. point sets with $\rho_{X,\mathcal{D}} \leq \rho_0$, as the error bound is solely expressed in terms of the fill distance $h_{X,\mathcal{D}}$. In the non-matching case, however, without quasi-uniformity, the ratio $\rho_{X,\mathcal{D}}$ may diverge, leading to a deterioration of the approximation order.*

Next, we want to discuss the interpolatory case, but now with unexpected noise in the data. Hence, instead of computing $\boldsymbol{m}_N$ using $\boldsymbol{v}(\boldsymbol{x}_j)$ we compute $\boldsymbol{m}_N$ using the data $\boldsymbol{v}(\boldsymbol{x}_j) + \boldsymbol{\epsilon}_j$. This leads to the following modified result.

Theorem 29 (Interpolation with Noise) *Under the assumptions of Theorem 27, let $\boldsymbol{y}_j = \boldsymbol{v}(\boldsymbol{x}_j) + \boldsymbol{\epsilon}_j$, $1 \le j \le N$, and let $\boldsymbol{m}_N$ be computed from Lemma 8 with $\sigma = 0$. Let $1 \le q \le \infty$ and $\gamma = \max\{2, q\}$. Then there are constants $h_0, C > 0$ such that for all X with $h_{X,\mathcal{D}} \le h_0$ the following bounds hold, where the expectation is taken with respect to the distribution of the errors $\boldsymbol{\epsilon}_j$.*

1. *If $\beta \ge \tau$ then*

$$\begin{aligned}\mathbb{E}\left[(\|\boldsymbol{v} - \boldsymbol{m}_N\|_{W_q^s(\mathcal{D})^d})\right] &\le C h_{X,\mathcal{D}}^{\tau - s - d(1/2-1/q)_+} \mathcal{N}_\tau(\boldsymbol{v}, \boldsymbol{m}) \\ &+ C\left(h_{X,\mathcal{D}}^{-s-d(1/2-1/q)_+ + d/2} \rho_{X,\mathcal{D}}^{\tau - d/2} + h_{X,\mathcal{D}}^{d/\gamma - s}\right)\left(\mathbb{E}\left[\sum_{j=1}^N \|\boldsymbol{\epsilon}_j\|_2^2\right]\right)^{1/2}\end{aligned}$$

for all $0 \le s \le \ell(\tau, q, d)$ with $\ell(\tau, q, d)$ given in Theorem 17 and $s \in \mathbb{N}_0$ if $q = \infty$.

2. *If $\beta < \tau$ then*

$$\begin{aligned}\mathbb{E}\left[(\|\boldsymbol{v} - \boldsymbol{m}_N\|_{W_q^s(\mathcal{D})^d})\right] &\le C h_{X,\mathcal{D}}^{\beta - s - d(1/2-1/q)_+} \rho_{X,\mathcal{D}}^{\tau-\beta} \mathcal{N}_\beta(\boldsymbol{v}, \boldsymbol{m}) \\ &+ C\left(h_{X,\mathcal{D}}^{-s-d(1/2-1/q)_+ + d/2} \rho_{X,\mathcal{D}}^{\tau - d/2} + h_{X,\mathcal{D}}^{d/\gamma - s}\right)\left(\mathbb{E}\left[\sum_{j=1}^N \|\boldsymbol{\epsilon}_j\|_2^2\right]\right)^{1/2}\end{aligned}$$

for all $0 \le s \le \ell(\beta, q, d)$ and $s \in \mathbb{N}_0$ if $q = \infty$.

Proof In this situation, we have $\boldsymbol{m}_N = \boldsymbol{m} - \boldsymbol{I}_X \boldsymbol{m} + \boldsymbol{I}_X \boldsymbol{v} + \boldsymbol{I}_X \boldsymbol{\mathcal{E}}$ with $\boldsymbol{\mathcal{E}} = (\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_N)$. Thus, the bound (21) on $\|\boldsymbol{v} - \boldsymbol{m}_N\|_{W_q^s(\mathcal{D})^d}$ now becomes

$$\|\boldsymbol{v} - \boldsymbol{m}_N\|_{W_q^s(\mathcal{D})^d} \le \|\boldsymbol{v} - \boldsymbol{I}_X \boldsymbol{v}\|_{W_q^s(\mathcal{D})^d} + \|\boldsymbol{m} - \boldsymbol{I}_X \boldsymbol{m}\|_{W_q^s(\mathcal{D})^d} + \|\boldsymbol{I}_X \boldsymbol{\mathcal{E}}\|_{W_q^s(\mathcal{D})^d}.$$

As we have already bounded the first two terms, we only need to derive a bound on the interpolant of the error. As the interpolant belongs to $H^\tau(\mathcal{D})^d$, we can use the sampling inequality in the form

$$\|\boldsymbol{I}_X \boldsymbol{\mathcal{E}}\|_{W_q^s(\mathcal{D})^d} \le C\left(h^{\tau - s - d(1/2-1/q)_+} \|\boldsymbol{I}_X \boldsymbol{\mathcal{E}}\|_{H^\tau(\mathcal{D})^d} + h_{X,\mathcal{D}}^{d/\gamma - s} \|\boldsymbol{I}_X \boldsymbol{\mathcal{E}}\|_{\ell_2(X)^d}\right). \tag{23}$$

On the one hand, we obviously have $\|\boldsymbol{I}_X \boldsymbol{\mathcal{E}}\|_{\ell_2(X)^d}^2 = \sum_{j=1}^N \|\boldsymbol{\epsilon}_j\|_2^2$. On the other hand, using again the equivalence of norms from Theorem 22, we see that

$$\|\boldsymbol{I}_X \boldsymbol{\mathcal{E}}\|_{H^\tau(\mathcal{D})}^2 \le C \boldsymbol{\mathcal{E}}^{\mathrm{T}} \boldsymbol{K}(X, X)^{-1} \boldsymbol{\mathcal{E}} \le C q_X^{-2\tau + d} \|\boldsymbol{\mathcal{E}}\|_2^2,$$

where the last inequality follows from (Fuselier, 2008b, Theorems 6 and 7), which state that there is a constant $C > 0$ such that the smallest eigenvalue of the interpolation matrix can be bounded as

$$\lambda_{\min}(\boldsymbol{K}(X, X)) \ge C q_X^{-d-2} \inf_{\|\boldsymbol{\omega}\|_2 \le 1/q_X} \widehat{\Phi}(\boldsymbol{\omega}).$$

Hence, inequality (23) can be bounded by

$$\begin{aligned}
\|\boldsymbol{I}_X\boldsymbol{\mathcal{E}}\|_{W_q^s(\mathcal{D})^d} &\le C h_{X,\mathcal{D}}^{\tau-s-d(1/2-1/q)_+} q_X^{-\tau+d/2}\|\boldsymbol{\mathcal{E}}\|_2 + C h_{X,\mathcal{D}}^{d/\gamma-s}\left(\sum_{j=1}^N \|\boldsymbol{\epsilon}_j\|_2^2\right)^{1/2} \\
&\le C(h_{X,\mathcal{D}}^{-s-d(1/2-1/q)_+ +d/2}\rho_{X,\mathcal{D}}^{\tau-d/2} + h_{X,\mathcal{D}}^{d/\gamma-s})\left(\sum_{j=1}^N \|\boldsymbol{\epsilon}_j\|_2^2\right)^{1/2}.
\end{aligned}$$

We obtain the result by combining the estimates from Theorem 27 and the above estimate. ∎

4.2 Error Analysis for the Predictive Mean using Approximation

In this section, we will assume that the measurements at $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ of a vector-valued function $\boldsymbol{v} : \mathcal{D} \to \mathbb{R}$ are contaminated, i.e. that we observe $\boldsymbol{y}_j = \boldsymbol{v}(\boldsymbol{x}_j) + \boldsymbol{\epsilon}_j$, $1 \le j \le N$. However, for the time being, we will not make any assumption on the distribution of the noise, even allowing no noise at all. The difference to the previous section is that we now use not the interpolation operator but rather the operator $\boldsymbol{Q}_{X,\lambda} : C(\mathcal{D}) \to H^\tau(\Omega; \mathrm{div})$ from Theorem 15 with a general smoothing parameter $\lambda > 0$ and the kernel $\boldsymbol{K}_{\mathrm{div}}$. The approximation operator indeed maps to $H^\tau(\Omega; \mathrm{div})$ by Theorem 22. Of course, the same holds true in the rotation-free case.

To apply the sampling inequality from Theorem 17, we need the following bounds which are the vector-valued version of similar bounds in the scalar-valued case.

Lemma 30 *Let the assumptions on the domain and kernel of Theorem 22 hold. Let $\boldsymbol{K} = \boldsymbol{K}_{\mathrm{div}}$ or $\boldsymbol{K} = \boldsymbol{K}_{\mathrm{curl}}$, and $\mathcal{H} = H^\tau(\mathcal{D}, \mathrm{div})$ or $\mathcal{H} = H^\tau(\mathcal{D}, \mathrm{curl})$, respectively, equipped with the inner product in which $\boldsymbol{K}$ is the reproducing kernel. Let $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ and $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$ be given by $\boldsymbol{y}_j = \boldsymbol{v}(\boldsymbol{x}_j) + \boldsymbol{\epsilon}_j$, $1 \le j \le N$, with a function $\boldsymbol{v} \in \mathcal{H}$ and including the possibility $\boldsymbol{\epsilon}_j = \boldsymbol{0}$. If $\boldsymbol{Q}_{X,\lambda}$ is the approximation operator from Theorem 15 and $\boldsymbol{\mathcal{E}} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N)$ then*

$$\begin{aligned}
\|\boldsymbol{Y} - (\boldsymbol{Q}_{X,\lambda}\boldsymbol{Y})|X\|_{\ell_2} &\le \left(\|\boldsymbol{\mathcal{E}}\|_{\ell_2}^2 + \lambda\|\boldsymbol{v}\|_{\mathcal{H}}^2\right)^{1/2}, \\
\|\boldsymbol{Q}_{X,\lambda}\boldsymbol{Y}\|_{\mathcal{H}} &\le \left(\lambda^{-1}\|\boldsymbol{\mathcal{E}}\|_{\ell_2}^2 + \|\boldsymbol{v}\|_{\mathcal{H}}^2\right)^{1/2}.
\end{aligned}$$

Proof As $\boldsymbol{Q}_{X,\lambda}\boldsymbol{Y}$ minimizes the penalized least-squares expression, we have

$$\begin{aligned}
\|\boldsymbol{Y} - (\boldsymbol{Q}_{X,\lambda}\boldsymbol{Y})|X\|_{\ell_2}^2 + \lambda\|\boldsymbol{Q}_{X,\lambda}\boldsymbol{Y}\|_{\mathcal{H}}^2 &= \sum_{j=1}^N \|\boldsymbol{y}_j - \boldsymbol{Q}_{X,\lambda}\boldsymbol{Y}(\boldsymbol{x}_j)\|_2^2 + \lambda\|\boldsymbol{Q}_{X,\lambda}\boldsymbol{Y}\|_{\mathcal{H}}^2 \\
&\le \sum_{j=1}^N \|\boldsymbol{y}_j - \boldsymbol{v}(\boldsymbol{x}_j)\|_2^2 + \lambda\|\boldsymbol{v}\|_{\mathcal{H}}^2 \\
&= \sum_{j=1}^N \|\boldsymbol{\epsilon}_j\|_2^2 + \lambda\|\boldsymbol{v}\|_{\mathcal{H}}^2 \\
&= \|\boldsymbol{\mathcal{E}}\|_{\ell_2}^2 + \lambda\|\boldsymbol{v}\|_{\mathcal{H}}^2,
\end{aligned}$$

which proves both statements. ■

We will also need the following auxiliary result for considering approximation with mismatched smoothness. To state it, we need to recall *band-limited* functions, which are functions having a compactly supported Fourier transform. For a parameter $\delta > 0$, we will use the spaces

$$\begin{aligned} \mathcal{B}^\delta &:= \{\boldsymbol{v} \in L_2(\mathbb{R}^d)^d : \operatorname{supp} \widehat{\boldsymbol{v}} \subseteq B_\delta[0]\}, \\ \widetilde{B}^\delta &:= \left\{\boldsymbol{v} \in \mathcal{B}_\delta : \int_{\mathbb{R}^d} \frac{\|\widehat{\boldsymbol{v}}(\boldsymbol{\omega})\|_2^2}{\|\boldsymbol{\omega}\|_2^2} d\boldsymbol{\omega} < \infty \right\}, \end{aligned}$$

where $B_\delta[0]$ denotes the closed unit ball in $\mathbb{R}^d$ with radius $\delta > 0$. Moreover, we need the sub-spaces $\widetilde{\mathcal{B}}^\delta_{\rm div}$ and $\widetilde{\mathcal{B}}^\delta_{\rm curl}$ of $\widetilde{\mathcal{B}}^\delta$, containing the divergence-free and rotation-free vector fields, respectively. For any $\boldsymbol{v} \in \widetilde{\mathcal{B}}^\delta_{\rm div}$ we can use the compact support of $\widehat{\boldsymbol{v}}$ in $B_\delta[0]$ and $\delta \geq 1$ to derive

$$\begin{aligned} (2\pi)^d \|\boldsymbol{v}\|^2_{\widetilde{H}^\tau(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} \frac{\|\widehat{\boldsymbol{v}}(\boldsymbol{\omega})\|_2^2}{\|\boldsymbol{\omega}\|_2^2}(1+\|\boldsymbol{\omega}\|_2^2)^{\tau+1} d\boldsymbol{\omega} \\ &= \int_{\|\boldsymbol{\omega}\|_2 \le \delta} \frac{\|\widehat{\boldsymbol{v}}(\boldsymbol{\omega})\|_2^2}{\|\boldsymbol{\omega}\|_2^2}(1+\|\boldsymbol{\omega}\|_2^2)^{\beta+1}(1+\|\boldsymbol{\omega}\|_2^2)^{\tau-\beta} d\boldsymbol{\omega} \\ &\le 2^{\tau-\beta}\delta^{2(\tau-\beta)} \int_{\|\boldsymbol{\omega}\|_2 \le \delta} \frac{\|\widehat{\boldsymbol{v}}(\boldsymbol{\omega})\|_2^2}{\|\boldsymbol{\omega}\|_2^2}(1+\|\boldsymbol{\omega}\|_2^2)^{\beta+1} d\boldsymbol{\omega} \end{aligned}$$

whenever $0 \le \beta \le \tau$. This is the first result of the following Lemma, which also summarizes Theorems 1 and 2 of Fuselier (2008a), noting that $t = 0$ is also possible there, as was the case in (Narcowich et al., 2006, Theorem 3.4).

Lemma 31 *1. If $\tau \ge \beta \ge 0$, $\delta \ge 1$ and $\boldsymbol{v} \in \widetilde{\mathcal{B}}^\delta_{\rm div}$ then*

$$\|\boldsymbol{v}\|_{\widetilde{H}^\tau(\mathbb{R}^d)^d} \le 2^{(\tau-\beta)/2}\delta^{\tau-\beta}\|\boldsymbol{v}\|_{\widetilde{H}^\beta(\mathbb{R}^d)^d}.$$

2. Let $\tau \ge \beta > d/2$. Given $\boldsymbol{v} \in \widetilde{H}^\tau(\mathbb{R}^d; \mathrm{div})$ and a point set $X \subseteq \mathbb{R}^d$ with separation radius q_X, there exists a function $\boldsymbol{v}_\delta \in \widetilde{\mathcal{B}}^\delta_{\rm div}$ with $\boldsymbol{v}_\delta|X = \boldsymbol{v}|X$ and

$$\|\boldsymbol{v} - \boldsymbol{v}_\delta\|_{\widetilde{H}^\beta(\mathbb{R}^d)^d} \le 5\kappa^{\beta-\tau} q_X^{\tau-\beta} \|\boldsymbol{v}\|_{\widetilde{H}^\tau(\mathbb{R}^d)^d},$$

with $\delta = \kappa/q_X$, where $\kappa \ge 1$ depends only on β and d.

3. The same statements hold for rotation-free functions.

We start our investigation on the approximation error by first falsely assuming that there is an error though the values are actually exact. This leads to the following result. As before, we assume that the prior mean has the same regularity as the unknown function.

Theorem 32 (Approximation without Noise) *Let Assumption 23 on the domain $\mathcal{D}$, Assumption 24 on the kernel $\boldsymbol{K}$ and Assumption 26 on the kernel $\boldsymbol{K}$, the target function $\boldsymbol{v}$ and mean $\boldsymbol{m}$ hold. Let $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subseteq \mathcal{D}$ and let $\boldsymbol{y}_j = \boldsymbol{v}(\boldsymbol{x}_j)$, $1 \le j \le N$. Finally, given $\lambda > 0$ let $\boldsymbol{m}_{N,\lambda}$ be the predictive, vector-valued mean, based on the information $(\boldsymbol{x}_j, \boldsymbol{y}_j)$, $1 \le j \le N$. Let $1 \le q \le \infty$ and $\gamma = \max\{2, q\}$. Then, there are constants $h_0, C > 0$ such that for all X with $h_{X,\mathcal{D}} \le h_0$ the following bounds hold.*

1. *If $\beta \ge \tau$ then*

$$\|\boldsymbol{v} - \boldsymbol{m}_{N,\lambda}\|_{W_q^s(\mathcal{D})^d} \le C \left(h_{X,\mathcal{D}}^{\tau - s - d(1/2-1/q)_+} + \sqrt{\lambda} h_{X,\mathcal{D}}^{d/\gamma - s} \right) \mathcal{N}_\tau(\boldsymbol{v}, \boldsymbol{m})$$

   for all $0 \le s \le \ell(\tau, q, d)$ with $\ell(\tau, q, d)$ given in Theorem 17 and $s \in \mathbb{N}_0$ if $q = \infty$.

2. *If $\beta < \tau$ then*

$$\begin{aligned} \|\boldsymbol{v} - \boldsymbol{m}_{N,\lambda}\|_{W_q^s(\mathcal{D})^d} &\le C h_{X,\mathcal{D}}^{\beta - s - d(1/2-1/q)_+} \rho_{X,\mathcal{D}}^{\tau-\beta} \mathcal{N}_\beta(\boldsymbol{v}, \boldsymbol{m}) \\ &\quad + C\sqrt{\lambda} h_{X,\mathcal{D}}^{\beta - \tau + d/\gamma - s} \rho_{X,\mathcal{D}}^{\tau-\beta} \mathcal{N}_\beta(\boldsymbol{v}, \boldsymbol{m}) \end{aligned}$$

   for all $0 \le s \le \ell(\beta, q, d)$ and $s \in \mathbb{N}_0$ if $q = \infty$.

Proof As we assume no noise, we have $\boldsymbol{m}_{N,\lambda} = \boldsymbol{m} - \boldsymbol{Q}_{X,\lambda}\boldsymbol{m} + \boldsymbol{Q}_{X,\lambda}\boldsymbol{v}$, which allows us to split the error again as

$$\|\boldsymbol{v} - \boldsymbol{m}_{N,\lambda}\|_{W_q^s(\mathcal{D})^d} \le \|\boldsymbol{v} - \boldsymbol{Q}_{X,\lambda}\boldsymbol{v}\|_{W_q^s(\mathcal{D})^d} + \|\boldsymbol{m} - \boldsymbol{Q}_{X,\lambda}\boldsymbol{m}\|_{W_q^s(\mathcal{D})^d}.$$

As we assume $\boldsymbol{v}$ and $\boldsymbol{m}$ to have the same regularity, we concentrate again on $\boldsymbol{v}$. Here, the sampling inequality from Theorem 17 and Lemma 30 with $\boldsymbol{\mathcal{E}} = \boldsymbol{0}$ immediately yield in the case $\beta \ge \tau$,

$$\begin{aligned} \|\boldsymbol{v} - \boldsymbol{Q}_{X,\lambda}\boldsymbol{v}\|_{W_q^s(\mathcal{D})^d} &\le C h^{\tau - s - d(1/2-1/q)_+} \|\boldsymbol{v} - \boldsymbol{Q}_{X,\lambda}\boldsymbol{v}\|_{H^\tau(\mathcal{D})^d} \\ &\quad + C h^{d/\gamma - s} \|\boldsymbol{v} - \boldsymbol{Q}_{X,\lambda}\boldsymbol{v}\|_{\ell_2(X)^d} \\ &\le C h^{\tau - s - d(1/2-1/q)_+} \|\boldsymbol{v}\|_{H^\tau(\mathcal{D})^d} \\ &\quad + C h^{d/\gamma - s} \sqrt{\lambda} \|\boldsymbol{v}\|_{H^\tau(\mathcal{D})^d}, \end{aligned}$$

where we have set $h = h_{X,\mathcal{D}}$.

For the second statement, we can again concentrate on $\boldsymbol{v}$, as the same arguments apply to $\boldsymbol{m}$. For the extended function $E_{\mathrm{div}}\boldsymbol{v} \in \widetilde{H}^\beta(\mathbb{R}^d; \mathrm{div})$, we choose an interpolatory band-limited function $\boldsymbol{v}_\delta \in \widetilde{B}_{\mathrm{div}}^\delta$ with $\delta = \kappa/q_X$, according to the second point of Lemma 31. Then, we start with

$$\|\boldsymbol{v} - \boldsymbol{Q}_{X,\lambda}\boldsymbol{v}\|_{W_q^s(\mathcal{D})^d} \le \|\boldsymbol{v} - \boldsymbol{v}_\delta\|_{W_q^s(\mathcal{D})^d} + \|\boldsymbol{v}_\delta - \boldsymbol{Q}_{X,\lambda}\boldsymbol{v}\|_{W_q^s(\mathcal{D})^d}. \tag{24}$$

To bound the first term, we use the sampling inequality from Theorem 17, noting $\boldsymbol{v}|X = \boldsymbol{v}_\delta|X$, yielding

$$\begin{aligned} \|\boldsymbol{v} - \boldsymbol{v}_\delta\|_{W_q^s(\mathcal{D})^d} &\le C h^{\beta - s - d(1/2-1/q)_+} \|\boldsymbol{v} - \boldsymbol{v}_\delta\|_{H^\beta(\mathcal{D})^d} \\ &\le C h^{\beta - s - d(1/2-1/q)_+} \|\boldsymbol{v}\|_{H^\beta(\mathcal{D})^d}, \end{aligned}$$

where the last inequality follows via

$$\|\boldsymbol{v}-\boldsymbol{v}_\delta\|_{H^\beta(\mathcal{D})^d} \le \|E_{\mathrm{div}}\boldsymbol{v}-\boldsymbol{v}_\delta\|_{\widetilde{H}^\beta(\mathbb{R}^d)^d} \le C\|E_{\mathrm{div}}\boldsymbol{v}\|_{\widetilde{H}^\beta(\mathbb{R}^d)^d} \le C\|\boldsymbol{v}\|_{H^\beta(\mathcal{D})^d},$$

where we have used Lemma 31 with $\beta=\tau$ and the boundedness of the extension operator.

For the second term in (24), we first note that we have $\boldsymbol{Q}_{X,\lambda}\boldsymbol{v}=\boldsymbol{Q}_{X,\lambda}\boldsymbol{v}_\delta$, as the band-limited function has the same values on X as the function and the approximation operator uses only these values. Thus, we can once again use the sampling inequality to derive

$$\begin{aligned}\|\boldsymbol{v}_\delta-\boldsymbol{Q}_{X,\lambda}\boldsymbol{v}_\delta\|_{W_q^s(\mathcal{D})^d} &\le h^{\tau-s-d(1/2-1/q)_+}\|\boldsymbol{v}_\delta-\boldsymbol{Q}_{X,\lambda}\boldsymbol{v}_\delta\|_{H^\tau(\mathcal{D})^d}\\ &\quad + Ch^{d/\gamma-s}\|\boldsymbol{v}_\delta-\boldsymbol{Q}_{X,\lambda}\boldsymbol{v}_\delta\|_{\ell_2(X)^d}.\end{aligned}$$

To further bound the norms on the right-hand side, we use Lemma 30 in combination with Lemma 31 and the equivalence of norms from Theorem 22. This yields on the one hand

$$\begin{aligned}\|\boldsymbol{v}_\delta-\boldsymbol{Q}_{X,\lambda}\boldsymbol{v}_\delta\|_{H^\tau(\mathcal{D})^d} &\le C\|\boldsymbol{v}_\delta\|_{H^\tau(\mathcal{D})^d} \le C\|\boldsymbol{v}_\delta\|_{\widetilde{H}^\tau(\mathbb{R}^d)^d} \le Cq_X^{\beta-\tau}\|\boldsymbol{v}_\delta\|_{\widetilde{H}^\beta(\mathbb{R}^d)^d}\\ &\le Ch^{\beta-\tau}\rho^{\tau-\beta}\|\boldsymbol{v}\|_{H^\beta(\mathcal{D})^d},\end{aligned}$$

where the last inequality follows from

$$\begin{aligned}\|\boldsymbol{v}_\delta\|_{\widetilde{H}^\beta(\mathbb{R}^d)^d} &\le \|\boldsymbol{v}_\delta-E_{\mathrm{div}}\boldsymbol{v}\|_{\widetilde{H}^\beta(\mathbb{R}^d)^d}+\|E_{\mathrm{div}}\boldsymbol{v}\|_{\widetilde{H}^\beta(\mathbb{R}^d)^d} \le C\|E_{\mathrm{div}}\boldsymbol{v}\|_{\widetilde{H}^\beta(\mathbb{R}^d)^d}\\ &\le C\|\boldsymbol{v}\|_{H^\beta(\mathcal{D})^d}.\end{aligned}$$

On the other hand, from Lemma 30 with $\mathcal{E}=0$ we have

$$\|\boldsymbol{v}_\delta-\boldsymbol{Q}_{X,\lambda}\boldsymbol{v}_\delta\|_{\ell_2(X)^d} \le C\sqrt{\lambda}\|\boldsymbol{v}_\delta\|_{H^\tau(\mathcal{D})} \le C\sqrt{\lambda}h^{\beta-\tau}\rho^{\tau-\beta}\|\boldsymbol{v}\|_{H^\beta(\mathcal{D})}.$$

■

Remark 33 *The above result shows that we can choose the* smoothing parameter λ *as a function of* $h_{X,\mathcal{D}}$ *to obtain the same convergence results as in the interpolatory case.*

Next, we discuss the approximation properties of the approximation operator in the case of noisy data in the most general situation, i.e. we do not make any assumptions on the noise.

Theorem 34 (Approximation with Noise) *Let the assumptions of Theorem 32 hold, except that now the data are given by* $\boldsymbol{y}_j=\boldsymbol{v}(\boldsymbol{x}_j)+\boldsymbol{\epsilon}_j$, $1\le j\le N$. *Again, given* $\lambda>0$, *let* $\boldsymbol{m}_{N,\lambda}$ *be the predictive, vector-valued mean, based on the information* $(\boldsymbol{x}_j,\boldsymbol{y}_j)$, $1\le j\le N$. *Let* $1\le q\le\infty$ *and* $\gamma=\max\{2,q\}$. *Then, there are constants* $h_0, C>0$ *such that for all* X *with* $h_{X,\mathcal{D}}\le h_0$ *the following bounds hold, where the expectation is taken with respect to the distribution of the errors* $\boldsymbol{\epsilon}_j$.

1. *If $\beta \geq \tau$ then*

$$\mathbb{E}\left[\|\boldsymbol{v}-\boldsymbol{m}_{N,\lambda}\|_{W_q^s(\mathcal{D})^d}\right] \leq C\left(h_{X,\mathcal{D}}^{\tau-s-d(1/2-1/q)_+}+\sqrt{\lambda}h_{X,\mathcal{D}}^{d/\gamma-s}\right)\mathcal{N}_\tau(\boldsymbol{v},\boldsymbol{m})$$
$$+C\left(\frac{1}{\sqrt{\lambda}}h_{X,\mathcal{D}}^{\tau-s-d(1/2-1/q)_+}+h_{X,\mathcal{D}}^{d/\gamma-s}\right)\mathbb{E}\left[\left(\sum_{j=1}^N\|\boldsymbol{\epsilon}_j\|_2^2\right)^{1/2}\right]$$

for all $0 \leq s \leq \ell(\tau,q,d)$ with $\ell(\tau,q,d)$ given in Theorem 17 and $s \in \mathbb{N}_0$ if $q=\infty$.

2. *If $\beta < \tau$ then*

$$\mathbb{E}\left[\|\boldsymbol{v}-\boldsymbol{m}_{N,\lambda}\|_{W_q^s(\mathcal{D})^d}\right] \leq Ch_{X,\mathcal{D}}^{\beta-s-d(1/2-1/q)_+}\rho_{X,\mathcal{D}}^{\tau-\beta}\mathcal{N}_\beta(\boldsymbol{v},\boldsymbol{m})$$
$$+C\sqrt{\lambda}h_{X,\mathcal{D}}^{\beta-\tau+d/\gamma-s}\rho_{X,\mathcal{D}}^{\tau-\beta}\mathcal{N}_\beta(\boldsymbol{v},\boldsymbol{m})$$
$$+C\left(\frac{1}{\sqrt{\lambda}}h_{X,\mathcal{D}}^{\tau-s-d(1/2-1/q)_+}+h_{X,\mathcal{D}}^{d/\gamma-s}\right)\mathbb{E}\left[\left(\sum_{j=1}^N\|\boldsymbol{\epsilon}_j\|_2^2\right)^{1/2}\right]$$

for all $0 \leq s \leq \ell(\beta,q,d)$ and $s \in \mathbb{N}_0$ if $q=\infty$.

Proof In this situation, we have $\boldsymbol{m}_{N,\lambda}=\boldsymbol{m}-\boldsymbol{Q}_{X,\lambda}\boldsymbol{m}+\boldsymbol{Q}_{X,\lambda}\boldsymbol{v}+\boldsymbol{Q}_{X,\lambda}\boldsymbol{\mathcal{E}}$, where again $\boldsymbol{\mathcal{E}}=(\boldsymbol{\epsilon}_1,\dots,\boldsymbol{e}_N)$. Thus, in addition to the error bound in Theorem 32 we have the additional term $\|\boldsymbol{Q}_{X,\lambda}\boldsymbol{\mathcal{E}}\|_{W_q^s(\mathcal{D})^d}$ that we need to bound. On the one hand Lemma 30 for the case $\boldsymbol{v}=\boldsymbol{0}$, together with Theorem 22, yields the bound $\|\boldsymbol{Q}_{X,\lambda}\boldsymbol{\mathcal{E}}\|_{H^\tau(\mathcal{D})^d} \leq \frac{C}{\sqrt{\lambda}}\|\boldsymbol{\mathcal{E}}\|_{\ell_2}$, and, on the other hand, also the bound

$$\|\boldsymbol{Q}_{X,\lambda}\boldsymbol{\mathcal{E}}|X\|_{\ell_2} \leq \|\boldsymbol{\mathcal{E}}-\boldsymbol{Q}_{X,\lambda}\boldsymbol{\mathcal{E}}|X\|_{\ell_2}+\|\boldsymbol{\mathcal{E}}\|_{\ell_2} \leq 2\|\boldsymbol{\mathcal{E}}\|_{\ell_2}.$$

Thus the sampling inequality yields in this case

$$\begin{aligned}\|\boldsymbol{Q}_{X,\lambda}\boldsymbol{\mathcal{E}}\|_{W_q^s(\mathcal{D})^d} &\leq Ch^{\tau-s-d(1/2-1/q)_+}\|\boldsymbol{Q}_{X,\lambda}\boldsymbol{\mathcal{E}}\|_{H^\tau(\mathcal{D})^d}\\ &\quad +Ch^{d/\gamma-s}\|\boldsymbol{Q}_{X,\lambda}\boldsymbol{\mathcal{E}}\|_{\ell_2(X)^d}\\ &\leq Ch^{\tau-s-d(1/2-1/q)_+}\frac{1}{\sqrt{\lambda}}\|\boldsymbol{\mathcal{E}}\|_{\ell_2}\\ &\quad +Ch^{d/\gamma-s}\|\boldsymbol{\mathcal{E}}\|_{\ell_2}.\end{aligned}$$

Together with the error bound from Theorem 32, this yields the stated error estimate. ■

Remark 35 *The above theorem is the result for vector-valued divergence- or rotation-free functions corresponding to (Wynne et al., 2021, Theorem 4) for the scalar-valued case. Though, we essentially achieve the same approximation orders, the proof of (Wynne et al., 2021, Theorem 4) seems problematic, as in formula (18) the authors essentially use a bound of the form*

$$\|I_Xf\|_{H^\tau(\mathcal{D})} \leq Cq_X^{\beta-\tau}\|f\|_{H^\beta(\mathcal{D})},$$

which they claim can be derived from the proof of (Narcowich et al., 2006, Theorem 4.2). We do not believe that this is the case, as the above bound would imply the inverse estimate

$$\|s\|_{H^\tau(\mathcal{D})} \le C q_X^{\beta-\tau} \|s\|_{H^\beta(\mathcal{D})}, \qquad s \in V_X,$$

which is still an open question for general bounded domains. Nonetheless, the results of (Wynne et al., 2021, Theorem 4) are correct, as our proof can essentially be given also in the scalar-valued case.

Remark 36 *Note that in the error bound of the latter theorem, there are apparently competing terms on how to choose the smoothing parameter λ. To cast more light on this, we look at the specific situation of $q = 2$ and $s = 0$. If we also assume quasi-uniform data sets then the error bound for the case $\beta < \tau$, becomes*

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{v} - \boldsymbol{m}_{N,\lambda}\|_{L_2(\mathcal{D})^d}\right] \;&\le\; C\left(h_{X,\mathcal{D}}^{\beta} + \sqrt{\lambda} h_{X,\mathcal{D}}^{\beta-\tau+d/2}\right)\mathcal{N}_\beta(\boldsymbol{v},\boldsymbol{m}) \\ &\quad + C\left(\frac{1}{\sqrt{\lambda}} h_{X,\mathcal{D}}^{\tau} + h_{X,\mathcal{D}}^{d/2}\right)\mathbb{E}\left[\left(\sum_{j=1}^N \|\boldsymbol{\epsilon}_j\|_2^2\right)^{1/2}\right].\end{aligned}$$

Thus, the competing terms can for example be balanced by choosing the smoothing parameter as $\sqrt{\lambda} = h_{X,\mathcal{D}}^{\tau-\beta/2}$, which yields

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{v} - \boldsymbol{m}_{N,\lambda}\|_{L_2(\mathcal{D})^d}\right] \;&\le\; C h_{X,\mathcal{D}}^{\min\{\beta,(\beta+d)/2\}}\mathcal{N}_\beta(\boldsymbol{v},\boldsymbol{m}) \\ &\quad + h_{X,\mathcal{D}}^{\min\{\beta/2,d/2\}}\mathbb{E}\left[\left(\sum_{j=1}^N \|\boldsymbol{\epsilon}_j\|_2^2\right)^{1/2}\right].\end{aligned}$$

For our final result, we assume that each noise $\boldsymbol{\epsilon}_j$ has an independent normal distribution with variance $\sigma^2 I_d$ and that we use the approximation operator with the correct parameter $\lambda = \sigma$.

Corollary 37 *Let the assumptions of Theorem 34 hold. Assume that $\boldsymbol{\epsilon}_j \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I_d)$, $1 \le j \le N$.*

1. *If $\beta \ge \tau$ then*

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{v} - \boldsymbol{m}_{N,\sigma}\|_{W_q^s(\mathcal{D})^d}\right] &\le C\left(h_{X,\mathcal{D}}^{\tau-s-d(1/2-1/q)_+} + \sigma h_{X,\mathcal{D}}^{d/\gamma-s}\right)\mathcal{N}_\tau(\boldsymbol{v},\boldsymbol{m}) \\ &\quad + C\rho_{X,\mathcal{D}}^{d/2}\left(h_{X,\mathcal{D}}^{\tau-s-d(1/2-1/q)_+-d/2} + \sigma h_{X,\mathcal{D}}^{d/\gamma-s-d/2}\right).\end{aligned}$$

   for all $0 \le s \le \ell(\tau, q, d)$ and $s \in \mathbb{N}_0$ if $q = \infty$.

2. *If $\beta < \tau$ then*

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{v} - \boldsymbol{m}_{N,\sigma}\|_{W_q^s(\mathcal{D})^d}\right] &\le C h_{X,\mathcal{D}}^{\beta-s-d(1/2-1/q)_+}\rho_{X,\mathcal{D}}^{\tau-\beta}\mathcal{N}_\beta(\boldsymbol{v},\boldsymbol{m}) \\ &\quad + C\sigma h_{X,\mathcal{D}}^{\beta-\tau+d/\gamma-s}\rho_{X,\mathcal{D}}^{\tau-\beta}\mathcal{N}_\beta(\boldsymbol{v},\boldsymbol{m}) \\ &\quad + C\rho_{X,\mathcal{D}}^{d/2}\left(h_{X,\mathcal{D}}^{\tau-s-d(1/2-1/q)_+-d/2} + \sigma h_{X,\mathcal{D}}^{d/\gamma-s-d/2}\right).\end{aligned}$$

   for all $0 \le s \le \ell(\beta, q, d)$ and $s \in \mathbb{N}_0$ if $q = \infty$.

Proof This immediately follows using the estimates from Theorem 34, together with $N \le C q_X^{-d} = C h_{X,\mathcal{D}}^{-d} \rho_{X,\mathcal{D}}^d$ and

$$\mathbb{E}\left[\left(\sum_{j=1}^{N} \|\boldsymbol{\epsilon}_j\|_2^2\right)^{1/2}\right] \le \left(\mathbb{E}\left[\sum_{j=1}^{N} \|\boldsymbol{\epsilon}_j\|_2^2\right]\right)^{1/2} = \sqrt{N}\sigma \ \le \ h_{X,\mathcal{D}}^{-d/2} \rho_{X,\mathcal{D}}^{d/2} \sigma.$$

■

Conclusion

We have introduced and characterised divergence- and rotation-free Gaussian processes. We have derived rigorous error estimates for the predictive means of such processes covering the cases of matching/non-matching smoothness and observations.

A possible next step in learning problems for example from fluid dynamics would be to follow Bilionis et al. (2013) and to use Gaussian processes with tensor product kernels to model velocity fields $\boldsymbol{v} : \mathcal{D} \times \mathcal{D}_T \times \mathcal{D}_p \to \mathbb{R}^3$ that not only depend on a spatial variable $\boldsymbol{x} \in \mathcal{D}$ but also on time $t \in \mathcal{D}_T$ and certain parameters $p \in \mathcal{D}_p$.

Acknowledgments

Quoc Thong Le Gia and Ian Sloan are grateful to the Australian Research Council for support under grant DP180100506. Holger Wendland is grateful for the support and hospitality of the Sydney Mathematical Research Institute (SMRI) during his visit in August/September 2023.

References

R. J. Adler. *An introduction to continuity, extrema, and related topics for general Gaussian processes*, volume 12 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1990.

M. A. Álvarez, L. Rosasco, and N. D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4:195–266, 2012.

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

R. Arcangéli, M. Cruz López de Silanes, and J. J. Torrens. Extension of sampling inequalities to Sobolev semi-norms of fraction order and derivative data. *Numer. Math.*, 121:587–608, 2012.

N. Aronszajn. Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68:337–404, 1950.

I. Bilionis, N. Zabaras, B. A. Konomi, and G. Lin. Multi-output separable Gaussian process: Towards an efficient fully Bayesian paradigm for uncertainty quantification. *J. Comput. Phys.*, 241:212–239, 2013.

E. V. Bonilla, K. Chain, and C. Williams. Multi-task gaussian process prediction. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

S. Brenner and L. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, New York, 3rd edition, 1994.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7:331–368, 2007.

A. Caponnetto, M. Pontil, C. Micchelli, and Y. Ying. Universal multi-task kernels. *J. Mach. Learn. Res.*, 9:1615–1646, 2008.

Claudio Carmeli, Ernesto De Vito, and Alessandro Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4: 377–408, 2006.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8:19–61, 2010.

E. Fuselier. Sobolev-type approximation rates for divergence-free and curl-free RBF interpolants. *Math. Comput.*, 77:1407–1423, 2008a.

E. Fuselier. Improved stability estimates and a characterization of the native space for matrix-valued RBFs. *Adv. Comput. Math.*, 29:269–290, 2008b.

P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman, Boston, 1985.

Hasan Gunes and Ulrich Rist. On the use of kriging for enhanced data reconstruction in a separated transitional flat-plate boundary layer. *Physics of Fluids*, 20:104109, 2008.

P. Li, Q. Wang, W. Zuo, and L. Zhang. Log-Euclidean kernels for sparse representation and dictionary learning. In *2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia*, pages 1601–1608, 2013.

C. Micchelli and M. Pontil. Kernels for multi–task learning. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

F. J. Narcowich and J. D. Ward. Generalized Hermite interpolation via matrix-valued conditionally positive definite functions. *Math. Comput.*, 63:661–687, 1994.

F. J. Narcowich, J. D. Ward, and H. Wendland. Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting. *Math. Comput.*, 74:643–763, 2005.

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

D. Nieman, B. Szabo, and H. van Zanten. Contraction rates for sparse variational approximations in gaussian process regression. *Journal of Machine Learning Research*, 23:1–26, 2022.

H. Owhadi. Gaussian process hydrodynamics. *Applied Mathematics and Mechanics*, 44: 1175–1198, 2023.

G. Parra and F. Tobar. Spectral mixture kernels for multi-output Gaussian processes. In I. Guyon, U. von Luxemburg, H. Wallach, R. Fergus, S. Vishawanathan, and R. Garnett, editors, *Advances in neural information processing systems 30 (NIPS 2017)*, 2017.

D. Pati, A. Bhattacharya, and G. Cheng. Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior. *Journal of Machine Learning Research*, 16:2837–2851, 2015.

George B. Pedrick. *Theory of Reproducing Kernels for Hilbert Spaces of Vector Valued Functions*. PhD thesis, University of Kansas, USA, 1957.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridte, Massachusetts, 2006.

M. Scheuerer, R. Schaback, and M. Schlather. Interpolation of spatial data—a stochastic or a deterministic problem? *European J. Appl. Math.*, 24(4):601–629, 2013.

B. Schölkopf and A. J. Smola. *Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts, 2002.

D. Schräder and H. Wendland. A high-order, analytically divergence-free discretization method for Darcy's problem. *Math. Comput.*, 80:263–277, 2011.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

A. M. Stuart and A. L. Teckentrup. Posterior consistency for Gaussian process approximation of Bayesian posterior distributions. *Math. Comput.*, 87:721–753, 2017.

A. L. Teckentrup. Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse problems. *SIAM/ASA J. Uncertainty Quantification*, 8:1310–1337, 2020.

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

N. Wahlström, M. Kok, T. B. Schön, and F. Gustafsson. Modeling magnetic fields using gaussian processes. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3522–3526, 2013.

H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, UK, 2005.

H. Wendland. Divergence-free kernel methods for approximating the Stokes problem. *SIAM J. Numer. Anal.*, 47:3158–3179, 2009.

H. Wendland and C. Rieger. Approximate interpolation with applications to selecting smoothing parameters. *Numer. Math.*, 101:643–662, 2005.

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

S. Zhou, D. Pati, T. Wang, Y. Yang, and R. J. Carroll. Gaussian processes with errors in variables: Theory and computation. *Journal of Machine Learning Research*, 24:1–53, 2023.