Reproducing Kernel Hilbert Spaces and Gaussian Process Regression

Nicholas Krämer

Last update: June 12, 2018

Abstract

This text is a summary of what I have learned about reproducing kernel Hilbert spaces and its connection to Gaussian process regression. I wrote it for myself in the context of the preparation for my master thesis about Gaussian processes and uncertainty quantification.

The goal was to understand a little bit more than the absolute minimum related to reproducing kernel Hilbert spaces and Gaussian process regression, mostly so that texts such as [ST18] will be accessible within a reasonable amount of work.

Contents

The text contains two parts: in the first part I summarise statements related to reproducing kernel Hilbert spaces and in the second part I do the same with Gaussian process regression. We will see that there is an abundance of connections between those two.

My main source for the first part is [Fas07]. Whenever a mathematically more precise treatment of the matter was necessary, I referred to [Wen05]. The content of the second part is close to what is presented in [RW06, Chapter 2, 4, 6]. I also relied on basic literature about measure theory, functional, and Fourier analysis; see [Dud02], [Alt02] or [SS07], just to name a few.

We sometimes abbreviate reproducing kernel Hilbert space by RKHS, radial basis function by RBF and Gaussian process by GP. By H' we denote the dual space of a vector space H and write $\mathrm{d}x := \mathrm{d}\lambda^d(x)$ for integration with respect to the Lebesgue measure.

Deliberately, we only consider real-valued functions. Most of the statements in section 2 and 3 admit straightforward extensions to complex-valued functions.

1. Scattered Data Interpolation

Reproducing kernel Hilbert spaces are a mathematical tool to analyse properties of kernel-based approximations such as scattered data interpolation. Scattered data interpolation is concerned with the following problem. Given some function f, a data set $X = \{x_1, ..., x_N\}$ and a kernel function $K(x, y)$, we want to approximate

$$f(x) \approx \mathcal{P}_f(x) = \sum_{j=1}^{N} c_j K(x, x_j). \tag{1}$$

Assuming that evaluations $f(X) := (f(x_1), ..., f(x_N))^T$ are available, this problem reduces to solving $Kc = f(X),\ K = (K(x_i, x_j))_{i,j=1}^N,\ c = (c_1, ..., c_N)^T$.One of the questions arising in this context is whether this problem has a unique solution. This is the case if and only if K is nonsingular, i.e. $\det K \neq 0$. Since hopefully this is independent of the pointset X, we have to figure out for which kernel functions K this is satisfied. This leads to the notion of positive definite functions.

2. Positive Definite Functions

Definition 2.1. A matrix $A \in \mathbb{R}^{N\times N}$ is *positive semidefinite* if for all $c \in \mathbb{R}^N$, its quadratic form $c^T Ac$ is nonnegative, $c^T Ac \geq 0$. If $c^T Ac = 0$ implies $c \equiv 0$, then we call *A positive definite.*

An important property of positive definite matrices is that a positive definite matrix has only positive eigenvalues and therefore a positive determinant. Thus, positive definite matrices are always invertible.

We want to make sure that the kernel matrix $K = (K(x_i, x_j))_{j=1}^N$ is positive definite on any pointset X.

Definition 2.2. A continuous function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is *positive definite* if for any N pairwise different points $x_1, ..., x_N \in \mathbb{R}^d$ and all $c = (c_1, ..., c_N) \in \mathbb{R}^N$,

$$\sum_{j=1}^{N}\sum_{i=1}^{N} c_i \Phi(x_i - x_j)c_j \geq 0.$$

If this being zero implies $c \equiv 0$, the function is *strictly positive definite.*

We say that a (kernel) function $K(x, y)$ is *stationary* if we can write it as $K(x, y) = \Phi(x-y)$. In light of the previous definition, this suggests using stationary kernels based on a strictly positive definite function for our approximation problem. In this case the kernel matrix K is positive definite, hence invertible.

There are many ways to characterise positive definite functions. One of the most popular ways is through Bochner's theorem which characterises positive definite functions in terms of the Fourier transform of a certain measure. It is an example for integral characterisations of positive definite functions.

We denote by $\langle \cdot, \cdot \rangle$ the scalar product in $\mathbb{R}^d$. By $\widehat{\mu}$ we denote the *Fourier transform* of a positive, finite measure μ which is given by

$$\widehat{\mu}(z) = \frac{1}{\sqrt{(2\pi)^d}} \int_{\mathbb{R}^d} e^{-i\langle z,\tau\rangle}\ \mathrm{d}\mu(\tau). \tag{2}$$

The same notation applies to the Fourier transform of functions, namely for a function $f \in L^1(\mathbb{R}^d)$ the Fourier transform $\widehat{f}$ is given by

$$\widehat{f}(z) = \frac{1}{\sqrt{(2\pi)^d}} \int_{\mathbb{R}^d} f(\tau)e^{-i\langle z,\tau\rangle}\ \mathrm{d}\tau. \tag{3}$$

The following theorem is stated as in [Fas07, Theorem 3.3]. This reference includes parts of the proof as well as further references.

Theorem 2.3 (Bochner). *A continuous function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if*

$$\varphi(z) = \widehat{\mu}(z) = \frac{1}{\sqrt{(2\pi)^d}} \int_{\mathbb{R}^d} e^{-i\langle z,\tau\rangle}\ \mathrm{d}\mu(\tau)$$

i.e. if it is the Fourier transform of a finite non-negative Borel measure μ on $\mathbb{R}^d$.

The measure μ is called the *spectral measure*, or *spectral distribution.* If μ is absolutely continuous with respect to the Lebesgue measure, its corresponding density is called the *spectral density.*

Using theorem 2.3 we can define positive definite functions.

Example 2.4 (Matérn Functions)**.** Let $\nu, \rho > 0$ and $S : \mathbb{R}^d \to [0, \infty]$ be given by

$$S(x) = \frac{\rho^d \Gamma(\nu + d/2)}{(2\nu\pi)^{d/2}\Gamma(\nu)}(1 + \rho^2\|x\|^2/(2\nu))^{-(\nu+d/2)}.$$

We use S as a spectral density to define a measure μ via $\mu(A) := \int_A S(y)\, \mathrm{d}\lambda^d(y)$, where λ^d stands for the d-dimensional Lebesgue measure and $A \in \mathscr{B}(\mathbb{R}^d)$ is an element in the Borel σ-algebra. The measure μ is a non-negative Borel measure and by

$$\int_{\mathbb{R}^d} S(y)\, \mathrm{d}\lambda^d(y) = \frac{\rho^d \Gamma(\nu + d/2)}{(2\nu\pi)^{d/2}\Gamma(\nu)} \int_{\mathbb{R}^d} (1 + \rho^2|y|^2/(2\nu))^{-(\nu+d/2)}\, \mathrm{d}\lambda^d(y) < \infty$$

it is finite. Hence we can apply theorem 2.3 and obtain the positive definite function

$$\varphi_{\nu,\rho}(r) = \frac{(\sqrt{2\nu}|r|/\rho)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\sqrt{2\nu}\|x\|/\rho)$$

where K_ν is the modified Bessel function of the second kind [Fas07]. The family of functions $(\varphi_{\nu,\rho})_{\nu,\rho}$ is known as the family of *Matérn functions.*

It can be shown that a Matérn function is strictly positive definite as the modified Bessel function of second kind is.

The parameter ν is called the *smoothness parameter* of $\varphi_{\nu,\rho}$. To understand that, note that $\varphi_{\nu,\rho}$ has regularity $C^k(\mathbb{R}^d)$ for $\nu = (d + k + 1)/2$. We call ρ *correlation length* of $\varphi_{\nu,\rho}$, it is also referred to as *characteristic length-scale* [RW06].

For half-integer values of $\nu = n + 1/2$, $n \in \mathbb{N}$ we can write the Matérn function as the product of an exponential and a polynomial of order n,

$$\varphi_{n+1/2}(x) = \exp\left(-\sqrt{2(n + 1/2)}x/\lambda\right) \frac{\Gamma(n + 1)}{\Gamma(2n + 1)} \sum_{j=0}^{n} \frac{(n + j)!}{j!(n - j)!} \left(\sqrt{8(n + 1/2)}x/\lambda\right)^{n-j}. \qquad (4)$$

In machine learning applications, the most common choices seem to be $\nu = 3/2$ and $\mu = 5/2$ [RW06, p. 85]. For $\nu \geq 7/2$, the kernel becomes rather smooth so that it becomes difficult to distinguish between finite and infinite values for ν; see [RW06, Chapter 4].

Example 2.5 (Exponential)**.** The exponential kernel with correlation length $\rho > 0$ is given by $\varphi(r) = e^{-r/\rho}$. On the one hand we can recover it from the Matérn functions, choosing $\nu = 1/2$. On the other hand, if we want to construct it using theorem 2.3, we can use the spectral density

$$S(x) = \frac{\rho^d \Gamma((d + 1)/2)}{\pi^{(d+1)/2}}(1 + \rho^2\|y\|^2)^{-(d+1)/2}$$

Later we will see that the exponential kernel defines a special stochastic process.

Example 2.6 (Gaussian)**.** The Gaussian function with correlation length $\rho > 0$ is given by $\varphi(r) = e^{-r^2\rho^{-2}/2}$. It is the limit of the corresponding Matérn function for $\nu \to \infty$.

An extension of theorem 2.3 is the so called Wiener-Khintchine theorem which uses theorem 2.3 to define a stationary random field with a covariance function via the Fourier transform of a finite non-negative measure; we refer to theorem 6.6.

We are interested in not only identifying positive, but also strictly positive definite functions because these yield a well-posed interpolation problem. One way to do that, restricted to so-called radial functions, is by Schoenberg's theorem.

Definition 2.7. A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is *radial* if there exists a univariate function $\varphi : [0, \infty) \to \mathbb{R}$ such that $\Phi(x) = \varphi(\|x\|)$ for some norm $\| \cdot \|$ on $\mathbb{R}^d$.

We call (kernel) functions which we can write as $K(x, y) = \varphi(\|x-y\|)$, *radial basis function.* Usually, but not exclusively, one uses the Euclidean norm $\| \cdot \|_{\ell^2}$ with respect to radial basis functions.

Example 2.8. The Matérn, exponential and Gaussian function are examples for a radial basis function via $K(x, y) = \varphi(\|x - y\|_{\ell^2})$.

Note that whereas kernels defines by $K(x, y) = \Phi(x - y)$ are translation invariant, radial basis functions are invariant with respect to translations and all rigid body motions, for example rotations. This isotropy is the reason why in statistics literature radial basis functions are referred to as *isotropic functions*.

The following is stated as in [Fas07, Theorem 3.8].

Theorem 2.9 (Schoenberg)**.** *A continuous function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is strictly positive definite and radial for all dimensions $d \in \mathbb{N}$ if the corresponding φ is of the form*

$$\varphi(r) = \int_0^\infty e^{-r^2 t^2} \, \mathrm{d}\mu(t)$$

for some finite non-negative Borel measure μ on $[0, \infty)$ which is not concentrated on the origin.

Using that theorem we can immediately see that the Gaussian kernel, $k(r) = e^{-r^2/\gamma^2}$, $\gamma > 0$, is strictly positive definite: as the point evaluation measure $\mu = \delta_{1/\gamma}$ is a finite non-negative Borel measure on $[0, \infty)$ and not concentrated at the origin. It satisfies the assumptions of Schoenberg's theorem. Therefore, by

$$\varphi(r) = \int_0^\infty e^{-r^2 t^2} \, \mathrm{d}\delta_{1/\gamma}(t) = e^{-r^2/\gamma^2} \tag{5}$$

we are dealing with a strictly positive radial basis function.

3. Reproducing Kernel Hilbert Spaces

We turn to reproducing kernel Hilbert spaces. Every vector space we consider will be a real vector space. For a pre-Hilbert space H we denote the corresponding inner product by $(\cdot, \cdot)_H$.

Definition 3.1. Let $\Omega \subseteq \mathbb{R}^d$ and $H = \{f : \Omega \to \mathbb{R}\}$ be a Hilbert space of functions over Ω. A *reproducing kernel* for H is a function $K : \Omega \times \Omega \to \mathbb{R}$ such that the following holds:

1. $K(\cdot, x) \in H$ for all $x \in \Omega$
2. $f(x) = (f, K(\cdot, x))_H$ for all $x \in \Omega$ and $f \in H$

The tuple (H, K) of Hilbert space and reproducing kernel is called *reproducing kernel Hilbert space.*

Such a reproducing kernel is always unique. To see that, assume that the reproducing property holds for two such kernels, $f(x) = (f, K(\cdot, x))_H = (f, \tilde{K}(\cdot, x))_H$. Then $(f, K(\cdot, x) - \tilde{K}(\cdot, x))_H = 0$ and choosing $f = K(\cdot, x) - \tilde{K}(\cdot, x)$ yields $\|K(\cdot, x) - \tilde{K}(\cdot, x)\|_H = 0$ for all $x \in \Omega$.

Under which circumstances can we find such a reproducing kernel?

Proposition 3.2. *Let H be a Hilbert space. There exists a unique reproducing kernel K for H if and only if the point evaluation δ_x given by $\delta_x(f) = f(x)$ is continuous for all $x \in \Omega$.*

Proof. Clearly δ_x is linear. Assume it is continuous, thus $\delta \in H'$. For all $x \in \Omega$ we can derive $f(x) = \delta_x(f) = (f, K_x)_H$ for $K_x \in H$ by Riesz' representation theorem. Setting $K(y, x) := K_x(y)$ we obtain a reproducing kernel.

Conversely, assume there exists such a reproducing kernel. Note that we have the estimate $|\delta_x(f)| = |(f, K_x)_H| \leq \|f\|_H \|K_x\|_H \leq M\|f\|_H$ for some $M > 0$, hence δ_x is continuous. □

This gives the intuition behind RKHS as the spaces in which we are able to consistently work with point evaluations. As a counterexample, $L^2(\Omega)$ is not such a space since the point evaluations are not continuous in $L^2(\Omega)$. In fact, a point evaluation functional is not even well-defined as an operator on $L^2(\Omega)$, since $L^2(\Omega)$ comprises of equivalence classes of functions.

There are some useful properties of reproducing kernels.

Proposition 3.3. *Let H be a reproducing kernel Hilbert space on some domain $\Omega \subseteq \mathbb{R}^d$ with reproducing kernel K. Then K satisfies the following properties*

1. $K(x,y) = (K(x,\cdot), K(\cdot,y))_H$
2. $K(x,y) = K(y,x)$
3. *Convergence in* $\|\cdot\|_H$ *implies pointwise convergence for all* $x \in \Omega$

Proof. The first one is a direct result of the reproducing property of K. The second one employs the first property and symmetry of the inner product. The third one results from

$$|f(x) - f_n(x)| = |(f - f_n, K(\cdot,x))_H| \leq \|f - f_n\|_H \|K\|_N \leq M\|f - f_n\|_H$$

for some $M > 0$. □

It seems natural to use these reproducing kernels for scattered data interpolation. Indeed, reproducing kernels are positive definite and under mild assumptions on the Hilbert space, strictly positive definite. So far, however, we have not specified what a positive definite kernel is.

Definition 3.4. Let $\Omega \subseteq \mathbb{R}^d$. A continuous function $K : \Omega \times \Omega \to \mathbb{R}$ is *positive definite* if for any N pairwise different points $x_1, ..., x_N \in \mathbb{R}^d$ and all $c = (c_1, ..., c_N) \in \mathbb{R}^N$,

$$\sum_{j=1}^{N} \sum_{i=1}^{N} c_i K(x_i, x_j) c_j \geq 0.$$

If this being zero implies $c \equiv 0$, the kernel is *strictly positive definite.*

Proposition 3.5. *Let H be a reproducing kernel Hilbert space with reproducing kernel K. Then K is positive definite. Moreover, K is strictly positive definite if the point evaluation functionals are linearly independent in H'.*

Proof. See [Fas07, Theorem 13.2] for details. □

On a side note, in the context of geometric group theory, in particular Kazhdan's Property (T), there exists a very similar concept called a *kernel of conditionally negative type* which basically uses the same definition. It is used to prove a theorem after Shalom which relates the property (T) to vanishing cohomology groups which then eventually leads to proving Gromov's polynomial growth theorem; see [BDV09].[1]

4. Native Spaces

Reproducing kernel Hilbert spaces are mainly a theoretical tool to analyse kernel-based approximations such as scattered data interpolation. In many settings, one starts by choosing a kernel function. For e.g. an error analysis it is important to know the RKHS corresponding to the chosen kernel. This leads to the notion of native spaces.

We say that a kernel K is symmetric if $K(x,y) = K(y,x)$ holds for all $x, y \in \Omega$. For example, stationary or reproducing kernels are symmetric.

Let K be a symmetric and strictly positive definite kernel. Define the space

$$H_K := \operatorname{span}\{K(\cdot,x),\ x \in \Omega\}. \tag{6}$$

If K was the reproducing kernel to some Hilbert space H such that H_K is included in H, we would notice that for $f, g \in H_K$,

$$(f,g)_H = \left(\sum_{i=1}^{N} c_i K(\cdot, x_i), \sum_{j=1}^{M} d_j K(\cdot, y_j) \right)_H = \sum_{i=1}^{N} \sum_{j=1}^{M} c_i d_j K(x_i, y_j) \tag{7}$$

[1] Thanks to Libby for providing that insight

holds. This motivates the definition of a bilinear form $(\cdot,\cdot)_{H_K}$ on H_K via

$$(f, g)_{H_K} = \sum_{i=1}^{N} \sum_{j=1}^{M} c_i d_j K(x_i, y_j). \tag{8}$$

This defines an inner product on H_K: Bilinearity is clear by construction, symmetry follows from symmetry of the kernel and positive definiteness is given since K is strictly positive definite. Therefore H_K is a pre-Hilbert space. It is easy to see that K is the reproducing kernel for H_K.

As an inner product, $(\cdot,\cdot)_{H_K}$ induces a norm $\|\cdot\|_{H_K}$ via $\|f\|_K = \sqrt{(f,f)_{H_K}}$. We call the closure of H_K with respect to that norm *native space* and denote it by $\mathcal{N}_K = \overline{H_K}$. Thus, $(\mathcal{N}_K, K)$ is a reproducing kernel Hilbert space.

If one works with so-called *conditionally positive definite kernels* instead of strictly positive definite kernels, the formal construction of the native space is a bit more subtle; see [Wen05] for more.

Native spaces contain all functions for which the kernel-based approximation “should perform well”. We would like to be able to identify these spaces for given kernel function K. We will see that this amounts to investigating the decay of the Fourier transform of a function which itself is related to the regularity of the function itself.

We will see two constructions of native spaces. The first one considers a native space associated with stationary kernels on $\mathbb{R}^d$. The second one works uses an eigendecomposition of Mercer kernels on compact domains to construct a native space.

Theorem 4.1. *Let $\Phi \in C\left(\mathbb{R}^d\right) \cap L^1\left(\mathbb{R}^d\right)$ be strictly positive definite. Then, the native space of $K(x,y) = \Phi(x-y)$ is given by*

$$\mathcal{N}_K := \left\{ f \in L^2\left(\mathbb{R}^d\right) \cap C\left(\mathbb{R}^d\right) : \ \frac{\widehat{f}}{\sqrt{\widehat{\Phi}}} \in L^2\left(\mathbb{R}^d\right) \right\}$$

where we equip this Hilbert space with the inner product

$$(f,g)_{\mathcal{N}_\Phi} = \frac{1}{\sqrt{(2\pi)^d}} \int_{\mathbb{R}^d} \frac{\widehat{f}(x)\widehat{g}(x)}{\widehat{\Phi}(x)} \, \mathrm{d}x.$$

If a function is square integrable we expect that it exhibits some kind of decay at infinity. In light of theorem 4.1 it is thus interesting to investigate the decay of the Fourier transform.

There are ways to formalise the correspondence between decay of Fourier transform and regularity; we refer to [SS07] for a detailed explanation of basic Fourier analysis in $\mathbb{R}^d$. For example, one can use the Riemann-Lebesgue lemma which says that if a function f is integrable, its Fourier transform vanishes at infinity. Thus, for a function $f \in W^{k,1}(\mathbb{R}^d)$ one can apply that to the weak derivatives as well, resulting in $\lim_{|x|\to\infty} \widehat{f^{(k)}}(x) = 0$, $f^{(k)}$ denotes the k-th weak derivative of f. The last step would be to prove a formula like $\widehat{f^{(k)}}(x) = x^k \widehat{f}(x)$. This, however, cannot be proven for general Sobolev functions on $\mathbb{R}^d$, $d > 1$, without restricting ourselves to certain cases such as the case where $f^{(k)}$ admits a certain decay itself. Therefore we keep the analysis at more of a hand waving level, trying to summarise the intuition behind the correspondence between decay of the Fourier transform and smoothness of a function.

The perhaps most natural way to understand this correspondence is to look at the definition of Sobolev spaces via the Fourier transform. For $2k > d$, one then denotes

$$H^k\left(\mathbb{R}^d\right) = \left\{ f \in L^2\left(\mathbb{R}^d\right) \cap C\left(\mathbb{R}^d\right) : \widehat{f}(\cdot)(1 + \|\cdot\|^2_{L^2(\Omega)})^{m/2} \in L^2\left(\mathbb{R}^d\right) \right\}. \tag{9}$$

This definition of the Sobolev space has one big advantage over the classical way involving weak derivatives: It does not require m to be an integer.

Sobolev spaces are the native spaces to many kernel functions. Observe that in order for the function $f \in L^2(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ to be in $H^k(\mathbb{R}^d)$, we require that $g(x) := \widehat{f}(x)(1 + \|x\|^2_{L^2(\Omega)})^{m/2}$ is square integrable. Think of it as requiring $\widehat{f}$ to decay as $o(\|x\|^m_{L^2(\Omega)})$ for $|x| \to \infty$, where we use the standard Landau notation for asymptotic behaviour.

Example 4.2. For the Matérn radial basis function given by $K_{\nu,\rho}(x,y) = \varphi_{\nu,\rho}(\|x-y\|_{\ell^2})$ with $\varphi_{\nu,\rho}$ as in equation (4), $\rho > 0$, and $\nu = \beta - d/2$, we have the Fourier transform

$$\widehat{\varphi}(\omega) = (1 + \|\omega\|^2)^{-\beta}, \quad 2\beta > d.$$

Together with theorem 4.1, this yields $\mathcal{N}_K = H^\beta\left(\mathbb{R}^d\right)$, $2\beta > d$. This is why sometimes the Matérn functions are referred to as *Sobolev splines* [Fas07].

As the Gaussian kernel is the limit of the Matérn kernel for $\beta \to \infty$, this also indicates that the native space for the Gaussian contains certain classes of smooth functions. In fact, one can show that it includes the space $H_a = \{f \in C(\mathbb{R}^d) \mid \operatorname{supp} \widehat{f} \subseteq [-a,a]\},\ a > 0$, which is called *space of band-limited functions* [Fas07, p. 110].

Not only the Matérn kernel has a Sobolev space as its native space. The following theorem is stated as in [SSS13, Theorem 2.1] and concludes our analysis of the interplay between regularity of a function, the decay of its Fourier transform and the native space corresponding to a kernel.

Theorem 4.3. *Let $\Phi \in L^1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ be positive definite. Assume there exist constants $c_1, c_2 > 0$, $c_1 \leq c_2$, such that its Fourier transform $\widehat{\Phi}$ satisfies*

$$c_1(1 + \|\xi\|^2)^{-\tau} \leq \widehat{\Phi}(\xi) \leq c_2(1 + \|\xi\|^2)^{-\tau}$$

for some $\tau > \frac{d}{2}$. Then the native space corresponding to Φ is the Sobolev space $H^\tau(\mathbb{R}^d)$ and the native space norm and the Sobolev norm are equivalent.

The next way of characterizing the native spaces is uses spectral analysis of Hilbert-Schmidt operators which are based on so-called Mercer kernels.

Let $\Omega \subseteq \mathbb{R}^d$ be a compact domain. Let $K : \Omega \times \Omega \to \mathbb{R}$ be continuous, strictly positive definite and symmetric. We define the operator

$$T_K : L^2(\Omega) \longrightarrow L^2(\Omega), \quad v \longmapsto T_K v, \quad T_K v(x) = \int_\Omega K(x,y)v(y)\ \mathrm{d}y. \tag{10}$$

As a Hilbert-Schmidt operator based on a sufficiently smooth kernel, T_K is compact; see [Alt02, 8.15] for details. By

$$(f, T_K g)_{L^2} = \int_\Omega f(x)(T_K g)(x)\ \mathrm{d}x = \int_\Omega f(x) \int_\Omega K(x,y)g(y)\ \mathrm{d}y\ \mathrm{d}x \tag{11}$$

and the theorem of Fubini-Tonelli,

$$\int_\Omega f(x) \int_\Omega K(x,y)v(y)\ \mathrm{d}y\ \mathrm{d}x = \int_\Omega g(y) \int_\Omega K(x,y)f(x)\ \mathrm{d}x\ \mathrm{d}y = (T_K f, g)_{L^2} \tag{12}$$

and therefore T_K is self-adjoint. By the spectral theorem for compact self-adjoint operators (e.g. [Alt02, 10.12]), we can find an eigendecomposition $(\lambda_k, \phi_k)_{k\in\mathbb{N}}$ such that

1. $\{\phi_k\}_{k\in\mathbb{N}}$ is an orthonormal system for $L^2(\Omega)$
2. For all $v \in L^2(\Omega)$, $T_K v(x) = \sum_{i\in I} \lambda_i (v, \phi_k)_{L^2(\Omega)} \phi_k(x)$ for some $I \subseteq \mathbb{N}$
3. $\{\lambda_k\}_{k\in\mathbb{N}}$ are pairwise disjoint eigenvalues with 0 being the only accumulation point.

This eigendecomposition is the key observation of the following characterisation.

Theorem 4.4 (Mercer)**.** *Let K, T_K, ϕ_k and λ_k be as before. Then every eigenfunction ϕ_k is continuous, $k \in \mathbb{N}$, and we have the representation of the kernel K as*

$$K(x,y) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x)\phi_k(y)$$

where the sum is absolutely and uniformly convergent.

We call a kernel satisfying the assumptions of theorem 4.4 a *Mercer kernel.* In its general form, Mercer's theorem requires K to be continuous and symmetric as well as T_K to be positive definite, that is

$$\int_{\Omega\times\Omega} K(x,y)v(x)v(y)\ \mathrm{d}x\ \mathrm{d}y \geq 0 \tag{13}$$

for all $v \in L^2(\Omega)$. In our case, with K being strictly positive definite, this is holds automatically.

Define the space $\mathcal{S} = \{f(x) = \sum_{i\in\mathbb{N}} f_i\phi_i(x)\}$. For $f, g \in \mathcal{S}$ we denote the bilinear form

$$(f,g)_{\mathcal{S}} := \sum_{k\in\mathbb{N}} \frac{f_k g_k}{\lambda_k}. \tag{14}$$

It is clearly an inner product on $\mathcal{S}$. One can show that $(\mathcal{S}, (\cdot,\cdot)_{\mathcal{S}})$ is a Hilbert space [Wen05]. Later it will become intuitive why this is the natural choice of an inner product on $\mathcal{S}$.

It is not only a Hilbert space, but it also has reproducing kernel K. To see that, observe that by the expansion for K given by Mercer's theorem, $K(\cdot,x) \in \mathcal{S}$ for all $x \in \Omega$. Furthermore,

$$(f, K(\cdot,x))_{\mathcal{S}} = \sum_{k\in\mathbb{N}} \frac{f_k\lambda_k\phi_k(x)}{\lambda_k} = \sum_{k\in\mathbb{N}} f_k\phi_k(x) = f(x) \tag{15}$$

for any $f \in \mathcal{S}$ and $x \in \Omega$ and therefore the reproducing property is satisfied as well. So far we only know that $\mathcal{S}$ contains the native space corresponding to K, $\mathcal{N}_K \subseteq \mathcal{S}$. We will see that $\mathcal{N}_k \neq \mathcal{S}$ holds.

What is the native space then? By the reproducing kernel property, it seems natural to stick to the eigenfunction expansions. According to the Picard criterion [Wen05],

$$T_K(L^2(\Omega)) = \left\{ f \in L^2(\Omega) : \ \sum_{i=1}^{\infty} \frac{|(f,\phi_k)_{L^2(\Omega)}|}{\lambda_k^2} < \infty \right\} \tag{16}$$

which relates the image of T_K to the set of functions with some rapid decrease of the generalized Fourier coefficients $(f,\phi_k)_{L^2(\Omega)}$. This again relates to f having a certain regularity; see below.

In a very informal manner, one can say that this space given by (16) is a lot smoother than what the native space of K actually contains.

In order to skip many tedious computations, we will now jump right to the correct conclusion, namely that $\mathcal{N}_K$ is the image of $T^{\frac{1}{2}}$, where we define this operator via the eigendecomposition of T. The following theorem as well as a formal proof is contained in [Wen05].

Theorem 4.5. *Let K be continuous, strictly positive definite and symmetric. Then the native space of K is given by*

$$\mathcal{N}_K = \left\{ f \in L^2(\Omega) : \ \sum_{k\in\mathbb{N}} |(f,\phi_k)_{L^2(\Omega)}|^2/\lambda_k < \infty \right\}$$

with scalar product $(f,g)_{\mathcal{N}_K} = \sum_{k\in\mathbb{N}} \frac{1}{\lambda_k}(f,\phi_k)_{L^2(\Omega)}(g,\phi_k)_{L^2(\Omega)}$.

Knowing the native space for K, we see that the initial choice of bilinear form $(\cdot,\cdot)_{\mathcal{S}}$ as in (14) is consistent with $(\cdot,\cdot)_{\mathcal{N}_K}$.

The coefficients $(f,\phi_k)_{L^2(\Omega)}$ can be thought of as generalised Fourier coefficients. To see that, note that for 2π-periodic functions in $\mathbb{R}$, the Fourier series expansion of a function is defined through the coefficients $c_n = \widehat{f}(n)$, $n \in \mathbb{Z}$. By $\widehat{f}(n) = \int_{-\pi}^{\pi} f(z)e^{-inz}\ \mathrm{d}z$, this resembles the scalar product $(f,\phi_k)_{L^2(\Omega)}$ for a Fourier basis as a choice of orthonormal basis. The same decay statements we made in the context of theorem 4.1 can be adapted to Fourier series representations [TVV03, p. 93], however, the formalisation is bound to a certain amount obstacles if one does not work with periodic functions.

5. Error Estimates in Native Spaces

Now we will answer the question of how well we can approximate a function with reproducing kernels. To this end, let $\Omega \subseteq \mathbb{R}^d$ be a compact domain and $X = \{x_1, ..., x_N\} \subseteq \Omega$ be a set of points in our domain. Let K be a continuous, symmetric, strictly positive kernel. Hence it is a reproducing kernel with some native space $\mathcal{N}_K$.

Assume $f \in \mathcal{N}_K$. We want to investigate approximation properties of

$$f(x) \approx \mathcal{P}_f(x) = \sum_{i=1}^{N} f(x_i)K(x,x_i). \tag{17}$$

In particular, we want to find a bound for $|f(x) - \mathcal{P}_f(x)|$ which then leads to an $L^\infty(\Omega)$ estimate. The derivation of bounds with respect to other L^p spaces is also possible [Fas07, Wen05].

If one works on a mesh, the size of the grid is controlled by the mesh-width. Its equivalent in scattered data approximation is called *fill distance* and defined as

$$h := h(X,\Omega) = \sup_{x\in\Omega} \min_{x_j \in X} \|x - x_j\|_{L^2(\Omega)}. \tag{18}$$

It is the diameter of the largest empty ball one can fit into the domain. For $(2^n+1)^d$ equally spaced points the fill distance is $h = 2^{-n}$, the one for a low discrepancy sequence such as Halton or Sobol points is approximately the same, $h \approx 2^{-n}$.

We would like to write our interpolation approximation in a way which allows an easier analysis. We use cardinal functions. A set of cardinal functions $\{u_1^*(x), ..., u_N^*(x)\}$ associated with a pointset X is defined by $u_j^*(x_i) = \delta_{ij}$. These are uniquely defined, as the linear problem $Au^*(x) = b(x)$ with $u^*(x) = (u_1^*(x), ..., u_N^*(x))^T$, $b(x) = (K(x,x_1), ..., K(x,x_N))^T$ and $A = (K(x_i,x_j))_{i,j=1}^N$ has a unique solution for a strictly positive definite kernel K. Here we call the matrix A in order to avoid confusion with the kernel function K. This also tells us that we can find the cardinal functions in $\operatorname{span}\{K(\cdot,x_j),\ j = 1,...,N\}$. The downdraw of this approach is that for every evaluation of these cardinal basis functions we have to solve a system. Using the cardinal functions, we rewrite

$$\mathcal{P}_f(x) = \sum_{i=1}^{N} c_i K(x,x_i) = \sum_{i=1}^{N} f_i u_i^*(x). \tag{19}$$

Before we can derive estimates we need to introduce some more tools.

Let $u : \mathbb{R}^d \to \mathbb{R}^N$ be any function. We denote the quadratic form $Q : \mathbb{R}^N \to \mathbb{R}$ associated with the kernel K by

$$Q(u(x)) = K(x,x) - 2\sum_{i=1}^{N} u_i(x)K(x,x_i) + \sum_{i,j=1}^{N} u_i(x)u_j(x)K(x_i,x_j). \tag{20}$$

Definition 5.1. Let $\Omega \subset \mathbb{R}^d$ and $K : \Omega \times \Omega \to \mathbb{R}$ be a continuous, strictly positive definite symmetric kernel. Let X be a pointset of pairwise distinct points in Ω. We define $P_{K,X}(x) = \sqrt{Q(u^*(x))}$ as the *power function* associated with X and K. By $u^*(x)$ we denote the vector consisting of the evaluations of the cardinal functions at $x \in \Omega$.

According to [Fas07], the power function has its name due to the resemblance to the power function in statistical learning theory as observed by R. Schaback; see texts by R. Schaback for more on the power function, e.g. [Sch93].

There are many possibilities to write the power function in a different way. Using the reproducing property of the kernel, one can for example show

$$P_{X,K}(x) = \left\| K(\cdot,x) - \sum_{j=1}^{N} u_j^*(x) K(\cdot,x_j) \right\|_{\mathcal{N}_K}. \tag{21}$$

Another way is to write

$$P_{X,K}(x) = \sqrt{K(x,x) - (b(x))^T A^{-1} b(x)}, \tag{22}$$

using the linear system $Au(x) = b(x)$ arising in the derivation of the cardinal basis functions. By positive definiteness of A, this gives rise to the error bound $0 \leq P_{X,K}(x) \leq \sqrt{K(x,x)}$.

Let us turn to deriving error estimates for the interpolation problem. Using the power function and the reproducing property of the kernel we obtain the generic error estimate

$$|f(x) - \mathcal{P}_f(x)| \leq \|f\|_{\mathcal{N}_K} P_{X,K}(x) \tag{23}$$

which only works for $f \in \mathcal{N}_K$. There are also estimates which do not require this assumption but in our basic analysis of the approximation we have to assume that.

The main benefit of this estimate is that we can decouple the analysis of the employed pointset and kernel through the power function from the (regularity of the) function we want to interpolate.

To turn this estimate into an error estimate involving the fill distance, we have to bound the power function. The way to do this is to use existence of an approximation scheme with a so-called local polynomial reproduction property and to combine this with a Taylor expansion of the kernel function.

We start by deriving the polynomial reproduction property. It is based on something called an *interior cone condition* which basically implies that whenever a domain satisfies this condition it contains balls of a controllable radius. As it is not of particular interest to us, we refer to [Fas07, Definition 14.2]. For a detailed presentation see [Wen05].

We denote by $\mathbb{P}_k$ the polynomials of degree less than or equal to k. The following is stated as in [Fas07, Theorem 14.4] which is summary of a series of results in [Wen05, Chapter 3].

Theorem 5.2. *Let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain and satisfy an interior cone condition. Let $X = \{x_1, ..., x_N\} \subseteq \mathbb{R}^d$ with fill distance $h_{X,\Omega}$. Then the cardinal functions defined via X and K provide a* local polynomial reproduction property *of degree $\ell \in \mathbb{N}$. That is, there exist positive constants h_0, c_1, c_2 such that if X satisfies $h_{X,\Omega} \leq h_0$, for all $x \in \Omega$ there exist numbers $\tilde{u}_1(x), ..., \tilde{u}_N(x)$ such that*

1. $\sum_{i=1}^{N} \tilde{u}_i(x) p(x_j) = p(x)$ *for all polynomials* $p \in \mathbb{P}_\ell\left(\mathbb{R}^d\right)$
2. $\sum_{i=1}^{N} |\tilde{u}_j(x)| \leq c_1$
3. $\tilde{u}_j(x) = 0$ *if* $\|x - x_j\|_{L^2(\Omega)} \leq c_2 h_{X,\Omega}$

The first statement is the polynomial reproduction property, the second one is related to the Lebesgue constant *of the interpolation operator and the third one shows locality.*

Using that local polynomial reproduction property, we can derive an error estimate for the interpolation. We use the usual multi-index notation for derivatives and Taylor approximations. By D_2^α we denote the α-th (weak) derivative with respect to the second component, $\alpha \in \mathbb{N}_0^d$.

Theorem 5.3. *Let $\Omega \in \mathbb{R}^d$ be bounded and satisfy an interior cone condition. Assume the kernel $K \in C^{2k}(\Omega \times \Omega)$ is symmetric and strictly positive definite. Assume $f \in \mathcal{N}_K$ with corresponding approximation $\mathcal{P}_f$. Then there exist positive constants h_0, C_1, C_2 such that if $h_{X,\Omega} \leq h_0$ we have*

$$|f(x) - \mathcal{P}_f(x)| \leq C_1 h_{X,\Omega}^k \sqrt{\mathbb{M}_K(x)} \|f\|_{\mathcal{N}_K}.$$

Here, we define

$$\mathbb{M}_K(x) := \max_{|\beta|=2k} \max_{w,z \in Z_x} |D_2^\beta K(w,z)|, \quad Z_x = \Omega \cap B(x, C_2 h_{X,\Omega}),$$

where $B(a,b)$ denotes the ball around a with radius b.

Proof. We only mention the important ideas behind the proof.

Using the previous generic error estimate we only have to bound $P_{X,K}$. In particular, we want to derive $P_{X,K}(x) \leq Ch^k_{X,\Omega}\sqrt{\mathbb{M}_K(x)}$. We use that the quadratic form Q is minimized at u^*. Combining that with a Taylor approximation of $K(x,\cdot)$ centered at x and the local polynomial reproduction we only need one more Taylor approximation and a lot of inequalities to obtain that bound; see [Fas07, Theorem 14.5] for details. □

In a nutshell, this theorem says that a kernel of smoothness C^{2k} gives approximation quality h^k, at least asymptotically. Therefore, for smooth kernel functions we have arbitrary approximation quality. Note however, that for smooth kernels the native spaces might become rather small which causes a problem because we require $f \in \mathcal{N}_K$ in order to use that estimate. A possible generalisation is to derive such estimates in general Sobolev spaces; see [Fas07, Chapter 15].

Of course the estimate depends strongly on $\mathbb{M}_K(x)$. For more on bounding that quantity see e.g. [Wen01].

6. Random Fields and Gaussian Processes

In the following part, we will turn our attention to Gaussian processes, in particular Gaussian process regression. After deriving some basics about Gaussian processes we will see that we can use a lot of the RKHS machinery to analyse properties of Gaussian process regression.

We start by stating the absolute basics about Gaussian processes. This goes hand in hand with some theory about random fields. Gaussian processes are specific stochastic processes which can be understood as generalisations of Gaussian random vectors.

We restrict ourselves to real-valued random fields on $\mathbb{R}^d$, for practical reasons. Further, we will skip a lot of real analysis and probability theory corresponding to (infinite) product measures; we refer to [Dud02] for that kind of background.

Definition 6.1. Let $D \subseteq \mathbb{R}^d$. A collection of real-valued random variables $(R(x))_{x\in D}$ is called a *random field.* If $d = 1$, this collection is called *stochastic process.*

We can associate the random field with the map $R : D \times \Omega \to \mathbb{R}$ given by $R(x,\omega) = R(x)(\omega)$. For any $\omega \in \Omega$, the function $R(\cdot,\omega) : D \to \mathbb{R}$ is called *realization* of R.

Definition 6.2. Let $(R(x))_{x\in D}$ be a random field. It is called a *second order random field* if for all $x \in D$, $R(x) \in L^2(\Omega)$. For a second order random field we define the *mean function* $m : D \to \mathbb{R}$, $m(x) = \mathbb{E}[R(x)]$ and the *covariance function* $C : \Omega \times \Omega \to \mathbb{R}$ which is defined through $C(x,y) = \mathbb{E}[(R(x) - m(x))(R(y) - m(y))]$.

Gaussian random fields are special cases of second order random fields and correspondingly, Gaussian processes are special cases of second order stochastic processes.

Recall that a convenient way to define multivariate Gaussian distributions is via the Fourier transform. We say that an $\mathbb{R}^d$-valued random vector X follows the multivariate Gaussian distribution, $X \sim \mathcal{N}(\mu,\Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ symmetric positive semidefinite, if its Fourier transform, also known as the *characteristic functional*, is given by $\varphi_X(y) = \exp\left(i\langle \mu, y\rangle - \frac{1}{2}\langle \Sigma y, y\rangle\right)$, $y \in \mathbb{R}^d$.

Definition 6.3. A Gaussian random field is a second order random field $(G(x))_{x\in D}$ for which any finite dimensional distribution is multivariate Gaussian. Formally, for $N \in \mathbb{N}$ and $x_1, ..., x_N \in D$, we have $(G(x_1), ..., G(x_N)) \sim \mathcal{N}(\mu,\Sigma)$ with $\mu \in \mathbb{R}^N$ and $\Sigma \in \mathbb{R}^{N\times N}$ symmetric positive semidefinite.

Corresponding to definition 6.1, a Gaussian process is a special case of definition 6.3 for $d = 1$. That is, a Gaussian process is a stochastic process such that any finite number of evaluations follows a multivariate Gaussian distribution.

It will be vital to understand that a Gaussian random field, respectively a Gaussian process, is uniquely defined by its mean and covariance function. This relies on a strong result in probability theory, due to Kolmogorov. It is sometimes called Kolmogorov's extension theorem. In a nutshell, it says that if we have a family of probability measures which is consistent with respect to taking subfamilies, we can extend it to a uniquely defined consistent measure. We refer to [Dud02] for a concise treatment of this theorem.

Proposition 6.4. *Let $m : D \to \mathbb{R}$ be a function and $C : D \times D \to \mathbb{R}$ be symmetric positive semidefinite. Then there exists a Gaussian random field with mean function m and covariance function C for which the law is uniquely determined on $(\mathbb{R}^D, \mathscr{B}(\mathbb{R})^D)$.*

Proof. We will only sketch the procedure of that proof: Let $x_1, ..., x_N \in D$, denote the corresponding covariance matrix C_J and mean vector m_J. Using the projection matrix $\pi_{J \to K}$ this setting satisfies the consistency condition of theorem Kolmogorov's extension theorem and using that theorem we get the corollary. □

Previously, in section 2, we announced that we can use theorem 2.3 to characterise certain stochastic processes. This is the Wiener-Khintchine theorem, which says that the covariance function of a stationary random field is the Fourier transform of a finite non-negative Borel measure and we can recover that measure by taking the inverse Fourier transform of the covariance function.

Definition 6.5. A random field $(b(x))_{x \in D}$, $D \subseteq \mathbb{R}^d$ is called *stationary* if for all $x, y, h \in D$ we have $\mathbb{E}[b(x+h)] = \mathbb{E}[b(x)]$ and $\mathbb{E}[b(x+h)b(y+h)] = \mathbb{E}[b(x)b(y)]$. It is called *strictly stationary* if $(b(x))_{x \in D}$ and $(b(x+h))_{x \in D}$ have the same finite-dimensional distributions for any $x, h \in D$.

Theorem 6.6 (Wiener-Khintchine)**.** *The following are equivalent:*

1. *There exists a real-valued stationary random field with bounded, continuous, stationary covariance function k.*
2. *$k : \mathbb{R}^d \to \mathbb{R}$ is the Fourier transform of a finite non-negative Borel measure ν on $\mathbb{R}^d$.*

Proof. This is a direct result of proposition 6.4 and theorem 2.3. □

Example 6.7 (Matérn Covariance)**.** In example 2.4, we constructed the Matérn function out of its spectral density. Using theorem 6.6, we take the inverse Fourier transform of the Matérn function and recover its spectral distribution. The Matérn covariance defines a real-valued stationary random field. It can be shown that this random field has sample paths which are $\lceil \nu \rceil - 1$ times differentiable [SWN13].

Example 6.8 (Ornstein-Uhlenbeck Process)**.** The Gaussian process defined by the exponential covariance $k(x) = e^{-|x|/\lambda}$ in dimension $d = 1$ is the *Ornstein-Uhlenbeck process.* Among others, it is used in financial modelling of interest rates.

This completes our quick survey of Gaussian random fields and Gaussian processes. From now on, according to the terminology employed in Gaussian process literature such as [RW06], we will exclusively use the term Gaussian process. This includes the case in which the collection of random variables might as well be a random field instead of a process.

We write $X \sim \mathrm{GP}(m, C)$ for a Gaussian process with mean function $m = m(\cdot)$ and covariance function $C = C(\cdot, \cdot)$.

7. Gaussian Process Interpolation and Regression

In this section we mostly rely on [RW06, Chapter 2] and [ST18], using a notation similar to the latter one.

Other than traditional regression approaches, Gaussian process regression assumes an underlying stochastic process from which one derives quantities to estimate that function.

Let $\Omega \subseteq \mathbb{R}^d$. Assume we are given points $x_1, ..., x_N \in \Omega$ and measurements $y_1, ..., y_N \in \mathbb{R}$ which correspond to evaluations of a function $f : \Omega \to \mathbb{R}$, $y_i \approx f(x_i)$, $i = 1, ..., N$. We want to use these measurements to reconstruct f. This already sounds very similar to the problem setting in scattered data interpolation, and in fact we will see a lot of similarities between these two concepts.

We are facing two different possibilities now. Firstly, we can assume that our measurements are exact in the sense that $y_i = f(x_i)$ for $i = 1, ..., N$. This is called *Gaussian process interpolation*, sometimes known as *Gaussian process emulation.*

Secondly, we can admit that our measurements are slightly wrong which leads to a regression approach as the more suitable choice. This is called *Gaussian process regression*, also known as *kriging* in geophysics[2]. We will see that the generalisation of interpolation to regression is very easy and straightforward.

7.1. Interpolation

We start by considering Gaussian process interpolation. We take a Bayesian point of view. Let f_0 be a Gaussian process with mean function $m : \Omega \to \mathbb{R}$ and a symmetric positive definite function $K : \Omega \times \Omega \to \mathbb{R}$ as a covariance function, $f_0 \sim \mathrm{GP}(m(\cdot), K(\cdot, \cdot))$. In statistical terms, this is called putting a Gaussian *prior distribution* on f as we are modelling it with such a Gaussian process. Note that f_0 is a stochastic process and not a function.

Denote the point set $X := (x_1, ..., x_N) \subseteq \Omega$. As f_0 is a Gaussian process, the vector given by $f_0(X) = (f_0(x_1), ..., f_0(x_N))$ is multivariate Gaussian, $f_0(X) \sim \mathcal{N}(m(X), K(X, X))$, where $m(X) = (m(x_1), ..., m(x_N))^T$ are the evaluations of the mean and $K(X, X) = (K(x_i, x_j))_{i,j \leq N}$ consists of evaluations of the covariance function at the point set X.

We want to be able to reproduce evaluations of f at new points $X_* := (x_{N+1}, ..., x_{N+M}) \subseteq \Omega$. We do that by improving our prior using the measurements y_i, $i = 1, ..., N$. In particular, we condition f_0 to be exact on X, $f_0(x_i) = y_i$, $i = 1, ..., N$.

The joint distribution of $f_0(X)$ and $f_0(X_*)$ is given by

$$\begin{pmatrix} f_0(X) \\ f_0(X_*) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(X) \\ m(X_*) \end{pmatrix}, \begin{pmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{pmatrix}\right) \tag{24}$$

where $K(X_*, X) = (K(x_{N+i}, x_j)_{i,j}$ is the matrix consisting of the evaluations of the covariance functions at the pointsets X_* and X. The other submatrices are defined in the same way. By conditioning onto $f_0(X) = y$ we obtain a new distribution

$$\mathbb{P}_{f_0(X_*)|f_0(X)=y} \equiv \mathcal{N}(\overline{m}(X_*), \overline{K}(X_*, X_*)) \tag{25}$$

where $\overline{m}(X_*) = m(X_*) - K(X_*, X)(K(X, X))^{-1}(m(X) - y)$ is the new mean and the matrix given by $\overline{K}(X_*, X_*) = K(X_*, X_*) - K(X_*, X)(K(X, X))^{-1}K(X, X_*)$ is the new covariance; see e.g. [RW06, Appendix A.2] for details. This construction works for all pointsets $X_* \subseteq \Omega$ and thus yields a new Gaussian process

$$f_N \sim \mathrm{GP}(m_N(\cdot), K_N(\cdot, \cdot)) \tag{26}$$

given by $m_N(Y) = \overline{m}(Y)$ and $K_N(Y, Y) = \overline{K}(Y, Y)$ for any $Y \subseteq \Omega$. It is sometimes called *predictive distribution* and corresponds to the *posterior distribution* in Bayesian statistics. In the context of Bayesian statistics, m_N is called the *conditional mean* and in the case of Gaussian distributions it coincides with the *maximum a posteriori estimator*, the mode of the posterior distribution.

[2] From what is being used in the literature, the terms kriging, Gaussian process interpolation, Gaussian process regression and Gaussian process emulation do not seem to be clearly distinguishable

We observe some interesting properties. First of all, f_N is exact at the measurements as by

$$m_N(X) = m(X) - K(X,X)(K(X,X))^{-1}(m(X) - y) = y \tag{27}$$

as well as by

$$K_N(X,X) = K(X,X) - K(X,X)(K(X,X))^{-1}K(X,X) = 0, \tag{28}$$

f_N hits the measurements without any (co)variance. Also, $K_N(X_*, X_*) \leq K(X_*, X_*)$ for any $X_* \subseteq \Omega$, since K is positive definite. The inequality is to be understood element wise.

The mean function m_N is a linear estimator for f. It can be written as a linear combination of covariance evaluations via

$$m_N(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i), \quad x \in \Omega, \tag{29}$$

for $\alpha = (\alpha_1, ..., \alpha_N)$ given by $\alpha = (K(X,X))^{-1}y$. In fact, it can be shown that it is the *best linear predictor* as it is the linear predictior with the smallest mean-squared error [ST18, p. 5].

Note that this is the exact same technique as the one being used in scattered data interpolation. We try to approximate a function by solving $K(X,X)\alpha = y$ and then construct the approximation as a linear combination of the coefficients and the kernel evaluations.

7.2. Regression

So far we have required y_i to be an exact measurement of $f(x_i)$. This might be too restrictive, so we want to incorporate a possible error into our model. We assume additive noise, with regard to which the measurements $y_1, ..., y_N$ behave like $y_i \approx f(x_i) + \xi_i,\ \xi_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, ..., N$ for some $\sigma > 0$. In this case, the prior f_0 is such that

$$f_0(X) \sim \mathcal{N}(m(X), K(X,X) + \sigma^2 I_N) \tag{30}$$

where I_N is the N-dimensional identity matrix. Incorporating that into the model, the joint probability distribution becomes

$$\begin{pmatrix} f_0(X) \\ f_0(X_*) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m(X) \\ m(X_*) \end{pmatrix}, \begin{pmatrix} K(X,X) + \sigma^2 I_N & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{pmatrix} \right) \tag{31}$$

for any $X_* \subseteq \Omega$. We can progress identically as in the previous part, therefore the predictive distribution becomes

$$f_N \sim \mathrm{GP}(m_N^\sigma(\cdot), K_N^\sigma(\cdot, \cdot)) \tag{32}$$

where $m_N^\sigma(X_*) = m(X_*) - K(X_*, X)(K(X,X) + \sigma^2 I_N)^{-1}(m(X) - y)$ is the new mean and $K_N^\sigma(X_*, X_*) = K(X_*, X_*) - K(X_*, X)(K(X,X) + \sigma^2 I_N)^{-1}K(X, X_*)$ is the new covariance.

We refer to [SSS13] for a more exhaustive treatment of the interplay between Gaussian process regression and scattered data interpolation. The authors discuss the question whether Gaussian process interpolation is a stochastic or a deterministic problem.

On a last note, there are many more connections between scattered data approximation, Bayesian inference and Gaussian processes. Among others, we refer to [RW06] and [ST18] for more on that subject.

8. RKHS and Calculus of Variations

As a conclusion we want show some interesting connections to other areas of mathematics. We we take a slightly different point of view. In particular, we will consider both regression and interpolation in the context of calculus of variations and see that they can be viewed as the solutions to certain regularised minimisation problems.

Due to the easier notation we go back to the noise-free setting and sometimes we will sketch how it works in the noisy setting. Either way, the generalisation is straightforward.

Recall that m_N as our estimator for f is given by

$$m_N(X_*) = m(X_*) - K(X_*, X)(K(X,X))^{-1}(m(X) - y). \tag{33}$$

For simplicity reasons we assume $m \equiv 0$, in which case our estimator is given by

$$m_N(X_*) = K(X_*, X)(K(X,X))^{-1}y. \tag{34}$$

We view this as a scattered data interpolation problem in which case we write

$$f(x) \approx m_N(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i), \tag{35}$$

where $\alpha = (\alpha_1, ..., \alpha_N)^T$ is given by solving $K(X,X)\alpha = y$. The noisy setting can be considered in the same way by solving $(K(X,X) + \sigma^2 I_N)\alpha = y$ instead.

Implicitly we are dealing with a variational problem. Let H be a reproducing kernel Hilbert space on Ω with reproducing kernel K. We will see that m_N is also a solution to the variational problem

$$m_N = \arg\min_{f \in H} \|f\|_H \quad \text{s.t.} \quad f(x_i) = y_i, \ i = 1, ..., N. \tag{36}$$

This is due to the so-called representer theorem. In the following we state a very basic version. For a more general version see for example [SC08, Theorem 5.5].

Theorem 8.1 (Representer Theorem)**.** *Let $\Omega \subseteq \mathbb{R}^d$ and H be a reproducing kernel Hilbert space on Ω with reproducing kernel K. Let $(X, y) \in (\Omega \times \mathbb{R})^N$ be the usual data set consisting of a pointset X and measurements y. Let $L : (\mathbb{R}^2)^N \to [0, \infty]$ be any function. Then the variational problem*

$$\arg\min_{f \in H} \|f\|_H + L((y_1, f(x_1)), ..., (y_N, f(x_N))) \tag{37}$$

has a solution in $\text{span}\{K(\cdot, x_i), \ i = 1, ..., N\}$.

Proof. The prove is rather simple and goes more or less as in the following sketch. Write f as the sum of its projections, $P_H f + P_{H^\perp} f$ with P_V denoting the projection onto a subspace V. Use a Pythagoras type orthogonality of the this projection to split $\|P_H f + P_{H^\perp} f\| = \|P_H f\| + \|P_{H^\perp} f\|$ and eventually it will become evident that the term $P_{H^\perp} f$ is irrelevant for the minimisation and therefore can be set to zero. □

Note that we basically allow every non-negative loss function L. Often the representer theorem is stated with respect to convex loss function which then guarantee the uniqueness of the minimum. We however only stated that there is at least one solution in that span of kernel evaluations.

That theorem can be used to relate Gaussian process interpolation to the problem given by (37). The idea is to write the condition $f(x_i) = y_i$, $i = 1, ..., N$ in terms of Lagrange multipliers which satisfy the conditions on L in the representer theorem. Therefore m_N as the solution to (37) can be written as $m_N(x) = \sum_{i=1}^N c_i K(x, x_i)$. By plugging that representation into the minimisation problem we can derive the first oder optimality condition as $K(X,X)c = f(X)$, $c = (c_1, ..., c_N)^T$.

Similarly, we can work with regression in the same way. To this end denote

$$L((y_1, f(x_1)), ..., (y_N, f(x_N))) = \frac{1}{\lambda} \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{38}$$

for which we can also apply the representer theorem. Again, writing the solution as the linear combination of kernel evaluations and deriving the first order optimality condition we obtain the coefficients, say $d = (d_1, ..., d_N)^T$, via $(K(X,X) + \lambda I_N)d = f(X)$. For $\lambda = \sigma^2$ this is the same as the Gaussian process regression.

References

[Alt02] H. Alt. *Lineare Funktionalanalysis.* Springer-Verlag Berlin Heidelberg New York. 2002.

[BDV09] B. Bekka, P. de la Harpe and A. Valette. *Kazhdan's Property (T).* Cambridge University Press. 2009. `http://www.math.harvard.edu/~ctm/home/text/books/bekka_harpe_valette/KazhdanTotal.pdf`

[Dud02] R. Dudley. *Real Analysis and Probability.* Cambridge University Press. 2002.

[Fas07] G. Fasshauer. *Meshfree Approximation Methods with Matlab.* Interdisciplinary Mathematical Sciences, Vol. 6. 2007.

[RW06] C. Rasmussen, C. Williams. *Gaussian Processes for Machine Learning.* MIT Press. 2006. `https://www.gaussianprocess.org/gpml`.

[SC08] I. Steinwart, A. Christmann *Support Vector Machines.* Springer-Verlag New York. 2008.

[Sch93] R. Schaback. *Comparison of Radial Basis Function Interpolants.* in K. Jetter, L.L. Schumaker and F. Utreras (eds.): "Multivariate Approximations: From CAGD to Wavelets". 1993.

[SS07] E.M. Stein, R. Shakarchi. *Fourier Analysis: An Introduction.* Princeton University Press. 2007.

[SSS13] M. Scheuerer, R. Schaback, M. Schlather. *Interpolation of spatial data - a stochastic or a deterministic problem?.* Cambridge University Press. 2013.

[ST18] A. Stuart, A. Teckentrup. *Posterior Consisteny for Gaussian Process Approximations of Bayesian Posterior Distributions.* Mathematics of Computation (87), 721-753. 2018.

[SWN13] T. J. Santner, B. J. Williams, W. I. Notz *The design and analysis of computer experiments.* Springer Science & Business Media. 2013.

[TVV03] H. ter Morsche, J. van den Berg, E. van de Vrie. *Fourier and Laplace Transforms.* Cambridge University Press. 2003.

[Wen01] H. Wendland. *Gaussian interpolation revisited.* Trends in Approximation Theory, K. Kopotun, T. Lyche and M. Neamtu (eds), Vanderbilt University Press, pp. 517-526. 2001.

[Wen05] H. Wendland. *Scattered Data Approximation.* Cambridge University Press. 2005.