
Lecture notes on Scientific Computing 2

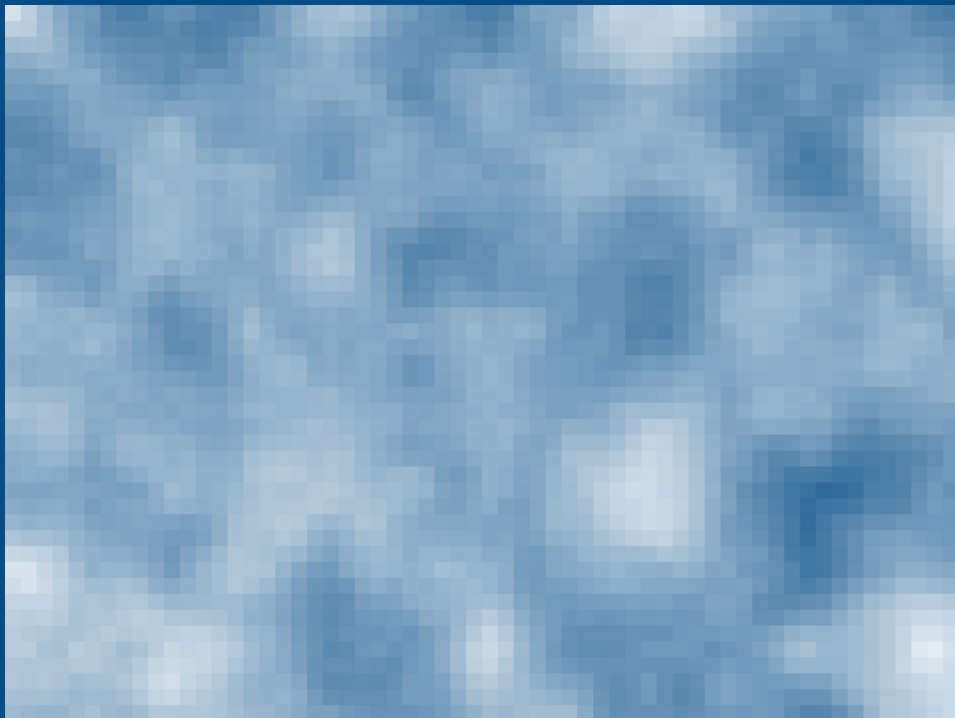
Written by
Manuel Hinz

mh@mssh.dev or s6mlhinz@uni-bonn.de

Lecturer

Prof. Dr. Jochen Garcke

garcke@ins.uni-bonn.de



Contents

Chapter 0 Manuel’s notes and introduction 2

0.1 Overview 2

0.1.1 Function approximation / Interpolation 2

0.1.2 Dimensionality reduction 3

Chapter 1 Kernel based methods 4

1.1 Kernels 4

1.1.1 Examples 4

1.1.2 Kernels in machine learning 5

1.1.3 Mercer kernels 6

1.1.4 Properties of kernels 7

1.2 Reproducing Kernel Hilbert Space (RKHS) 10

Chapter 0:

Manuel's notes and introduction

Warning

These are unofficial lecture notes written by a student. They are messy, will almost surely contain errors, typos and misunderstandings and may not be kept up to date! I do however try my best and use these notes to prepare for my exams. Feel free to email me any corrections to mh@mssh.dev or s6mlhinz@uni-bonn.de.
Happy learning!

General Information

- Ecampus: [Ecampus link](#)
- Basis: [Basis link](#)
- Website: <https://ins.uni-bonn.de/teachings/ss-2024-440-v3e2-wissenschaftlich/>
- Time slot(s): Tuesday 10-12 and Thursday 08-10
- Exams: Oral, unless more than 50 people take the exam
- Deadlines: tbd
- Two topics:
 - Kernel based methods for function approximation
 - Nonlinear dimensionality reduction / manifold learning / latent space embeddings
- Official lecture notes for most of the lectures
- Exercises are a mix of theory (proofs, (counter-)examples) and programming tasks

Start of lecture 01
(09.04.23)

0.1 Overview

We begin with a quick overview of the two parts of the lecture:

0.1.1 Function approximation / Interpolation

Consider $x_i \in \mathbb{R}^d$, $\hat{f}_i \in \mathbb{R}$:

$$\{(x_i, \hat{f}_i)\}_{i=1}^N.$$

Aim: Find a “good” function f such that

$$f(x_i) = \hat{f}_i, \quad i = 1, \dots, N$$

To compute f , we can make use of a discrete representation of f using **Ansatzfunctions** $\{b_j\}_{j=1}^N$:

$$f(x) = \sum_{j=1}^N c_j b_j(x).$$

Here we assume the same number of data and functions b_j .

For interpolation, we can solve this via:

$$BC = \hat{F}$$

Kernel functions that are centered at the locations x_j turn out to be a good choice:

$$b_j(x) = k(x_j, x)$$

which gives

$$f(x) = \sum_{j=1}^N c_j k(x_j, x).$$

We will also consider approximation instead of interpolation

$$f(x_i) \approx \hat{f}_i.$$

This is in particular relevant in machine learning, where one usually assumes, and actually has noise and measurement errors in the given data.

Example: Assess credit risk

Example: Chemistry / energy of molecules. This needs a kernel on graphs

Example: Time series. This needs a kernel on time series

Remark. We will also see that kernels relate to similarity measures and therefore to distances (dissimilarity).

Lagrangian Interpolation does not work great for a lot of points and higher dimensions

Careful not to discriminate, credit risk should be independent of neighbourhood for example!

Topics in part 1:

- What are kernels and their properties
- Reproducing Kernel Hilbert spaces as the function space in which we are working
- Function interpolation and their approximation properties
- Generalized interpolation for solving partial differential equations
- Kernel methods for prediction in machine learning, representer theorem and regularization
- Gaussian Process Regression and Support Vector Machines

0.1.2 Dimensionality reduction

Distances and similarities are a key aspect of the second topic of the course:

Dimensionality reduction for high-dimensional data

The key idea is to find a “good” low dimensional representation (called embedding), such that chosen properties in high dimensions are approximately preserved.

- Linear dimensionality reduction (numerical linear algebra)
- Nonlinear dimensionality reduction (numerical linear algebra with Non-euclidean geometry)
- Dimensionality reduction with neural networks and other nonlinear function representations

Chapter 1:

Kernel based methods

1.1 Kernels

Definition (Gaussian kernel). The **gaussian kernel** is a prime example of a kernel:

$$k(x, y) := \exp(-\alpha \|x - y\|_2^2) = \phi(\|x - y\|_2)$$

for all $x, y \in \mathbb{R}^d$ where α is a scaling parameter.

Definition 1.1. Let Ω be an arbitrary nonempty set. A function $k : \Omega \times \Omega \rightarrow \mathbb{R}$ is called **kernel** on Ω . We call k a **symmetric kernel** if

$$k(x, y) = k(y, x)$$

for all $x, y \in \Omega$.

Definition 1.2. A function $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is said to be **radial** if there exists a function $\phi : [0, \infty] \rightarrow \mathbb{R}$ such that

$$\Phi(x) = \phi(\|x\|_2)$$

for all $x \in \mathbb{R}^d$. Such a function is traditionally called a **radial basis function (rbf)**.

1.1.1 Examples

Example ((Inverse) multiquadratics). **Multiquadratics** are of the form

$$\phi(r) = (1 + \alpha r^2)^\beta$$

for positive β , while **inverse multiquadratics** have a $\beta < 0$.

Example (Polyharmonic kernels). **Polyharmonic kernels** are of the form

$$\phi(r) = r^\beta \log(|r|)$$

where $\beta \in 2\mathbb{Z}$.

The special case $\beta = 2$ is the so-called **thin-plate spline**. It relates to the partial differential equation that describes the bending of thin plates.

While the previous examples were monotone kernels (as a function of r), these are not!

Example (Wendland's kernels). **Wendland's kernels** are of the form

$$\phi_{a,1} := (1 - r)_+^{(a+1)} (1 + (a+1)r)$$

with the **cut-off function**

$$(x)_+ := \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Remark. There are also non radial kernels:

Translation-invariant or **stationary** kernels are functions of differences:

$$k(x, y) = \Phi(x - y).$$

For periodic setups, we have the **Dirichlet kernel** as an example:

$$D(\phi) := \frac{1}{2} + \sum_{j=1}^N \cos(j\phi) = \frac{\sin\left(\left(n + \frac{1}{2}\right)\phi\right)}{2 \sin\left(\frac{\phi}{2}\right)}.$$

This is applied to differences $\phi = \alpha - \beta$ of angles or 2π -periodic arguments and is an important tool for Fourier series theory.

There are so called **zonal kernels**, for working on a sphere, where the kernel can be represented as a function of an angle. An example are functions of inner products, such as

$$k(x, y) = \exp(x^\top y).$$

Remember, $x^\top y$ is the (scaled) cosine of the angle between the two vectors.

Remark. We will see that a kernel k on Ω defines a function $k(x, \cdot)$ for all fixed $x \in \Omega$. The space

$$\mathcal{K}_0 := \text{span}\{k(x, \cdot) \mid x \in \Omega\}$$

can for example be used as a so called trial space in meshless methods for solving partial differential equations.

Remark. Kernels can always be restricted to subsets without losing essential properties. This easily allows kernels on embedded manifolds, e.g. the sphere.

Remark. Most of this works for complex kernels too.

1.1.2 Kernels in machine learning

In machine learning the data $x \in \Omega$ can be quite diverse and without (much) structure on first glance. For example consider images, text documents, customers, graphs, ...

Here, one views the kernel as a **similarity measure**, i.e.

$$k : \Omega \times \Omega \rightarrow \mathbb{R}$$

return a number $k(x, y)$ describing the similarity of two patterns x and y .

To work with general data, we first need to represent it in a Hilbert space \mathcal{F} , the so-called **feature space**. One considers the (application dependent) **feature map**

$$\Phi : \Omega \rightarrow \mathcal{F}.$$

The map describes each $x \in \Omega$ by a collection of **features** which are characteristic for a x and capture the essentials of elements of Ω . Since we are now in \mathcal{F} we can work with linear techniques. In particular we can use the scalar product in \mathcal{F} of two elements of Ω represented by their features:

$$\langle \Phi(x), \Phi(y) \rangle_{\mathcal{F}} =: k(x, y)$$

and define a kernel that way.

Remark. Given a kernel, neither the feature map nor the feature space are unique, as the following example shows:

Example. Let $\Omega = \mathbb{R}$, $k(x, y) = x \cdot y$. A feature map, with feature space $\mathcal{F} = \mathbb{R}$ is given by the identity map.

In \mathbb{R}^d , we can work with the standard scalar product

Reminder: A Hilbert space is a complete vector space with a scalar product

Such a construction can be made for any arbitrary kernel, therefore every kernel has many different feature spaces

But, the map $\Phi : \Omega \rightarrow \mathbb{R}^2$ defined by

$$\Phi(x) := (x/\sqrt{2}, x/\sqrt{2})$$

is also a feature map given the same k !

The following two examples show how one can handle non-euclidean origin spaces:

Example (Kernels on a set of documents). Consider a collection of documents. We represent each document as a **bag of words** and describe a bag as a vector in a space in which each dimension is associated with a term from the set of words, i.e. the dictionary. The feature map is

that is a set of frequencies of (chosen) words

$$\Phi(t) := (wf(w_1, t), wf(w_2, t), \dots, wf(w_d, t)) \in \mathbb{R}^d$$

where $wf(w_i, t)$ is the frequency of word w_i in document t .

A simple kernel is the vector space kernel

$$k(t_1, t_2) = \langle \Phi(t_1), \Phi(t_2) \rangle = \sum_{j=1}^d wf(w_j, t_1) wf(w_j, t_2).$$

Natural extensions to this kernel take e.g. word order, relevance or semantics into account, which can be achieved by using matrices in the scalar product:

$$k(t_1, t_2) = \langle S\Phi(t_1), S\Phi(t_2) \rangle = \Phi^\top(t_1) S^\top S \Phi(t_2).$$

Example (Graph kernels). Another non-euclidean data object are graphs, where the class of **random walk kernels** can be defined. These are based on the idea that given a pair of graphs, one performs random walks on both and counts the number of matching walks. With \tilde{A}_\times the adjacency matrix of the **direct product graph** of the two involved graphs, one defines:

$$k(G, H) := \sum_{j=1}^{N_G} \sum_{k=1}^{N_H} \sum_{l=1}^{\infty} \lambda_l [\tilde{A}_\times^l]_{j,k}.$$

More generally, one can define a **random walk graph kernel** k as

$$k(G, H) := \sum_{k=0}^{\infty} \lambda_k q_\times^T W_\times^k p_\times,$$

where W_\times is the **weight matrix** of the direct product graph, q_\times^T is the **stopping probability** on the direct product graph, and p_\times is the initial product distribution on the direct product graph.

1.1.3 Mercer kernels

More generally, one can consider kernels of the **Hilbert-Schmidt** or **Mercer** form

$$k(x, y) = \sum_{i \in I} \lambda_i \varphi_i(x) \varphi_i(y),$$

with certain functions $\varphi_i : \Omega \rightarrow \mathbb{R}$, certain positive **weights** λ_i and an index set I such that the following **summability condition** holds for all $x \in \Omega$:

$$k(x, x) = \sum_{i \in I} \lambda_i \varphi_i(x)^2 < \infty \quad (1)$$

Remark. Such kernels arise in machine learning if the functions φ_i each describe a feature of x and the feature space is the weighted l_2 -space of sequences with indices in I :

$$l_{2,I,\lambda} := \left\{ \{\xi_i\}_{i \in I} : \sum_{i \in I} \lambda_i \xi_i^2 < \infty \right\}.$$

This expansion also occurs when kernels generating positive operators are expanded into eigenfunctions on Ω . Such kernels can be views as arising from generalized convolutions. Generally kernels have three major application fields:

- Convolutions
- Trial spaces
- Covariances

We are mainly concerned with the last two.

Start of lecture 02
(11.04.24)

1.1.4 Properties of kernels

Consider an arbitrary set $X = \{x_1, \dots, x_N\}$ of N **distinct** elements of Ω and a symmetric Kernel K on $\Omega \times \Omega$.

N is the number of data points (always!)

$$f(x) = \sum_{j=1}^N a_j k(x_j, x), x \in \Omega$$

Remark. The set of $k(x_j, \cdot)$ might not be linear independent!

For X we construct the symmetric $N \times N$ Kernel matrix

$$K = K_{X,X} = (k(x_j, x_k))_{1 \leq j, k \leq N}$$

and obtain the interpolation problem

$$\hat{f}_k = f(x_k) = \sum_{j=1}^N a_j k(x_j, x_k)$$

in matrix form

$$K_{X,X} a = \hat{F}$$

Remark. With kernels, we will see that this is indeed solvable, because our matrix is symmetric and positive definite.

Definition 1.3. A Kernel on $\Omega \times \Omega$ is **symmetric and positive semidefinite**, if all Kernel matrices for all finite sets of distinct elements of Ω are symmetric and positive definite

semidefinite and definite have conflicting definitions in the literature!

Theorem 1.4. 1. Kernels arising from **feature maps** via

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

are positive semidefinite.

2. **Hilbert-Schmidt** or **Mercer Kernels**

$$k(x, y) = \sum_{i \in I} \varphi_i(x) \varphi_i(y)$$

are positive semidefinite.

Proof. 1. K is a **Gram(-ian)** matrix

2.

$$\begin{aligned} a^\top K a &= \sum_{j,k=1}^N a_j a_k k(x_j, x_k) = \sum_{j,k=1}^N a_j a_k \sum_{i \in I} \varphi_i(x_j) \varphi_i(x_k) \\ &= \sum_{i \in I} \lambda_i \sum_{j=1}^N a_j \varphi_i(x_j) \sum_{k=1}^N a_k \varphi_i(x_k) = \sum_{i \in I} \lambda_i \left(\sum_{j=1}^N a_j \varphi_i(x_j) \right)^2 \geq 0 \end{aligned}$$

A Gram matrix, is a matrix whose entries are given by inner products $K_{i,j} = \langle v_i, v_j \rangle$

□

Theorem 1.5. Let K be a symmetric positive semidefinite (spsd) Kernel on Ω . Then

1. $k(x, x) \geq 0$ for all $x \in \Omega$
2. $|k(x, y)|^2 \leq k(x, x)k(y, y)$ for all $x, y \in \Omega$
3. $2|k(x, y)|^2 \leq k(x, x)^2 + k(y, y)^2$ for all $x, y \in \Omega$
4. Any finite linear combination spsd Kernels with nonnegative coefficients gives a spsd Kernel. If any of these kernels is positive definite, and its coefficient is positive, then the combination of kernels is positive definite.
5. The product of two spsd kernels is spsd.
6. The product of two spd kernels is spd.

Proof. 1.: Use the set $\{x\}$ in Definition 1.3.

2.: Consider K of $\{x, y\}$. The determinant of such a positive semidefinite matrix is nonnegative, therefore

$$k(x, x)k(y, y) - k(x, y)^2 \geq 0$$

3.: $2ab \leq a^2 + b^2$ for all $a, b \in \mathbb{R}_0^+$. Therefore this follows from 2.

4.: Expand $a^\top K a$ to see this.

5.: Follows from Lemma 1.6.

6.: Follows from Lemma 1.6 and a bit more linear algebra. \square

Lemma 1.6 (Schur's Lemma). For two matrices A, B , the matrix C with elements

$$C_{jk} = A_{jk}B_{jk}$$

is called the **Schur product** or **Hardarmard product**. The Schur product of two psd matrices is psd.

Proof. Decompose $A = S^\top D S$ with S an orthogonal matrix and $D = \text{diag}(\lambda_1, \dots, \lambda_n)$ a diagonal matrix with $\lambda_i \geq 0$ the eigenvalues of A .

For all $q \in \mathbb{R}^N$ we look at

$$\begin{aligned} q^\top C q &= \sum_{j,k} q_j q_k a_{jk} b_{jk} = \sum_{j,k=1}^N q_j q_k \sum_{m=1}^N \lambda_m S_{jm} S_{km} \\ &= \sum_{m=1}^N \lambda_m \sum_{j,k=1}^N \underbrace{q_j S_{jm}}_{P_{j,m}} \underbrace{q_k S_{km}}_{P_{k,m}} b_{jk} = \sum_{m=1}^N \sum_{j,k=1}^N \underbrace{P_{jm} P_{km} b_{jk}}_{\geq 0 \text{ since } B \text{ is psd}} \geq 0 \end{aligned}$$

\square

Remark. Note that we only considered symmetric matrices, the above also holds if one of the matrices is not symmetric, but positive definite instead.

Remark. Our overall aim is to go from kernels to a **reproducing kernel Hilbert space (RKHS)**. Therefore we define candidate spaces and a bilinear form in a way we would expect them.

Definition. For spsd K we define

$$H := \text{span}\{k(x, \cdot) \mid x \in \Omega\}.$$

In the same way

$$L := \text{span}\{\delta_x \mid x \in \Omega, \delta_x : H \rightarrow \mathbb{R}\}$$

the linear space of all finite linear combinations of pointevaluation functionals actions on functions of H , where

$$\delta_x(f) = f(x).$$

It is important that elements of L act on elements of H ! These two spaces are paired in some sense.

We can, by definition, write all Elements from L and H as

$$\lambda_{a,X} := \sum_{j=1}^N a_j \delta_{x_j}$$

$$f_{a,X} := \sum_{j=1}^N a_j k(x_j, x) = \lambda_{a,X}^{(y)} k(x, y)$$

with $a \in \mathbb{R}^n, X = \{x_1, \dots, x_N\} \subset \Omega$ any arbitrary finite subset of Ω .

Remark. From $f_{a,X} = 0$ or $\lambda_{a,X} = 0$ it does not follow that $a = 0$!

We now define a bilinear form on L

$$\langle \lambda_{a,X}, \lambda_{b,Y} \rangle_L := \sum_{j=1}^M \sum_{k=1}^N a_j b_k k(x_j, x_k) = \lambda_{a,X}^{(x)} \lambda_{b,Y}^{(y)} k(x, y) = \lambda_{a,X}(f_{b,Y})$$

Added remark. One has to be a bit careful here: $\lambda_{a,X}^{(x)} \lambda_{b,Y}^{(y)} k(x, y)$ does not mean point wise multiplication, but concatenation:

$$\lambda_{a,X}^{(x)} \lambda_{b,Y}^{(y)} k(x, y) = \lambda_{a,X}^{(x)} (\lambda_{b,Y}^{(y)} k(x, y))$$

This is well-defined, since it is based on the actions of the functional and not the specific representation.

We can observe that the bilinear form is psd, since the kernel matrices have this property.

$$|\lambda_{a,X}(f_{b,Y})| = |\langle \lambda_{a,X}, \lambda_{b,Y} \rangle_L| \leq \|\lambda_{a,X}\|_L \|\lambda_{b,Y}\|_L \quad (\star)$$

Theorem 1.7. If K is spsd Kernel on Ω , the bilinear form $\langle \cdot, \cdot \rangle_L$ is positive definite in the space L of functionals defined on H . This L is a pre-Hilbert-space.

Proof. $0 = \langle \lambda_{a,X}, \lambda_{a,X} \rangle_L$ for $a \in \mathbb{R}^n, X = \{x_1, \dots, x_N\} \subset \Omega$.

Then by (\star) we have $\lambda_{a,X} = 0$ as a functional on H . □

Theorem 1.8. The mapping $R : \lambda_{a,X} \mapsto f_{a,X} = \lambda_{a,X}(k(\cdot, y))$ is linear and bijective from L onto H . Thus

$$\langle f_{a,X}, f_{b,Y} \rangle_H := \langle \lambda_{a,X}, \lambda_{b,Y} \rangle_L = \langle R(\lambda_{a,X}), R(\lambda_{b,Y}) \rangle_H$$

is an inner product on H . R acts as the Riesz map.

Proof. Linearity is obvious. If $f_{b,Y} = R(\lambda_{b,Y}) \in H$ vanishes, the definition of $\langle \cdot, \cdot \rangle_L$ implies that $\lambda_{b,Y}$ is orthogonal to all of L . Due to Theorem 1.7 it is zero. The Riesz property comes from the definition of $\langle \cdot, \cdot \rangle_L$:

$$\lambda_{a,X}(f_{b,Y}) = \langle \lambda_{a,X}, \lambda_{b,Y} \rangle_L = \langle f_{a,X}, f_{b,Y} \rangle_H = \langle R(\lambda_{a,X}), f_{b,Y} \rangle$$

□

Specializing to $\lambda_{1,x}$, i.e. to a point $x \in \Omega$, we get:

$$\begin{aligned} \langle \lambda_{1,x}, \lambda_{b,Y} \rangle_L &= \lambda_{1,x}(f_{b,Y}) = \delta_x(f_{b,Y}) = f_{b,Y}(x) \\ &= \langle R(\lambda_{1,x}), R(\lambda_{b,Y}) \rangle_H = \langle R(\lambda_{1,x}), f_{b,Y} \rangle_H = \langle k(x, \cdot), f_{b,Y} \rangle_H \end{aligned}$$

In other words, for all $f \in H, x \in \Omega$, we have

$$f(x) = \underline{\delta_x(f)} = \langle f, R(\delta_x) \rangle_H = \langle f, k(x, \cdot) \rangle_H$$

which is the so-called **reproduction equation** for values of functions from the inner product.

There might be different representations of elements in L, H . While the representation is not unique, the element is

Here we use that the functionals in L are restricted to functions in H

Added remark. In this lecture (\star) refers to the reproduction equation.

For $f = k(\cdot, y)$, we set $k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle_H$. We furthermore can observe $\forall f \in H, x \in \Omega$:

$$|\delta_x(f)| = |f(x)| = |\langle f, k(x, \cdot) \rangle_H| \leq \|f\|_H \|k(x, \cdot)\|_H = \|f\|_H \sqrt{K(x, x)}$$

and

$$\langle \delta_x, \delta_y \rangle_L = \langle k(x, \cdot), k(y, \cdot) \rangle_H = k(x, y) \forall x, y \in \Omega$$

$$\|\delta_x - \delta_y\|_L^2 = \|\delta_x\|_L^2 - 2\langle \delta_x, \delta_y \rangle + \|\delta_y\|_L^2 = k(x, x) - 2\langle k(x, \cdot), k(y, \cdot) \rangle_H + k(y, y)$$

is a **distance** on Ω :

$$\text{dist}(x, y) := \|\delta_x - \delta_y\|_L = \sqrt{k(x, x) - 2\langle k(x, \cdot), k(y, \cdot) \rangle_H + k(y, y)}.$$

We see that for all $x, y \in \Omega$

$$|f(x)f(y)| \leq \|f\|_H \|\delta_x - \delta_y\|_L = \|f\|_H \text{dist}(x, y)$$

and therefore all functions in H are continuous with respect to this distance.

Theorem 1.9. Each symmetric positive definite kernel k on a set Ω is the **reproducing kernel** of a Hilbert space called the **native space** $\mathcal{H} = \mathcal{N}_k$ of the kernel. This Hilbert space is unique and it is a space of functions on Ω . The kernel k fulfills

$$\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x) \quad (2)$$

Proof. Citation: The existence of native spaces follows from standard Hilbert space arguments, see e.g. chapter 11 from the lecture notes of Schaback. \square

\mathcal{H} can be constructed as the closure of H

Added remark. The good ideas are from Schaback, the errors are from me, Prof. Garcke

The errors in this script are largely due to me :)

Proof of uniqueness:

If k is a reproducing kernel in a different Hilbert space T , we observe

$$\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} = k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle_T$$

which shows that the inner products coincide on H . Since T is a Hilbert space, it must contain the closure \mathcal{N}_k of H as a closed subspace. For T to be larger than \mathcal{H} non-zero element $f \in T$ must exist that is orthogonal to \mathcal{N}_k and in particular to H . We observe

$$f(x) = \langle f, k(x, \cdot) \rangle_T = 0 \quad \forall x \in \Omega.$$

which is a contradiction to $f \neq 0$, because of (2) for T .

Dual spaces:

$\delta_x : \mathcal{N}_k \rightarrow \mathbb{R}, f \mapsto f(x)$ for all $f \in \mathcal{N}_k, x \in \Omega$.

The dual space \mathcal{N}_k^* of \mathcal{N}_k is again a Hilbert space.

$$\begin{aligned} R : \mathcal{N}_k^* &\rightarrow \mathcal{N}_k \\ \lambda(f) &= \langle f, R(\lambda) \rangle_{\mathcal{N}_k} \forall f \in \mathcal{N}_k, \lambda \in \mathcal{N}_k^* \\ \langle \lambda, \mu \rangle_{\mathcal{N}_k^*} &= \langle R(\lambda), R(\mu) \rangle_{\mathcal{N}_k} \forall \lambda, \mu \in \mathcal{N}_k^* \end{aligned}$$

Also via the reproducing equation 2

$$\delta_x(f) = \langle f, k(x, \cdot) \rangle_{\mathcal{N}_k} \forall f \in \mathcal{N}_k, x \in \Omega.$$

So $k(x, \cdot)$ is the **Riesz representer** $R(\delta_x)$ of δ_x

$$\begin{aligned} \langle \delta_x, \delta_y \rangle_{\mathcal{N}_k^*} &= \langle R(\delta_x), R(\delta_y) \rangle_{\mathcal{N}_k} = k(x, y) & \forall x, y \in \Omega \\ \|\delta_x\|_{\mathcal{N}_k^*} &= \|k(x, \cdot)\|_{\mathcal{N}_k} = \sqrt{k(x, x)} & \forall x \in \Omega \\ \lambda(f) &= \langle f, \lambda^* k(x, \cdot) \rangle_{\mathcal{N}_k} & \forall f \in \mathcal{N}_k, \lambda \in \mathcal{N}_k^* \end{aligned}$$

so that $\lambda^* k(x, \cdot)$ is the Riesz representer of λ .

1.2 Reproducing Kernel Hilbert Space (RKHS)

Definition 1.10. A Hilbert space \mathcal{H} of functions on a set Ω with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is called **RKHS** if there is a kernel function $k : \Omega \times \Omega \rightarrow \mathbb{R}$ with $k(x, \cdot) \in \mathcal{H}$ for all $x \in \Omega$ and the reproducing kernel property

$$f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}} \quad \forall x \in \Omega, f \in \mathcal{H}$$

This directly implies

$$k(y, x) = \langle k(y, \cdot), k(x, \cdot) \rangle_{\mathcal{H}} = \langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} = k(x, y).$$

For positive semi-definiteness one can use a Gram matrix argument or take any $X = \{x_1, \dots, x_N\} \subset \Omega$ and $a \in \mathbb{R}^n$

$$\begin{aligned} \sum_{j,k=1}^N a_j a_k k(x_j, x_k) &= \sum_{j,k=1}^N a_j a_k \langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} \\ &= \left\langle \sum_{j=1}^N a_j k(x_j, \cdot), \sum_{k=1}^N a_k k(x_k, \cdot) \right\rangle_{\mathcal{H}} \\ &= \left\| \sum_{j=1}^N a_j k(x, \cdot) \right\|_{\mathcal{H}}^2 \geq 0 \end{aligned}$$

Theorem 1.11. Each Hilbert space \mathcal{H} of real valued functions on some set Ω with point evaluation functionals

$$\delta_x : f \mapsto f(x) \quad \forall f \in \mathcal{H}$$

is a RKHS with a unique positive definite kernel k on Ω . The kernel is uniquely defined by providing the Riesz representers of the (continuous) point evaluation functionals. The space \mathcal{H} is the native space of the kernel.

Proof. Under the given hypothesis, there must be a Riesz representer of δ_x . By the definition of the Riesz map it takes the form $k(x, \cdot) \in \mathcal{H}$ satisfying the reproduction equation 2.

In other words, any such Hilbert space has a symmetric positive definite kernel.

The final statement follows from theorem 1.9, because the native space and \mathcal{H} are Hilbert spaces that contain all $k(x, y)$. \square

Theorem 1.12. If a Hilbert (sub-)space of functions on Ω has a finite orthogonal basis v_1, \dots, v_N the reproducing kernel is

$$k_N(x, \cdot) = \sum_{j=1}^N v_j(x) v_j(\cdot) \quad \forall x \in \Omega$$

In case of a subspace we have

$$\sum_{j=1}^N |v_j(x)|^2 = k_N(x, x) \leq k(x, x) \quad \forall x \in \Omega$$

Which in some sense means that larger dimensions of the subspace can't add too much to the norm

Proof. The kernel must have a representation in the ONB

$$k_N(x, \cdot) = \sum_{j=1}^N \langle k_N(x, \cdot), v_j \rangle_{\mathcal{H}} v_j(\cdot) \stackrel{(2)}{=} \sum_{j=1}^N v_j(x) v_j(\cdot)$$

For the subspace,

$$\begin{aligned} k_N(x, x) &= \langle k_N(x, \cdot), k_N(x, \cdot) \rangle_{\mathcal{H}} \\ &\stackrel{\text{Hilbert subspace}}{=} \langle k_N(x, \cdot), k(x, \cdot) \rangle_{\mathcal{H}} \\ &\leq \sqrt{k_N(x, x)} \sqrt{k(x, x)} \quad \forall x \in \Omega \end{aligned}$$

\square

Added remark. *The subspace property does not hold for arbitrary Hilbert spaces, this tells us that a RKHS is really not the same as a normal Hilbert space!*