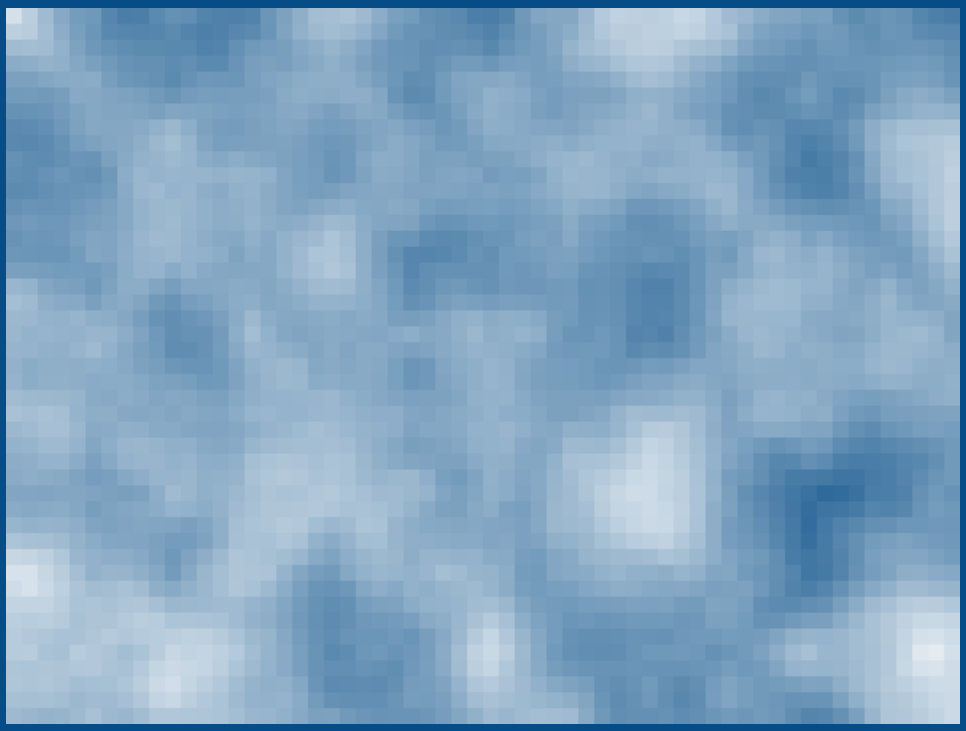
Lecture notes on Scientific Computing 2

Written by
Manuel Hinz
mh@mssh.dev or s6mlhinz@uni-bonn.de

Lecturer
Prof. Dr. Jochen Garcke
garcke@ins.uni-bonn.de



University of Bonn
Summer semester 2024
Last update: April 23, 2024

Contents

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

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

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

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: Asses credit risk
Example: Chemistry / energy of molecules. This needs a kernel on graphs
Example: Time series. This needs a kernel on time series

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

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

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:*

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

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 $k : \Omega \times \Omega \to \mathbb{R}$ 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 $\Phi : \mathbb{R}^d \to \mathbb{R}$ is said to be **radial** if there exists a function $\phi : [0,\infty] \to \mathbb{R}$ 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*

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

*with the **cut-off function***

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

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:*

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

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*

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

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

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

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

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.

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

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

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**

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

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

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.*

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

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

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

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

$$\Phi(t) \coloneqq (wf(w_1, t), wf(w_2, t), \ldots, wf(w_d, t)) \in \mathbb{R}^d$$

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:

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

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:*

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

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

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

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

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

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:*

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

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

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

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

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

2.

$$a^\intercal K a = \sum_{j,k=1}^{N} a_j a_k k(x_j, x_k) = \sum_{j,k=1}^{N} a_j a_k \sum_{i \in I} \varphi_i(x) \varphi_i(y)$$

$$= \sum_{i \in I} \lambda_i \sum_{j=1}^{N} a_j \varphi_i(x_j) \sum_{k=1}^{N} a_k \varphi_i(x_k) = \sum_{i \in I} \lambda_i \left(\sum_{j=1}^{N} a_j \phi_i(x_j) \right)^2 \ge 0$$

□

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 positve 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 $a^\intercal K a$ to see this.
5.: Follows from Lemma 1.6.
6.: Follows from Lemma 1.6 and a bit more linear algebra. □

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 $A = S^\intercal D S$ with S an orthogonal matrix and $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ a diagonal matrix with $\lambda_i \geq 0$ the eigenvalues of A.
For all $q \in \mathbb{R}^N$ we look at

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

□

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 expec them.*

Definition. *For spsd K we define*

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

In the same way

$$L \coloneqq span\{\delta_x \mid x \in \Omega, \delta_x : H \to \mathbb{R}\}$$

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

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

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

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 $a = 0$!*

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

We now define a bilinear form on L

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

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.

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

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. □

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

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

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:

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

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

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

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

Start of lecture 03 (16.04.24)

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*

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

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

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

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

The erros 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

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

which is a contradiction to $f \neq 0$, because of (2) for T.
Dual spaces:
$\delta_x : \mathcal{N}_k \to \mathbb{R}, f \mapsto f(x)$ 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.

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

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

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

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 $k : \Omega\times\Omega \to \mathbb{R}$ with $k(x,\cdot) \in \mathcal{H}$ forall $x \in \Omega$ and the reproducing kernel property*

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

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 $a \in \mathbb{R}^n$

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

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

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

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 th 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 $k(x,y)$. □

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*

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

In case of a subspace we have

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

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

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

For the subspace,

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

□

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!*

Start of lecture 04 (18.04.24)

Remember: Kernels of the Mercer form

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

with the summability condition

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

Then observe

$$\begin{aligned} |f(x)| &= \left| \sum_{i\in I} \langle f, \varphi_i\rangle_{\mathcal{H}} \phi_i(x) \right| \\ &\leq \sum_{i\in I} \left| \frac{\langle f,\varphi_i\rangle_{\mathcal{H}}}{\sqrt{\lambda_i}} \right| |\varphi_i(x)|\sqrt{\lambda_i} \\ &\leq \sqrt{\sum_{i\in I} \frac{\langle f,\varphi_i\rangle_{\mathcal{H}}}{\sqrt{\lambda_i}}} \sqrt{\underbrace{\sum_{i\in I} \varphi_i^2(x)\lambda_i}_{<\infty}} \end{aligned}$$

$$\mathcal{H} := \left\{ f \in \mathcal{H} : \|f\|_\lambda^2 = \sum_{i\in I} \frac{\langle f,\varphi_i\rangle_{\mathcal{H}}}{\sqrt{\lambda_i}} \right\}$$

$$\langle f, g\rangle_\lambda = \sum_{i\in I} \frac{\langle f,\varphi_i\rangle_{\mathcal{H}} \langle g,\varphi_i\rangle_{\mathcal{H}}}{\lambda_i} \qquad \forall f,g \in \mathcal{H}_\lambda \tag{3}$$

Using 3 as the kernel, we have to check if all $f_x := k(x,\cdot) \in \mathcal{H}_\lambda$.
Observe

$$\langle f_x, \varphi\rangle_{\mathcal{H}} = \lambda_i \varphi_i(x)$$

and

$$\sum_{i\in I} \frac{\langle f_x,\varphi_i\rangle_{\mathcal{H}}^2}{\lambda_i} = \sum_{i\in I} \lambda_i \varphi_i^2(x) < \infty$$

to see $f_x \in \mathcal{H}_\lambda$.

Check the reproduction equation

$$\begin{aligned} \langle f, k(x,\cdot)\rangle_\lambda &= \sum_{i\in I} \frac{\langle f,\varphi_i\rangle_{\mathcal{H}} \langle k(x,\cdot),\varphi_i\rangle_{\mathcal{H}}}{\lambda_i} \\ &= \sum_{i\in I} \frac{\langle f,\varphi_i\rangle \lambda_i \varphi_i(x)}{\lambda_i} = f(x) \qquad \forall x \in \Omega \end{aligned}$$

The kernel therefore reproduces on $\mathcal{H}_\lambda$. The proves theorem **??**.
If a Hilbert space of functions on Ω has a countable ONB $\{\varphi_i\}_{i\in I}$, each summability condition (**) leads to a reproducing mercer kernel (*) for a suitable subspace of functions with continuous point evaluations.

Corollary 1.13. *The spaces $\mathcal{H}_\lambda$ defined above are the natives spaces of the corresponding Mercer kernels.*

Example (Trigometric polynomials)**.** *Consider the space of trigometric polynomials* $\frac{1}{\sqrt{2}}, \cos(nx), \sin(nx), n \in \mathbb{N}$ *which are* ***orthonormal*** *in the inner product*

$$\langle f, g \rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t)g(t)dt.$$

With $I = (0,0) \cup (\mathbb{N}, 0) \cup (0, \mathbb{N})$

$$\varphi_i(x) = \begin{cases} \frac{1}{\sqrt{2}} & i = (0,0) \\ \cos(nx) & i = (n,0), n \geq 1 \\ \sin(nx) & i = (n,0), n \geq 1 \end{cases}$$

So for $f \in \mathcal{H}$

$$f = \sum_{i \in I} \langle f, \varphi_i \rangle_{\mathcal{H}} \varphi_i.$$

All φ_i *are uniformly bounded, so the summability condition does hold when the weights are summable.*
Fixing some $m \geq 1$*, we define*

$$\lambda_i = \begin{cases} 1 & i = (0,0) \\ n^{-2m} & otherwise \end{cases}$$

We set the Mercer kernel

$$k_{2m}(x,y) \coloneqq \frac{1}{\sqrt{2}} + \sum_{n=1}^{\infty} n^{-2m} (\cos(nx)\cos(ny) + \sin(nx)\sin(ny))$$

this can also be tought of as an ***extension kernel***

This can be rewritten with the usual trigonometric rules

One can see $K''_{2m} = K_{2m-2}$*, so* $K_2 m$ *piecewise polynomial of degree* $2m$*, which is* $2m-2$ *times differentiable.*

1.2.1 Kernels for subspaces

Let us fix a nonempty set $X \subset \Omega$ and look at the closed subspace

$$\mathcal{H}_X \coloneqq \overline{\text{span}\{k(x,\cdot) | x \in X\}} \subseteq \mathcal{H}$$

Projector for $\mathcal{H}$ to the closed subspace $\mathcal{H}_0$: $\pi_0 : \mathcal{H} \to \mathcal{H}_0$ with properties

This is NOT $\{\mathcal{H}_0\}$, but a generic subspace

- $\pi_0^2 = \pi_0$
- π_0 gives unique best approximation in $\mathcal{H}_0$, $u \mapsto u_0$
- $u_0 \perp u - u_0$
- $\text{Id} - \pi_0$ projects onto the orthogonal complement $\mathcal{H}_0^{\perp} = \{u \in \mathcal{H} \mid \langle u, v \rangle_{\mathcal{H}} \forall v \in \mathcal{H}_0\}$
- $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_0^{\perp}$

Theorem 1.14. *Let* $\mathcal{H}_0$ *be a closed subspace of* $\mathcal{H}$ *with reproducing kernel* k_0 *and let* $\pi_0 : \mathcal{H} \to \mathcal{H}_0$ *be the projection onto* $\mathcal{H}_0$*.*
The subspace kernel is

$$k_0(x,\cdot) = \pi_0 k(x,\cdot)$$

for all $x \in \Omega$*. The reproducing kernel for the orthogonal complement* $\mathcal{H}_0^{\perp}$ *is* $k - k_0$*.*

Proof. $\text{Id} = \pi_0 + (\text{Id} - \pi_0) = \pi_0 + \pi_0^{\perp}$.
Thus $f(x) = (\pi_0 \circ f)(x) + (\pi_0^{\perp} \circ f)(x)$ inserted into the reproducing equation

$$\begin{aligned} f(x) &= \langle f, k(x,\cdot) \rangle_{\mathcal{H}} \\ &= \langle \pi_0 f + \pi_0^{\perp} f, \pi_0 k(x,\cdot) + \pi_0^{\perp} k(x,\cdot) \rangle_{\mathcal{H}} \\ &= \langle \pi_0 f, \pi_0 k(x,\cdot) \rangle + \langle \pi_0^{\perp} f, \pi_0^{\perp} k(x,\cdot) \rangle \end{aligned}$$

Using $f \in \mathcal{H}_0$ and $f \in \mathcal{H}_0^\perp$ eliminates on part of the sum each and the statements follow. □

Remark. *Orthogonal space decompositions correspond to additive kernel decompositions using the appropriate projections.*

Theorem 1.15. *Let $X \subseteq \Omega$ be nonempty. For the closed subspace $\mathcal{H}_X$ it holds*

$$\mathcal{H}_X^\perp = \{f \mid f \in \mathcal{H} : f(X) = \{0\}\}.$$

This is a statement we would expect for our later use case of approximation and interpolation

Proof. If $f(X) = \{0\}$, then $\langle f, v\rangle_\mathcal{H} = 0 \forall v \in \mathcal{H}_X$.

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

since $f \in \mathcal{H}_X^\perp$ by the reproduction equation and conversely we set for $f \in \mathcal{H}_X^\perp$ that $f(X) = \{0\}$. □

With π_X the projector from $\mathcal{H}$ to $\mathcal{H}_X$ we denote

$$f_X \coloneqq \pi_X(f).$$

Standard results from Hilbert space theory gives us

Theorem 1.16. *Each function $f \in \mathcal{H}$ has an orthogonal decomposition*

$$f = f_X + f_{X^\perp}$$

with $f_X \in \mathcal{H}_X$ and $f_{X^\perp} \in \mathcal{H}_X^\perp$. In particular each $f \in \mathcal{H}$ has an interpolant $f_X \in \mathcal{H}_X$ recovering the values of f on X. Additionally

$$\|f - f_X\|_\mathcal{H} = \inf_{g \in \mathcal{H}_X} \|f - g\|_\mathcal{H}$$

and

$$\|f_X\|_\mathcal{H} = \inf_{\substack{g \in \mathcal{H}: \\ \forall x \in X}} \|g\|_\mathcal{H} = \inf_{v \in \mathcal{H}_{X^\perp}} \|f - v\|_\mathcal{H}$$

Corollary 1.17. *The interpolant $f_X \in \mathcal{H}_X$ to a function f on X is at the same time the best approximation to f from all functions in $\mathcal{H}_X$.*

This is just the previous theorem in words

Corollary 1.18. *The interpolant $f_X \in \mathcal{H}_X$ to a function f on X minimizes the norm under all interpolants from the full space $\mathcal{H}$.*

This property is usefull, if the norm encodes smoothness as well. Penalizing unnecessary changes of the function

Corollary 1.19. *For all sets $X \subseteq Y \subseteq \Omega$ and $f \in \mathcal{H}$ we have*

$$\|f_X\|_\mathcal{H} \leq \|f_Y\|_\mathcal{H} \leq \|f\|_\mathcal{H}$$

and

$$\|f\|_\mathcal{H} \geq \|f - f_X\|_\mathcal{H} \geq \|f - f_y\|_\mathcal{H}$$

where for completeness we define $f_\emptyset = 0, f_{\emptyset^\perp} = f$ and $\mathcal{H}_\emptyset = \{0\}$ with $\mathcal{H}_{\emptyset^\perp} = \mathcal{H}$.

Consider only $f(\cdot) = k(x,\cdot)$ for a fixed $f, x \in \Omega$.

Start of lecture 05 (23.04.24)

Definition 1.20. *The function*

$$P_X(x) \coloneqq \|k(x,\cdot) - k_X(x,\cdot)\|_\mathcal{H} \qquad x \in \Omega$$

*is called **power function** w.r.t. the set X and the kernel k.*
*A different definition goes with the **error functional** $\epsilon_{X,x} \in \mathcal{H}^*$*

$$\epsilon_{X,x} f \mapsto f(x) - (\Pi_X(f))(x).$$

The power function is then defined as $P_X(x) \coloneqq \|\epsilon_{X,x}\|_{\mathcal{H}^*}$.

Theorem 1.21. *The two definitions for the power function are equivalent. P_X has the following properties*

1. $P_X(x) = 0 \ \forall x \in X$
2. $P_\emptyset(x)^2 = k(x,x) \ \forall x \in \Omega$
3. $P_\Omega(x) = 0 \ \forall x \in \Omega$
4. $0 = P_\Omega(x) \leq P_Y(x) \leq P_X(x) \leq P_\emptyset(x)$ *for* $X \subseteq Y \subseteq \Omega$
5. $P_X(x) = \inf_{g \in \mathcal{H}_X} \|k(x,\cdot) - g\|_\mathcal{H}$
6. $P_X(x) = \sup_{f \in \mathcal{H}, \|f\|_\mathcal{H} \leq 1, f(X) = \{0\}} f(x) \ \forall x \in \Omega$
7. $\forall x \in \Omega, f \in \mathcal{H}$

$$|f(x) - f_X(x)| = |f_X^\intercal(x)| \leq P_X(x) \|f_X^\intercal(x)\|_\mathcal{H} = P_X(x) \|f - f_X\|_\mathcal{H} \leq P_X(x) \|f\|_\mathcal{H}$$

Added remark. *General approximation goal: Split the approximation error into an error of the space and an error of the function (similarly to SC1).*
One aim is to generalize 7. to not rely on a specific point.

Proof. Due to $\langle \epsilon_{X,x}, \epsilon_{X,x} \rangle_{\mathcal{H}^*} = \langle R(\epsilon_{X,x}), R(\epsilon_{X,x}) \rangle_\mathcal{H}$ we have to show that the Riesz representer of $\delta_x \circ \Pi_X$ is $K_X(x,\cdot)$.

$$\begin{aligned}
\langle f, R(\delta_x \circ \Pi_X) \rangle &= \delta_x \circ \Pi_X(f) = f_X(x) = \langle f_X, k(x,\cdot) \rangle_\mathcal{H} \\
&= \langle f_X, k_X(x,\cdot) + k_{X^\intercal}(x,\cdot) \rangle_\mathcal{H} \\
&\overset{f_X \perp K_{X^\intercal}}{=} \langle f_X, k_X(x,\cdot) \rangle_\mathcal{H} = \langle f - f_{X^\intercal}, k_X(x,\cdot) \rangle_\mathcal{H} \\
&\overset{f_{X^\intercal} \perp K_X}{=} \langle f, k_X(x,\cdot) \rangle_\mathcal{H}
\end{aligned}$$

Proof of 7.:

$$\begin{aligned}
f(x) - f_X(x) = f_{X^\intercal}(x) &= \langle f_{X^\intercal}, k(x,\cdot) \rangle_\mathcal{H} \\
&= \langle f_{X^\intercal}, k(x,\cdot) - \underbrace{k_X(x,\cdot)}_{f_{X^\intercal} \perp} k_X(x,\cdot) \rangle \\
&\overset{\text{C.S.}}{\leq} \cdots = \|f_{X^\intercal}\| \mathcal{H} P_X(x)
\end{aligned}$$

Proof of 6.:
We see from the first inequality

$$P_X(x) \geq \sup_{\|f_{X^\intercal}\|_\mathcal{H} \leq 1} |f_{X^\intercal}(x)|$$

and equality must hold for the representer of $\epsilon_{X,x}$. □

Notice the connection to operator norm approaches to 6.

Remark. *Consider the subspace $\mathcal{H}_X^* = \overline{span\{\delta_x \mid x \in X\}}$ of the dual space of $\mathcal{H}$. Then 5. can equivalently be given as*

$$P_X(x) = \inf_{\lambda \in \mathcal{H}_X^*} \|\delta_x\|_{\mathcal{H}^\intercal} \tag{4}$$

Consider the interpolation of $f(x) = k(x,\cdot) \in \mathcal{H}$.
For $x \in \Omega$ we get for the interpolant in $\mathcal{H}_X$

$$k(x_k, x) = \sum_{j=1}^{N} u_j(x) k(x_j, x_k) \qquad 1 \leq k \leq N \tag{5}$$

which has solution coefficients $u_j(x)$ as a function on Ω.

Added remark. *If the kernel matrix is invertible, u_j is either 0 or 1. See Lagrange interpolation? In our setting it is enough to know that it exists, but might not be unique.*

Theorem 1.22. *If the kernel matrix is non-singular the u_j from 4 are $\in \mathcal{H}_X$ and there is a Lagrange basis $u_j(x_k) = \delta_{jk}, 1 \leq j, k \leq N$.*
In general it still holds

$$f_X(x) = \sum_{j=1}^{N} u_j(x) f(x_j)$$

In the formula the influence of X and f are separated.

This is sometimes called quasi-interpolation

Proof. The first statement follows from 5. From the second:

$$\begin{aligned} f_X(x) &= \sum_{k=1}^{N} a_k k(x_k, x) \\ &= \sum_{k=1}^{N} a_k \sum_{j=1}^{N} u_j(x) k(x_j, x_k) \\ &= \sum_{j=1}^{N} a_j(x) \underbrace{\sum_{k=1}^{N} a_k k(x_j, x_k)}_{= f_{a,X} = f(x_j) \forall x_j \in X} = \sum_{j=1}^{N} u_j(x) f(x_j) \end{aligned}$$

□

Theorem 1.23. *The power function has the following explicit representation:*

$$P_X(x) = k(x,x) - 2\sum_{j=1}^{N} u_j(x) k(x_j, x) + \sum_{j=1}^{N}\sum_{k=1}^{N} u_j(x) u_k(x) k(x_j, x_k) = k(x,x) - k_X(x,x)$$

Proof. For $K_X \in \mathcal{H}_X$ $k_X(x,z) = \sum_{j=1}^{N} u_j(x) k(x_j, z)$

$$\begin{aligned} P_X^2(x) &= \langle k(x,\cdot) - k_X(x,\cdot), k(x,\cdot) - k_X(x,\cdot) \rangle_{\mathcal{H}} \\ &= k(x,x) - 2\langle k(x,\cdot), \sum_{j=1}^{N} u_j(x) k(x_j,\cdot) \rangle_{\mathcal{H}} + \sum_{j=1}^{N} \underbrace{\sum_{k=1}^{N}}_{a} u_j(x) \underbrace{u_k(x) k(x_j, x_k)}_{\text{with a } = k(x, x_j)} \\ &= k(x,x) - \underbrace{\sum_{j=1}^{N} u_j(x) k(x_j, x)}_{k_X(x,x)} \end{aligned}$$

□

Consider $f_X(x) = \sum_{j=1}^{N} u_j(x) f(x_j)$ the interpolant on X.
Let us also consider arbitrary estimation formulas

$$(x, f) \mapsto \sum_{j=1}^{N} v_j(x) f(x_j)$$

with no assumptions on the scalars v_j. For fixed x we get for the error functional
$f \mapsto f(x) - \sum_{j=1}^{N} v_j(x) f(x_j) = \left(\delta_x \sum_{j=1}^{N} v_j(x) \delta_{x_j}\right)(f)$.
Ad for optimal estimation for all $f \in \mathcal{H}$, we should choose v_j to minimize the following expression:

$$V_{X,v}(x) \coloneqq \|\delta_x - \sum_{j=1}^{N} v_j(x) \delta_{x_j}\|_{\mathcal{H}^*}.$$

Remember the dual form of the fifth property 4:

$$P_X(x) = \inf_{\lambda \in \mathcal{H}^*} \|\delta_x - \lambda\|_{\mathcal{H}^*}$$

we also saw that the function u_j are the solution.
We also see that the optimal error, in the worst case sense, is described to be the power function.

Theorem 1.24. *In the above sense, kernel based approximation yields the best linear estimation of unknown function values $f(x)$ from known function values $f(x_j)$ at points x_j.*

List of Lectures