Scattered Data Approximation

Fabian Bosshard

Contents

Preface

This document contains unofficial student-made notes for the course Scattered Data Approximation taught by Michael Multerer with the assistance of Jacopo Quizi in Winter 2025/2026 at the Università della Svizzera italiana. It is mainly based on [1]. Section 4 is summarized from [2]. The textbooks used in the course were [3, 4]. If you spot an error, please report it to fabianlucasbosshard@gmail.com. The LaTeX source is available at `https://github.com/fabianbosshard/usi-informatics-course-summaries`.

References

[1] Michael Multerer. Scattered Data Approximation. Università della Svizzera italiana, 2025. URL: `https://muchip.github.io/resources/SDA_notes.pdf`.

[2] Tilo Arens et al. Mathematik. 5th ed. Springer Spektrum, 2022. URL: `https://link.springer.com/book/10.1007/978-3-662-64389-1`.

[3] Gregory Fasshauer. Meshfree Approximation Methods with MATLAB. 1st ed. World Scientific, 2007. URL: `https://doi.org/10.1142/6437`.

[4] Holger Wendland. Scattered Data Approximation. 1st ed. Cambridge University Press, 2005. URL: `https://doi.org/10.1017/CBO9780511617539`.

1 Matrix Basics

1.1 Derivatives

For a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$ we use the Jacobian (numerator-layout) convention

$$\frac{\partial f}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times n}$$

i.e. a row vector. The column-gradient is obtained by transposition: $\nabla_{\boldsymbol{x}} f = \left(\frac{\partial f}{\partial \boldsymbol{x}}\right)^\top$

Linear form. For $f(\boldsymbol{x}) = \boldsymbol{b}^\top \boldsymbol{x}$ with $\boldsymbol{b} \in \mathbb{R}^n$,

$$\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{b}^\top \boldsymbol{x}\right) = \boldsymbol{b}^\top$$

Quadratic form. Let $\underline{\boldsymbol{A}} \in \mathbb{R}^{n\times n}$ and define $f(\boldsymbol{x}) = \boldsymbol{x}^\top \underline{\boldsymbol{A}}\, \boldsymbol{x}$. Writing

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i\, A_{ij}\, x_j,$$

we differentiate componentwise. For $k \in \{1, \dots, n\}$,

$$\begin{aligned}
\frac{\partial f}{\partial x_k} &= \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} x_i\, A_{ij}\, x_j\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}\, \frac{\partial}{\partial x_k}(x_i x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}\left(\frac{\partial x_i}{\partial x_k}\, x_j + x_i\, \frac{\partial x_j}{\partial x_k}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}\left(\delta_{ik}\, x_j + x_i\, \delta_{jk}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}\, \delta_{ik}\, x_j + \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}\, x_i\, \delta_{jk} \\
&= \sum_{j=1}^{n}\left(\sum_{i=1}^{n} A_{ij}\, \delta_{ik}\right) x_j + \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} A_{ij}\, \delta_{jk}\right) \\
&= \sum_{j=1}^{n} A_{kj}\, x_j + \sum_{i=1}^{n} x_i\, A_{ik} \\
&= (\underline{\boldsymbol{A}}\boldsymbol{x})_k + (\underline{\boldsymbol{A}}^\top \boldsymbol{x})_k \\
&= (\underline{\boldsymbol{A}}\boldsymbol{x} + \underline{\boldsymbol{A}}^\top \boldsymbol{x})_k \\
&= ((\underline{\boldsymbol{A}} + \underline{\boldsymbol{A}}^\top)\boldsymbol{x})_k
\end{aligned}$$

where δ_{ik} denotes the Kronecker delta.

Collecting these components into a row vector gives

$$\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{x}} &= \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \\
&= \begin{bmatrix} ((\underline{\boldsymbol{A}} + \underline{\boldsymbol{A}}^\top)\boldsymbol{x})_1 & \cdots & ((\underline{\boldsymbol{A}} + \underline{\boldsymbol{A}}^\top)\boldsymbol{x})_n \end{bmatrix} \\
&= ((\underline{\boldsymbol{A}} + \underline{\boldsymbol{A}}^\top)\boldsymbol{x})^\top \\
&= \boldsymbol{x}^\top (\underline{\boldsymbol{A}} + \underline{\boldsymbol{A}}^\top)^\top \\
&= \boldsymbol{x}^\top (\underline{\boldsymbol{A}} + \underline{\boldsymbol{A}}^\top)
\end{aligned}$$

If furthermore $\underline{\boldsymbol{A}} = \underline{\boldsymbol{A}}^\top$, then

$$\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^\top \underline{\boldsymbol{A}}\, \boldsymbol{x}\right) = 2\boldsymbol{x}^\top \underline{\boldsymbol{A}}$$

Squared norm with a matrix. Let $\underline{\boldsymbol{A}} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$. For

$$f(\boldsymbol{x}) = \|\underline{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b}\|^2 = (\underline{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b})^\top (\underline{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b})$$

expand:

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top \underline{\boldsymbol{A}}^\top \underline{\boldsymbol{A}}\, \boldsymbol{x} - 2\, \boldsymbol{b}^\top \underline{\boldsymbol{A}}\, \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{b}$$

Differentiate using the previous rules:

$$\frac{\partial f}{\partial \boldsymbol{x}} = \boldsymbol{x}^\top \Big(\underline{\boldsymbol{A}}^\top \underline{\boldsymbol{A}} + (\underline{\boldsymbol{A}}^\top \underline{\boldsymbol{A}})^\top \Big) - 2\, \boldsymbol{b}^\top \underline{\boldsymbol{A}}$$

Since $\underline{\boldsymbol{A}}^\top \underline{\boldsymbol{A}}$ is symmetric, $(\underline{\boldsymbol{A}}^\top \underline{\boldsymbol{A}})^\top = \underline{\boldsymbol{A}}^\top \underline{\boldsymbol{A}}$, hence

$$\frac{\partial}{\partial \boldsymbol{x}} \|\underline{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b}\|^2 = 2\, \boldsymbol{x}^\top \underline{\boldsymbol{A}}^\top \underline{\boldsymbol{A}} - 2\, \boldsymbol{b}^\top \underline{\boldsymbol{A}} = 2\, (\underline{\boldsymbol{A}}\boldsymbol{x} - \boldsymbol{b})^\top \underline{\boldsymbol{A}}$$

1.2 Inverse

1.2.1 Properties

$(s\underline{\boldsymbol{A}})^{-1} = \frac{1}{s} \underline{\boldsymbol{A}}^{-1}$

1.2.2 Woodbury

says that the inverse of a rank-k correction of some matrix can be computed by doing a rank-k correction to the inverse of the original matrix.

Let $\underline{\boldsymbol{A}} \in \mathbb{R}^{n \times n}$ invertible, $\underline{\boldsymbol{U}} \in \mathbb{R}^{n \times k}$, $\underline{\boldsymbol{C}} \in \mathbb{R}^{k \times k}$ invertible, $\underline{\boldsymbol{V}} \in \mathbb{R}^{k \times n}$. (Recall that the product $\underline{\boldsymbol{U}}\underline{\boldsymbol{C}}\underline{\boldsymbol{V}}$ has rank at most k.)

Then

$$\boxed{(\underline{\boldsymbol{A}} + \underline{\boldsymbol{U}}\underline{\boldsymbol{C}}\underline{\boldsymbol{V}})^{-1} = \underline{\boldsymbol{A}}^{-1} - \underline{\boldsymbol{A}}^{-1}\underline{\boldsymbol{U}}\big(\underline{\boldsymbol{C}}^{-1} + \underline{\boldsymbol{V}}\underline{\boldsymbol{A}}^{-1}\underline{\boldsymbol{U}}\big)^{-1}\underline{\boldsymbol{V}}\underline{\boldsymbol{A}}^{-1}}$$

2 Introduction

We have data values y_i (e.g., measurements or samples) together with the corresponding data sites $\boldsymbol{x}_i$. We want a model s_f that matches the data according to a certain criterion.

We distinguish 2 cases, depending on our knowledge of the underlying data-generating process f:

- f is unknown, we seek a model s_f that allows extrapolation to unseen data sites. goodness of fit may be defined in various ways.
- f is known, for example as the solution operator of a PDE, the goal is to build a surrogate model s_f that is cheaper to evaluate than f itself.

scattered: data sites not located on a uniform grid

data sites: $X := \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subset \Omega \subset \mathbb{R}^d$

We start from the interpolation problem, which aims at exactly matching a given set of data:

Problem 2.1 (Scattered data interpolation)**.** Given data $(\boldsymbol{x}_i, y_i)$, $i = 1, \dots, N$, with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, find a continuous function s_f such that

$$s_f(\boldsymbol{x}_i) = y_i, \quad i = 1, \dots, N \tag{2.1}$$

◀

A common solution to Problem 2.1 is to assume that s_f is a linear combination of certain functions φ_j:

$$s_f(\boldsymbol{x}) = \sum_{j=1}^{N} c_j \varphi_j(\boldsymbol{x}) \tag{2.2}$$

Applying the interpolation condition (2.1) to (2.2) yields

$$\underline{\boldsymbol{A}}\boldsymbol{c} = \boldsymbol{y} \tag{2.3}$$

where the generalized Vandermonde matrix is given by

$$\underline{\boldsymbol{A}} := [\varphi_j(\boldsymbol{x}_i)]_{i,j=1}^{N} = \begin{bmatrix} \varphi_1(\boldsymbol{x}_1) & \cdots & \varphi_N(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(\boldsymbol{x}_N) & \cdots & \varphi_N(\boldsymbol{x}_N) \end{bmatrix} \tag{2.4}$$

Problem 2.1 is well-posed (i.e. a solution exists and is unique), iff $\underline{\boldsymbol{A}}$ is non-singular.

Example 2.1 (Polynomial interpolation)**.** Given $x_1, \dots, x_N \in \mathbb{R}$ and $y_1, \dots, y_N \in \mathbb{R}$, find a polynomial $p \in \Pi_{N-1} := \operatorname{span}\{1, x, \dots, x^{N-1}\}$ such that $p(x_i) = y_i$, $i = 1, \dots, N$. With respect to the monomial basis $\varphi_j(x) = x^{j-1}$, $j = 1, \dots, N$, the Vandermonde matrix is given by

$$\underline{\boldsymbol{A}} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^{N-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_N & \cdots & x_N^{N-1} \end{bmatrix} \in \mathbb{R}^{N \times N}$$

It can be shown that

$$\det(\underline{\boldsymbol{A}}) = \prod_{1 \le i < j \le N} (x_j - x_i)$$

Therefore, we have $\det(\underline{\boldsymbol{A}}) \neq 0$ whenever $x_i \neq x_j$ for $i \neq j$, which implies the well-posedness of the polynomial interpolation problem. ◀

Generalizing the ideas of Example 2.1 yields the following concept:

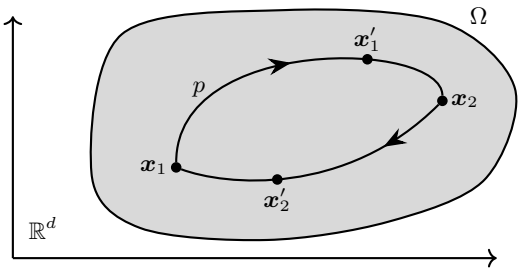
Definition 2.2 (Haar space)**.** Let $V \subset C(\Omega)$ be a finite-dimensional function space with basis $\{\varphi_1, \dots, \varphi_N\}$. V is a Haar space, iff $\det(\underline{\boldsymbol{A}}) \neq 0$ for every set of mutually distinct points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \Omega$, where $\underline{\boldsymbol{A}} = [\varphi_j(\boldsymbol{x}_i)]_{i,j=1}^{N}$ is the corresponding generalized Vandermonde matrix from (2.4). ◀

We have the following negative result:

Theorem 2.1 (Mairhuber-Curtis)**.** If $\Omega \subset \mathbb{R}^d$, $d > 1$, contains an interior point, then there exist no Haar spaces except for one-dimensional ones. ◁

Proof (Contradiction)**.** Let $d > 1$ and assume that there exists a Haar space $V \subset C(\Omega)$ with basis $\{\varphi_1, \dots, \varphi_N\}$ where $N > 1$. Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \Omega$ be a set of mutually distinct interior points. By assumption, $\det(\underline{\boldsymbol{A}}) \neq 0$.

Now consider a simple closed path p connecting only $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$:



We can interchange the positions of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, effectively swapping the first two rows of $\underline{\boldsymbol{A}}$, by continuously moving along p. This in turn changes the sign of $\det(\underline{\boldsymbol{A}})$.

Since $\varphi_j(\cdot)$, p and $\det(\cdot)$ are all continuous functions, there must exist $\boldsymbol{x}_1', \boldsymbol{x}_2'$ on p such that $\det(\underline{\boldsymbol{A}}') = 0$ for the corresponding generalized Vandermonde matrix $\underline{\boldsymbol{A}}'$, which contradicts the assumption that V is a Haar space. □

Theorem 2.1 implies that if we want to have a well-posed multivariate interpolation problem, we cannot choose the basis in advance, as we did in Example 2.1. Instead, the basis should depend on the data locations $\boldsymbol{x}_i$.

Example 2.2 (Interpolation by distance matrices)**.** Given data sites $x_1, \dots, x_N \in \mathbb{R}$ and values $y_1, \dots, y_N \in \mathbb{R}$, we make the ansatz

$$s_f(x) = \sum_{j=1}^{N} c_j |x - x_j|$$

which amounts to the linear spline interpolant if we solve (2.1). For $d > 1$, this can be generalized to:

$$s_f(\boldsymbol{x}) = \sum_{j=1}^{N} c_j \|\boldsymbol{x} - \boldsymbol{x}_j\|_2$$

◀

The basis functions in Example 2.2 are examples of:

Definition 2.3 (Radial basis function)**.** A function $K : \mathbb{R}^d \to \mathbb{R}$ is called radial, iff there exists a univariate function $k : [0, \infty) \to \mathbb{R}$ such that $K(\boldsymbol{x}) = k(r)$, where $r := \|\boldsymbol{x}\|$ and $\|\cdot\|$ is any norm on $\mathbb{R}^d$. We say that $\varphi_j(\boldsymbol{x}) = K(\boldsymbol{x} - \boldsymbol{x}_j)$, $j = 1, \dots, N$, are RBFs. ◀

Definition 2.4. A function $K : \mathbb{R}^d \to \mathbb{R}$ is called positive semi-definite, iff the generalized Vandermonde matrix $\underline{\boldsymbol{A}} = [\varphi_j(\boldsymbol{x}_i)]_{i,j=1}^N = [K(\boldsymbol{x}_i - \boldsymbol{x}_j)]_{i,j=1}^N$ is symmetric positive semi-definite any mutually distinct $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$ and any $N \in \mathbb{N}$. It is called positive definite, iff $\underline{\boldsymbol{A}}$ is symmetric positive definite. ◀

Example 2.3 (Matérn kernels)**.** The Matérn kernels

$$k_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\ell} \right)^\nu B_\nu \left(\frac{\sqrt{2\nu} r}{\ell} \right), \quad \nu, \ell > 0$$

where B_ν is the modified Bessel function of the second kind of order ν, are positive definite. In particular, $k_{1/2}(r) = \exp(-r/\ell)$ and $k_\infty(r) = \exp(-r^2/(2\ell^2))$. ◀

Fact 2.2.

(1) If $K_1, \dots, K_n$ are positive semi-definite and $c_l \geq 0$, $l = 1, \dots, n$, then $K = \sum_{l=1}^n c_l K_l$ is also positive semi-definite. If at least one K_l is positive definite and $c_l > 0$, then K is positive definite.

(2) If K is positive semi-definite, then $K(\boldsymbol{0}) \geq 0$.

(3) If K is positive semi-definite, then $K(\boldsymbol{x}) = K(-\boldsymbol{x})$.

(4) Any positive semi-definite function is bounded, i.e., $|K(\boldsymbol{x})| \leq K(\boldsymbol{0})$.

(5) If K is positive semi-definite with $K(\boldsymbol{0}) = 0$, then $K \equiv 0$.

(6) The product of positive (semi-)definite functions is positive (semi-)definite. ◁

For the corresponding Vandermonde matrix $\underline{\boldsymbol{A}} = [K(\boldsymbol{x}_i - \boldsymbol{x}_j)]_{i,j=1}^N$, Fact 2.2 can be interpreted as follows. **(1)** states that if $\underline{\boldsymbol{A}}_1, \ldots, \underline{\boldsymbol{A}}_n$ are symmetric positive semi-definite and $c_l \geq 0$, then $\underline{\boldsymbol{A}} = \sum_{l=1}^n c_l \underline{\boldsymbol{A}}_l$ is symmetric positive semi-definite. If at least one $\underline{\boldsymbol{A}}_l$ is symmetric positive definite and $c_l > 0$, then $\underline{\boldsymbol{A}}$ is symmetric positive definite. **(2)** means that the diagonal entries satisfy $A_{kk} \geq 0$ for all k. **(3)** means that $\underline{\boldsymbol{A}}$ is symmetric, i.e., $A_{ij} = A_{ji}$. **(4)** means that $|A_{ij}| \leq A_{kk}$ for all i, j, k. **(5)** means that $\underline{\boldsymbol{A}} = \underline{\boldsymbol{0}}$ if $A_{kk} = 0$. **(6)** corresponds to the fact that the entrywise product of symmetric positive (semi-)definite matrices is symmetric positive (semi-)definite.

3 Scattered data interpolation with polynomial precision

Often it is desirable that an approximation can represent polynomials exactly. As we have seen in Theorem 2.1, in this case the interpolation points need to be chosen carefully.

3.1 Unisolvency

We define $\boldsymbol{x}^{\boldsymbol{\alpha}} := x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ for $\boldsymbol{x} \in \mathbb{R}^d$ and a *multi-index* $\boldsymbol{\alpha} \in \mathbb{N}_0^d$.

Definition 3.1 (q-unisolvent)**.** We call a set $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \mathbb{R}^d$ q-unisolvent, iff the only polynomial $p \in \Pi_q^d := \operatorname{span}\{\boldsymbol{x}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{N}_0^d, \|\boldsymbol{\alpha}\|_1 \leq q\}$ interpolating zero data on X is $p \equiv 0$. This means that the matrix

$$\underline{\boldsymbol{P}} := \left[\boldsymbol{x}_i^{\boldsymbol{\alpha}}\right]_{\substack{i=1,\ldots,N \\ \|\boldsymbol{\alpha}\|_1 \leq q}} \in \mathbb{R}^{N \times m_q} \tag{3.1}$$

has full column rank $m_q := \dim \Pi_q^d = \binom{q+d}{d}$. ◀

Remark 3.1. For $d = 1$, a set $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}$ is q-unisolvent iff $N \geq m_q = q + 1$ and the points are pairwise distinct (a nonzero univariate polynomial of degree $\leq q$ has at most q distinct roots).

For general $d \geq 1$, the condition $N \geq m_q = \dim \Pi_q^d$ is necessary: if $N < m_q$, then $\underline{\boldsymbol{P}} \in \mathbb{R}^{N \times m_q}$ cannot have full column rank, hence there exists a nonzero $p \in \Pi_q^d$ such that $p(\boldsymbol{x}_i) = 0$ for all $\boldsymbol{x}_i \in X$.

In contrast to the one-dimensional case, for $d > 1$ the condition $N \geq m_q$ is not sufficient. The reason is that a nonzero multivariate polynomial may vanish on infinitely many points (e.g. along an algebraic curve or surface). Concretely, if there exists a nonzero $p \in \Pi_q^d$ with $X \subset Z(p) := \{\boldsymbol{x} \in \mathbb{R}^d : p(\boldsymbol{x}) = 0\}$, then X is not q-unisolvent. ◀

3.2 Interpolation with polynomial precision

Taking polynomials into account for the approximation gives rise to a specific version of Problem 2.1:

$$s_f(\boldsymbol{x}) = \sum_{j=1}^{N} c_j \varphi_j(\boldsymbol{x}) + \sum_{k=1}^{m_q} d_k p_k(\boldsymbol{x}) \tag{3.2}$$

for a basis $\{p_1, \ldots, p_{m_q}\}$ of Π_q^d and $q \geq 0$.

Enforcing the interpolation conditions $s_f(\boldsymbol{x}_i) = y_i$ for $i = 1, \ldots, N$ leads to a linear system of N equations for $N + m_q$ unknowns. To determine the remaining m_q coefficients, we add the additional conditions

$$\sum_{j=1}^{N} c_j p_k(\boldsymbol{x}_j) = 0 \quad \text{for } k = 1, \ldots, m_q \tag{3.3}$$

Remark 3.2. The constraints (3.3) prevent a non-unique decomposition between the RBF part and the polynomial part of the ansatz (3.2). Without (3.3), one can often trade off polynomial pieces between the two parts in (3.2) and still hit the same data values (non-uniqueness). (3.3) forces the RBF coefficient vector $\boldsymbol{c}$ to be orthogonal (in the discrete sense) to all polynomials in Π_q^d, so the polynomial component is carried purely by $\boldsymbol{d}$. ◀

Introducing the matrices

$$\underline{\boldsymbol{A}} := \left[\varphi_j(\boldsymbol{x}_i)\right]_{i,j=1}^{N} \in \mathbb{R}^{N \times N} \quad \text{and} \quad \underline{\boldsymbol{P}} := \left[p_j(\boldsymbol{x}_i)\right]_{\substack{i=1,\ldots,N \\ j=1,\ldots,m_q}} \in \mathbb{R}^{N \times m_q}$$

as well as the vectors $\boldsymbol{c} := [c_1, \ldots, c_N]^\top$, $\boldsymbol{d} := [d_1, \ldots, d_{m_q}]^\top$ yields the saddle point system

$$\begin{bmatrix} \underline{\boldsymbol{A}} & \underline{\boldsymbol{P}} \\ \underline{\boldsymbol{P}}^\top & \underline{\boldsymbol{0}} \end{bmatrix} \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \tag{3.4}$$

Since we now solve the augmented system (3.4) instead of (2.3), we don't need K to be positive definite anymore - it's enough for K to be conditionally positive definite of order $(q + 1)$ (Definition 3.2), as we will see in Theorem 3.2.

3.3 Conditionally positive definite functions

Definition 3.2. A function $K : \mathbb{R}^d \to \mathbb{R}$ is called conditionally positive definite of order $(q+1)$, iff for any mutually distinct points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$ and any $N \in \mathbb{N}$ the generalized Vandermonde matrix $\underline{\boldsymbol{A}} = [K(\boldsymbol{x}_i - \boldsymbol{x}_j)]_{i,j=1}^N$ satisfies

$$\boldsymbol{c}^\top \underline{\boldsymbol{A}} \boldsymbol{c} \geq 0 \quad \text{for any } \boldsymbol{c} \in \operatorname{kern}(\underline{\boldsymbol{P}}^\top) \tag{3.5}$$

where $\underline{\boldsymbol{P}} = [\boldsymbol{x}_i^{\boldsymbol{\alpha}}]_{i=1,\dots,N,\, \|\boldsymbol{\alpha}\|_1 \leq q}$ is the polynomial Vandermonde matrix from Definition 3.1, Equation 3.1. It is called strictly conditionally positive definite, iff equality in (3.5) only holds for $\boldsymbol{c} = \boldsymbol{0}$. ◀

We have the following relation between conditionally positive definite functions.

Fact 3.1. A function that is (strictly) conditionally positive definite of order $(q+1)$ is also (strictly) conditionally positive definite of any higher order. In particular, a function that is (strictly) conditionally positive definite of order 1 is (strictly) conditionally positive definite of any order. ◁

Proof. Let $r \geq q$. Since $\Pi_q^d \subseteq \Pi_r^d$, the columns of $\underline{\boldsymbol{P}}_q$ form a subset of the columns of $\underline{\boldsymbol{P}}_r$. So a vector that is orthogonal to all columns of $\underline{\boldsymbol{P}}_r$ is also orthogonal to all columns of $\underline{\boldsymbol{P}}_q$. Consequently, $\operatorname{kern}(\underline{\boldsymbol{P}}_r^\top) \subseteq \operatorname{kern}(\underline{\boldsymbol{P}}_q^\top)$. □

3.4 Well-posedness of the saddle point system

Definition 3.1 and Definition 3.2 now allow us to state the conditions under which the saddle point system (3.4) is well-posed:

Theorem 3.2. Let $K : \mathbb{R}^d \to \mathbb{R}$ be strictly conditionally positive definite of order $(q+1)$ and let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ be q-unisolvent. Then, the linear system (3.4) is uniquely solvable. ◁

Proof. To prove the assertion, we show that the kernel of the matrix in (3.4) consists of only the zero vector $\boldsymbol{0}_{N+m_q}$:

$$\begin{bmatrix} \underline{\boldsymbol{A}} & \underline{\boldsymbol{P}} \\ \underline{\boldsymbol{P}}^\top & \underline{\boldsymbol{0}} \end{bmatrix} \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0}_N \\ \boldsymbol{0}_{m_q} \end{bmatrix} \Leftrightarrow \begin{cases} \underline{\boldsymbol{A}}\boldsymbol{c} + \underline{\boldsymbol{P}}\boldsymbol{d} = \boldsymbol{0}_N \\ \underline{\boldsymbol{P}}^\top \boldsymbol{c} = \boldsymbol{0}_{m_q} \end{cases}$$

To find $\boldsymbol{c}$, we multiply the top equation by $\boldsymbol{c}^\top$ from the left, yielding

$$\boldsymbol{c}^\top \underline{\boldsymbol{A}} \boldsymbol{c} + \boldsymbol{c}^\top \underline{\boldsymbol{P}} \boldsymbol{d} = 0$$

From the bottom equation, we have $\underline{\boldsymbol{P}}^\top \boldsymbol{c} = \boldsymbol{0} \Leftrightarrow \boldsymbol{c}^\top \underline{\boldsymbol{P}} = \boldsymbol{0}^\top$. Consequently, we infer $\boldsymbol{c}^\top \underline{\boldsymbol{A}} \boldsymbol{c} = 0$. Since K is strictly conditionally positive definite of order $(q+1)$, this implies $\boldsymbol{c} = \boldsymbol{0}_N$.

To find $\boldsymbol{d}$, we insert $\boldsymbol{c} = \boldsymbol{0}_N$ into the top equation, yielding

$$\underline{\boldsymbol{A}}\boldsymbol{c} + \underline{\boldsymbol{P}}\boldsymbol{d} = \underline{\boldsymbol{P}}\boldsymbol{d} = \boldsymbol{0}_N$$

By the q-unisolvency of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, the matrix $\underline{\boldsymbol{P}}$ has full column rank m_q. Therefore, $\boldsymbol{d} = \boldsymbol{0}_{m_q}$ is the only solution to the top block of (3.4). □

Remark 3.3. Definition 3.2 and Theorem 3.2 are a special, finite-dimensional, instance of the inf-sup- or Ladyzhenskaya-Babuška-Brezzi (LBB) condition, which guarantees the well-posedness of (infinite-dimensional) saddle point problems. ◀

Example 3.4. The generalized multiquadrics $K(\boldsymbol{x}) = (1 + \|\boldsymbol{x}\|^2)^\beta$, $0 < \beta \notin \mathbb{N}$, are strictly conditionally positive definite of order $\lceil \beta \rceil$.

The radial powers $K(\boldsymbol{x}) = \|\boldsymbol{x}\|^\beta$, $0 < \beta \notin 2\mathbb{N}$ are strictly conditionally positive definite of order $\lceil \beta/2 \rceil$. This means that the distance functions from Example 2.2 are conditionally positive definite of order 1.

Duchon's thin plate splines $K(\boldsymbol{x}) = \|\boldsymbol{x}\|^{2\beta} \log(\|\boldsymbol{x}\|)$, $\beta \in \mathbb{N}^*$, are strictly conditionally positive definite of order $\beta + 1$. ◀

4 Functional Analysis

4.1 Norm, Completeness, Inner product

Definition 4.1 (Norm)**.** If V is a vector space over $\mathbb{F}$[1], then a function $\|\cdot\| : V \to \mathbb{R}_{\geq 0}$ is called a norm if

- $\|x\| = 0 \Leftrightarrow x = 0$ (positive definite)
- $\|\lambda x\| = |\lambda|\|x\|$ (homogeneous)
- $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)

for all $x, y \in V$ and $\lambda \in \mathbb{F}$. A vector space on which a norm is defined is called a normed space. ◀

The function spaces used in functional analysis are mostly infinite-dimensional. Therefore, the norm and the notion of convergence alone are not sufficient to obtain powerful results.

A normed space can be completed by adding a limit for every Cauchy sequence that does not converge, until all Cauchy sequences possess a limit. In this way one passes from a normed space X to its *completion* $\overline{X}$. If it is not clear which norm is meant, this is indicated as an exponent:

$$\overline{X}^{\|\cdot\|_X} \tag{4.1}$$

If Y is the completion of X, one says that X is dense in Y.

Theorem 4.1. Let $(X, \|\cdot\|_X)$ be a normed space. There exists a Banach space $(\overline{X}, \|\cdot\|_{\overline{X}})$ called completion of X and an injective mapping $J : X \to \overline{X}$ such that

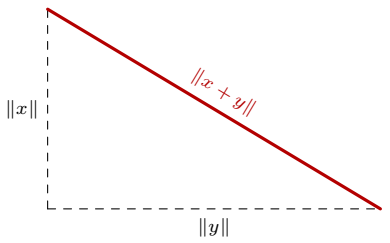
$$J(v + w) = J(v) + J(w), \quad J(\lambda v) = \lambda J(v), \quad \|J(v)\|_{\overline{X}} = \|v\|_X$$

for all $v, w \in X$. This completion is unique up to isometry (rotation). ◁

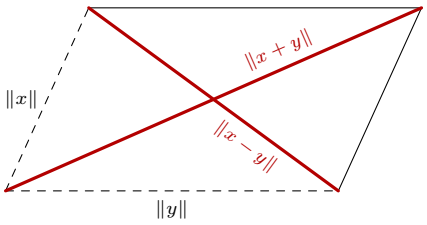
Definition 4.2 (Inner product)**.** If V is a vector space over $\mathbb{F}$ ($\mathbb{R}$ or $\mathbb{C}$), then a function $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{F}$ is called an inner product if

- $\langle x, y\rangle = \overline{\langle y, x\rangle}$ (conjugate symmetry)
- $\langle \lambda x + y, z\rangle = \lambda\langle x, z\rangle + \langle y, z\rangle$ (linear in first argument)
- $\langle x, x\rangle \geq 0$ with equality if and only if $x = 0$ (positive definite)

for all $x, y, z \in V$ and $\lambda \in \mathbb{F}$: A vector space on which an inner product is defined is called an inner product space. ◀



Pythagorean formula



Parallelgram law

Theorem 4.2 (Parallelgram law)**.** For any x, y in an inner product space, we have

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2 \tag{4.2}$$

◁

[1] henceforth, $\mathbb{F}$ is short for $\mathbb{R}$ or $\mathbb{C}$

Theorem 4.3 (Pythagorean formula)**.** For any x, y in an inner product space with $x \perp y$ (i.e. they are orthogonal, i.e. $\langle x, y\rangle = 0$), we have

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 \tag{4.3}$$

◁

Theorem 4.4 (Cauchy-Schwarz inequality)**.** For any x, y in an inner product space, we have

$$\boxed{|\langle x, y\rangle| \leq \|x\| \cdot \|y\|} \tag{4.4}$$

with equality if and only if x and y are linearly dependent. ◁

Definition 4.3 (Hilbert space)**.** An inner product space that is complete with respect to the norm induced by the inner product is called a Hilbert space. ◀

4.2 Bases in Hilbert spaces

In linear algebra the concept of a (Hamel) basis is central. For a Hamel basis of a vector space, every vector can be represented as a *finite* linear combination of basis vectors. Since linear algebra mostly deals with finite-dimensional spaces, this is no restriction there.

The spaces of interest in functional analysis, however, are all *infinite*-dimensional. While the general statement still holds that such a space has a Hamel basis (that is, a linearly independent, spanning (i.e. every vector can be represented as a finite linear combination) subset), every such Hamel basis has *infinitely* many elements. Moreover: Those spaces that possess a *countable* Hamel basis are not complete (e.g. the space of trigonometric polynomials). One can show that in an infinite-dimensional Banach space a Hamel basis is always uncountable. This disqualifies the concept of a Hamel basis for functional analysis.

The idea of a Schauder basis is to drop the requirement that the representation of a vector uses finitely many elements. Instead we represent the elements as series whose partial sums are linear combinations of the basis vectors.

Definition 4.4 (Orthonormal basis)**.** A subset B of a Hilbert space $\mathcal{H}$ is called an orthogonal system if any two distinct elements of B are orthogonal, i.e., $\langle x, y\rangle = 0$ if $x \neq y$.

If there exists in a Hilbert space $\mathcal{H}$ an orthogonal system

$$B = \{e_k \in \mathcal{H} : k \in \mathbb{N}\}$$

such that for every $f \in \mathcal{H}$ the representation

$$f = \sum_{k=1}^{\infty} \langle f, e_k\rangle e_k$$

holds, then B is called an orthonormal basis of $\mathcal{H}$. ◀

In Hilbert spaces orthonormal bases replace the standard bases from linear algebra. That the elements of a Orthonormal basis actually have norm 1 can be checked easily:

$$e_n = \sum_{k=1}^{\infty} \langle e_n, e_k\rangle e_k = \langle e_n, e_n\rangle e_n = \|e_n\|^2 e_n \quad \Longrightarrow \quad \|e_n\| = 1$$

Thus we have:

$$\langle e_k, e_n\rangle = \delta_{kn} = \begin{cases} 1, & k = n \\ 0, & k \neq n \end{cases} \tag{4.5}$$

Theorem 4.5. The representation of a vector f in a Hilbert space with respect to an orthonormal basis is unique. ◁

Proof. Consider the representation

$$f = \sum_{k=1}^{\infty} c_k e_k$$

for some coefficients (c_k). Fix some index $n \in \mathbb{N}$. Then for every $N \geq n$ we have

$$c_n = \sum_{k=1}^{N} c_k \delta_{kn} = \sum_{k=1}^{N} c_k \langle e_k, e_n \rangle = \left\langle \sum_{k=1}^{N} c_k e_k, e_n \right\rangle$$

Letting N tend to infinity yields

$$c_n = \langle f, e_n \rangle$$

where we used the continuity of the inner product in its first argument[2] □

This uniqueness of the representation means that we have the technique of comparing coefficients at our disposal.

Example 4.1 (Fourier series)**.** The prototype of such an orthonormal basis is the set

$$\varphi_k = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}kx}, \quad k \in \mathbb{Z},\ x \in (-\pi, \pi)$$

in the Hilbert space $L^2(-\pi, \pi)$. Here the series representation

$$f = \sum_{k=-\infty}^{\infty} \langle f, \varphi_k \rangle \varphi_k$$

is the Fourier series expansion of f, where $\langle f, g \rangle = \int_{-\pi}^{\pi} f(x)\overline{g(x)}\,\mathrm{d}x$. ◀

With arguments similar to the uniqueness of the representation above one proves Parseval's identity

$$\boxed{\|f\|^2 = \sum_{k=1}^{\infty} |\langle f, e_k \rangle|^2} \tag{4.6}$$

Remark 4.2. In Definition 4.4 we implicitly built in that the orthonormal basis is countable. This is not necessary, but the spaces with countable orthonormal bases, called separable spaces, are the most important in practice. ◀

4.3 Operators

4.3.1 Linear bounded operators are continuous

Definition 4.5 (Boundedness, continuity)**.** Let U and V be normed spaces. An operator $\mathcal{A} : U \to V$ is called *bounded* if there exists a constant $M_{\mathcal{A}} > 0$ such that

$$\|\mathcal{A}u\|_V \leq M_{\mathcal{A}} \|u\|_U$$

for all $u \in U$. An operator $\mathcal{A} : U \to V$ is called *continuous* if

$$\lim_{n\to\infty} (\mathcal{A}u_n) = \mathcal{A}\left(\lim_{n\to\infty} u_n\right)$$

for every sequence (u_n).

The set of linear bounded operators from U to V is denoted by $B(U, V)$. ◀

A linear, bounded operator is continuous:

$$\|\mathcal{A}u_n - \mathcal{A}u\|_V \overset{\text{linear}}{=} \|\mathcal{A}(u_n - u)\|_V \overset{\text{bounded}}{\leq} M_{\mathcal{A}} \|u_n - u\|_U$$

where (u_n) is a sequence in U converging to u. The converse also holds: A linear continuous operator is bounded.

The boundedness of an operator depends essentially on the underlying space and the norm used (see Example 4.4).

[2] i.e. $\lim_{m\to\infty} \langle f_m, g \rangle = \langle \lim_{m\to\infty} f_m, g \rangle$

Caution 4.3. A bounded operator is something completely different from a bounded function. For a bounded function, the image is a bounded set. For a bounded operator, every bounded set is mapped to a bounded set. The image need not be bounded. Among linear bounded operators only the zero operator has a bounded image. ◀

Example 4.4 (Differential operator)**.** Consider the differential operator

$$\frac{\mathrm{d}}{\mathrm{d}t} : C^1([0,1]) \to C([0,1]).$$

We can turn both spaces into normed spaces by equipping them with the maximum norm, i.e.

$$\|f\|_\infty := \max_{t\in[0,1]} |f(t)|.$$

In this case the differential operator is not bounded, as can be seen from the sequence (f_n) with $f_n(t) = \cos(nt)$. We have $\|f_n\|_\infty = 1$ for all n, but $\|f_n'\|_\infty = \max_{t\in[0,1]} |n\sin(nt)|$, which tends to infinity as $n \to \infty$.

If, however, we equip $C^1([0,1])$ with the norm

$$\|f\|_{C^1} := \|f\|_\infty + \|f'\|_\infty$$

then for every $f \in C^1([0,1])$ we trivially have

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t} f \right\|_\infty = \|f'\|_\infty \le \|f\|_{C^1}.$$

Thus the differential operator is bounded with $M_{\frac{\mathrm{d}}{\mathrm{d}t}} = 1$. ◀

4.3.2 Operator norm

The set $B(U,V)$ is itself again a vector space and we can also equip it with a norm:

Definition 4.6 (Operator norm)**.** For a linear bounded operator $\mathcal{A} \in B(U,V)$ we define

$$\|\mathcal{A}\| := \sup_{u\in U\setminus\{0\}} \frac{\|\mathcal{A}u\|_V}{\|u\|_U} = \sup_{\|u\|_U=1} \|\mathcal{A}u\|_V \tag{4.7}$$

◀

If V is a Banach space, then $B(U,V)$ is also a Banach space. The Operator norm has all the properties of a Norm. In addition, for the composition $\mathcal{B}\mathcal{A}$ of two linear bounded operators $\mathcal{A} : U \to V$ and $\mathcal{B} : V \to W$ the estimate

$$\|\mathcal{B}\mathcal{A}\| \le \|\mathcal{B}\|\,\|\mathcal{A}\|$$

holds.

4.3.3 Bounded inverse

Complete spaces have, among other things, the advantage that they ensure well-behaved properties of linear bounded operators. Consider the situation where we want to solve an operator equation

$$\mathcal{A}x = y \tag{4.8}$$

If the operator $\mathcal{A}$ is injective, then this equation has for every y from the image of $\mathcal{A}$ a unique solution, i.e. the operator $\mathcal{A}$ possesses an inverse $\mathcal{A}^{-1}$. For the numerical solution of the problem it would, however, be useful if this inverse were also continuous, because then small errors in the right-hand side (the data) would change the solution of the equation only slightly.

Theorem 4.6 (Bounded inverse)**.** If U and V are Banach spaces and $\mathcal{A} : U \to V$ a bijective linear bounded operator, then the inverse $\mathcal{A}^{-1} : V \to U$ is also bounded (and hence continuous). ◁

Unfortunately, the assumptions in Theorem 4.6 are often not fulfilled in practice. For compact operators, the image is never a complete space and hence their inverse is always unbounded.

4.3.4 Continuous extension of an operator

We consider a linear bounded operator $\mathcal{A} \in B(U, V)$, where U and V are normed spaces. As already mentioned, many important results require complete spaces (i.e. Banach spaces).

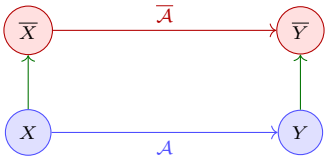
If U and V are not complete, they can be enlarged to their completions (4.1). Consider a Cauchy sequence (u_n) from U. Then

$$\|\mathcal{A}u_n - \mathcal{A}u_m\| = \|\mathcal{A}(u_n - u_m)\| \leq \|\mathcal{A}\| \, \|u_n - u_m\|$$

Thus $(\mathcal{A}u_n)$ is a Cauchy sequence in V. The Cauchy sequence (u_n) has a limit in $u \in \overline{U}$, the Cauchy sequence $(\mathcal{A}u_n)$ a limit $v \in \overline{V}$. We set

$$\overline{\mathcal{A}}u = \mathcal{A}\left(\lim_{n\to\infty} u_n\right) = \lim_{n\to\infty} \mathcal{A}u_n = v$$

and obtain an operator $\overline{\mathcal{A}} : \overline{U} \to \overline{V}$ with $\overline{\mathcal{A}}u = \mathcal{A}u$ for all $u \in U$. In other words: We extend $\mathcal{A}$ continuously from U to $\overline{U}$. This extension is unique and the norm of the operator does not change, i.e. $\|\overline{\mathcal{A}}\| = \|\mathcal{A}\|$.



Usually, however, one uses the same symbol for the continuous extension as for the original operator, i.e. one writes $\mathcal{A}$ again for $\overline{\mathcal{A}}$.

4.3.5 Neumann series

Theorem 4.7 (Perturbation lemma)**.** If V is a Banach space and $\mathcal{A} : V \to V$ a linear bounded operator with $\|\mathcal{A}\| < 1$, then

$$(\mathrm{id}_V - \mathcal{A})^{-1} = \sum_{k=0}^{\infty} \mathcal{A}^k$$

where the series converges in the operator norm. It holds that

$$\left\|(\mathrm{id}_V - \mathcal{A})^{-1}\right\| \leq \frac{1}{1 - \|\mathcal{A}\|}$$

◁

4.4 Functionals and distributions

Definition 4.7 (Functional)**.** If V is a vector space over $\mathbb{F}$, then a mapping $\varphi : V \to \mathbb{F}$ is called a functional. If φ is linear, it is called a linear functional or also a linear form. ◀

Example 4.5. When solving a linear operator equation of the form (4.8) one can proceed by minimizing the norm of the residual, or equivalently its square

$$\|\mathcal{A}x - y\|^2$$

This is a nonlinear functional that has the minimum 0 for a possible solution.

Sometimes minimizing only the residual leads to instabilities. One possibility for improvement is the introduction of an additional penalty term

$$\|\mathcal{A}x - y\|^2 + \lambda\|x\|^2$$

with a suitably chosen constant $\lambda > 0$ (regularization). ◀

4.4.1 Distributions

can be regarded as a generalization of functions.

We denote by $C_0^\infty(\mathbb{R})$ the space of infinitely differentiable functions with compact support (i.e. there exists a compact interval $I \subset \mathbb{R}$ such that $f(x) = 0$ for all $x \notin I$). The space $C_0^\infty(\mathbb{R})$ is also called the space of test functions.

A sequence (φ_k) of test functions converges to $\varphi \in C_0^\infty(\mathbb{R})$ if there exists a compact interval $I \subset \mathbb{R}$ such that for every order of derivative $n \in \mathbb{N}_0$

$$\|\varphi_k^{(n)} - \varphi^{(n)}\|_{I,\infty} \xrightarrow{k\to\infty} 0$$

and $\varphi_k(x) = 0$ for all $x \notin I$. On the compact interval I every sequence of derivatives of the (φ_k) therefore converges uniformly to the corresponding derivative of φ. Outside of I all these functions vanish.

We cannot describe this notion of convergence by means of a norm.[3] However, we can consider functionals $\psi : C_0^\infty(\mathbb{R}) \to \mathbb{C}$ that are continuous with respect to this notion of convergence. From $\varphi_k \to \varphi$ we therefore require $\psi(\varphi_k) \to \psi(\varphi)$. The set D of all such continuous functionals is called the space of distributions.

Example 4.6 (δ-distribution)**.** A very simple distribution is the δ-distribution defined by

$$\delta(\varphi) = \varphi(0), \quad \varphi \in C_0^\infty(\mathbb{R}).$$

If (φ_k) is a sequence of test functions with $\varphi_k \to \varphi \in C_0^\infty(\mathbb{R})$, then there is uniform convergence on every compact interval that contains zero. Hence

$$\lim_{k\to\infty} \delta(\varphi_k) = \lim_{k\to\infty} \varphi_k(0) = \varphi(0) = \delta(\varphi).$$

This shows that δ is a continuous functional on the space of test functions. ◀

4.4.2 Regular distributions

Definition 4.8 (Regular distribution)**.** A function that is integrable over every compact interval is called locally integrable. For every locally integrable function $f : \mathbb{R} \to \mathbb{C}$ we obtain a corresponding distribution by

$$\psi_f(\varphi) := \int_{-\infty}^{\infty} f(x)\varphi(x)\,\mathrm{d}x, \quad \varphi \in C_0^\infty(\mathbb{R})$$

We also write

$$\psi_f(\varphi) = \langle f, \varphi \rangle, \quad \varphi \in C_0^\infty(\mathbb{R})$$

Two different locally integrable functions yield different distributions. Distributions that can be represented in this way are called regular distributions. ◀

Example 4.7 (Heaviside function)**.** The function $H : \mathbb{R} \to \mathbb{R}$,

$$H(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geq 0, \end{cases}$$

is also called the Heaviside function. Since it is piecewise continuous, it is integrable over every compact interval $I \subseteq \mathbb{R}$. Thus we can define a distribution ψ_H by

$$\psi_H(\varphi) = \int_{-\infty}^{\infty} H(x)\varphi(x)\,\mathrm{d}x, \quad \varphi \in C_0^\infty(\mathbb{R})$$

The integral exists because the domain of integration is in fact only the compact interval outside of which φ vanishes, and the integrand is piecewise continuous there. The continuity of ψ_H with respect to the convergence in $C_0^\infty(\mathbb{R})$ can be shown using the Lebesgue dominated convergence theorem. ◀

By identifying the locally integrable functions with the regular distributions they generate, we obtain an embedding of locally integrable functions into the distributions. Since all classical function spaces contain only locally integrable functions, we thereby recover, for example, the continuous or the continuously differentiable functions in the distributions.

[3] $C_0^\infty(\mathbb{R})$ is therefore not a normed space, and in particular not a Banach space or an inner product space. The problem is that we consider infinitely many derivatives.

Caution 4.8. Not every distribution is regular! ◀

The δ-distribution is the classical example of a distribution that cannot be represented by a locally integrable function. Nevertheless, especially among practitioners, the notations

$$\delta(\varphi) = \langle \delta, \varphi \rangle = \int_{-\infty}^{\infty} \delta(x)\varphi(x)\,\mathrm{d}x = \varphi(0)$$

are common. Note, however, that in the penultimate notation this is by no means an integral in the Lebesgue sense, but merely the evaluation of the functional δ at the point φ!

The integral notation has yet another background, and this is closely linked to the fact that the δ-distribution is suitable for modeling instantaneous impulse transfer.

4.4.3 Derivative of distributions

The remarkable fact about distributions is that we can generalize many properties of classical functions and transfer them to distributions.

For a continuously differentiable function f we consider the regular distribution generated by f'. We apply this distribution to a test function φ that vanishes outside an interval $I = (a, b)$. By integration by parts we obtain

$$\begin{aligned}\langle f', \varphi \rangle &= \int_I f'(x)\varphi(x)\,\mathrm{d}x \\ &= [f(x)\varphi(x)]_a^b - \int_I f(x)\varphi'(x)\,\mathrm{d}x \\ &= -\int_{-\infty}^{\infty} f(x)\varphi'(x)\,\mathrm{d}x = -\langle f, \varphi' \rangle\end{aligned}$$

since $\varphi(a) = \varphi(b) = 0$. Note that the notation in the last line is justified, because with φ the derivative φ' is again a test function.

Caution 4.9. An antiderivative of a test function is in general no longer a test function. ◀

For every continuously differentiable function f the equation

$$\langle f', \varphi \rangle = -\langle f, \varphi' \rangle, \quad \varphi \in C_0^\infty(\mathbb{R}),$$

holds, which establishes a relation between the distribution given by f' and that given by f. Conversely, this relation characterizes function and derivative.

With this equation we can generalize the concept of the derivative: We use it to define a derivative in the distributional sense for an *arbitrary distribution*.

Definition 4.9 (Distributional derivative)**.** If d is a distribution, then its derivative d' is defined by

$$\langle d', \varphi \rangle := -\langle d, \varphi' \rangle, \quad \varphi \in C_0^\infty(\mathbb{R})$$ ◀

Example 4.10.

- The Heaviside function H is not continuously differentiable. However, we can determine its derivative in the distributional sense. For $\varphi \in C_0^\infty(\mathbb{R})$ we have

$$\begin{aligned}\langle H', \varphi \rangle &= -\langle H, \varphi' \rangle = -\int_{-\infty}^{\infty} H(x)\varphi'(x)\,\mathrm{d}x \\ &= -\int_0^{\infty} \varphi'(x)\,\mathrm{d}x = -[\varphi(x)]_0^{\infty} = \varphi(0) = \langle \delta, \varphi \rangle\end{aligned}$$

  Thus the derivative of the Heaviside function is the δ-distribution.

- The derivative of the δ-distribution can be written down immediately. For $\varphi \in C_0^\infty(\mathbb{R})$ we have

$$\langle \delta', \varphi \rangle = -\langle \delta, \varphi' \rangle = -\varphi'(0)$$ ◀

There are no restrictions in the definition of the distributional derivative. Every distribution is differentiable in the distributional sense arbitrarily many times.

4.5 Operators in Hilbert spaces

Due to the inner product, in Hilbert spaces it is possible to adopt many statements and arguments from analytic geometry. However, because of the infinite dimension of function spaces there are also differences in the chains of argument.

4.5.1 Orthogonal projection

Let $\mathcal{H}$ be an arbitrary Hilbert space and $U \subset \mathcal{H}$ a closed subspace. U is therefore itself again a Hilbert space. We now want to know whether for an arbitrary $x \in \mathcal{H}$ there exists a $\hat{u} \in U$ that has minimal distance to x, i.e.

$$\|x - \hat{u}\| \leq \|x - u\| \quad \forall u \in U \tag{4.9}$$

We now show that such a $\hat{u}$ exists and is unique.

To this end we construct two sequences recursively. We set $a_0 = 0$ and $b_0 = \|x - u_0\|$ for some $u_0 \in U$. We define

$$c_n = \frac{a_{n-1} + b_{n-1}}{2}$$

and check whether $c_n \leq \|x - u\|$ holds for all $u \in U$. If yes, we set $a_n = c_n$ and $b_n = b_{n-1}$, otherwise $a_n = a_{n-1}$ and $b_n = c_n$.

(a_n) is monotonically increasing and bounded above by b_0, (b_n) is monotonically decreasing and bounded below by $a_0 = 0$. By the monotone convergence criterion both sequences are convergent, and with a proof by contradiction one can also show that they converge to the same limit $\rho \geq 0$, which we call the distance of x to U.

Analogous to the construction of (b_n) we also obtain a sequence (u_n) from U with $\|x - u_n\| \to \rho$. The parallelogram identity now yields

$$\begin{aligned}
\|u_n - u_m\|^2 &= \|u_n - x - (u_m - x)\|^2 \\
&= 2\|u_n - x\|^2 + 2\|u_m - x\|^2 - 4\left\|\frac{u_n + u_m}{2} - x\right\|^2 \\
&\leq 2\|u_n - x\|^2 + 2\|u_m - x\|^2 - 4\rho^2
\end{aligned}$$

where the right-hand side converges to 0 as $n, m \to \infty$. Thus (u_n) is Cauchy and hence has a limit $\hat{u} \in U$. This proves the existence of a best approximation (4.9), because by our construction we have

$$\|x - \hat{u}\| = \rho \leq \|x - u\| \tag{4.10}$$

for all $u \in U$.

Moreover, $x - \hat{u}$ is orthogonal to U. Indeed, choose an arbitrary $v \in U$ and set in (4.10) $u = \hat{u} + \alpha v \in U$, then

$$\begin{aligned}
\|x - \hat{u}\|^2 &\leq \|x - \hat{u} - \alpha v\|^2 \\
&= \|x - \hat{u}\|^2 - 2\,\mathrm{Re}(\overline{\alpha}\langle x - \hat{u}, v\rangle) + |\alpha|^2\|v\|^2
\end{aligned} \tag{4.11}$$

where the right-hand side is a quadratic function in $\alpha \in \mathbb{F}$ that attains its minimum at

$$\alpha = \frac{\langle x - \hat{u}, v\rangle}{\|v\|^2}$$

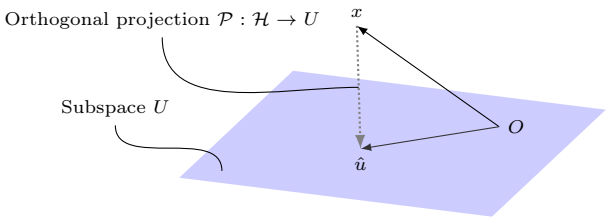
(4.11) must hold for all $\alpha \in \mathbb{F}$ and in particular also for this minimizer. Inserting this choice of α into (4.11) we obtain

$$\|x - \hat{u}\|^2 \leq \|x - \hat{u}\|^2 - \frac{|\langle x - \hat{u}, v\rangle|^2}{\|v\|^2}$$

which is only possible if the numerator is zero. Since $v \in U$ was arbitrary, it follows that

$$\langle x - \hat{u}, v\rangle = 0 \tag{4.12}$$

for all $v \in U$. The best approximation $\hat{u}$ is also unique, as can be shown using the Pythagorean theorem.



Thus, by the assignment $x \mapsto \hat{u}$ we have determined a mapping $\mathcal{P} : \mathcal{H} \to U$.

Theorem 4.8 (Orthogonal projection)**.** If U is a closed subspace of a Hilbert space $\mathcal{H}$, then there exists a $\mathcal{P} : \mathcal{H} \to U$ with the property

$$\|x - \mathcal{P}x\| \leq \|x - u\|$$

for all $u \in U$. The operator $\mathcal{P}$ is linear and bounded with $\|\mathcal{P}\| = 1$. Moreover,

$$\langle x - \mathcal{P}x, u \rangle = 0$$

for all $u \in U$. ◁

From Theorem 4.8 it follows that one can carry out an orthogonal decomposition in Hilbert spaces. If U is a closed subspace of $\mathcal{H}$, one defines the orthogonal complement

$$U^{\perp} := \{v \in \mathcal{H} : \langle v, u \rangle = 0 \quad \forall u \in U\}$$

which is likewise a closed subspace of $\mathcal{H}$. For $x \in \mathcal{H}$, for example, $x - \mathcal{P}x \in U^{\perp}$. Thus every x can be written uniquely as

$$x = u + v$$

with $u \in U$ and $v \in U^{\perp}$. Here $u = \mathcal{P}x$ and $v = x - \mathcal{P}x$.

4.5.2 Continuous linear forms

Theorem 4.9 (Riesz representation theorem)**.** In a Hilbert space $\mathcal{H}$ for every continuous linear form $\varphi : \mathcal{H} \to \mathbb{F}$ there exists exactly one $z \in \mathcal{H}$ such that

$$\varphi(f) = \langle f, z \rangle \tag{4.13}$$

for all $f \in \mathcal{H}$. Conversely, for every $z \in \mathcal{H}$ a continuous linear form is given by $\langle \cdot, z \rangle$.

In more operator-theoretic terms: There is an isometric isomorphism

$$J : \mathcal{H}' \to \mathcal{H} \tag{4.14}$$

such that

$$\boxed{\varphi(f) = \langle f, (J\varphi) \rangle} \tag{4.15}$$

for all $f \in \mathcal{H}$ and all $\varphi \in \mathcal{H}'$, where $\mathcal{H}'$ is the dual space of $\mathcal{H}$, i.e. the space of all continuous linear forms on $\mathcal{H}$. ◁

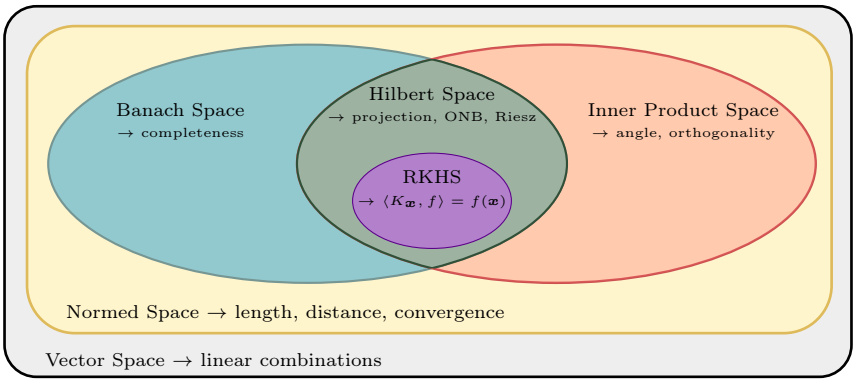
For the proof one uses the fact that the null space N of a continuous linear form φ, $N = \{f \in \mathcal{H} \mid \varphi(f) = 0\}$, forms a closed subspace and its orthogonal complement $N^{\perp}$ has dimension 1. From an element of $N^{\perp}$ the vector z belonging to φ can be constructed.

5 Reproducing kernel Hilbert spaces

Definition 5.1 (RKHS)**.** Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be real Hilbert space of functions $f : \Omega \to \mathbb{R}$. A function $K : \Omega \times \Omega \to \mathbb{R}$ is a reproducing kernel for $\mathcal{H}$, iff

1. $K(\boldsymbol{x}, \cdot) \in \mathcal{H}$ for all $\boldsymbol{x} \in \Omega$ ("kernel sections live in the space")
2. $\langle K(\boldsymbol{x}, \cdot), f\rangle_{\mathcal{H}} = f(\boldsymbol{x})$ for all $f \in \mathcal{H}$ and all $\boldsymbol{x} \in \Omega$ (reproducing property)

If $\mathcal{H}$ exhibits a reproducing kernel, we call it a reproducing kernel Hilbert space. ◀



Theorem 5.1. The reproducing kernel of a RKHS is unique. ◁

Proof. Let K_1 and K_2 be two reproducing kernels of $\mathcal{H}$. By 5.1.2, for all $\boldsymbol{x} \in \Omega$ and $f \in \mathcal{H}$,

$$\langle K_1(\boldsymbol{x}, \cdot), f\rangle_{\mathcal{H}} = f(\boldsymbol{x}) = \langle K_2(\boldsymbol{x}, \cdot), f\rangle_{\mathcal{H}}$$

Hence,

$$\langle K_1(\boldsymbol{x}, \cdot), f\rangle_{\mathcal{H}} - \langle K_2(\boldsymbol{x}, \cdot), f\rangle_{\mathcal{H}} = \langle K_1(\boldsymbol{x}, \cdot) - K_2(\boldsymbol{x}, \cdot), f\rangle_{\mathcal{H}} = 0 \tag{5.1}$$

This means that $\langle K_1(\boldsymbol{x}, \cdot) - K_2(\boldsymbol{x}, \cdot), f\rangle_{\mathcal{H}} = 0$ for all $f \in \mathcal{H}$. Since $K_1(\boldsymbol{x}, \cdot) - K_2(\boldsymbol{x}, \cdot) \in \mathcal{H}$ by 5.1.1, we can set $f = K_1(\boldsymbol{x}, \cdot) - K_2(\boldsymbol{x}, \cdot)$ in (5.1) to obtain

$$\|K_1(\boldsymbol{x}, \cdot) - K_2(\boldsymbol{x}, \cdot)\|_{\mathcal{H}}^2 = 0$$

so $K_1(\boldsymbol{x}, \cdot) - K_2(\boldsymbol{x}, \cdot) = 0_{\mathcal{H}}$ and therefore $K_1(\boldsymbol{x}, \cdot) = K_2(\boldsymbol{x}, \cdot)$ for all $\boldsymbol{x} \in \Omega$. Then, for every $\boldsymbol{x}, \boldsymbol{y} \in \Omega$,

$$K_1(\boldsymbol{x}, \boldsymbol{y}) = \big(K_1(\boldsymbol{x}, \cdot)\big)(\boldsymbol{y}) = \big(K_2(\boldsymbol{x}, \cdot)\big)(\boldsymbol{y}) = K_2(\boldsymbol{x}, \boldsymbol{y})$$

hence $K_1 = K_2$. □

Recall the definition of the point evaluation functional $\delta_{\boldsymbol{x}} : \mathcal{H} \to \mathbb{R}$, $f \mapsto \delta_{\boldsymbol{x}}(f) = f(\boldsymbol{x})$.

The existence of a reproducing kernel is equivalent to the point evaluation functional $\delta_{\boldsymbol{x}}$ being continuous for every $\boldsymbol{x} \in \Omega$, i.e., there exists $M_{\boldsymbol{x}} > 0$ such that

$$|\delta_{\boldsymbol{x}} f| = |f(\boldsymbol{x})| \leq M_{\boldsymbol{x}} \|f\|_{\mathcal{H}} \quad \text{for all } f \in \mathcal{H}$$

This means that $\delta_{\boldsymbol{x}}$ is contained in the dual space $\mathcal{H}'$ of $\mathcal{H}$. By the Riesz representation theorem, there exists $(J\delta_{\boldsymbol{x}}) \in \mathcal{H}$ such that $\langle (J\delta_{\boldsymbol{x}}), f\rangle_{\mathcal{H}} = f(\boldsymbol{x})$ for all $f \in \mathcal{H}$, i.e., $(J\delta_{\boldsymbol{x}})(\boldsymbol{y})$ is the reproducing kernel.

Theorem 5.2 (Properties of Reproducing Kernels)**.** Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be a RKHS. Then

1. $K(\boldsymbol{x}, \boldsymbol{y}) = \langle K(\boldsymbol{y}, \cdot), K(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$
2. $K(\boldsymbol{x}, \boldsymbol{y}) = K(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$
3. Convergence in $\mathcal{H}$ implies pointwise convergence, i.e., if $\|f_n - f\|_{\mathcal{H}} \to 0$ as $n \to \infty$, then $|f_n(\boldsymbol{x}) - f(\boldsymbol{x})| \to 0$ for all $\boldsymbol{x} \in \Omega$. ◁

Proof. Theorem 5.2.1 follows from Definition 5.1 if we set $f = K(\boldsymbol{y}, \cdot)$.

Theorem 5.2.2 follows from Theorem 5.2.1 and the symmetry of the inner product.

Theorem 5.2.3 follows from Definition 5.1.2 and the Cauchy-Schwarz inequality:

$$|f_n(\boldsymbol{x}) - f(\boldsymbol{x})| \overset{5.1.2}{=} |\langle K(\boldsymbol{x}, \cdot), f_n - f\rangle_{\mathcal{H}}| \leq \|K(\boldsymbol{x}, \cdot)\|_{\mathcal{H}} \|f_n - f\|_{\mathcal{H}}$$ □

Remark 5.1. We have $\|K(\boldsymbol{x},\cdot)\|_{\mathcal{H}} = \sqrt{K(\boldsymbol{x},\boldsymbol{x})}$. ◀

The reproducing kernel of an RKHS is positive definite in the sense of Definition 2.4, if we replace $K(\boldsymbol{x}_i - \boldsymbol{x}_j)$ by $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ everywhere.

Definition 5.2. Let $K : \Omega \times \Omega \to \mathbb{R}$ be a kernel. We say that K is positive definite, iff the kernel matrix

$$\underline{\boldsymbol{K}} := [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^N = \begin{bmatrix} K(\boldsymbol{x}_1,\boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_1,\boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ K(\boldsymbol{x}_N,\boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_N,\boldsymbol{x}_N) \end{bmatrix} \tag{5.2}$$

is positive semi-definite for any mutually distinct $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and any $N \in \mathbb{N}$. It is called strictly positive definite if the kernel matrix is positive definite. ◀

The following lemma is a direct consequence of the reproducing property 5.1.2 and characterizes the kernel matrix.

Lemma 5.3. Suppose that $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ is a RKHS with reproducing kernel $K : \Omega \times \Omega \to \mathbb{R}$. Given a set of mutually distinct points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, we have

$$\underline{\boldsymbol{K}} = \langle \boldsymbol{\Phi}, \boldsymbol{\Phi}^\top \rangle_{\mathcal{H}} := \begin{bmatrix} \langle\varphi_1,\varphi_1\rangle_{\mathcal{H}} & \cdots & \langle\varphi_1,\varphi_N\rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle\varphi_N,\varphi_1\rangle_{\mathcal{H}} & \cdots & \langle\varphi_N,\varphi_N\rangle_{\mathcal{H}} \end{bmatrix} \tag{5.3}$$

where $\varphi_i := \Phi(\boldsymbol{x}_i) \in \mathcal{H}$ with $\Phi : \Omega \to \mathcal{H}$, $\boldsymbol{x} \mapsto K(\boldsymbol{x},\cdot)$ is the canonical feature map[4]. The vector $\boldsymbol{\Phi} := [\varphi_1, \ldots, \varphi_N]^\top \in \mathcal{H}^N$ is called the canonical feature vector. ◁

Lemma 5.3 corresponds more or less the kernel trick.

Theorem 5.4. Suppose that $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ is a RKHS with reproducing kernel $K : \Omega \times \Omega \to \mathbb{R}$. Then, K is positive definite. Moreover, K is strictly positive definite iff $\delta_{\boldsymbol{x}_1}, \ldots, \delta_{\boldsymbol{x}_N}$ are linearly independent for any choice of mutually distinct $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \in \Omega$ and any $N \in \mathbb{N}$. ◁

The evaluation functionals $\delta_{\boldsymbol{x}_1}, \ldots, \delta_{\boldsymbol{x}_N}$ are *linearly independent* if no nontrivial linear combination of them produces the zero functional in $\mathcal{H}'$, i.e.,

$$\sum_{i=1}^N c_i \delta_{\boldsymbol{x}_i} = 0_{\mathcal{H}'} \quad \Longrightarrow \quad c_1 = \cdots = c_N = 0$$

where $0_{\mathcal{H}'}$ is the zero vector of the dual space $\mathcal{H}'$, i.e., the functional that maps every vector $f \in \mathcal{H}$ to the scalar $0 \in \mathbb{R}$. So the LHS above means $\sum_{i=1}^N c_i \delta_{\boldsymbol{x}_i}(f) = 0$ for all $f \in \mathcal{H}$.

Proof. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ be mutually distinct points in Ω and let $\boldsymbol{c} \in \mathbb{R}^N$ with $\boldsymbol{c} \neq \boldsymbol{0}$. For the kernel matrix (5.3), we have

$$\boldsymbol{c}^\top \underline{\boldsymbol{K}} \boldsymbol{c} = \boldsymbol{c}^\top \langle \boldsymbol{\Phi}, \boldsymbol{\Phi}^\top \rangle_{\mathcal{H}} \boldsymbol{c} = \langle \boldsymbol{c}^\top \boldsymbol{\Phi}, \boldsymbol{\Phi}^\top \boldsymbol{c} \rangle_{\mathcal{H}} = \langle \boldsymbol{\Phi}^\top \boldsymbol{c}, \boldsymbol{\Phi}^\top \boldsymbol{c} \rangle_{\mathcal{H}} = \|\boldsymbol{\Phi}^\top \boldsymbol{c}\|_{\mathcal{H}}^2 \geq 0 \tag{5.4}$$

i.e. K is positive definite.

To show the second part, assume that K is not strictly positive definite. Hence, there exists a vector $\boldsymbol{c} \in \mathbb{R}^N \setminus \{\boldsymbol{0}\}$ such that $\boldsymbol{c}^\top \underline{\boldsymbol{K}} \boldsymbol{c} = 0$. (5.4) implies $\|\boldsymbol{\Phi}^\top \boldsymbol{c}\|_{\mathcal{H}}^2 = 0$ and thus $\boldsymbol{\Phi}^\top \boldsymbol{c} \equiv 0_{\mathcal{H}}$. So,

$$\begin{aligned} 0 = \langle f, 0_{\mathcal{H}} \rangle_{\mathcal{H}} = \langle f, \boldsymbol{\Phi}^\top \boldsymbol{c} \rangle_{\mathcal{H}} &= \left\langle f, \sum_{i=1}^N c_i K(\boldsymbol{x}_i, \cdot) \right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^N c_i \langle f, K(\boldsymbol{x}_i,\cdot)\rangle_{\mathcal{H}} = \sum_{i=1}^N c_i f(\boldsymbol{x}_i) = \sum_{i=1}^N c_i \delta_{\boldsymbol{x}_i}(f) \end{aligned}$$

for all $f \in \mathcal{H}$. Consequently, we have

$$\left\| \sum_{i=1}^N c_i \delta_{\boldsymbol{x}_i} \right\|_{\mathcal{H}'} \overset{(4.7)}{=} \sup_{0 \neq f \in \mathcal{H}} \frac{\left| \sum_{i=1}^N c_i \delta_{\boldsymbol{x}_i}(f) \right|}{\|f\|_{\mathcal{H}}} = 0 \quad \overset{\text{Def 4.1}}{\Longrightarrow} \quad \sum_{i=1}^N c_i \delta_{\boldsymbol{x}_i} \equiv 0_{\mathcal{H}'}$$

which implies the linear dependence of the point evaluation functionals $\delta_{\boldsymbol{x}_i}(f)$, $i = 1, \ldots, N$, since $\boldsymbol{c} \neq \boldsymbol{0}$. The converse direction follows analogously. □

[4] in Machine Learning jargon

The reverse statement of Theorem 5.4 is also correct: Each strictly positive definite kernel can be associated to an RKHS, its native space. Motivated by the fact that for $f = \sum_{i=1}^{N} c_i K(\boldsymbol{x}_i, \cdot)$ we have $\|f\|_{\mathcal{H}}^2 = \boldsymbol{c}^\top \underline{\boldsymbol{K}} \boldsymbol{c}$[5], we define the linear space

$$H_K(\Omega) := \left\{ \sum_{i=1}^{N} c_i K(\boldsymbol{x}_i, \cdot) : c_i \in \mathbb{R},\ \boldsymbol{x}_i \in \Omega,\ N \in \mathbb{N} \right\} \tag{5.5}$$

equipped with the bilinear form

$$\langle f, g \rangle_K = \left\langle \sum_{i=1}^{N} c_i K(\boldsymbol{x}_i, \cdot), \sum_{j=1}^{M} d_j K(\boldsymbol{y}_j, \cdot) \right\rangle_K := \boldsymbol{c}^\top \left[K(\boldsymbol{x}_i, \boldsymbol{y}_j) \right]_{i=1,j=1}^{N,M} \boldsymbol{d} \tag{5.6}$$

where $M = N = \infty$ is possible.

Theorem 5.5. If $K : \Omega \times \Omega \to \mathbb{R}$ is symmetric and strictly positive definite, then the bilinear form $\langle \cdot, \cdot \rangle_K$ defined in (5.6) defines an inner product on $H_K(\Omega)$. Moreover, $H_K(\Omega)$ is a pre-Hilbert space with reproducing kernel K. ◁

Proof. Symmetry follows directly from the symmetry of K. Positive definiteness follows from the strict positive definiteness of K according to

$$\langle f, f \rangle_K = \langle \boldsymbol{\Phi}^\top \boldsymbol{c}, \boldsymbol{\Phi}^\top \boldsymbol{c} \rangle_K = \boldsymbol{c}^\top \underline{\boldsymbol{K}}\, \boldsymbol{c} > 0$$

for all $f = \boldsymbol{\Phi}^\top \boldsymbol{c} = \sum_{i=1}^{N} c_i K(\boldsymbol{x}_i, \cdot) \neq 0_{H_K(\Omega)}$. Hence $\langle \cdot, \cdot \rangle_K$ is an inner product.

Finally, for any $\boldsymbol{y} \in \Omega$, the reproducing property is obtained by

$$\langle K(\boldsymbol{y}, \cdot), f \rangle_K = 1 \left[K(\boldsymbol{y}, \boldsymbol{x}_j) \right]_{j=1}^{N} \boldsymbol{c} = \sum_{j=1}^{N} c_j K(\boldsymbol{y}, \boldsymbol{x}_j) = \sum_{j=1}^{N} c_j K(\boldsymbol{x}_j, \boldsymbol{y}) = f(\boldsymbol{y})$$

for all $f = \sum_{j=1}^{N} c_j K(\boldsymbol{x}_j, \cdot) \in H_K(\Omega)$. □

Theorem 5.5 provides that $(H_K(\Omega), \langle \cdot, \cdot \rangle_K)$ is a pre-Hilbert space, hence it is not necessarily complete. But as we know from Theorem 4.1, each normed space exhibits a completion that is unique up to isometry.

Definition 5.3 (Native Space)**.** The completion $\mathcal{N}_K(\Omega) := \overline{H_K(\Omega)}^{\|\cdot\|_K}$ with respect to the norm $\|f\|_K := \sqrt{\langle f, f \rangle_K}$ is called the native space of K. ◀

5.1 Spectral POV

Another characterization of the native space is given by the eigenfunctions of the linear operator

$$T_K : L^2(\Omega) \to L^2(\Omega), \quad (T_K v)(\boldsymbol{x}) := \int_\Omega K(\boldsymbol{x}, \boldsymbol{y}) v(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}$$

Fact 5.6 (Mercer)**.** Let $K \in C(\Omega \times \Omega)$ be a continuous and positive definite kernel. Then, there holds

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\boldsymbol{x}) \phi_i(\boldsymbol{y}),$$

where $\{(\lambda_i, \phi_i)\}_{i=1}^{\infty}$ are the eigen-pairs of the compact operator T_K. ◁

The previous fact allows for a spectral characterization of the native space. To this end, we endow

$$\mathcal{H} := \left\{ f : \Omega \to \mathbb{R} : f = \sum_{i=1}^{\infty} c_i \phi_i, c_i \in \mathbb{R} \right\}$$

with the inner product

$$\langle f, g \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^{\infty} c_i \phi_i, \sum_{i=1}^{\infty} d_i \phi_i \right\rangle_{\mathcal{H}} := \sum_{i=1}^{\infty} \frac{c_i d_i}{\lambda_i} = \sum_{i=1}^{\infty} \frac{\langle f, \phi_i \rangle_{L^2(\Omega)} \langle g, \phi_i \rangle_{L^2(\Omega)}}{\lambda_i}$$

[5] i.e. motivated by the kernel trick

There holds

$$\begin{aligned}\langle K(\boldsymbol{x},\cdot), f\rangle_{\mathcal{H}} &= \sum_{i=1}^{\infty} \frac{\langle K(\boldsymbol{x},\cdot),\phi_i\rangle_{L^2(\Omega)}\langle f,\phi_i\rangle_{L^2(\Omega)}}{\lambda_i} \\ &= \sum_{i=1}^{\infty} \frac{\langle \sum_{j=1}^{\infty}\lambda_j\phi_j(\boldsymbol{x})\phi_j(\cdot),\phi_i\rangle_{L^2(\Omega)}\langle f,\phi_i\rangle_{L^2(\Omega)}}{\lambda_i} \\ &= \sum_{i=1}^{\infty} \frac{\lambda_i\phi_i(\boldsymbol{x})\langle f,\phi_i\rangle_{L^2(\Omega)}}{\lambda_i} = \sum_{i=1}^{\infty} c_i\phi_i(\boldsymbol{x}) = f(\boldsymbol{x})\end{aligned}$$

Consequently, we may set $\langle\cdot,\cdot\rangle_{\mathcal{N}_K(\Omega)} := \langle\cdot,\cdot\rangle_{\mathcal{H}}$ and obtain

$$\mathcal{N}_K(\Omega) = \left\{ f \in L^2(\Omega) : \langle f, f\rangle_{\mathcal{N}_K(\Omega)} < \infty \right\}.$$

6 Approximation results

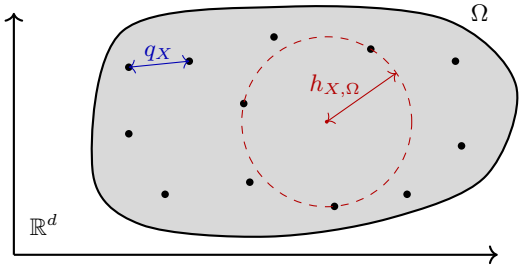
Definition 6.1. Given $\Omega \subset \mathbb{R}^d$ and $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subset \Omega$, we introduct the fill distance

$$h_{X,\Omega} := \sup_{\boldsymbol{x} \in \Omega} \min_{\boldsymbol{x}_i \in X} \|\boldsymbol{x} - \boldsymbol{x}_i\|_2 \tag{6.1}$$

and the separation distance

$$q_X := \min_{\boldsymbol{x}_i \neq \boldsymbol{x}_j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \tag{6.2}$$

We call X quasi-uniform, iff there exists a constant $c \geq 1$ such that $q_X/c \leq h_{X,\Omega} \leq c q_X$. ◀



We start by introducing the concept of *Lagrange bases*. To this end, we recall the kernel matrix

$$\underline{\boldsymbol{K}} := [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^N$$

from Definition 5.2 and the feature vector

$$\boldsymbol{\Phi}(\boldsymbol{x}) = [K(\boldsymbol{x}_i, \boldsymbol{x})]_{i=1}^N = [\varphi_1(\boldsymbol{x}), \dots, \varphi_N(\boldsymbol{x})]^\top$$

Further, we denote the canonical basis in $\mathbb{R}^N$ by $\boldsymbol{e}_1, \dots, \boldsymbol{e}_N$.

Theorem 6.1. Let K be a strictly positive definite kernel. Then, for any mutually distinct points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, the *Lagrange basis* is given by

$$\ell_j(\boldsymbol{x}) := \sum_{k=1}^N c_k^{(j)} K(\boldsymbol{x}_k, \boldsymbol{x}) = \boldsymbol{c}^{(j)} \boldsymbol{\Phi}(\boldsymbol{x})$$

with $\boldsymbol{c}^{(j)} := \boldsymbol{e}_j^\top \underline{\boldsymbol{K}}^{-1}$, $j = 1, \dots, N$. The functions ℓ_j satisfy $\ell_j(\boldsymbol{x}_i) = \delta_{ij}$. ◁

Proof. There holds

$$\ell_j(\boldsymbol{x}_i) = \boldsymbol{c}^{(j)} \boldsymbol{\Phi}(\boldsymbol{x}_i) = \boldsymbol{e}_j^\top \underline{\boldsymbol{K}}^{-1} \underline{\boldsymbol{K}}_{:,i} = \boldsymbol{e}_j^\top \underline{\boldsymbol{K}}^{-1} \underline{\boldsymbol{K}} \boldsymbol{e}_i = \boldsymbol{e}_j^\top \boldsymbol{e}_i = \delta_{ij}$$

since $\boldsymbol{\Phi}(\boldsymbol{x}_i)$ is the i-th column of $\underline{\boldsymbol{K}}$. □

Given a function $f : \Omega \to \mathbb{R}$ we can write its interpolant according to

$$s_f(\boldsymbol{x}) = \sum_{j=1}^N f(\boldsymbol{x}_j) \ell_j(\boldsymbol{x})$$

where $\boldsymbol{x} \in \Omega$.

To derive an error estimate in terms of the fill distance, an important tool is the

Definition 6.2 (power function)**.** Let $\Omega \subset \mathbb{R}^d$ and $K : \Omega \times \Omega \to \mathbb{R}$ be a continuous and strictly positive definite kernel. Given any set $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ of mutually distinct points, the power function $P_{K,X} : \Omega \to \mathbb{R}$ is defined as

$$P_{K,X}(\boldsymbol{x}) := \left\| K(\boldsymbol{x}, \cdot) - \sum_{j=1}^N \ell_j(\boldsymbol{x}) K(\boldsymbol{x}_j, \cdot) \right\|_{\mathcal{N}_K(\Omega)} \tag{6.3}$$

where $\boldsymbol{x} \in \Omega$ and ℓ_j are the Lagrange basis functions from Theorem 6.1. ◀

The power function reflects how well the observations at X can approximate or predict the value at a new point.

A direct calculation yields the following

Fact 6.2. Under the same assumptions as in Definition 6.2, we have

$$P_{K,X}(\boldsymbol{x}) = \sqrt{K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{x})^\top \underline{\boldsymbol{K}}^{-1} \boldsymbol{\Phi}(\boldsymbol{x})} \tag{6.4}$$

for all $\boldsymbol{x} \in \Omega$, and, hence, $0 \leq P_{K,X}(\boldsymbol{x}) \leq \sqrt{K(\boldsymbol{x},\boldsymbol{x})}$. ◁

The point-wise approximation error can be bounded by the power function.

Theorem 6.3. Under the same assumptions as in Definition 6.2, we have

$$|f(\boldsymbol{x}) - s_f(\boldsymbol{x})| \leq P_{K,X}(\boldsymbol{x}) \|f\|_{\mathcal{N}_K(\Omega)}$$

for every $f \in \mathcal{N}_K(\Omega)$ and $\boldsymbol{x} \in \Omega$. ◁

Proof. Using the reproducing property from Definition 5.1, we obtain

$$s_f(\boldsymbol{x}) = \sum_{j=1}^N f(\boldsymbol{x}_j)\ell_j(\boldsymbol{x}) = \sum_{j=1}^N \langle K(\boldsymbol{x}_j,\cdot), f\rangle_{\mathcal{N}_K(\Omega)} \ell_j(\boldsymbol{x}) = \left\langle \sum_{j=1}^N \ell_j(\boldsymbol{x}) K(\boldsymbol{x}_j,\cdot),\, f \right\rangle_{\mathcal{N}_K(\Omega)}$$

and thus

$$\begin{aligned}
|f(\boldsymbol{x}) - s_f(\boldsymbol{x})| &= \left| \langle K(\boldsymbol{x},\cdot), f\rangle_{\mathcal{N}_K(\Omega)} - \left\langle \sum_{j=1}^N \ell_j(\boldsymbol{x}) K(\boldsymbol{x}_j,\cdot),\, f \right\rangle_{\mathcal{N}_K(\Omega)} \right| \\
&= \left| \left\langle K(\boldsymbol{x},\cdot) - \sum_{j=1}^N \ell_j(\boldsymbol{x}) K(\boldsymbol{x}_j,\cdot),\, f \right\rangle_{\mathcal{N}_K(\Omega)} \right| \\
&\overset{(4.4)}{\leq} \left\| K(\boldsymbol{x},\cdot) - \sum_{j=1}^N \ell_j(\boldsymbol{x}) K(\boldsymbol{x}_j,\cdot) \right\|_{\mathcal{N}_K(\Omega)} \cdot \|f\|_{\mathcal{N}_K(\Omega)} \\
&= P_{K,X}(\boldsymbol{x}) \cdot \|f\|_{\mathcal{N}_K(\Omega)}
\end{aligned}$$

by the Cauchy-Schwarz inequality. □

Fact 6.4. Let $\Omega \subset \mathbb{R}^d$ satisfy an interior cone condition, i.e., there exists an angle $\alpha > 0$ such that the interior angle at every corner of Ω is bigger than α. Let $K \in C^{2k}(\Omega \times \Omega)$ be a strictly positive definite kernel. Then there exist constants $C_K, h_0 > 0$ such that

$$P_{K,X}(\boldsymbol{x}) \leq C_K h_{X,\Omega}^k$$

whenever $h_{X,\Omega} \leq h_0$ ◁

Combining Theorem 6.3 and Fact 6.4 yields the final error estimate

Theorem 6.5. Let $\Omega \subset \mathbb{R}^d$ satisfy an interior cone condition, $K \in C^{2k}(\Omega \times \Omega)$ be a strictly positive definite kernel and $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be a set of mutually distinct points. Then, we have for every $f \in \mathcal{N}_K(\Omega)$

$$|f(\boldsymbol{x}) - s_f(\boldsymbol{x})| \leq C_K h_{X,\Omega}^k \|f\|_{\mathcal{N}_K(\Omega)}$$

for all $\boldsymbol{x} \in \Omega$ whenever $h_{X,\Omega} \leq h_0$. ◁

6.1 Connection to Gaussian processes

Let $(\mathcal{S}, \mathcal{F}, \mathrm{P})$ be a probability space.[6]

Definition 6.3 (Random variable)**.** A *random variable* is a measurable map $X : \mathcal{S} \to \mathbb{R}$. ◀

Consider a vector $\boldsymbol{Z}$ split into two parts $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, where

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \underline{\boldsymbol{\Sigma}}_{11} & \underline{\boldsymbol{\Sigma}}_{12} \\ \underline{\boldsymbol{\Sigma}}_{21} & \underline{\boldsymbol{\Sigma}}_{22} \end{bmatrix} \right) \tag{6.5}$$

The conditional distribution of $\boldsymbol{Z}_2$ given $\boldsymbol{Z}_1$ is also Gaussian:

$$\boldsymbol{Z}_2 \mid \boldsymbol{Z}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}_2 + \underline{\boldsymbol{\Sigma}}_{21} \underline{\boldsymbol{\Sigma}}_{11}^{-1} (\boldsymbol{Z}_1 - \boldsymbol{\mu}_1),\ \underline{\boldsymbol{\Sigma}}_{22} - \underline{\boldsymbol{\Sigma}}_{21} \underline{\boldsymbol{\Sigma}}_{11}^{-1} \underline{\boldsymbol{\Sigma}}_{12} \right) \tag{6.6}$$

Definition 6.4 (Random function)**.** A *random function* is a map $f : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$. For fixed $\boldsymbol{x} \in \mathbb{R}^d$, $f(\boldsymbol{x}, \cdot)$ is a random variable; for fixed $\omega \in \mathcal{S}$, $f(\cdot, \omega)$ is a deterministic function. ◀

Definition 6.5 (Gaussian process)**.** A collection of random variables $\{f(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathbb{R}^d\}$ is called a *Gaussian process* if for all $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^d$,

$$[f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_m)]$$

is multivariate Gaussian. It is characterized by its mean $\mu(\boldsymbol{x}) = \mathrm{E}[f(\boldsymbol{x})]$ and covariance function $k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}(f(\boldsymbol{x}), f(\boldsymbol{x}'))$. ◀

Consider a Gaussian process with zero mean and covariance function k. Suppose we have observed $\boldsymbol{y} := [y_1, \ldots, y_{n-1}]^\top$ at points $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n-1}\}$. We want to predict the value of f at a new point $\boldsymbol{x}_\star := \boldsymbol{x}_n$, $f_\star := f(\boldsymbol{x}_\star)$.

The covariance matrix is given by $\underline{\boldsymbol{K}} = \left[k(\boldsymbol{x}_i, \boldsymbol{x}_j)\right]_{i,j=1}^n$ and it can be partitioned as

$$\underline{\boldsymbol{K}} = \begin{bmatrix} \underline{\boldsymbol{K}}_{mm} & \underline{\boldsymbol{K}}_{m\star} \\ \underline{\boldsymbol{K}}_{\star m} & \underline{\boldsymbol{K}}_{\star\star} \end{bmatrix} \tag{6.7}$$

where $\underline{\boldsymbol{K}}_{mm} \in \mathbb{R}^{(n-1)\times(n-1)}$ is the covariance on X, $\underline{\boldsymbol{K}}_{m\star} \in \mathbb{R}^{(n-1)\times 1}$ contains the covariances between X and $\boldsymbol{x}_\star$, and $\underline{\boldsymbol{K}}_{\star\star} \in \mathbb{R}$ is the variance at $\boldsymbol{x}_\star$.

Then:

(i) The conditional distribution of $f_\star$ given the observations $\boldsymbol{y}$ is Gaussian with

$$f_\star \mid \boldsymbol{y} \sim \mathcal{N}\left(\underline{\boldsymbol{K}}_{\star m} \underline{\boldsymbol{K}}_{mm}^{-1} \boldsymbol{y},\ \underline{\boldsymbol{K}}_{\star\star} - \underline{\boldsymbol{K}}_{\star m} \underline{\boldsymbol{K}}_{mm}^{-1} \underline{\boldsymbol{K}}_{m\star} \right)$$

where the covariance term is the Schur complement (7.1) of $\underline{\boldsymbol{K}}_{mm}$ in $\underline{\boldsymbol{K}}$.

(ii) Let $P_{k,X}$ be the power function from Definition 6.2. Using the explicit formula (6.4),

$$P_{k,X}(\boldsymbol{x}_\star)^2 = k(\boldsymbol{x}_\star, \boldsymbol{x}_\star) - \boldsymbol{\Phi}(\boldsymbol{x}_\star)^\top \underline{\boldsymbol{K}}_{mm}^{-1} \boldsymbol{\Phi}(\boldsymbol{x}_\star) = \underline{\boldsymbol{K}}_{\star\star} - \underline{\boldsymbol{K}}_{\star m} \underline{\boldsymbol{K}}_{mm}^{-1} \underline{\boldsymbol{K}}_{m\star}$$

implying that the power function evaluated at $\boldsymbol{x}_\star$ is exactly the conditional standard deviation (posterior uncertainty) of the Gaussian process at $\boldsymbol{x}_\star$ given the observations at X.

[6] Usually, the sample space is denoted by Ω, but we already used this symbol for domains in $\mathbb{R}^d$. Since the event space $\mathcal{F}$ is a *σ-algebra*, (i.e., a collection of subsets of $\mathcal{S}$ that contains the empty set, is closed under complementation and countable unions), it is sometimes denoted by Σ. By using $\mathcal{F}$ here, we avoid confusion with the notation for the covariance matrix.

7 Numerical methods

In this chapter, we focus on the situation where the basic or kernel function under consideration is positive definite. As a consequence, the generalized Vandermonde or kernel matrix $\underline{\boldsymbol{K}} = \left[K(\boldsymbol{x}_i, \boldsymbol{x}_j)\right]_{i,j=1}^{N} \in \mathbb{R}^{N\times N}$ is positive semi-definite. We need the following

Lemma 7.1. Let $\underline{\boldsymbol{A}}$ be a symmetric and positive semi-definite matrix. Then, the Schur complement of $\underline{\boldsymbol{A}}_{1,1}$ in $\underline{\boldsymbol{A}}$

$$\underline{\boldsymbol{S}} := \underline{\boldsymbol{A}}_{2,2} - \underline{\boldsymbol{A}}_{2,1}\underline{\boldsymbol{A}}_{1,1}^{-1}\underline{\boldsymbol{A}}_{1,2} \tag{7.1}$$

is well defined for any block partitioning of

$$\underline{\boldsymbol{A}} = \begin{bmatrix} \underline{\boldsymbol{A}}_{1,1} & \underline{\boldsymbol{A}}_{1,2} \\ \underline{\boldsymbol{A}}_{2,1} & \underline{\boldsymbol{A}}_{2,2} \end{bmatrix}$$

for which $\underline{\boldsymbol{A}}_{1,1}^{-1}$ exists. Moreover, $\underline{\boldsymbol{A}}_{1,1}$ is always symmetric and positive semi-definite, while $\underline{\boldsymbol{S}}$ is symmetric and positive semi-definite. ◁

Proof. Let $\left[\begin{smallmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{smallmatrix}\right] \in \mathbb{R}^N$ be partitioned similarly to $\underline{\boldsymbol{A}}$. Since

$$\begin{bmatrix} \underline{\boldsymbol{A}}_{1,1} & \underline{\boldsymbol{A}}_{1,2} \\ \underline{\boldsymbol{A}}_{2,1} & \underline{\boldsymbol{A}}_{2,2} \end{bmatrix} = \underline{\boldsymbol{A}} = \underline{\boldsymbol{A}}^\top = \begin{bmatrix} \underline{\boldsymbol{A}}_{1,1}^\top & \underline{\boldsymbol{A}}_{2,1}^\top \\ \underline{\boldsymbol{A}}_{1,2}^\top & \underline{\boldsymbol{A}}_{2,2}^\top \end{bmatrix}$$

we obtain

$$\underline{\boldsymbol{A}}_{1,1} = \underline{\boldsymbol{A}}_{1,1}^\top, \quad \underline{\boldsymbol{A}}_{2,2} = \underline{\boldsymbol{A}}_{2,2}^\top, \quad \underline{\boldsymbol{A}}_{1,2} = \underline{\boldsymbol{A}}_{2,1}^\top$$

Consequently, $\underline{\boldsymbol{A}}_{1,1}$ is symmetric and there holds

$$0 \le \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{0}\end{bmatrix}^\top \underline{\boldsymbol{A}} \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{0}\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{0}\end{bmatrix}^\top \begin{bmatrix}\underline{\boldsymbol{A}}_{1,1}\boldsymbol{x}\\ \underline{\boldsymbol{A}}_{2,1}\boldsymbol{x}\end{bmatrix} = \boldsymbol{x}^\top \underline{\boldsymbol{A}}_{1,1}\boldsymbol{x}$$

Therefore, $\underline{\boldsymbol{A}}_{1,1}$ is positive semi-definite. In fact, it is even positive definite as $\underline{\boldsymbol{A}}_{1,1}^{-1}$ exists by assumption. Furthermore, there holds

$$\underline{\boldsymbol{S}}^\top = \underline{\boldsymbol{A}}_{2,2}^\top - \underline{\boldsymbol{A}}_{1,2}^\top \underline{\boldsymbol{A}}_{1,1}^{-\top} \underline{\boldsymbol{A}}_{2,1}^\top = \underline{\boldsymbol{A}}_{2,2} - \underline{\boldsymbol{A}}_{2,1}\underline{\boldsymbol{A}}_{1,1}^{-1}\underline{\boldsymbol{A}}_{1,2} = \underline{\boldsymbol{S}}$$

Finally, we consider $\left[\begin{smallmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{smallmatrix}\right]$ with $\boldsymbol{x} = -\underline{\boldsymbol{A}}_{1,1}^{-1}\underline{\boldsymbol{A}}_{1,2}\boldsymbol{y}$. This yields

$$\begin{aligned} 0 &\le \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}^\top \underline{\boldsymbol{A}} \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}^\top \begin{bmatrix}\underline{\boldsymbol{A}}_{1,1}\boldsymbol{x} + \underline{\boldsymbol{A}}_{1,2}\boldsymbol{y}\\ \underline{\boldsymbol{A}}_{2,1}\boldsymbol{x} + \underline{\boldsymbol{A}}_{2,2}\boldsymbol{y}\end{bmatrix} \\ &= \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}^\top \begin{bmatrix} -\underline{\boldsymbol{A}}_{1,2}\boldsymbol{y} + \underline{\boldsymbol{A}}_{1,2}\boldsymbol{y} \\ -\underline{\boldsymbol{A}}_{2,1}\underline{\boldsymbol{A}}_{1,1}^{-1}\underline{\boldsymbol{A}}_{1,2}\boldsymbol{y} + \underline{\boldsymbol{A}}_{2,2}\boldsymbol{y}\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{0}\\ \underline{\boldsymbol{S}}\boldsymbol{y}\end{bmatrix} = \boldsymbol{y}^\top \underline{\boldsymbol{S}}\boldsymbol{y} \end{aligned}$$

which yields the semi-definiteness of $\underline{\boldsymbol{S}}$. □

Given a positive semi-definite matrix $\underline{\boldsymbol{A}}$, successively reducing the Schur complement by setting

$$\underline{\boldsymbol{A}}_1 := \underline{\boldsymbol{A}}, \quad \boldsymbol{\ell}_i := \frac{1}{\sqrt{a_{\pi(i),\pi(i)}^{(i)}}} \boldsymbol{a}_{:,\pi(i)}^{(i)}, \quad \underline{\boldsymbol{A}}_{i+1} := \underline{\boldsymbol{A}}_i - \boldsymbol{\ell}_i\boldsymbol{\ell}_i^\top$$

for a permutation π of the set $\{1,\dots,N\}$ leads to a representation

$$\underline{\boldsymbol{A}} = \sum_{i=1}^{\operatorname{rank}(\underline{\boldsymbol{A}})} \boldsymbol{\ell}_i\boldsymbol{\ell}_i^\top$$

given that all pivots $a_{\pi(i),\pi(i)}^{(i)}$ are non-zero. In this case, also all matrices $\underline{\boldsymbol{A}}_i$, $i = 1,\dots,\operatorname{rank}(\underline{\boldsymbol{A}})$ are positive semi-definite. This can be seen by introducing the permutation matrix $\underline{\boldsymbol{P}} := \left[\boldsymbol{e}_{\pi(1)},\dots,\boldsymbol{e}_{\pi(N)}\right]^\top$ and considering the matrix $\underline{\boldsymbol{P}}\underline{\boldsymbol{A}}\underline{\boldsymbol{P}}^\top$ in Lemma 7.1.

Remark 7.1. For $\pi(i) = i$, we obtain the well known Cholesky decomposition. ◀

Lemma 7.2. Let $\underline{\boldsymbol{A}}$ be a symmetric and positive semi-definite matrix. Then, there holds

$$|a_{i,j}| \le \sqrt{a_{i,i}\, a_{j,j}}$$

for all $i,j = 1,\dots,N$. ◁

Proof. The positive semi-definiteness of the Schur complement established by Lemma 7.1 holds true for any pivot element $a_{i,i}$, $i = 1, \dots, N$. In particular, all diagonal elements of the Schur complement have to be non-negative, which implies

$$0 \leq a_{j,j} - \frac{a_{i,j}^2}{a_{i,i}} \quad \text{or} \quad |a_{i,j}| \leq \sqrt{a_{i,i}\, a_{j,j}}$$

as claimed. □

A direct consequence of Lemma 7.2 is that the largest element of a positive semi-definite matrix is always located on the diagonal, i.e.,

$$|a_{i,j}| \leq \sqrt{a_{i,i}\, a_{j,j}} \leq \frac{a_{i,i} + a_{j,j}}{2} \leq \max_{i=1,\dots,N} a_{i,i}.$$

Therefore, if all diagonal elements are zero, the matrix has to be the zero matrix. This motivates the pivoted version of the Cholesky decomposition, Algorithm 1, which greedily removes the largest element from the Schur complement.

Definition 7.1 (Biorthogonal basis)**.** Let $\underline{\boldsymbol{L}}, \underline{\boldsymbol{B}} \in \mathbb{R}^{N \times m}$. We say that the column vectors of $\underline{\boldsymbol{L}}$ and $\underline{\boldsymbol{B}}$ form a biorthogonal basis of two subspaces $\mathcal{L}, \mathcal{B} \subseteq \mathbb{R}^N$ if

$$\underline{\boldsymbol{B}}^\top \underline{\boldsymbol{L}} = \underline{\boldsymbol{I}}_m$$

i.e., each pair of basis vectors satisfies $\boldsymbol{b}_i^\top \boldsymbol{\ell}_j = \delta_{ij}$ where $i, j = 1, \dots, m$ and δ_{ij} denotes the Kronecker delta.

In contrast to an orthonormal basis, the vectors within each set $\{\boldsymbol{\ell}_i\}$ or $\{\boldsymbol{b}_i\}$ need not be mutually orthogonal; only the cross-orthogonality between both sets is required. ◀

Remark 7.2. In the context of Algorithm 1, the matrix $\underline{\boldsymbol{B}}$ can be interpreted as the dual basis to $\underline{\boldsymbol{L}}$ with respect to the Euclidean inner product, since $\underline{\boldsymbol{B}}^\top \underline{\boldsymbol{L}} = \underline{\boldsymbol{I}}_m$ implies that each $\boldsymbol{b}_i$ extracts the coefficient of $\boldsymbol{\ell}_i$ from any linear combination of the columns of $\underline{\boldsymbol{L}}$. Hence, for any $\boldsymbol{v} \in \operatorname{span}(\underline{\boldsymbol{L}})$, the coefficient vector in this basis is given by $\boldsymbol{c} = \underline{\boldsymbol{B}}^\top \boldsymbol{v}$, $\boldsymbol{v} = \underline{\boldsymbol{L}}\boldsymbol{c}$. This duality ensures numerical stability when orthogonality of $\underline{\boldsymbol{L}}$ cannot be preserved. ◀

Algorithm 1 Pivoted Cholesky Decomposition

Require: symmetric and positive semi-definite matrix $\underline{\boldsymbol{K}} \in \mathbb{R}^{N \times N}$, tolerance $\varepsilon \geq 0$
Ensure: low-rank approximation $\underline{\boldsymbol{K}} \approx \underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top$ and biorthogonal basis $\underline{\boldsymbol{B}}$ such that $\underline{\boldsymbol{B}}^\top \underline{\boldsymbol{L}} = \underline{\boldsymbol{I}}_m$

```
Initialization: m := 1, d := diag(K), L := [], B := [], err := ‖d‖_1
while err > ε do
    determine π(m) := arg max_{1≤i≤N} d_i
    compute ℓ_m := 1/√(d_π(m)) (K − L L^⊤) e_π(m) and b_m := 1/√(d_π(m)) (I − B L^⊤) e_π(m)
    L := [L, ℓ_m], B := [B, b_m]
    d := d − ℓ_m ⊙ ℓ_m                          ▷ ⊙ denotes the Hadamard product
    err := ‖d‖_1, m := m + 1
```

By the previous considerations (Lemma 7.1, Remark 7.1, Lemma 7.2), the pivoting strategy in Algorithm 1 amounts to a total pivoting, which always eliminates the largest entry of the Schur complement (7.1). Moreover, it also computes the biorthogonal basis associated to $\underline{\boldsymbol{L}}$.

Fact 7.3. For any $\varepsilon \in \mathbb{R}_{\geq 0}$, Algorithm 1 computes $N \times m$-matrices $\underline{\boldsymbol{B}}$ and $\underline{\boldsymbol{L}}$ with $m \leq \operatorname{rank}(\underline{\boldsymbol{K}})$ such that $\underline{\boldsymbol{K}} - \underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top$ is psd and

$$\begin{aligned} \operatorname{trace}(\underline{\boldsymbol{K}} - \underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top) &\leq \varepsilon \\ \underline{\boldsymbol{B}}^\top \underline{\boldsymbol{L}} &= \underline{\boldsymbol{I}}_m \\ \underline{\boldsymbol{K}}\underline{\boldsymbol{B}} &= \underline{\boldsymbol{L}} \end{aligned}$$

as can be proven by induction. ◁

Corollary 7.4. Let $\underline{\boldsymbol{U}} = [\boldsymbol{b}_{\pi(1),:}^\top, \dots, \boldsymbol{b}_{\pi(m),:}^\top]^\top \in \mathbb{R}^{m \times m}$ be the first m rows[7] of the matrix that is obtained if we permute the rows of $\underline{\boldsymbol{B}}$ by π. Then $\underline{\boldsymbol{U}}\underline{\boldsymbol{U}}^\top = [k_{\pi(i),\pi(j)}]_{i,j=1,\dots,m}^{-1}$. ◁

[7] These are exactly the nonzero rows of $\underline{\boldsymbol{B}}$.

Proof. W.l.o.g., we assume that π is the identity permutation, i.e., $\pi(m) = m$. Let

$$\underline{\boldsymbol{K}} = \begin{bmatrix} \underline{\boldsymbol{K}}_{1,1} & \underline{\boldsymbol{K}}_{1,2} \\ \underline{\boldsymbol{K}}_{2,1} & \underline{\boldsymbol{K}}_{2,2} \end{bmatrix}, \qquad \underline{\boldsymbol{L}} = \begin{bmatrix} \underline{\boldsymbol{L}}_1 \\ \underline{\boldsymbol{L}}_2 \end{bmatrix} \begin{matrix} \} m \\ \} N-m \end{matrix}$$

where $\underline{\boldsymbol{K}}_{1,1}, \underline{\boldsymbol{L}}_1 \in \mathbb{R}^{m \times m}$. We have $\underline{\boldsymbol{K}}_{1,1} = \underline{\boldsymbol{L}}_1 \underline{\boldsymbol{L}}_1^\top$. Furthermore, Fact 7.3 yields

$$\underline{\boldsymbol{B}}^\top \underline{\boldsymbol{L}} = \begin{bmatrix} \underline{\boldsymbol{U}} \\ \underline{\boldsymbol{0}}_{(N-m)\times m} \end{bmatrix}^\top \begin{bmatrix} \underline{\boldsymbol{L}}_1 \\ \underline{\boldsymbol{L}}_2 \end{bmatrix} = \begin{bmatrix} \underline{\boldsymbol{U}}^\top & \underline{\boldsymbol{0}} \end{bmatrix} \begin{bmatrix} \underline{\boldsymbol{L}}_1 \\ \underline{\boldsymbol{L}}_2 \end{bmatrix} = \underline{\boldsymbol{U}}^\top \underline{\boldsymbol{L}}_1 = \underline{\boldsymbol{I}}_m$$

which shows $\underline{\boldsymbol{U}}^\top = \underline{\boldsymbol{L}}_1^{-1}$ or $\underline{\boldsymbol{U}} = \underline{\boldsymbol{L}}_1^{-\top}$. Combining this with the previous argument yields

$$\underline{\boldsymbol{K}}_{1,1}^{-1} = \left(\underline{\boldsymbol{L}}_1 \underline{\boldsymbol{L}}_1^\top\right)^{-1} = \underline{\boldsymbol{L}}_1^{-\top} \underline{\boldsymbol{L}}_1^{-1} = \underline{\boldsymbol{U}} \underline{\boldsymbol{U}}^\top$$

□

Remark 7.3. The matrix $\underline{\boldsymbol{U}}$ is upper triangular since $\underline{\boldsymbol{B}}$ is constructed such that $\boldsymbol{b}_{i,j} = 0$ for all $i < j$ in Algorithm 1. ◀

Remark 7.4. The well known Nyström method for the low-rank approximation of kernel matrices randomly selects data sites $\boldsymbol{x}_{\pi(1)}, \dots, \boldsymbol{x}_{\pi(m)}$ and computes the approximation

$$\underline{\boldsymbol{K}} \approx \left[K(\boldsymbol{x}_i, \boldsymbol{x}_{\pi(j)})\right]_{\substack{i=1,\dots,n \\ j=1,\dots,m}} \left[K(\boldsymbol{x}_{\pi(i)}, \boldsymbol{x}_{\pi(j)})\right]_{i,j=1,\dots,m}^{-1} \left[K(\boldsymbol{x}_{\pi(i)}, \boldsymbol{x}_j)\right]_{\substack{i=1,\dots,m \\ j=1,\dots,n}}$$

The previous corollary shows that this is equivalent to a pivoted Cholesky decomposition with pivots $\pi(1), \dots, \pi(m)$. ◀

Corollary 7.5. Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be an RKHS. Given the canonical freature vector $\boldsymbol{\Phi}(\boldsymbol{x}) := [K(\boldsymbol{x}_i, \boldsymbol{x})]_{i=1}^N$, the Newton basis

$$\boldsymbol{N}(\boldsymbol{x}) := \underline{\boldsymbol{B}}^\top \boldsymbol{\Phi}(\boldsymbol{x}) = \sum_{i=1}^m \left[\underline{\boldsymbol{B}}^\top\right]_{:,\pi(i)} \varphi_{\pi(i)}(\boldsymbol{x}) = \sum_{i=1}^m \left[\underline{\boldsymbol{B}}_{\pi(i),:}\right]^\top K(\boldsymbol{x}_{\pi(i)}, \boldsymbol{x}) \tag{7.2}$$

forms an orthonormal system in $\mathcal{H}$, i.e., $\langle N_i, N_j \rangle_{\mathcal{H}} = \delta_{ij}$ for $i, j = 1, \dots, m$, where $m = \operatorname{rank}(\underline{\boldsymbol{B}})$. ◁

Proof. There holds

$$\langle \boldsymbol{N}, \boldsymbol{N}^\top \rangle_{\mathcal{H}} = \underline{\boldsymbol{B}}^\top \langle \boldsymbol{\Phi}, \boldsymbol{\Phi}^\top \rangle_{\mathcal{H}} \underline{\boldsymbol{B}} = \underline{\boldsymbol{B}}^\top \underline{\boldsymbol{K}} \underline{\boldsymbol{B}} = \underline{\boldsymbol{B}}^\top \underline{\boldsymbol{L}} = \underline{\boldsymbol{I}}_m$$

by the third part of Fact 7.3. □

Remark 7.5. We have

$$\operatorname{span}\{N_1, \dots, N_m\} = \operatorname{span}\{K(\boldsymbol{x}_{\pi(1)}, \cdot), \dots, K(\boldsymbol{x}_{\pi(m)}, \cdot)\} \subset \operatorname{span}\{\varphi_1, \dots, \varphi_N\}$$ ◀

The orthogonal projection of a function $f \in \mathcal{H}$ onto the subspace spanned by $N_1, \dots, N_m$ is computed by

$$\mathcal{P} f := \sum_{i=1}^m N_i \langle N_i, f \rangle_{\mathcal{H}} \overset{5.1.2}{=} \underline{\boldsymbol{N}}^\top \underline{\boldsymbol{B}}^\top \boldsymbol{f} = \boldsymbol{\Phi}^\top \underline{\boldsymbol{B}} \underline{\boldsymbol{B}}^\top \boldsymbol{f} \tag{7.3}$$

where $\boldsymbol{f} = [f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_N)]^\top$. In particular, there holds

$$[(\mathcal{P} f)(\boldsymbol{x}_i)]_{i=1}^N = \underline{\boldsymbol{K}} \underline{\boldsymbol{B}} \underline{\boldsymbol{B}}^\top \boldsymbol{f} = \underline{\boldsymbol{L}} \underline{\boldsymbol{B}}^\top \boldsymbol{f}$$

Given a (low-rank) factorization of the kernel matrix $\underline{\boldsymbol{K}}$, we can directly compute the least square solution to the linear system

$$\underline{\boldsymbol{K}} \boldsymbol{c} = \boldsymbol{f}$$

Theorem 7.6. Let $\underline{\boldsymbol{K}} \approx \underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top$ be the pivoted Cholesky decomposition of the kernel matrix $\underline{\boldsymbol{K}}$. A minimum norm solution of the problem

$$\left\| \underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top \boldsymbol{x} - \boldsymbol{f} \right\|_2 \to \min$$

is given by

$$\boldsymbol{x}^\dagger = \underline{\boldsymbol{L}} \left(\underline{\boldsymbol{L}}^\top \underline{\boldsymbol{L}} \right)^{-2} \underline{\boldsymbol{L}}^\top \boldsymbol{f} \tag{7.4}$$

The cost for the computation of $\boldsymbol{x}^\dagger$ is $\mathcal{O}(Nm^2)$ where $m = \operatorname{rank}(\underline{\boldsymbol{L}})$. ◁

Proof. The Gaussian normal equations read

$$\underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top \underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top \boldsymbol{x} = \underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top \boldsymbol{b}$$

Inserting $\boldsymbol{x}^\dagger$ from (7.4) yields

$$\underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top \underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top \underline{\boldsymbol{L}} \left(\underline{\boldsymbol{L}}^\top \underline{\boldsymbol{L}} \right)^{-2} \underline{\boldsymbol{L}}^\top \boldsymbol{b} = \underline{\boldsymbol{L}}\underline{\boldsymbol{L}}^\top \boldsymbol{b}$$

which shows that $\boldsymbol{x}^\dagger$ solves the Gaussian normal equations and is consequently a minimum norm solution. □

Remark 7.6. The matrix $(\underline{\boldsymbol{L}}^\top \underline{\boldsymbol{L}})^{-2}$ has condition number $(\kappa(\underline{\boldsymbol{L}}))^4$ and therefore easily becomes ill-conditioned. To mitigate this, one may compute the QR-decomposition $\underline{\boldsymbol{L}} = \underline{\boldsymbol{Q}}\underline{\boldsymbol{R}}$. Then, there holds $\underline{\boldsymbol{L}}^\top \underline{\boldsymbol{L}} = \underline{\boldsymbol{R}}^\top \boldsymbol{Q}^\top \underline{\boldsymbol{Q}}\underline{\boldsymbol{R}} = \underline{\boldsymbol{R}}^\top \underline{\boldsymbol{R}}$. The action of $(\underline{\boldsymbol{L}}^\top \underline{\boldsymbol{L}})^{-2}$ can thus be computed by solving two linear systems for $\underline{\boldsymbol{R}}$ and $\underline{\boldsymbol{R}}^\top$ respectively.

Since these matrices are assumed to be relatively small, a robust solver is given by the (pseudo-)inverse based on the singular value decomposition $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$. There holds

$$\left(\underline{\boldsymbol{L}}^\top \underline{\boldsymbol{L}} \right)^{-2} = \left(\underline{\boldsymbol{R}}^\top \underline{\boldsymbol{R}} \right)^{-2} = \left(\underline{\boldsymbol{V}}\underline{\boldsymbol{\Sigma}}^2 \underline{\boldsymbol{V}}^\top \right)^{-2} = \underline{\boldsymbol{V}}\underline{\boldsymbol{\Sigma}}^{-4} \underline{\boldsymbol{V}}^\top.$$

This particularly allows to threshold small values in $\underline{\boldsymbol{\Sigma}}^{-4}$. ◀

8 Optimal recovery

We consider the following

Problem 8.1 (Optimal recovery)**.** Given values $f_i := \lambda_i(f)$, $i = 1, \dots, N$, where $\{\lambda_1, \dots, \lambda_N\}$ is a set of linearly independent linear functionals (called information functionals), how can we best approximate the value $\lambda(f)$ of a known functional λ for an unknown function f? ◀

In the Hilbert space setting, the set of all functions that are consistent with the given data

$$\mathcal{A} := \{g \in \mathcal{H} : \lambda_i(g) = f_i, i = 1, \dots, N\}$$

forms an *affine subspace* of $\mathcal{H}$. We are looking for the "least biased" among these functions.

As we will see in Theorem 8.4, the solution to this problem is given by the minimum-norm interpolant, i.e., the function $g^\star \in \mathcal{H}$ with

$$\lambda_i(g^\star) = \lambda_i(f), \quad i = 1, \dots, N \tag{8.1}$$

and

$$\|g^\star\|_{\mathcal{H}} = \min_{g \in \mathcal{H}:(8.1)} \|g\|_{\mathcal{H}}$$

Geometrically, this corresponds to the orthogonal projection of the origin $0_{\mathcal{H}}$ onto the affine subspace $\mathcal{A}$.

We present three corresponding optimality results for radial basis function interpolation. As a preparation, we require two lemmata.

Lemma 8.1. Let K be a strictly positive definite kernel. Then

$$\langle s_f,\, s_f - g\rangle_{\mathcal{N}_K(\Omega)} = 0 \tag{8.2}$$

for all interpolants $g \in \mathcal{N}_K(\Omega)$ with $g(\boldsymbol{x}_i) = f(\boldsymbol{x}_i)$ for $i = 1, \dots, N$. ◁

This is consistent with the geometric interpretation of s_f being the orthogonal projection of $0_{\mathcal{H}}$ onto the affine subspace $\mathcal{A}$.

Proof. There holds

$$\begin{aligned}
\langle s_f,\, s_f - g\rangle_{\mathcal{N}_K(\Omega)} &= \left\langle \sum_{j=1}^N c_j K(\boldsymbol{x}_j, \cdot),\, s_f - g \right\rangle_{\mathcal{N}_K(\Omega)} \\
&= \sum_{j=1}^N c_j \langle K(\boldsymbol{x}_j, \cdot),\, s_f - g\rangle_{\mathcal{N}_K(\Omega)} \\
&= \sum_{j=1}^N c_j \big(s_f(\boldsymbol{x}_j) - g(\boldsymbol{x}_j)\big) = 0
\end{aligned}$$

since s_f and g both interpolate f. □

Lemma 8.2. Let K be a strictly positive definite kernel. Then

$$\langle f - s_f,\, h\rangle_{\mathcal{N}_K(\Omega)} = 0 \tag{8.3}$$

for all $h \in \operatorname{span}\{\varphi_1, \dots, \varphi_N\}$, where $\varphi_k(\cdot) := K(\boldsymbol{x}_k, \cdot)$. ◁

This means that the error $f - s_f$ is orthogonal to the finite-dimensional space spanned by the basis functions. Geometrically, this means that s_f corresponds to the orthogonal projection of f onto $\operatorname{span}\{\varphi_1, \dots, \varphi_N\}$.

Proof.

$$\begin{aligned}
\langle f - s_f, h\rangle_{\mathcal{N}_K(\Omega)} &= \left\langle f - s_f, \sum_{k=1}^N \alpha_k \varphi_k \right\rangle_{\mathcal{N}_K(\Omega)} \\
&= \sum_{k=1}^N \alpha_k \langle \varphi_k, f - s_f\rangle_{\mathcal{N}_K(\Omega)} \\
&= \sum_{k=1}^N \alpha_k \underbrace{\big(f(\boldsymbol{x}_k) - s_f(\boldsymbol{x}_k)\big)}_{=0} = 0
\end{aligned}$$

since s_f interpolates f. □

A straightforward consequence is a Pythagorean theorem.

Corollary 8.3. There holds

$$\|f\|^2_{\mathcal{N}_K(\Omega)} = \|f - s_f\|^2_{\mathcal{N}_K(\Omega)} + \|s_f\|^2_{\mathcal{N}_K(\Omega)}$$

◁

Proof.

$$\begin{aligned}
\|f - s_f\|^2_{\mathcal{N}_K(\Omega)} &= \langle f - s_f, f - s_f\rangle_{\mathcal{N}_K(\Omega)} \\
&= \langle f, f\rangle_{\mathcal{N}_K(\Omega)} - 2\langle f, s_f\rangle_{\mathcal{N}_K(\Omega)} + \langle s_f, s_f\rangle_{\mathcal{N}_K(\Omega)} \\
&= \langle f, f\rangle_{\mathcal{N}_K(\Omega)} - 2\langle f, s_f\rangle_{\mathcal{N}_K(\Omega)} + 2\langle s_f, s_f\rangle_{\mathcal{N}_K(\Omega)} - \langle s_f, s_f\rangle_{\mathcal{N}_K(\Omega)} \\
&= \|f\|^2_{\mathcal{N}_K(\Omega)} - 2\underbrace{\langle f - s_f, s_f\rangle_{\mathcal{N}_K(\Omega)}}_{\overset{(8.3)}{=}0} - \|s_f\|^2_{\mathcal{N}_K(\Omega)} \\
&= \|f\|^2_{\mathcal{N}_K(\Omega)} - \|s_f\|^2_{\mathcal{N}_K(\Omega)}
\end{aligned}$$

since $s_f \in \operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$. □

Theorem 8.4 (Optimality I)**.** Let K be a strictly positive definite kernel. Then, given the values $f_1, \ldots, f_N$,

$$\|s_f\|_{\mathcal{N}_K(\Omega)} = \min_{\substack{g \in \mathcal{N}_K(\Omega) \\ g(\boldsymbol{x}_j) = f_j}} \|g\|_{\mathcal{N}_K(\Omega)} \tag{8.4}$$

i.e., the interpolant s_f is the minimum-norm interpolant. ◁

Proof.

$$\begin{aligned}
\|s_f\|^2_{\mathcal{N}_K(\Omega)} &= \langle s_f,\, s_f\rangle_{\mathcal{N}_K(\Omega)} \\
&= \langle s_f,\, s_f - g + g\rangle_{\mathcal{N}_K(\Omega)} \\
&= \underbrace{\langle s_f,\, s_f - g\rangle_{\mathcal{N}_K(\Omega)}}_{\overset{(8.2)}{=}0} + \langle s_f,\, g\rangle_{\mathcal{N}_K(\Omega)} \\
&= \langle s_f,\, g\rangle_{\mathcal{N}_K(\Omega)}
\end{aligned}$$

By the Cauchy-Schwarz inequality (4.4),

$$\|s_f\|^2_{\mathcal{N}_K(\Omega)} \le \|s_f\|_{\mathcal{N}_K(\Omega)}\, \|g\|_{\mathcal{N}_K(\Omega)}$$

Dividing by $\|s_f\|_{\mathcal{N}_K(\Omega)}$ yields the assertion. □

Theorem 8.5 (Optimality II)**.** Let K be a strictly positive definite kernel. Then s_f is the best approximation to $f \in \mathcal{N}_K(\Omega)$ within $\operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$, i.e.,

$$\|f - s_f\|_{\mathcal{N}_K(\Omega)} \le \|f - g\|_{\mathcal{N}_K(\Omega)}$$

for all $g \in \operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$. ◁

Proof. Since $g - s_f \in \operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$, we have from (8.3) that

$$\begin{aligned}
\|f - s_f\|^2_{\mathcal{N}_K(\Omega)} &= \langle f - s_f,\, f - g + g - s_f\rangle_{\mathcal{N}_K(\Omega)} \\
&= \langle f - s_f,\, f - g\rangle_{\mathcal{N}_K(\Omega)} + \underbrace{\langle f - s_f,\, g - s_f\rangle_{\mathcal{N}_K(\Omega)}}_{\overset{(8.3)}{=}0} \\
&= \langle f - s_f,\, f - g\rangle_{\mathcal{N}_K(\Omega)}
\end{aligned}$$

By the Cauchy-Schwarz inequality (4.4),

$$\|f - s_f\|^2_{\mathcal{N}_K(\Omega)} \le \|f - s_f\|_{\mathcal{N}_K(\Omega)}\, \|f - g\|_{\mathcal{N}_K(\Omega)}$$

which we divide by $\|f - s_f\|_{\mathcal{N}_K(\Omega)}$ to obtain the assertion. □

Remark 8.1. The previous two optimality theorems also hold for strictly conditionally positive definite kernels, given that the point set $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ is unisolvent. ◀

We state the last optimality theorem in the context of quasi-interpolation without proof.

Theorem 8.6 (Optimality III)**.** Let K be strictly conditionally positive definite with respect to $P \subset C(\Omega)$ and assume that X is P-unisolvent. Then for any fixed $\boldsymbol{x} \in \Omega$

$$\sup_{\substack{f \in \mathcal{N}_K(\Omega) \\ \|f\|_{\mathcal{N}_K(\Omega)}=1}} \left| f(\boldsymbol{x}) - \sum_{j=1}^{N} f(\boldsymbol{x}_j)\, \ell_j(\boldsymbol{x}) \right| \;\leq\; \sup_{\substack{g \in \mathcal{N}_K(\Omega) \\ \|g\|_{\mathcal{N}_K(\Omega)}=1}} \left| g(\boldsymbol{x}) - \sum_{j=1}^{N} g(\boldsymbol{x}_j)\, c_j \right|$$

for any choice $c_1, \ldots, c_N \in \mathbb{R}$ with $\boldsymbol{c}^\top \underline{\boldsymbol{P}} = \boldsymbol{0}$. ◁

9 Least squares approximation

As we have shown in the previous chapter, the kernel interpolation solves a constraint optimization problem. We adopt this perspective here, but make the more general assumption that our ansatz is of the form

$$s_{f,m} := \sum_{j=1}^{m} c_j K(\tilde{\boldsymbol{x}}_j, \cdot)$$

where $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_m$ are not necessarily contained in X. Then, we are looking for a vector $\boldsymbol{c} \in \mathbb{R}^m$ which minimizes the quadratic form

$$\frac{1}{2}\boldsymbol{c}^\top \underline{\boldsymbol{Q}}\boldsymbol{c}$$

for some symmetric and positive definite matrix $\underline{\boldsymbol{Q}}$, subject to the linear constraints

$$\underline{\boldsymbol{A}}\boldsymbol{c} = \boldsymbol{f}$$

with the generalized Vandermonde matrix $\underline{\boldsymbol{A}} \in \mathbb{R}^{N\times m}$.

Remark 9.1. If $\underline{\boldsymbol{Q}}$ is chosen as the kernel Gram matrix on the centers,

$$\underline{\boldsymbol{Q}} = \left[K(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)\right]_{i,j=1}^{m}$$

then the quadratic form $\boldsymbol{c}^\top \underline{\boldsymbol{Q}}\boldsymbol{c}$ coincides with the squared RKHS (native space) norm of the kernel expansion $s_{f,m}$, as can be seen from

$$\begin{aligned}
\|s_{f,m}\|^2_{\mathcal{N}_K(\Omega)} &= \left\| \sum_{j=1}^{m} c_j K(\tilde{\boldsymbol{x}}_j, \cdot) \right\|^2_{\mathcal{N}_K(\Omega)} \\
&= \left\langle \sum_{i=1}^{m} c_i K(\tilde{\boldsymbol{x}}_i, \cdot), \sum_{j=1}^{m} c_j K(\tilde{\boldsymbol{x}}_j, \cdot) \right\rangle_{\mathcal{N}_K(\Omega)} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \left\langle K(\tilde{\boldsymbol{x}}_i, \cdot), K(\tilde{\boldsymbol{x}}_j, \cdot)\right\rangle_{\mathcal{N}_K(\Omega)} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j K(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j) \\
&= \boldsymbol{c}^\top \underline{\boldsymbol{Q}}\boldsymbol{c}
\end{aligned}$$

where we used the basic properties of Inner products and RKHSs

Hence minimizing $\frac{1}{2}\boldsymbol{c}^\top \underline{\boldsymbol{Q}}\boldsymbol{c}$ selects, among all expansions satisfying the constraints, the interpolant of minimal native space norm. Typically, this leads to a smooth interpolant. ◀

This constraint optimization problem is solved by minimizing

$$\mathcal{L}(\boldsymbol{c}, \boldsymbol{\lambda}) := \frac{1}{2}\boldsymbol{c}^\top \underline{\boldsymbol{Q}}\boldsymbol{c} - \boldsymbol{\lambda}^\top(\underline{\boldsymbol{A}}\boldsymbol{c} - \boldsymbol{f})$$

with the Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^N$. The unique minimum of $\mathcal{L}(\boldsymbol{c}, \boldsymbol{\lambda})$ is obtained from the solution of the saddle point formulation

$$\begin{bmatrix} \underline{\boldsymbol{Q}} & -\underline{\boldsymbol{A}}^\top \\ \underline{\boldsymbol{A}} & \underline{\boldsymbol{0}} \end{bmatrix} \begin{bmatrix} \boldsymbol{c} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{f} \end{bmatrix}$$

The solution is obtained by block Gaussian elimination in accordance with

$$\begin{aligned}
\boldsymbol{\lambda} &= \left(\underline{\boldsymbol{A}}\underline{\boldsymbol{Q}}^{-1}\underline{\boldsymbol{A}}^\top\right)^\dagger \boldsymbol{f} \\
\boldsymbol{c} &= \underline{\boldsymbol{Q}}^{-1}\underline{\boldsymbol{A}}^\top \boldsymbol{\lambda}
\end{aligned}$$

In the particular case that $m = N$, $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i$, and $\underline{\boldsymbol{A}} = \underline{\boldsymbol{Q}} = \underline{\boldsymbol{K}}$, we find

$$\boldsymbol{c} = \boldsymbol{\lambda} = \underline{\boldsymbol{K}}^{-1}\boldsymbol{f}$$

as in the previous chapter. However, the presented approach is more general as it also considers the cases $N < m$ (underdetermined least squares) and $N > m$ (overdetermined least squares), where the matrix $\underline{\boldsymbol{Q}}$ takes the role of a regularization term.

In the case that $\underline{\boldsymbol{Q}} = \left[K(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)\right]_{i,j=1}^m$ represents the native space norm of the interpolant, we obtain the least-squares problem

$$\min_{\boldsymbol{c}\in\mathbb{R}^m} \frac{1}{2}\|\underline{\boldsymbol{A}}\boldsymbol{c} - \boldsymbol{f}\|_2^2 + \frac{\omega}{2}\|s_{f,m}\|_{\mathcal{N}_K(\Omega)}^2$$

The ridge parameter ω controls the tradeoff between the smoothness and the fit of $s_{f,m}$. Finally, if we choose $m = N$, $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i$, and $\underline{\boldsymbol{A}} = \underline{\boldsymbol{Q}} = \underline{\boldsymbol{K}}$, this minimization problem becomes

$$\min_{\boldsymbol{c}\in\mathbb{R}^N} \frac{1}{2}\|\underline{\boldsymbol{K}}\boldsymbol{c} - \boldsymbol{f}\|_2^2 + \frac{\omega}{2}\boldsymbol{c}^\top \underline{\boldsymbol{K}}\boldsymbol{c}$$

The first order optimality condition reads

$$\underline{\boldsymbol{K}}^2\boldsymbol{c} + \omega\underline{\boldsymbol{K}}\boldsymbol{c} = \underline{\boldsymbol{K}}(\underline{\boldsymbol{K}} + \omega\underline{\boldsymbol{I}})\boldsymbol{c} = \underline{\boldsymbol{K}}\boldsymbol{f}$$

If $\underline{\boldsymbol{K}}$ has a trivial kernel, this equation is satisfied iff

$$(\underline{\boldsymbol{K}} + \omega\underline{\boldsymbol{I}})\boldsymbol{c} = \boldsymbol{f}$$

10 Support vector machines

10.1 Binary classification and linear separability

Given N points $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subset \mathbb{R}^d$ and labels $y_i \in \{-1, 1\}$ for $i = 1, \dots, N$, we introduce the sets

$$\begin{aligned} X_+ &:= \{\boldsymbol{x}_i \in X : y_i = 1\} \\ X_- &:= \{\boldsymbol{x}_i \in X : y_i = -1\} \end{aligned} \tag{10.1}$$

We are interested in solving the

Problem 10.1 (Binary classification problem)**.** Given the two sets X_+ and X_- from (10.1), find a function $f : \mathbb{R}^d \to \mathbb{R}$ such that $f(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in X_+$ and $f(\boldsymbol{x}) < 0$ for all $\boldsymbol{x} \in X_-$. ◀

In the easiest case, the two sets can be split by a separating hyper-plane.

Definition 10.2 (Linear separability)**.** The sets X_+ and X_- are called linearly separable, iff there exists a separating hyper-plane

$$H = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{n}^\top \boldsymbol{x} = m\}$$

such that $\boldsymbol{n}^\top \boldsymbol{x} > m$ iff $\boldsymbol{x} \in X_+$ and $\boldsymbol{n}^\top \boldsymbol{x} \leq m$ iff $\boldsymbol{x} \in X_-$. ◀

10.2 Hard-margin support vector machines

If X_+ and X_- are linearly separable, it is sufficient to determine an affine map

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

whose zero levelset

$$\mathcal{S} := \{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}) = 0\}$$

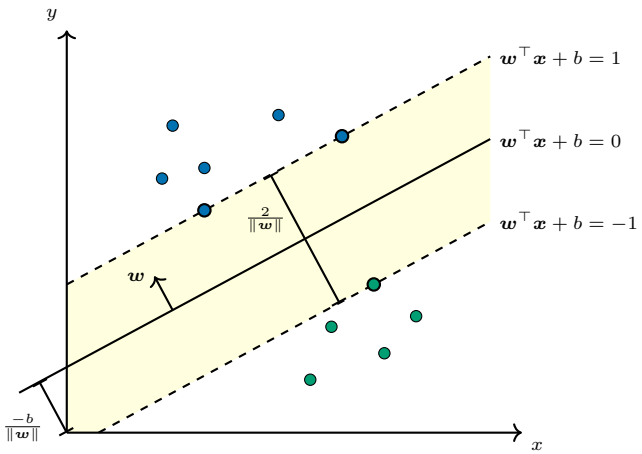
serves as separator. More precisely, we wish to determine a vector $\boldsymbol{w} \in \mathbb{R}^d$ and a threshold b such that the following two separation conditions are satisfied:

$$\begin{aligned} \boldsymbol{w}^\top \boldsymbol{x}_i + b &\geq 1, \quad \text{if } y_i = 1 \\ \boldsymbol{w}^\top \boldsymbol{x}_i + b &\leq -1, \quad \text{if } y_i = -1 \end{aligned} \tag{10.2}$$

These conditions can be summarized according to

$$y_i\big(\boldsymbol{w}^\top \boldsymbol{x}_i + b\big) \geq 1 \tag{10.3}$$

for all $i = 1, \dots, N$.



Given that X_+ and X_- are linearly separable, typically there exists more than one solution. Therefore, we aim to find a separator that maximizes the distance from X_+ and X_- (maximal margin). In this case, there exist points $\boldsymbol{x}_+ \in X_+$, $\boldsymbol{x}_- \in X_-$, such that

$$\boldsymbol{w}^\top \boldsymbol{x}_+ + b = 1, \qquad \boldsymbol{w}^\top \boldsymbol{x}_- + b = -1 \tag{10.4}$$

which are called support vectors. Taking any pair of such points, we have

$$\boldsymbol{w}^\top(\boldsymbol{x}_+ - \boldsymbol{x}_-) = 2 \tag{10.5}$$

From this, we obtain the separator by solving the maximization problem

$$\frac{1}{\|\boldsymbol{w}\|_2}\boldsymbol{w}^\top(\boldsymbol{x}_+ - \boldsymbol{x}_-) = \frac{2}{\|\boldsymbol{w}\|_2} \to \max \tag{10.6}$$

The latter es equivalent to the minimization problem

$$\frac{1}{2}\|\boldsymbol{w}\|_2^2 = \frac{1}{2}\sum_{i=1}^{d} w_i^2 \to \min \tag{10.7}$$

Imposing the separation conditions finally yields the constrained optimization

Problem 10.3 (primal)**.**

$$\min_{[\boldsymbol{w},b]^\top \in \mathbb{R}^{d+1}} \frac{1}{2}\sum_{i=1}^{d} w_i^2$$

such that

$$y_i\big(\boldsymbol{w}^\top \boldsymbol{x}_i + b\big) \geq 1$$

for all $i = 1, \ldots, N$. ◀

A solution $[\boldsymbol{w}^\star, b^\star]^\top$ to Problem 10.3 gives rise to the hard margin SVM classifier according to

$$c(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b) \tag{10.8}$$

Fact 10.1. Let $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset \mathbb{R}^d$ with labels $y_i \in \{-1, 1\}$ for $i = 1, \ldots, N$ be given. If the sets X_+ and X_- are non-empty and linearly separable, then the optimization Problem 10.3 has a unique solution $[\boldsymbol{w}^\star, b^\star]^\top$ with $\boldsymbol{w}^\star \neq \boldsymbol{0}$. ◁

10.3 Lagrangian formulation and KKT conditions

To solve the optimization Problem 10.3, we introduce the N non-negative Lagrange multipliers λ_i, $i = 1, \ldots, N$, and consider the Lagrange functional

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^{N} \lambda_i\big(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1\big) \tag{10.9}$$

The constrained optimization Problem 10.3 is now equivalent to the unconstrained one

$$\min_{[\boldsymbol{w},b]^\top \in \mathbb{R}^{d+1}} \max_{\boldsymbol{\lambda} \in \mathbb{R}^N} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\lambda}) \tag{10.10}$$

which is called the primal problem.

Minimizing $\mathcal{L}$ with respect to $\boldsymbol{w}$ and b yields the first order optimality conditions

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{w}}\mathcal{L} &= \boldsymbol{w} - \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x}_i = \boldsymbol{0} \\ \frac{\partial}{\partial b}\mathcal{L} &= -\sum_{i=1}^{N} \lambda_i y_i = 0 \end{aligned} \tag{10.11}$$

In addition, we have to satisfy the complementarity conditions

$$\lambda_i \geq 0, \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 \geq 0, \quad \lambda_i\big(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1\big) = 0 \tag{10.12}$$

for $i = 1, \ldots, N$.

Equations (10.11) and (10.12) are known as Karush–Kuhn–Tucker conditions (KKT). They are necessary and sufficient for the existence of an optimal solution. Particularly, (10.12) ensures that either $\boldsymbol{x}_i$ is lying on the hyperplane $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1$ or $\lambda_i = 0$.

10.4 Dual problem and primal–dual correspondence

Inserting (10.11) into (10.9) eliminates the variables $\boldsymbol{w}, b$ according to

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j =: -f(\boldsymbol{\lambda}) \tag{10.13}$$

Problem 10.4 (dual)**.**

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^N} f(\boldsymbol{\lambda})$$

such that

$$\sum_{i=1}^{N} \lambda_i y_i = 0, \quad \lambda_i \geq 0$$

for all $i = 1, \ldots, N$. This is called the dual problem to Problem 10.3. ◀

Vice versa, wa can solve Problem 10.3 by solving Problem 10.4 and inserting (10.11). Therefore, we have the following

Theorem 10.2. Let $\boldsymbol{\lambda}^\star \in \mathbb{R}^N$ be a solution to the dual Problem 10.4. Setting

$$\boldsymbol{w}^\star := \sum_{i=1}^{N} \lambda_i^\star y_i \boldsymbol{x}_i$$

and choosing $b^\star$ such that

$$y_i\big((\boldsymbol{w}^\star)^\top \boldsymbol{x}_i + b^\star\big) = 1$$

for any $i \in \{1, \ldots, N\}$ with $\lambda_i \neq 0$ yields the solution $[\boldsymbol{w}^\star, b^\star]^\top$ to Problem 10.3. ◁

A solution to the dual problem exists whenever the conditions of Fact 10.1 are satisfied.

Remark 10.1. Since f in Problem 10.4 is not strictly convex, the solution may not be unique. The optimization problem can efficiently be solved by the active set method. ◀

10.5 Kernel support vector machines

We finish this chapter by considering the situation that the sets X_+ and X_-, cf. (10.1), are not linearly separable. In this case, we replace the inner product $\boldsymbol{x}^\top \boldsymbol{y}$ in (10.8) by the inner product of the canonical feature map, i.e.,

$$\langle K(\boldsymbol{x}, \cdot), K(\boldsymbol{y}, \cdot) \rangle_{\mathcal{H}} = K(\boldsymbol{x}, \boldsymbol{y})$$

for all $\boldsymbol{x}, \boldsymbol{y} \in X$.

In what follows, we assume that the data (φ_i, y_i), $i = 1, \ldots, N$, is linearly separable in the RKHS $\mathcal{H}$. Then there exist

$$w \in \mathcal{H}_X := \operatorname{span}\{K(\boldsymbol{x}_1, \cdot), \ldots, K(\boldsymbol{x}_N, \cdot)\} \subset \mathcal{H}, \qquad b \in \mathbb{R}$$

such that

$$y_i\big(\langle w, K(\boldsymbol{x}_i, \cdot) \rangle_{\mathcal{H}} + b\big) \geq 1$$

for all $i = 1, \ldots, N$.

Remark 10.2. The kernel interpolant

$$w(\boldsymbol{x}) = \sum_{i=1}^{N} y_i \ell_i(\boldsymbol{x})$$

obviously satisfies $w(\boldsymbol{x}_i) = y_i$ for $i = 1, \ldots, N$ and, hence, the above inequality for $b = 0$. Therefore, the existence of a solution is guaranteed whenever the kernel is strictly positive definite. ◀

Analogously to the linear case, the weight is obtained by solving the optimization problem

$$\min_{(w,b)\in\mathcal{H}_X\times\mathbb{R}} \frac{1}{2}\|w\|_{\mathcal{H}}^2 \tag{10.14}$$

such that the constraints above are satisfied. The existence and uniqueness of a solution is obtained analogously to Fact 10.1. Considering the dual problem yields the optimization problem

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^N} \frac{1}{2}\sum_{i,j=1}^{N} \lambda_i\lambda_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) - \sum_{i=1}^{N}\lambda_i \tag{10.15}$$

with the constraints

$$\sum_{i=1}^{N}\lambda_i y_i = 0, \quad \lambda_i \geq 0$$

for all $i = 1,\dots,N$.

Given a solution $\boldsymbol{\lambda}^\star$, we retrieve $w^\star$ via

$$w^\star = \sum_{i=1}^{N}\lambda_i^\star y_i K(\boldsymbol{x}_i,\cdot)$$

and $b^\star$ by a choice such that

$$y_i\big(\langle w, K(\boldsymbol{x}_i,\cdot)\rangle_{\mathcal{H}} + b^\star\big) = 1$$

for any $i \in \{1,\dots,N\}$ with $\lambda_i \neq 0$. The classifier is then finally given by

$$c(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{N}\lambda_i^\star y_i K(\boldsymbol{x}_i,\boldsymbol{x}) + b^\star\right)$$