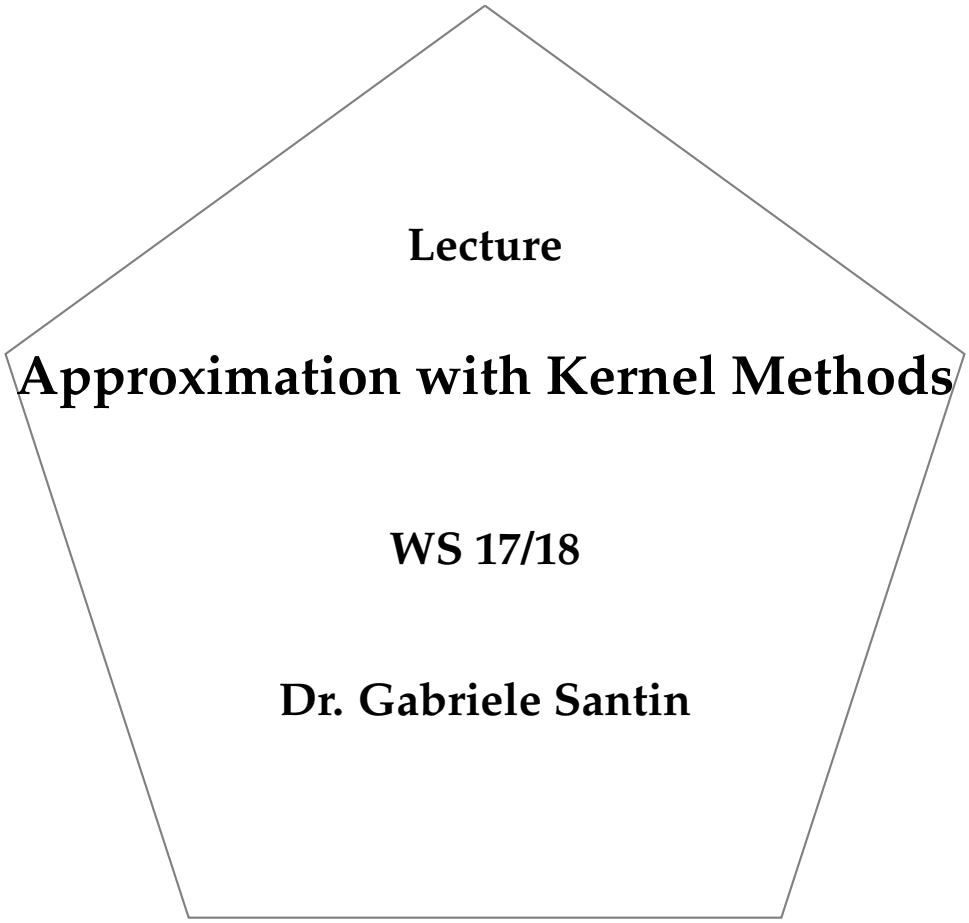

Lecture

Approximation with Kernel Methods

WS 17/18

Dr. Gabriele Santin

.

These notes are a collection of the material presented in the lecture "Approximation with Kernel Methods", WiSe 2017/2018, and are intended only as support material for the students attending the lecture.
A large part of the content is based on the lecture notes written by Prof. B. Haasdonk for the same lecture in WiSe 2015/2016, and in any case none of the content of this document has to be intended as original material.
Please, report any mistake found in these notes to G. Santin (santinge@mathematik.uni-stuttgart.de)

Organization

Contact:

- G. Santin, santinge@mathematik.uni-stuttgart.de
- Consultation hours on request

ILIAS/C@MPUS Site:

- Registration via C@MPUS
- Lecture material: announcements and info, lecture notes, demos from lecture
- Exercises: exercise sheets, electronic submission

Exercises:

- Participation by first electronic submission
- 4 theory exercises, 1 week solution time, 3 programming exercises, $3-4$ weeks solution time
- One exercise discussion every two weeks, on Wed.
- Corrected as "Votier-Übung"
- Single solution, no group submission
- Successful exercises if $\geq 50\%$, plus ≥ 1 presentation. Programming exercises should be running (any language is ok), and results should be explained
- First exercise published Mon. 16.10., submission Mon. 23.10., discussion Wed. 25.10.

Exam:

- The official responsible of the lecture/exam is Prof. B. Haasdonk
- Successful exercises give 0.3 grading bonus in the exam
- The exam will be oral, 30 minutes
- Exam date and registration: to be announced, via C@MPUS

Other data:

- $3+1$ SWS

- Knowledge on Numerik I is a good prerequisite, no other requested

Basic references:

- H. Wendland. Scattered Data Approximation, volume 17 of Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2005
- B. Schölkopf and A. J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, 2002

Further texts:

- G. E. Fasshauer. Meshfree Approximation Methods with MATLAB, volume 6 of Interdisciplinary Mathematical Sciences. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2007. With 1 CD-ROM (Windows, Macintosh and UNIX)
- I. Steinwart and A. Christmann. Support Vector Machines. Information Science and Statistics. Springer, New York, 2008

Contents

1. Introduction and motivation

This chapter serves as a collection of motivations for the study of kernel-based methods and as an introduction of the topic covered in the course. The presentation is intentionally not always rigorous, since the various details will be discussed in the following chapters.

1.1 Definitions of kernel and positivity classes

We start by defining the fundamental object of this course.

Definition 1.1. *Let Ω be a nonempty set. A real valued kernel on Ω is a symmetric function $K : \Omega \times \Omega \to \mathbb{R}$. A complex valued kernel on Ω is an hermitian function $K : \Omega \times \Omega \to \mathbb{C}$.*

The distinction between real and complex valued kernels is important, since they satisfy different symmetry properties and they lead to a different analysis. But we will use complex valued kernel only for theoretical results, so we will always assume that kernels are real valued if not explicitly stated.

An important fact is that Ω can be a general set (e.g., a set of strings or graphs or images). This is particularly interesting in the case of pattern analysis, while the main results for interpolation applies to $\Omega \subset \mathbb{R}^d$.

We will consider special classes of kernels, defined as follows.

Definition 1.2 (Positivity classes)**.** *Let Ω be a nonempty set. For all $N \in \mathbb{N}$ and for a set of N pairwise distinct elements $X_N := \{x_i\}_{i=1}^N \subset \Omega$ define the kernel matrix (or Gramian matrix) $A := A_{K,X_N} \in \mathbb{R}^{N\times N}$ as $A := [K(x_i, x_j)]_{i,j=1}^N$.*

A kernel K on Ω is positive definite (PD) on Ω if for all $N \in \mathbb{N}$, for any set of N pairwise distinct elements $X_N := \{x_i\}_{i=1}^N \subset \Omega$ the kernel matrix is positive semidefinite, i.e., for all vectors $\alpha := \{\alpha_i\}_{i=1}^N \in \mathbb{R}^N$ it holds

$$\alpha^T A\alpha = \sum_{i,j=1}^{N} \alpha_i\alpha_j K(x_i, x_j) \geq 0.$$

The kernel is strictly positive definite (SPD) if the kernel matrix is positive definite, i.e., the inequality holds with "$>$" when $\alpha \neq 0$.

Notation warning: some authors (including H. Wendland) denote as "positive definite" what we call here "strictly positive definite". This is because positive definite kernels are useful mainly in pattern analysis, so sometimes there is no need to stress the difference if dealing only with interpolation or PDE solution.

There is also another interesting positivity class, namely conditionally (strictly) positive definite kernels (C(S)PD), but we will introduce it later.

1.2 Why we are interested in kernels

We look at two class of problems which lead quite naturally to the use of kernels.

1.2.1 Multivariate scattered (or meshless) interpolation

Problem 1.3. *Let $\Omega \subset \mathbb{R}^d$. Given a set $X_N := \{x_i\}_{i=1}^N \subset \Omega$ of pairwise distinct points and target values $\{f_i\}_{i=1}^N \subset \mathbb{R}$, find a continuous function $s : \Omega \to \mathbb{R}$ such that $s(x_i) = f_i$, $1 \leq i \leq N$.*

Here *multivariate* means that we want to deal with generic $d \geq 1$, while *meshless* or *scattered* means that the points X_N can be arbitrarily unstructured (e.g., they don't need to form a grid).

The problem is well understood if $d = 1$ (univariate interpolation). In this case different techniques are possible: e.g. (see Numerik I) Problem 1.3 can be solved by polynomial interpolation of degree $N - 1$. The process works as follows: we can fix an approximation space $V := \mathbb{P}_{N-1}$ (which is a linear, N-dimensional space with $V \subset \mathcal{C}(\mathbb{R})$) and look for an interpolant $s \in V$. If we fix a basis $\{\phi_j\}_{j=1}^N$ of V, e.g. the monomial basis $\phi_j(x) := x^{j-1}$, $1 \leq j \leq N$, we have that

$$s(x) := \sum_{j=1}^{N} \alpha_j \phi_j(x), \tag{1.1}$$

where the coefficients $\{\alpha_j\}_{j=1}^N$ can be computed from the N interpolation conditions of Problem 1.3, that is

$$s(x_i) = f_i, 1 \leq i \leq N. \tag{1.2}$$

The ansatz (1.1) and the interpolation conditions (1.2) can be put together to obtain a linear system

$$A_{\phi,X_N}\alpha := \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_N(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \dots & \phi_N(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_1(x_N) & \phi_2(x_N) & \dots & \phi_N(x_N) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_N \end{bmatrix}.$$

In this case (polynomial interpolation in $d = 1$ using the monomial basis) the *interpolation matrix* A_{ϕ,X_N} is the Vandermonde matrix, i.e.,

$$A_{\phi,X_N} = \begin{bmatrix} 1 & x_1 & \dots & x_1^{N-1} \\ 1 & x_2 & \dots & x_2^{N-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & \dots & x_N^{N-1} \end{bmatrix}$$

which is known to be invertible for any arbitrary set X_N of pairwise distinct points.

Putting all together: in the case $d = 1$ it is possible to solve Problem 1.3. In details: it is possible to do the following

- fix a suitable linear space $V \subset \mathcal{C}(\Omega)$
- prove that there exists a unique $s \in V$ with $s(x_i) = f_i$, $1 \leq i \leq N$, for all possible X_N and $\{f_i\}_{i=1}^N$
- the function s can be computed by solving a linear system
- the linear system has a unique solution since A_{Φ,X_N} is invertible for all X_N.

This space V is the prototype of an Haar space as in the following definition.

Definition 1.4 (Haar space). *Let $\Omega \subset \mathbb{R}^d$ contain at least N points and $V \subset \mathcal{C}(\Omega)$ be an N-dimensional linear space. Then V is called an Haar space of dimension N on Ω if for arbitrary distinct points $\{x_i\}_{i=1}^N \subset \Omega$ and arbitrary $\{f_i\}_{i=1}^N \subset \mathbb{R}$, there exist a unique function $s \in V$ with $s(x_i) = f_i$ for $1 \leq i \leq N$.*

We first see that this notion indeed corresponds to the above discussion (Property (ii) is stated for completeness).

Proposition 1.5. *Under the assumptions of Definition 1.4, the following are equivalent:*

i) V is an N-dimensional Haar space on Ω;

ii) Every $v \in V \setminus \{0\}$ has at most $N-1$ distinct zeros;

iii) For any set of pairwise distinct points $X_N \subset \Omega$ and any basis $\{\phi_j\}_{j=1}^N$ of V, the interpolation matrix A_{ϕ,X_N} is invertible.

Proof. (i)⇒(ii) Assume that V is an Haar space and that there exists $v \in V\setminus\{0\}$ with N distinct zeros $X_N \subset \Omega$, and define $u := 0$ ($u \in V$ since V is a linear space). Then both u and v are distinct interpolants in V with target values $f_i := 0$, $1 \leq i \leq 0$, which is a contradiction to the definition of Haar space.

(ii)⇒(iii) Assume there exists a set $X_N \subset \Omega$ of pairwise distinct points and a basis $\{\phi_j\}_{j=1}^N$ of V such that A_{ϕ,X_N} is singular. Then there exists a vector $\alpha \in \mathbb{R}^N \setminus \{0\}$ such that $A_{\phi,X_N}\alpha = 0$. Since $\alpha \neq 0$, the function $v(x) := \sum_{j=1}^N \alpha_j \phi_j(x)$ is not the zero function, but $v(x_i) = \sum_{j=1}^N \alpha_j \phi_j(x_i) = (A_{\phi,X_N}\alpha)_i = 0$ (i-th row), $1 \leq i \leq N$, so v has N distinct zeros.

(iii)⇔(i) A_{ϕ,X_N} is nonsingular if and only if $A_{\phi,X_N}\alpha = b$ has a unique solution α for $b := [f_i]_{i=1}^N$, if and only if $\sum_{j=1}^N \alpha_j \phi_j(x_i) = f_i$ has a unique solution α, if and only if $s(x) := \sum_{j=1}^N \alpha_j \phi_j(x)$ is the unique interpolant.

□

Coming back to Problem (1.3), it seems natural to ask: is there any Haar space in dimension $d > 1$? How to characterize them? How to choose a suitable one? Indeed, such spaces don't even exist!

Theorem 1.6 (Mairhuber-Curtis). *Let $\Omega \subset \mathbb{R}^d$, $d > 1$, be a set with nonempty interior (i.e., there exists $x_0 \in \Omega$ and $\epsilon > 0$ such that $B(x_0, \epsilon) \subset \Omega$). Then there exist no Haar space of dimension $N > 1$ on Ω.*

Proof. The proof shows that for any N-dimensional space $V \subset \mathcal{C}(\Omega)$ there exists a set of pairwise distinct points $X_N \subset \Omega$ such that the interpolation matrix is singular, contradicting Property (iii) of Proposition (1.5). To show this, assume $V := \text{ span } \{\phi_1, \dots, \phi_N\} \subset \mathcal{C}(\Omega)$ is an Haar space of dimension N on Ω and consider a set of pairwise distinct points $X_N \subset \Omega$ such that $x_1, x_2 \subset B(x_0, \epsilon)$. Since V is an Haar space, by property (iii) of Proposition 1.5 we have $\det(A_{\phi, X_N}) \neq 0$.

Noe define continuous and simple curves (i.e., no self intersections) $\gamma_1, \gamma_2 : [0, 1] \to B(x_0, \epsilon)$ with $\gamma_1(0) = x_1$, $\gamma_1(1) = x_2$, $\gamma_2(0) = x_2$, $\gamma_2(1) = x_1$, and such that $\gamma_1(t) \neq \gamma_2(t), \gamma_i(t) \neq x_3, \dots, x_N$, $t \in [0, 1]$. Then $X_N(t) := \{\gamma_1(t), \gamma_2(t), x_3, \dots, x_N\}$ are distinct for all $t \in [0, 1]$.

Define as $D(t) := \det(A_{\phi, X_N(t)})$ the determinant of the corresponding matrix. We have $D(0), D(1) \neq 0$ by Property (ii), but $D(0)D(1) < 0$ (first two rows are permuted), and $D(t)$ is continuous in t, then there exists $\bar{t} \in [0, 1]$ with $D(\bar{t}) = 0$. So $X_N(\bar{t})$ is a set of pairwise distinct points with $D(\bar{t}) = 0$, thus V is not an Haar space. □

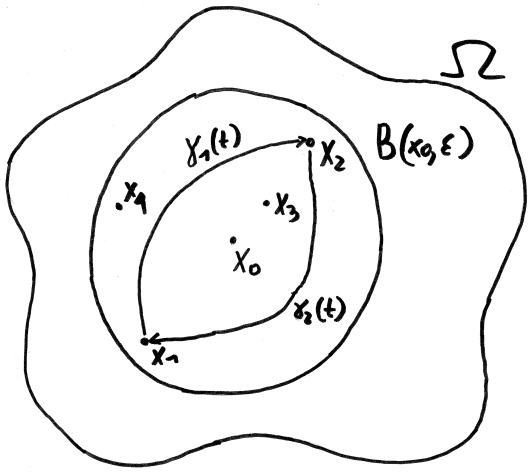

Figure 1.1: Illustration of the proof of the Mairhuber-Curtis Theorem

This means that the space V can not be chosen a-priori, i.e., it needs to be dependent on the particular points X_N. Here is where kernels come into play: we can consider a continuous and strictly positive definite kernel $K : \Omega \times \Omega \to \mathbb{R}$ and use the second variable of the kernel to generate a data-dependent basis and set as

$$V := V(X_N) := \text{ span } \{\phi_j(x) := K(x, x_j), 1 \leq j \leq N\}.$$

The interpolant is now of the form

$$s(x) := \sum_{j=1}^{N} \alpha_j \phi_j(x) = \sum_{j=1}^{N} \alpha_j K(x, x_j),$$

and the interpolation matrix is

$$A_{K,X_N} := \begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_N(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \dots & \phi_N(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_1(x_N) & \phi_2(x_N) & \dots & \phi_N(x_N) \end{bmatrix} = \begin{bmatrix} K(x_1, x_1) & \dots & K(x_1, x_N) \\ K(x_2, x_2) & \dots & K(x_2, x_N) \\ \vdots & \vdots & \vdots \\ K(x_N, x_1) & \dots & K(x_N, x_N) \end{bmatrix}.$$

Thank to the definition 1.2 of strictly positive definite kernel, this matrix is positive definite. This means (we will see this in more details) that it is also always invertible.

Theorem 1.7 (Kernel interpolation is well defined)**.** *Let $\Omega \subset \mathbb{R}^d$ and K a SPD kernel on Ω. Given any set $X_N := \{x_i\}_{i=1}^N \subset \Omega$ of pairwise distinct points and target values $\{f_i\}_{i=1}^N \subset \mathbb{R}$, there exists a unique kernel interpolant*

$$s(x) := \sum_{j=1}^{N} \alpha_j K(x, x_j),$$

with $s(x_i) = f_i$ for $1 \leq i \leq N$.

1.2.2 Mapping linear algorithms in high dimensional spaces

Problem 1.8. *Let Ω be a set. Consider a set of data $X_N := \{x_i\}_{i=1}^N \subset \Omega$ with $X_N := X_+ \cup X_-$, representing elements of two classes (positive and negative). Find a function*

$$f(x) = \begin{cases} 1, & x \in X_+ \\ -1, & x \in X_- \end{cases}$$

which classifies the points in the two classes.

If $\Omega \subset \mathbb{R}^d$ and the points are linearly separable, there exists a function $f(x) := (w, x) + b$ with $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that

$$f(x) := (w, x) + b \begin{cases} > 0, & x \in X_+ \\ < 0, & x \in X_- \end{cases}. \tag{1.3}$$

This function is named a classifier, and it is not unique. We will study an algorithm called linear Support Vector Machine (linear SVM) that computes w, b efficiently and such that the "separation margin" between the two classes is maximized.

Observe that, if X_N is large enough and the points are not all linear dependent, we can assume that $w := \sum_{j=1}^{N} \alpha_j x_j$. Thus

$$f(x) := (w, x) + b = \sum_{j=1}^{N} \alpha_j (x_i, x_j) + b, \tag{1.4}$$

i.e., the classifier can be expressed in terms of inner products between the data. This is the case of any other linear algorithm, if w can be expressed in this way.

What can we do in the case the two classes are not linearly separable? A possible solution is to map the data into an higher dimensional space, i.e., consider a feature map $\phi : \Omega \to H$ into an Hilbert space H called the feature space, and apply the same classification algorithm in this new space.

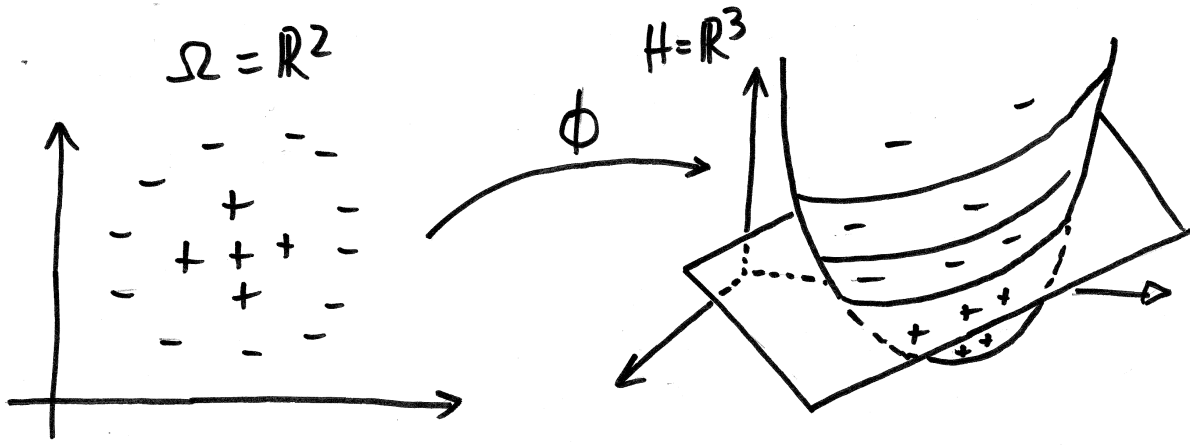

Figure 1.2: Example of a feature map from $\Omega := \mathbb{R}^2$ to the feature space $H := \mathbb{R}^3$.

The hope is that the data are linearly separable in the high dimensional space. This means that we now consider $w \in H$, $b \in \mathbb{R}$ and the classifier (1.3) is now of the form

$$f(x) := (w, \phi(x))_H + b.$$

If we use the same assumption on w, we have $w := \sum_{j=1}^{N} \alpha_j \phi(x_j)$, then (1.4) becomes

$$f(x) := (w, x)_H + b = \sum_{j=1}^{N} \alpha_j (\phi(x_i), \phi(x_j))_H + b. \tag{1.5}$$

The same algorithm (in this case linear SVM) can be applied here, but considering inner products $(\phi(x_i), \phi(x_j))_H$ of the transformed data. The nice thing is that the function $K(x, y) := (\phi(x), \phi(y))_H$ is indeed a kernel!

Theorem 1.9 (Kernels induced by feature maps)**.** *Let Ω be a set, H an Hilbert space and $\phi : \Omega \to H$ a feature map. Then the function $K(x, y) := (\phi(x), \phi(y))_H$ is a positive definite kernel.*

Proof. It suffices to verify the condition of definition 1.2: let $N \in \mathbb{N}$, $X_N \in \Omega$ pairwise distinct and $\alpha \in \mathbb{R}^N$, then

$$\sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) = \sum_{i,j=1}^{N} \alpha_i \alpha_j (\phi(x_i), \phi(x_j))_H = \left(\sum_{i}^{N} \alpha_i \phi(x_i), \sum_{j=1}^{N} \alpha_j \phi(x_j) \right)_H$$
$$= \left\| \sum_{i=1}^{N} \alpha_i \phi(x_i) \right\|_H^2 \geq 0,$$

since the norm is positive definite.

Observe that strict positive definiteness can not be concluded because $\{\phi(x_i)\}_{j=1}^{N}$ can be in general linearly dependent. □

This means that the mapping of algorithms expressed in terms of inner products to higher dimensional spaces leads naturally to the use of positive definite kernels.

Some consequences and comments on the Theorem:

- If we directly consider $K(x, y) = (\phi(x), \phi(x))_H$, we see that there is no need to assume that $\Omega \subset \mathbb{R}^d$, but instead we can define a feature map, then also a kernel, on general sets Ω
- If $K(x, y)$ is known explicitly, the inner products in the high dimensional space can be computed implicitly by just evaluating K, which is in general a much cheaper operation (called the kernel trick).
- At this stage there is no need to assume that the space H is an Hilbert space, since the proof would work for any inner product space. But we will see in the following that, if a feature map exists, then there exists also another one with values in an Hilbert space. So this assumption is not restrictive.

1.3 Questions to be addressed

This preliminary discussion motivates some questions that will be addressed during this course:

- How can kernels be constructed and characterized?
- What are common properties of all kernels?

- What is the relation between general kernels and kernels obtained from feature maps?
- The kernel interpolation, so far, is just a "join the dots" operation: what kind of functions can be really approximated? How well can functions be approximated? (and the same for classification)
- How can this approach be adapted to solve PDEs?
- What kind of algorithms can be used to efficiently solve kernel approximation problems?
- How to choose in practice a suitable kernel, its parameters, the data/points?
- How to measure the performances of an algorithm? (test/train, cross-validation, ...)

2. Basic properties and examples of kernels

We recall the positivity classes seen in the previous chapter.

Definition 2.1 (Positivity classes)**.** *Let Ω be a nonempty set. For all $N \in \mathbb{N}$ and for a set of N pairwise distinct elements $X_N := \{x_i\}_{i=1}^N \subset \Omega$ define the kernel matrix (or Gramian matrix) $A := A_{K,X_N} \in \mathbb{R}^{N\times N}$ as $A := [K(x_i, x_j)]_{i,j=1}^N$.*

A kernel K on Ω is positive definite (PD) on Ω if for all $N \in \mathbb{N}$, for any set of N pairwise distinct elements $X_N := \{x_i\}_{i=1}^N \subset \Omega$ the kernel matrix is positive semidefinite, i.e., for all vectors $\alpha := \{\alpha_i\}_{i=1}^N \in \mathbb{R}^N$ it holds

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

The kernel is strictly positive definite (SPD) if the kernel matrix is positive definite, i.e., the inequality holds with "$>$" when $\alpha \neq 0$.

2.1 Criteria for (S)PD

We see now some basic criteria to check if a kernel is (S)PD. They are properties of the associated kernel matrix, while we will later see more sophisticated criteria to directly conclude (strict) positive definiteness from the kernel itself.

We start by recalling the following fact from linear algebra.

Theorem 2.2 (LDU decomposition)**.** *Let $A \in \mathbb{R}^{N\times N}$. Assume that the leading principal submatrices of A are non singular, i.e., $\det(A_n) \neq 0$ for all $1 \leq n \leq N$, where $A_n := [A_{ij}]_{ij=1}^n$.*

Then there exists a unique LDU decomposition

$$A = LDU$$

with L lower triangular, U upper triangular, D diagonal, $\mathrm{diag}(L) = \mathrm{diag}(U) = [1, \dots, 1]^T$ and $D_{ii} \neq 0$, $1 \leq i \leq N$.

If A is also symmetric, then $U = L^T$, i.e. $A = LDL^T$.

Proof. See Linear Algebra courses or [2, Section 4.1]. The idea is roughly to apply Gauss elimination in a proper way. □

Proposition 2.3 (Criteria for positive definiteness)**.** *Let $A \in \mathbb{R}^{N\times N}$ be a symmetric matrix. The following are equivalent:*

i) A is PD, i.e., $\alpha^T A\alpha > 0$ for all $\alpha \in \mathbb{R}^N$;

ii) A has a unique Cholesky decomposition, i.e., $A = LL^T$ with L lower triangular with $L_{ii} > 0$;

iii) The eigenvalues $\{\lambda_i\}_{i=1}^N$ of A are positive;

iv) The leading principal minors of A are positive, i.e., $\det(A_n) > 0$ for all $1 \leq n \leq N$, where $A_n := [A_{ij}]_{ij=1}^n$. In particular A is invertible with $\det(A) > 0$.

Proof. First, consider the eigen-decomposition $A = V\Lambda V^T$, with $V, \Lambda \in \mathbb{R}^{N\times N}$, $V^TV = VV^T = I$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N)$, and denote as V_i the columns of V (this decomposition exists because A is symmetric).

(iii)⇒(i) Since the columns of V are an o.n.b. of $\mathbb{R}^N$, $0 \neq \alpha = \sum_{i=1}^N V_i\beta_i$ with $\beta \in \mathbb{R}^N \setminus \{0\}$, i.e., $\alpha = V\beta$. Then

$$\begin{aligned}\alpha A\alpha &= \left(\beta^TV^T\right) A \left(V\beta\right) = \beta^TV^T \left(V\Lambda V^T\right) V\beta = \beta^T \left(V^TV\right) \Lambda \left(V^TV\right) \beta \\ &= \beta^T\Lambda\beta = \sum_{i=1}^N \beta_i^2\lambda_i,\end{aligned}$$

thus $\alpha A\alpha > 0$ if $\lambda_i > 0$.

(i)⇒(iii) $\alpha^TA\alpha > 0$ for all $\alpha \neq 0$, and in particular for $\alpha = V_i$ we obtain $0 < V_i^TAV_i = \lambda_iV_i^TV_i = \lambda_i$.

(i)⇒(iv) $\alpha^TA\alpha > 0$ for all $\alpha \neq 0$, and in particular for $\alpha := [\alpha_n^T, 0^T]^T$ with $\alpha_n \in \mathbb{R}^n$ we have $0 < \alpha^TA\alpha = \alpha_n^TA_n\alpha_n$ so A_n is PD. Then the determinant $\det(A_n) = \prod_{i=1}^n \lambda_i(A_n) > 0$ by (iii).

We then have

(ii)⇒(i) L is invertible since $\det(L) = L_{11} \ldots L_{NN} > 0$, so $L^T\alpha \neq 0$ if $\alpha \neq 0$. It follows that

$$\alpha^TA\alpha = (\alpha^TL)(L^T\alpha) = \|L^T\alpha\|^2 > 0.$$

(iv)⇒(ii) By Theorem 2.2, each A_n has a unique LDU decomposition, so in particular ($n = N$) $A = LDL^T$ is unique. We prove that D has positive diagonal, from which it follows that $A = (L\sqrt{D})(\sqrt{D}L^T)$ is a Cholesky decomposition of A, since $L\sqrt{D}$ is a lower triangular matrix with positive diagonal.

To see this, we first prove that the principal $n \times n$ submatrix of L and D give the unique LDU decomposition of A_n, for all $1 \leq n \leq N-1$.

Indeed, for all n ("$*$" denotes a generic vector or scalar)

$$\begin{aligned}\begin{bmatrix} A_{n-1} & * \\ * & * \end{bmatrix} = A_n = L_nD_nU_n &= \begin{bmatrix} L_{n-1} & 0 \\ * & * \end{bmatrix} \begin{bmatrix} D_{n-1} & 0 \\ 0^T & * \end{bmatrix} \begin{bmatrix} L_{n-1}^T & * \\ 0^T & * \end{bmatrix} \\ &= \begin{bmatrix} L_{n-1}D_{n-1} & 0 \\ * & * \end{bmatrix} \begin{bmatrix} L_{n-1}^T & * \\ 0^T & * \end{bmatrix} = \begin{bmatrix} L_{n-1}D_{n-1}L_{n-1}^T & * \\ * & * \end{bmatrix},\end{aligned}$$

so $A_{n-1} = L_{n-1}D_{n-1}L_{n-1}^T$ is a LDU decomposition of A_{n-1} (and so it is the unique one).

It follows that $\det(A_n) = \det(D_n) = D_{11}D_{22}\dots D_{nn}$, and since $\det(A_n) > 0$ by hypothesis, we have

$$\begin{array}{lll} 0 < \det(A_1) & = D_{11} & \Rightarrow \quad D_{11} > 0 \\ 0 < \det(A_2) & = D_{11}D_{22} & \Rightarrow \quad D_{22} > 0 \\ \vdots & \vdots & \vdots \quad \vdots \\ 0 < \det(A) & = D_{11}\dots D_{NN} & \Rightarrow \quad D_{NN} > 0. \end{array}$$

Finally, to prove that this Cholesky decomposition is unique, assume there exists another one with matrix L'. Since L' has positive diagonal, we can rewrite $A = L'L'^T$ as

$$A_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \frac{L'_{21}}{L'_{11}} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \frac{L'_{N1}}{L'_{11}} & \frac{L'_{N2}}{L'_{22}} & \dots & 1 \end{bmatrix} \begin{bmatrix} L'^2_{11} & 0 & \dots & 0 \\ 0 & L'^2_{22} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & L'^2_{NN} \end{bmatrix} \begin{bmatrix} 1 & \frac{L'_{12}}{L'_{11}} & \dots & \frac{L'_{1N}}{L'_{11}} \\ 0 & 1 & \dots & \frac{L'_{2N}}{L'_{22}} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix},$$

so we obtain another LDU decomposition, which is unique, so $L' = L$.

□

Remark 2.4. *This Proposition almost applies also for positive semidefinite matrices by replacing $>$ with $\geq$, and in this case* (i)⇔(iii)⇒(iv)*. But the matrix*

$$A := \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$$

satisfies $e_2^T A e_2 < 0$ even if the leading principal minors are non-negative.

In the case of PD kernels, hence of positive semidefinite matrices, (iv) *can instead be replaced by the following:*

iv) The principal minors of A are non negative, i.e., $\det(\tilde{A}) \geq 0$ for all principal submatrix $\tilde{A}$ of A, where a principal submatrix is a matrix obtained by removing the same rows and columns.

Indeed, using this new characterization the matrix A of the above example can be proven to be not positive semidefinite since the bottom-right submatrix has negative determinant.

Moreover, a decomposition $A = LL^T$ still exists, but it is not unique and it possible that $L_{ii} = 0$ for some i.

Recall that we have already seen another characterization of PD kernels in Theorem 1.9, which does not require any computation with the kernel matrix.: If there exists an Hilbert space H and a feature map $\phi : \Omega \to H$ such that $K(x, y) = (\phi(x), \phi(y))_H$, then K is a PD kernel.

2.2 Properties of (S)PD kernels

The following are useful properties of any PD kernel.

Proposition 2.5. *Let $K : \Omega \times \Omega \to \mathbb{R}$ be a PD kernel, then*

i) $K(x,x) \geq 0$ *for all* $x \in \Omega$ *(non-negativity of the diagonal)*

ii) $K(x,y)^2 \leq K(x,x)K(y,y)$ *(Cauchy-Schwarz inequality).*

If K is SPD then "$>$" holds.

Proof. (i) Let $N := 1$, $X_1 := \{x\}$, $\alpha := 1$. Positive definiteness implies $\alpha A \alpha = \alpha^2 K(x,x) \geq 0$, so $K(x,x) \geq 0$ for all x.

(ii) Use $N := 2$, $X_2 := \{x, y\}$, i.e.,

$$A = \begin{bmatrix} K(x,x) & K(x,,y) \\ K(y,x) & K(y,y) \end{bmatrix},$$

so $\det(A) = K(x,x)K(y,y) - K(x,y)^2 \geq 0$ (by Property (iv)).
In both cases, the same argument proves "$>$" if K is SPD. □

We will also need the fact that kernel matrices of not pairwise distinct points are still positive semidefinite.

Proposition 2.6. *Let K be PD or SPD and X_N be a set of points, not necessarily pairwise distinct. Then the kernel matrix $A := [K(x_i, x_j)]_{i,j=1}^N$ is positive semidefinite.*

Proof. If the points are pairwise distinct the statement follows from the definition of PD and SPD kernels.

If instead there are duplicated points, we use point iv) of Remark 2.4 and prove that every principal minor of A has non negative determinant. Indeed, every principal minor is defined by removing from A rows and columns with indexes $I \subset \{1, \ldots, N\}$. If I is such that $X_N(I) := \{x_i \in X_N : i \notin I\}$ are pairwise distinct, then since the submatrix is the kernel matrix of $X_N(I)$, it is positive semidefinite. If instead $X_N(I)$ are not pairwise distinct, then the submatrix contains two equal columns, so its determinant is zero.

□

2.3 Basic operations on kernels

It is often useful to obtain a new kernel by performing basic operations on other kernels. Moreover, these operations can be used to prove that a kernel is PD, if it is possible to prove that it is generated from operations on other kernels which are known to be PD. We collect in the following Proposition some of them.

Proposition 2.7 (Basic operations on kernels). *Let Ω be a set, $K_1, K_2 : \Omega \times \Omega \to \mathbb{R}$ be PD kernels, $f : \Omega \to \Omega$, $g : \Omega \to \mathbb{R}$, $\Omega' \subset \Omega$, and x, y generic elements in Ω. Then the following are PD kernels:*

i) $K : \Omega' \times \Omega' \to \mathbb{R}$ *with* $K := (K_1)_{|_{\Omega' \times \Omega'}}$ *(PD on Ω')*

ii) $K(x, y) := K_1(x, y) + K_2(x, y)$

iii) $K(x, y) := aK_1(x, y)$ *if* $a \geq 0$

iv) $K(x, y) := K_1(x, y)K_2(x, y)$

v) $K(x, y) := \exp(K_1(x, y))$

vi) $K(x, y) := K_1(f(x), f(y))$

vii) $K(x, y) := g(x)g(y)$

viii) $K(x, y) := g(x)K_1(x, y)g(y)$

ix) $K(x, y) := h(K_1(x, y))$ *with* $h(z) := \sum_{i=0}^{\infty} a_i z^i$, $a_i \geq 0$, *with radius of convergence* $\rho > 0$ *and* $|K_1(x, y)| \leq \rho$.

Proof. The points (ii) – (v) are exercises.

(i) It follows just by applying Definition 1.2.

(vi) For a set $X_N \subset \Omega$, define $X'_N := \{x'_i := f(x_i), x_i \in X_N\}$. Then

$$A := [K_1(f(x_i), f(x_j))]_{ij=1}^N = [K_1(x'_i, x'_j)]_{ij=1}^N$$

is positive semidefinite (even in the case X_N are not pairwise distinct, see Proposition 2.6).

(vii) Here

$$A = \begin{bmatrix} g(x_1)g(x_1) & \dots & g(x_1)g(x_N) \\ g(x_2)g(x_1) & \dots & g(x_2)g(x_2) \\ \vdots & \vdots & \vdots \\ g(x_N)g(x_1) & \dots & g(x_N)g(x_N) \end{bmatrix} = \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_N) \end{bmatrix} [g(x_1), \dots, g(x_N)] =: \bar{g}\bar{g}^T,$$

thus $\alpha^T A \alpha = \alpha^T \bar{(}\bar{g}\bar{g}^T)\alpha = (\alpha^T \bar{g})(\bar{g}^T \alpha) = (\bar{g}^T \alpha)^2 \geq 0$.

(viii) It follows by combining (iv) and (vii).

(ix) The kernel obtained from the finite sum $K_m(x, y) := \sum_{i=0}^m a_i K_1(x, y)^i$ is PD thanks to (ii), (iii), (iv). Define as A_m the corresponding kernel matrix. Then $\alpha^T A \alpha = \lim_{m\to\infty} \alpha^T A_m \alpha \geq 0$, since the argument of the limit is positive and the series converges for $|K_1(x, y)| \leq \rho$.

□

Remark 2.8. *To obtain strictly PD kernels one needs to assume the following:*

i) K_1 *is SPD*

ii) K_1 *or* K_2 *is SPD*

iii) K_1 *is SPD and* $a > 0$

iv) K_1 *and* K_2 *are SPD*

vi) K_1 *is SPD and* f *is injective*

viii) K_1 *is SPD and* $g(x) \neq 0$ *for all* x

One other particular way to generate kernels is to combine kernels defined on lower dimensional spaces. This is useful for example when the data are tuples where each component represents a different object.

Proposition 2.9 (Kernels for product spaces)**.** *Let* $d \geq 1$. *Consider the set* $\Omega := \Omega_1 \times \Omega_2 \times \cdots \times \Omega_d$ *for arbitrary sets* Ω_l *and denote* $x \in \Omega$ *as* $x := (x^{(l)})_{l=1}^d$ *with* $x^{(l)} \in \Omega_l$.
Let $K_l : \Omega_l \times \Omega_l \to \mathbb{R}$ *be a (S)PD kernel on* Ω_l.
Then the following are (S)PD kernels on Ω*:*

i) $K(x, y) := \sum_{l=1}^d K_l\left(x^{(l)}, y^{(l)}\right)$.

ii) $K(x, y) := \prod_{l=1}^d K_l\left(x^{(l)}, y^{(l)}\right)$

Proof. Denote as A_l the kernel matrix of K_l.

(i) For all i, j we have $K(x_i, y_j) := \sum_{l=1}^d K_l\left(x_i^{(l)}, y_j^{(l)}\right)$, so we can apply Property (ii) of Proposition (2.7). Observe that the kernel matrix satisfies $A = A_1 + A_2 + \cdots + A_N$.

(ii) With the same idea we obtain $A = A_1 \circ A_2 \circ \cdots \circ A_N$ (Hadamard or pointwise product). The result that the Hadamard product of positive definite matrices is positive definite is known as Schur's Product Theorem.

□

2.4 Examples of kernels

With the tools of the previous sections, we can now consider some relevant examples of kernels and prove that they are positive definite. We first consider some examples for $\Omega \subset \mathbb{R}^d$.

Proposition 2.10. *We have the following for* $\Omega \subset \mathbb{R}^d$*:*

i) $K(x,y) := (x,y)$ *(linear kernel) is PD*

ii) $K(x,y) := ((x,y)+a)^p$, $a \geq 0$, $p \in \mathbb{N}$ *(polynomial kernel) is PD*

iii) $K(x,y) := \exp(-\varepsilon^2\|x-y\|^2)$, $\varepsilon > 0$ *(Gaussian kernel) is SPD.*

Proof. (i) There exist a feature map, so we can use Theorem 1.9: With $H := \mathbb{R}^d$, $\phi(x) := x$ we have $K(x,y) = (\phi(x), \phi(y))_H$.

(ii) The kernel $K(x,y) := a$ is PD since $\alpha A \alpha = a\left(\sum_{i=1}^d \alpha_i\right)^2 \geq 0$. Thus the polynomial kernel can be obtained as sums and products of PD kernels (since the linear kernel is PD) and from Proposition 2.7 we conclude that it is PD.

(iii) We prove the statement for $d = 1$. For $d > 1$ we have

$$\exp(-\varepsilon^2\|x-y\|^2) = \exp\left(-\sum_{i=1}^d \varepsilon^2 (x^{(i)} - y^{(i)})^2\right) = \prod_{i=1}^d \exp\left(-\varepsilon^2 (x^{(i)} - y^{(i)})^2\right),$$

so from the case $d = 1$ it follows that the Gaussian is a product of SPD kernels, hence SPD by Property (iv) of Proposition 2.7 and the corresponding Remark.

For $d = 1$, recall (see Analysis I/II/III) that the Fourier transform is defined as

$$F(f)(\omega) := \int_{\mathbb{R}} e^{-ix\omega} f(x) dx, \ \omega \in \mathbb{R}.$$

We first prove that the Gaussian kernel is the Fourier transform of a positive function. Indeed, if $f_\sigma(x) := e^{-\frac{x^2}{2\sigma^2}}$ it can be proven that $F(f_\sigma)(\omega) = \sqrt{2\pi}\,\sigma e^{-\frac{1}{2}\omega^2\sigma^2}$. Taking $\sigma := \sqrt{2}\varepsilon$, we obtain $F(f_\sigma)(\omega) = 2\sqrt{\pi}\varepsilon e^{-\varepsilon^2\omega^2}$, so for any $\omega_i, \omega_i \in \mathbb{R}$

$$K(\omega_j, \omega_l) := e^{-\varepsilon^2(\omega_j - \omega_l)^2} = \frac{1}{2\sqrt{\pi}\varepsilon} F(f_\sigma)(\omega_j - \omega_l).$$

Now we prove PD: For $X_N := \{\omega_1, \ldots, \omega_N\} \subset \mathbb{R}^d$ and $0 \neq \alpha \in \mathbb{R}^N$ we have

$$\begin{aligned}
\alpha^T A \alpha &= \sum_{j,l=1}^N \alpha_j \alpha_l K(\omega_j, \omega_l) = \frac{1}{2\sqrt{\pi}\varepsilon} \sum_{j,l=1}^N \alpha_j \alpha_l F(f_\sigma)(\omega_j - \omega_l) \\
&= \frac{1}{2\sqrt{\pi}\varepsilon} \sum_{j,l=1}^N \alpha_j \alpha_l \int_{\mathbb{R}} e^{-ix(\omega_j - \omega_l)} f_\sigma(x) dx \\
&= \frac{1}{2\sqrt{\pi}\varepsilon} \int_{\mathbb{R}} \sum_{j,l=1}^N \alpha_j \alpha_l e^{-ix(\omega_j - \omega_l)} f_\sigma(x) dx
\end{aligned}$$

and using the fact that $e^{-ix(\omega_j - \omega_l)} = (e^{-ix\omega_j})\,\overline{(e^{-ix\omega_l})}$ we obtain

$$\alpha^T A \alpha = \frac{1}{2\sqrt{\pi}\varepsilon} \int_{\mathbb{R}} \left| \sum_{j=1}^N \alpha_j e^{-ix\omega_j} \right|^2 f_\sigma(x) dx \geq 0,$$

since both the functions in the integral are non negative.

To prove SPD, observe that for $\{w_j\}_{j=1}^N$ pairwise distinct, the functions $\{e^{-ix\omega_j}\}_{j=1}^N$ are linearly independent, so the argument of the integral is strictly positive and then the kernel is SPD.

□

Remark 2.11 (Gaussian kernel)**.** *The Gaussian kernel is an example of a Radial Basis Function (RBF) kernel. Indeed, it can be written as*

$$K(x,y) := \Phi(\varepsilon\|x-y\|)$$

with $\Phi : [0,\infty) \to \mathbb{R}$ and $\varepsilon > 0$ a shape parameter. We will study in details this kind of kernels, and we will see that the proof of SPD for general RBFs is very similar to the one for the Gaussian (so most of the work is done already).

These kernels are well studied (almost all the error analysis that we will see applies mainly to RBF kernels). Moreover, they are the most used kernels in practical applications, since they are really easy to compute. Indeed, one can compute a distance matrix $D \in \mathbb{R}^{N\times N}$ with $D_{ij} := \|x_i - x_j\|$, and obtain the kernel matrix just as $A = \Phi(\varepsilon D)$ (see demo in ILIAS).

Remark 2.12 (Polynomial kernel)**.** *The PD of the Polynomial kernel can be proven also by finding a suitable feature map. For example for $d = 2$, $p = 2$ the kernel is*

$$\begin{aligned} K(x,y) &= ((x,y)+a)^2 = (x^{(1)}y^{(1)} + x^{(2)}y^{(2)} + a)^2 \\ &= (x^{(1)})^2(y^{(1)})^2 + (x^{(2)})^2(y^{(2)})^2 + a^2 + 2x^{(1)}y^{(1)}x^{(2)}y^{(2)} + 2ax^{(1)}y^{(1)} + 2ax^{(2)}y^{(2)}, \end{aligned}$$

and a feature map $\phi : \mathbb{R}^2 \to \mathbb{R}^6$ is

$$\phi(x) = \left[a, \sqrt{2a}\, x^{(1)}, \sqrt{2a}\, x^{(2)}, (x^{(1)})^2, (x^{(2)})^2, \sqrt{2}\, x^{(1)}x^{(2)}\right]^T,$$

where 6 is the dimension of the space $\mathbb{P}_2(\mathbb{R}^2)$ of polynomial of degree $p = 2$ in $d = 2$ variables.

In general the kernel is

$$K(x,y) := ((x,y)+a)^p = \left(\sum_{i=1}^d x^{(i)}y^{(i)} + a\right)^p$$

which is a d-variate polynomial of degree p. In general it contains not all the monomial terms but only the ones with multi index $j := (j_1,\dots,j_d) \in J$, for a certain set $J \subset \mathbb{N}_0^d$. If $a > 0$ it contains all the monomials, so $m := |J| = \binom{d+p}{d} = \dim(\mathbb{P}_p(\mathbb{R}^d))$.

The kernel can be written as

$$K(x,y) = \sum_{j\in J} a_j x^j y^j,$$

for some positive numbers $\{a_j\}_{j\in J}$, and a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^m$ is

$$\phi(x) := [\sqrt{a_1}x^{j_1}, \dots, \sqrt{a_m}x^{j_m}]^T.$$

Observe that using the kernel instead of the feature map is very convenient (kernel trick, see remark after Theorem 1.9). Indeed, in this way we work with d dimensional instead of $\binom{d+p}{d}$ dimensional vectors.

This feature map representation proves also that the polynomial kernel is not SPD in general: e.g., if X_N contains N pairwise distinct points and $m > N$, then $\{\phi(x_i)\}_{i=1}^N$ can not be linear dependent. So the kernel is only positive definite (see proof of Theorem 1.9).

We then consider some examples of kernels for structured data.

Example 2.13 (Bag-of-words kernel)**.** *Let Ω be the set of all finite strings over an alphabet $\Sigma := \{\sigma_1, \sigma_2, \dots, \sigma_m\}$, i.e.,*

$$\Omega := \{\sigma_1, \sigma_2, \dots, \sigma_m, \sigma_1\sigma_2, \dots, \sigma_1\sigma_2\sigma_3, \dots\}.$$

Let $D = \{w_i\}_{i=1}^d$ be a dictionary of d elements, and for a finite string $x \in \Omega$ define the function $f_i : \Omega \to \mathbb{N}_0$ as

$$f_i(x) := \text{“number of occurrences of world } w_i \text{ in } x\text{”}.$$

Then the bag-of-words kernel is defined as

$$K(x, y) := \sum_{i=1}^{d} f_i(x) f_i(y)$$

and it is PD since it has the feature map $\phi : \Omega \to \mathbb{N}_0^d$, $\phi(x) := [f_1(x), \dots, f_d(x)]^T$.

This kernel can be used for text processing, e.g., for spam filtering if D is a set of “dangerous” words and $x \in \Omega$ is the text of an email.

Observe that $K(x, y)$ is the number of common words of the strings x, y, counted with multiplicities, so it is a “similarity measure” between two strings.

Example 2.14 (Tree kernel)**.** *Let Ω be a finite set of trees.*

Let $S := \{t_i\}_{i=1}^d$ be the set of all subtrees of trees in Ω. For a tree $x \in \Omega$ define the function $f_i : \Omega \to \mathbb{N}_0$ as

$$f_i(x) := \text{“number of occurrences of subtree } t_i \text{ in } x\text{”}.$$

Then the tree kernel is defined as

$$K(x, y) := \sum_{i=1}^{d} f_i(x) f_i(y)$$

and it is PD since it has the feature map $\phi : \Omega \to \mathbb{N}_0^d$, $\phi(x) := [f_1(x), \dots, f_d(x)]^T$.

This kernel can be used e.g. for natural language processing, where the trees are parse trees representing the syntactic structure of a sentence, or to process html files.

Observe that $K(x, y)$ is the number of common subtrees of the trees x, y, counted with multiplicities, so it is a “similarity measure” between two trees.

3. Kernels and Hilbert spaces

We start now to study a class of Hilbert spaces which are strictly connected with PD kernels.

First recall Theorem 1.7.

Theorem 3.1 (Kernel interpolation is well defined)**.** *Let $\Omega \subset \mathbb{R}^d$ and K a SPD kernel on Ω. Given any set $X_N := \{x_i\}_{i=1}^N \subset \Omega$ of pairwise distinct points and target values $\{f_i\}_{i=1}^N \subset \mathbb{R}$, there exists a unique kernel interpolant*

$$s(x) := \sum_{j=1}^{N} \alpha_j K(x, x_j),$$

with $s(x_i) = f_i$ for $1 \leq i \leq N$.

The idea is to analyze the set of functions for which kernel interpolation "works". This means that, in the case the data values $\{f_i\}_{i=1}^N$ come from the sampling of an unknown function $f : \Omega \to \mathbb{R}$, i.e., $f_i = f(x_i)$ for $x_i \in X_N$, we want to know when s is a good approximation of f, provided that we have sufficiently many data X_N. This will lead us to the study of error analysis and convergence results for this type of approximation.

It will turn out that this set of functions is indeed an Hilbert space, with an inner product defined by the kernel.

In this space it will also be possible to analyze in more details the process of approximation, in the sense that a precise functional characterizations can be connected to the computation of s.

We start by defining a particular class of Hilbert spaces and discussing some of its properties.

3.1 Reproducing kernel Hilbert spaces

Definition 3.2 (RKHS)**.** *Let Ω be a nonempty set, $\mathcal{H}$ an Hilbert space of functions $f : \Omega \to \mathbb{R}$ with inner product $(\cdot, \cdot)_{\mathcal{H}}$. Then $\mathcal{H}$ is called a Reproducing Kernel Hilbert Space on Ω (RKHS) if there exists a function $K : \Omega \times \Omega \to \mathbb{R}$ (the reproducing kernel) such that*

i.) $K(\cdot, x) \in \mathcal{H}$ for all $x \in \Omega$,

ii.) $(f, K(\cdot, x))_{\mathcal{H}} = f(x)$ for all $x \in \Omega$, for all $f \in \mathcal{H}$ (reproducing property).

This definition could seem a bit artificial, but RKHS are quite common spaces, as shown by the following examples.

Example 3.3 (Finite dimensional Hilbert spaces)**.** *Let Ω be a nonempty set and $\mathcal{H}$ be an Hilbert space of functions $f : \Omega \to \mathbb{R}$, with $\dim(\mathcal{H}) = N < \infty$. Let $\{v_j\}_{j=1}^N$ be an orthonormal basis of $\mathcal{H}$. Then $\mathcal{H}$ is a RKHS on Ω with kernel*

$$K(x,y) := \sum_{j=1}^{N} v_j(x)v_j(y),\ x,y \in \Omega.$$

Indeed, for all $x \in \Omega$ the function $K(x,\cdot) = \sum_{j=1}^N v_j(x)v_j(\cdot)$ is clearly an element of $\mathcal{H}$ because it is a linear combination (with x-dependent coefficients) of basis elements, and for all $f(\cdot) := \sum_{i=1}^N c_i v_i(\cdot) \in \mathcal{H}$, it holds

$$\begin{aligned}(f, K(\cdot,x))_{\mathcal{H}} &= \left(\sum_{i=1}^{N} c_i v_i, \sum_{j=1}^{N} v_j(x)v_j\right)_{\mathcal{H}} = \sum_{i,j=1}^{N} c_i v_j(x)\,(v_i,v_j)_{\mathcal{H}} \\ &= \sum_{i,j=1}^{N} c_i v_j(x)\delta_{ij} = \sum_{i=1}^{N} c_i v_i(x) = f(x).\end{aligned}$$

Example 3.4 (The Sobolev space $H_0^1((0,1))$)**.** *Consider the Sobolev space* [1]

$$H_0^1((0,1)) := \{f : [0,1] \to \mathbb{R}, f \in L_2((0,1)), f' \in L_2((0,1)), f(0) = f(1) = 0\}$$

where f' is the weak derivative of f, equipped with the inner product

$$(f,g)_{H_0^1((0,1))} := \int_0^1 f'(y)g'(y)dy.$$

Then $H_0^1((0,1))$ is a RKHS on $(0,1)$ with reproducing kernel the Brownian bridge kernel

$$K(x,y) := \min(x,y) - xy = \begin{cases} y(1-x), & y \le x \\ x(1-y), & y > x \end{cases}.$$

Indeed, it holds $K(\cdot,y) \in L_2((0,1))$,

$$\partial_y K(x,y) = \begin{cases} (1-x), & y \le x \\ -x, & y > x \end{cases} \in L_2((0,1))$$

and $K(0,y) = K(1,y) = 0$, so $K(\cdot,y) \in H_0^1((0,1))$ for all $y \in \Omega$. Moreover for all $f \in H_0^1((0,1))$

$$\begin{aligned}(f, K(\cdot,x))_{H_0^1((0,1))} &= \int_0^1 f'(y)\partial_y K(x,y)dy = \int_0^x f'(y)(1-x)dy + \int_x^1 f'(y)(-x)dy \\ &= \int_0^x f'(y)dy - x\int_0^1 f'(y)dy = f(x) - f(0) - x\,(f(1) - f(0)) \\ &= f(x).\end{aligned}$$

[1] For a function $f \in L_2$, it makes no sense to prescribe a pointwise value as we do in the definition. But it can be proven that, when $H_0^1((0,1))$ is properly defined, it holds $H_0^1((0,1)) \subset C((0,1))$, so the assertion $f(0) = f(1) = 0$ makes sense.

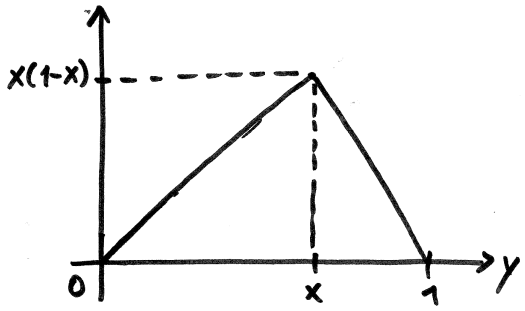

Figure 3.1: Brownian bridge kernel for one given $x \in (0, 1)$.

3.1.1 Properties

We see some basic properties of RKHS, which will be useful later.

Proposition 3.5. *Let $\mathcal{H}$ be a RKHS on Ω with reproducing kernel K. Let $N, M \in \mathbb{N}$, $\alpha \in \mathbb{R}^N$, $\beta \in \mathbb{R}^M$, $X_N, Y_M \subset \Omega$, and define the functions*

$$f(x) := \sum_{i=1}^{N} \alpha_i K(x, x_i), \quad g(x) := \sum_{j=1}^{M} \beta_j K(x, y_j), \quad x \in \Omega.$$

Then we have the following:

i) $f, g \in \mathcal{H}$,

ii) $(f, g)_{\mathcal{H}} = \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \beta_j K(x_i, y_j)$.

Proof. The two properties follow from the definition of RKHS:

(i) By Property i of Definition 3.2 we have $K(\cdot, x_i) \in \mathcal{H}$ for all $x_i \in X_N$. Since $\mathcal{H}$ is an Hilbert space, it is in particular a linear space and thus it contains all finite linear combinations of its elements, so $f, g \in \mathcal{H}$.

(ii) The inner product of f, g is well defined since $f, g \in \mathcal{H}$. We use Property ii of Definition 3.2 and linearity of the inner product to obtain

$$\begin{aligned}(f, g)_{\mathcal{H}} &= \left(\sum_{i=1}^{N} \alpha_i K(\cdot, x_i), \sum_{j=1}^{M} \beta_j K(\cdot, y_j) \right)_{\mathcal{H}} = \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \beta_j \left(K(\cdot, x_i), K(\cdot, y_j) \right)_{\mathcal{H}} \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \beta_j K(x_i, y_j).\end{aligned}$$

□

The following is the first connection between RKHS and PD kernels.

Theorem 3.6 (Reproducing kernels are PD kernels)**.** *Let $\mathcal{H}$ be a RKHS with reproducing kernel K. Then K is unique and it is a positive definite kernel.*

Proof. Taking $f(\cdot) := K(\cdot, y)$, $g(\cdot) := K(\cdot, y)$, we get from (ii) of Proposition 3.5 that $K(x, y) = (f, g)_{\mathcal{H}} = (g, f)_{\mathcal{H}} = K(y, x)$, so K is a kernel according to Definition 1.1.

To prove PD of K we check Definition 1.2. For $X_N \subset \Omega$ pairwise distinct and $\alpha \in \mathbb{R}^N$, $\alpha \neq 0$ we use Property (ii) of Proposition 3.5 to obtain

$$\begin{aligned}\alpha^T A \alpha = \sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) &= \left(\sum_{i=1}^{N} \alpha_i K(\cdot, x_i), \sum_{j=1}^{N} \alpha_j K(\cdot, x_j) \right)_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^{N} \alpha_i K(\cdot, x_i) \right\|_{\mathcal{H}}^2 \geq 0.\end{aligned}$$

Observe that we can not conclude > 0 since $\{K(\cdot, x_i)\}_{i=1}^N$ can be linearly dependent in general, so K is not SPD in general.

Assume now K_1, K_2 are two reproducing kernels of $\mathcal{H}$. From i of Definition 3.2 we have $K_1(\cdot, x), K_2(\cdot, y) \in \mathcal{H}$ for all $x, y \in \Omega$. Since K_1, K_2 are both reproducing kernels of $\mathcal{H}$, they both satisfy the reproducing property, so for all $x, y \in \Omega$ we obtain

$$K_1(x, y) = (K_1(\cdot, y), K_2(\cdot, x))_{\mathcal{H}} = K_2(x, y).$$

□

3.1.2 Characterization

We recall the following theorem.

Theorem 3.7 (Riesz representation theorem for Hilbert spaces)**.** *Let $\mathcal{H}$ be an Hilbert space and denote as $\mathcal{H}'$ its dual, i.e., the set of linear and continuous functionals $\lambda : \mathcal{H} \to \mathbb{R}$ with norm*

$$\|\lambda\|_{\mathcal{H}'} = \sup_{f \in \mathcal{H}, f \neq 0} \frac{|\lambda(f)|}{\|f\|_{\mathcal{H}}}.$$

Then for all $\lambda \in \mathcal{H}'$ there exists a unique $v_\lambda \in \mathcal{H}$ (the Riesz representer of λ) such that

$$\lambda(f) = (v_\lambda, f)_{\mathcal{H}} \text{ for all } f \in \mathcal{H}.$$

Moreover, $\|\lambda\|_{\mathcal{H}'} = \|v_\lambda\|_{\mathcal{H}}$.

It is now possible to completely characterize Hilbert spaces which are RKHS.

Proposition 3.8 (Properties of RKHS)**.** *Let Ω be a nonempty set and $\mathcal{H}$ an Hilbert space of functions $f : \Omega \to \mathbb{R}$, then*

i) $\mathcal{H}$ is a RKHS if and only if the point evaluation functionals are continuous (i.e., for all $x \in \Omega$ the functional $\delta_x : \mathcal{H} \to \mathbb{R}$, $\delta_x(f) := f(x)$ satisfy $\delta_x \in \mathcal{H}'$).

If $\mathcal{H}$ is a RKHS with kernel K, then

ii) $K(\cdot, x)$ is the Riesz-representer of the functional $\delta_x \in \mathcal{H}'$

iii) K is strictly PD if and only if $\{\delta_x, x \in \Omega\}$ are linearly independent

iv) $|f(x)| \leq \sqrt{K(x,x)}\, \|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$, $x \in \Omega$. In particular $\|f\|_{\mathcal{H}} = 0$ implies $f(x) = 0$ for all $x \in \Omega$

v) Convergence in $\mathcal{H}$ implies pointwise convergence (i.e., if $f \in \mathcal{H}, \{f_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$ and $\lim_{n\to\infty} \|f - f_n\|_{\mathcal{H}} = 0$, then $\lim_{n\to\infty} |f(x) - f_n(x)| = 0$ for all $x \in \Omega$).

Proof. (i) Assume $\mathcal{H}$ is a RKHS with kernel K. Then using (iv)

$$|\delta_x(f)| = |f(x)| \leq \sqrt{K(x,x)}\, \|f\|_{\mathcal{H}},$$

so δ_x is bounded since $\|\delta_x\|_{\mathcal{H}'} \leq C_x := \sqrt{K(x,x)}$, and thus it is continuous..

Assume instead that $\delta_x \in \mathcal{H}'$ for all $x \in \Omega$. Then by Theorem 3.7 there exists a Riesz representer $v_{\delta_x} \in \mathcal{H}$. If we define $K(\cdot, x) := v_{\delta_x}(\cdot) \in \mathcal{H}$, then K satisfies the two properties of Definition 3.2 since clearly $v_{\delta_x} \in \mathcal{H}$ and $(f, v_{\delta_x})_{\mathcal{H}} = f(x)$ for all $x \in \Omega$ and $f \in \mathcal{H}$ by definition of Riesz representer. So $\mathcal{H}$ has a reproducing kernel, thus it is a RKHS.

(ii) The reproducing property implies that $(f, K(\cdot, x))_{\mathcal{H}} = f(x)$ for all $x \in \Omega$, $f \in \mathcal{H}$. Since $K(\cdot, x) \in \mathcal{H}$, it is the unique Riesz representer of δ_x.

(iii) We first show that a finite set of linear functionals is linearly independent if and only if their Riesz representers are linearly independent.

Let $\lambda_1, \ldots, \lambda_N \in \mathcal{H}'$ and $v_{\lambda_1}, \ldots, v_{\lambda_N} \in \mathcal{H}'$ be their Riesz representers. The functionals are linearly dependent if and only if there exists $\alpha \in \mathbb{R}^N$ such that $\lambda := \sum_{i=1}^N \alpha_i \lambda_i = 0$ in $\mathcal{H}'$, i.e. $\lambda(f) = 0$ for all $f \in \mathcal{H}$. This is true if and only if

$$0 = \lambda(f) = \sum_{i=1}^N \alpha_i \lambda_i(f) = \sum_{i=1}^N \alpha_i \left(v_{\lambda_i}, f\right)_{\mathcal{H}} = \left(\sum_{i=1}^N \alpha_i v_{\lambda_i}, f\right)_{\mathcal{H}},$$

i.e., if and only if $\sum_{i=1}^N \alpha_i v_{\lambda_i} = 0$ in $\mathcal{H}$, i.e., if and only if the Riesz representers are linearly dependent.

Now $\{\delta_x, x \in \Omega\}$ are linearly independent if and only if their Riesz representers are linearly independent, i.e., (Property (ii)) if and only if $\{K(\cdot, x), x \in \Omega\}$ are linearly independent.

In this case in the proof of Theorem 3.6 one can conclude $\alpha^T A\alpha > 0$, since

$$\left\| \sum_{i=1}^{N} \alpha_i K(\cdot, x_i) \right\|_{\mathcal{H}}^2 > 0,$$

so K is SPD.

(iv) By the Cauchy-Schwarz inequality and the reproducing property we have for all $x \in \Omega$ and $f \in \mathcal{H}$ that

$$|f(x)| = |(f, K(\cdot, x))_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|K(\cdot, x)\|_{\mathcal{H}}$$

where

$$\|K(\cdot, x)\|_{\mathcal{H}} = (K(\cdot, x), K(\cdot, x))_{\mathcal{H}}^{1/2} = \sqrt{K(x, x)}.$$

(v) Since $(f - f_n) \in \mathcal{H}$ for all n, using (iv) we obtain for all $x \in \Omega$ that

$$|f(x) - f_n(x)| \leq \sqrt{K(x, x)}\, \|f - f_n\|_{\mathcal{H}} \to 0 \text{ as } n \to \infty,$$

and since $\lim_{n\to\infty} \|f - f_n\|_{\mathcal{H}} = 0$ we have $\lim_{n\to\infty} |f(x) - f_n(x)| = 0$.

□

Example 3.9 (L_2 is not a RKHS)**.** *Using Property* (ii) *of Theorem 3.8 we can prove that $L_2(\mathbb{R})$ is not a RKHS (it holds the same for any L_2 space).*

Indeed, for $\varepsilon > 0$ consider the function $f_\varepsilon(x) := (1 - |\varepsilon x|)_+$. It holds $\delta_0(f_\varepsilon) = f_\varepsilon(0) = 1$ for all ε, but

$$\|f_\varepsilon\|_{L_2(\mathbb{R})}^2 = \int_{\mathbb{R}} f_\varepsilon^2(x) dx = \frac{2}{3\varepsilon},$$

so in particular $f_\varepsilon \in L_2(\mathbb{R})$.

It follows that δ_0 is not bounded (thus not continuous), since

$$\|\delta_0\|_{L_2(\mathbb{R})'} = \sup_{f \in L_2(\mathbb{R}), f \neq 0} \frac{|f(0)|}{\|f\|_{L_2(\mathbb{R})}} \geq \frac{|f_\varepsilon(0)|}{\|f_\varepsilon\|_{L_2(\mathbb{R})}} = \sqrt{\frac{3\varepsilon}{2}},$$

and $\lim_{\varepsilon\to\infty} \sqrt{3\varepsilon/2} = \infty$.

3.2 The native space of a PD kernel

When solving approximation problems, we work in the opposite direction, i.e., we have a kernel and we would like to know what space of functions can be approximated.

We see in the following Theorem that we can indeed start from a given PD kernel K and identify and construct a space that is associated to it.

Moreover, it would be possible to deduce properties of the functions in this space by looking at properties of the kernel. This is a useful indication in practical applications to choose a proper kernel for a given approximation task.

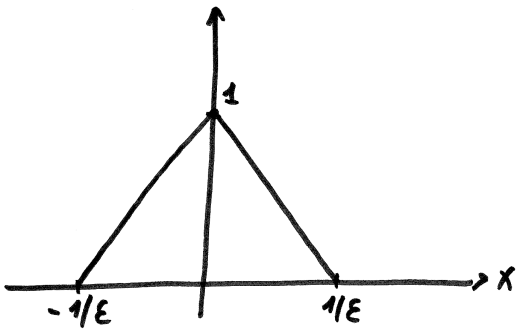

Figure 3.2: The function f_ε.

Theorem 3.10 (RKHS from kernels – Moore - Aronszajn)**.** *Let Ω be a nonempty set and $K : \Omega \times \Omega \to \mathbb{R}$ a positive definite kernel. Then there exists a unique RKHS $\mathcal{H}_K(\Omega)$ with reproducing kernel K.*

Some remarks before the proof:

- In the approximation literature, the RKHS space of a kernel is usually called the native space of the kernel. We will use this notion from now on.
- We also change the notation to stress the dependence on the kernel: the native space of K on Ω will be denoted as $\mathcal{H}_K$ or even $\mathcal{H}_K(\Omega)$. Notice that in some texts it is denoted as $\mathcal{N}_K(\Omega)$.

Proof. The construction is guided by Proposition 3.5, where we proved that any RKHS needs to contain functions of a certain form, and the inner product is defined in a certain way for those functions. We start from this observation and construct the full space $\mathcal{H}_K(\Omega)$.

First, following Property (i) of Proposition 3.5 we consider the set $\mathcal{H}_0$ of finite linear combinations of kernels centered in points of Ω, i.e.,

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

It is a set of functions $\Omega \to \mathbb{R}$, and a generic $f \in \mathcal{H}_0$ can be written as

$$f(x) = \sum_{j=1}^{N} \alpha_j K(x, x_j) \tag{3.1}$$

for some $N \in \mathbb{N}$, $X_N := \{x_j\}_{j=1}^N \subset \Omega$ and $\alpha \in \mathbb{R}^N$. We can assume without loss of generality that the points X_N are pairwise distinct, otherwise it is enough so sum the

coefficients α_j corresponding to two equal points, and that all the α_j are non zero, otherwise we can remove them from the sum.

Observe that the representation (3.1) does not need to be unique, i.e., it could be that there exists $N, M \in \mathbb{N}$, $X_N := \{x_j\}_{j=1}^N, Y_M := \{y_i\}_{i=1}^M \subset \Omega$ and $\alpha \in \mathbb{R}^N, \beta \in \mathbb{R}^M$ such that

$$f(x) = \sum_{j=1}^{N} \alpha_j K(x, x_j) = \sum_{i=1}^{M} \beta_i K(x, y_i) \quad \text{for all } x \in \Omega.$$

Second, following Property (ii) of Proposition 3.5 we can define a map $\mathcal{B} : \mathcal{H}_0 \times \mathcal{H}_0 \to \mathbb{R}$ as

$$\mathcal{B}(f, g) := \mathcal{B}\left(\sum_{j=1}^{N} \alpha_j K(x, x_j), \sum_{i=1}^{M} \beta_i K(x, y_i)\right) := \sum_{j=1}^{N}\sum_{i=1}^{M} \alpha_j \beta_i K(x_j, y_i). \tag{3.2}$$

This map is well defined, i.e., it is independent of the particular representation of f, g. Indeed, from the definition of g we have

$$\mathcal{B}(f, g) = \sum_{j=1}^{N}\sum_{i=1}^{M} \alpha_j \beta_i K(x_j, y_i) = \sum_{j=1}^{N} \alpha_j \sum_{i=1}^{M} \beta_i K(x_j, y_i) = \sum_{j=1}^{N} \alpha_j g(x_j),$$

so it depends only on the values of g, not on its particular representation. The same holds for f with the same argument.

Moreover, $\mathcal{B}$ is symmetric and bilinear from the definition and because K is symmetric. It is also positive definite because K is PD and thus

$$\mathcal{B}(f, f) = \sum_{i,j=1}^{N} \alpha_j \alpha_i K(x_j, x_i) = \alpha^T A \alpha \geq 0.$$

This means that $\mathcal{B}$ is a semi inner product on $\mathcal{H}_0$, so we can introduce the notation $(f, g)_{\mathcal{H}_0} := \mathcal{B}(f, g)$ for $f, g \in \mathcal{H}_0$. Recall that this implies that $\|f\|_{\mathcal{H}_0} := \sqrt{(f, f)_{\mathcal{H}_0}}$ is a seminorm and so it satisfies the Cauchy-Schwarz inequality.

We can now prove that K acts as a reproducing kernel on $\mathcal{H}_0$ w.r.t. $(\cdot, \cdot)_{\mathcal{H}_0}$, i.e., that it satisfies the two properties of Definition 3.2. Indeed, from the definition of $\mathcal{H}_0$ we have $K(\cdot, x) \in \mathcal{H}_0$ for all $x \in \Omega$, and we can also compute the inner product with any $f \in \mathcal{H}$ and obtain by definition (3.2) of $(\cdot, \cdot)_{\mathcal{H}_0}$

$$(f, K(\cdot, x))_{\mathcal{H}_0} = \sum_{j=1}^{N} \alpha_j K(x_j, x) = f(x).$$

What remains to have an inner product is to prove that it is non degenerate, i.e., $(f, f)_{\mathcal{H}_0} = 0 \Rightarrow f = 0$ for all $f \in \mathcal{H}_0$. To see this, we use the reproducing property and the Cauchy-Schwarz inequality to obtain

$$|f(x)| = \left|(f, K(\cdot, x))_{\mathcal{H}_0}\right| \leq \|f\|_{\mathcal{H}_0} \|K(\cdot, x)\|_{\mathcal{H}_0} \quad \text{for all } x \in \Omega,$$

thus $\|f\|_{\mathcal{H}_0} = 0$ implies $f(x) = 0$ for all $x \in \Omega$.

So we have that $(\cdot,\cdot)_{\mathcal{H}_0}$ is an inner product on $\mathcal{H}_0$, i.e., $\mathcal{H}_0$ is a pre-Hilbert space with inner product $(\cdot,\cdot)_{\mathcal{H}_0}$ and norm $\|\cdot\|_{\mathcal{H}_0}$.

This means that the completion of $\mathcal{H}_0$ with respect to $\|\cdot\|_{\mathcal{H}_0}$ is an Hilbert space. We denote it as $\mathcal{H}_K(\Omega)$ and its inner product as $(\cdot,\cdot)_{\mathcal{H}_K(\Omega)}$. In detail:

$$\mathcal{H}_K(\Omega) := \{f \mid \exists\, \{f_n\}_n \subset \mathcal{H}_0 \text{ Cauchy sequence with } \lim_{n\to\infty} f_n = f\},$$

and, for $f := \lim_{n\to\infty} f_n$, $g := \lim_{m\to\infty} g_m$,

$$(f,g)_{\mathcal{H}_K(\Omega)} := \Big(\lim_{n\to\infty} f_n, \lim_{m\to\infty} g_m\Big)_{\mathcal{H}_K(\Omega)} = \lim_{n\to\infty}\lim_{m\to\infty} (f_n, g_m)_{\mathcal{H}_0}\,.$$

Observe that of course $\mathcal{H}_0 \subset \mathcal{H}_K(\Omega)$ by construction, and an element f of $\mathcal{H}_0$ can be obtained as limit of the Cauchy sequence $\{f_n\}_n$ with $f_n := f$ for all n.

What remains to prove is that the elements of $\mathcal{H}_K(\Omega)$ are functions, and that K is still a reproducing kernel on $\mathcal{H}_K(\Omega)$. To see this, observe that for $f \in \mathcal{H}_K(\Omega)$ and for any Cauchy sequence $\{f_n\}_n$ that converges to f, we have

$$|f_n(x) - f_m(x)| \le \|f_n - f_m\|_{\mathcal{H}_K(\Omega)} \sqrt{K(x,x)} \quad \text{for all } x \in \Omega,$$

so $\{f_n(x)\}_n$ is a Cauchy sequence in $\mathbb{R}$, which is complete, so we can define $f(x) := \lim_{n\to\infty} f_n(x)$, and thus f is a pointwise defined function $f : \Omega \to \mathbb{R}$.

Moreover, the kernel K is a reproducing kernel on $\mathcal{H}_K(\Omega)$ since by construction $K(\cdot,x) \in \mathcal{H}_0 \subset \mathcal{H}_K(\Omega)$, and for all $f \in \mathcal{H}_K(\Omega)$ the reproducing property holds since

$$\begin{aligned} f(x) := \lim_{n\to\infty} f_n(x) &\underset{f_n\in\mathcal{H}_0}{=} \lim_{n\to\infty} (f_n, K(\cdot,x))_{\mathcal{H}_0} \\ &\underset{f_n \text{ converges in } \mathcal{H}_0}{=} \Big(\lim_{n\to\infty} f_n, K(\cdot,x)\Big)_{\mathcal{H}_0} \underset{f_n\to f}{=} (f, K(\cdot,x))_{\mathcal{H}_K(\Omega)}. \end{aligned}$$

Finally, $\mathcal{H}_K(\Omega)$ is unique. Indeed, if $\mathcal{H}$ is another RKHS on Ω with reproducing kernel K, we can use again Proposition 3.5 and prove that $\mathcal{H}_0 \subset \mathcal{H}$ and that the inner product is defined by $\mathcal{B}$, and then we can use again this proof to conclude $\mathcal{H} = \mathcal{H}_K(\Omega)$.

□

Remark 3.11. *Two remarks on the proof:*

- *The representation* (3.1) *is unique if K is SPD (exercise).*
- *We need to check that the elements f or the completion are in fact functions, because it can happen that they are something else. Consider for example the case of the completion of the space of continuous functions $C((0,1))$ w.r.t. the inner product $(f,g) := \int_0^1 f(x)g(x)dx$. In this case the completion is $L_2((0,1))$, and the elements of this space are equivalence classes, not pointwise defined functions.*

3.2.1 Consequences on K

We have seen in Theorem 1.9 that feature maps $\phi : \Omega \to \mathbb{R}$ induce PD kernels via $K(x, y) := (\phi(x), \phi(y))_H$. Indeed, all PD kernels are of this type.

Proposition 3.12 (Kernel feature map)**.** *Let K be a PD kernel on Ω. Then the kernel is defined by the feature map $\phi : \Omega \to \mathcal{H}_K(\Omega)$, $\phi(x) := K(\cdot, x)$, which is called the kernel feature map.*

Proof. Theorem 3.10 guarantees the existence of $\mathcal{H}_K(\Omega)$ and that $\phi(x) := K(\cdot, x) \in \mathcal{H}_K(\Omega)$ for all $x \in \Omega$. Then from the reproducing property we have

$$(\phi(x), \phi(y))_{\mathcal{H}_K(\Omega)} = (K(\cdot, x), K(\cdot, y))_{\mathcal{H}_K(\Omega)} = K(x, y).$$

□

3.2.2 Consequences on Ω

The existence of a native space for a given kernel allows to define some structure on the set Ω, even when Ω has no own structure (e.g., it could be a set of strings in the case of the kernel of Example 2.13).

Proposition 3.13 (Metric on Ω)**.** *Let K be a PD kernel on a set Ω. Then we can define a pseudo metric on Ω as*

$$d_K(x, y) := \|K(\cdot, x) - K(\cdot, y)\|_{\mathcal{H}_K(\Omega)} = \sqrt{K(x, x) - 2K(x, y) + K(y, y)}.$$

It is a metric if K is SPD.

Proof. By the definition, d_K is positive, symmetric and satisfies the triangle inequality. Moreover $x = y$ implies $K(\cdot, x) = K(\cdot, y)$ so $d_K(x, y) = 0$, so it is a pseudo metric.

We can conclude that it is a metric if $d_K(x, y) = 0$ implies $x = y$, and this is the case if K is SPD, since in this case $K(\cdot, x)$ and $K(\cdot, y)$ are linearly independent (Proposition 3.8, Property iii).

The second equality is derived using the reproducing property of Definition 3.2:

$$\begin{aligned}\|K(\cdot, x) - K(\cdot, y)\|^2_{\mathcal{H}_K(\Omega)} &= (K(\cdot, x) - K(\cdot, y), K(\cdot, x) - K(\cdot, y))_{\mathcal{H}_K(\Omega)} \\ &= (K(\cdot, x), K(\cdot, x))_{\mathcal{H}_K(\Omega)} - 2\,(K(\cdot, x), K(\cdot, y))_{\mathcal{H}_K(\Omega)} + (K(\cdot, y), K(\cdot, y))_{\mathcal{H}_K(\Omega)} \\ &= K(x, x) - 2K(x, y) + K(y, y).\end{aligned}$$

□

3.2.3 Consequences on $\mathcal{H}_K(\Omega)$

Thanks to the construction of Theorem 3.10, it is also possible to deduce properties of the elements of $\mathcal{H}_K(\Omega)$ just by looking at properties of the kernel. This is very useful when solving approximation problems, since one can anticipate the smoothness of the approximant by choosing a suitable kernel.

Proposition 3.14 (Native space and smoothness)**.** *Assume K is SPD on Ω. Then the following holds:*

i) If $\dim(\Omega) = \infty$ then $\dim(\mathcal{H}_K(\Omega)) = \infty$.

ii) Every $f \in \mathcal{H}_K(\Omega)$ is Lipschitz continuous w.r.t. d_K.

iii) If moreover $\Omega \subset \mathbb{R}^d$ is open and $K \in C^{2k}(\Omega \times \Omega)$ for $k \in \mathbb{N}$, then $\mathcal{H}_K(\Omega) \subset C^k(\Omega)$. In particular, for all multiindex $a := (a_1, \dots, a_d)$ with $|a| := a_1 + a_2 + \cdots + a_d \leq k$ it holds

$$D^a f(x) := \partial^{a_1}_{x^{(1)}} \partial^{a_2}_{x^{(2)}} \dots \partial^{a_d}_{x^{(d)}} f(x) = (f, D^a_2 K(\cdot, x))_{\mathcal{H}_K(\Omega)}, \tag{3.3}$$

where the subscript 2 means that we differentiate w.r.t. the second argument.

Proof. The first two properties are consequences of the last propositions:

(i) Let $\dim(\Omega) = N \leq \infty$, so we can select X_N pairwise distinct points in Ω. Since K is SPD, the functions $\{K(\cdot, x_i)\}_{i=1}^N$ are linearly independent (Proposition 3.8, Property iii), and by Theorem 3.10 we have $\{K(\cdot, x_i)\}_{i=1}^N \subset \mathcal{H}_K(\Omega)$. It follows that $\dim(\mathcal{H}_K(\Omega)) \geq \dim(\Omega)$, so $\dim(\Omega) = \infty$ implies $\dim(\mathcal{H}_K(\Omega)) = \infty$.

(ii) Using Proposition 3.13 and the Cauchy-Schwarz inequality we have

$$\begin{aligned} |f(x) - f(y)| &= \left|(f, K(\cdot, x) - K(\cdot, y))_{\mathcal{H}_K(\Omega)}\right| \leq \|f\|_{\mathcal{H}_K(\Omega)} \|K(\cdot, x) - K(\cdot, y)\|_{\mathcal{H}_K(\Omega)} \\ &= \|f\|_{\mathcal{H}_K(\Omega)} d_K(x, y), \end{aligned}$$

which proves Lipschitz continuity.

For the smoothness, we use the fact that, if $K \in C^{2k}(\Omega \times \Omega)$ and $|a| \leq k$, then $D^a_2 K(\cdot, x)$ exists and is in $\mathcal{H}_K(\Omega)$. This will be proven later, when dealing with PDEs.

(iii) We prove the formula 3.3 by induction on $|a|$, which proves existence and continuity of the derivatives of f. For $|a| = 0$ formula 3.3 is just the reproducing property. For $|a| > 0$, we can assume e.g. that $a_1 > 0$ and define $b := (a_1 - 1, a_2 \dots, a_d)$. Denoting as e_1 the first unit vector in $\mathbb{R}^d$, we have by the induction hypothesis

and by definition of partial derivative that

$$\begin{aligned} D^a f(x) &= \lim_{h\to 0} \frac{1}{h}\left(D^b f(x+he_1) - D^b f(x)\right) \\ &= \lim_{h\to 0} \frac{1}{h}\left[\left(f, D_2^b K(\cdot, x+he_1)\right)_{\mathcal{H}_K(\Omega)} - \left(f, D_2^b K(\cdot, x)\right)_{\mathcal{H}_K(\Omega)}\right] \\ &= \lim_{h\to 0} \frac{1}{h}\left(f, D_2^b K(\cdot, x+he_1) - D_2^b K(\cdot, x)\right)_{\mathcal{H}_K(\Omega)} \\ &= \left(f, \lim_{h\to 0} \frac{1}{h}\left(D_2^b K(\cdot, x+he_1) - D_2^b K(\cdot, x)\right)\right)_{\mathcal{H}_K(\Omega)} \\ &= \left(f, D_2^a K(\cdot, x)\right)_{\mathcal{H}_K(\Omega)}. \end{aligned}$$

The same holds for any other b with $|b| < k$.

□

3.2.4 Basic operations on $\mathcal{H}_K(\Omega)$

Finally, we show that some of the basic operations on PD kernels discussed in Proposition 2.7 have a corresponding version in terms of $\mathcal{H}_K(\Omega)$.

Proposition 3.15 (Basic operations on kernels - $\mathcal{H}_K(\Omega)$)**.** *Let K, K_1, K_2 be PD kernels on $\Omega \times \Omega \to \mathbb{R}$, $g : \Omega \to \mathbb{R} \setminus \{0\}$, $a > 0$. Then we have the following:*

i) $K(x,y) := g(x)g(y)$ *results in* $\mathcal{H}_K(\Omega) = \text{span}\{g\}$.

ii) $K(x,y) := aK_1(x,y)$ *results in* $\mathcal{H}_K(\Omega) = \mathcal{H}_{K_1}(\Omega)$.

iii) $K(x,y) := g(x)K_1(x,y)g(y)$ *results in* $\mathcal{H}_K(\Omega) = g\mathcal{H}_{K_1}(\Omega) := \{gf | f \in \mathcal{H}_{K_1}(\Omega)\}$ *(weighted RKHS).*

iv) $K(x,y) := K_1(x,y) + K_2(x,y)$ *results in* $\mathcal{H}_K(\Omega) = \mathcal{H}_{K_1}(\Omega) + \mathcal{H}_{K_2}(\Omega)$.

v) $K_1 \ll K_2$ *results in* $\mathcal{H}_{K_1}(\Omega) \subset \mathcal{H}_{K_2}(\Omega)$. *Here $\ll$ is the partial order on the set of positive definite kernels on Ω defined by*

$$K_1 \ll K_2 \quad \Leftrightarrow \quad K := K_2 - K_1 \text{ is PD}$$

Proof. Exercise. □

4. Interpolation in native spaces

Since we are dealing with interpolation, we assume in this Chapter that K is a SPD kernel and $\Omega \subset \mathbb{R}^d$. With these assumptions, we have from Theorem 1.7 that for every $X_N \subset \Omega$ pairwise distinct and $\{f_i\}_{i=1}^N \subset \mathbb{R}$ the kernel interpolant with data points X_N and data values $\{f_i\}_{i=1}^N$ exists and is unique.

When the data come from the sampling of an unknown function f, i.e., $f_i := f(x_i)$, $1 \leq i \leq N$, we denote the interpolant as s_f instead of s to specify the dependence on f.

Working with the native space $\mathcal{H}_K(\Omega)$, we can reformulate the process of kernel interpolation as follows.

Proposition 4.1 (Interpolation in the native space)**.** *For any set X_N of pairwise distinct points, the linear space*

$$V(X_N) := \textit{ span } \{K(\cdot, x_i), x_i \in X_N\}$$

is an N-dimensional subspace of $\mathcal{H}_K(\Omega)$, and the kernel interpolant s_f is an element of $V(X_N)$.

Moreover, if $f \in \mathcal{H}_K(\Omega)$, we have

$$s_f = \Pi_{V(X_N)}(f),$$

where $\Pi_{V(X_N)}$ is the orthogonal projection from $\mathcal{H}_K(\Omega)$ to $V(X_N)$.

Proof. The fact that $V(X_N) \subset \mathcal{H}_K(\Omega)$ follows from the construction of $\mathcal{H}_K(\Omega)$ in Theorem 3.10, and $s_f \in V(X_N)$ by definition of s_f.

To prove that $s_f = \Pi_{V(X_N)}(f)$, we need to prove that $f - s_f$ is orthogonal to $V(X_N)$, i.e., since $\{K(\cdot, x_i)\}_{i=1}^N$ is a basis of $V(X_N)$, that

$$(f - s_f, K(\cdot, x_i))_{\mathcal{H}_K(\Omega)} = 0, \ 1 \leq i \leq N.$$

But from the reproducing property we obtain

$$\begin{aligned}(f - s_f, K(\cdot, x_i))_{\mathcal{H}_K(\Omega)} &= (f, K(\cdot, x_i))_{\mathcal{H}_K(\Omega)} - (s_f, K(\cdot, x_i))_{\mathcal{H}_K(\Omega)} \\ &= f(x_i) - s_f(x_i) = 0, \ \ 1 \leq i \leq N,\end{aligned}$$

since s_f interpolates f in the points X_N. □

Since the interpolant is the orthogonal projection of f into $V(X_N)$, we can rewrite the properties of orthogonal projection in the following way.

Corollary 4.2 (Orthogonal decomposition)**.** *Let $X_N \subset \Omega$ be a set of pairwise distinct points. A function $g \in \mathcal{H}_K(\Omega)$ is orthogonal to $V(X_N) \subset \mathcal{H}_K(\Omega)$ (i.e., $g \in V(X_N)^\perp$) if and only if $g(x_i) = 0$ for all $x_i \in X_N$.*

Moreover, each $f \in \mathcal{H}_K(\Omega)$ can be uniquely decomposed as $f = s_f + (f - s_f)$ with

$$\begin{aligned}
&s_f \in V(X_N),\ s_f(x_i) = f(x_i) \text{ for all } x_i \in X_N \\
&(f - s_f) \in V(X_N)^\perp,\ (f - s_f)(x_i) = 0 \text{ for all } x_i \in X_N \\
&(f, f - s_f)_{\mathcal{H}_K(\Omega)} = 0 \\
&\|f\|^2_{\mathcal{H}_K(\Omega)} = \|s_f\|^2_{\mathcal{H}_K(\Omega)} + \|s - s_f\|^2_{\mathcal{H}_K(\Omega)}.
\end{aligned}$$

Proof. It is a general property of orthogonal projections that any $f \in \mathcal{H}_K(\Omega)$ can be uniquely decomposed in $f = g + g^\perp$ with $g = \Pi_{V(X_N)}(f)$ and $g^\perp = \Pi_{V(X_N)^\perp}(f)$. It follows that $(g, g^\perp)_{\mathcal{H}_K(\Omega)} = 0$ and thus

$$\begin{aligned}
\|f\|^2_{\mathcal{H}_K(\Omega)} &= (f, f)_{\mathcal{H}_K(\Omega)} = (g + g^\perp, g + g^\perp)_{\mathcal{H}_K(\Omega)} \\
&= \|g\|^2_{\mathcal{H}_K(\Omega)} + 2(g, g^\perp)_{\mathcal{H}_K(\Omega)} + \|g^\perp\|^2_{\mathcal{H}_K(\Omega)} \\
&= \|g\|^2_{\mathcal{H}_K(\Omega)} + \|g^\perp\|^2_{\mathcal{H}_K(\Omega)}.
\end{aligned}$$

Since we proved that $s_f = \Pi_{V(X_N)}(f) = g$, we have also $g^\perp = f - g = f - s_f$, so we obtain all the results in the Proposition.

□

4.1 Optimality of kernel interpolation

From the properties of orthogonal projections, we can deduce interesting optimality properties of kernel interpolation.

This is interesting because one could try to find other ways to approximate the function $f \in \mathcal{H}_K(\Omega)$. On one hand, still considering the ansatz

$$s_f(x) := \sum_{j=1}^{N} \alpha_j K(x, x_j) \in V(X_N),$$

it would be possible to determine the coefficient vector α by something different from interpolation, e.g., by linear least squares.

On the other hand, it would be possible to keep the interpolation conditions $s_f(x_i) = f(x_i)$, but instead avoid using the same ansatz, i.e., $s_f \notin V(X_N)$.

In both cases, we prove that we would obtain a worse approximation of f.

Proposition 4.3 (Interpolation gives best approximation)**.** *The kernel interpolant is the unique best approximation of $f \in \mathcal{H}_K(\Omega)$ from the space $V(X_N)$, i.e.,*

$$\|f - s_f\|_{\mathcal{H}_K(\Omega)} = \min_{s \in V(X_N)} \|f - s\|_{\mathcal{H}_K(\Omega)},$$

and the minimum is unique.

Proof. It follows from the fact that s_f is the orthogonal projection of f into $V(X_N)$ (Proposition 4.1), and orthogonal projections give the unique best approximation. □

Proposition 4.4 (Minimal norm interpolation)**.** *Consider the space of interpolating functions in $\mathcal{H}_K(\Omega)$, i.e.,*

$$S := \{s \in \mathcal{H}_K(\Omega) : s(x_i) = f(x_i), 1 \leq i \leq N\}.$$

Then

$$\|s_f\|_{\mathcal{H}_K(\Omega)} = \min_{s \in S} \|s\|_{\mathcal{H}_K(\Omega)},$$

and the minimum is unique.

Proof. Clearly $s_f \in S$ as it is an interpolant, so it is a candidate to be the minimal norm interpolant. Moreover, s_f is the unique interpolant of f in $V(X_N)$ by construction (since the interpolant exists and it is unique).

Now we assume $s \in S$, and we prove that $\|s\|_{\mathcal{H}_K(\Omega)} \geq \|s_f\|_{\mathcal{H}_K(\Omega)}$. Indeed, if $s \in V(X_N)$ we have immediately $s = s_f$ since s_f is the unique interpolant of f in $V(X_N)$.

Otherwise, thanks to Corollary 4.2, we can uniquely decompose s as

$$s = g + g^{\perp}$$

with $g(x_i) = s(x_i) = f(x_i)$ and $g \in V(X_N)$, thus $g = s_f$. Moreover,

$$\|s\|^2 = \|g\|^2_{\mathcal{H}_K(\Omega)} + \|g^{\perp}\|^2_{\mathcal{H}_K(\Omega)},$$

so

$$\|s_f\|^2_{\mathcal{H}_K(\Omega)} = \|g\|^2_{\mathcal{H}_K(\Omega)} = \|s\|^2_{\mathcal{H}_K(\Omega)} - \|g^{\perp}\|^2_{\mathcal{H}_K(\Omega)} \leq \|s\|^2_{\mathcal{H}_K(\Omega)}.$$

□

Remark 4.5. *Some remarks on the two propositions:*

- *Proposition 4.4 proves that we can solve a potentially infinite dimensional optimization problem (since $\dim(S) = \infty$ when $\dim(\mathcal{H}_K(\Omega)) = \infty$) by solving a finite dimensional linear system $A\alpha = b$.*
- *So far we considered the approximants obtained by removing one of the two conditions defining s_f, and we obtained in both cases that s_f is a better approximation of f. We could in principle try to compare s_f with the solution obtained by removing both the conditions, i.e., $s_f \notin V(X_N)$ and not satisfying the interpolation conditions. But in this case the best approximation is*

  $$f = \arg\min_{s \in \mathcal{H}_K(\Omega)} \|f - s\|,$$

  which is clearly the best possible approximation of f, but of course can not be computed using only the data values $\{f(x_i)\}_{i=1}^N$.

4.2 Simple bounds

Again using the properties of the orthogonal projection, we can obtain a first result that proves that both the norm of the error and the norm of the interpolant are small if the norm of $f \in \mathcal{H}_K(\Omega)$ is small. Moreover, we have that increasing the set of interpolation points we get a smaller error.

Proposition 4.6. *Let $f \in \mathcal{H}_K(\Omega)$, $X_N \subset \Omega$ a set of pairwise distinct points and s_f the unique kernel interpolant of f on X_N. Let moreover $Y \subset \Omega$ be a set of pairwise distinct points with $X_N \subset Y$, and let s_Y be the kernel interpolant of f on Y.*

Then we have $V(X_N) \subset V(Y)$ and

$$\|s_f\|_{\mathcal{H}_K(\Omega)} \leq \|s_Y\|_{\mathcal{H}_K(\Omega)} \leq \|f\|_{\mathcal{H}_K(\Omega)},$$
$$\|f - s_Y\|_{\mathcal{H}_K(\Omega)} \leq \|f - s_f\|_{\mathcal{H}_K(\Omega)} \leq \|f\|_{\mathcal{H}_K(\Omega)}.$$

Proof. We have $X_N \subset Y$ so

$$V(X_N) := \text{ span } \{K(\cdot, x_i), x_i \in X_N\} \subset V(Y) := \text{ span } \{K(\cdot, y_i), y_i \in Y\}.$$

Using Corollary 4.2 we obtain

$$\|s_f\|_{\mathcal{H}_K(\Omega)} = \sqrt{\|f\|^2_{\mathcal{H}_K(\Omega)} - \|f - s_f\|^2_{\mathcal{H}_K(\Omega)}} \leq \|f\|_{\mathcal{H}_K(\Omega)},$$
$$\|f - s_f\|_{\mathcal{H}_K(\Omega)} = \sqrt{\|f\|^2_{\mathcal{H}_K(\Omega)} - \|s_f\|^2_{\mathcal{H}_K(\Omega)}} \leq \|f\|_{\mathcal{H}_K(\Omega)}.$$

Since s_Y is the interpolant of f on Y, using the same argument proves $\|s_Y\|_{\mathcal{H}_K(\Omega)} \leq \|f\|_{\mathcal{H}_K(\Omega)}$. Moreover, since $X_N \subset Y$, s_Y is also an interpolant of f on X_N. Then, with the notation of Proposition 4.4 we have $s_Y \in S$, thus, again from the same proposition,

$$\|s_f\|_{\mathcal{H}_K(\Omega)} \leq \|s_Y\|_{\mathcal{H}_K(\Omega)}.$$

For the second part we use instead Proposition 4.3. We proved $V(X_N) \subset V(Y)$, and s_Y is the kernel interpolant of f from $V(Y)$. So we obtain

$$\|f - s_Y\|_{\mathcal{H}_K(\Omega)} \leq \|f - s_f\|_{\mathcal{H}_K(\Omega)}.$$

□

4.3 General error bounds and the power function

We move now to general error analysis of the interpolation process. General here means that we derive results that hold for any SPD kernel, while we will obtain more specific results and convergence rates for specific classes of kernels. In particular, smoother kernels lead to native spaces of smoother functions (Proposition 3.14), so we should expect a faster convergence in that case (as it is for polynomial interpolation, for example).

The formulation of the results is easier if we consider the following basis of $V(X_N)$.

Proposition 4.7 (Cardinal or Lagrange basis)**.** *There exists a cardinal or Lagrange basis of $V(X_N)$, i.e., a basis $\{\ell_j\}_{j=1}^N$ of $V(X_N)$ such that*

$$\ell_j(x_i) = \delta_{ij} \quad 1 \leq i, j \leq N.$$

Moreover, the kernel interpolant of $f \in \mathcal{H}_K(\Omega)$ can be expressed as

$$s_f(x) = \sum_{j=1}^{N} f(x_j)\ell_j(x). \tag{4.1}$$

Proof. The existence of the Lagrange basis is a general result from Numerik I, but we shortly see the proof.

Since kernel interpolation is well defined for any vector of values $b := [f_i]_{i=1}^N$, we define ℓ_j as the unique kernel interpolant on X_N with data $b := e_j$ (the j-th coordinate vector). By definition of kernel interpolation, this means that

$$\ell_j \in V(X_N) \quad \text{and} \quad \ell_j(x_i) = (e_j)_i = \delta_{ij}, \quad 1 \leq i, j \leq N.$$

We now prove that they are a basis. Since they are N elements in $V(X_N)$, it suffices to prove that they are linearly independent. To see this, we take $\alpha \in \mathbb{R}^N$ and assume $\sum_{j=1}^N \alpha_j \ell_j(x) = 0$ for all $x \in \Omega$. For all $1 \leq i \leq N$, taking $x := x_i$ we obtain

$$0 = \sum_{j=1}^{N} \alpha_j \ell_j(x_i) = \sum_{j=1}^{N} \alpha_j \delta_{ij} = \alpha_i,$$

so they are linearly independent.

To prove that formula (4.1) is correct, we define $g(x) := \sum_{j=1}^N f(x_j)\ell_j(x)$. We have $g \in V(X_N)$ since $\{\ell_j\}_{j=1}^N$ is a basis of $V(X_N)$, and, for all $1 \leq i \leq N$,

$$g(x_i) = \sum_{j=1}^{N} f(x_j)\ell_j(x_i) = f(x_i),$$

so $g = s_f$ by uniqueness of kernel interpolant. □

We now introduce the anticipated error bound for the interpolation of $f \in \mathcal{H}_K(\Omega)$ with data points X_N. The motivation of the following reasoning is that we would like to obtain some quantity, which we denote as $P_{X_N}(x)$, such that we can bound the pointwise interpolation error as

$$|f(x) - s_f(x)| \leq P_{X_N}(x) \, \|f\|_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega,$$

and we would like this bound to be optimal, i.e., $P_{X_N}(x)$ is the smallest possible number such that the inequality holds for a fixed $x \in \Omega$.

This means that, for $f \in \mathcal{H}_K(\Omega)$, $f \neq 0$, we can look for the minimal number $P_{X_N}(x)$ such that

$$\frac{|f(x) - s_f(x)|}{\|f\|_{\mathcal{H}_K(\Omega)}} \leq P_{X_N}(x) \quad \text{for all } x \in \Omega.$$

It would be possible to just take this as a definition, i.e., define $P_{X_N}(x)$ as

$$P_{X_N}(x) := \sup_{f \in \mathcal{H}_K(\Omega), f \neq 0} \frac{|f(x) - s_f(x)|}{\|f\|_{\mathcal{H}_K(\Omega)}}.$$

This definition is correct (as we will see in Proposition 4.12), but it makes things more complicated. Instead, we take the following definition.

Definition 4.8 (Power function)**.** *Let $X_N \subset \Omega$ pairwise distinct. The power function is a function $P_{X_N} : \Omega \to \mathbb{R}$ defined as*

$$P_{X_N}(x) = \left\| K(\cdot, x) - \sum_{j=1}^{N} K(\cdot, x_j)\ell_j(x) \right\|_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega.$$

With this definition, we have the following fundamental theorem, which proves that the power function gives the desired error bound and that it can actually be computed.

Theorem 4.9 (Power function error bound – Schaback)**.** *Let $X_N \subset \Omega$ pairwise distinct and $f \in \mathcal{H}_K(\Omega)$. Then it holds*

$$|f(x) - s_f(x)| \leq P_{X_N}(x) \ \|f\|_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega. \tag{4.2}$$

Moreover, the power function can be computed as

$$P_{X_N}(x) = \sqrt{K(x,x) - 2\sum_{j=1}^{N} \ell_j(x)K(x, x_j) + \sum_{i,j=1}^{N} \ell_j(x)\ell_i(x)K(x_j, x_i)}. \tag{4.3}$$

Remark 4.10. *Some remarks before the proof:*

- *Since the Lagrange basis can be computed, the formula* (4.3) *really defines a computable quantity (although this is not the best (most stable) way to compute the power function). This means that the power function can be used also numerically and it provides good indications on the interpolation error (see demo in ILIAS).*
- *The error bound* (4.2) *is very nice in the sense that it splits the error in two terms, one depending only on the function f, and the other one depending only on K, Ω and X_N. This is somehow similar to the error bound for polynomial interpolation in $d = 1$ (see Numerik I), which gives for $f \in C^{n+1}(\mathbb{R})$*

$$|f(x) - p_n(x)| \leq \frac{1}{(n+1)!} \left| \prod_{i=0}^{n} (x - x_i) \right| \left\| f^{(n+1)} \right\|_\infty \quad \text{for all } x \in [a, b].$$

- *As in the case of polynomial interpolation, where we obtained convergence estimates by investigating the decay rate of the term* $\frac{1}{(n+1)!}\left|\prod_{i=0}^{n}(x-x_i)\right|$ *depending on the points* $\{x_i\}_{i=0}^{n}$ *and the interval* $[a,b]$*, we will obtain here convergence estimates by investigating the decay of the power function depending on* X_N*,* Ω*,* K*.*

Proof. We first prove the error bound (4.2). For $f \in \mathcal{H}_K(\Omega)$, using the formulation (4.1) of the interpolant w.r.t. the Lagrange basis and the reproducing property of the kernel K we obtain

$$s_f(x) = \sum_{j=1}^{N} \ell_j(x) f(x_j) = \sum_{j=1}^{N} \ell_j(x)(f, K(\cdot, x_j))_{\mathcal{H}_K(\Omega)} = \left(f, \sum_{j=1}^{N} \ell_j(x) K(\cdot, x_j) \right)_{\mathcal{H}_K(\Omega)}$$

and also

$$f(x) = (f, K(\cdot, x))_{\mathcal{H}_K(\Omega)}.$$

So using bi-linearity of the inner product and the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}
|f(x) - s_f(x)| &= \left| (f, K(\cdot, x))_{\mathcal{H}_K(\Omega)} - \left(f, \sum_{j=1}^{N} \ell_j(x) K(\cdot, x_j) \right)_{\mathcal{H}_K(\Omega)} \right| \\
&= \left| \left(f, K(\cdot, x) - \sum_{j=1}^{N} \ell_j(x) K(\cdot, x_j) \right)_{\mathcal{H}_K(\Omega)} \right| \\
&\leq \|f\|_{\mathcal{H}_K(\Omega)} \left\| K(\cdot, x) - \sum_{j=1}^{N} \ell_j(x) K(\cdot, x_j) \right\|_{\mathcal{H}_K(\Omega)} \\
&= \|f\|_{\mathcal{H}_K(\Omega)} P_{X_N}(x),
\end{aligned}$$

where we used Definition 4.8 of the power function.

We can now prove the formula (4.3) using the fact that $(K(\cdot, x), K(\cdot, y))_{\mathcal{H}_K(\Omega)} = K(x, y)$ (by the reproducing property) :

$$\begin{aligned}
P_{X_N}(x)^2 &= \left\| K(\cdot, x) - \sum_{j=1}^{N} \ell_j(x) K(\cdot, x_j) \right\|_{\mathcal{H}_K(\Omega)}^2 \\
&= \left(K(\cdot, x) - \sum_{j=1}^{N} \ell_j(x) K(\cdot, x_j), K(\cdot, x) - \sum_{i=1}^{N} \ell_i(x) K(\cdot, x_i) \right)_{\mathcal{H}_K(\Omega)} \\
&= K(x, x) - 2 \sum_{j=1}^{N} \ell_j(x) K(x, x_j) + \sum_{i,j=1}^{N} \ell_j(x) \ell_i(x) K(x_j, x_i).
\end{aligned}$$

□

4.3.1 Properties of the power function

We see now some alternative characterization of the power function, which will be useful to prove properties of this error estimation. Moreover, we prove also that the informal definition at the beginning of this section was indeed correct.

We first need the following result on the Lagrange basis.

Proposition 4.11. *The cardinal (or Lagrange) basis of $V(X_N)$ can be obtained by the following formula*

$$\ell_j(x) = \sum_{i=1}^{N} \left(A^{-1}\right)_{ij} K(x, x_i), \quad 1 \leq j \leq N. \tag{4.4}$$

Moreover, for all $x \in \Omega$ the interpolant of the function $f_x := K(\cdot, x)$ is

$$s_{f_x}(\cdot) = \sum_{j=1}^{N} \ell_j(x) K(\cdot, x_j). \tag{4.5}$$

Proof. The first part is a result proven in Numerik I, but we see a short proof. Since the formula (4.4) defines a function in $V(X_N)$, and since the Lagrange basis is unique by Proposition 4.11, we just need to check that with this new formula it holds $\ell_j(x_k) = \delta_{kj}$:

$$\ell_j(x_k) = \sum_{i=1}^{N} (A^{-1})_{ij} K(x_k, x_i) = \sum_{i=1}^{N} (A^{-1})_{ij} A_{ki} = \left(AA^{-1}\right)_{kj} = \delta_{kj}.$$

Now, we take $f_x := K(\cdot, x)$ for $x \in \Omega$ and consider the interpolant s_{f_x}, which is of the form

$$s_{f_x}(\cdot) = \sum_{j=1}^{N} \alpha_j(x) K(\cdot, x_j),$$

where the coefficients $\alpha(x) \in \mathbb{R}^N$ are such that $s_{f_x}(x_k) = f_x(x_k) = K(x_k, x)$ for $1 \leq k \leq N$. We prove that the coefficients are $\alpha_j(x) = \ell_j(x)$. Indeed, we can substitute the formula for the Lagrange basis and obtain

$$\begin{aligned} s_{f_x}(x_k) = \sum_{j=1}^{N} \ell_j(x) K(x_k, x_j) &= \sum_{j=1}^{N} \left(\sum_{i=1}^{N} \left(A^{-1}\right)_{ij} K(x, x_i) \right) K(x_k, x_j) \\ \sum_{i=1}^{N} K(x, x_i) \sum_{j=1}^{N} \left(A^{-1}\right)_{ij} K(x_k, x_j) &= \sum_{i=1}^{N} K(x, x_i) \sum_{j=1}^{N} \left(A^{-1}\right)_{ij} A_{kj} \\ = \sum_{i=1}^{N} K(x, x_i) \delta_{ik} &= K(x, x_k). \end{aligned}$$

Since the interpolant is unique, we can conclude that $s_{f_x}(\cdot) = \sum_{j=1}^{N} \ell_j(x) K(\cdot, x_j)$. □

We can now prove the alternative ways to define the power function.

Proposition 4.12 (Characterization of the power function). *For every $x \in \Omega$, we have*

i) $P_{X_N}(x) = \|f_x - s_{f_x}(\cdot)\|_{\mathcal{H}_K(\Omega)}$ *with* $f_x := K(\cdot, x)$

ii) $P_{X_N}(x) = \sqrt{K(x,x) - \sum_{j=1}^N K(x, x_j)\ell_j(x)}$

iii) $P_{X_N}(x) = \sup_{f \in \mathcal{H}_K(\Omega), f \neq 0} \frac{|f(x) - s_f(x)|}{\|f\|_{\mathcal{H}_K(\Omega)}}$.

Proof. **i** From Definition 4.8 the power function is $P_{X_N}(x) = \left\| K(\cdot, x) - \sum_{j=1}^N K(\cdot, x_j)\ell_j(x) \right\|_{\mathcal{H}_K(\Omega)}$, and from Proposition 4.11 $s_{f_x}(\cdot) = \sum_{j=1}^N \ell_j(x) K(\cdot, x_j)$, so the results follows immediately.

ii We have from equation (4.3) in Theorem 4.9 that the power function can be computed as

$$P_{X_N}(x)^2 = K(x,x) - 2\sum_{j=1}^N \ell_j(x) K(x, x_j) + \sum_{i,j=1}^N \ell_j(x)\ell_i(x) K(x_j, x_i),$$

and we can simplify the last term using Proposition 4.11:

$$\sum_{i,j=1}^N \ell_j(x)\ell_i(x) K(x_j, x_i) = \sum_{j=1}^N \ell_j(x) \sum_{i=1}^N \ell_i(x) K(x_j, x_i) = \sum_{j=1}^N \ell_j(x) s_{f_x}(x_j)$$
$$= \sum_{j=1}^N \ell_j(x) K(x_j, x),$$

since $s_{f_x}(x_j) = f_x(x_j) = K(x, x_j)$.

It follows that $P_{X_N}(x)^2 = K(x,x) - \sum_{j=1}^N \ell_j(x) K(x, x_j)$.

iii From Theorem 4.9 we know already that for all $x \in \Omega$ it holds

$$|f(x) - s_f(x)| \leq P_{X_N}(x) \ \|f\|_{\mathcal{H}_K(\Omega)}, \tag{4.6}$$

so

$$\sup_{f \in \mathcal{H}_K(\Omega), f \neq 0} \frac{|f(x) - s_f(x)|}{\|f\|_{\mathcal{H}_K(\Omega)}} \leq P_{X_N}(x).$$

We just need to find a function for which equality holds,

This works by taking $f := f_x - s_{f_x} = K(\cdot, x) - \sum_{j=1}^N \ell_j(x) K(\cdot, x_j)$. Indeed, the interpolant of f is $s_f = 0$ from Corollary 4.2 (since $f(x_i) = f_x(x_i) - s_{f_x}(x_i) = 0$). It follows that the left hand side of equation (4.6) can be written, using Property ii, as

$$|f_x(x) - s_{f_x}(x)| = K(x,x) - \sum_{j=1}^N \ell_j(x) K(x, x_j) = P_{X_N}(x)^2$$

and the right hand side, using Property i, is

$$P_{X_N}(x)\ \|f\|_{\mathcal{H}_K(\Omega)} = P_{X_N}(x)\,\|f_x - s_{f_x}\|_{\mathcal{H}_K(\Omega)} = P_{X_N}(x)^2.$$

□

Using this new characterizations, we can prove that the error bound provided by the power function is good, in the sense that it is bounded, it is exact on the interpolation points, and it is decreasing when the number of interpolation points increases.

Proposition 4.13 (Properties of the power function)**.** *Let $X_N \subset \Omega$ pairwise distinct. Then we have*

i) $P_{X_N}(x) \leq \sqrt{K(x,x)}$ *for all $x \in \Omega$ (the error is bounded for all $x \in \Omega$)*

ii) $P_{X_N}(x) = 0$ *if and only if $x \in X_N$ (the bound is exact in the interpolation points)*

iii) If $Y \subset \Omega$ is finite and $X_N \subsetneq Y$, then $P_Y(x) < P_{X_N}(x)$ for all $x \in \Omega \setminus X_N$ (the error is strictly decreasing if the interpolation set is increasing)

Proof. We use Property (i) of Proposition 4.12, i.e., $P_{X_N}(x) = \|f_x - s_{f_x}\|_{\mathcal{H}_K(\Omega)}$ with $f_x := K(\cdot, x)$.

i From Proposition 4.6 the norm of the interpolation error is smaller than the norm of f. Thus

$$P_{X_N}(x) = \|f_x - s_{f_x}\|_{\mathcal{H}_K(\Omega)} \leq \|f_x\|_{\mathcal{H}_K(\Omega)} = \sqrt{(K(\cdot,x), K(\cdot,x))_{\mathcal{H}_K(\Omega)}} = \sqrt{K(x,x)}.$$

ii $P_{X_N}(x) = 0$ if and only if $\|f_x - s_{f_x}\|_{\mathcal{H}_K(\Omega)} = 0$ if and only if $f_x = s_{f_x}$, and the interpolation is exact if and only if $f_x \in V(X_N)$. Since $f_x = K(\cdot, x)$, this is possible if and only if $x_i \in X_N$, because $K(\cdot, x)$ is linearly independent from $\{K(\cdot, x_i)\}_{i=1}^N$ if $x \notin X_N$.

iii From Proposition 4.6, interpolation on a larger set gives smaller error, so $P_Y(x) < P_{X_N}(x)$.

□

Remark 4.14. *Observe that Property ii does not mean that the interpolation error can be zero only on the interpolation points. Indeed, the error bound is*

$$|f(x) - s_f(x)| \leq P_{X_N}(x)\ \|f\|_{\mathcal{H}_K(\Omega)} \quad \textit{for all } x \in \Omega,$$

so the power function is only an upper bound on the error. It can happen that the left hand side is zero for some $x \in \Omega$, while the right hand side is not.

The result only means that the upper bound on the right is guaranteed to be zero (so, exact) when $x \in X_N$.

Finally, we see another characterization of the power function. It would be the key element in the error bound we will prove in Section 4.5. Observe that in the book [1] this is directly used as a definition of the power function, and the other definitions are derived from this one.

Proposition 4.15 (Power function and quadratic form). *For $X_N \subset \Omega$ pairwise distinct and $x \in \Omega$, consider the vectors*

$$b(x) := [K(x, x_1), K(x, x_2), \dots, K(x, x_N)]^T \in \mathbb{R}^N, \quad u^*(x) := [\ell_1(x), \ell_2(x), \dots, \ell_N(x)]^T \in \mathbb{R}^N.$$

For $u \in \mathbb{R}^N$ define the quadratic form

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

Then it holds

$$P_{X_N}(x) = \sqrt{Q(u^*(x))} = \min_{u \in \mathbb{R}^N} \sqrt{Q(u)}.$$

Proof. If we substitute the vector $u^*(x)$ in the definition of Q, we obtain from equation 4.3 in Theorem 4.9 that $P_{X_N}(x) = \sqrt{Q(u^*(x))}$.

We can then consider the quadratic form Q and minimize it w.r.t. $u \in \mathbb{R}^N$: We have

$$d_u(Q(u)) = d_u\left(K(x,x) - 2u^T b(x) + u^T A u\right) = 2Au - 2b(x).$$

Thus $d_u(Q(u)) = 0$ if and only if $Au = b(x)$, i.e., if and only if $u = A^{-1}b(x)$, i.e., from the definition of $b(x)$ and Proposition 4.11

$$u_j = \sum_{i=1}^{N} (A^{-1})_{ij} K(x, x_i) = \ell_j(x).$$

From the definition of $u^*(x)$, this means that $u^*(x)$ minimizes Q. □

4.4 General stability bounds

We conclude this part on general results (i.e., applicable to any SPD kernel) by stating a result on the stability of the interpolation process. This result is indeed a bound on the largest and smaller eigenvalues of the kernel matrix.

Theorem 4.16 (Condition number of kernel matrix). *Let $X_N \subset \Omega$ be pairwise distinct, A the corresponding kernel matrix and $\lambda_{\min}$, $\lambda_{\max}$ be its minimal and maximal eigenvalues. For every $x_i \in X_N$, consider the power functions $P_{X_N \setminus \{x_j\}}$ of the sets $X_N \setminus \{x_j\}$.*

Then the condition number of A is $\kappa(A) = \lambda_{\max}/\lambda_{\min}$, and it holds

$$\lambda_{\max} \leq N \max_{1 \leq j \leq N} K(x_j, x_j)$$

and

$$\lambda_{\min}(A) \leq \min_{1 \leq j \leq N} P_{X_N \setminus \{x_j\}}(x_j)^2 \tag{4.7}$$

Proof. Since A is symmetric positive definite, its singular values are the eigenvalues, so $\kappa(A) = \lambda_{\max}/\lambda_{\min}$.

For the maximal eigenvalue we use the Gershgorin circle theorem: for all $1 \leq i \leq N$, we have

$$|\lambda_{\max} - A_{ii}| \leq \sum_{j=1, j\neq i}^{N} |A_{ij}|,$$

thus

$$\lambda_{\max} \leq N \max_{1\leq i,j\leq N} |A_{ij}| = N \max_{1\leq i,j\leq N} |K(x_i, x_j)| \leq N \max_{1\leq j\leq N} |K(x_j, x_j)|,$$

since $K(x, y)^2 \leq K(x, x)K(x, y)$ for all $x, y \in \Omega$ from Proposition 2.5.

For the minimal eigenvalue we use the characterization of the power function of Proposition 4.15. We consider for simplicity the case $x_j = x_N$ and use the notation $X_{N-1} := X_N \setminus \{x_N\}$. Moreover, we denote as $A_{N-1}$the kernel matrix of X_{N-1}, and as $b_{N-1}(x)$ and $u^*_{N-1}(x)$ the vectors of Proposition 4.15 corresponding to the points X_{N-1}. We have

$$\begin{aligned} P_{X_{N-1}}(x_N)^2 &= (u^*_{N-1}(x_N))^T A_{N-1} u^*_{N-1}(x_N) - 2(u^*_{N-1}(x_N))^T b_{N-1}(x_N) + K(x_N, x_N) \\ &= \left[(u^*_{N-1}(x_N))^T, -1\right] \begin{bmatrix} A_{N-1} & b_{N-1}(x_N) \\ b_{N-1}(x_N)^T & K(x_N, x_N) \end{bmatrix} \begin{bmatrix} u^*_{N-1}(x_N) \\ -1 \end{bmatrix}. \end{aligned}$$

Since $b_{N-1}(x_N) := [K(x_N, x_1), K(x_N, x_2), \ldots, K(x_N, x_N)]^T$, we have that the matrix in the bilinear form is exactly A, i.e., the kernel matrix of the full set of points X_N. Then it holds

$$\begin{aligned} P_{X_{N-1}}(x_N)^2 &= \left[(u^*_{N-1}(x_N))^T, -1\right] A \begin{bmatrix} u^*_{N-1}(x_N) \\ -1 \end{bmatrix} \geq \lambda_{\min} \left\| \left[(u^*_{N-1}(x_N))^T, -1\right]^T \right\|_2^2 \\ &\geq \lambda_{\min} \left(\left\| u^*_{N-1}(x_N) \right\|_2^2 + 1 \right) \geq \lambda_{\min}, \end{aligned}$$

since $u^T A u \geq \lambda_{\min} \|u\|_2^2$ for all $u \in \mathbb{R}^N$. □

Remark 4.17. *The relation* (4.7) *between the power function and the minimal eigenvalue has been known for a lot of time as trade-off principle, because it shows that it is impossible to obtain good approximation error (i.e., small power function) and in the same time to have also good conditioning of the kernel matrix (i.e., large minimal eigenvalue).*

But this is true only if the interpolant is computed by solving the full linear system $A\alpha = b$, while this trade-off is not present if one computes the interpolant in other ways.

This is the only way to compute the interpolant that we have seen so far, but in the next we will see some algorithms that try to mitigate the problem by avoiding the solution of the full linear system.

4.5 Error bounds

We can finally obtain the full error bounds. As in the case of polynomial interpolation, the plan is to start from the error bound

$$|f(x) - s_f(x)| \leq P_{X_N}(x) \, \|f\|_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega,$$

and to obtain bounds on the power function by investigating the dependence on the kernel K, the points X_N and the set Ω, and by identifying "good" kernels, points, and sets.

4.5.1 Interpolation points

The quality of the points can be measured by looking at how well distributed they are. In dimension $d = 1$ it is easy: we have an interval $[a, b]$ and points $a \leq x_1 < x_2 < \cdots < x_N \leq b$ with distances $h_i := x_i - x_{i-1}$, $2 \leq i \leq N$, and $h_0 := x_1 - a$, $h_{N+1} := b - x_N$, and we can define the grid size as

$$h := \max_{0 \leq i \leq N+1} h_i.$$

To have good interpolation, we should consider a set of points X_N with small grid size, otherwise there are "large holes" in $[a, b]$. This idea can be generalized to higher dimensions $d \geq 1$ by considering instead the fill distance, which is the radius of the largest ball B with center in Ω, and such that $B \cap \Omega$ does not contain any point of X_N.

Definition 4.18 (Fill distance). *Let $X_N \subset \Omega$. The fill distance of X_N in Ω is defined as*

$$h_N := h_{X_N,\Omega} := \sup_{x \in \Omega} \min_{x_j \in X_N} \|x - x_j\|_2^2$$

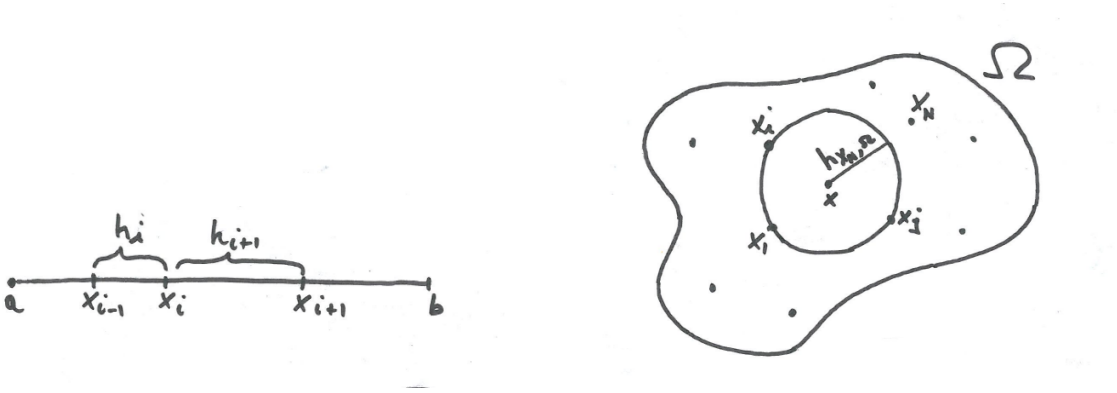

Figure 4.1: Grid size and fill distance.

4.5.2 Interpolation set

On Ω, instead, we need the following assumption.

Definition 4.19 (Interior cone condition). *A set $\Omega \subset \mathbb{R}^d$ satisfies an interior cone condition if there is a radius $r > 0$ and an angle $\theta \in (0, \pi/2)$ such that the cone $C(r, \theta)$ of radius r and angle θ can be centered at any point in Ω and rotated such that $C(r, \theta) \subset \bar{\Omega}$.*

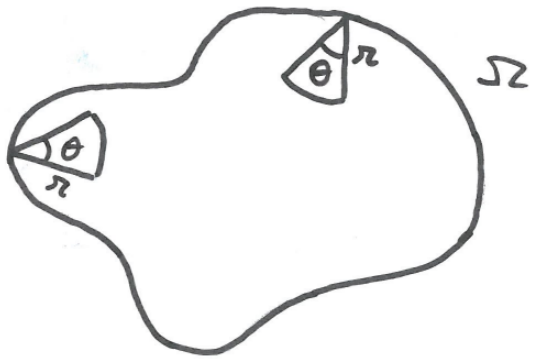

Figure 4.2: Set satisfying a cone condition with radius r and angle θ.

This condition guarantees the existence of a stable local polynomial reproduction (Theorem 3.14 in [5]). It just follows from geometrical assumptions on Ω, and it is not necessarily related to kernels.

Proposition 4.20. *Assume $\Omega \subset \mathbb{R}^d$ is bounded and satisfies an interior cone condition with angle $\theta \in (0, \pi/2)$ and radius $r > 0$. Let $m \in \mathbb{N}$. Then there exist constants $c_1, c_2, h_0 > 0$ such that for all X_N with fill distance $h_N \leq h_0$ and for all $x \in \Omega$ there exists $\tilde{u}(x) := [\tilde{u}_1(x), \dots \tilde{u}_N(x)]^T \in \mathbb{R}^N$ such that*

i) $\sum_{j=1}^N \tilde{u}_j(x) p(x_j) = p(x)$ *for all polynomial p of degree m on $\mathbb{R}^d$*

ii) $\sum_{j=1}^N |\tilde{u}_j(x)| \leq c_1$

iii) $\tilde{u}_j(x) = 0$ *if* $\|x - x_j\|_2 > c_2 h_N$.

4.5.3 Error bound

We finally have the error bound.

Theorem 4.21 (Error bound). *Let $\Omega \subset \mathbb{R}^d$ be open, bounded, and satisfying a cone condition. Assume that $X_N \subset \Omega$ has fill distance h_N. Let $k \in \mathbb{N}$ and $K : \Omega \times \Omega \to \mathbb{R}$ be a SPD kernel with $K \in C^{2k}(\Omega \times \Omega)$. Let $a \in \mathbb{N}_0^d$ with $|a| \leq k$.*

Then there exists constants $C, h_0 > 0$, independent of $x, K, f \in \mathcal{H}_K(\Omega)$, such that if $h_N \leq h_0$ it holds

$$|D^a f(x) - D^a s_f(x)| \leq C\, C_K\, h_{X_N}^{k-|a|}\, \|f\|_{\mathcal{H}_K(\Omega)} \quad \textit{for all } x \in \Omega. \tag{4.8}$$

The constant C_K is defined as

$$C_K^2 := \max_{\substack{b,c\, \in \mathbb{N}_0^d \\ |b|+|c|=2k}} \left(\max_{y,z \in \Omega \cap B(x, ch_N)} \left| D_1^b D_2^c K(z,y) \right| \right)$$

where $c = c_2$ of Proposition 4.20 if $|a| = 0$.

Proof. We see only a sketch of the proof for the case $|a| = 0$. The complete proof can be found in [5], Theorem 11.13, or [1], Theorem 14.5 (only for the case $|a| = 0$).

The idea is to start from the bound of Proposition 4.9, i.e.,

$$|f(x) - s_f(x)| \leq P_{X_N}(x)\ \|f\|_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega.$$

Then the goal is to find a bound on the power function. We can use Proposition 4.15, i.e.,

$$P_{X_N}(x)^2 = Q(u^*(x)) = \min_{u \in \mathbb{R}^N} Q(u).$$

In particular, for all $x \in \Omega$, we have

$$P_{X_N}(x)^2 \leq Q(\tilde{u}(x)),$$

where $\tilde{u}(x)$ is the vector of Proposition 4.20, which exists because Ω satisfies an interior cone condition. This means

$$\begin{aligned} P_{X_N}(x)^2 \leq Q(\tilde{u}(x)) &= K(x,x) - 2\sum_{j=1}^{N} \tilde{u}_j(x) K(x, x_j) + \sum_{i,j=1}^{N} \tilde{u}_j(x)\tilde{u}_i(x) K(x_j, x_i) \\ &= K(x,x) - 2\sum_{j \in J(x)} \tilde{u}_j(x) K(x, x_j) + \sum_{i,j \in J(x)}^{N} \tilde{u}_j(x)\tilde{u}_i(x) K(x_j, x_i), \end{aligned}$$

where $J(x)$ is the set of indexes such that $\tilde{u}_j(x) \neq 0$ from Property (iii). This means that all the x_j that appear in the sums are at most at distance $\|x - x_j\|_2 \leq c_2 h_N$ from x.

The rest of the proof is to compute Taylor expansions up to order k of the terms $K(x, x_j)$ and $K(x_i, x_j)$, which can be done since the kernel is in $C^{2k}(\Omega \times \Omega)$.

Then the final bound can be obtained by using the polynomial reproduction property (i) of Proposition 4.20 to sum the polynomial parts of the Taylor expansions, and by bounding the remainders using (ii) of Proposition 4.20 and the constant C_K. □

Remark 4.22. *Some remarks on the Theorem:*

- *The error bound states that one should have "well distributed" points X_N (i.e., small fill distance) to obtain good approximation.*

- *The error bound tells us that, by kernel interpolation, not only one obtains pointwise convergence, but also convergence of all the derivatives up to a certain order.*
- *As can be expected, the rate of convergence is reduced of a factor 1 for every derivative. This is reasonable as the smoothness of the target function is reduced every time a derivative is taken.*
- *This is a worst-case error bound, i.e., it applies to all functions $f \in \mathcal{H}_K(\Omega)$. In practice, it is possible to have single functions for which the convergence is faster.*
- *The fill distance is a dimension dependent quantity, since to have a fill distance h_N in $\Omega \subset \mathbb{R}^d$, one needs to have $N = \mathcal{O}(h_N^{-d})$ points.*

Finally, we have seen the idea of the proof for the case $|a| = 0$. The more general case follows from the same argument by using the following generalization of the power function. It is completely similar to the case of the power function defined above, but it makes things more complicated. So we just introduce its definition in the following, and we just prove that it is a good definition for an upper bound.

Proposition 4.23 (Power function for derivatives)**.** *Let $X_N \subset \Omega$ pairwise distinct and $f \in \mathcal{H}_K(\Omega)$. Let $k \in \mathbb{N}$ and $K \in C^{2k}(\Omega \times \Omega)$, Let $a \in \mathbb{N}_0^d$ with $|a| \leq k$. Then a generalized power function can be defined as*

$$P^a_{X_N}(x) := \left\| D_2^a K(\cdot, x) - \sum_{j=1}^N K(\cdot, x_j) D^a \ell_j(x) \right\|_{\mathcal{H}_K(\Omega)} . \tag{4.9}$$

and it holds

$$|D^a f(x) - D^a s_f(x)| \leq P^a_{X_N}(x) \|f\|_{\mathcal{H}_K(\Omega)}, \quad \text{for all } x \in \Omega. \tag{4.10}$$

Moreover

$$P^a_{X_N}(x) = \sqrt{Q^a(D^a u^*(x))} = \min_{u \in \mathbb{R}^N} \sqrt{Q^a(u)}.$$

with

$$Q^a(u) := D_1^a D_2^a K(x, x) - 2 \sum_{j=1}^N u_j D_1^a K(x, x_j) + \sum_{i,j=1}^N u_j u_i K(x_j, x_i).$$

Proof. We only see the proof of the error bound. We have from Property iii of Proposition 3.14 that, if $\Omega \subset \mathbb{R}^d$ is open and $K \in C^{2k}(\Omega \times \Omega)$, then for all $a \in \mathbb{N}_0^d$ with $|a| \leq k$ it holds

$$D^a f(x) = (f, D_2^a K(\cdot, x))_{\mathcal{H}_K(\Omega)} .$$

Moreover, we have

$$D^a s_f(x) = D^a \left(\sum_{j=1}^N \ell_j(x) f(x_j) \right) = \sum_{j=1}^N D^a \ell_j(x) f(x_j) = \left(f, \sum_{j=1}^N D^a \ell_j(x) K(\cdot, x_j) \right)_{\mathcal{H}_K(\Omega)}$$

Then we can conclude by Cauchy- Schwarz:

$$\begin{aligned}
|D^a f(x) - D^a s_f(x)| &= \left| (f, D^a K(\cdot, x))_{\mathcal{H}_K(\Omega)} - \left(f, \sum_{j=1}^{N} D^a \ell_j(x) K(\cdot, x_j) \right)_{\mathcal{H}_K(\Omega)} \right| \\
&= \left| \left(f, D^a K(\cdot, x) - \sum_{j=1}^{N} D^a \ell_j(x) K(\cdot, x_j) \right)_{\mathcal{H}_K(\Omega)} \right| \\
&\leq \|f\|_{\mathcal{H}_K(\Omega)} P^a_{X_N}(x).
\end{aligned}$$

□

5. Translational invariant and RBF kernels

We consider now a special classes of kernels, i.e., translational invariant kernels. They are of great interest in approximation for several reasons. Indeed, they include as a special case Radial Basis Function (RBF) kernels, which are practically the most used kernels for interpolation, as they are easy to implement in terms of the distance matrix (see Remark 2.11). Moreover, the analysis in this case has very strong connections to the theory of Fourier transform and Sobolev spaces.

Definition 5.1 (Translational invariant and RBF kernels)**.** *A kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is translational invariant if*

$$K(x, y) = K(x + z, y + z) \text{ for all } x, y, z \in \mathbb{R}^d.$$

It is radial (or radially invariant) if

$$K(x, y) = K(z, w) \text{ for all } x, y, z, w \in \mathbb{R}^d \text{ with } \|x - y\|_2 = \|z - w\|_2.$$

5.1 Characterization of translational invariant and radial kernels

These two classes of kernels can be represented in a more convenient form as follows.

Proposition 5.2 (Characterization of translational and radial invariance)**.** *A kernel K is translational invariant if and only if there exists $\Phi : \mathbb{R}^d \to \mathbb{R}$ such that $K(x, y) = \Phi(x-y)$ for all $x, y \in \mathbb{R}^d$.*

A kernel K is radial if and only if there exists $\Phi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ such that $K(x, y) = \Phi(\|x - y\|_2)$ for all $x, y \in \mathbb{R}^d$.

Proof. We see the two cases separately:

(Transl.) If $K(x, y) = \Phi(x - y)$ clearly it holds $K(x + z, y + z) = \Phi(x + z - y - z) = \Phi(x - y) = K(x, y)$, so K is translational invariant.

Assume instead $K(x, y) = K(x + z, y + z)$ for all $x, y, z \in \mathbb{R}^d$. For a fixed $x_0 \in \mathbb{R}^d$, set $\Phi(x) := K(x_0, x_0 - x)$. Using $z := -x + x_0$, it follows that

$$K(x, y) = K(x - x + x_0, y - x + x_0) = K(x_0, x_0 - (x - y)) = \Phi(x - y).$$

The function Φ is independent of the particular choice of x_0, since for any $x_0' \in \mathbb{R}^d$ it holds

$$\Phi(x) := K(x_0, x_0 - x) = K(x_0 + (x_0' - x_0), x_0 - x + (x_0' - x_0)) = K(x_0', x_0' - x).$$

(Rotat.) If $K(x,y) = \Phi(\|x-y\|_2)$ clearly it holds $K(x,y) = K(z,w)$ for all $x,y,z,w \in \mathbb{R}^d$ with $\|x-y\|_2 = \|z-w\|_2$, so K is radial.

Assume instead $K(x,y) = K(z,w)$ for all $x,y,z,w \in \mathbb{R}^d$ with $\|x-y\|_2 = \|z-w\|_2$. For fixed $x_0 \in \mathbb{R}^d$, set $\Phi(r) := K(x_0, x_0 + re_1)$, with $e_1 \in \mathbb{R}^d$ the first unit vector. Using $z := x_0$ and $w := x_0 + \|x-y\|_2 e_1$ we have $\|x-y\|_2 = \|z-w\|_2$, so it follows that

$$K(x,y) = K(x_0, x_0 + \|x-y\|_2 e_1) = \Phi(\|x-y\|_2).$$

Again, the definition is independent of x_0, e_1: if $x_0', v \in \mathbb{R}^d$ and $\|v\|_2 = 1$, we have

$$\Phi(r) := K(x_0, x_0 + re_1) = K(x_0', x_0' + rv),$$

since $\|x_0 + re_1 - x_0\|_2 = \|x_0' + rv - x_0'\|_2$.

□

Remark 5.3. *Some remarks on the definitions:*

- *A radial kernel is usually called a Radial Basis Function (RBF) kernel. Sometimes there is confusion of notions between Φ and K, which are both called RBF.*
- *It is clear that RBF kernels are in particular translational invariant, as $K(x+z, y+z) = \Phi(\|x+z-y-z\|_2) = \Phi(\|x-y\|_2) = K(x,y)$.*
- *RBF kernels are usually defined up to a scaling factor $\varepsilon > 0$, i.e., they are defined as $K(x,y) := \Phi(\varepsilon\|x-y\|_2)$. This parameter is called shape parameter, and it does not change the radial invariance, but allows to control the "support" of the kernel.*
- *The representation of K in terms of Φ is useful because it allows to deduce properties of K by looking at the possibly simpler function Φ (which is univariate in the case of RBF kernels). Moreover, in the case of RBF kernels, it makes also implementation easier, as we have seen with the Gaussian kernel.*
- *We will see how to obtain PD or SPD of K from some properties of Φ. In the case K is a positive or strictly positive definite kernel, the function Φ will be called a positive or strictly positive definite function. In the case of RBF kernels, since the function Φ is univariate, it is called positive or strictly positive definite function on $\mathbb{R}^d$ if the kernel $K(x,y) = \Phi(\|x-y\|_2)$ is PD or SPD for $x,y \in \mathbb{R}^d$.*

5.2 Translational invariant PD kernels and Fourier transform

We will now provide a connection between PD of translational invariant kernels and the Fourier transform of the function Φ.

First, recall the definition of the Fourier transform (see Analysis I/II/III).

Definition 5.4 (Fourier transform). *For $f \in L_1(\mathbb{R}^d)$ we define the Fourier transform of f as*

$$\hat{f}(\omega) := Ff(\omega) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{-ix^T\omega} dx, \ \ \omega \in \mathbb{R}^d.$$

and the inverse Fourier transform as

$$\check{f}(\omega) := \left(F^{-1}f(\omega) :=\right) (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{ix^T\omega} dx, \ \ \omega \in \mathbb{R}^d.$$

The following is the more general characterization of positive definiteness of K in terms of Φ, and it is considered a fundamental theorem in the theory of kernels. We just see its statement.

Theorem 5.5 (Bochner). *Let Φ be a continuous function. Then $K(x, y) := \Phi(x - y)$ is positive definite if and only if Φ is the Fourier transform of a non-negative, finite, Borel measure μ, i.e.,*

$$\Phi(x) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-ix^T\omega} d\mu(\omega), \ x \in \mathbb{R}^d.$$

5.2.1 A more simple characterization

There is a much more simple version of this theorem under further assumptions on the function Φ, and it will apply to all kernels of interest.

Recall that, in the case of the Gaussian kernel, we have proven in Proposition 2.10 that it is strictly positive definite by proving that it is the Fourier transform of a positive function f. This means that it is the Fourier transform of the measure $d\mu(\omega) := f(\omega)d\omega$, which is finite and non-negative. This idea can be generalized.

First, we need to recall some properties of the Fourier transform (see again Analysis I/II/III or [5, Chapter 5]).

Proposition 5.6 (Properties of Fourier transform). *Let $f \in L_1(\mathbb{R}^d)$*

i) $\hat{f}$ is continuous.

ii) If $f(x) = f(-x)$ for all $x \in \mathbb{R}^d$, then $\hat{f}(x) \in \mathbb{R}$.

iii) If also $\hat{f} \in L_1(\mathbb{R}^d)$ then $F^{-1}(\hat{f}) = f$.

iv) If $g \in L_1(\mathbb{R}^d)$ then $\int_{\mathbb{R}^d} f(x)\hat{g}(x)dx = \int_{\mathbb{R}^d} \hat{f}(x)g(x)dx$.

v) If $a \in \mathbb{R}^d$, then $F(f)(x - a) = e^{-ix^T a}\hat{f}(x)$.

We now see some properties of translational invariant PD kernels that will be useful later, when computing their Fourier transforms.

Proposition 5.7 (Properties of PD translational invariant kernels). *Let $K(x, y) = \Phi(x - y)$ be a PD translational invariant kernel on $\mathbb{R}^d$. Then for all $x \in \mathbb{R}^d$ it holds*

i) $\Phi(x) = \Phi(-x)$, *thus* $\hat{\Phi}(x) \in \mathbb{R}$.

ii) $\Phi(0) \geq 0$.

iii) $|\Phi(x)| \leq \Phi(0)$.

iv) $\Phi(0) = 0$ *implies* $\Phi = 0$.

Proof. The first property is a direct consequence of the definition and of symmetry of K:

i $\Phi(x) = \Phi(x - 0) = K(x, 0) = K(0, x) = \Phi(0 - x) = \Phi(-x)$.

The other ones follow from the general properties of PD kernels proven in Proposition 2.5

ii $0 \leq K(x, x) = \Phi(x - x) = \Phi(0)$.

iii $|\Phi(x)| = |K(x, 0)| \leq \sqrt{K(x, x)}\sqrt{K(0, 0)} = \sqrt{\Phi(x - x)}\sqrt{\Phi(0 - 0)} = \Phi(0)$.

iv It follows from point iii.

□

Finally, we need the following tool. The idea is to have a proper formalization of the concept that "the δ function is the identity of the convolution", i.e.,

$$\text{“}(f * \delta)(x) = \int_{\mathbb{R}^d} f(y)\delta(y - x)dy = f(x)\text{”},$$

(which is well defined only in the sense of distributions). A possible formalization is the concept of approximation by convolution (see again Analysis I/II/III or [5, Chapter 5]).

Proposition 5.8 (Approximation by convolution)**.** *Let* $m \in \mathbb{N}$ *and* $g_m(x) := (m/\pi)^{d/2} e^{-m\|x\|_2^2}$, $x \in \mathbb{R}^d$. *Then the following hold:*

i) $\int_{\mathbb{R}^d} g_m(x) = 1$.

ii) $\hat{g}_m(x) = (2\pi)^{-d/2} e^{-\|x\|_2^2/(4m)}$.

iii) $F\left(F(g_m)\right)(x) = g_m(x)$.

iv) $f(x) = \lim_{m\to\infty}(f * g_m)(x) := \lim_{m\to\infty} \int_{\mathbb{R}^d} f(y)g_m(y - x)dy$ *if* f *is slowly increasing, i.e., there exists* $n \in \mathbb{N}$ *such that* $f(x) = \mathcal{O}(\|x\|_2^n)$ *for* $\|x\|_2 \to \infty$.

Finally, we can state the Corollary of Theorem 5.5.

Corollary 5.9. *Let* $\Phi : \mathbb{R}^d \to \mathbb{R}$, $\Phi \in L_1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$. *Then* $K(x, y) := \Phi(x - y)$ *is PD if and only if* Φ *is bounded and* $\hat{\Phi}(\omega) \geq 0$ *for all* $\omega \in \mathbb{R}^d$. *It is SPD if and only if* $\hat{\Phi}(\omega) > 0$ *for all* $\omega \in \mathbb{R}^d$.

Proof. We prove only one implication, i.e, that Φ with these properties results in K PD. The other one can be found in [5], Theorem 6.11.

If we prove that $\hat{\Phi} \in L_1(\mathbb{R}^d)$, we can conclude by Property (iii) of Proposition 5.6 that

$$\Phi(x) = F^{-1}(\hat{\Phi})(x) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{ix^T\omega} \hat{\Phi}(\omega) d\omega.$$

This is enough to prove PD or SPD of K. Indeed, we can use the same proof as in Proposition 2.10 for the Gaussian kernel, i.e., for a set X_N of pairwise distinct points, we have

$$\begin{aligned}
\alpha^T A \alpha &= \sum_{j,l=1}^{N} \alpha_j \alpha_l K(x_j, x_l) = \sum_{j,l=1}^{N} \alpha_j \alpha_l \Phi(x_j - x_l) \\
&= (2\pi)^{-d/2} \sum_{j,l=1}^{N} \alpha_j \alpha_l \int_{\mathbb{R}^d} e^{i(x_j - x_l)^T \omega} \hat{\Phi}(\omega) d\omega \\
&= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \left(\sum_{j,l=1}^{N} \alpha_j \alpha_l e^{i(x_j - x_l)^T \omega} \right) \hat{\Phi}(\omega) d\omega \\
&= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \left| \sum_{j=1}^{N} \alpha_j e^{ix_j^T \omega} \right|^2 \hat{\Phi}(\omega) d\omega.
\end{aligned}$$

Since X_N are pairwise distinct, we have $\left| \sum_{j=1}^{N} \alpha_j e^{ix_j^T \omega} \right|^2 > 0$. Thus $\alpha^T A \alpha \geq 0$ if $\hat{\Phi} \geq 0$, and $\alpha^T A \alpha > 0$ if $\hat{\Phi} > 0$.

To prove that $\hat{\Phi} \in L_1(\mathbb{R}^d)$, we use the approximation by convolution (iv) of Proposition 5.8, which can be applied because $|\Phi(x)| \leq \Phi(0)$ ((iii) of Proposition 5.7), so the function is slowly increasing. We have

$$\begin{aligned}
(2\pi)^{d/2} \Phi(0) &= (2\pi)^{d/2} \lim_{m \to \infty} (\phi * g_m)(0) = (2\pi)^{d/2} \lim_{m \to \infty} \int_{\mathbb{R}^d} \Phi(y) g_m(y - 0) dy \\
&= (2\pi)^{d/2} \lim_{m \to \infty} \int_{\mathbb{R}^d} \Phi(y) g_m(y) dy.
\end{aligned}$$

Now we can use Property (iv) of Proposition 5.6, since $F\left(F(g_m)\right)(x) = g_m(x)$ by (iii) of Proposition 5.8:

$$(2\pi)^{d/2} \Phi(0) = (2\pi)^{d/2} \lim_{m \to \infty} \int_{\mathbb{R}^d} \Phi(y) g_m(y) dy = (2\pi)^{d/2} \lim_{m \to \infty} \int_{\mathbb{R}^d} \hat{\Phi}(y) \hat{g}_m(y) dy,$$

and use the Monotone Convergence Theorem to exchange the integral and the limit

$$\begin{aligned}
(2\pi)^{d/2} \Phi(0) &= (2\pi)^{d/2} \lim_{m \to \infty} \int_{\mathbb{R}^d} \hat{\Phi}(y) \hat{g}_m(y) dy = (2\pi)^{d/2} \int_{\mathbb{R}^d} \lim_{m \to \infty} \left(\hat{\Phi}(y) \hat{g}_m(y) \right) dy \\
&= \int_{\mathbb{R}^d} \hat{\Phi}(y) \left(\lim_{m \to \infty} (2\pi)^{d/2} \hat{g}_m(y) \right) dy = \int_{\mathbb{R}^d} \hat{\Phi}(y) dy,
\end{aligned}$$

since $\lim_{m\to\infty}(2\pi)^{d/2}\hat{g}_m(x) = 1$ for all $x \in \mathbb{R}^d$ by (ii) of Proposition 5.8.

It follows that $\int_{\mathbb{R}^d} \hat{\Phi}(y)dy = (2\pi)^{d/2}\Phi(0)$, so $\hat{\Phi} \in L_1(\mathbb{R}^d)$ since Φ is bounded by hypothesis. □

Remark 5.10. *The direction we proved is the most interesting for our aims, as we will be able to prove that a kernel is (S)PD by proving that Φ is bounded and $\hat{\Phi}(x) \geq 0$. The other direction is useful only if one wants to construct a new kernel starting from a particular Φ.*

5.3 Sobolev spaces and native spaces

We move now to the connection between native spaces of translational invariant kernels and Sobolev spaces.

We first recall that the Sobolev space $H^s(\mathbb{R}^d)$ has the following representation. Observe that it is well defined also for $s \notin \mathbb{N}$, and it holds $H^{\lceil s \rceil}(\mathbb{R}^d) \subset H^s(\mathbb{R}^d) \subset H^{\lfloor s \rfloor}(\mathbb{R}^d)$.

Proposition 5.11 (Sobolev spaces)**.** *There is the following representation of the Sobolev space $H^s(\mathbb{R}^d)$:*

$$H^s(\mathbb{R}^d) := \left\{ f \in L_2(\mathbb{R}^d) : \hat{f}(\cdot)(1 + \|\cdot\|_2^2)^{s/2} \in L_2(\mathbb{R}^d) \right\}$$

with the inner product

$$(f, g)_{H^s(\mathbb{R}^d)} := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \hat{f}(\omega)\overline{\hat{g}(\omega)}(1 + \|\omega\|_2^2)^s d\omega.$$

Moreover, it holds $H^s(\mathbb{R}^d) \subset C(\mathbb{R}^d)$ if $s > d/2$.

We have seen in Chapter 3 the construction of the native space $\mathcal{H}_K(\Omega)$. In the case $\Omega = \mathbb{R}^d$ and K is translational invariant, there is also an alternative definition of this space, which makes use of the Fourier transform.

Theorem 5.12 (Native spaces via Fourier transform)**.** *Assume $\Phi \in L_1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$ is a real valued, strictly positive definite function and let $K(x, y) := \Phi(x - y)$ for all $x, y \in \mathbb{R}^d$.*

Define

$$\mathcal{G} := \left\{ f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \frac{\hat{f}}{\sqrt{\hat{\Phi}}} \in L_2(\mathbb{R}^d) \right\}$$

and the bilinear form

$$(f, g)_{\mathcal{G}} := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\Phi}(\omega)} d\omega.$$

Then $\mathcal{G} = \mathcal{H}_K(\mathbb{R}^d)$, in the sense that the sets coincide and $(\cdot, \cdot)_{\mathcal{G}} = (\cdot, \cdot)_{\mathcal{H}_K(\mathbb{R}^d)}$.

Proof. We see only a sketch of the proof. The complete one can be found in Theorem 10.12 in [5].

First, observe that the definition is well posed, since Φ SPD implies $\hat{\Phi} > 0$ (Corollary 5.9).

Moreover, we have from the proof of Corollary 5.9 that $\hat{\Phi} \in L_1(\mathbb{R}^d)$. Since $\frac{\hat{f}}{\sqrt{\hat{\Phi}}} \in L_2(\mathbb{R}^d)$ by assumption, we have that also $\hat{f} \in L_1(\mathbb{R}^d)$ (by Cauchy-Schwarz):

$$\int_{\mathbb{R}^d} |\hat{f}(\omega)| d\omega = \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|}{\sqrt{\hat{\Phi}(\omega)}} \sqrt{\hat{\Phi}(\omega)} d\omega \leq \left(\int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\Phi}(\omega)} d\omega \right)^{1/2} \left(\int_{\mathbb{R}^d} |\hat{\Phi}(\omega)| d\omega \right)^{1/2} < \infty.$$

In particular, from iii of Proposition 5.6 this implies that

$$f(x) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \hat{f}(\omega) e^{i\omega^T x} d\omega \quad \text{for all } x \in \Omega.$$

The step that we skip is to prove that $\mathcal{G}$ with inner product $(\cdot, \cdot)_{\mathcal{G}}$ is a Hilbert space of functions on $\mathbb{R}^d$.

We only prove the reproducing property. We use the fact that $\hat{\Phi} \in \mathbb{R}$ ((i) in Proposition 5.7) and that $F(f)(x - a) = e^{-ia^T x} \hat{f}(x)$ ((v) in proposition 5.6):

$$\begin{aligned}
(f, K(\cdot, x))_{\mathcal{G}} &:= (f, \Phi(\cdot - x))_{\mathcal{G}} := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\hat{f}(\omega) \overline{\hat{\Phi}(\omega - x)}}{\hat{\Phi}(\omega)} d\omega \\
&= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\hat{f}(\omega) \overline{\hat{\Phi}(\omega) e^{-i\omega^T x}}}{\hat{\Phi}(\omega)} d\omega = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\hat{f}(\omega) \hat{\Phi}(\omega) e^{i\omega^T x}}{\hat{\Phi}(\omega)} d\omega \\
&= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \hat{f}(\omega) e^{i\omega^T x} d\omega = f(x).
\end{aligned}$$

Now, since $\mathcal{G}$ with inner product $(\cdot, \cdot)_{\mathcal{G}}$ is a Hilbert space, and since K is the reproducing kernel with respect to $(\cdot, \cdot)_{\mathcal{G}}$, it follows that $\mathcal{G} = \mathcal{H}_K(\mathbb{R}^d)$ by Theorem 3.6. □

We conclude with the following corollary, which is indeed the main result in this section.

Corollary 5.13 (Sobolev spaces and native spaces)**.** *Let $\Phi \in L_1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$. Assume that, for a given $s > d/2$, there exists constants $c_1, c_2 > 0$ such that*

$$c_1(1 + \|\omega\|_2^2)^{-s} \leq \hat{\Phi}(\omega) \leq c_2(1 + \|\omega\|_2^2)^{-s} \quad \textit{for all } \omega \in \mathbb{R}^d. \tag{5.1}$$

Let $K(x, y) = \Phi(x - y)$. Then $\mathcal{H}_K(\mathbb{R}^d)$ is norm equivalent to $H^s(\mathbb{R}^d)$, i.e., they are equal as sets and, for all $f \in \mathcal{H}_K(\mathbb{R}^d)$, the norms satisfy

$$c_2^{-1} \|f\|_{H^s(\mathbb{R}^d)} \leq \|f\|_{\mathcal{H}_K(\mathbb{R}^d)} \leq c_1^{-1} \|f\|_{H^s(\mathbb{R}^d)}.$$

Proof. First, Φ is SPD thanks to Corollary 5.9. Indeed, it is bounded because it is in $L_1(\mathbb{R}^d) \cap C(\mathbb{R}^d)$, and $\hat{\Phi} > 0$ because of the condition (5.1).

Now we prove that $f \in \mathcal{H}_K(\mathbb{R}^d)$ if and only if $f \in H^s(\mathbb{R}^d)$. By Theorem 5.12, we have $f \in \mathcal{H}_K(\mathbb{R}^d)$ if and only if

$$f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \text{ and } \frac{\hat{f}}{\sqrt{\hat{\Phi}}} \in L_2(\mathbb{R}^d).$$

By the assumption (5.1), this is true if and only if

$$f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) \text{ and } \hat{f}(\cdot)(1 + \|\cdot\|_2^2)^{s/2} \in L_2(\mathbb{R}^d).$$

Thanks to Theorem 5.11, this means that $f \in H^s(\mathbb{R}^d)$ (since $s > d/2$ implies $H^s(\mathbb{R}^d) \subset C(\mathbb{R}^d)$).

For all $f, g \in \mathcal{H}_K(\mathbb{R}^d)$, we have from Theorem 5.12 that

$$(f, g)_{\mathcal{H}_K(\mathbb{R}^d)} = (f, g)_{\mathcal{G}} := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\Phi}(\omega)} d\omega.$$

We can then use the lower bound in condition (5.1) to obtain

$$\begin{aligned}(f, g)_{\mathcal{H}_K(\mathbb{R}^d)} = (f, g)_{\mathcal{G}} &:= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\Phi}(\omega)} d\omega \\ &\leq (2\pi)^{-d/2} \int_{\mathbb{R}^d} \hat{f}(\omega)\overline{\hat{g}(\omega)} c_1^{-1} (1 + \|\omega\|_2^2)^s d\omega \\ &= c_1^{-1} (f, g)_{H^s(\mathbb{R}^d)},\end{aligned}$$

using the norm $(f, g)_{H^s(\mathbb{R}^d)}$ of Theorem 5.11. The other direction of the norm equivalence works in the same way. □

Remark 5.14. *Two remarks on this part:*

- *The result of the corollary can be extended to subsets $\Omega \subset \mathbb{R}^d$ if the boundary of Ω is regular enough. We don't consider a formal statement here.*
- *In the case of RBF kernels, the function Φ is even univariate. This allows to simplify Bochner Theorem and Corollary 5.9 even further (Theorem 6.18 in [5]).*

We conclude this section by showing some examples of functions Φ such that the kernel is a SPD radial basis function kernel. The shape parameter ε is omitted, but all the kernels can be defined as $K(x, y) := \Phi(\varepsilon(x - y))$. A more complete list of kernels of this type can be found in Appendix D of [1].

The Matérn kernels are called Sobolev kernels, as their Fourier transform is exactly of the form of Corollary 5.13, without the need of constants c_1, c_2.

Instead, in the case of the Gaussian and (Generalized) Inverse Multiquadrics, the decay of the Fourier transform is faster than any polynomial, so Corollary 5.13 does not apply.

Name	Φ	$\hat{\Phi}$
Gaussian	$\Phi(x) = e^{-\|x\|_2^2}$	$\hat{\Phi}(\omega) = \frac{1}{\sqrt{2}} e^{-\|\omega\|_2^2/4}$
BasicMatérn $\beta = \frac{d+1}{2}$	$\Phi(x) = e^{-\|x\|_2}$	$\hat{\Phi}(\omega) = (1 + \|\omega\|_2^2)^{-\beta}$
Linear Matérn $\beta = \frac{d+3}{2}$	$\Phi(x) = (1 + \|x\|_2) e^{-\|x\|_2}$	$\hat{\Phi}(\omega) = (1 + \|\omega\|_2^2)^{-\beta}$
Quadratic Matérn $\beta = \frac{d+5}{2}$	$\Phi(x) = (3 + 3\|x\|_2 + \|x\|_2^2) e^{-\|x\|_2}$	$\hat{\Phi}(\omega) = (1 + \|\omega\|_2^2)^{-\beta}$
Inverse Multiquadrics (IMQ)	$\Phi(x) = 1/\left(\sqrt{1 + \|x\|_2^2}\right)$	a positive Bessel function
Generalized Inv. Multiq. $\beta > 0$	$\Phi(x) = (1 + \|x\|_2^2)^{-\beta}$	a positive Bessel function

5.4 Compactly supported RBF kernels

We discuss now a particular class of RBF kernels, known as Wendland kernels. They have some special features, which make them very attractive:

- They are RBF kernels, so they can be easily implemented via distance matrices.
- They have compact support, so the kernel matrix can be sparse.
- They are polynomials inside their support, so they are easy to compute and fast to evaluate.
- They have native spaces which are norm equivalent to Sobolev spaces (of integer order in odd space dimension).

5.4.1 Remarks on compactly supported kernels

The construction of compactly supported kernels need to consider the following reasoning.

- The following can be proven: Assume $\Phi : [0, \infty) \to \mathbb{R}$ is continuous and SPD on every $\mathbb{R}^d$ (i.e., $K(x, y) := \Phi(\|x - y\|_2)$ is SPD for $x, y \in \mathbb{R}^d$, for all d). If there exists $r_0 \in [0, \infty)$ such that $\Phi(r_0) = 0$, then $\Phi = 0$.
- This seems to suggest that radial and compactly supported SPD kernels cannot exist.
- Instead, we have the following: A continuous and compactly supported function $\Phi : [0, \infty) \to \mathbb{R}$ cannot be SPD on every $\mathbb{R}^d$.

- Moreover, we have seen in (i) of Proposition 2.7 that if a kernel K is SPD on Ω, it is also SPD on $\Omega' \subset \Omega$. This implies in particular: If $\Phi : [0, \infty) \to \mathbb{R}$ is SPD on $\mathbb{R}^d$, then it is also SPD on all $\mathbb{R}^{d'}$, with $d' \leq d$.
- Putting all together: If a continuous and compactly supported function $\Phi : [0, \infty) \to \mathbb{R}$ is SPD on $\mathbb{R}^d$, there exists a maximal d' such that Φ is SPD on $\mathbb{R}^{d'}$ for all $d \leq d'$, but not on $\mathbb{R}^d$ with $d > d'$.

5.4.2 Wendland kernels

The Wendland kernels satisfy this requirement, and they start by considering the following function.

Proposition 5.15 (Truncated power)**.** *Let $l \in \mathbb{N}$. The function $\Phi_l : \mathbb{R}_{\geq 0} \to \mathbb{R}$ defined as*

$$\Phi_l(r) := (1-r)_+^l := \begin{cases} (1-r)^l, & \text{if } r \leq 1 \\ 0, & \text{if } r > 1 \end{cases} \tag{5.2}$$

Is strictly positive definite on $\mathbb{R}^d$ if $l \geq \lfloor d/2 \rfloor + 1$ (i.e., the kernel $K(x, y) := \Phi_l(\|x - y\|_2)$ is SPD on $\mathbb{R}^d$).

Proof. Theorem 6.20 in [5]. □

Remark 5.16. *Some comments:*

- *For $l \in \mathbb{N}$, the kernel $K(x, y) := \Phi_l(\|x - y\|_2)$ is C^0 but not differentiable in $x = y$ (i.e., in $r = 0$).*
- *The goal is to generalize to*

$$\Phi(r) = \begin{cases} p(r), & \text{if } r \leq 1 \\ 0, & \text{if } r > 1, \end{cases}$$

  with p a polynomial of minimal order, in a way such that the kernel is still SPD, the differentiability is increased, and the evaluation is still fast.
- *The idea is to integrate the function Φ_l in a proper way, by introducing an integral operator. It will increases the degree of a factor 2. This operation is sometimes called dimension walk.*

The integration operation is defined as follows.

Definition 5.17 (Wendland functions)**.** *For a function $f : [0, \infty) \to \mathbb{R}$ such that $t \to tf(t) \in L_1([0, \infty))$, define*

$$(If)(r) := \int_r^\infty tf(t)dt, \quad r \geq 0.$$

Let $d \in \mathbb{N}$ and define $l := \lfloor d/2 \rfloor + k + 1$. Then the Wendland functions are defined as

$$\Phi_{d,k} := I^k \Phi_l = I(I(\dots(\Phi_l))).$$

With this operation, one obtains the Wendland kernels, which indeed satisfy all the requirements that we are looking for.

Theorem 5.18 (Wendland kernels). *For $d, k \in \mathbb{N}$, the Wendland kernels are defined as $K(x, y) := \Phi_{d,k}(\|x - y\|_2)$.*

They are SPD on $\mathbb{R}^d$, and $K \in C^{2k}(\mathbb{R}^d \times \mathbb{R}^d)$. Moreover, the functions $\Phi_{d,k}$ have the representation

$$\Phi_{d,k}(r) = \begin{cases} p_{d,k}(r), & \text{if } r \leq 1 \\ 0, & \text{if } r > 1, \end{cases}$$

with $p_{d,k}$ a polynomial of degree $\lfloor d/2 \rfloor + 3k + 1$. Moreover, $p_{d,k}$ are of minimal degree for a given d and k, and they are unique up to constants.

Proof. Theorem 9.13 in [5]. □

The explicit form of the polynomials can be easily computed from Definition 5.17. We see some examples here.

Proposition 5.19 (Explicit form of Wendland kernels). *Let $d, k \in \mathbb{N}$ and $l := \lfloor d/2 \rfloor + k + 1$. There exists constants $c_{d,k} > 0$ such that*

$$\begin{aligned} \Phi_{d,0}(r) &= (1 - r)_+^l \\ \Phi_{d,1}(r) &= c_{d,1} (1 - r)_+^{l+1} [(l + 1)r + 1] \\ \Phi_{d,2}(r) &= c_{d,2} (1 - r)_+^{l+2} [(l + 1)(l + 3)r^2 + 3(l + 2)r + 3] \end{aligned}$$

Proof. We use Definition 5.17 and Proposition 5.15, i.e., $\Phi_{d,k} := I^k \Phi_l$.

For $k = 0$ there is no integration, so $\Phi_{d,0}(r) = \Phi_l(r)$ for all d.

For $k = 1$ we consider only $r \in [0, 1]$:

$$\begin{aligned} \Phi_{d,1}(r) := I\Phi_l(r) &:= \int_r^\infty t(1 - t)_+^l dt = \int_r^1 t(1 - t)^l dt \\ &= \left[\frac{-(1 - t)^{l+1}((l + 1)t + 1)}{2 + 3l + l^2} \right]_{t=r}^{t=1} = \frac{(1 - r)^{l+1}((l + 1)r + 1)}{2 + 3l + l^2}, \end{aligned}$$

while $\Phi_{d,1}(r) = 0$ for $r > 0$, so we get the formula in the statement.

The same for $k = 2$ and $r \in [0, 1]$:

$$\begin{aligned} \Phi_{d,2}(r) := I^2\Phi_l(r) = I\Phi_{d,1}(r) &= \int_r^1 t \, \frac{(1 - t)^{l+1}((l + 1)t + 1)}{2 + 3l + l^2} dt \\ &= \left[\frac{(1 - t)^{l+2}((l + 1)(l + 3)t^2 + 3(l + 2)t + 3)}{(l + 1)(l + 2)(l + 3)(l + 4)} \right]_{t=r}^{t=1} \\ &= \frac{(1 - r)^{l+2}((l + 1)(l + 3)r^2 + 3(l + 2)r + 3)}{(l + 1)(l + 2)(l + 3)(l + 4)}, \end{aligned}$$

while $\Phi_{d,2}(r) = 0$ for $r > 0$, so we get the formula in the statement.

□

Finally, these kernels are included in the framework of Section 5.3, so we can characterize their native spaces.

Theorem 5.20 (Native space of Wendland kernels)**.** *Let $d, k \in \mathbb{N}$, with $d \geq 3$ if $k = 0$.*

There exists constants $c_1, c_2 > 0$ depending only on d, k such that

$$c_1(1 + \|\omega\|_2^2)^{-d/2-k-1/2} \leq \hat{\Phi}_{d,k}(\omega) \leq c_2(1 + \|\omega\|_2^2)^{-d/2-k-1/2} \quad \text{for all } \omega \in \mathbb{R}^d.$$

In particular, the native space on $\mathbb{R}^d$ of the Wendland kernel $K(x, y) := \Phi_{d,k}(x - y)$ is norm equivalent to the Sobolev space $H^{d/2+k+1/2}(\mathbb{R}^d)$.

Remark 5.21. *Observe the following things:*

- *The exponent is an integer in odd space dimension.*
- *One can expect a space of smoothness C^k, because of $K \in C^{2k}$ (Theorem 3.14). But we get the additional factor $d/2 + 1/2$ (in the sense of weak derivatives).*

5.5 Error bounds revisited

We now come back to the error bound for interpolation of Theorem 4.21, and give a refined version in the case of some RBF kernels.

The idea to obtain these refined error estimates is to carefully bound the constant C_K appearing in the error bound, which depends on the derivatives of the kernel. The hypotheses of the following theorems are exactly the ones of Theorem 4.21, but we write them again for completeness.

Theorem 5.22 (Convergence order for translational invariant kernels)**.** *Let $K(x, y) := \Phi(x - y)$ be a SPD, translational invariant kernel on $\mathbb{R}^d$.*

Assume that $\Phi \in C_\nu^{2k}(\mathbb{R}^d)$, i.e., $\Phi \in C^{2k}(\mathbb{R}^d)$ and, for $a \in \mathbb{N}_0^d$ with $|a| = 2k$, $D^a\Phi(x) = \mathcal{O}(\|x\|_2^\nu)$ for $\|x\|_2 \to 0$.

Let $\Omega \subset \mathbb{R}^d$ be open, bounded, and satisfying a cone condition. Assume that $X_N \subset \Omega$ has fill distance h_N. Let $a \in \mathbb{N}_0^d$ with $|a| \leq k$.

Then there exists constants $C, h_0 > 0$, independent of $x, K, f \in \mathcal{H}_K(\Omega)$, such that if $h_N \leq h_0$ it holds

$$|D^a f(x) - D^a s_f(x)| \leq C\, h_{X_N}^{k+\nu/2-|a|}\, \|f\|_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega.$$

In particular we have the following cases for Wendland, Gaussian and Inverse Multiquadric kernels.

Theorem 5.23 (Convergence order for Wendland kernels)**.** *Let $K(x, y) := \Phi_{d,k}(\|x - y\|_2)$ be a Wendland kernel with $k, d \in \mathbb{N}$.*

Let $\Omega \subset \mathbb{R}^d$ be open, bounded, and satisfying a cone condition. Assume that $X_N \subset \Omega$ has fill distance h_N. Let $a \in \mathbb{N}_0^d$ with $|a| \leq k$.

Then there exists constants $C, h_0 > 0$, independent of $x, K, f \in \mathcal{H}_K(\Omega)$, such that if $h_N \leq h_0$ it holds

$$|D^a f(x) - D^a s_f(x)| \leq C\, h_{X_N}^{k+1/2-|a|}\, \|f\|_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega.$$

Theorem 5.24 (Convergence order for Gaussian and IMQ kernels)**.** *Let $K(x,y) := \Phi(\|x-y\|_2)$ be the Gaussian or Inverse Multiquadric.*

Let $\Omega \subset \mathbb{R}^d$ be open, bounded, and satisfying a cone condition. Assume that $X_N \subset \Omega$ has fill distance h_N. Let $a \in \mathbb{N}_0^d$.

Then for all $l \in \mathbb{N}$ with $l \geq |a|$ there exists constants $C(l), h_0(l) > 0$ such that if $h_N \leq h_0(l)$ it holds

$$|D^a f(x) - D^a s_f(x)| \leq C(l)\, h_N^{l-|a|}\, \|f\|_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega.$$

In the case $|a| = 0$, even spectral convergence holds, i.e., there exists $c, C > 0$ such that

$$|f(x) - s_f(x)| \leq C\, e^{-\frac{c}{h_N}}\, \|f\|_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega.$$

Remark 5.25 (Convergence order for Wendland kernels)**.** *The convergence rate of Theorem 5.23 are obtained by keeping ε constant, hence the support of the RBF kernel is fixed. This means that, when the fill distance is reduced, the kernel becomes no more compactly supported, so the benefits on the computational side are lost. Nevertheless, in practical applications one usually solves an interpolation problem with a fixed set of interpolation points X_N, so the support can be chosen accordingly.*

Remark 5.26 (Quasi-optimal convergence order)**.** *In the case of Wendland kernels, the native space is a Sobolev space $H^s(\mathbb{R}^d)$, with $s = d/2 + k + 1/2$. From the theorem above, we obtain convergence of order $\gamma := k + 1/2$ for $|a| = 0$. This means convergence of order $\gamma = s - d/2$.*

We have seen that, to have a fill distance h_N, we need to consider $N = \mathcal{O}(h_N^{-1/d})$ points. A sequence of sets $\{X_N\}_{N\in\mathbb{N}}$ such that $h_N \leq cN^{-1/d}$ for all $N \in \mathbb{N}$ is called asymptotically uniform (so, very well distributed).

In this case, substituting in the error bound this h_N, we obtain an order of convergence

$$h_{X_N}^{\gamma} = h_{X_N}^{s-d/2} \leq c' N^{-\frac{s}{d}+\frac{1}{2}}.$$

Here, it is evident that the convergence order is dimension-dependent. But also, it is difficult to do more than this: It can be proven that in a Sobolev space $H^s(\mathbb{R}^d)$, the best possible convergence (using any method, not necessarily kernels) is of order $N^{-s/d}$.

6. Algorithms for kernel interpolation

We have seen so far that kernel interpolation is well defined in arbitrary space dimension and using arbitrary pairwise distinct interpolation points. Moreover, we have studied error analysis in the case the target function comes from the native space of the kernel, and that such space can be closely related to certain Sobolev spaces in some cases.

The goal of this part is to present some algorithms that allow to compute the kernel interpolant (or an approximant) in an efficient and stable way. They will make use of the theory of Reproducing Kernel Hilbert Spaces that we have studied in the previous chapters, but they all results in actual algorithms that can be implemented. Some comments:

- This is a selection of possible algorithms. Many others exist, but the following ones present a reasonable set of ideas that can be found also in other methods.
- The following methods all have advantages and disadvantages, which will be discussed. In general, when choosing a particular method, one should think to use the method that better fits the particular application.
- All the methods avoid the solution of the linear system $A\alpha = b$.
- Regularized interpolation considers weaker interpolation conditions and improves the condition number of the kernel matrix.
- Partition of Unity method divides the interpolation problem on Ω into several interpolation problems on subsets $\{\Omega^{(i)}\}_{i=1}^{M}$, and then combines the resulting interpolants.
- Greedy kernel interpolation selects a small subset $X_n \subset X_N$ of the interpolation points and solve the interpolation problem restricted to X_n.

6.1 General considerations

We first see two tools that are very useful in practice, and which are independent of the particular method.

6.1.1 Train/validation/test sets

In many methods the approximation depends on some parameters, which need to be chosen to obtain good results. An example is the shape parameter $\varepsilon > 0$ in a RBF

kernel, but also other parameters can be present. Moreover, there is the need to test the quality of the approximant.

In practical applications, the target function f is unknown, so it cannot be used to check if the approximation is good. All we know is the set of interpolation points $X_N \subset \Omega$ and the corresponding data values, which we denote here as $F_N := \{f_i\}_{i=1}^N$. In this case, the most common approach is to split the sets into train, validation and test sets.

This means the following: first permute the two sets X_N, F_N, then fix numbers N_{tr}, N_{val}, N_{te} such that $N = N_{tr} + N_{val} + N_{te}$, and define a partition of X_N, F_N as

$$\begin{array}{llll} X_{tr} & := \{x_i, 1 \le i \le N_{tr}\} & F_{tr} & := \{f_i, 1 \le i \le N_{tr}\} \\ X_{val} & := \{x_i, N_{tr}+1 \le i \le N_{tr}+N_{val}\} & F_{val} & := \{f_i, N_{tr}+1 \le i \le N_{tr}+N_{val}\} \\ X_{te} & := \{x_i, N_{tr}+N_{val}+1 \le i \le N\} & F_{te} & := \{f_i, N_{tr}+N_{val}+1 \le i \le N\}. \end{array}$$

The idea is to use the validation set X_{val} to validate (i.e., choose) the parameters, and the test set X_{te} to test the error. Having disjoint sets allows to have a fair way to test the algorithm.

For the process, we also need an error function that returns the error of the interpolant s_f evaluated on a generic set of points $X := \{x_i\}_i$ w.r.t. the exact values $F := \{f_i\}_i$. We denote as $|X|$ the number of elements of X. Examples are the maximal error and the Root Mean Square Error (RMSE) defined as

$$E(s_f, X, F) := \max_{1 \le i \le |X|} |s_f(x_i) - f_i|, \quad E(s_f, X, F) := \sqrt{\frac{1}{|X|} \sum_{i=1}^{|X|} (s_f(x_i) - f_i)^2}.$$

Then, one decides a set of possible parameters $\{\varepsilon_1, \dots, \varepsilon_{N_\varepsilon}\}$ that has to be checked. A common choice is to take them logarithmically equally spaced, since the correct scale is not known in advance, in general.

The training and validation process is described in Algorithm 1. We denote as $s_f(\varepsilon_i)$ the interpolant obtained with parameter ε_i.

Algorithm 1 Validation and test

1: Input: $X_{tr}, X_{val}, X_{te}, F_{tr}, F_{val}, F_{te}, \{\varepsilon_1, \dots, \varepsilon_{N_\varepsilon}\}$
2: **for** $i = 1, \dots, N_\varepsilon$ **do**
3: Compute interpolant $s_f(\varepsilon_i)$ with data (X_{tr}, F_{tr})
4: Compute error $e_i := E(s_f(\varepsilon_i), X_{val}, F_{val})$
5: **end for**
6: Choose parameter $\bar{\varepsilon} := \arg\min e_i$
7: Compute interpolant $s_f(\bar{\varepsilon})$ with data $(X_{tr} \cup X_{val}, F_{tr} \cup F_{val})$
8: Compute error $\bar{e} = E(s_f(\bar{\varepsilon}), X_{te}, F_{te})$
9: Output: interpolant $s_f(\bar{\varepsilon})$, optimal parameter $\bar{\varepsilon}$, test error $\bar{e}$

A more advanced way to realize the same idea is k-fold cross validation. To have an even better selection of the parameters, one can repeat the validation step (lines

2-6 in the previous algorithm) by changing the validation set at each step. To do so, we don't select a validation set (so $N = N_{tr} + N_{te}$), and instead consider a partition of X_{tr}, F_{tr} into $k \in \{1, \dots, N\}$ disjoint subsets, all approximately of the same size:

$$\begin{array}{llll} X_{tr} & := \{x_i, 1 \le i \le N_{tr}\} := \cup_{i=1}^{k} X_i & F_{tr} & := \{f_i, 1 \le i \le N_{tr}\} := \cup_{i=1}^{k} F_i \\ X_{te} & := \{x_i, N_{tr}+1 \le i \le N_{tr}+N_{te}\} & F_{te} & := \{f_i, N_{tr}+1 \le i \le N_{tr}+N_{te}\} \end{array}$$

In the validation step, each of the the X_i is used as a validation set, and the validation is repeated for all i. The error e_i for the parameter ε_i is then defined as the average error over all these permutations, as described in Algorithm 2.

In the case $k = N$, k-fold cross validation is called Leave One Out Cross Validation (LOOCV).

Algorithm 2 k-fold cross validation and test

Input: $X_{tr} = \cup_{i=1}^{k} X_i, X_{te}$, $F_{tr} = \cup_{i=1}^{k} F_i, F_{te}$, $\{\varepsilon_1, \dots, \varepsilon_{N_\varepsilon}\}$
for $i = 1, \dots, N_\varepsilon$ **do**
  for $j = 1, \dots, k$ **do**
    Compute interpolant $s_f(\varepsilon_i)$ with data $(\cup_{i \neq j} X_i, \cup_{i \neq j} F_i)$
    Compute error $e^{(j)} := E(s_f(\varepsilon_i), X_j, F_j)$
  end for
  $e_i := \text{mean}\{e^{(j)}, 1 \le j \le k\}$
end for
Choose parameter $\bar{\varepsilon} := \arg\min e_i$
Compute interpolant $s_f(\bar{\varepsilon})$ with data (X_{tr}, F_{tr})
Compute error $\bar{e} = E(s_f(\bar{\varepsilon}), X_{te}, F_{te})$
Output: interpolant $s_f(\bar{\varepsilon})$, optimal parameter $\bar{\varepsilon}$, test error $\bar{e}$

6.1.2 Vector valued functions

A second thing that can be applied to all the methods is the following. We have seen so far how to approximate functions $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$. It is easy (at least in the computational sense) to extend the process to the approximation of vector-valued functions $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}^q$, $q \ge 1$.

We still consider parwise distinct data points $X_N \subset \Omega$, but in this case the data values are vectors $f(x_i) \in \mathbb{R}^q$. It is still possible to construct an interpolant of the form

$$s_f(x) = \sum_{j=1}^{N} \alpha_j K(x, x_j), \quad x \in \Omega,$$

but now considering coefficient vectors $\alpha_j \in \mathbb{R}^q$. Imposing the interpolation conditions results in a linear system $A\alpha = b$, where A is still the same kernel matrix, but

instead α and b are $N \times q$ matrices defined as

$$\alpha := \begin{bmatrix} \vdots \\ \alpha_j^T \\ \vdots \end{bmatrix} \in \mathbb{R}^{N \times q}, \; b := \begin{bmatrix} \vdots \\ f(x_i)^T \\ \vdots \end{bmatrix} \in \mathbb{R}^{N \times q}.$$

The existence of a unique solution is still guaranteed by the fact that A is a positive definite matrix.

This is a very simple way to deal with vector-valued functions. It is an instance of the use of matrix-valued kernels, which are a generalization of the kernels that we have seen so far. A large part of the theory can be extended also to cover this case.

6.2 Regularized interpolation

Some ideas:

- The interpolant is obtained by requiring that $s_f(x_i) = f_i$, $1 \leq i \leq N$. But we have seen that the resulting linear system can be very ill-conditioned.
- Moreover, if the data are affected by noise, i.e., $f_i = f(x_i) + \eta$, it makes no sense to require exact interpolation.
- The interpolation conditions can be written also as the minimization of the quantity $\sum_{i=1}^{N} (f_i - s_f(x_i))^2$. The idea is to relax this condition, in the sense that we still require this quantity to be small, but we add a term that penalizes solutions with large norm.
- Also in this case, the solution can be found by solving a linear system, which will be better conditioned than the usual one.

We give the following definition.

Definition 6.1 (Regularized interpolant)**.** *Let Ω be a nonempty set, K a PD kernel on $\Omega \times \Omega$. Let $X_N \subset \Omega$ be pairwise distinct, $\{f_i = f(x_i)\}_{i=1}^N \subset \mathbb{R}$, and let $\lambda \geq 0$ be a regularization parameter.*

A regularized interpolant s_f^λ is a solution of

$$s_f^\lambda(\cdot) := \operatorname*{arg\,min}_{s \in \mathcal{H}_K(\Omega)} \left[\sum_{i=1}^{N} (f_i - s(x_i))^2 + \lambda \|s\|_{\mathcal{H}_K(\Omega)}^2 \right].$$

The following theorem characterizes exactly the solution(s) of this problem.

Theorem 6.2 (Representer Theorem)**.** *In the setting of Definition 6.1 and if $f \in \mathcal{H}_K(\Omega)$, there exists a regularized interpolant (i.e., a solution of the minimization problem) of the form*

$$s_f^\lambda(\cdot) = \sum_{j=1}^{N} \alpha_j K(\cdot, x_j),$$

where the vector of coefficients $\alpha \in \mathbb{R}^N$ is a solution of the linear system

$$(A + \lambda I)\alpha = b, \quad b_i = f(x_i). \tag{6.1}$$

Moreover, if K is SPD this is the unique solution of the minimization problem

Proof. Define the functional $J : \mathcal{H}_K(\Omega) \to \mathbb{R}$

$$J(s) := \sum_{i=1}^{N} (f_i - s(x_i))^2 + \lambda \|s\|^2_{\mathcal{H}_K(\Omega)}.$$

We first prove that for every $s \in \mathcal{H}_K(\Omega)$ there exists $g \in V(X_N)$ such that $J(g) \leq J(s)$. To see this, using Corollary (4.2) we write $s \in \mathcal{H}_K(\Omega)$ as

$$s = g + g^\perp, \ g \in V(X_N), g^\perp \in V(X_N)^\perp,$$

where $g^\perp(x_i) = 0$ for all $x_i \in X_N$, so $s(x_i) = g(x_i) + g^\perp(x_i) = g(x_i)$, and $\|s\|^2_{\mathcal{H}_K(\Omega)} = \|g\|^2_{\mathcal{H}_K(\Omega)} + \|g^\perp\|^2_{\mathcal{H}_K(\Omega)}$. So we can obtain the following (since $\lambda \geq 0$):

$$\begin{aligned} J(g) &= \sum_{i=1}^{N} (f_i - g(x_i))^2 + \lambda \|g\|^2_{\mathcal{H}_K(\Omega)} = \sum_{i=1}^{N} (f_i - s(x_i))^2 + \lambda \|g\|^2_{\mathcal{H}_K(\Omega)} \\ &= \sum_{i=1}^{N} (f_i - s(x_i))^2 + \lambda \|s\|^2_{\mathcal{H}_K(\Omega)} - \lambda \|g^\perp\|^2_{\mathcal{H}_K(\Omega)} = J(s) - \lambda \|g^\perp\|^2_{\mathcal{H}_K(\Omega)} \leq J(s). \end{aligned}$$

So we can restrict the minimization over $V(X_N)$, i.e., consider only functions

$$s := \sum_{j=1}^{N} \alpha_j K(\cdot, x_j),$$

for some unknown $\alpha \in \mathbb{R}^N$. For these functions it holds

$$s(x_i) = \sum_{j=1}^{N} \alpha_j K(x_i, x_j) = (A\alpha)_i \quad \text{and} \quad \|s\|^2_{\mathcal{H}_K(\Omega)} = \sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) = \alpha^T A \alpha.$$

This means that the functional J depends only on $\alpha \in \mathbb{R}^N$ and can be written as

$$\begin{aligned} J(\alpha) &= \|A\alpha - b\|_2^2 + \lambda \alpha^T A \alpha = (A\alpha - b)^T (A\alpha - b) + \lambda \alpha^T A \alpha \\ &= \alpha^T A^T A \alpha - 2\alpha^T A^T b + b^T b + \lambda \alpha^T A \alpha. \end{aligned}$$

We can minimize it by solving a finite dimensional optimization problem. Indeed, its derivative w.r.t. α is (A is symmetric)

$$d_\alpha(J(\alpha)) = 2A^T A\alpha - 2A^T b + 2\lambda A\alpha = 2A(A\alpha - b + \lambda\alpha),$$

so $d_\alpha(J(\alpha)) = 0$ if and only if

$$A\left(A + \lambda I\right)\alpha = Ab,$$

which is satisfied by α such that $\left(A + \lambda I\right)\alpha = b$ (also for PD kernels). If K is SPD then A is invertible, so this is the only solution. □

Remark 6.3. *Some comments on this result:*

- *This is a generalization of interpolation, in the sense that for $\lambda = 0$ we obtain exact interpolation when K is SPD.*
- *On the other hand, the parameter $\lambda \geq 0$ improves the condition number of the linear system, and thus the stability of the solution. Indeed, the condition number of $A + \lambda I$ is*
$$\kappa(\lambda) := \frac{\lambda_{\min}(A + \lambda I)}{\lambda_{\max}(A + \lambda I)} = \frac{\lambda_{\min}(A) + \lambda}{\lambda_{\max}(A) + \lambda}$$
and this is a strictly decreasing function of λ, with $\kappa(0) = \kappa(A)$ and $\lim_{\lambda \to \infty} \kappa(\lambda) = 1$.
- *Regularized interpolation with $\lambda > 0$ allows to solve approximation problems also with PD kernels. Indeed, the matrix 6.1 is invertible since the minimal eigenvalue is*
$$\lambda_{\min}(A) + \lambda \geq \lambda > 0$$
since by Remark 2.4 $\lambda_{\min}(A) \geq 0$ for K PD.
- *From the proof it follows also that $J(s_f^\lambda) \leq J(f)$. So in this sense regularized interpolation is a form of unrestricted best approximation, in the same spirit of the discussion in Section 4.1. Indeed, we removed both the requirement $s_f^\lambda \in V(X_N)$ (which we obtain back from the Representer Theorem) and $s_f^\lambda(x_i) = f(x_i)$.*
- *The parameter λ is usually also chosen via k-fold cross validation.*
- *Especially in the Engineering / applied sciences literature, kernel approximation is generally referred exactly to this method. In particular, if interpolation ($\lambda = 0$) is a better solution, it is selected by a proper cross validation.*
- *On the other hand, this method does not reduce the size of the linear system to be solved, so one should consider to do something else, or in addition, in the case N is very large.*
- *The functional J can be written as the sum of a risk functional L (the term with the squared error) and a regularization functional R (the term with the squared norm). It is possible to define other regularized approximant by changing these functionals, and the Representer Theorem applies to a wide class of them (see Chapters 3,4 of [3]). Each pair of functionals results in a different way to determine the coefficients α (like 6.1 here).*
- *The idea of computing an approximant by minimizing a weighted sum of a regularization and loss functional is used in many other fields, and usually is referred to as Tikhonov regularization (see e.g. lectures on Inverse Problems). In particular, the regularization functionals can be defined to enforce particular structure in the approximant (for example total variation minimization for image reconstruction).*
- *There is a Demo in ILIAS demonstrating the behavior of the method.*

6.3 Partition of unity method

Some ideas:

- When the number N of points is large, it is difficult to solve the full linear system. Also just storing the full matrix can be very expensive.
- In many cases, the target function f can have local features, so it could be not a good idea to compute a global solution.
- The idea of the Partition of Unity Method (PUM) is to divide the interpolation problem over Ω into several smaller problems defined on local domains $\Omega^{(j)}$, $1 \leq j \leq M$. On each $\Omega^{(j)}$, one considers only the interpolation points $X_j := X_N \cap \Omega^{(j)}$ and solves the associated interpolation problem. The global solution is then obtained by combining the local solutions.
- The straightforward approach of just dividing Ω into disjoint sets would fail, as the global approximant would be in general discontinuous (Figure 6.1).

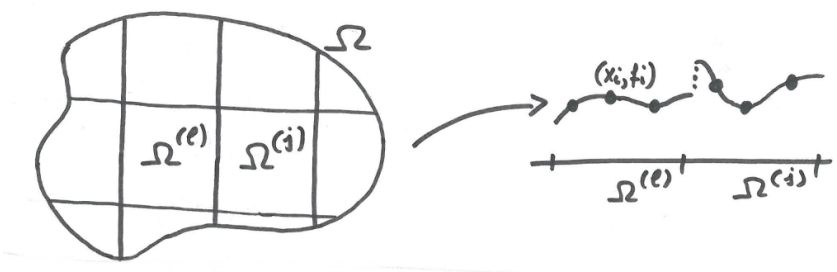

Figure 6.1: A wrong approach to PUM.

The method works as follows. First, we consider an open covering of Ω.

Definition 6.4 (Open covering of Ω)**.** *Let $\Omega \subset \mathbb{R}^d$. A collection of sets $\left\{\Omega^{(j)}\right\}_{j=1}^{M}$, $M \in \mathbb{N}$, is an open covering of Ω if*

i) The $\Omega^{(j)}$ are open,

ii) They form a covering of Ω, i.e., $\Omega \subset \bigcup_{j=1}^{M} \Omega^{(j)}$,

iii) They are partially overlapping, i.e., for all $1 \leq j \leq M$, $\Omega^{(j)} \cap \left(\bigcup_{i \neq j} \Omega^{(j)}\right) \neq \emptyset$.

Based on the covering, we can define local weight functions. These weights are chosen to be a partition of unity in the following sense.

Definition 6.5 (Partition of unity)**.** *A set of continuous functions $w_j : \Omega \to \mathbb{R}$, $1 \leq j \leq M$,$M \in \mathbb{N}$, is a partition of unity w.r.t. the covering $\left\{\Omega^{(j)}\right\}_{j=1}^{M}$ if*

i) $w_j(x) \geq 0$ *for all* $x \in \Omega$,

ii) $\operatorname{supp}(w_j) \subset \Omega^{(j)}$ *(in particular, for all* $x \in \Omega$ *only a small number of* $w_j(x)$ *is nonzero),*

iii) $\sum_{j=1}^{M} w_j(x) = 1$ *for all* $x \in \Omega$.

With an open covering and a partition of unity we can properly define the PUM-interpolant.

Definition 6.6 (PUM-interpolant)**.** *Let* $\Omega \subset \mathbb{R}^d$*, and consider pairwise distinct data points* $X_N \subset \Omega$ *and data values* $\{f(x_i)\}_{i=1}^N \subset \mathbb{R}$*. Let* K *be a SPD kernel on* Ω*.*

For $M \in \mathbb{N}$*, consider an open covering* $\{\Omega^{(j)}\}_{j=1}^M$ *of* Ω *and a partition of unity* $\{w_j\}_{j=1}^M$ *w.r.t. the covering.*

For $1 \leq j \leq M$*, define* $X_j := X_N \cap \Omega^{(j)}$ *and denote as* $s_f^{(j)}$ *the local interpolant, i.e., the kernel interpolant with data points* X_j *and data values* $\{f(x_i), x_i \in X_j\}$*.*

Then the PUM- interpolant is defined as

$$s_f(x) := \sum_{j=1}^{M} w_j(x) s_f^{(j)}(x) \quad \text{for all } x \in \Omega. \tag{6.2}$$

We now can prove that this way of defining the interpolant works well, in the sense that the PUM-interpolant satisfies the interpolation conditions and it is a smooth function.

Proposition 6.7 (Properties of PUM-interpolant)**.** *The PUM-interpolant* s_f *is a global interpolant of* f *on the points* X_N*, i.e.,*

$$s_f(x_i) = f(x_i) \quad \text{for all} \quad x_i \in X_N.$$

Moreover, if $k.m \in \mathbb{N}$*,* $K \in C^k(\Omega \times \Omega)$ *and* $w_j \in C^m(\Omega)$ *for all* $1 \leq j \leq M$*, we have*

$$s_f \in C^{\min(k,m)}(\Omega).$$

Proof. We first compute the value of $s_f(x_i)$ for $x_i \in X_N$. Using Definition 6.6 we have

$$s_f(x_i) = \sum_{j=1}^{M} w_j(x_i) s_f^{(j)}(x_i).$$

Then using (ii) of Definition 6.5 we can restrict the sum over $j : x_i \in \Omega^{(j)}$, and for these j we have by Definition 6.6 that $s_f^{(j)}(x_i) = f(x_i)$. So we can conclude that

$$\begin{aligned} s_f(x_i) &= \sum_{j=1}^{M} w_j(x_i) s_f^{(j)}(x_i) = \sum_{j : x_i \in \Omega^{(j)}} w_j(x_i) s_f^{(j)}(x_i) = \sum_{j : x_i \in \Omega^{(j)}} w_j(x_i) f(x_i) \\ &= f(x_i) \sum_{j=1}^{M} w_j(x_i) = f(x_i), \end{aligned}$$

where we used in the last step the fact that $\sum_{j=1}^{M} w_j(x) = 1$ (by (iii) of Proposition 6.5).

Now, since $s_f^{(j)}$ is a finite linear combination of K, and $K \in C^k(\Omega \times \Omega)$, we have that $s_f^{(j)} \in C^k(\Omega)$.

Since we assumed that $w_j \in C^m(\Omega)$, it follows by the definition (6.2) of PUM-interpolant that $s_f \in C^{\min(k,m)}(\Omega)$. □

The most common way to obtain an open covering and a partition of unity is as follows:

- $\underline{\Omega^{(j)}}$: Consider points $\{c_j\}_{j=1}^{M} \subset \Omega$ on a regular grid and a radius $r > 0$. Define $\Omega^{(j)} := B(c_j, r)$ (open ball). If r is large enough, it holds $\Omega \subset \cup_{j=1}^{M} \Omega^{(j)}$ and the local domains are overlapping.
- $\underline{w_j}$: For some $k \in \mathbb{N}$, consider the Wendland kernel $W(x, y) := \Phi_{d,k}(\|x - y\|_2/r)$ (we use the notation W to distinguish it from the kernel used to compute the interpolant). Define for all $1 \leq j \leq M$

$$w_j(x) := \frac{W(x, c_j)}{\sum_{i=1}^{M} W(x, c_i)}.$$

These w_j satisfy all the requirements of Definition 6.5:

i) w_j is non negative and continuous, by definition of Wendland kernels.

ii) The support of w_j is the ball of center c_j and radius r, i.e., $\Omega^{(j)}$, by definition of Wendland kernels.

iii) For all $x \in \Omega$, by construction $\sum_{j=1}^{M} w_j(x) = \sum_{j=1}^{M} \frac{W(x,c_j)}{\sum_{i=1}^{M} W(x,c_i)} = 1$.

Moreover, this partition satisfies $w_j \in C^{2k}(\Omega)$, as we have seen in Theorem 5.18.

Remark 6.8. *Some comments on this method:*

- *The computation of the local interpolants can be performed in parallel, so the construction of the PUM interpolant is potentially very efficient.*
- *When the PUM interpolant is evaluated in a new point $x \in \Omega$, only a small number of terms in the sum* (6.2) *are non zero, so also the evaluation of the interpolant is potentially very efficient.*
- *To have a good efficiency it is fundamental to have a fast way to decide, for a given point $x \in \Omega$, what are the subdomains such that $x \in \Omega^{(j)}$. In this way the evaluation can really involve only the evaluation of the desired local interpolants. To do so, usually special data structures are used to organize the centers c_j of the covering.*
- *Nevertheless, the approach that we have seen requires to consider a number of local domains of radius r of the order $M = \mathcal{O}\left(\left(\frac{diam(\Omega)}{r}\right)^d\right)$, so the method is rarely used in high dimensions.*

- *Under further assumptions on $\left\{\Omega^{(j)}\right\}_{j=1}^{M}$ and $\{w_j\}_{j=1}^{M}$, it can be proven that the PUM interpolant has exactly the same convergence order of standard interpolation (i.e. Theorem 5.22 holds with different constants). These additional assumptions are satisfied by the particular constructions based on balls and Wendland kernels. The theorem is stated below, without proof. The assumptions are quite technical, but the idea is to guarantee that all the sets $\Omega^{(j)}$ are regular, in the sense that they satisfy a cone condition independently of j, and that the functions w_j have bounded derivatives, again independently of j.*

- *There is a Demo in ILIAS demonstrating the behavior of the method.*

Theorem 6.9 (Convergence order of PUM interpolation)**.** *Let $K(x,y) := \Phi(x-y)$ be a SPD, translational invariant kernel on $\mathbb{R}^d$.*

Assume that $\Phi \in C_\nu^{2k}(\mathbb{R}^d)$, i.e., $\Phi \in C^{2k}(\mathbb{R}^d)$ and, for $a \in \mathbb{N}_0^d$ with $|a| = 2k$, $D^a\Phi(x) = \mathcal{O}(\|x\|_2^\nu)$ for $\|x\|_2 \to 0$.

Let $\Omega \subset \mathbb{R}^d$ be open and bounded and let $X_N \subset \Omega$ be pairwise distinct. Let $\{\Omega^{(j)}\}_{j=1}^{M}$ be a open covering of Ω which is regular for (Ω, X_N), i.e., additionally to the requirements of Definition 6.4, assume that

- *For each $x \in \Omega$, the number of sets $\Omega^{(j)}$ such that $x \in \Omega^{(j)}$ is bounded by a constant C_Ω.*

- *There exist a constant $C_r > 0$ and an angle $\theta \in (0, \pi/2)$ such that for each $1 \le j \le M$, $\Omega^{(j)} \cap \Omega$ satisfies a cone condition with angle $\theta \in (0, \pi/2)$ and radius $r = C_r h_{X_N,\Omega}$.*

Moreover, let $\{w_j\}_{j=1}^{M}$ be a partition of unity which is k-stable for $\{\Omega^{(j)}\}_{j=1}^{M}$, i.e., additionally to the requirements of Definition 6.5, assume that

- *for every $a \in \mathbb{N}_0^d$ with $|a| \le k$ there exists a constant $C_\alpha > 0$ such that, for all $1 \le j \le M$, it holds*

$$\|D^a w_j\|_{L_\infty(\Omega^{(j)})} \le C_\alpha \operatorname{diam}\left(\Omega^{(j)}\right)^{-|a|}$$

Then there exists constants $C, h_0 > 0$, independent of $x, K, f \in \mathcal{H}_K(\Omega)$, such that for all $a \in \mathbb{N}_0^d$ with $|a| \le k$ it holds

$$|D^a f(x) - D^a s_f(x)| \le C\, h_{X_N}^{k+\nu/2-|a|}\, \|f\|_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega,$$

where s_f is the PUM-interpolant of f.

Proof. See Theorem 15.19 in [5]. □

6.4 Greedy kernel interpolation

Some ideas:

- We have seen that the full kernel matrix is in general ill-conditioned. This suggests that a good idea could be to select a small submatrix that approximates A in some sense, and solve only the reduced linear system. Selecting a submatrix of A means to select a subset of interpolation points.
- On the other hand, it is interesting to find a small subset of interpolation points X_n such that the interpolation error is small. This could be applied in theory to select good points from Ω, or in practice to select good points from a set X_N of given data.
- This selection is in general very expensive, also in the case the points X_n are selected from a finite set X_N, $n \leq N$. Indeed, finding the optimal choice of points is a combinatorial problem, which has a very large computational cost.
- The idea of greedy algorithms is to perform this selection incrementally, i.e., adding one point at a time. Instead of optimizing the full selection of X_n, at each iteration only the "best new point" is selected, based on some error indicator.
- The resulting interpolant, based only on the set X_n, has the great advantage to be fast to evaluate, because it is defined by a linear combination of only n (instead of N) terms. This is particularly useful if the interpolant needs to be evaluated multiple times, e.g. to obtain predictions of the values of the unknown function f.

The general structure of the algorithm is as follows. For the moment, we consider a generic selection rule $\mu : \Omega \to \mathbb{R}$, i.e., an error indicator that tells us what point to select. We will be more specific later, and define different selection rules. To make explicit the dependence of the interpolant on the current set X_n, we denote as s_n, instead of s_f, the interpolant based on the points X_n.

Definition 6.10 (Kernel greedy interpolation)**.** *Let $\Omega \subset \mathbb{R}^d$ and K be a SPD kernel on Ω. Let $f \in \mathcal{H}_K(\Omega)$.*

Define $X_0 := \emptyset$, $V(X_0) := \{0\}$, $s_0 := 0$, and, for $n \geq 1$,

- *Select $x_n := \underset{x \in \Omega \setminus X_{n-1}}{\arg\max}\ \eta(x)$*
- *Define $X_n := X_{n-1} \cup \{x_n\}$ and $V(X_n) :=$ span $\{K(\cdot, x_i), x_i \in X_n\}$*
- *Compute $s_n := \Pi_{V(X_n)}(f)$.*

The algorithm is terminated when $\eta(x_n) \leq \tau$, for τ a given tolerance.

Remark 6.11. *Two comments:*

- *It could be that the maximizer of η is not unique. In this case only one of the multiple points is selected and included in X_n.*

- *In practical implementations, the new point is selected from a large set of points X_N, not from Ω. This means that the first step is replaced by*

$$x_n := \operatorname*{arg\,max}_{x \in X_N \setminus X_{n-1}} \eta(x).$$

6.4.1 The Newton basis

Before seeing how to select a "good" new point, i.e., how to define η, we see how to perform the computation of the interpolant efficiently.

The problematic point is that the interpolants on X_n and X_{n-1} are defined as

$$s_{n-1}(\cdot) = \sum_{j=1}^{n-1} \alpha_j^{(n-1)} K(\cdot, x_j), \quad s_n(\cdot) = \sum_{j=1}^{n} \alpha_j^{(n)} K(\cdot, x_j),$$

where the coefficients are the solution of two different linear systems, involving the kernel matrix on X_{n-1} and on X_n. Thus, in general, $\alpha_j^{(n-1)} \neq \alpha_j^{(n)}$ for all $1 \leq j \leq n-1$. This means that all the coefficients need to be recomputed when adding a new point.

A solution to avoid the full recomputation is to use a different basis of $V(X_n)$, which is called Newton basis. The name comes from the analogy with the Newton basis of polynomial interpolation. In this case, if we have two sets $X_{n-1} \subset X_n \subset [a, b]$ and consider the two interpolants e.g. in monomial form, we have also in this case that all the coefficients are in general different, i.e.,

$$p_{n-1}(x) = \sum_{j=1}^{n-1} \alpha_j^{(n-1)} x^{j-1}, \quad p_n(x) = \sum_{j=1}^{n} \alpha_j^{(n)} x^{j-1}.$$

Instead, one can consider the polynomial Newton basis on X_n defined as

$$v_j(x) := \prod_{i=1}^{j-1} (x - x_i) \quad 1 \leq j \leq n$$

and the resulting interpolant can be written as

$$p_{n-1}(x) = \sum_{j=1}^{n-1} c_j v_j(x), \quad p_n(x) = \sum_{j=1}^{n} c_j v_j(x) = p_{n-1}(x) + c_n v_n(x),$$

where now the coefficients $c_1, \ldots, c_{n-1}$ does not change when adding a new point, and only the new basis element v_n and the coefficient c_n are computed. This is due to the fact that $v_n(x_i) = 0$ for $1 \leq i < n$, so the interpolation conditions on the points X_{n-1} are not changed by adding v_n in the linear combination.

There exists an analogous for kernel interpolation.

Definition 6.12 (Newton basis)**.** *Let $X_N \subset \Omega$ be pairwise distinct. The Newton basis $\{v_j\}_{j=1}^N$ of $V(X_N)$ is defined as the Gram-Schmidt orthonormalization of $\{K(\cdot, x_i)\}_{i=1}^N$, i.e.,*

$$
\begin{aligned}
v_1(x) &:= \frac{K(x, x_1)}{\|K(\cdot, x_1)\|_{\mathcal{H}_K(\Omega)}} = \frac{K(x, x_1)}{\sqrt{K(x_1, x_1)}} \\
\tilde{v}_n(x) &:= K(x, x_n) - \sum_{j=1}^{n-1} (K(\cdot, x_n), v_j)_{\mathcal{H}_K(\Omega)} \, v_j(x) = K(x, x_n) - \sum_{j=1}^{n-1} v_j(x_n) v_j(x) \\
v_n(x) &:= \frac{\tilde{v}_n(x)}{\|\tilde{v}_n\|_{\mathcal{H}_K(\Omega)}}, \quad 1 < n \leq N.
\end{aligned}
$$

Remark 6.13. *As is clear from the definition, the Newton basis is dependent of the ordering of the points in X_N (because the Gram-Schmidt procedure is). This will not be a problem in greedy algorithms, as the order of the points will be determined by the selection procedure.*

This basis is indeed what we are looking for.

Proposition 6.14 (Properties of the Newton basis)**.** *Let $X_N \subset \Omega$ and $\{v_j\}_{j=1}^N$ be the corresponding Newton basis. Then it holds:*

i) The Newton basis is an orthonormal basis of $V(X_N)$.

ii) It is nested, i.e., for all $1 < n \leq N$,

$$V(X_{n-1}) = \textit{ span } \{v_1, \ldots, v_{n-1}\}, \quad V(X_n) = \textit{ span } \{v_1, \ldots, v_n\}.$$

iii) For all $1 < n \leq N$ it holds $v_n \in V(X_{n-1})^\perp$ so $v_n(x_i) = 0$ for all $1 \leq i < n$.

iv) $v_n(x_n) = P_{X_{n-1}}(x_n)$ for all $1 < n \leq N$.

Proof. We have the following:

(i) The basis is constructed via a Gram-Schmidt procedure, so it is orthonormal, i.e., for all $1 \leq i, j \leq N$ it holds

$$(v_i, v_j)_{\mathcal{H}_K(\Omega)} = \delta_{ij}.$$

(ii) This also follows from the Gram-Schmidt construction. It can also be seen by induction since clearly

$$v_1 = \frac{K(x, x_1)}{\sqrt{K(x_1, x_1)}} \in V(X_1)$$

and, if $\{v_j\}_{j=1}^{n-1} \in V(X_{n-1})$, we have also by definition

$$\tilde{v}_n(x) := K(x, x_n) - \sum_{j=1}^{n-1} v_j(x_n) v_j(x) \in V(X_n),$$

so the basis is nested.

(iii) Since $(v_n, v_j)_{\mathcal{H}_K(\Omega)} = \delta_{nj}$ from (i), and span $\{v_1, \dots, v_{n-1}\} = V(X_{n-1})$ from (ii), we have $v_n \in V(X_{n-1})^\perp$. So $v_n(x_i) = 0$ for $1 \le i < n$ from Corollary 4.2.

(iv) This part requires the knowledge of the fact that

$$P_{X_{n-1}}(x)^2 = K(x,x) - \sum_{j=1}^{n-1} v_j(x)^2,$$

which holds for a general orthonormal basis (as the Newton basis is), and which will be proven in the next Proposition.

We start by computing $\tilde{v}_n(x_n)$ using the definition:

$$\tilde{v}_n(x_n) = K(x_n,x_n) - \sum_{j=1}^{n-1} v_j(x_n)v_j(x_n) = P_{X_{n-1}}(x_n)^2.$$

Then, again using the definition of $\tilde{v}_n$, we have

$$\begin{aligned}
\|\tilde{v}_n\|^2 &= (\tilde{v}_n, \tilde{v}_n)_{\mathcal{H}_K(\Omega)} \\
&= \left(K(\cdot, x_n) - \sum_{j=1}^{n-1} v_j(x_n)v_j, K(\cdot,x_n) - \sum_{i=1}^{n-1} v_i(x_n)v_i \right)_{\mathcal{H}_K(\Omega)} \\
&= K(x_n,x_n) - 2\sum_{j=1}^{n-1} v_j(x_n)(K(\cdot,x_n), v_j)_{\mathcal{H}_K(\Omega)} + \sum_{i,j=1}^{n-1} v_i(x_n)v_j(x_n)(v_i,v_j)_{\mathcal{H}_K(\Omega)} \\
&= K(x_n,x_n) - 2\sum_{j=1}^{n-1} v_j(x_n)^2 + \sum_{j=1}^{n-1} v_i(x_n)v_j(x_n)\delta_{ij} \\
&= P_{X_{n-1}}(x_n)^2.
\end{aligned}$$

So $v_n := \frac{\tilde{v}_n}{\|\tilde{v}_n\|_{\mathcal{H}_K(\Omega)}} = P_{X_{n-1}}(x_n)$.

□

6.4.2 Interpolation with the Newton basis

As we did for the interpolant s_n, to simplify the notation we use in the following the notation $P_n := P_{X_n}$ for the power function on X_n, assuming to have a sequence $\emptyset =: X_0 \subset \dots X_n \subset \dots X_N$.

The following result applies to one single set X_n and to a general orthonormal basis of $V(X_n)$.

Proposition 6.15 (Interpolation with orthonormal bases)**.** *Let $X_n \subset \Omega$ be pairwise distinct and consider the subspace $V(X_n) \subset \mathcal{H}_K(\Omega)$. Let $\{v_j\}_{j=1}^n$ be an orthonormal basis of*

$V(X_n)$, i.e., $(v_i, v_j)_{\mathcal{H}_K(\Omega)} = \delta_{ij}$. Let $f \in \mathcal{H}_K(\Omega)$. Then, the interpolant s_n on X_n with data $\{f(x_i)\}_{i=1}^n$ can be computed as

$$s_n(x) = \sum_{j=1}^{n} (f, v_j)_{\mathcal{H}_K(\Omega)} v_j(x), \tag{6.3}$$

and the power function of X_n as

$$P_n(x)^2 = K(x,x) - \sum_{j=1}^{n} v_j(x)^2.$$

Proof. From Proposition 4.1 we know that the interpolant with points X_n is the orthogonal projection into $V(X_n)$. Since $\{v_j\}_{j=1}^N$ is an o.n.b. of $V(X_n)$ (Proposition 6.14), it follows that

$$s_n(x) = \sum_{j=1}^{n} (f, v_j)_{\mathcal{H}_K(\Omega)} v_j(x).$$

Moreover, defining $f_x := K(\cdot, x)$, we know from (i) of Proposition 4.12 that

$$P_n(x) := P_{X_n}(x) = \|f_x - s_{f_x}(\cdot)\|_{\mathcal{H}_K(\Omega)}.$$

Using the form (6.3) and the reproducing property we can compute

$$s_{f_x}(\cdot) = \sum_{j=1}^{n} (f_x, v_j)_{\mathcal{H}_K(\Omega)} v_j(\cdot) = \sum_{j=1}^{n} (K(\cdot, x), v_j)_{\mathcal{H}_K(\Omega)} v_j(\cdot) = \sum_{j=1}^{n} v_j(x) v_j(\cdot).$$

It follows that

$$\begin{aligned}
P_n(x)^2 = \|f_x - s_{f_x}(\cdot)\|^2_{\mathcal{H}_K(\Omega)} &= \left\| K(\cdot, x) - \sum_{j=1}^{n} v_j(x) v_j(\cdot) \right\|^2_{\mathcal{H}_K(\Omega)} \\
&= \left(K(\cdot, x) - \sum_{j=1}^{n} v_j(x) v_j(\cdot), K(\cdot, x) - \sum_{i=1}^{n} v_i(x) v_i(\cdot) \right)^2_{\mathcal{H}_K(\Omega)} \\
&= K(x,x) - 2 \sum_{j=1}^{n} v_j(x) (v_j, K(\cdot, x))_{\mathcal{H}_K(\Omega)} + \sum_{i,j=1}^{n} v_i(x) v_j(x) (v_i, v_j)_{\mathcal{H}_K(\Omega)} \\
&= K(x,x) - \sum_{j=1}^{n} v_j(x)^2.
\end{aligned}$$

□

Now we can finally see that, using the Newton basis, we can add a new point and update the interpolant without recomputing all the coefficients, but only adding a new basis element and computing a new coefficient, exactly as in the polynomial case.

Proposition 6.16 (Iterative interpolation with the Newton basis)**.** *Let $f \in \mathcal{H}_K(\Omega)$, $\emptyset =: X_0 \subset \dots X_n \subset \dots X_N \subset \Omega$. For all $n \geq 1$, define the residual as*

$$r_0 := f, \quad r_n := f - s_n.$$

Then

i) We have the update formulas

$$\begin{aligned} s_n(x) &= \sum_{j=1}^{n} (f, v_j)_{\mathcal{H}_K(\Omega)} v_j(x) = s_{n-1}(x) + (f, v_n)_{\mathcal{H}_K(\Omega)} v_n(x) \\ P_n(x)^2 &= K(x,x) - \sum_{j=1}^{n} v_j(x)^2 = P_{n-1}(x)^2 - v_n(x)^2 \\ r_n(x) &= r_{n-1}(x) - (f, v_n)_{\mathcal{H}_K(\Omega)} v_n(x). \end{aligned}$$

ii) For the coefficient of the basis v_n, it holds $(f, v_n)_{\mathcal{H}_K(\Omega)} := \frac{r_{n-1}(x_n)}{P_{n-1}(x_n)}$.

In particular, all the updates can be obtained efficiently by reusing previously computed quantities.

Proof. (i) We show the result for s_n, and the others are completely analogous. From Proposition 6.14, we have that the Newton basis $\{v_j\}_{j=1}^{n-1}$ is an orthonormal basis of $V(X_{n-1})$ and $\{v_j\}_{j=1}^{n}$ is an orthonormal basis of $V(X_n)$. Then we can use Proposition 6.15 to obtain

$$\begin{aligned} s_{n-1}(x) &= \sum_{j=1}^{n-1} (f, v_j)_{\mathcal{H}_K(\Omega)} v_j(x) \\ s_n(x) &= \sum_{j=1}^{n} (f, v_j)_{\mathcal{H}_K(\Omega)} v_j(x), \end{aligned}$$

thus

$$s_n(x) = s_{n-1}(x) + (f, v_n)_{\mathcal{H}_K(\Omega)} v_n(x).$$

(ii) From definition 6.12 we have

$$(f, v_n) = \left(f, \frac{\tilde{v}_n}{\|\tilde{v}_n\|_{\mathcal{H}_K(\Omega)}} \right)_{\mathcal{H}_K(\Omega)} = \frac{1}{\|\tilde{v}_n\|_{\mathcal{H}_K(\Omega)}} (f, \tilde{v}_n)_{\mathcal{H}_K(\Omega)},$$

where

$$\tilde{v}_n(x) = K(x, x_n) - \sum_{j=1}^{n-1} v_j(x_n) v_j(x).$$

We proved in the proof of Proposition 6.14 that $\|\tilde{v}_n\|_{\mathcal{H}_K(\Omega)} = P_{n-1}(x_n)$. So we are done if we prove that $(f, \tilde{v}_n)_{\mathcal{H}_K(\Omega)} = r_{n-1}(x_n)$. Using the definition of $\tilde{v}_n$, r_{n-1}, the reproducing property and the formula for s_n, we have

$$\begin{aligned}(f, \tilde{v}_n)_{\mathcal{H}_K(\Omega)} &= \left(f, K(\cdot, x_n) - \sum_{j=1}^{n-1} v_j(x_n) v_j\right)_{\mathcal{H}_K(\Omega)} \\ &= f(x_n) - \sum_{j=1}^{n-1} v_j(x_n)(f, v_j)_{\mathcal{H}_K(\Omega)} \\ &= f(x_n) - s_{n-1}(x_n) \\ &= r_{n-1}(x_n).\end{aligned}$$

□

6.4.3 Selection rules and error

We can now formulate more precisely the selection rules/error indicators η used in the algorithm. Each one leads to a different approximant (since the interpolation points are different) and so the error analysis will be different.

Definition 6.17 (Selection rules)**.** *We have the following selection rules*

- *P-greedy:* $x_n := \underset{x_n \in \Omega \setminus X_{n-1}}{\arg\max}\ P_{n-1}(x)$
- *f-greedy:* $x_n := \underset{x_n \in \Omega \setminus X_{n-1}}{\arg\max}\ |r_{n-1}(x)|$
- *f/P-greedy:* $x_n := \underset{x_n \in \Omega \setminus X_{n-1}}{\arg\max}\ \frac{|r_{n-1}(x)|}{P_{n-1}(x)}$

Remark 6.18. *Some comments*

- *Recall that in practical implementation, the selection is only over a large but finite set X_N, not all Ω.*
- *They are all well defined rules, and we have seen how to efficiently compute the quantities to be maximized (i.e., P_n, r_n). Moreover, $P_{n-1}(x) \neq 0$ if $x \notin X_{n-1}$ (Proposition 4.13). This motivates the restriction of the selection on $\Omega \setminus X_{n-1}$.*
- *When adding a new point x_n, the first two error indicators at the next iteration satisfy $\eta(x_n) = 0$. Indeed, $P_{X_n}(x_n) = 0$ since $x_n \in X_n$ (Proposition 4.13) and $r_n(x_n) = f(x_n) - s_n(x_n) = 0$ since s_n is the interpolant of f on X_n. The third one is "$0/0$", but we excluded in the definition to select x_n from X_{n-1}.*

- *The motivation for the first two selections is clear: they try to put to* 0 *an upper bound on the error (*P*-greedy) or the maximal error itself (*f*-greedy). The third one is unclear at the moment, but we see in the next Proposition that it is indeed locally optimal, i.e., it guarantees the maximal reduction of the error.*

Proposition 6.19 (Convergence of greedy interpolation)**.** *For* $n \in \mathbb{N}$*, let* $X_n \subset \Omega$ *be the sequence of points selected by the greedy algorithm with any of the selection rules of Definition 6.17. Let* $f \in \mathcal{H}_K(\Omega)$ *and* $r_n := f - s_n$ *(with* $r_0 := f$*). Then it holds*

$$\|f - s_n\|^2_{\mathcal{H}_K(\Omega)} = \|f - s_{n-1}\|^2_{\mathcal{H}_K(\Omega)} - \left(\frac{r_{n-1}(x_n)}{P_{n-1}(x_n)}\right)^2.$$

In particular, for all $n \in \mathbb{N}$ *it holds*

i) $\|f - s_n\|_{\mathcal{H}_K(\Omega)} \leq \|f - s_{n-1}\|_{\mathcal{H}_K(\Omega)}$ *(the error is non increasing w.r.t.* $\|\cdot\|_{\mathcal{H}_K(\Omega)}$*)*

ii) $\|f - s_n\|_{\mathcal{H}_K(\Omega)} < \|f - s_{n-1}\|_{\mathcal{H}_K(\Omega)}$ *if* $r_{n-1}(x_n) \neq 0$ *(the error is decreasing w.r.t* $\|\cdot\|_{\mathcal{H}_K(\Omega)}$*, if the algorithm selects a point such that the pointwise error is not already zero)*

Moreover

iii) f/P*-greedy is locally optimal w.r.t* $\|\cdot\|_{\mathcal{H}_K(\Omega)}$*, i.e., for all* $n \in \mathbb{N}$ *the interpolant* s_n *obtained with* x_n *selected by* f/P*-greedy satisfies*

$$\|f - s_n\|^2_{\mathcal{H}_K(\Omega)} \leq \|f - \Pi_{X_{n-1}\cup\{x\}}(f)\|^2_{\mathcal{H}_K(\Omega)} \quad \text{for all } x \in \Omega.$$

Proof. We have from Proposition 6.16 that, for any selection rule,

$$r_n(x) = r_{n-1}(x) - (f, v_n)_{\mathcal{H}_K(\Omega)} v_n(x).$$

Using the orthonormality properties of the Newton basis we can compute

$$\begin{aligned}
\|r_n\|^2_{\mathcal{H}_K(\Omega)} &= \big(r_{n-1} - (f, v_n)_{\mathcal{H}_K(\Omega)} v_n, r_{n-1} - (f, v_n)_{\mathcal{H}_K(\Omega)} v_n\big)_{\mathcal{H}_K(\Omega)} \\
&= \|r_{n-1}\|^2_{\mathcal{H}_K(\Omega)} + (f, v_n)^2_{\mathcal{H}_K(\Omega)} \|v_n\|^2_{\mathcal{H}_K(\Omega)} - 2(f, v_n)_{\mathcal{H}_K(\Omega)} (r_{n-1}, v_n)_{\mathcal{H}_K(\Omega)} \\
&= \|r_{n-1}\|^2_{\mathcal{H}_K(\Omega)} + (f, v_n)^2_{\mathcal{H}_K(\Omega)} - 2(f, v_n)_{\mathcal{H}_K(\Omega)} (r_{n-1}, v_n)_{\mathcal{H}_K(\Omega)}.
\end{aligned}$$

Moreover, $r_{n-1} = f - s_{n-1}$, so $r_{n-1}(x_j) = 0$ for $1 \leq j \leq n-1$. In particular, $r_{n-1} \in V(X_{n-1})$ from Corollary 4.2, so $(r_{n-1}, v_j)_{\mathcal{H}_K(\Omega)} = 0$ for $1 \leq j \leq n-1$, since $v_j \in V(X_{n-1})$ from Proposition 6.14. So using the definition of $\tilde{v}_n$ we have

$$(r_{n-1}, \tilde{v}_n)_{\mathcal{H}_K(\Omega)} = (r_{n-1}, K(\cdot, x_n))_{\mathcal{H}_K(\Omega)} - \sum_{j=1}^{n-1} v_j(x_n)(r_{n-1}, v_j)_{\mathcal{H}_K(\Omega)} = r_{n-1}(x_n),$$

thus

$$(r_{n-1}, v_n)_{\mathcal{H}_K(\Omega)} = \left(r_{n-1}, \frac{\tilde{v}_n}{\|\tilde{v}_n\|_{\mathcal{H}_K(\Omega)}}\right)_{\mathcal{H}_K(\Omega)} = \frac{r_{n-1}(x_n)}{P_{n-1}(x_n)}.$$

Since we proved in (ii) of Proposition 6.16 that also $(f, v_n)_{\mathcal{H}_K(\Omega)} = \frac{r_{n-1}(x_n)}{P_{n-1}(x_n)}$, we can conclude that

$$\begin{aligned}\|r_n\|^2_{\mathcal{H}_K(\Omega)} &= \|r_{n-1}\|^2_{\mathcal{H}_K(\Omega)} + (f, v_n)^2_{\mathcal{H}_K(\Omega)} - 2(f, v_n)_{\mathcal{H}_K(\Omega)}(r_{n-1}, v_n)_{\mathcal{H}_K(\Omega)} \\ &= \|r_{n-1}\|^2_{\mathcal{H}_K(\Omega)} - \left(\frac{r_{n-1}(x_n)}{P_{n-1}(x_n)}\right)^2.\end{aligned}$$

From this formula, points (i) and (ii) follows directly, since all the terms are positive.

Also (iii) follows from this formula, since it holds for all possible new point x_n, and f/P-greedy selects exactly the point that maximizes the term $\frac{r_{n-1}(x_n)}{P_{n-1}(x_n)}$ (Definition 6.17). □

Remark 6.20. *Some remarks on the result:*

- *Property* (ii) *guarantees that the error for f- and f/P-greedy is strictly decreasing until exact convergence. Indeed, from Definition 6.17 they always select a point x_n where $r_{n-1}(x_n) \neq 0$, unless there is no point $x \in \Omega \setminus X_{n-1}$ such that $r_{n-1}(x) \neq 0$. But this means that $r_{n-1} = 0$ in Ω, i.e., $s_{n-1} = f$, so the interpolant is exact.*

- *The error is decreasing w.r.t. the norm of $\mathcal{H}_K(\Omega)$. This implies that also the maximum error is decreasing, but it does not need to be monotonically decreasing (i.e., it can have some oscillations).*

- *We see now also a result on convergence rate of the algorithm for P- and f/P-greedy. The two results are quite different, because the research topic is still quite ongoing. In particular, the convergence rate for P-greedy is much faster than the one for f/P-greedy, although experimentally is clear that the opposite holds.*

- *Both convergence rates are based on the number n of points, and not on the fill distance as in the other convergence results that we have seen. This is because, in general, it can not be expected that the points selected by the algorithm have a small fill distance, especially for f- and f/P-greedy, which select points which are good for one single target function $f \in \mathcal{H}_K(\Omega)$.*

Theorem 6.21 (Convergence rates)**.** *Assume the greedy selection is done over Ω. In the following c, c_1, c_2 are constants independent of $n \in \mathbb{N}$.*

For P-greedy it holds the following:

- *For SPD kernels which generate Sobolev spaces $H^k(\Omega)$ for a given k (in the sense of Corollary 5.13), it holds for all $n \in \mathbb{N}$ and $f \in \mathcal{H}_K(\Omega)$*

$$\|f - f_n\|_{L_\infty(\Omega)} \leq cn^{-k/d+1/2}\|f\|_{\mathcal{H}_K(\Omega)}.$$

- *For the Gaussian and Inverse Multiquadric kernels it holds for all $n \in \mathbb{N}$ and $f \in \mathcal{H}_K(\Omega)$*

$$\|f - f_n\|_{L_\infty(\Omega)} \leq c_1 e^{-c_2 n^{\frac{1}{d}}}\|f\|_{\mathcal{H}_K(\Omega)}.$$

For f/P-greedy, we define for $M > 0$ the set

$$\mathcal{H}_K(\Omega)^M := \{f \in \mathcal{H}_K(\Omega), f = \sum_j \alpha_j K(\cdot, x_j) : \sum_j |\alpha_j| \leq M\}.$$

Then for $f \in \mathcal{H}_K(\Omega)^M$ it holds for all $n \in \mathbb{N}$

$$\|f - f_n\|_{\mathcal{H}_K(\Omega)} \leq M \left(1 + \frac{n}{\sup\limits_{x\in\Omega} K(x,x)}\right)^{-1/2}.$$

6.4.4 Implementation

To conclude this part, we see how to practically implement the algorithm. We assume to run it over a large, given set X_N and to know the values $\{f(x_i)\}_{i=1}^N$. So far we have seen the following:

- The interpolant $s_n(x)$ requires (f, v_n) and $v_n(x)$.
- (f, v_n) requires $r_{n-1}(x_n)$ and $P_{n-1}(x_n)$.
- $P_{n-1}(x)$ requires $K(x,x)$ and $v_j(x)$.
- The residual $r_n(x_j)$ requires $f(x_j)$ and $s_n(x_j)$.
- The selection rules require $P_n(x)$, $r_n(x)$.

Thus, all we miss to have a complete computation is a way to compute $v_j(x)$ for $x \in \Omega$.

Proposition 6.22 (Computation of the Newton basis)**.** *Let $X_N \subset \Omega$ and let $\{v_j\}_{j=1}^N$ be the Newton basis. Then there exists an upper triangular and invertible matrix $B \in \mathbb{R}^{N\times N}$ such that for $1 \leq j \leq N$ it holds*

$$v_j(x) = \sum_{i=1}^N B_{ij} K(x, x_i) \quad \textit{for all } x \in \Omega. \tag{6.4}$$

Moreover, $(B^{-T})_{ij} = v_j(x_i)$ and B^{-T} is the Cholesky factor of A, i.e.,

$$A = B^{-T} B^{-1}.$$

Proof. Since both $\{K(\cdot, x_j)\}_{j=1}^n$ and $\{v_j\}_{j=1}^n$ are bases of $V(X_n)$ for all $1 \leq n \leq N$, there exists an invertible matrix of change of basis such that (6.4) holds.

Morever, B is upper triangular because the n-th Newton basis is an element of $V(X_n)$, so it depends only on $K(\cdot, x_i)$, $1 \leq i \leq n$, i.e.,

$$v_j(x) = \sum_{i=1}^N B_{ij} K(x, x_i) = \sum_{i=1}^j B_{ij} K(x, x_i),$$

so $B_{ij} = 0$ for $i > j$.

Moreover, by orthonormality of the Newton basis it holds

$$\begin{aligned}\delta_{ij} = (v_i, v_j)_{\mathcal{H}_K(\Omega)} &= \left(\sum_{l=1}^{N} B_{li} K(\cdot, x_l), \sum_{m=1}^{N} B_{mj} K(\cdot, x_m) \right)_{\mathcal{H}_K(\Omega)} \\ &= \sum_{l,m=1}^{N} B_{li} B_{mj} \left(K(\cdot, x_l), K(\cdot, x_m) \right)_{\mathcal{H}_K(\Omega)} \\ &= \sum_{l,m=1}^{N} B_{li} B_{mj} K(x_l, x_m) = (B^T A B)_{ij},\end{aligned}$$

i.e., $B^T A B = I$. Since B is invertible, this implies $A = B^{-T} B^{-1}$.

We can then evaluate $v_j(x_l)$ using (6.4) for $1 \leq j, l \leq N$ and obtain

$$v_j(x_l) = \sum_{i=1}^{N} B_{ij} K(x_l, x_i) = (AB)_{lj} = (B^{-T})_{ij}.$$

If we define $L := B^{-T}$, we have that L is lower triangular (B is upper triangular), $A = B^{-T} B^{-1} = L L^T$, and, from (iv) of Proposition 6.14, we have

$$L_{ii} = (B^{-T})_{ii} = v_i(x_i) = P_{i-1}(x_i) > 0$$

since the power function is positive. This means that this is the unique Cholesky factorization of A (Proposition ii). □

Remark 6.23. *Some final comments on this method:*

- *If n is too large, greedy algorithms behave more or less like the interpolant computed on the full set X_N. In particular, they are not a solution for ill conditioning if n is too large.*
- *They are usually worse than PUM if the target function f has local features and $d = 1, 2, 3$.*
- *On the other hand, they work very well in high space dimension.*
- *About the selection rules: f/P-greedy gives usually the faster convergence, but becomes unstable for relatively small n. P-greedy is practically slower, as it ignores f to select the points, but it is very stable. Usually f-greedy is the best option, as it is sort of intermediate between the other two.*
- *There is a demo in ILIAS, with an implementation of the different selection rules. The practical implementation is just a pivoted Cholesky factorization of the kernel matrix A of the full set X_N, where the pivoting rule is determined by the particular greedy selection rule.*

7. Solution of Partial Differential Equations

We start a new part, with the goal of approximating the solution of PDEs instead of approximating functions from pointwise values.

The general goal will be, for $\Omega \subset \mathbb{R}^d$ an open and bounded set, to find u such that

$$\begin{aligned} Lu(x) &= f(x),\ x \in \Omega \\ Bu(x) &= g(x),\ x \in \partial\Omega, \end{aligned} \tag{7.1}$$

where L is a linear differential operator and B a linear boundary-value operator. The typical example is

$$\begin{aligned} \Delta u(x) &= f(x),\ x \in \Omega \\ u(x) &= g(x),\ x \in \partial\Omega. \end{aligned}$$

We will see the following methods:

- Approximation by collocation: it is a generalization of the ideas seen so far, in the same setting. It comes from looking for an approximate solution which satisfies the PDE in some given points. Depending on the ansatz, we will have
  - symmetric collocation. This is the theoretically motivated one, for which we will get the full theory.
  - non-symmetric collocation. It can be obtained with minor modifications of the symmetric method, so it is worth having a look. It is also much used and have some computational advantages, but it is weaker in the theoretical motivations.

  They both can be analyzed in the framework of generalized interpolation.
- For interpolation and pattern analysis, kernel methods are (among) the state of the art methods. For this kind of applications, instead, there are usually better alternatives, such as the Finite Element Method. Nevertheless, kernel-based approximation of PDE is promising in high dimension or for complex geometries, included problems on manifolds. In particular, they do not require the use of a mesh (so they are usually called meshless methods).
- We will see also a third method, called RBF-Finite differences (RBF-FD). It has a clear formulation and clear advantages, but it is not well studied theoretically, yet. Nevertheless, it works very well, and also outperforms standard methods in some cases. So it is worth studying.

7.1 Generalized interpolation

We start by studying generalized interpolation problems, which are not necessarily related to the solution of PDEs. But still we change the notation with respect to the previous chapters and denote as $u \in \mathcal{H}_K(\Omega)$ the unknown target function, to distinguish it from the function f which plays the role of the right hand side in (7.1).

So far, we have considered interpolation problems, i.e., we know u through pointwise values $\{u(x_i)\}_{i=1}^N$ on $X_N \subset \Omega$ and we require the interpolant to satisfy $s_u(x_i) = u(x_i)$, $1 \leq i \leq N$. Another way to formulate the problem is to say that we have the set of functionals

$$\Lambda_N := \{\delta_{x_1}, \dots, \delta_{x_N}\}$$

and that we require $\delta_{x_i}(s_u) = \delta_{x_i}(u)$, $1 \leq i \leq N$. These functionals are linear and continuous ((i) of Proposition 3.8), and also linearly independent ((iii) of Proposition 3.8).

We can generalize the idea and assume to have general linear and continuous, linearly independent data functionals $\Lambda_N := \{\lambda_1, \dots, \lambda_N\} \subset \mathcal{H}_K(\Omega)'$, and assume that we know the unknown function u via the evaluations $\{\lambda_i(f)\}_{i=1}^N$. For example

- If $\lambda_i := \delta_{x_i}$ we have the usual pointwise interpolation data.
- If $\lambda_i := \delta_{x_i} \circ D^a$ for a multiindex $a \in \mathbb{N}_0^d$, we have additionally information on the derivative in the point x_i. This will be the case in the solution of PDEs.
- When data are of these two types, the problem is called an Hermite-Birkhoff problem.
- Other cases are possible, for example $\lambda_i(u) := \int_B u(x)dx$ with $B \subset \Omega$ or $\lambda_i := \frac{1}{M}\sum_{j=1}^M \delta_{y_j}$ for $Y_M \subset \Omega$.

Given Λ_N, a generalized interpolant s_u of u is defined as follows. We formulate the problem in $\mathcal{H}_K(\Omega)$, but the definition is valid in any other Hilbert space.

Definition 7.1 (Generalized interpolation)**.** *Let $\Omega \subset \mathbb{R}^d$ and K a SPD kernel on Ω.*

Let $u \in \mathcal{H}_K(\Omega)$ and $\Lambda_N := \{\lambda_1, \dots, \lambda_N\} \subset \mathcal{H}_K(\Omega)'$, and assume that the set of data $\{\lambda_i(u)\}_{i=1}^N$ is known.

A function $s_u \in \mathcal{H}_K(\Omega)$ is a generalized interpolant of u with respect to the data functionals Λ_N if it satisfies the generalized interpolation conditions

$$\lambda_i(s_u) = \lambda_i(u) \;\; 1 \leq i \leq N. \tag{7.2}$$

Remark 7.2. *This definition is really a generalization of the usual interpolation, because when the functional are $\lambda_i := \delta_{x_i}$ for $x_i \in X_N$, the conditions* (7.2) *are*

$$s_u(x_i) = \lambda_i(s_u) = \lambda_i(u) = u(x_i) \;\; 1 \leq i \leq N,$$

which are the usual interpolation conditions.

As in the case of standard interpolation, we need to choose a proper finite ansatz for s_u, i.e., select a linear and N-dimensional subspace $V_N := \text{span}\,\{v_1, \ldots, v_N\} \subset \mathcal{H}_K(\Omega)$ and require that

$$s_u(x) := \sum_{j=1}^{N} \alpha_j v_j(x) \;\; x \in \Omega.$$

In this way, using the linearity of the data functionals Λ_N, we are able to rewrite the generalized interpolation conditions (7.2) as

$$\lambda_i(s_u) = \sum_{j=1}^{N} \alpha_j \lambda_i(v_j) = \lambda_i(u) \;\; 1 \leq i \leq N.$$

These conditions can be written, like in the usual interpolation case, as a linear system

$$A_\Lambda \alpha = b,$$

where now $A_\Lambda \in \mathbb{R}^{N\times N}$, $(A_\Lambda)_{ij} := \lambda_i(v_j)$, and $b \in \mathbb{R}^N$, $b_i := \lambda_i(u)$.

The choice of the ansatz will make the difference between the symmetric and non-symmetric approach when solving PDEs, and we will be able to prove that A_Λ is invertible in the symmetric case, so a unique solution exists.

7.1.1 Optimal recovery

If a particular ansatz is used, the generalized interpolation problem can be solved. We first see how to compute this solution, and then we will prove that it is the optimal one in some sense. The result can be formulated in a general Hilbert space H of functions on $\Omega \subset \mathbb{R}^d$.

Proposition 7.3 (Computation of a generalize interpolant)**.** *Let H be an Hilbert space of functions on $\Omega \subset \mathbb{R}^d$ and let $\Lambda_N := \{\lambda_1, \ldots, \lambda_N\} \subset H'$ be a set of linear and continuous, linearly independent functionals. Let $v_j \in H$ be the Riesz representer of λ_j, $1 \leq j \leq N$. Assume that the data $\lambda_j(u)$ are know for an unknown function $u \in H$. Consider the ansatz*

$$s_u(x) := \sum_{j=1}^{N} \alpha_j v_j(x) \;\; x \in \Omega.$$

Then there exists a unique vector of coefficients $\alpha \in \mathbb{R}^N$ such that

$$\lambda_i(s_u) = \lambda_i(u) \;\; 1 \leq i \leq N.$$

It is the solution of the linear system

$$A_\Lambda \alpha = b, \tag{7.3}$$

where $b \in \mathbb{R}^N$, $b_i := \lambda_i(u)$, and $A_\Lambda \in \mathbb{R}^{N\times N}$, $(A_\Lambda)_{ij} := (v_i, v_j)_H$, is positive definite.

Proof. First, the Riesz representers exist since Λ_N are linear and continuous, so s_u is well defined.

If we impose the interpolation conditions, by linearity and by definition of Riesz representers we obtain that

$$\lambda_i(s_u) = \lambda_i\left(\sum_{j=1}^{N} \alpha_j v_j\right) = \sum_{j=1}^{N} \alpha_j \lambda_i\left(v_j\right) = \sum_{j=1}^{N} \alpha_j \left(v_i, v_j\right)_H,$$

so $\alpha \in \mathbb{R}^N$ needs to satisfy $A_\Lambda \alpha = b$, with $(A_\Lambda)_{ij} := (v_i, v_j)_H$.

In particular, A_Λ is symmetric. Moreover, we have seen that also the Riesz representers are linearly independent if Λ_N are linearly independent (proof of (iii) of Proposition 3.8). This means that the matrix A_Λ is also positive definite, since for all $\alpha \in \mathbb{R}^N \setminus \{0\}$ we have

$$\alpha^T A_\Lambda \alpha = \sum_{i,j=1}^{N} \alpha_i \alpha_j (v_i, v_j)_H = \left(\sum_{i}^{N} \alpha_i v_i, \sum_{j=1}^{N} \alpha_j v_j\right)_H = \left\|\sum_{i=1}^{N} \alpha_i v_i\right\|_H^2 > 0.$$

So there exists a unique solution s_u using this ansatz. □

As in the case of standard interpolation, there exist in general other ways to obtain a function that interpolates the data in this generalized sense. As we did in Section 4.1 for standard interpolation, we prove now that the solution of Proposition 7.3 is optimal in the following sense.

Definition 7.4 (Optimal recovery)**.** *Let H be an Hilbert space, $\Lambda_N := \{\lambda_1, \ldots, \lambda_N\} \subset H'$ be a set of linear and continuous, linearly independent functionals. Assume that, for a target function $u \in H$, we know the data $\lambda_j(u)$. Then the optimal recovery problem is to find a function $s_u \in H$ such that*

$$s_u := \arg\min \{\|s\|_H : s \in H, \lambda_i(s) = \lambda_i(u), 1 \le i \le N\}$$

Proposition 7.5 (Solution of optimal recovery problem)**.** *Let H be an Hilbert space, $\Lambda_N := \{\lambda_1, \ldots, \lambda_N\} \subset H'$ be a set of linear and continuous, linearly independent functionals. Let $v_j \in H$ be the Riesz representer of λ_j, $1 \le j \le N$. Assume that the data $\lambda_j(u)$ are know for an unknown function $u \in H$.*

Then the generalized interpolant $s_u \in H$ of Proposition 7.3 satisfies the following:

i) It is the orthogonal projection into $V_N :=$ span $\{v_1, \ldots, v_N\}$.

ii) It is the unique solution of the optimal recovery problem.

iii) It is the best approximation of u from V_N, i.e.,

$$\|u - s_u\|_H = \min\{\|u - s\|_H : s \in V_N\}.$$

Proof. We have from Proposition 7.3 that s_u satisfies $\lambda_i(u) = \lambda_i(s_u)$ for $1 \le i \le N$.

(i) Since v_i are linearly independent, they are a basis of V_N. So it suffices to use the fact that

$$(u - s_u, v_i)_H = (u, v_i)_H - (s_u, v_i)_H = \lambda_i(u) - \lambda_i(s_u) = 0,$$

to conclude that s_u is the orthogonal projection of u in V_N.

(iii) Since s_u is the orthogonal projection in V_N, it is also the best approximation from V_N.

(ii) As we did for interpolation, we can consider another solution $s \in H$ of the generalized interpolation problem, i.e., a function which satisfies $\lambda_i(s) = \lambda_i(u)$, and prove that $\|s_u\|_H \leq \|s\|_H$. This follows with the same argument as in the proof of Proposition 4.4, i.e., by decomposing s as $s = g + g^\perp$ with $g \in V_N$ and $g^\perp \in V_N^\perp$.

□

Remark 7.6. *Some comments on the Proposition:*

- *This is exactly what we have done for standard interpolation, just seen in another way. Indeed, if the functionals are $\lambda_i = \delta_{x_i}$ for some points, generalized interpolation is exactly standard interpolation.*
- *The reason for considering the native space of a SPD kernel K instead of a general Hilbert space, in the case of interpolation, is that we need the functionals to be linear and continuous (i.e., if and only if K is PD) and linearly independent (i.e., K is SPD). Both facts have been proven in Proposition 3.8.*
- *In this case, the last Proposition is exactly what we have seen in the case of standard interpolation: in this case the Riesz representer of λ_j is $K(\cdot, x_j)$, so the ansatz is*

$$s_u(x) = \sum_{j=1}^{N} \alpha_j K(x, x_j).$$

The interpolation conditions are $s_u(x_i) = \lambda_i(s_u) = \lambda_i(u) = u(x_i) \;\; 1 \leq i \leq N$. The matrix A_Λ is defined as

$$(A_\Lambda)_{ij} = (v_i, v_j)_{\mathcal{H}_K(\Omega)} = (K(\cdot, x_i), K(\cdot, x_j))_{\mathcal{H}_K(\Omega)} = K(x_i, x_j),$$

so it is the usual kernel matrix. The results on optimality are the same that we have seen in the case of standard interpolation in Section 4.1.

Example 7.7 (Generalized interpolation)**.** *We see an example of application of Proposition 7.3 using the native space $\mathcal{H}_K(\Omega)$ of a SPD kernel K. Assume that we want to approximate a function $u : \Omega \to \mathbb{R}$ with $\Omega := [a, b] \subset \mathbb{R}$, and that we are given, as usual, a set of points $X_{N-1} \subset \Omega$ and the values $\{u(x_i)\}_{i=1}^{N-1}$.*

This means that we can define an interpolant s_u by requiring that

$$\lambda_i(s_u) = \lambda_i(u) \;\; 1 \leq i \leq N-1,$$

where $\lambda_i := \delta_{x_i}$. This is just what we have done for interpolation.

Moreover, assume that we know that the target function f has a given mean $c := \int_\Omega f(x)dx$, and we want to include this condition to improve the interpolant.

A possible way to do so is to consider a discretization $Y_M \subset \Omega$, $M \in \mathbb{N}$, with $X_N \cap Y_M = \emptyset$, and approximate the integral as

$$c := \int_\Omega f(x)dx \approx \frac{1}{M}\sum_{m=1}^{M} f(y_m).$$

This gives us another linear functional $\lambda_N := \frac{1}{M}\sum_{m=1}^{M}\delta_{y_m}$, and we can add the condition that $\lambda_N(s_u) = \lambda_N(u) = c$. This functional is linear and continuous on $\mathcal{H}_K(\Omega)$, since it is a linear combination of points evaluations, which are linear and continuous on $\mathcal{H}_K(\Omega)$.

If we define $\Lambda_N := \{\lambda_1, \ldots, \lambda_N\}$, we have:

- *$\Lambda_N \subset \mathcal{H}_K(\Omega)'$.*
- *Λ_N are linearly independent, since $X_{N-1} \cap Y_M = \emptyset$ (it would be enough to assume that Y_M contains a point that is not contained in X_{N-1}), and we have seen that in the native space of a SPD kernel $\{\delta_x : x \in \Omega\}$ are linearly independent.*

Thus we can apply Proposition 7.3 and use the ansatz

$$s_u(x) := \sum_{j=1}^{N} \alpha_j v_j(x),$$

where v_j is the Riesz representer of λ_j. i.e.,

$$v_j = \begin{cases} K(\cdot, x_j), & 1 \leq j \leq N-1 \\ \frac{1}{M}\sum_{m=1}^{M} K(\cdot, y_m), & j = N, \end{cases}$$

i.e.,

$$s_u(x) := \sum_{j=1}^{N-1} \alpha_j K(x, x_j) + \alpha_N \frac{1}{M}\sum_{m=1}^{M} K(x, y_m).$$

From the proposition, we have that the linear system to solve to find $\alpha \in \mathbb{R}^N$ has a matrix A_Λ with $(A_\Lambda)_{ij} = (v_i, v_j)_{\mathcal{H}_K(\Omega)}$. This means

$$(A_\Lambda)_{ij} = \begin{cases} (K(\cdot, x_i), K(\cdot, x_j))_{\mathcal{H}_K(\Omega)} = K(x_i, x_j), & 1 \leq i, j \leq N-1 \\ \left(K(\cdot, x_i), \frac{1}{M}\sum_{m=1}^{M} K(\cdot, y_m)\right)_{\mathcal{H}_K(\Omega)} = \frac{1}{M}\sum_{m=1}^{M} K(x_i, y_m), & 1 \leq i \leq N-1, j = N \\ \left(\frac{1}{M}\sum_{m=1}^{M} K(\cdot, y_m), K(\cdot, x_j)\right)_{\mathcal{H}_K(\Omega)} = \frac{1}{M}\sum_{m=1}^{M} K(x_j, y_m), & i = N, 1 \leq j \leq N-1 \\ \left(\frac{1}{M}\sum_{n=1}^{M} K(\cdot, y_n), \frac{1}{M}\sum_{m=1}^{M} K(\cdot, y_m)\right)_{\mathcal{H}_K(\Omega)} = \frac{1}{M^2}\sum_{n,m=1}^{M} K(y_n, y_m), & i, j = N. \end{cases}$$

Observe that we obtained for all i, j that

$$(v_i, v_j)_{\mathcal{H}_K(\Omega)} = \lambda_i^x \lambda_j^y K(x, y),$$

where the superscript x or y means that the functional is applied to the corresponding variable. This will be true also in the general case.

If we divide the functionals into $D := \{\lambda_1, \dots, \lambda_{N-1}\}$ and $M := \{\lambda_N\}$, we obtain that A_Λ has a block structure

$$A_\Lambda = \begin{bmatrix} A_{D,D} & A_{D,M} \\ A_{D,M}^T & A_{M,M} \end{bmatrix}$$

where $A_{D,D} \in \mathbb{R}^{(N-1)\times(N-1)}$, $A_{D,M} \in \mathbb{R}^{(N-1)\times 1}$, $A_{D,D} \in \mathbb{R}^{1\times 1}$. Each block contains the evaluations of the functionals M and D in the order specified by the subscript. Also this will happen in the general case.

Again from the proposition, we have that this matrix is invertible, so a unique solution exists.

7.1.2 Linear functionals and SPD kernels

To use the generalized interpolant of Proposition 7.3 in the case $H = \mathcal{H}_K(\Omega)$, we need to:

- Compute the Riesz representer v_λ of a general linear functional $\lambda \in \mathcal{H}_K(\Omega)'$ to construct the ansatz for s_u.
- For the Riesz representers $v_\lambda, v_\mu \in \mathcal{H}_K(\Omega)$ of $\lambda, \mu \in \mathcal{H}_K(\Omega)'$, compute the inner product $(v_\lambda, v_\mu)_{\mathcal{H}_K(\Omega)}$ to construct the invertible matrix A_Λ.

We see in the next proposition that things work as in the case of point-evaluation functionals discussed in the previous example.

Proposition 7.8 (Linear functionals in $\mathcal{H}_K(\Omega)'$)**.** *Let K be a SPD kernel on $\Omega \neq \emptyset$. Let $\lambda, \mu \in \mathcal{H}_K(\Omega)'$ be linear and continuous functionals. Then the following holds:*

i) $\lambda^y K(\cdot, y) \in \mathcal{H}_K(\Omega)$ *(as a function of the first variable).*

ii) $\lambda(f) = (f, \lambda^y K(\cdot, y))_{\mathcal{H}_K(\Omega)}$ *for all $f \in \mathcal{H}_K(\Omega)$, i.e., $\lambda^y K(\cdot, y)$ is the Riesz representer of the functional λ.*

iii) $(v_\lambda, v_\mu)_{\mathcal{H}_K(\Omega)} = \lambda^x \mu^y K(x, y)$.

Proof. Since $\lambda \in \mathcal{H}_K(\Omega)'$, by Theorem 3.7 there exists a unique Riesz representer $v_\lambda \in \mathcal{H}_K(\Omega)$ such that $\lambda(f) = (f, v_\lambda)_{\mathcal{H}_K(\Omega)}$ for all $f \in \mathcal{H}_K(\Omega)$.

Now $f_x(y) := K(x, y)$ is a function of $\mathcal{H}_K(\Omega)$ for all $x \in \Omega$ by definition of reproducing kernel, as a function of the variable y. So we can apply λ to f_x w.r.t. the free variable y. We use the Riesz representer v_λ and the reproducing property to obtain that

$$\lambda^y K(x, y) = \lambda(f_x) = (f_x, v_\lambda)_{\mathcal{H}_K(\Omega)} = (K(\cdot, x), v_\lambda)_{\mathcal{H}_K(\Omega)} = v_\lambda(x).$$

This proves that $v_\lambda(\cdot) = \lambda^y K(\cdot, y)$, so (ii) is satisfied. Since by definition the Riesz representer is a function of $\mathcal{H}_K(\Omega)$, also (i) is proven.

To prove the third points, we use the previous computations and directly obtain

$$(v_\lambda, v_\mu)_{\mathcal{H}_K(\Omega)} = \lambda(v_\mu) = \lambda(\mu^y K(\cdot, y)) = \lambda^x \mu^y K(x, y).$$

□

If we go back to Proposition 7.3, the previous proposition tells us that there exists a generalized interpolant of the form

$$s_u(x) = \sum_{j=1}^{N} \alpha_j v_j(x) = \sum_{j=1}^{N} \alpha_j \lambda_j^y K(x, y),$$

and that the matrix A_Λ is defined as

$$(A_\Lambda)_{ij} = (v_i, v_j)_{\mathcal{H}_K(\Omega)} = \lambda_i^x \lambda_j^y K(x, y),$$

i.e., everything works as in the example.

7.2 Symmetric collocation

We come back to the solution of the PDE 7.1. The plain application of the previous results on generalized interpolation leads directly to the formulation of symmetric collocation.

The idea is the following. To define an approximate solution $s_u \in \mathcal{H}_K(\Omega)$ of the PDE

$$\begin{aligned} Lu(x) &= f(x),\ x \in \Omega \\ Bu(x) &= g(x),\ x \in \partial\Omega, \end{aligned}$$

we consider a set of collocation points $X_N \subset \bar{\Omega}$, and we require that s_u satisfies the equations in these points, i.e.,

$$\begin{aligned} (Ls_u)(x_i) &= (Lu)(x_i) = f(x_i),\ x_i \in \Omega \\ (Bs_u)(x_i) &= (Bu)(x_i) = g(x_i),\ x_i \in \partial\Omega. \end{aligned}$$

These conditions can be rewritten by defining functionals e.g. $\lambda_i := \delta_{x_i} \circ L$ and requiring $\lambda_i(s) = \lambda_i(u) = f(x_i)$. So symmetric collocation fits into the theory of generalized interpolation.

Definition 7.9 (Symmetric collocation)**.** *Let $\Omega \subset \mathbb{R}^d$ be open and bounded and consider linear differential operators L, B and the problem*

$$\begin{aligned} Lu(x) &= f(x),\ x \in \Omega \\ Bu(x) &= g(x),\ x \in \partial\Omega. \end{aligned}$$

For $N \in \mathbb{N}$, consider a set of points $X_N \subset \bar{\Omega}$ and divide them into N_{in} interior points and N_{bd} boundary points, i.e.,

$$X_{in} := X_N \cap \Omega = \{x_1, \dots, x_{N_{in}}\}, \quad X_{bd} := X_N \cap \partial\Omega = \{x_{N_{in}+1}, \dots, x_N\}.$$

Define the set of linear functionals $\Lambda_N := \{\lambda_i, \dots, \lambda_N\}$ as

$$\lambda_i := \begin{cases} \delta_{x_i} \circ L, & x_i \in X_{in} \\ \delta_{x_i} \circ B, & x_i \in X_{bd}. \end{cases}$$

Assume that K is a SPD kernel on Ω such that Λ_N are continuous and linear independent on $\mathcal{H}_K(\Omega)$.

Then the solution by symmetric collocation is defined as the generalized interpolant with data functionals Λ_N, i.e.,

$$\begin{aligned} s_u(x) &= \sum_{j=1}^{N} \alpha_j \lambda_j^y K(x,y) \\ &= \sum_{j=1}^{N_{in}} \alpha_j (\delta_{x_j} \circ L)^y K(x,y) + \sum_{j=N_{in}+1}^{N} \alpha_j (\delta_{x_j} \circ B)^y K(x,y), \end{aligned}$$

where $\alpha \in \mathbb{R}^N$ is the unique solution of the linear system with matrix $(A_\Lambda)_{ij} = (v_i, v_j)_{\mathcal{H}_K(\Omega)}$, i.e.,

$$\begin{bmatrix} A_{LL} & A_{LB} \\ A_{LB}^T & A_{BB} \end{bmatrix} \alpha = \begin{bmatrix} b_L \\ b_B \end{bmatrix},$$

with $A_{LL} \in \mathbb{R}^{N_{in} \times N_{in}}$, $A_{LB} \in \mathbb{R}^{N_{in} \times N_{bd}}$, $A_{BB} \in \mathbb{R}^{N_{bd} \times N_{bd}}$ and

$$\begin{aligned} (A_{LL})_{ij} &= (\delta_{x_i} \circ L)^x (\delta_{x_j} \circ L)^y K(x,y), \quad x_i, x_j \in X_{in} \\ (A_{LL})_{ij} &= (\delta_{x_i} \circ L)^x (\delta_{x_j} \circ B)^y K(x,y), \quad x_i \in X_{in}, x_j \in X_{bd} \\ (A_{LL})_{ij} &= (\delta_{x_i} \circ B)^x (\delta_{x_j} \circ B)^y K(x,y), \quad x_i, x_j \in X_{bd}. \end{aligned}$$

and $b_L \in \mathbb{R}^{N_{in}}$, $b_B \in \mathbb{R}^{N_{bd}}$, $(b_L)_i = f(x_i)$, $(b_B)_i = g(x_i)$.

What is missing is how to find a kernel K such that Λ_N are continuous and linear independent on $\mathcal{H}_K(\Omega)$. We will see in details how to do so in the case L (and B) are of the following type.

Definition 7.10 (Linear differential operator)**.** *Let $\Omega \subset \mathbb{R}^d$ be open and bounded and let $k \in \mathbb{N}$. An operator $L : C^k(\Omega) \to C(\Omega)$ is a linear differential operator of order k with constant coefficients if there exists coefficients $c_a \in \mathbb{R}$ for $a \in \mathbb{N}_0^d$ such that*

$$L := \sum_{|a| \le k} c_a D^a.$$

7.2.1 Differential functionals

What we are left to prove to use symmetric collocation on a linear PDE with L, B as in Definition 7.10 and $k \in \mathbb{N}$ is that, for a properly chosen kernel, it holds that the functionals of type $\lambda_i := \delta_{x_i} \circ D^a$, for $a \in \mathbb{N}_0^d$, $|a| \leq k$, are continuous in $\mathcal{H}_K(\Omega)$, and that they are linearly independent for properly chosen collocation points.

In the following, we use again the notation $\lambda^y K(x,y)$ to indicate the the functional λ is applied to the second variable, but also the notation $D_2^a K(x,y)$, which again denotes that D^a is applied to the second variable. The reason for the different notation is that, to be precise, we should write $(\delta_y \circ D^a)^z K(x,z)$ do denote the same thing.

We recall a part of Proposition 3.14.

Proposition 7.11 (Native space and smoothness)**.** *Let $k \in \mathbb{N}$. Assume $\Omega \subset \mathbb{R}^d$ is open, K is SPD on Ω and $K \in C^{2k}(\Omega \times \Omega)$ for $k \in \mathbb{N}$. Then $\mathcal{H}_K(\Omega) \subset C^k(\Omega)$ and in particular, for all multiindex $a \in \mathbb{N}_0^d$ with $|a| \leq k$ and for all $f \in \mathcal{H}_K(\Omega)$, it holds*

$$D^a f(x) = (f, D_2^a K(\cdot, x))_{\mathcal{H}_K(\Omega)}. \tag{7.4}$$

To prove this result, we assumed that $D_2^a K(\cdot, x) \in \mathcal{H}_K(\Omega)$ for all $x \in \Omega$ and $|a| \leq k$. Under this assumption, setting $c_x := \|D_2^a K(\cdot, x)\|_{\mathcal{H}_K(\Omega)}$ we have by Cauchy-Schwarz that

$$|D^a f(x)| = \left|(f, D_2^a K(\cdot, x))_{\mathcal{H}_K(\Omega)}\right| \leq \|f\|_{\mathcal{H}_K(\Omega)} \|D_2^a K(\cdot, x)\|_{\mathcal{H}_K(\Omega)} = c_x \|f\|_{\mathcal{H}_K(\Omega)},$$

so $\lambda := \delta_x \circ D^a$ is bounded. So to guarantee that $\lambda := \delta_x \circ D^a$ is linear and continuous on the native space $\mathcal{H}_K(\Omega)$ of a kernel, it suffices to take K SPD and $K \in C^{2k}(\Omega \times \Omega)$.

For completeness, we also see a proof of the fact that $D_2^a K(\cdot, x) \in \mathcal{H}_K(\Omega)$ for all $x \in \Omega$ and $|a| \leq k$, even if this was not discussed in the lecture. We first need the following.

Proposition 7.12 (Difference quotient)**.** *Let $k \in \mathbb{N}$, let $\Omega \subset \mathbb{R}^d$ be an open set, and let $f \in C^k(\Omega)$. Let $a \in \mathbb{N}_0^d$ with $|a| \leq k$ and $h > 0$. For $d = 1$ (i.e., $a \in \mathbb{N}_0$), define the difference quotient as*

$$\Delta_{a,h} f(x) = \frac{1}{h^a} \sum_{j=0}^{a} (-1)^{a-j} \binom{a}{j} f(x + jh),$$

and, for $d > 1$ (i.e., $a := (a_1, \ldots, a_d) \in \mathbb{N}_0^d$), as

$$\Delta_{a,h} f(x) := \Delta_{a_1,h}^{x^{(1)}} \cdot \cdots \cdot \Delta_{a_d,h}^{x^{(d)}} f(x),$$

where the superscript $x^{(j)}$ means that the difference is applied to the j-th component of x.

Then it holds

$$\lim_{h \to 0} \Delta_{a,h} f(x) = D^a f(x) \quad \text{for all } x \in \Omega, |a| \leq k.$$

We then can prove the result.

Proposition 7.13 (Derivative of the kernel are in $\mathcal{H}_K(\Omega)$)**.** *Let $k \in \mathbb{N}$. Assume $\Omega \subset \mathbb{R}^d$ is open, K is SPD on Ω and $K \in C^{2k}(\Omega \times \Omega)$ for $k \in \mathbb{N}$. Then for all multiindex $a \in \mathbb{N}_0^d$ with $|a| \leq k$ and for all $x \in \Omega$ we have $D_2^a K(\cdot, x) \in \mathcal{H}_K(\Omega)$.*

Proof. First, since $K \in C^{2k}(\Omega \times \Omega)$, the function $f_a(\cdot) := D_2^a K(\cdot, y)$ is a well defined function on Ω for all $a \in \mathbb{N}_0^d$ with $|a| \leq k$. Moreover, $f_a \in C^k(\Omega)$.

According to Theorem 3.10, to prove $f_a \in \mathcal{H}_K(\Omega)$ we need to prove that f_a is the limit of a Cauchy sequence in $\mathcal{H}_0$, with

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

The idea is to construct the Cauchy sequence using $\Delta_{a,h}$ of Proposition 7.12.

For $n \in \mathbb{N}$, we define f_n as $f_n(\cdot) := \Delta_{a,1/n}^y K(\cdot, y)$. The proposition gives that

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \Delta_{a,1/n}^y K(x, y) = D_2^a K(x, y) = f_a(x) \quad \text{for all } x \in \Omega.$$

We see the case $d = 1$ for simplicity, but the same holds for $d > 1$.

If $d = 1$, f_n has the form

$$f_n(\cdot) = \Delta_{a,1/n}^y K(\cdot, y) = \frac{1}{(1/n)^a} \sum_{j=0}^{a} (-1)^{a-j} \binom{a}{j} K\left(\cdot, y + \frac{j}{n}\right).$$

Since Ω is open, if $y \in \Omega$ and n is large enough, it also holds that all the points $x_j := y + \frac{j}{n}$, $0 \leq j \leq a$ are in Ω. This means that $f_n \in \mathcal{H}_0$, as there exist a finite set $X_N \subset \Omega$ and coefficients $\{\alpha_j(n)\}_{j=1}^N$ s.t.

$$f_n(\cdot) = \sum_{j=0}^{N} \alpha_i(n) K(\cdot, x_j).$$

The sequence $\{f_n\}_{n\in\mathbb{N}}$ is Cauchy: we have

$$\begin{aligned}
(f_n, f_m)_{\mathcal{H}_K(\Omega)} &= \left(f_m, \sum_{j=1}^{N} \alpha_j(n) K(\cdot, x_j) \right)_{\mathcal{H}_K(\Omega)} = \sum_{j}^{a} \alpha_j(n) f_m(x_j) \\
&= \frac{1}{(1/n)^a} \sum_{j=0}^{a} (-1)^{a-j} \binom{a}{j} f_m\left(x + \frac{j}{n}\right) \\
&= \Delta_{a,1/n} f_m(x) = \Delta_{a,1/n}^x \Delta_{a,1/m}^y K(x, y).
\end{aligned}$$

It follows that

$$\begin{aligned}
\lim_{m,n\to\infty} (f_n, f_m)_{\mathcal{H}_K(\Omega)} &= \lim_{m,n\to\infty} \Delta_{a,1/n}^x \Delta_{a,1/m}^y K(x, y) \\
&= \lim_{n\to\infty} \Delta_{a,1/n}^x f_a(x) = D^a f_a(x) = D_1^a D_2^a K(x, y),
\end{aligned}$$

since $f_a \in C^k(\Omega)$ (because $K \in C^{2k}(\Omega)$). Since the two difference operators act on the two variables separately, we can exchange the order of the two limits (in n and in m). In particular, it also holds for $n = m$. Thus we have

$$\|f_n - f_m\|^2_{\mathcal{H}_K(\Omega)} = \|f_n\|^2_{\mathcal{H}_K(\Omega)} + \|f_m\|^2_{\mathcal{H}_K(\Omega)} - 2(f_n, f_m)_{\mathcal{H}_K(\Omega)},$$

so

$$\lim_{m,n\to\infty} \|f_n - f_m\|^2_{\mathcal{H}_K(\Omega)} = \left(D_1^a D_2^a K(x,y)\right)^2 + \left(D_1^a D_2^a K(x,y)\right)^2 - 2\left(D_1^a D_2^a K(x,y)\right)^2 = 0,$$

so $\{f_n\}_{\mathcal{H}_K(\Omega)}$ is a Cauchy sequence.

Since $\{f_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{H}_0$, there exists a function $f \in \mathcal{H}_K(\Omega)$ with $\lim_{n\to\infty} \|f - f_n\|_{\mathcal{H}_K(\Omega)} = 0$. By reproducing property, and since $\{f_n\}_{n\in\mathbb{N}}$ is a convergent sequence in $\mathcal{H}_K(\Omega)$, we can compute

$$\begin{aligned} f(x) &= (f, K(\cdot, x))_{\mathcal{H}_K(\Omega)} = \left(\lim_{n\to\infty} f_n, K(\cdot, x)\right)_{\mathcal{H}_K(\Omega)} = \lim_{n\to\infty} (f_n, K(\cdot, x))_{\mathcal{H}_K(\Omega)} \\ &= \lim_{n\to\infty} f_n(x) = D_2^a K(x, y), \end{aligned}$$

so $D_2^a K(x, y) \in \mathcal{H}_K(\Omega)$. □

Remark 7.14. *This is the first time we see why we need $K \in C^{2k}(\Omega \times \Omega)$ to obtain $\mathcal{H}_K(\Omega) \subset C^k(\Omega)$. Indeed, the condition $K \in C^k(\Omega \times \Omega)$ guarantees that the function $D_2^a K(\cdot, y)$ is a well defined continuous function, for $|a| \leq k$. To have also that $D_2^a K(\cdot, y) \in \mathcal{H}_K(\Omega)$, we need to assume that k additional smooth derivatives exist, i.e., $K \in C^{2k}(\Omega \times \Omega)$.*

Finally, we have the following result for the linear independence of the functionals.

Theorem 7.15 (Linear independence of differential functionals)**.** *Let K be a SPD translational invariant kernel on $\mathbb{R}^d$, i.e., $K(x, y) := \Phi(x - y)$ for all $x, y \in \mathbb{R}^d$. Let $k \in \mathbb{N}$ and assume that $\Phi \in L_1(\mathbb{R}^d) \cap C^{2k}(\mathbb{R}^d)$.*

Let $a_1, \ldots, a_N \in \mathbb{N}_0^d$ with $|a_i| \leq k$, and let $X_N \subset \mathbb{R}^d$. Assume that $a_i \neq a_j$ if $x_i = x_j$.

Then the functionals $\Lambda_N := \{\lambda_1, \ldots, \lambda_N\}$, $\lambda_i := \delta_{x_i} \circ D^{a_i}$, are linearly independent on $\mathcal{H}_K(\mathbb{R}^d)$.

Proof. We assume by contradiction that there exists coefficients c_j such that $\sum_{j=1}^N c_j \lambda_j = 0$. Since $\mathcal{H}_K(\mathbb{R}^d)$ by Proposition 7.13, we have also that their Riesz representers are linearly dependent, and in particular

$$\left\| \sum_{j=1}^N c_j \lambda_j^y K(\cdot, y) \right\|_{\mathcal{H}_K(\mathbb{R}^d)} = 0.$$

We can use now the characterization of the inner product of $\mathcal{H}_K(\mathbb{R}^d)$ in terms of Fourier transform (Theorem 5.12). It states that if $f \in \mathcal{H}_K(\mathbb{R}^d)$, it holds

$$\|f\|^2_{\mathcal{H}_K(\mathbb{R}^d)} = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\Phi}(\omega)} d\omega.$$

Since $K(x,y) = \Phi(x-y)$, we can simplify the application of the functional λ_i as

$$\lambda_i^y K(\cdot, y) = (\delta_{x_i} \circ D^{a_i})^y K(\cdot, y) = (\delta_{x_i} \circ D^{a_i})^y \Phi(\cdot - y) = (-1)^{|a_i|}(D^{a_i}\Phi)(\cdot - x_i).$$

So we can compute the Fourier transform of $\lambda_i^y K(\cdot, y)$. We use (v) of Proposition 5.6 (Fourier transform and translation) and the fact that

$$F(D^a f)(\omega) = (i\omega)^a (Ff)(\omega).$$

We obtain

$$\begin{aligned} F(\lambda_i^y K(\cdot, y))(\omega) &= F\left((-1)^{|a_i|}(D^{a_i}\Phi)(\cdot - x_i)\right)(\omega) = (-1)^{|a_i|} F\left((D^{a_i}\Phi)(\cdot - x_i)\right)(\omega) \\ &= (-i\omega)^{a_i} F\left(\Phi(\cdot - x_i)\right)(\omega) \\ &= (-i\omega)^{a_i} e^{-ix_i^T\omega} F(\Phi)(\omega). \end{aligned}$$

This means that

$$\begin{aligned} 0 = \left\| \sum_{j=1}^N c_j \lambda_j^y K(\cdot, y) \right\|^2_{\mathcal{H}_K(\mathbb{R}^d)} &= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\left| \sum_{j=1}^N c_j (-i\omega)^{a_i} e^{-ix_i^T\omega} \hat{\Phi}(\omega) \right|^2}{\hat{\Phi}(\omega)} d\omega \\ = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \left| \sum_{j=1}^N c_j (-i\omega)^{a_i} e^{-ix_i^T\omega} \right|^2 \hat{\Phi}(\omega) d\omega \end{aligned}$$

Since K is SPD, $\hat{\Phi} > 0$. So we need to have that

$$\sum_{j=1}^N c_j (-i\omega)^{a_i} e^{-ix_i^T\omega} = 0.$$

This can be shown to hold if and only if $c_j = 0$ for all $1 \le j \le N$. The proof is technical (and it does not add too much to this topic), so we omit it. It can be found in Theorem 16.4 in [5]. □

Remark 7.16. *Observe that the theorem proves linear independence on $\mathcal{H}_K(\mathbb{R}^d)$, but this implies linear independence also on $\mathcal{H}_K(\Omega)$, for any $\Omega \subset \mathbb{R}^d$ (as long as $X_N \subset \Omega$). Otherwise, since $\mathcal{H}_K(\Omega) \subset \mathcal{H}_K(\mathbb{R}^d)$, we would have a contradiction.*

7.2.2 Computation of derivatives for RBF kernels

In the case the SPD kernel is radial, i.e., $K(x,y) := \Phi(\|x-y\|_2)$ for $\Phi : \mathbb{R}_{\ge 0} \to \mathbb{R}$, the computation of derivatives is simple, as it reduces to the computation of the univariate derivatives of Φ. We assume in the following that $\Phi : \mathbb{R}_{\ge 0} \to \mathbb{R}$ is smooth enough to compute the derivatives that we will consider.

For example if $d = 2$ we consider a point $x := (x_1, x_2)^T \in \mathbb{R}^2$ (we use this notation only in this section, instead of $(x^{(1)}, x^{(2)})$ as we did before) and let $r(x_1, x_2) := \|x\|_2 = \sqrt{x_1^2 + x_2^2}$. For $i = 1, 2$ we have

$$\partial_{x_i} r(x_1, x_2) = \frac{x_i}{\sqrt{x_1^2 + x_2^2}} = \frac{x_i}{r}.$$

Using the chain rule we have for example

$$\partial_{x_i} \Phi(\|x\|) = \Phi'(r)\, \partial_{x_i} r(x_1, x_2) = \frac{x_i}{r}\, \Phi'(r),$$

and for the second order derivative we have

$$\begin{aligned}
\partial^2_{x_1^2} \Phi(\|x\|) &= \frac{x_1^2}{r^2}\, \Phi''(r) + \frac{x_2^2}{r^3}\, \Phi'(r) \\
\partial^2_{x_2^2} \Phi(\|x\|) &= \frac{x_2^2}{r^2}\, \Phi''(r) + \frac{x_1^2}{r^3}\, \Phi'(r) \\
\partial^2_{x_1 x_2} \Phi(\|x\|) &= \frac{x_1 x_2}{r^2}\, \Phi''(r) - \frac{x_1 x_2}{r^3}\, \Phi'(r).
\end{aligned}$$

In particular, the Laplacian and double Laplacian (which are also radial) can be obtained as

$$\begin{aligned}
\left(\partial^2_{x_1^2} + \partial^2_{x_2^2}\right) \Phi(\|x\|) &= \Phi''(r) + \frac{1}{r}\, \Phi'(r) \\
\left(\partial^4_{x_1^4} + \partial^4_{x_1^2 x_2^2} + \partial^4_{x_2^4}\right) \Phi(\|x\|) &= \Phi''''(r) + \frac{2}{r}\, \Phi'''(r) - \frac{1}{r^2}\, \Phi''(r) + \frac{1}{r^3}\, \Phi'(r)
\end{aligned}$$

This means, for example, that the Laplacian of the kernel can be obtained as

$$\Delta^x K(x, y) = \Phi''(\|x - y\|_2) + \frac{1}{\|x - y\|_2}\, \Phi'(\|x - y\|_2).$$

7.2.3 Error analysis - ideas

We give here some ideas on the error analysis for symmetric collocation. The complete discussion is quite involved, but the main ideas are similar to the ones of the error analysis for interpolation. The full discussion can be found in Chapter 16 of [5].

Assuming that the functionals

$$\{\delta_x \circ L, x \in X_N\} \tag{7.5}$$

are linearly independent over $\mathcal{H}_K(\Omega)$, we have from Proposition 7.3 that the matrix $A_\Lambda := (v_i, v_j)_{\mathcal{H}_K(\Omega)}$ is positive definite.

If we assume that the condition is true for all $x \in \Omega$, i.e.,

$$\{\delta_x \circ L, x \in \Omega\}$$

are linearly independent, we have as a consequence that (7.5) holds for all possible $X_N \subset \Omega$ pairwise distinct.

Moreover, we have from Proposition 7.8 that

$$(v_i, v_j)_{\mathcal{H}_K(\Omega)} = \lambda_i^x \lambda_j^y K(x,y) = (\delta_{x_i} \circ L)^x (\delta_{x_j} \circ L)^y K(x,y).$$

So we can define a function $K_L : \Omega \times \Omega \to \mathbb{R}$ as

$$K_L(w,z) := (\delta_w \circ L)^x (\delta_z \circ L)^y K(x,y),$$

and we have that K_L is a symmetric and strictly positive definite kernel.

The same holds for K_B under the assumption that $\{\delta_x \circ B,\ x \in \Omega\}$ is linearly independent over $\mathcal{H}_K(\Omega)$.

In particular, both kernels have a well defined power function.

With some work (which we do not see here) the error

$$|u(x) - s_u(x)|, \quad x \in \bar{\Omega},$$

can be bounded using the standard bound of Theorem 4.9 in terms of the power function of K_L or K_B, depending on $x \in \Omega$ or $x \in \partial\Omega$.

Then, the estimate of Theorem 4.21 can be applied to find bounds on the error. They will depend on the fill distances $h_{X_{in},\Omega}$ and $h_{X_{bd},\partial\Omega}$, and on the smoothness of the kernels K_L, K_B.

Remark 7.17. *Some final comment on symmetric collocation*

- *We have seen that symmetric collocation is a particular case of generalized interpolation, with functionals defined by evaluating the PDE in the collocation points.*
- *To guarantee that the functionals are continuous and linearly independent, we have seen that it is enough to use a translational invariant kernel K with $K \subset C^{2k}(\Omega \times \Omega)$ (for a PDE of order k).*
- *To compute the ansatz and the matrix, we need only to compute derivatives of the kernel, which can be obtained in closed form.*
- *On the other hand, for a PDE of order k, we need to compute the derivatives of the kernel up to order $2k$, which can be tedious also with RBF kernels.*
- *This formulation fits into the theory of generalized interpolation, so in particular we have the existence of a unique solution and some optimality property. This is particularly good e.g. in the case of kernel associated with Sobolev spaces: in this case Proposition 7.5 tells us that we are computing the solution of minimal Sobolev norm (up to a constant, see Corollary 5.13) among all the solution which satisfy the generalized interpolation conditions.*

- *On the other hand, the problems on ill-conditioning of the linear system occur also in this case, as for standard interpolation. So the numerical solution is often not as good as the one obtained with other methods. But also, the algorithms that we have seen in Chapter 6 can be applied here (with small modifications).*
- *There is also a possible risk connected with this optimality property: If we consider a PDE with multiple solutions, e.g., $\Delta u = f$ (without boundary conditions), the above theory guarantees that for any set of collocation points there exists a unique solution s_u by symmetric collocation. It will be the one of minimal norm, but it is not necessarily desirable to have a unique solution of a problem with multiple ones.*
- *With this formulation, nonlinear PDEs can simply not be solved. Indeed, we need the Riesz representers, which doesn't exists for nonlinear functionals.*

7.3 Non symmetric collocation

There is also a much easier approach to the solution of PDEs with kernel collocation. The only difference is that, instead of using the ansatz of Proposition 7.3, we use a simpler one. This simplification comes at the price that the matrix A_Λ is no more guaranteed to be invertible, and all the optimality properties of Proposition 7.5 are lost. But experimentally it is quite difficult to find collocation points such that the matrix is singular, and usually this method works as good as symmetric collocation. This is also known as method of Kansa.

The idea is still to find an approximate solution $s_u \in \mathcal{H}_K(\Omega)$ of the PDE

$$\begin{aligned} Lu(x) &= f(x),\ x \in \Omega \\ Bu(x) &= g(x),\ x \in \partial\Omega \end{aligned}$$

by considering a set of collocation points $X_N \subset \bar{\Omega}$, and require that s_u satisfies the equations in these points, i.e.,

$$\begin{aligned} (Ls_u)(x_i) &= (Lu)(x_i) = f(x_i),\ x_i \in \Omega \\ (Bs_u)(x_i) &= (Bu)(x_i) = g(x_i),\ x_i \in \partial\Omega. \end{aligned}$$

But now we totally ignore the use of the correct ansatz from generalized interpolation.

We simply use the ansatz for standard interpolation, i.e.,

$$s_u(x) := \sum_{j=1}^{N} \alpha_j K(x, x_j) \quad x \in \bar{\Omega}.$$

Imposing the generalized interpolation conditions results in

$$\lambda_i(s_u) = (\delta_{x_i} \circ L)(s_u) = \sum_{j=1}^{N} \alpha_j (\delta_{x_i} \circ L)^x K(x, x_j) \;\; 1 \leq i \leq N_{in}$$

$$\lambda_i(s_u) = (\delta_{x_i} \circ B)(s_u) = \sum_{j=1}^{N} \alpha_j (\delta_{x_i} \circ B)^x K(x, x_j) \;\; 1 + N_{in} \leq i \leq N.$$

These conditions can be written as a linear system $A_\Lambda \alpha = b$, where b is as before, but now A_Λ is of the form

$$A_\Lambda := \begin{bmatrix} A_L \\ A_B \end{bmatrix}$$

with $A_L \in \mathbb{R}^{N_{in} \times N}$, $A_B \in \mathbb{R}^{N_{in} \times N}$, and

$$(A_L)_{ij} = (\delta_{x_i} \circ L)^x K(x, x_j), \quad x_i \in X_{in}, x_j \in X_N$$
$$(A_B)_{ij} = (\delta_{x_i} \circ B)^x K(x, x_j), \quad x_i \in X_{bd}, x_j \in X_N.$$

in particular, the matrix is still $N \times N$, but not symmetric nor positive definite. In general, there is no guarantee that it is even invertible.

Remark 7.18. *Some comments:*

- *This method is clearly much simpler: The ansatz only requires the kernel and not the derivatives, the linear system has a simpler matrix (only one application of the functional, not two).*
- *It is easier to implement: Only one application of the operators on K is required, so there are less derivatives to compute.*
- *It does not require the use of a kernel $K \in C^{2k}(\Omega)$ for a problem of order k. This means that also the solution s_u can be less smooth, that is good if we are trying to solve a PDE that has not too smooth solutions.*
- *Experimentally, it is usually difficult to find points where A_Λ is singular (but of course it may happen).*
- *Experimentally, it has similar accuracy than symmetric collocation (different studies have shown a slightly better behavior of one or the other method).*
- *There exists optimization or even greedy algorithms that can be applied to this method. They work by selecting a subset of the collocation points (i.e., a submatrix of A_Λ) that is guaranteed to be invertible.*
- *In the case that these optimizations are used, there exists error and convergence results also for this case.*
- *The method can be extended in principle to non linear PDEs.*

8. Support Vector Machines

Richiamo lezione precedente

Di nuovo dividere lo spazio in decision regions

Fare introduzione a parole (compreso idea di support vector - non importa cosa c'e' fuori)

Storia: '63 – '92

Vapnik (Deep Learning Come from the Devil?)(Facebook AI Research)

8.1 Linearly separable datasets and separating hyperplanes

We start by analyzing the case of $X_N \subset \mathbb{R}^d$.

We assume in the following that the dataset is such that X_N is pairwise distinct, and both the positive and negative classes are nonempty, i.e., there exist $x_i \in X_N$ such that $y_i = +1$, and the same for $y_i = -1$.

Definition 8.1 (Linearly separable dataset)**.** *A dataset $D := (X_N, \{y_i\}_{i=1}^N) \subset \Omega \times \{-1, 1\}^N$, with $\Omega \subset \mathbb{R}^d$, is linearly separable in $\mathbb{R}^d$ if there exist $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that*

$$\left.\begin{array}{ll} (w, x_i) + b > 0, & \forall i : y_i = +1 \\ (w, x_i) + b < 0, & \forall i : y_i = -1 \end{array}\right\} \Rightarrow y_i \left((w, x_i) + b\right) > 0, \quad 1 \leq i \leq N. \tag{8.1}$$

In this case, the hyperplane $H := H(w, b) := \{x \in \mathbb{R}^d : (w, x_i) + b = 0\}$ is called a separating hyperplane.

The task of classification is to obtain a separating hyperplane by computing suitable w, b. Observe that there are two things that makes this task not yet exactly stated: First, there exist in general infinitely many hyperplanes. Second, any fixed hyperplane is defined up to scaling of w, b, i.e., for any constant $c > 0$, we have $H(w, b) = H(cw, cb)$. To fix both problems, we see the following.

Proposition 8.2 (Distance between X_N and $H(w, b)$)**.** *Let $D := (X_N, \{y_i\}_{i=1}^N) \subset \Omega \times \{-1, 1\}^N$ with $\Omega \subset \mathbb{R}^d$ be a linearly separable in $\mathbb{R}^d$ dataset and $H(w, b)$ a separating hyperplane. Then the margin $\gamma > 0$ can be computed as*

$$\gamma := \textit{ dist } (X_N, H(w, b)) = \min_{1 \leq i \leq N} \frac{y_i \left((w, x_i) + b\right)}{\|w\|} \tag{8.2}$$

and $\gamma > 0$.

Proof. For a given $x_i \in X_N$, we compute dist $(x_i, H(w, b))$ as the solution of the optimization problem

$$\begin{aligned} &\min_{x \in \mathbb{R}^d} \|x_i - x\|^2 \\ &s.t. (w, x) + b = 0. \end{aligned}$$

The Lagrangian is

$$L(x,\lambda) := \|x_i - x\|^2 + \lambda((w,x)+b) = (x_i,x_i) - 2(x_i,x) + (x,x) + \lambda(w,x) + \lambda b,$$

so

$$\begin{aligned}\partial_x L(x,\lambda) &= -2x_i + 2x + \lambda w = 0\\ \partial_\lambda L(x,\lambda) &= (w,x) + b = 0.\end{aligned}$$

In particular $x = x_i - \frac{\lambda}{2}w$ and

$$0 = (w,x) + b = (w,x_i) - \frac{\lambda}{2}\|w\|^2 + b,$$

i.e.,

$$\lambda = \frac{2}{\|w\|^2}((w,x_i)+b)).$$

Thus

$$\|x - x_i\| = \left\|x_i - x_i - \frac{\lambda}{2}w\right\| = \left\|\frac{(w,x_i)+b}{\|w\|^2}w\right\| = \frac{|(w,x_i)+b|}{\|w\|}.$$

So we have, since $H(w,b)$ is a separating hyperplane, that

$$\text{dist }(X_N, H(w,b)) := \min_{1\le i\le N} \text{dist }(x_i, H(w,b)) = \min_{1\le i\le N} \frac{|(w,x_i)+b|}{\|w\|} = \min_{1\le i\le N} \frac{y_i((w,x_i)+b)}{\|w\|}$$

Assume now $\gamma = 0$. Then there exists $x_i \in X_N$ such that

$$y_i\left((w,x_i)+b\right) = 0,$$

which contradicts the definition of linearly separable dataset. □

An hyperplane H has non unique representation, since we can scale w, b. So it is customary to require the following additional constraint:

Definition 8.3 (Canonical separating hyperplane)**.** *Let D, $H(w,b)$ as above. We call $H(w,b)$ a canonical separating hyperplane if w, b are scaled such that*

$$\gamma = \textit{ dist }(X_N, H(w,b)) = \frac{1}{\|w\|}, \tag{8.3}$$

i.e.,

$$\min_{1\le i\le N} y_i\left((w,x_i)+b\right) = 1. \tag{8.4}$$

[Plot solution] [remark on independence from other points][remark on $(-w,-b)$: different decision

Now we have a unique representation for any given hyperplane, which fixes the second of the two problems. The first one, i.e., choosing the one hyperplane among possibly infinitely many, is solved by looking at the maximum margin separating hyperplane, in the sense of the following definition. It is a meaningful requirement, since it is the "fairest" hyperplane, in the sense that it is equally distant from the positive and negative class.

Definition 8.4 (Maximum margin classifier)**.** *A maximum margin classifier is a classifier that provides a separating hyperplane with the largest possible margin γ, i.e., that maximizes the distance $distX_N, H(w,b)$. In particular, the distance to the two classes is equal.*

Proposition 8.5. *A max. margin, canonical separating hyperplane satisfies, for all x_i positive and x_j negative*

$$\gamma = \frac{1}{\|w\|} \leq \frac{1}{2}\|x_i - x_j\|. \tag{8.5}$$

Proof.

$$(w, x_i) + b = +1$$
$$(w, x_j) + b = -1$$

thus

$$2 = (w, x_i - x_j) \leq \|w\|\|x_i - x_j\|,$$

i.e.

$$\gamma = \frac{1}{\|w\|} \leq \frac{1}{2}\|x_i - x_j\|. \tag{8.6}$$

□

8.1.1 Linear, hard margin SVM in primal form

We can now formulate the SVM optimization problem, which is precisely the problem of finding w, b such that $H(w,b)$ is a maximal margin canonical separating hyperplane. We remark that this is the problem in primal form. The actual algorithm will come to the solution of the dual formulation of the problem.

Definition 8.6 (Primal form of linear, hard margin SVM)**.** *Let $D := (X_N, \{y_i\}_{i=1}^N) \subset \Omega \times \{-1,1\}^N$ with $\Omega \subset \mathbb{R}^d$ be a linearly separable dataset. Then the following optimization problem is called hard margin SVM in primal form:*

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2}\|w\|_2^2 \tag{8.7}$$
$$s.t.\ y_i\left((w, x_i) + b\right) \geq 1, \quad 1 \leq i \leq N. \tag{8.8}$$

If (w^*, b^*) is a solution, the hard margin SVM classifier is defined as

$$s(x) := \mathit{sign}\ \left((w^*, x) + b^*\right). \tag{8.9}$$

Remark 8.7.
- *This problem is clearly the problem of finding a maximal margin separating hyperplane, according to definition and proposition . The maximization of $\frac{1}{\|w\|}$ is replaced by the equivalent minimization of $\frac{1}{2}\|w\|^2$ just for mathematical convenience.*
- *The optimization problem is quadratic with linear constraints. In particular, it is a convex optimization problem.*
- *The assumption that D is linearly separable guarantees that the feasible set is non empty.*
- *To be precise, it should be formulated as* $\inf$*, not* $\min$*. But we will see that a minimum exists.*
- *We see in the following that the hyperplane is also in canonical form.*

Proposition 8.8 ((w^*, b^*) define a canonical separating hyperplane)**.** *Let (w^*, b^*) be a solution of* (8.41)*. Then $H(w, b)$ is a canonical hyperplane.*

Proof. We need to prove that, if (w^*, b^*) is a minimizer, then it holds

$$\begin{aligned}(w^*, x_i) + b^* &= +1 \text{ for some } i \text{ s.t. } y_i = +1\\ (w^*, x_i) + b^* &< -1 \text{ for some } i \text{ s.t. } y_i = -1,\end{aligned}$$

since for any other i it holds, thanks to the constraints, that $y_i\left((w, x_i) + b\right) > 1$.

The argument works as follows: we prove that, if there exists a minimizer (w^*, b^*) such that

$$\begin{aligned}(w^*, x_i) + b^* &= +1 \text{ for some } i \text{ s.t. } y_i = +1\\ (w^*, x_i) + b^* &< -1 \text{ for all } i \text{ s.t. } y_i = -1\end{aligned}$$

or viceversa for $+1, -1$ exchanged, then we can construct another minimizer such that $y_i\left((w, x_i) + b\right) > 1$, $1 \le i \le N$. In this case, we find another solution (w', b') with $\|w'\| < \|w^*\|$, which contradicts the minimality of w^*.

Thus, first assume that

$$\begin{aligned}(w^*, x_i) + b^* &= +1 \text{ for some } i \text{ s.t. } y_i = +1\\ (w^*, x_i) + b^* &< -1 \text{ for all } i \text{ s.t. } y_i = -1.\end{aligned}$$

If $\delta > 0$ is small enough, then $(w^*, \bar{b})$ with $\bar{b} := b^* + \delta$ satisfies

$$\begin{aligned}(w^*, x_i) + \bar{b} &> +1 \text{ for all } i \text{ s.t. } y_i = +1\\ (w^*, x_i) + \bar{b} &< -1 \text{ for all } i \text{ s.t. } y_i = -1,\end{aligned}$$

i.e., $(w^*, \bar{b})$ is still a feasible point, with minimal $\|w^*\|$, and such that $y_i\left((w^*, x_i) + \bar{b}\right) > 1$, $1 \le i \le N$.

In the case that (w^*, b^*) is a minimal solution such that $y_i\left((w^*, x_i) + b^*\right) > 1$ for $1 \leq i \leq N$, there exists $\lambda \in (0, 1)$ such that

$$y_i\left((\lambda w^*, x_i) + \lambda b^*\right) = \lambda y_i\left((w^*, x_i) + b^*\right) > 1, \;\; 1 \leq i \leq N.$$

It follows that $(\lambda w^*, \lambda b^*)$ satisfies the constraints, but, since $\lambda \in (0, 1)$, it holds

$$\|\lambda w^*\| = \lambda \|w^*\| < \|w^*\|,$$

which contradicts the minimality of w^*. □

Finally, the problem has a unique solution.

Theorem 8.9 (Existence and uniqueness of solutions)**.** *Let $D := (X_N, \{y_i\}_{i=1}^N) \subset \Omega \times \{-1, 1\}^N$ with $\Omega \subset \mathbb{R}^d$ be a linearly separable dataset such that both the positive and negative classes are nonempty.*

Then there exists a unique solution (w^*, b^*) of the hard margin SVM in primal form, and $w^* \neq 0$.

Proof. It is a strictly convex minimization problem over a convex set, so what can go wrong is that the solution is an inf, not a min.

Denote as f^* the inf of the target function over the feasible set.

Assume that $\{(w_j, b_j)\}_{j\in\mathbb{N}}$ is a minimizing sequence, i.e, (w_j, b_j) satisfy the constraints and $\lim_{j\to\infty} \|w_j\| = f^*$, with $\{\|w_j\|\}_{j\in\mathbb{N}}$ decreasing.

We can assume that the associated separating hyperplanes $H(w_j, b_j)$ are all canonical (it is just a matter of normalization).

Since the hyperplanes are canonical, we have that

$$0 \leq \frac{1}{\|w\|} = \gamma \leq \frac{1}{2}\|x_i - x_j\|,$$

i.e., there exists $\gamma' > 0$ such that

$$\frac{2}{\|x_i - x_j\|} \leq \|w\| \leq \frac{1}{\gamma'},$$

(first by canonical hyperplane, second by linear separability). This implies that, if a minimizer w^* exists, it satisfies $\|w^*\| > 0$.

Moreover, this proves that the sequence $\{\|w_j\|\}_{j\in\mathbb{N}}$ is bounded, so (Bolzano-Weierstrass Theorem) there exists a convergent subsequence $\{w_{j_k}\}_{k\in\mathbb{N}}$ with $\lim_{k\to\infty} \|w_{j_k}\| = f^*$.

Let $w^* := \lim_{k\to\infty} w_{j_k}$. Since $(\cdot, \cdot)$ and $\|\cdot\|$ are continuous, also w^* is a feasible point. Moreover, also $\|\cdot\|$ is continuous, so it holds

$$\|w^*\| = \left\|\lim_{k\to\infty} w_{j_k}\right\| = \lim_{k\to\infty} \|w_{j_k}\| = f^*,$$

i.e., w^* is a minimizer.

Assume now (w', b') is another minimizer. We need to have $\|w^*\| = \|w'\|$.

If $w^* \neq w'$, for all $\lambda \in [0,1]$, the vector $w_\lambda := \lambda w^* + (1-\lambda)w'$ satisfies the constraints since, for all $1 \leq i \leq N$,

$$\begin{aligned} & y_i\left((w_\lambda, x_i) + b\right) = \lambda y_i\left((w^*, x_i) + b\right) + (1-\lambda) y_i\left((w', x_i) + b\right) \\ \geq\ & \lambda + (1-\lambda) = 1. \end{aligned}$$

Moreover, unless $w^* = w'$, we have $\|w_\lambda\| < \lambda\|w^*\| + (1-\lambda)\|w'\| = \|w^*\| = \|w'\|$ by strict convexity. So also $w' = w^*$.

If $b^* \neq b'$, assume $b^* < b'$. Since, for all $1 \leq i \leq N$, both satisfy

$$\begin{aligned} y_i\left((w^*, x_i) + b^*\right) &\geq 1 \\ y_i\left((w^*, x_i) + b'\right) &\geq 1. \end{aligned}$$

Since $b^* < b'$, we have

$$\begin{aligned} y_i\left((w^*, x_i) + b^*\right) &> 1 \quad \text{for all } i \text{ such that } y_i = -1 \\ y_i\left((w^*, x_i) + b'\right) &> 1 \quad \text{for all } i \text{ such that } y_i = +1. \end{aligned}$$

Thus we can consider the pair $(w^*, \frac{b^*+b'}{2})$. It satisfies

$$y_i\left((w^*, x_i) + \frac{b^*+b'}{2}\right) = \frac{1}{2} y_i\left((w^*, x_i) + b^*\right) + \frac{1}{2} y_i\left((w^*, x_i) + b'\right) > \frac{1}{2} + \frac{1}{2} = 1,$$

so it is feasible. Since the inequality is strictly satisfied, we can shrink w^* and obtain a better solution, i.e., there exists $\lambda \in (0,1)$ such that $(\lambda w^*, \frac{b^*+b'}{2})$ is still feasible, but $\|\lambda w^*\| < \|w^*\|$, which contradicts the optimality of w^*

□

Remark 8.10. *We have now a well defined optimization problem with $d+1$ variables and N constraints. We have seen that it has a unique solution, so we could be happy with this. Nevertheless, it is convenient to derive an equivalent formulation of the problem, which will have N variables and $N+1$ constraints. The reason this new formulation is convenient is*

- *It allows to define the concept of support vectors, and to deduce sparsity of SVM representation. This will also allow efficient predictions.*
- *There will be a very efficient algorithm (better that generic quadratic optimization) derived from this formulation.*
- *Most important, it allows to introduce kernels via feature maps in an efficient way and implicitly, so that we can transform the algorithm into a nonlinear one.*

$$\min_{w \in H, b \in \mathbb{R}} \frac{1}{2}\|w\|_2^2 \tag{8.10}$$

$$s.t.\ y_i\left((w, \phi(x_i))_H + b\right) \geq 1, \quad 1 \leq i \leq N. \tag{8.11}$$

prediction:

$$s(x) := \text{ sign } \left((w^*, \phi(x))_H + b^*\right). \tag{8.12}$$

We need some results on convex optimization to derive this dual formulation.

8.1.2 Convex optimization

Sch. Chapter 6.

Definition 8.11 (Convex optimization with linear constraints)**.** *Let $F \subset \mathbb{R}^d$ be convex and nonempty, $f : F \to \mathbb{R}$ convex, $g_i, h_j : F \to \mathbb{R}$ affine functions for $1 \leq i \leq m$, $1 \leq j \leq l$, $l, m \in \mathbb{N}$.*

We look at the solution $x^* \in F$ of the problem

$$\begin{cases} \min_{x \in F} f(x) \\ s.t.\ g_i(x) \leq 0, \ \ 1 \leq i \leq m \\ \quad h_j(x) = 0, \ \ 1 \leq j \leq l \end{cases} \tag{8.13}$$

Remark: - any local optimum is global
- the set of minimizers is a convex set
- if f strictly convex, the solution is unique (if it exists)
Generalization of Lagrange multipliers for inequality constraints:

Definition 8.12 (Karush-Kuhn-Tucker points)**.** *Consider the problem defined in Definition 8.11. For $\mu \in \mathbb{R}^m$, $\lambda \in \mathbb{R}^l$, define the Lagrangian / Lagrange function*

$$L(x, \mu, \lambda) := f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{l} \lambda_j h_j(x).$$

A point $(x^*, \mu^*, \lambda^*) \in \mathbb{R}^{d+m+l}$ is a KKT point if

$$\partial_x L(x^*, \mu^*, \lambda^*) = \nabla_x f(x^*) + \sum_{i=1}^{m} \mu_i^* \nabla_x g_i(x^*) + \sum_{j=1}^{l} \lambda_j^* \nabla_x h_j(x^*) = 0 \tag{8.14}$$

$$g_i(x^*) \leq 0, \ 1 \leq i \leq m \tag{8.15}$$

$$h_j(x^*) = 0, \ 1 \leq j \leq l \tag{8.16}$$

$$\mu_i^* \geq 0, \ 1 \leq i \leq m \tag{8.17}$$

$$\mu_i^* g_i(x^*) = 0, \ 1 \leq i \leq m. \tag{8.18}$$

Theorem 8.13 (KKT points and optimality)**.** *Consider the problem defined in Definition 8.11, with $f \in C^1(F)$. We have the following:*

i) If $x^* \in F$ is a minimizer, then there exists $\mu^* \in \mathbb{R}^m$, $\lambda^* \in \mathbb{R}^l$ such that $(x^*, \mu^*, \lambda^*) \in \mathbb{R}^{d+m+l}$ is a KKT point (necessary conditions).

ii) If $(x^*, \mu^*, \lambda^*) \in \mathbb{R}^{d+m+l}$ is a KKT point, then x^* is a minimizer (sufficient conditions).

8.1.3 Linear, hard margin SVM in dual form

Definition 8.14 (Dual form of linear, hard margin SVM)**.** *Let $D := (X_N, \{y_i\}_{i=1}^N) \subset \Omega \times \{-1, 1\}^N$ with $\Omega \subset \mathbb{R}^d$ be a linearly separable dataset. Then the following optimization problem is called hard margin SVM in dual form:*

$$\max_{\alpha \in \mathbb{R}^N} \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_k (x_i, x_j) \tag{8.19}$$

$$s.t. \sum_{i=1}^N \alpha_i y_i = 0 \tag{8.20}$$

$$\alpha_i \geq 0, \;\; 1 \leq i \leq N \tag{8.21}$$

Proposition 8.15 (Dual and primal SVM)**.** *Let $\alpha^* \in \mathbb{R}^N$ be a solution of* (8.19)*. Define $w* \in \mathbb{R}^d$, $b^* \in \mathbb{R}$ by*

$$w^* := \sum_{i=1}^N \alpha_i^* y_i x_i \tag{8.22}$$

and b^* such that, for an arbitrary i with $\alpha_i^* \neq 0$, it holds

$$y_i((w^*, x_i) + b^*) = 1. \tag{8.23}$$

Then (w^*, b^*) is the unique primal, hard margin SVM (8.41)*.*

Proof. The strategy is to apply the KKT theorem, part 2. In this case the optimization problem is the SVM primal

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{2} \|w\|_2^2 \tag{8.24}$$

$$s.t. \; y_i \left((w, x_i) + b\right) \geq 1, \;\; 1 \leq i \leq N. \tag{8.25}$$

which is of the form of definition

$$\begin{cases} \min_{x \in F} f(x) \\ s.t. \; g_i(x) \leq 0, \;\; 1 \leq i \leq m \\ \quad h_j(x) = 0, \;\; 1 \leq j \leq l \end{cases} \tag{8.26}$$

with $x = (w, b)$, $F = R^{d+1}$, $l = 0$, $m = N$, $g_i(x) := 1 - y_i((w, x_i) + b)$. Moreover, f is convex, F nonempty and g_i affine, so we are in the scenario of the definition.

It follows that the Lagrangian

$$L(x, \mu, \lambda) := f(x) + \sum_{i=1}^m \mu_i g_i(x) + \sum_{j=1}^l \lambda_j h_j(x)$$

is

$$L(w, b, \alpha) := \frac{1}{2} \|w\|^2 + \sum_{i=1}^N \alpha_i (1 - y_i((w, x_i) + b)).$$

We can use KKT theorem since f is continuously differentiable. This means that we prove that (w^*, b^*, α^*) obtained from the dual SVM is a KKT point. It follows from the theorem that x^* is a minimizer of the primal SVM (which we proved to be also unique).

So we need to check the KKT conditions (h is not present)

$$\partial_x L(x^*, \mu^*, \lambda^*) = 0 \tag{8.27}$$
$$g_i(x^*) \leq 0,\ 1 \leq i \leq m \tag{8.28}$$
$$\mu_i^* \geq 0,\ 1 \leq i \leq m \tag{8.29}$$
$$\mu_i^* g_i(x^*) = 0,\ 1 \leq i \leq m. \tag{8.30}$$

We see them:

- $\boldsymbol{\mu_i^* \geq 0,\ 1 \leq i \leq m}$: In this case it is $\alpha_i^* \geq 0$ for $1 \leq i \leq N$, which is satisfied because it is one of the constraints definition 8.19.

- $\boldsymbol{\partial_x L(x^*, \mu^*, \lambda^*) = 0}$: In this case it is

$$0 = \nabla_w L(w^*, b^*, \alpha^*) = w^* - \sum_{i=1}^N y_i \alpha_i^* x_i \Leftrightarrow w = \sum_{i=1}^N y_i \alpha_i^* x_i$$
$$0 = \nabla_b L(w^*, b^*, \alpha^*) = -\sum_{i=1}^N y_i \alpha_i^* \Leftrightarrow \sum_{i=1}^N y_i \alpha_i^* = 0.$$

  The first is satisfied because of the definition of w^*, the second because of the constraints.

- $\boldsymbol{\mu_i^* g_i(x^*) = 0,\ 1 \leq i \leq m}$: In this case $\alpha_i^*(1 - y_i((w^*, x_i) + b^*)) = 0$ for $1 \leq i \leq N$. We prove that, if $\alpha_i^* \neq 0$, then $y_i((w^*, x_i) + b^*) = 1$.

  First, observe that, if $\alpha_i^* \neq 0$, then there exists also $\alpha_j^* \neq 0$ with $y_j = -y_i$. Indeed, from the constraint $\alpha_i \geq 0$ we have $\alpha_i > 0$. Since the second constraint is $\sum_{i=1}^N \alpha_i y_i = 0$, then there exists $j \neq i$ with $y_i = -y_j$ and $\alpha_j^* \neq 0$.

  Assume that the index i is the one used for the normalization to define b^*, i.e., $y_i((w^*, x_i) + b^*) = 1$. We have the following:

  - If $\alpha_j^* \neq 0$ and $y_j = -y_i$: we prove that $(w^*, x_j) = (w^*, x_i) - \frac{2}{y_i}$. From this it follows:

$$y_j\left((w^*, x_j) + b^*\right) = -y_i((w^*, x_j) + b^*) = -y_i\left((w^*, x_i) - \frac{2}{y_i} + b^*\right)$$
$$= -y_i\left((w^*, x_i) + b^*\right) + 2 = 1.$$

  - If $\alpha_j^* \neq 0$ and $y_j = y_i$: we prove that $(w^*, x_j) = (w^*, x_i)$. From this it follows:

$$y_j\left((w^*, x_j) + b^*\right) = y_i((w^*, x_j) + b^*) = y_i((w^*, x_i) + b^*) = 1.$$

- $\boldsymbol{g_i(x^*) \leq 0,\ 1 \leq i \leq m}$: In this case it is $y_i((w^*, x_i) + b^*) \geq 1$ for $1 \leq i \leq N$. If $\alpha_i \neq 0$, it follows from the previous point. Assume $\alpha_i = 0$. If we prove that there exists $j \neq i$ with $y_i = -y_j$, $\alpha_j \neq 0$, and such that $y_j(w^*, x_i) \leq y_j(w^*, x_j) - 2$, we have

$$\begin{aligned} y_i((w^*, x_i) + b^*) &= -y_j((w^*, x_i) + b^*) = -y_j(w^*, x_i) - y_j b^* \\ &\geq -y_j(w^*, x_j) + 2 - y_j b^* = -y_j((w^*, x_j) + b^*) + 2 = 1. \end{aligned}$$

For simplicity, we collect the three assumptions in a separate proposition, so the proof is complete.

□

Proposition 8.16. *Assume that the index i is the one used for the normalization to define b^*, i.e., $y_i((w^*, x_i) + b^*) = 1$. Then*

i) If $\alpha_j^* \neq 0$ and $y_j = -y_i$, then it holds that $(w^*, x_j) = (w^*, x_i) - \frac{2}{y_i}$.

ii) If $\alpha_j^* \neq 0$ and $y_j = y_i$, then it holds that $(w^*, x_j) = (w^*, x_i)$.

On the other hand, assume that i is such that $\alpha_i = 0$. Then

iii) There exists $j \neq i$ with $y_i = -y_j$, $\alpha_j \neq 0$, and such that $y_j(w^*, x_i) \leq y_j(w^*, x_j) - 2$.

Proof. Assume α^* is a solution. Assume $v \in \mathbb{R}^N$ is a feasible direction, i.e., there exists $\delta > 0$ such that for all $t \in [-\delta, \delta]$, we have $\alpha^* + tv$ feasible. Then by setting $J(\alpha) := \sum_{i=1}^N \alpha_i - \sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j (x_i, x_j)$, we have $\nabla_{\alpha^*} J(\alpha) \cdot v = 0$, where

$$\begin{aligned} \partial_{\alpha_l} J(\alpha^*) &= 1 - \sum_{j=1}^n y_l \alpha_j^* y_j (x_l, x_j) = 1 - y_l \left[\sum_{j=1}^n \alpha_j^* y_j (x_l, x_j) \right] \\ &= 1 - y_l \left(\sum_{j=1}^n \alpha_j^* y_j x_j, x_l \right) = 1 - y_l \left(w^*, x_l \right). \end{aligned}$$

Now if i is such that $y_i((w^*, x_i) + b^*) = 1$ and $\alpha_j^* \neq 0$ and $y_j = -y_i$, the vector $v := e_i + e_j$ is a feasible direction. Indeed, taking $\delta := \min(\alpha_i^*, \alpha_j^*)$, we have $\alpha^* + tv \geq 0$ (since only α_i^*, α_j^* are affected, and by definition of δ), and

$$\sum_{l=1}^N (\alpha^* + tv)_l y_l = \sum_{l=1}^N \alpha^* y_l + t \sum_{l=1}^N v_l y_l = 0 + t y_i + t y_j = 0,$$

since $\sum_{l=1}^N \alpha^* y_l = 0$ by constraints.

It follows that

$$\begin{aligned} 0 = \nabla_\alpha J(\alpha^*) \cdot v &= \sum_{l=1}^N (1 - y_l \left(w^*, x_l \right)) v_l \\ &= 1 - y_i \left(w^*, x_i \right) + 1 - y_j \left(w^*, x_j \right), \end{aligned}$$

thus

$$y_j\,(w^*, x_j) = 2 - y_i\,(w^*, x_i)$$

and since $y_i = -y_j$ we conclude that

$$(w^*, x_j) = (w^*, x_i) - \frac{2}{y_i}.$$

If instead i is as above (i.e., such that $y_i((w^*, x_i) + b^*) = 1$) and $\alpha_j^* \neq 0$ and $y_j = y_i$, the same idea applies with $v := e_i - e_j$, and $0 = \nabla_\alpha J(\alpha^*) \cdot v$ implies that $(w^*, x_j) = (w^*, x_i)$. □

Theorem 8.17 (Existence of solution)**.** *Let $D := (X_N, \{y_i\}_{i=1}^N) \subset \Omega \times \{-1, 1\}^N$ with $\Omega \subset \mathbb{R}^d$ be a linearly separable dataset such that both the positive and negative classes are nonempty.*

Then there exists a solution $\alpha^* \in \mathbb{R}^N$ of the hard margin SVM in dual form.

Proof. We have to show that there exists a solution $\alpha^* \in \mathbb{R}^N$ of the problem ref, which we can rewrite as

$$\min_{\alpha \in \mathbb{R}^N} \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j (x_i, x_j) - \sum_{i=1}^N \alpha_i \tag{8.31}$$

$$s.t. \sum_{i=1}^N \alpha_i y_i = 0 \tag{8.32}$$

$$\alpha_i \geq 0, \quad 1 \leq i \leq N. \tag{8.33}$$

First, there exists feasible solutions. Indeed, under the present assumptions we proved in Theorem 8.9 that there exists a unique solution (w^*, b^*) of the primal problem. Moreover, using Theorem **??**, part (ii), we have that there exists (a not necessarily unique) $\alpha^* \in \mathbb{R}^N$, such that (w^*, b^*, α^*) satisfies the KKT conditions. Finally, we proved in the proof of Proposition 8.15 that the KKT conditions for our problem implies in particular that $\sum_{i=1}^N \alpha_i y_i = 0$ and $\alpha_i \geq 0$, $1 \leq i \leq N$. This means that there exist feasible points.

Then the existence of a solution follows by the same argument of Theorem 8.9, using a minimizing sequence. The argument works because the constraints are continuous, and the objective function is in this case

$$f(\alpha) := \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j (x_i, x_j) - \sum_{i=1}^N \alpha_i,$$

which is continuous and convex. □

Remark 8.18. *Some comments:*

- *Observe that to obtain uniqueness in the previous proposition, we would need to have that the objective function f is strictly convex (as we did in the proof of Theorem 8.9 to prove uniqueness of w^*). But f is not strictly convex in general. Indeed, f can be written as*

$$f(\alpha) := \frac{1}{2}\alpha^T Q\alpha - u^T\alpha, \;\; Q_{ij} := y_i y_j (x_i, x_j), u := [1, 1, \ldots, 1]^T,$$

and the matrix Q is in general positive semidefinite, since X_N can be linearly dependent (and they are for sure if $N > d$). A sufficient condition for uniqueness is that X_N is linearly independent in $\mathbb{R}^d$, which is not the case in general.

- *Even if the dual solution α^* is non unique, we have that the solutions are a convex set. Moreover, and more important, we proved that any solution α^* allows to define w^*, b^* such that (w^*, b^*) is the unique solution of the primal problem, i.e., the separating hyperplane is unique and the classifier $s(x) := \; sign \; ((w^*, x) + b^*)$ is uniquely defined.*

- *Choice of i for b^*. We defined b^* such that, for an arbitrary i with $\alpha_i^* \neq 0$, it holds*

$$y_i((w^*, x_i) + b^*) = 1. \tag{8.34}$$

We proved that if $\alpha_i^* \neq 0$, then $y_i((w^*, x_i) + b^*) = 1$.

- *We also proved that, for all $1 \leq i \leq N$, we have that if $\alpha_i^* \neq 0$, then $y_i((w^*, x_i) + b^*) = 1$, and that if $y_i((w^*, x_i) + b^*) > 1$, then $\alpha_i^* = 0$. This means that α_i^* can be nonzero if and only if x_i is on the boundary. Such x_i are called support vectors [Figure]. Observe that, thanks to non uniqueness of solution, we can have $y_i((w^*, x_i) + b^*) = 1$ and $\alpha_i = 0$.*

  This is a fundamental property in view of the definition of the classifier,

$$s(x) := \; sign \; ((w^*, x) + b^*) = \; sign \; \left(\sum_{i=1}^{N} \alpha_i y_i (x_i, x) + b^* \right),$$

  because it means that s can be evaluated using a possibly very small sum.

- *zero training error because of linear separability. i.e., if D is the training dataset and $s(x) := \; sign \; ((w^*, x) + b^*))$, we have $R_{emp}(s, D) = 0$.*

- *The KKT condition will be the central tool for deriving the Sequential Minimal Optimization algorithm (SMO). So we rewrite them here*

  - $\alpha_i^* \geq 0$,
  - $w^* = \sum_{i=1}^N y_i \alpha_i^* x_i$ *and* $\sum_{i=1}^N \alpha_i y_i = 0$
  - $\alpha_i^*(1 - y_i((w^*, x_i) + b^*)) = 0$,
  - $y_i((w^*, x_i) + b^*) \geq 1$,

Example 8.19 (Non uniqueness of solutions)**.** *A minimal example is the following. Consider X_4 with [Figure].*

Intuitively, the unique solution of the primal SVM is $w^* = (-1, 0)$ and $b^* = 0$. This defines the correct classifier

$$s(x) = \textit{ sign } ((w^*, x) + b^*) = \textit{ sign } \left(-x^{(1)}\right)$$

and it is a canonical hyperplane since $y_i((w^*, x_i) + b^*) = 1$ for some i (for all i, in this case).

Since $N = 4$, we have $\alpha^* \in \mathbb{R}^4$. Solutions are

$$\alpha^* = [1/4, 1/4, 1/4, 1/4], \alpha^* = [1/2, 1/2, 0, 0]\alpha^* = [0, 0, 1/2, 1/2].$$

(and any convex combination).

To see this formally, we need to show that α^* satisfies the KKT conditions, i.e.,

- $\alpha_i^* \geq 0$,
- $w^* = \sum_{i=1}^4 y_i \alpha_i^* x_i$ *and* $\sum_{i=1}^4 \alpha_i y_i = 0$
- $\alpha_i^*(1 - y_i((w^*, x_i) + b^*)) = 0$,
- $y_i((w^*, x_i) + b^*) \geq 1$,

The condition is not necessary: you can try $x_1 = (-1, 0)$, $x_2 = (1, 0)$, with $y_1 = 1 = -y_2$.

8.2 Nonlinear hard margin SVM

Now we formulated the SVM training and prediction in a form that involves x_i only via inner products. We can then map the data with a feature map $\Phi : \Omega \to H$, and replace every inner product (x, x_i) with $K(x, x_i) = (\Phi(x), \Phi(x_i))_H$.

Definition 8.20 (Nonlinear hard margin SVM)**.** *Let Ω be an arbitrary nonempty set and let $K : \Omega \times \Omega \to \mathbb{R}$ be a positive definite kernel.*

Let $N \in \mathbb{N}$ and $D := (X_N, \{y_i\}_{i=1}^N) \subset \Omega \times \{-1, 1\}^N$. Assume furthermore that $D := (\{\Phi(x_i)\}_{i=1}^N, \{y_i\}_{i=1}^N) \subset H \times \{-1, 1\}^N$ is linearly separable in H, i.e., there exist $w \in H$ and $b \in \mathbb{R}$ such that

$$y_i \left((w, \Phi(x_i))_H + b\right) \geq 1, \quad 1 \leq i \leq N.$$

Then the following optimization problem is called nonlinear hard margin SVM:

$$\min_{\alpha \in \mathbb{R}^N} \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_k K(x_i, x_j) - \sum_{i=1}^N \alpha_i \tag{8.35}$$

$$s.t. \sum_{i=1}^N \alpha_i y_i = 0 \tag{8.36}$$

$$\alpha_i \geq 0, \quad 1 \leq i \leq N. \tag{8.37}$$

The corresponding classifier is defined as

$$s(x) := \textit{sign}\ (f(x))\,. \tag{8.38}$$

where

$$f(x) = \sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*$$

and b^* is such that

$$f(x_i) = y_i$$

for some i with $\alpha_i^* > 0$.

Remark 8.21.

- *The definition is another way to write the following: $w^* = \sum_{i=1}^{N} \alpha_i^* y_i \Phi(x_i)$, so*

$$(w^*, \Phi(x))_H + b^* = \sum_{i=1}^{N} \alpha_i^* y_i (\Phi(x_i), \Phi(x))_H + b^* = \sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^* = f(x).$$

  The definition of b^* is: for i such that $\alpha_i^* \neq 0$, let b^* be such that

$$1 = y_i((w^*, \Phi(x_i))_H + b^*) = y_i(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*) = y_i f(x), \tag{8.39}$$

  i.e., $f(x_i) = y_i$.

- *It is difficult in general to understand a priori if a dataset will be linearly separable after mapping into the feature space. Nevertheless, we know that a solution exists if and only if the dataset is linearly separable. Thus one option is to try to apply the algorithms, and it will converge if and only if the data are actually linearly separable. Observe that this does not imply that the interpolation solution is the SVM solution, and in general it can be much less sparse.*

- *The dataset is always linearly separable. Indeed, it is enough to define $f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)$ as the interpolant of X_N with target values $\{y_i\}$. With $b^* = 0$.*

- *The same issue with non unique solutions is present here in the case K is PD but not SPD. Indeed, the matrix Q of the quadratic form is now the kernel matrix A, which can be singular.*

- *Instead, in the case of SPD kernels, the solution is unique.*

8.3 Nonlinear soft margin SVM

Motivation: we want to work with non linear separable data. Also, in some case it can be beneficial to relax the condition.

This means that we allow some margin violation, i.e., we accept that some point is not correctly classified.

This is done by using the Hinge loss:

$$L(x, y, f) := \max(0, 1 - yf(x)).$$

[Plot] With $f(x) := (w, x) + b$. Observe that, when x is correctly classified, we have $L(x_i, y_i, f(x_i)) = 0$. The problems arise when y_i and $f(x_i)$ have different sign (wrong classification), or when $y_i((w, x) + b) \leq 1$ (i.e., correct classification but within the margin). We allow $L(x_i, y_i, f(x_i)) > 0$, but we want to keep it small.

The balance between exact classification and non exact one is controlled by a positive parameter $C > 0$.

Ideally, we would like to minimize

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N} L(x_i, y_i, f(x_i)), \tag{8.40}$$

but this is a nonsmooth problem. So, we can instead replace $L(x_i, y_i, f(x_i))$ with an upper bound and minimize it. That is, we introduce slack variables $\{\xi_i\}_{i=1}^N$. We require that they are an upper bound, i.e.,

$$\begin{aligned}
&\xi_i \geq L(x_i, y_i, f(x_i)) = \max(0, 1 - y_i f(x_i)) \geq 1 - y_i f(x_i) = 1 - y_i((w, x) + b)\\
&\xi_i \geq 0.
\end{aligned}$$

Definition 8.22 (Primal form of linear, soft margin SVM)**.** *Let $D := (X_N, \{y_i\}_{i=1}^N) \subset \Omega\times\{-1, 1\}^N$ with $\Omega \subset \mathbb{R}^d$ be a dataset with nonempty classes. Then the following optimization problem is called soft margin SVM in primal form:*

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}, \xi\in\mathbb{R}^N} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N} \xi_i \tag{8.41}$$

$$s.t.\ y_i\left((w, x_i) + b\right) \geq 1 - \xi_i,\ \ 1 \leq i \leq N \tag{8.42}$$

$$\xi \geq 0,\ 1 \leq i \leq N. \tag{8.43}$$

Remark 8.23.

- *THis is still a quadratic problem with linear constraints*
- *If a point is correctly classified and at a distance greater than γ from the separating hyperplane, then $\xi_i = 0$. If it within the margin (correcly or not) it has $\xi_i > 0$ (less or greater than 1 depending on correct or incorrect).*

We can do the same process as before to derive a dual problem, prove via KKT conditions that it is equivalent to the primal, and then introduce a kernel. The final result is the following.

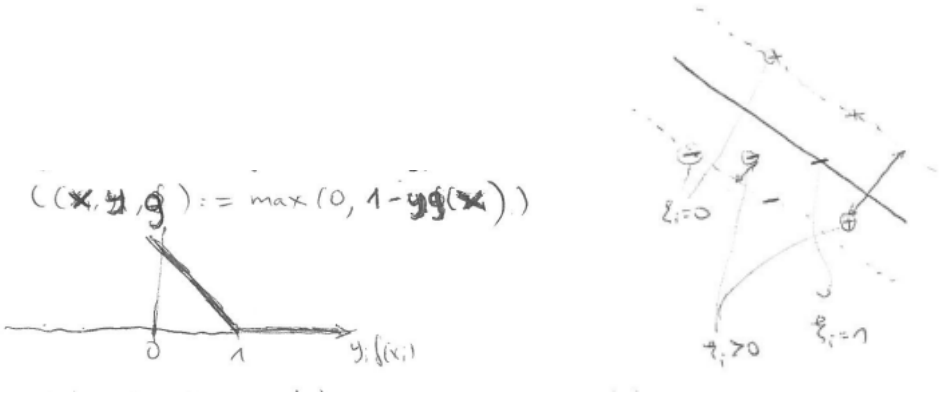

Figure 8.1: The function f_ε.

Definition 8.24 (Nonlinear soft margin SVM)**.** *Let Ω be an arbitrary nonempty set and let $K : \Omega \times \Omega \to \mathbb{R}$ be a positive definite kernel.*

Let $N \in \mathbb{N}$ and $D := (X_N, \{y_i\}_{i=1}^N) \subset \Omega \times \{-1, 1\}^N$ be a dataset with nonempty classes.. Let $C > 0$ be a regularization parameter.

Then the following optimization problem is called nonlinear soft margin SVM:

$$\min_{\alpha \in \mathbb{R}^N} \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i \tag{8.44}$$

$$s.t. \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{8.45}$$

$$0 \leq \alpha_i \leq C, \;\; 1 \leq i \leq N. \tag{8.46}$$

The corresponding classifier is defined as

$$s(x) := \; sign \; (f(x)) \, . \tag{8.47}$$

where

$$f(x) = \sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*$$

and b^* is such that

$$f(x_i) = y_i$$

for some i with $\alpha_i^* \in (0, C)$.

Remark 8.25.

- *The only modification is the box constraints, i.e., $0 \leq \alpha_i \leq C$. The intuition is that without C, the solution would then to produce $\alpha_i \to \infty$.*
- *In particular, $C \to \infty$ is hard margin.*

- *Now the support vectors are more complex:*
  - *If $y_i f(x_i) > 1$ then $\alpha_i = 0$ not a support vector.*
  - *If $y_i f(x_i) < 1$ then $\alpha_i = C$ bounded support vector (margin violation)*
  - *If $y_i f(x_i) = 1$ then $\alpha_i \in [0, C]$ (i.e., still non zero inside the margin)*
  - *If $\alpha_i \in (0, C)$, then $y_i f(x_i) = 1$. (i.e., bounded and non zero only if on the margin, unbounded SV)*
- *If there exist no bounded support vector, i.e., $\alpha_i < C$ for all i, then the data are linearly separable. It this case soft =hard. Moreover, increasing C will not change the solution.*
- *If D is the training dataset and $s(x)$ is the soft margin SVM classifier with $C > 0$, we have $R_{emp}(s, D) \leq N_{BSV}/N$, where $N_{BSV} := \{i : \alpha_i = C\}$. Indeed, we have misclassification if and only if $y_i f(x_i) < 0$.*

Semiparametric representer theorem
Derive alternative representation of primal via Hinge loss
KKT holds

8.4 Efficient implementation

We see now ... solution:

- In principle, the problem to be solved is a linearly constrained, convex quadratic program. Thus, it can be solved by standard solvers (like `quadprog` in Matlab)
- The problem can be Nevertheless difficult to solve for large amount of data. There is a specialized algorithm, which is the topic of this section, that works by splitting the problem into many as small as possible subproblems. It is called Sequential Minimal Optimization (SMO) and was introduced by John Platt in 1998 at Microsoft Research.
- There are standard and very efficient implementations of this algorithm, and in general of SVM-related algorithms.
  - LIBSVM, from National Taiwan University (https://www.csie.ntu.edu.tw/ cjlin/libsvm/ Reference implementation. There are interfaces for almost any programming language.
  - Both Matlab and Python have built-in solvers: `fitcsvm` (https://de.mathworks.com/he and `svm` module in Scikit-learn (http://scikit-learn.org/stable/modules/svm.html) (both use SMO, the second directly LIBSVM)
  - liquidSVM (http://www.isa.uni-stuttgart.de/software/)

8.4.1 Computation remarks

- To compute b, instead of picking a random index i such that $\alpha_i = C$, it is common to average over all i such that $\alpha_i = C$.
- remarks on cross validation, usually using as error function the empirical risk

8.4.2 Sequential Minimal Optimization

- It is an iterative method, i.e., an initial guess for $\alpha \in \mathbb{R}^N$ is improved until convergence
- The update is made such that the minimal possible number of entries of α are affected. This makes very large problems easy to solve.
- To decide what entries are to be updated, the algorithm looks at entries that do not satisfy the KKT conditions.

Define

$$J(\alpha_1, \alpha_2, \ldots, \alpha_N) := \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i.$$

which has to be minimized under the constraints

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq C, \;\; 1 \leq i \leq N.$$

- We want to find the minimal number of indexes that can be changed at each iteration. Thanks to the first constraint, it can not be 1:

$$y_j \alpha_j = -\sum_{k \neq j} \alpha_i y_i$$

but with two it works:

$$y_j \alpha_j + y_i \alpha_i = -\sum_{k \neq i,j} \alpha_i y_i$$

- We can restrict the target function to only the two indexes: in this case, it can be solved analytically
- Heuristic: if at least one of the two violates the KKT, the objective is strictly decreased
- In practice: within tolerance
- The first guess is $\alpha^{(0)} \in \mathbb{R}^N$, which is feasible.

Algorithm 3 SMO

Input: $X_N, \{y_i\}_{i=1}^N, C > 0$
Set $\alpha^{(0)} := 0$, $r := 0$
while α does not satisfy KKT conditions **do**
    Choose index i that violates the KKT conditions
    Choose another index $j \neq i$
    Set $\alpha_k^{(r+1)} := \alpha_k^{(r)}$ for $k \neq i, j$
    Set $J_{ij}(\alpha_i, \alpha_j) := J(\alpha_1^{(r)}, \ldots, \alpha_{i-1}^{(r)}, \alpha_i, \alpha_{i+1}^{(r)}, \ldots, \alpha_{j-1}^{(r)}, \alpha_j, \alpha_{j+1}^{(r)} \ldots, \alpha_N^{(r)})$
    Set $R := -\sum_{k \neq i,j} y_k \alpha_k^{(r)}$
    Solve $(\alpha_i^{(r+1)}, \alpha_j^{(r+1)}) = \arg\min\{J_{ij}(\alpha_i, \alpha_j) : y_i\alpha_i + y_j\alpha_j = R, \;\; 0 \leq \alpha_i, \alpha_j \leq C\}$
end while

Solution of the problem. We assume $i, j = 1, 2$, i.e., we have to solve

$$(\alpha_1^{(r+1)}, \alpha_2^{(r+1)}) = \arg\min\{J_{12}(\alpha_1, \alpha_2) : y_1\alpha_1 + y_2\alpha_2 = R, \;\; 0 \leq \alpha_1, \alpha_2 \leq C\}$$

Reparametrize as one-dimensional problem. We assume:

$$\alpha_1(\beta) := \alpha_1^{(r)} + \beta$$
$$\alpha_1(\beta) := \alpha_2^{(r)} - y_1 y_2 \beta$$

They satisfy the constraint since $\alpha^{(r)}$ satisfies the constraints:

$$\begin{aligned} y_1\alpha_1 + y_2\alpha_2 &= y_1\alpha_1^{(r)} + y_2\alpha_2^{(r)} + y_1\beta - y_1 y_2^2 \beta = y_1\alpha_1^{(r)} + y_2\alpha_2^{(r)} + y_1\beta - y_1\beta \\ &= 1\alpha_1^{(r)} + y_2\alpha_2^{(r)} = -\sum_{k \neq 1,2} y_k \alpha_k^{(r)} = R. \end{aligned}$$

Now the objective can be rewritten as a one-dimensional quadratic function:

$$J_{12}(\alpha_1(\beta), \alpha_2(\beta)) := J(\alpha_1(\beta), \alpha_2(\beta), \alpha_3^{(r)}, \ldots, \alpha_N^{(r)}) = a\beta^2 + b\beta + c$$

Thus $d_\beta J_{12}(\alpha_1(\beta), \alpha_2(\beta)) = 2a\beta + b$, i.e.,

$$\tilde{\beta} := \arg\min_{\beta \in \mathbb{R}} J_{12}(\alpha_1(\beta), \alpha_2(\beta)) = -\frac{b}{2a}.$$

Now we have to check that $\alpha_1(\tilde{\beta}), \alpha_2(\tilde{\beta})$ satisfy the remaining constraints, i.e., $0 \leq \alpha_1(\beta), \alpha_2(\beta) \leq C$.

$$\alpha_1^{(r)} + \beta \geq 0 \Leftrightarrow \beta \geq -\alpha_1^{(r)}$$
$$\alpha_1^{(r)} + \beta \leq C \Leftrightarrow \beta \leq C - \alpha_1^{(r)}$$
$$\alpha_2^{(r)} - y_1 y_2 \beta \geq 0 \Leftrightarrow \beta \begin{cases} \leq y_1 y_2 \alpha_2^{(r)} = \alpha_2^{(r)} & \text{if } y_1 = y_2 \\ \geq y_1 y_2 \alpha_2^{(r)} = -\alpha_2^{(r)} & \text{if } y_1 = -y_2 \end{cases}$$
$$\alpha_2^{(r)} - y_1 y_2 \beta \leq C \Leftrightarrow \beta \begin{cases} \geq y_1 y_2 (\alpha_2^{(r)} - C) = \alpha_2^{(r)} - C & \text{if } y_1 = y_2 \\ \leq y_1 y_2 (\alpha_2^{(r)} - C) = C - \alpha_2^{(r)} & \text{if } y_1 = -y_2 \end{cases}$$

This can be written as $\tilde{\beta} \in [\beta_{\min}, \beta_{\max}]$ [PLOT], with

$$\beta_{\min} := \begin{cases} \max\{-\alpha_1^{(r)}, \alpha_2^{(r)} - C\} & \text{if } y_1 = y_2 \\ \max\{-\alpha_1^{(r)}, -\alpha_2^{(r)}, C - \alpha_2^{(r)}\} & \text{if } y_1 = -y_2 \end{cases}$$

$$\beta_{\max} := \begin{cases} \min\{C - \alpha_1^{(r)}, \alpha_2^{(r)}\} & \text{if } y_1 = y_2 \\ \min\{C - \alpha_1^{(r)}, C - \alpha_2^{(r)}\} & \text{if } y_1 = -y_2. \end{cases}$$

So we can clip the result and define

$$\beta^* := \begin{cases} \tilde{\beta} & \text{if } \tilde{\beta} \in [\beta_{\min}, \beta_{\max}] \\ \beta_{\min} & \text{if } \tilde{\beta} \leq \beta_{\min} \\ \beta_{\max} & \text{if } \tilde{\beta} \geq \beta_{\max} \end{cases}$$

and finally

$$\alpha_1^{(r+1)} := \alpha_1(\beta^*), \quad \alpha_2^{(r+1)} := \alpha_2(\beta^*).$$

8.5 Multiclass classification

We assume now that $D := \left(X_N, \{y_i\}_{i=1}^N\right)$ with now $y_i \in \{1, \ldots, n_c\}$, $n_c \in \mathbb{N}$ is the number of classes. The idea is to obtain a multiclass classifier by combining different binary classifiers:

- One-versus-all Iterate over all the classes. For each class $j \in \{1, \ldots, n_c\}$, define a new dataset with same X_N, but

  $$y_i^{(j)} := \begin{cases} +1, & \text{if } y_i = j \\ -1, & \text{if } y_i \neq j \end{cases}$$

  Train a binary classifier $s^{(j)}(x) = \text{ sign } \left(f^{(j)}(x)\right)$ on the dataset $D^{(j)} := \left(X_N, \{y_i^{(j)}\}_{i=1}^N\right)$.

  For classification, assign $x \in \Omega$ to the class $j^* \in \{1, \ldots, n_c\}$ such that

  $$j^* := \underset{1 \leq j \leq n_c}{\arg\max}\, f^{(j)}(x)$$

  Pros: only n_c problems must be solved. Cons: each is of size N.

- One-versus-one Iterate over all the disjoint pairs of classes. For each pair $(i, j) \in \{1, \ldots, n_c\}^2$, $i < j$, define a new dataset with

  $$X^{(i,j)} := \{x_k \in X_N : y_k = i \text{ or } y_k = j\}$$

  and

  $$y_k^{(i,j)} := \begin{cases} +1, & \text{if } y_k = i \\ -1, & \text{if } y_k = j \end{cases}$$

Train a binary classifier $s^{(i,j)}(x) = \text{ sign } \left(f^{(i,j)}(x)\right)$ on the dataset $D^{(i,j)} := \left(X^{(i,j)}, \{y_k^{(i,j)}\}_{i=1}^{N(i,j)}\right)$. For classification, assign $x \in \Omega$ by majority vote, i.e., for a given point x all the classifier are evaluated, giving M predicted classes. The multiclass classifier then decides for the class that is most frequently predicted (ties can be resolved e.g. by choosing the smaller class number). Pros: each problem involves only a small number $N(i,j)$ of points, ideally $N(i,j) = \frac{2N}{n_c}$ points. Cons: $M := \frac{n_c(n_c-1)}{2}$ classifiers have to be trained. Usually less requirements in terms of memory.

9. Unsupervised learning

Mention SVR

- Unsupervised learning: we are given only some data $X_N \subset \Omega$, and no labels or numerical values/functions evaluations at X_N.
- The goal is to extract meaningful information from X_N, to do predictions for a new $x \in \Omega$.
- All the following algorithms can be formulated for $\Omega \subset \mathbb{R}^d$ and in terms of inner products. The usual kernel trick, with $K : \Omega \times \Omega \to \mathbb{R}$ a positive definite kernel, allow to transform them to nonlinear algorithms that work for arbitrary Ω (provided we can define a kernel on Ω).
- When the kernel is used, all the information we need about the data if the kernel matrix A. This simplifies implementation and data representation.
- Demos in ILIAS.

9.1 Novelty / outlier detection

- Data $X_N \subset \Omega \subset \mathbb{R}^d$
- Goal: find $s : \Omega \to \{-1, 1\}$ to decide if a new point $x \in \Omega$ is typical with respect to X_N, then $s(x) = +1$, or if it is an outlier or novel point, then $s(x) = -1$.
- Idea: find the minimal sphere that enclose the data, and declare a new point an outlier if it is not in the sphere.

Definition 9.1 (Minimal enclosing sphere, primal formulation)**.** *Let $X_N \subset \Omega \subset \mathbb{R}^d$. The minimal enclosing sphere is defined as*

$$\min_{c \in \mathbb{R}^d, r > 0} r^2 \tag{9.1}$$

$$s.t. \|x_i - c\|^2 \leq r^2, \ 1 \leq i \leq N. \tag{9.2}$$

The decision function is defined as

$$s(x) := \ sign \ \left(r^2 - \|x - c\|^2\right).$$

We can first derive the dual problem.

The Lagrangian is

$$L(c,r,\alpha) = r^2 + \sum_{i=1}^{n} \alpha_i \left(\|x_i - c\|^2 - r^2 \right) = \left(1 - \sum_{i=1}^{N} \alpha_i \right) r^2 + \sum_{i=1}^{N} \alpha_i \left(\|x_i\|^2 - 2(x_i, c) + \|c\|^2 \right)$$

thus

$$\partial_r L(c,r,\alpha) = 2r \left(1 - \sum_{i=1}^{N} \alpha_i \right)$$
$$\partial_c L(c,r,\alpha) = -2 \sum_{i=1}^{N} \alpha_i x_i + 2 \sum_{i=1}^{N} \alpha_i c$$

so we can take

$$\sum_{i=1}^{N} \alpha_i = 1$$
$$-2 \sum_{i=1}^{N} \alpha_i x_i + 2 \sum_{i=1}^{N} \alpha_i c = -2 \sum_{i=1}^{N} \alpha_i x_i + 2c \Rightarrow c = \sum_{i=1}^{N} \alpha_i x_i.$$

(i.e., the center is a linear combination of the data). Substituting into L: using $\sum_{i=1}^{N} \alpha_i = 1$ and $c = \sum_{i=1}^{N} \alpha_i x_i$

$$\begin{aligned} L(c,r,\alpha) &= \left(1 - \sum_{i=1}^{N} \alpha_i \right) r^2 + \sum_{i=1}^{N} \alpha_i \left(\|x_i\|^2 - 2(x_i, c) + \|c\|^2 \right) \\ &= \sum_{i=1}^{N} \alpha_i \|x_i\|^2 - 2 \sum_{i=1}^{N} \alpha_i (x_i, c) + \sum_{i=1}^{N} \alpha_i \|c\|^2 \\ &= \sum_{i=1}^{N} \alpha_i \|x_i\|^2 - 2 \left(\sum_{i=1}^{N} \alpha_i x_i, c \right) + \|c\|^2 \sum_{i=1}^{N} \alpha_i \\ &= \sum_{i=1}^{N} \alpha_i \|x_i\|^2 - \|c\|^2 \\ &= \sum_{i=1}^{N} \alpha_i \|x_i\|^2 - \sum_{i,j=1}^{N} \alpha_i \alpha_j (x_i, x_j). \end{aligned}$$

From KKT:

$$\alpha_i \left(\|x_i - c\|^2 - r^2 \right) = 0, 1 \leq i \leq N$$

thus $\|x_i - c\|^2 < r^2$ implies $\alpha_i = 0$. In this sense, only the points on the boundary can be "support vectors".

For $\alpha_i \neq 0$, define $r := \|x_i - c\|$.

Definition 9.2 (Minimal enclosing sphere, dual formulation)**.** *Let $X_N \subset \Omega$. Le $K : \Omega \times \Omega \to \mathbb{R}$ be a PD kernel on Ω. The minimal enclosing sphere in dual form is defined as*

$$\max_{\alpha \in \mathbb{R}^N} \sum_{i=1}^{N} \alpha_i (x_i, x_i) - \sum_{i,j=1}^{N} \alpha_i \alpha_j (x_i, x_j) \tag{9.3}$$

$$s.t. \sum_{i=1}^{N} \alpha_i = 0, \tag{9.4}$$

$$\alpha_i \geq 0 \quad 1 \leq i \leq N. \tag{9.5}$$

The decision function is defined as

$$f(x) := r^2 - \|x - c\|^2 = r^2 - \|x\|^2 - \sum_{i,j=1}^{N} \alpha_i \alpha_j (x_i, x_j) + 2 \sum_{i=1}^{N} \alpha_i (x, x_i).$$

and

$$s(x) := \textit{sign } (f(x))$$

with $r > 0$ such that for x_i with $\alpha_i \neq 0$ it holds $f(x_i) = 0$.

The nonlinear version is immediate.

Definition 9.3 (Nonlinear Minimal enclosing sphere, dual formulation)**.** *Let $X_N \subset \Omega$. Le $K : \Omega \times \Omega \to \mathbb{R}$ be a PD kernel on Ω. The minimal enclosing sphere in dual form is defined as*

$$\max_{\alpha \in \mathbb{R}^N} \sum_{i=1}^{N} \alpha_i K(x_i, x_i) - \sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) \tag{9.6}$$

$$s.t. \sum_{i=1}^{N} \alpha_i = 0, \tag{9.7}$$

$$\alpha_i \geq 0 \quad 1 \leq i \leq N. \tag{9.8}$$

The decision function is defined as

$$f(x) := r^2 - K(x, x) - \sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) + 2 \sum_{i=1}^{N} \alpha_i K(x, x_i).$$

and

$$s(x) := \textit{sign } (f(x))$$

with $r > 0$ such that for x_i with $\alpha_i \neq 0$ it holds $f(x_i) = 0$.

It is also possible to derive a soft-margin version similar of the SVM case.

9.2 Feature extraction/Principal component analysis (PCA)

- Data $X_N \subset \Omega \subset \mathbb{R}^d$
- Goal: find characteristic feature of the data.
- Idea: find vectors $\{v_i\}_{i=1}^n$, define the characterizing features as $f_i(x) := (x, v_i)$. This is equivalent to find a change of coordinates, where the first coordinates are the most relevant for the given dataset.
- This works as data compression if $n \leq d$.

A linear version can be defined as follows: we look for a vector v_1 that maximizes the alignment with the data, i.e.,

$$v_1 := \arg\max_{v \in \mathbb{R}^d} \sum_{i=1}^{N} \frac{(x_i, v)}{\|v\|^2}$$

at step $k > 1$, we define

$$v_k := \arg\max_{\substack{v \in \mathbb{R}^d \\ (v,v_j)=0, 1 \leq j < k}} \sum_{i=1}^{N} \frac{(x_i, v)}{\|v\|^2}$$

If we define the matrix $X \in \mathbb{R}^{N \times d}$ with $X := \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix}$, this can be reformulated as

$$v_k := \arg\max_{\substack{v \in \mathbb{R}^d \\ (v,v_j)=0, 1 \leq j < k}} \frac{\|Xv\|^2}{\|v\|^2} = \arg\max_{\substack{v \in \mathbb{R}^d \\ (v,v_j)=0, 1 \leq j < k}} \frac{v^T X^T X v}{v^T v},$$

i.e., the directions v_k are the eigenvalues of the matrix $X^T X \in \mathbb{R}^{d \times d}$ (or the singular values of X). Since $X^T X$ is symmetric and positive semidefinite, there exists $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ such that $X^T X v_i = \lambda_i v_i$. The eigenvalues can be used to measure how relevant a feature is, i.e., one can consider only $n \leq N$ with $\lambda_n > \tau$.

We need to find a kernelizable version, i.e., to find $\{\alpha_{ij} : 1 \leq i \leq d, 1 \leq j \leq N\}$ such that

$$v_i = \sum_{j=1}^{N} \alpha_{ij} x_j = X^T \alpha_i.$$

if $\alpha_{ij} = (\alpha_i)_j$. In this way the feature functions can be computed as $f_i(x) = (v_i, x) = \sum_{j=1}^{N} \alpha_{ij}(x, x_i)$.

This can be solved as follows: let $\alpha_i \in \mathbb{R}^N$ be an eigenvector of the matrix $XX^T \in \mathbb{R}^{N \times N}$, i.e., $XX^T \alpha_i = \lambda_i \alpha_i$. Then we have

$$X^T X v_i = X^T X X^T \alpha_i = X^T \left(XX^T \alpha_i\right) = X^T (\lambda_i \alpha_i) = \lambda_i X^T \alpha_i = \lambda_i v_i.$$

Definition 9.4 (Kernel PCA). *Let $X_N \subset \Omega$. Le $K : \Omega \times \Omega \to \mathbb{R}$ be a PD kernel on Ω. Let $n \leq N$.*

Let $\alpha_i := [\alpha_{ij}]_{j=1}^N$ be the eigenvectors of the kernel matrix A, sorted accordingly to the eigenvectors $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$. The the features of the kernel PCA are defined as

$$f_i(x) := \sum_{j=1}^{N} \alpha_{ij} K(x, x_j).$$

Remark 9.5.
- *The linear PCA can provide only d linear features, while the kernel one up to N nonlinear ones.*
- *Using a SPD kernel one can guarantee that all the N provide $f_i(x) \neq 0$ for all $x \in \Omega$, since all the eigenvalues are non zero.*
- *The solution can be very difficult, since one needs to compute an eigendecomposition of the full and dense kernel matrix. Approximate solvers can be used if only a few features are needed.*

9.3 Clustering

- Data $X_N \subset \Omega \subset \mathbb{R}^d$
- Goal: group the data into $k \in \mathbb{N}$, $k \leq N$ disjoint sets of "similar" data (clusters).
- Idea: find a set of points (centers) $\{\mu_j\}_{j=1}^k \subset \Omega$ and assign each point to the closest one.
- The problem is hard, and finding a global optimum is in general infeasible. Instead, there is an iterative algorithm that works well.

We denote the i-th cluster as C_i, i.e.,

$$C_i := \{x \in X_N : \|x - \mu_i\|^2 \leq \|x - \mu_j\|, j \neq i\}.$$

Ideally, we would like to find $\{\mu_j\}_{j=1}^k \subset \Omega$ such that

$$\min_{\{\mu_j\}_{j=1}^k \subset \Omega} \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x - \mu_i\|^2$$

Decision function:

$$s(x) := \arg\min_{1 \leq j \leq N} \|x - \mu_i\|^2$$

Definition 9.6 (k-means clustering). *Algorithm here.*

Remark 9.7.

- *The choice of the initial centers can be made in other ways, and it strongly affects the results.*
- *The number k of clusters can be unclear in general. A common way to find a good one is to run the algorithm for various k and define*

$$J(k) := \frac{1}{k} \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x - \mu_j\|.$$

  This quantity is decreasing in k, and one can find a suitable k when $J(k)$ slows down.
- *problems with empty clusters.*

Bibliography

[1] G. E. Fasshauer. Meshfree Approximation Methods with MATLAB, volume 6 of Interdisciplinary Mathematical Sciences. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2007. With 1 CD-ROM (Windows, Macintosh and UNIX).

[2] G. H. Golub and C. F. Van Loan. Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[3] B. Schölkopf and A. J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, 2002.

[4] I. Steinwart and A. Christmann. Support Vector Machines. Information Science and Statistics. Springer, New York, 2008.

[5] H. Wendland. Scattered Data Approximation, volume 17 of Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2005.