Methodological approaches in Machine Learning: Kernel Methods

A. N. Yannacopoulos

October 15, 2025

Contents

1 Basic idea

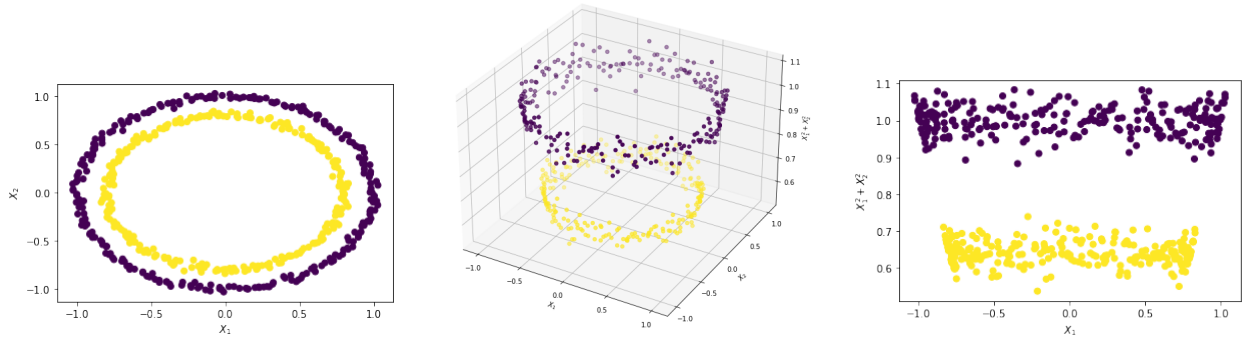
A method for going from the linear representation of data to the nonlinear representation of data is that of kernel methods. Assume that the original representation of the data are in terms of features $x = (\chi_1, \cdots, \chi_n) \in \mathbb{R}^n$, that are assumed to represent each datum as a point in the n-dimensional Euclidean space. Often this representation is not a good representation because the geometry of the data is not well represented in the Euclidean space.

Example 1.1 (The concentric circles data set)**.** As an example consider two sets of data $\mathbf{X} = \{x_1, \cdots, x_N\}$ and $\mathbf{X}' = \{x'_1, \cdots, x'_N\}$, generated by the models $x_i = (r_i \cos(2\pi\theta), r_i \sin(2\pi))$ and $X'_i = (r'_i \cos(2\pi\theta'), r'_i \sin(2\pi\theta'))$ respectively, with $r_i \sim N(10, 1)$, $r'_i \sim N(20, 1)$ and $\theta_i, \theta'_i \sim U[0, 1]$. These two data sets are situated on two concentric circles (fuzzy circles) and are clearly better represented in a nonlinear fashion and not in Euclidean space. For example, these two data sets are impossible to separate linearly in terms of a straight line in $\mathbb{R}^2$, eventhough they clearly represent data sets of different properties. Similar cases occur often in realistic biological data sets (see Fig. 1(a)).

Example 1.2 (The XOR classification problem)**.** Consider the XOR classification problem in which you have 4 points in $\mathbb{R}^2$,

$$\begin{aligned} B_1 = (-1, 1)) \quad B_2 = (1, -1), \\ R_1 = (-1, -1) \quad R_2 = (1, 1) \end{aligned} \tag{1}$$

These points are characterized as belonging into two distinct classes, B signifying blue and R signifying red, respectively. Clearly, these four points cannot be separated by a linear separation



(a) Original feature space (b) Transformed feature space (c) Transformed feature space

Figure 1: Concentric data set in different coordinates

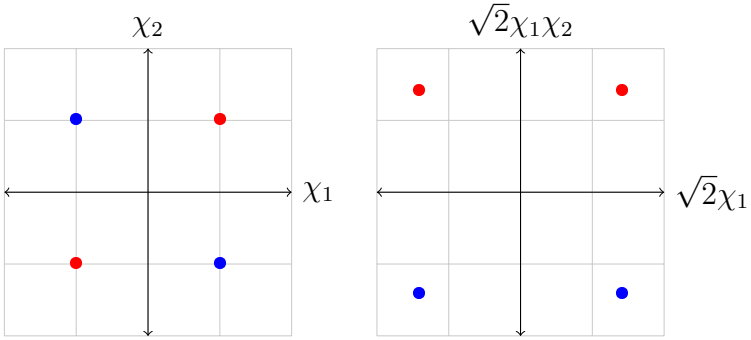
rule. There is no straight line in $\mathbb{R}^2$ that may separate the red from the blue points (see Fig. 2(a)).

Such situations be may nicely resolved by appropriate nonlinear mapping $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ to a new set of coordinates $z = (\zeta_1, \cdots, \zeta_m)$ with $m > n$. The mapping Φ can be written in terms of coordinates as $\Phi = (\Phi_1, \cdots \Phi_m)$, with $Z_i = \Phi(x) = \Phi_i(\chi_1, \cdots, \chi_n)$, $i = 1, \cdots, m$. By mapping our original data set $X = \{x_1, \cdots, x_N\} \subset \mathbb{R}^n$ to the new data set $Z = \{z_1, \cdots, z_N\} \subset \mathbb{R}^m$, in terms of the nonlinear coordinate transform Φ, by $z_i = \Phi(x_i)$, $i = 1, \cdots, N$, we obtain a new data set Z in the higher dimensional space $\mathbb{R}^m$. We will then treat this new data set instead of the original data set X, hoping that in the new higher dimensional space the data will be better represented and questions such as for example classification in terms of linear models will be more easily treated in the new representation.

Example 1.3 (Example 1.1 revisited)**.** As an illustrating example consider the data set consisting of the two concentric circles introduced above in Example 1.1, and consider the transformation from $x = (\chi_1, \chi_2)$ to $z = (\zeta_1, \zeta_2, \zeta_3) = (\sqrt{2}\chi_1\chi_2, \chi_1^2, \chi_2^2)$. The data in the new coordinates are easily seen to be separated by a 2 dimensional plane (see Fig. 1).

Example 1.4 (Example 1.2 revisited)**.** The data in the XOR Example 1.2 can be separated using a nonlinear trasformation of the original data from $\mathbb{R}^2$ to higher dimensions e.g. in $\mathbb{R}^3$. Consider the transformation from $x = (\chi_1, \chi_2)$ to $z = (\zeta_1, \zeta_2, \zeta_3) = (\sqrt{2}\chi_1\chi_2, \chi_1, \chi_2)$. The data in the new coordinates are easily seen to be separated by a 2 dimensional plane.

This high dimensional transformation by which we achieve separation is by no means unique. For example the nonlinear transformation from $x = (\chi_1, \chi_2)$ to $z = (\zeta_1, \zeta_2, \zeta_3, \zeta_4, \zeta_5, \zeta_6) = (\chi_1^2, \chi_2^2, \sqrt{2}\chi_1\chi_2, \sqrt{2}c\chi_1, \sqrt{2}c\chi_2, c)$ for some constant c (e.g., $c = 1$) will also work just fine for linear separation. Note, that for this particular choice, setting $z = \Phi(x)$ we have that $\Phi(x)\Phi(x')^T = (\chi_1\chi_1' + \chi_2\chi_2' + c)^2$. This observation is important for what follows, and is crucial to the computational issues related to determining the nonlinear transformation from the original variables x to the new variables z, in which the separation is feasible (see Fig. 2).



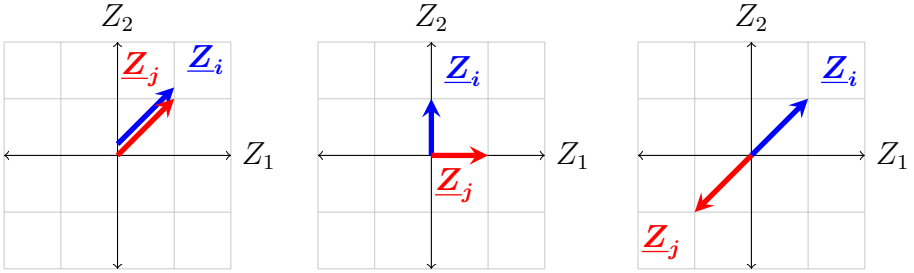
(a) Original feature space (b) Transformed feature space

Figure 2: XOR data and their nonlinear transformation

Recall that our main objective is to establish connectivities between the different data points, or better yet **measures of affinity** between the various data points. Moreover, it is our secret hope that upon trasnforming the data to the new representation these measurees of affinity may be better perceived, i.e., revealed more clearly than in the original representation.

Choosing measures of affinity may be an art and a science on its own. Different types of data call for different measures of affinity, hence the choice depends on the nature of the data or put more mathematically, on the nature of the set the data reside in. For example, different types of measures of affinity can be chosen if our data reside in a space supporting a linear structure (i.e. a vector space) whereas the same measures of affinity may be nonsensical if our data are such that they reside in a general metric space where linear structures cannot be supported (as for example data on a manifold, data for covariance matrices, data for probability measures, data for shapes, etc).

Sticking, for the time being, to the case that the data reside in a vector space X, we recall that the data can in general be understood as points in the vector space X. Assuming moreover, the the transformation Φ sends the data from the original feature vector space X to a new space Z which is again a vector space, the transformation Φ may be perceived as sending points of X to points of Z. Being elements of these vector spaces, each point can be perceived as a position vector, stemming from an origin 0, corresponding to the zero vector, to the corresponding point. So, working e.g., in the new representation $z = \Phi(x) \in Z$, for any datum i, we may understand it as a point $z_i \in Z$, determined in terms of the position vector z_i (we use the same notation for both for simplicity). Consider now to different datum points z_i, z_j, using the above interpretation. Our simple geometric intuition, entails that if z_i and z_j are very close together then these two vectors will be almost colinear, i.e. there exists $\lambda \geq 0$ such that $z_i \simeq \lambda z_j$ (clearly $\lambda = +1$ makes the most sense, but allowing for more relaxed sense of similarity $\lambda \geq 0$ is also possible). If z_i and z_j are uncorrelated, then we would expect that these two vectors are orthogonal to each other. Finally, if z_i and z_j are opposite to each other, i.e., negatively correlated, then we expect the existence of $\lambda < 0$ such that $z_i \simeq \lambda z_j$. All the three considerations above, can be interpreted through the "angle" formed between the



(a) Close "positive" affinity (b) No affinity (c) Close "negative" affinity

Figure 3: Different levels of data affinity

corresponding vectors z_i and z_j; this being an angle θ such that $\theta \simeq 0$ in the first case, $\theta \simeq \frac{\pi}{2}$ in the second case and $\theta \simeq \pi$ in the third case (see Fig. 3 for an illustrative depiction of this concept for the case of a two dimensional feature space).

A quick trip back to linear algebra reminds us that in a vector space the concept of angle between two vectors can be generalized in any dimension using the concept of the inner product between them. In particular the angle θ between two vectors x_i and x_j is defined by

$$\begin{aligned} \cos\theta &:= \frac{\langle x_i, x_j \rangle}{\|x_i\| \, \|x_j\|} \\ \|x_i\| &= (\langle x_i, x_i \rangle)^{1/2}, \\ \|x_j\| &= (\langle x_j, x_j \rangle)^{1/2}. \end{aligned} \tag{2}$$

This is a definition, motivated by the relevant definition of the dot product (which is a special case of the inner product) in $\mathbb{R}^2$ or $\mathbb{R}^3$.

Continuing along the above lines we now pose the following question:

> Having transformed the original data from the (low dimensional) representation as x to the new (higher dimensional) representation $z = \Phi(x)$, what could be a reasonable measure of affinity between the data points $z_i = \Phi(x_i)$ in the new representation?
>
> Choosing the "angle" as a measure of affinity we wish to monitor the quantify
>
> $$\langle z_i, z_j \rangle = \langle \Phi(x_i), \Phi(x_j) \rangle =: K(x_i, x_j). \tag{3}$$

Kernel methods, exploit this observation, while at the same time considering the question of how computationally demanding would be the need to perform the nonlinear coordinate change explicitly by calculating all the new coordinates $(\zeta_1, \cdots, \zeta_m)$ $(m > n)$ for each one the data points x_i (N in total). Since typically both the number of data N, as well as the dimensionality of the space onto which the original data are mapped into m are large, this consideration is very important. As can be seen, for most practical situations, we will not require the full

transformation $x_i \to z_i = \Phi(x_i)$, $i = 1, \cdots, N$, but rather only knowledge of the real numbers $K_{ij} := \langle z_i, z_j \rangle = \langle \Phi(x_i), \Phi(x_i) \rangle$ for $i, j = 1, \cdots N$. If we can devise a methodology of calculating directly the matrix containing K_{ij}, without having to pass through the intermediate step of calculating $\Phi(x_i)$.

As an important final observation we note that in principle there is no real need for the original feature space X to have the structure of a vector space (i.e. a linear structure). For example, in many important applications of interest, spaces of symbols, or words, can be considered (and in fact this is common in genomic applications). On the other hand, Z is always considered as a vector space, and in particular a complete inner product space (a Hilbert space) We can still proceed in pretty much the same fashion as long as the new space, in which the data point will reside after the transformation, Z, is a vector space, and in particular an inner product space. This allows us, after the transformation, to treat as linear data that can be inherently nonlinear! As a result of that we can apply (after a suitable transformation) linear methods for data that are initially residing in nonlinear structures. To emphasize the fact that the target space Z must be a linear space with an inner product structure[1] we will use the notation H instead of Z

Notation 1.5. In what follows we will:

- use the notation x and z for data points residing in the spaces X and Z respectively,
- use the notation x_i, x_j etc when more than two data points in X are required (resp. z_i, z_j for data points in Z),
- use the notation $\mathbf{X}$ and $\mathbf{Z}$ for data sets i.e., $\mathbf{X} = \{x_1, \cdots, x_N\} \subset X$ and $\mathbf{Z} = \{z_1, \cdots, z_N\} \subset Z$, respectively.
- replace Z with H for the target space to emphasize that this is a complete inner product space (a Hilbert space) and use the notation f, g or h for the elements of H.

2 Kernel functions

We will use the following terminology and notation that is common for kernel methods. .

- X original feature space $dim(X) = n$ (low dimensional). While in the introduction it was assumed that $X = \mathbb{R}^n$ this is no longer needed, X can be any finite dimensional space, not necessarily a vector space. The structure and dimensionality of X may possibly induce difficulties in the representation of the data.
- H new feature space $dim(H) = m$ (high dimensional - possibly infinite dimensional) — Easier (linear) representation
- $\Phi : X \to H$ feature map, transformation $x \to z = \Phi(x)$ leads to better representation.

[1] Recall that a complete inner product space is called a Hilbert space.

The important question is

What is this map Φ going to be?

Luckily, the theory works fine most of the times for some pre- prescribed mappings Φ, that are constructed in terms of stock kernel functions.

Definition 2.1. We will call a kernel function K, a function $K : X \times X \to \mathbb{R}$, satisfying the following properties:

1. K is symmetric, i.e. $K(x, x') = K(x', x)$ for any $x, x' \in X$,
2. The Gram matrix generated by K is positive semidefinite, i.e. for any set $\{x_1, \cdots, x_N\} \subset X$ and any $\xi = (\xi_1, \cdots, \xi_N) \in \mathbb{R}^N$ it holds that

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

The matrix $\mathbf{K} = (K(x_i, x_j))_{i,j=1}^{N} \in \mathbb{R}^{N\times N}$ is called the Gram matrix for the given set.

It is important to note the following:

- Given any mapping $\Phi : X \to H$, then $K : X \times X \to \mathbb{R}$, defined by $K(x, x') = \langle \Phi(x), \Phi(x') \rangle_H$ is a kernel function in the above sense.
- A kernel function $K : X \times X \to \mathbb{R}$ can be used to generate a mapping $\Phi : X \to H$ that can be used for better data representation (additional assumptions required).

We will elaborate on the above issues in what follows.

Example 2.2 (Typical examples of kernel functions)**.** We provide some typical examples of kernel functions.

1. Polynomial kernels: Polynomial kernels are kernel functions of the form

$$K(x_i, x_j) = (\langle x_i, x_j \rangle_{\mathbb{R}^n} + c)^p, \tag{4}$$

where $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$ is the inner product in $X = \mathbb{R}^n$ (possibly the dot product, but not necessarily), $c \in \mathbb{R}$ is a constant (parameter of the kernel function), and $p \in \mathbb{N}$ is a natural number, which is the degree of the polynomial.

2. Gaussian kernels: A very popular example of kernel function are Gaussian kernels

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right), \quad \sigma > 0, \tag{5}$$

where $\|\cdot\| = \|\cdot\|_{\mathbb{R}^n}$ is the Euclidean distance on $X = \mathbb{R}^n$ and $\sigma > 0$ is an arbitrary parameter. This example can be generalized in cases where X is a general vector space, other than $\mathbb{R}^n$.

3. Sigmoid kernels: These are kernels of the form

$$K(x_i, x - j) = \tanh(c_1\langle x_i, x_j \rangle_{\mathbb{R}^n} + c_2), \tag{6}$$

with the same notation as in item 1 above.

3 From feature map $\Phi : X \to H$ to kernel functions

We first consider how we can obtain kernel functions from feature maps:

Proposition 3.1. *Given a feature map $\Phi : X \to H$, we can obtain the kernel function $K : X \times X \to \mathbb{R}$ defined by*

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_H, \ \ \forall \ x, x' \in X, \tag{7}$$

where by $\langle \cdot, \cdot \rangle_H$ we denote the inner product in the Hilbert space H.

Proof. We will show that the function $K : X \times X \to \mathbb{R}$ defined as in (7) satisfies the properties in Def. 2.1.

Symmetry is immediate by the properties of the inner product.

To check positivity, take $\{x_1, \cdots, x_N\} \subset X$ and $\xi = (\xi_1, \cdots, \xi_N) \in \mathbb{R}^N$ arbitrary and construct $h = \sum_{i=1}^N \xi_i \Phi(x_i) \in H$. By the properties of the inner product

$$\begin{aligned} 0 \leq \langle h, h \rangle_H = \langle \sum_{i=1}^N \xi_i \Phi(x_i), \sum_{i=1}^N \xi_i \Phi(x_i) \rangle_H \\ = \langle \sum_{j=1}^N \xi_j \Phi(x_j), \sum_{i=1}^N \xi_i \Phi(x_i) \rangle_H = \sum_{i=1}^N \sum_{j=1}^N \langle \Phi(x_i), \Phi(x_j \rangle_H = \sum_{i=1}^N \sum_{j=1}^N K(x_i, x_j) \xi_i \xi_j, \end{aligned} \tag{8}$$

and the proof is complete. □

Of course one can construct new kernel functions from existing ones as follows:

Proposition 3.2 (New kernels from old ones)**.** *The following hold:*

(i) If K_1, K_2 are kernel functions and $\lambda_1, \lambda_2 \geq 0$, then so is $\lambda_1 K_1 + \lambda_2 K_2$.

(ii) If K_1, K_2 are kernel functions then so is $K_1 K_2$.

(iii) If $(K_n)_{n \in \mathbb{N}}$ is a sequence of kernel functions and the limit exists then $K := \lim_n K_n$ is a kernel function.

(iv) If K is a kernel function and $f : X \to \mathbb{R}$ is any function then the new function $K_f : X \times X \to \mathbb{R}$, defined by

$$K_f(x, x') = f(x) f(x') K(x, x'), \ \ \forall \ x, x' \in X,$$

is a kernel function.

Proof. (i) Symmetry is easy. For positive definiteness, set $\mathbf{K_1}, \mathbf{K_2}$ the Gram matrices for any $\{x_1, \cdots, x_N\} \subset X$. These are positive definite. The matrix $\mathbf{K} = \lambda_1 \mathbf{K}_1 + \lambda_2 \mathbf{K}_2$ is the Gram matrix for the function $K = \lambda_1 K_1 + \lambda_2 K_2$. This is clearly positive definite.

(ii) To show that we observe that if $\mathbf{K}_1$ is the Gram matrix for K_1 and $\mathbf{K}_2$ is the Gram matrix for K_2 then the Gram matrix $\mathbf{K}$ for $K = K_1 K_2$ is $\mathbf{K} = \mathbf{K_1} \circ \mathbf{K_2}$, the Hadamard product of the two matrices. From linear algebra we know that the Hadamard product of positive matrices is a positive matrix.

(iii) We will prove this by contradiction. Assume not. Hence, for some choice $\{x_1, \cdots, x_N\} \subset X$, $\xi = (\xi_1, \cdots, \xi_N) \in \mathbb{R}^N$, it holds that

$$\sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i, x_j) \xi_i \xi_j = -\epsilon, \;\; \epsilon > 0.$$

Take any $n \in \mathbb{N}$. By the positivity of K_n we have that

$$\sum_{i=1}^{N} \sum_{j=1}^{N} K_n(x_i, x_j) \xi_i \xi_j \geq 0.$$

Subtracting these two we obtain

$$\sum_{i=1}^{N} \sum_{j=1}^{N} (K_n(x_i, x_j) - K(x_i, x_j)) \xi_i \xi_j \geq \epsilon > 0,$$

and passing to the limit as $n \to \infty$ we obtain a contradiction.

(iv) By the positivity of K we have that $\sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i, x_j) \xi_i \xi_j \geq 0$ for every $\xi = (\xi_1, \cdots, \xi_N) \in \mathbb{R}^N$, hence also for $(f(x_1)\xi_1, \cdots, f(x_N)\xi_N) \in \mathbb{R}^N$. Positivity of K_f follows by this observation. □

Finally for kernels that can be expressed in terms of differences

$$K(x, x') = \phi(x - x'), \;\; \forall \; x, x' \in X, \tag{9}$$

a general characterization can be obtained in terms of the celebrated Bochner theorem.

Theorem 3.3 (Bochner)**.** *A function defined as in* (9) *with ϕ continuous, is positive definite if and only if ϕ is the characteristic function of a random variable $Y : \Omega \to X$, where $(\Omega, \mathcal{F}, \mu)$ is a probability space. That means that ϕ can be expressed in terms of the Fourier transform*

$$\phi(x) = \int \exp(ix \cdot y) d\mu(y),$$

or assuming that μ has density function g,

$$\phi(x) = \int \exp(ix \cdot y) g(y) dy.$$

If ϕ is an even function then K is a kernel function.

Using Bochner's theorem we may easily construct kernel functions. Moreover, it allows us to draw useful and intuitive connections with probability theory.

4 From kernels to feature maps

We now consider the problem of going from kernels to feature maps. This is covered by the celebrated Aronsajn theorem Aronszajn (1950).

Under additional assumptions we can see that any kernel function can be expressed in terms of such a mapping Φ, thus showing that any kernel function $K : X \times X \to \mathbb{R}$ generates a mapping $\Phi : X \to H$, for an appropriate Hilbert space H called the reproducing kernel Hilbert space (RKHS). This is guaranteed by a celebrated theorem of Aronszajn (1950) according to which for any kernel $K : X \times X$ there exists a Hilbert space H and a mapping $\Phi : X \to H$ such that $K(x, x') = \langle \Phi(x), \Phi(x') \rangle_H$ for every $x, x' \in H$. The Hilbert space H consists of all functions of the form $\sum_{i=1}^{\ell} a_i K(x_i, \cdot)$ for $i \in \mathbb{N}$, $a_i \in \mathbb{R}$, $x_i \in X$, $i = 1, \cdots, \ell$, as well as limits of such linear combinations.

We require the following definition of reproducing kernel Hilbert spaces (RKHS). For our needs a RKHS H is a set of functions f, that will be considered as possible models for our data, i.e., possible models $y = f(x)$. Clearly, being a set of functions, this set is a possibly infinite dimensional function space.

Definition 4.1 (Reproducing Kernel Hilbert Spaces (RKHS))**.** A Hilbert space H is called a reproducing Hilbert space if for any function $f : X \to \mathbb{R}$, $f \in H$, the mapping $x \mapsto f(x)$ is a continuous map (considered as a map $L : H \to \mathbb{R}$).

An RKHS is connected with a kernel function $K : X \times X \to \mathbb{R}$, which is called the reproducing kernel of H. In terms of this function we have the following two conditions (that can also be considered as an alternative definition for RKHS)

Definition 4.2 (RKHS II)**.** A Hilbert space $H = H_K$ consisting of functions $f : X \to \mathbb{R}$ is called a reproducing Hilbert space (RKHS) with reproducing kernel $K : X \times X \to \mathbb{R}$, if

1. For each $x \in X$, $K(x, \cdot) \in H$ (by $K(x, \cdot)$ we denote the mapping $x \to K(x, \cdot) =: \varphi_x(\cdot)$, such that $\varphi_x(x') = K(x, x')$ for all $x' \in X$).

2. For each $f \in H$, and $x \in X$,

$$f(x) = \langle f, K(x, \cdot) \rangle_H = \langle f, \varphi_x \rangle_H, \text{ (Reproducing property).} \tag{10}$$

What type of functions are included in a RKHS? By construction, it is rather easy to characterize the elements of this space.

Any element $f \in H$ is of the form

$$f(\cdot) = \sum_{i=1}^{N} a_i K(x_i, \cdot), \tag{11}$$

for some set $\{x_1, \cdots, x_N\} \subset X$, and some N, and $a_i \in \mathbb{R}$, or the limit as $N \to \infty$ of such a function.

This is extremely convenient, as it allows us to construct our model using expressions of the form (11) using as the set $\{x_1, \cdots, x_N\} \subset X$ our available data concerning the observed features. Of course, we should take into account an important caveat of this approach: The expansion (11) is fine as long as we have some guarantee that the models under consideration are indeed in the RKHS $H = H_K$. If not, then constructing models such as (11) may not be sufficient to capture the "real" connection between the features $x \in X$ and the response $y \in \mathbb{R}$. To answer to this point, it is important to choose kernels K and corresponding RKHS $H = H_K$ such that H_K is a function space as general as possible!

We now provide a version of Aronsjan's theorem.

Theorem 4.3. *For every kernel $K : X \times X \to \mathbb{R}$, there exists a Hilbert space $H = H_K$ (a RKHS with kernel K) and a mapping $\Phi : X \to H$ such that*

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_H, \ \ \forall \ x, x' \in X.$$

The space H_K can be characterized as the functions of the form (11) *and their limits in terms of the norm $\| \cdot \| = \| \cdot \|_H$, defined by $\|f\|_H = (\langle f, f \rangle)^{1/2}$ for every $f \in H$.*

Proof. (Sketch) We will show that given a kernel K we can construct a Hilbert space with the reproducing kernel property as above.

We will consider the space

$$\mathbb{K}_N := \{\sum_{i=1}^{N} a_i K(x_i, \cdot) \mid N \in \mathbb{N}, \ x_i \in X, \ a_i \in \mathbb{R}, \ i = 1, \cdots, N\},$$

for some $N \in \mathbb{N}$, and the space

$$\mathbb{K} = \bigcup_{N \in \mathbb{N}} \mathbb{K}_N.$$

In plain english, $\mathbb{K}$ is the function space consisting of functions of the form (11) for some $N \in \mathbb{N}$.

Clearly, $\mathbb{K}$ is a vector space, which is turned into an inner product space using the inner product

$$\langle f, g \rangle_{\mathbb{K}} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} a_i b_j K(x_i, z_j), \ \ \forall \ f, g \in \mathbb{K}, \text{ such that}$$

$$f(\cdot) = \sum_{i=1}^{N_1} a_i K(x_i, \cdot), \ \ \{x_1, \cdots, x_{N_1}\} \subset X,$$

$$g(\cdot) = \sum_{i=1}^{N_2} b_i K(z_j, \cdot), \ \ \{z_1, \cdots, z_{N_2}\} \subset X.$$

Note that (writing the above double sums as iterated sums)

$$\langle f, g \rangle_{\mathbb{K}} = \sum_{i=1}^{N_1} a_i g(x_i) = \sum_{j=1}^{N_2} b_j f(z_j).$$

It is easy to check that $\langle \cdot, \cdot \rangle_{\mathbb{K}}$ is indeed an inner product. Linearity and symmetry are immediate. Positivity follows from the positivity of K, since for any $f \in \mathbb{K}$,

$$\langle f, f \rangle_{\mathbb{K}} = \sum_{i=1}^{N_1} \sum_{j=1}^{N_1} K(x_i, x_j) a_i a_j = a^T \mathbf{K} a \geq 0.$$

It is important to note that by the very definition of this inner product,

$$\langle f, K(x, \cdot) \rangle_{\mathbb{K}} = \sum_{i=1}^{N_1} a_i K(x_i, x) = f(x), \tag{12}$$

which is the reproducing property.

We emphasize that, even though $\mathbb{K}$ already has many of the desired properties, it is still not the required Hilbert space. This is because $\mathbb{K}$ is not complete i.e., it may be an inner product space but it is not yet a Hilbert space! The non completeness of $\mathbb{K}$ stems from the fact that it is closed under finite linear combinations but not when limits of sequences of elements $\mathbb{K}$ are considered (with respect to the norm $\| \cdot \|_{\mathbb{K}} = (\langle \cdot, \cdot \rangle_{\mathbb{K}})^{1/2}$ induced by the inner product $\langle \cdot, \cdot \rangle_{\mathbb{K}}$).

To go from $\mathbb{K}$ to the corresponding reproducing kernel Hilbert space $H = H_K$ we work as follows: We construct H_K as the closure of $\mathbb{K}$ in the norm $\| \cdot \|_{\mathbb{K}}$, i.e., as the elements of $\mathbb{K}$, including any limit of sums of the form $f_N(\cdot) = \sum_{i=1}^{N} a_i K(x_i, \cdot)$ as $N \to \infty$, in the norm $\| \cdot \|_{\mathbb{K}}$.

We consider the inner product space $(\mathbb{K}, \| \cdot \|_{\mathbb{K}})$, with $\|f\|_{\mathbb{K}} = \sqrt{\langle f, f \rangle_{\mathbb{K}}}$. Moreover, consider $x \in X$ fixed, and the sequence $(f_n)_{n \in \mathbb{N}} \subset \mathbb{K}$ which is Cauchy with respect to the norm $\| \cdot \|_{\mathbb{K}}$. Note that f_n and f_m belong to some K_N for which by (12) the reproducing property holds. Hence,

$$(f_n(x) - f_m(x))^2 = \langle f_n - f_m, K(x, \cdot) \rangle_K^2 \leq \|K(x, \cdot)\|_{\mathbb{K}}^2 \, \|f_n - f_m\|_{\mathbb{K}}^2 = K(x, x)^2 \|f_n - f_m\|_{\mathbb{K}}^2, \tag{13}$$

where we also used the Cauchy-Schwarz inequality and the fact that $K(x, \cdot) \in H_K$. Since (by assumption) $\|f_n - f_m\|_{\mathbb{K}} \to 0$ as $n, m \to \infty$, we obtain from (13) that $|f_n(x) - f_m(x)| \to 0$ as $n, m \to \infty$, for any $x \in X$, hence $(f_n(x))_{n \in \mathbb{N}} \subset \mathbb{R}$ is Cauchy and by the completeness of $\mathbb{R}$, it has a limit. Let us define $f(x) := \lim_{n \to \infty} f_n(x)$, pointwise (i.e. for every $x \in X$). This means that the closure of $\mathbb{K}$ in the norm $\| \cdot \|_{\mathbb{K}}$, includes functions f such as above.

We then define H_K as the closure of $\mathbb{K}$ in the norm $\| \cdot \|_{\mathbb{K}}$. This can be turned into a Hilbert space using the inner product $\langle \cdot, \cdot \rangle_{H_K}$ defined as

$$\begin{aligned} \langle f, g \rangle_{H_K} := \lim_{n \to \infty} \langle f_n, g_n \rangle_{\mathbb{K}}, \;\; \forall \; f, g \in H_K, \\ f = \lim_{n \to \infty} f_n, \;\; (f_n)_{n \in \mathbb{N}} \subset \mathbb{K}, \\ g = \lim_{n \to \infty} g_n, \;\; (g_n)_{n \in \mathbb{N}} \subset \mathbb{K}. \end{aligned}$$

Note that by definition, since $f \in H_K$ it has to be the limit of a sequence $(f_n)_{n \in \mathbb{N}} \subset \mathbb{K}$. It is easy to show that $\langle \cdot, \cdot \rangle_{H_K}$ is an inner product in H_K, by noting that $\langle \cdot, \cdot \rangle_{\mathbb{K}}$ is an inner product in $\mathbb{K}$ and passing to the limit. Moreover, by the very definition of H_K, it is evident that $\mathbb{K}$ is

dense in H_K (meaning that any element $f \in H_K$ can be approximated as the limit of a sequence $(f_n)_{n\in\mathbb{N}} \subset \mathbb{K}$ (now in the norm $\|\cdot\|_{H_K}$).

What remains to show that $(H_K, \|\cdot\|_{H_K})$ is a Hilbert space, is to show that it is complete with respect to this norm. In other words, we must show that any Cauchy sequence $(f_n)_{n\in\mathbb{N}} \subset H_K$ has a limit in H_K. This follows by density arguments.

□

The reproducing property allows us to construct the mapping $\Phi : X \to H$ from the knowledge of the kernel. Indeed, consider the mapping Φ defined for any $x \in X$ by $\Phi(x) := K(x, \cdot) \in H$,

$$\Phi : X \to H = H_K, \;\; X \ni x \mapsto K(x, \cdot) =: \Phi(x) \in H = H_K.$$

Then applying the reproducing property (10) to the function $f(\cdot) = K(x', \cdot)$ we have that

$$f(x) \overset{(10)}{=} \langle f, K(x, \dot{)}\rangle_H = \langle K(x', \cdot), K(x, \dot{)}\rangle_H = K(x', x),$$

where for the last equality we used once more the reproducing property (10), but now for the element $K(x', \cdot) \in H = H_K$. The above, using the symmetry of the inner product and of the kernel function implies that

$$K(x, x') = \langle K(x, \cdot), K(x', \cdot)\rangle_H = \langle \Phi(x), \Phi(x')\rangle_H,$$

where in the last equality we used the definition of Φ.

The above argument shows that given a kernel function $K : X \times X \to \mathbb{R}$, and the corresponding RKHS $H = H_K$ defined as above, we may "read off" the corresponding nonlinear feature transformation $\Phi : X \to H = H_K$ in terms of the map $x \mapsto \Phi(x) := K(x, \cdot)$ where $K(x, \cdot)$ is understood as a function being an element of the function space $H = H_K$.

The kernel trick: Consider the transformation $X \ni x \to z \in H$, where X is the original feature space and H is the new feature space where we hope that the data are better and more easily represented. We will work with the new transformed data $z = \Phi(x)$, and for most data science applications we will require to calculate $\langle z, z'\rangle_H$, which can be considered as a measure of " affinity " of the transformed data $z = \Phi(x)$ and $z' = \Phi(x')$. On account of the high dimensionality of the new feature space, this calculation (needed for the calculation of the Gram matrix in the new feature space) is very expensive. However, if the tranformation Φ is related to a RKHS, i.e. $H = H_K$, then,

$$\langle z, z'\rangle_{H_K} = \langle \Phi(x), \Phi(x')\rangle_{H_K} = K(x, x'),$$

so the inner products in H can be calculated directly in terms of the calculation of the kernel function K on the data points x, x' in the original (low dimensional!) feature space X. This is clearly a lot cheaper than performing the inner product $\langle z, z'\rangle_{H_K}$ directly. This is one of the

reasons why out of all possible data transformations we favour those that are compatible with a RKHS structure! For such cases the Gram matrix of the transformed data (in $H = H_K$),

$$\{x_1, \cdots, x_N\} \mapsto \{z_1, \cdots z_N\} = \{\Phi(x_1), \cdots, \Phi(x_N)\},$$

is expressed in terms of the Gram matrix

$$\mathbf{K} = (K_{ij})_{i,j=1}^N, \quad K_{ij} = K(x_i, x_j).$$

This matrix (which can be very easily calculated in terms of the original data) is essentially all we require for most of our kernel algorithms calculations. Most of the time we will not ever need to look at the data in the new space! (eventhough the calculations and the methodology is made possible because we work in this new space).

We present some examples of popular kernel functions as well as the mappings Φ they generate.

1. The linear kernel $K(x, x') = x^T x' = x \cdot x' = \langle x, x \rangle_{\mathbb{R}^n}$, corresponding to the feature mapping $\Phi(x) = z = x$ (the identity map). This can be seen in two ways. (a) By directly expressing the kernel function K in terms of $K(x, x') = \langle \Phi(x), \Phi(x') \rangle = \langle x, x' \rangle = x^T x'$, where $\Phi(x) = x$ for every $x \in X$. (b) By considering the functions $K(x, \cdots)$ defined by $K(x, x') = x^T x'$ for any $x' \in X$. To fully specify this function $K(x, \cdot)$ for any $x \in X$ we need to specify only x (since then the function is created by the inner product $K(x, \cdot) = \langle x, \cdot \rangle$ so we can identify this function with x. In this sense $\Phi(x) = K(x, \cdot) \simeq x$, i.e. Φ can be identified with the identity mapping.

2. The polynomial kernels $K(x, x') = (c_1 + c_2 x^T x')^d = (c_1 + c_2 \langle x, x' \rangle)^d$ for $c_1 \geq 0$, $c_2, d \in \mathbb{N}$ corresponding to a higher dimensional polynomial map. For example in the case where $c_1, c_2 = 1$ and $d = 2$ for $n = 2$ we obtain the mapping $x = (x_1, x_2) \mapsto z = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2) \in \mathbb{R}^6$. This can be seen again in two ways. (a) By directly noting that

$$K(x, x') = (1 + x^T x')^2 = 1 + 2x^T x' + (x^T x')^2 = \langle \Phi(x), \Phi(x') \rangle_{\mathbb{R}^6}$$

for $\Phi(x) = \Phi(\chi_1, \chi_2) = (1, \sqrt{2}\chi_1, \sqrt{2}\chi_2, , \chi_1^2, \chi_2^2, \sqrt{2}\chi_1\chi_2) \in \mathbb{R}^6$, so that Φ can be considered as a mapping $\Phi : \mathbb{R}^2 \to \mathbb{R}^6$. (b) By considering for any $x \in X = \mathbb{R}^2$ the functions $K(x, \cdot)$ defined by

$$x' \mapsto K(x, x') = (1 + x^T x')^2 = 1 + 2x^T x' + (x^T x')^2$$
$$= 1 + 2\chi_1\chi_1' + 2\chi_2\chi_2' + \chi_1^2(\chi_1')^2 + \chi_2^2(\chi_2')^2 + 2\chi_1\chi_2\chi_1'\chi_2',$$

which is second order polynomial in $x' = (\chi_1', \chi_2')$. This polynomial can be uniquely determined by 6 coefficients depending on $x = (\chi_1, \chi_2)$ and in particular by the vector $z = (1, \sqrt{2}\chi_1, \sqrt{2}\chi_2, , \chi_1^2, \chi_2^2, \sqrt{2}\chi_1\chi_2) \in \mathbb{R}^6$. Since the function $K(x, \cdot)$ is uniquely determined by the vector z as defined above, we can identify $\Phi(x) = K(x, \cdot) \simeq z$.

3. The Gaussian kernel $K(x,x') = \exp(-\gamma\|x-x'\|^2)$, for $\gamma > 0$. The fact that this kernel is expressed in terms of inner products in the space X can be seen by expressing $\|x-x'\|^2 = x^Tx' - 2x^Tx' + (x')^Tx'$. This kernel produces a map to an infinite dimensional feature space H.

For all the above cases the Gram matrix, $K = (K(x_i,x_j))$, $i,j = 1,\cdots,N$ for a data set $\{x_1,\cdots,x_N\} \subset X$ can be directly calculated in terms of the inner product in the original space X which is easier to handle.

5 Mercer's theorem

For certain types of kernel functions, the construction of the corresponding RKHS can be made explicit in terms of an eigenvalue problem. This construction is based on an important result in functional analysis, known as Mercer's theorem.

Theorem 5.1 (Mercer)**.** *Let $(X,\mathcal{F},\mu)$ be a measure space, with X compact, set $K : X \to X \to \mathbb{R}$ be a continuous positive definite kernel function such that*

$$\int_X \int_X |K(x,x')|^2 d\mu(x) d\mu(x') < \infty,$$

and consider the corresponding integral operator $T_K : L^2(X,\mathcal{F},\mu) \to L^2(X,\mathcal{F},\mu)$ defined by

$$f \mapsto T_K f =: g, \;\; g(x) := \int_X K(x,x') f(x') d\mu(x').$$

(i) The operator T_K is a compact operator that admits a denumerable set $\{(\lambda_j,\phi_j),\ j \in \mathbb{N}\}$ of eigenvalues and eigenfunctions, i.e. solutions to the problem

$$T_K\phi_j = \lambda_j\phi_j, \;\; j \in \mathbb{N},$$

with $\lambda_j \geq 0$. Using a standard Gram-Schmidt orthogonalization procedure, we can turn this set into an orthonormal set (assume from now on).

(ii) The sequence can be turned into a basis of the corresponding Hilbert space and in particular we can use the expansion

$$K(x,x') = \sum_{j=1}^{\infty} \lambda_j \phi_j(x)\phi_j(x'), \;\; \forall\ x,x' \in X,$$

wth the above sum being uniformly convergent.

(iii) In this case we may construct the corresponding RKHS $H = H_K$ explicitly in terms of

$$H_K := \{\sum_{j=1}^{\infty} c_j\phi_j, \;:\; z_j \in \mathbb{R}, \text{ such that } \sum_{j=1}^{\infty} \frac{|c_j|^2}{\lambda_j} < \infty\},$$

$$\|z\|^2_{H_K} = \sum_{j=1}^{\infty} \frac{|c_j|^2}{\lambda_j}, \;\; \forall\ z = \sum_{j=1}^{\infty} c_j\phi_j \in H_K,$$

where $z \in H_K$ is to be understood as a function $z \equiv f : X \to \mathbb{R}$. In terms of this notation (and assuming that we have performed the orthonalization step) the inner product can be expressed as

$$\langle f_1, f_2 \rangle_{H_K} = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} \langle f_1, \phi_j \rangle \langle f_2, \phi_j \rangle,$$

$$\langle f_i, \phi_j \rangle = \int_X f_i(x) \phi_j(x) d\mu(x), \;\; i = 1, 2.$$

6 The representer theorem

One of the major reasons why we use RKHS in learning problems is the celebrated representer theorem, which allows us ro obtain a convenient and easy representation of learning problems in this setting.

Theorem 6.1 (Representer theorem)**.** *Consider the set of data $\{(x_i, y_i),\ i = 1, \cdots, N\} \subset X \times \mathbb{R}$, a kernel function $K : X \times X \to \mathbb{R}$ and the corresponding RKHS $H = H_K$. Consider also the loss function*

$$\mathcal{E}(f) := \frac{1}{N} \sum_{i=1}^{N} \mathcal{E}_i(y_i, f(x_i)) + R(\|f\|_{H_K}), \tag{14}$$

where $R(\|f\|_{H_K})$ is a regularization term, and $R : [0, \infty) \to \mathbb{R}$ is a strictly increasing function, and $\mathcal{E}_i$ is an arbitrary loss function.

Then, the minimizer of $\mathcal{E}(f)$ in H_K admits a representation of the form

$$f^*(\cdot) = \sum_{i=1}^{N} a_i K(x_i, \cdot), \;\; a_i \in \mathbb{R},$$

with $a_i \in \mathbb{R}$, to be obtained numerically in terms of an optimization problem in $\mathbb{R}^N$.

Proof. (Sketch) We consider the orthogonal decomposition of H_K is terms of the finite dimensional space $H_{K,N} = span\{K(x_i, \cdot),\ i = 1, \cdots, N\}$ and its orthogonal complement $H_{K,N}^{\perp} = \{f \in H_K \;:\; \langle f, g \rangle = 0,\ \forall\, g \in H_{K,N}\}$. Note that $H_{K,N}$ is finite dimensional and that any $f \in H_K$ can be expressed as $f = f_{K,N} + f^{\perp}$ where $f_{K,N} = P_{K,N} f \in H_{K,N}$ (the orthogonal projecion of f in $H_{K,N}$) and $f^{\perp} \in H_{K.N}^{\perp}$. Clearly, $f_{K,N}^{\perp}$ can be interpreted as the error we get by approximating f by $f_{K,N}$.

We note that

$$\|f\|_{H_K}^2 = \|f_{K,N}\|_{H_K}^2 + \|f_{K,N}^{\perp}\|_{H_K}^2 \geq \|f_{K,N}\|_{H_K}^2 \quad \Longrightarrow \quad \|f_{K,N}\|_{H_K} \leq \|f\|_{H_K} \tag{15}$$

and since R is an increasing function, we obtain

$$R(\|f_{K,N}\|_{H_K}) \leq R(\|f\|_{H_K}), \tag{16}$$

We now consider the first part of the loss function. This consists of a sum of $\mathcal{E}_i(y_i, f(x_i))$. We recall the reproducing property, according to which

$$f(x_i) = \langle f, K(x_i, \cdot)\rangle = f_{K,N}(x_i), \;\; \forall\, f \in H_K, \tag{17}$$

since $\langle f, K(x_i, \cdot)\rangle$ is in fact the projector of f onto $H_{K,N}$.

Putting the above in the loss function we see that

$$\mathcal{E}(f_{K,N}) \leq \mathcal{E}(f), \;\; \forall\, f \in H_K. \tag{18}$$

Therefore, $f_{K,N}$is a minimizer of the loss function (14).

If moreover, the regularization function R is strictly increasing, then for any $f \in H_K \setminus H_{K,N}$ it holds that

$$\|f_{K,N}\|_{H_K} < \|f\|_{H_K} \quad \Longrightarrow \quad R(\|f_{K,N}\|_{H_K}) < R(\|f\|_{H_k}), \tag{19}$$

and using the same argument as above we get that

$$\mathcal{E}(f_{K,N}) < \mathcal{E}(f), \;\; \forall\, f \in H_K. \tag{20}$$

So the only minimizers of $\mathcal{E}$ will reside in the finite dimensional space $H_{K,N}$. □

The representer theorem is important for various reasons:

(i) It guarantees that a suitable machine learning model for our data is a nonlinear model of the form

$$f^*(\cdot) = \sum_{i=1}^{N} a_i K(x_i, \cdot) = \sum_{i=1}^{N} a_i \psi_i(\cdot), \;\; a_i \in \mathbb{R},$$

i.e. in terms of the set of "basis functions" $\{\psi_i, \; i = 1, \cdots, N\}$ that are determined in terms of the kernel function K. Importantly, the basis functions ψ_i are determined (shaped) by the data themselves since $\psi_i(\cdot) := K(x_i, \cdot)$. In this respect, for each data set we may obtain different looking basis functions, depending on the training data, even though the kernel function K is always the same.

(ii) Naturally, by changing the RKHS H_K in which we consider our model, we will obtain different models (i.e. we will employ different kernel functions K). Intimately connected with that is the regularization term, the nature of which is determined in terms of the choice of norm for the RKHS, $\|\cdot\|_{H_K}$. Different norms impose different qualitative features on the chosen model, on properties such as e.g. smoothness of the model etc.

(iii) Another important consequence of the representer theorem is that it greatly simplifies the nature of the learning problem, in terms of the optimization problem involved. The original problem of choosing $f^* \in \arg\min_{f \in H} \mathcal{E}(f)$ is an infinite dimensional problem, that present both theoretical (mathematical) as well as computational difficulties. The mathematical difficulty is that the generalization of the Weierstrass maximum theorem for infinite dimensional spaces is tricky, as the characterization of compact sets in such spaces is not a straighforward (although not impossible!) task.

Since the representer theorem guarantees that the solution to problem

$$\min_{f\in H_K} \mathcal{E}(f) := \min_{f\in H_K} \frac{1}{N}\sum_{i=1}^{N} \mathcal{E}_i(y_i, f(x_i)) + R(\|f\|_{H_K}), \tag{21}$$

admits a solution $f^*(\cdot)$ among the functions of the form

$$f_N(\cdot; a) = \sum_{i=1}^{N} a_i K(x_i, \cdot), \quad a = (a_1, \cdots, a_N) \in \mathbb{R}^N,$$

in the sense that $f^*(\cdot) = f_N(\cdot, a^*)$ for a suitable choice of the parameter $a = a^* \in \mathbb{R}^N$.

This implies that we can restrict (with no loss of generality) our search for minimizers of (21) among the finite dimensional set

$$H_{K,N} = \text{span}\{K(\cdot, x_i) \ : \ i = 1, \cdots, N\} = \{f_N(\cdot; a) = \sum_{i=1}^{N} a_i K(\cdot, x_i), \quad a = (a_1, \cdots, a_N) \in \mathbb{R}^N\}, \tag{22}$$

which in turn can be identified with the parameter set

$$\Theta_N = \{a = (a_1, \cdots, a_N),\ a_i \in \mathbb{R},\ i = 1, \cdots, N\} \subset \mathbb{R}^N. \tag{23}$$

By restricting the minimization problem (21) in $H_{K,N}$ (respectively Θ_N), we can express the original optimization problem (21) as

$$\min_{f\in H_K} \mathcal{E}(f) = \min_{f\in H_{K,N}} \mathcal{E}(f) = \min_{a\in\Theta_N} \mathcal{E}(f_N(\cdot; a)) =: \min_{a\in\Theta_N} F_N(a), \tag{24}$$

where

$$a \mapsto F_N(a) := \mathcal{E}(f_N(\cdot; a)), \quad F_N : \Theta_N \subset \mathbb{R}^N \to \mathbb{R}, \tag{25}$$

is a function defined on the finite dimensional set $\Theta_N \subset \mathbb{R}^N$.

Clearly, the equivalent formulation of problem (21) in terms of the parametrized problem as in the far right of (24), i.e., the minimization of F_N over a suitable subset of $\mathbb{R}^N$ allows us to reduce the problem of minimizing over a function space, to an equivalent problem of minimizing over a subset of the parameter space $\mathbb{R}^N$ a problem that can be handled with standard minimization algorithms, such as (stochastic) gradient descent, ADAM or even more accurate schemes such as e.g., quasi Newton schemes.

We close this section by the following important question:

> The representer theorem guarantees that as long as we only consider minimizing among the set of models in the function space H_K, we can reduce this problem to an optimization problem in $\mathbb{R}^N$, where N is the number of available data.
>
> As long as we are happy with restricting ourselves in the universe of models in H_K this is the full story, however, a crucial question is whether the universe of models H_K is sufficient for our needs.
>
> This imposes the important question of "how big" is H_K as compared to the set of functions containing **any** possible model for our data.

Hence, what we do need is to characterize particular reproducing Hilbert spaces H_K related to various choices of kernel function K, or show that these spaces are dense in more general classes of function spaces that may be perceived as models for our data, for example the class of continuous functions, or other classes such as the class of smooth functions etc.

7 Which spaces of functions are reproducing kernel Hilbert spaces?

Our instantaneous choice for a model for a data set $\{(x_i, y_i) \ : \ x_i \in X, \ y_i \in Y, \ i = 1, \cdots, N\}$, would be a continuous function $f : X \to Y$, such that $f(x_i) \simeq y_i$ for all $i = 1, \cdots, N$. The request for continuity of f is very natural and is related to the stability of the model. Continuity of f guarantees that if two inputs $x, x' \in X$ are "close" then so would be the corresponding responses $y = f(x)$ and $y' = f(x')$. This is a reasonable request related to the robustness of the model f.

We will denote by $C(X; Y)$ the space of continuous functions $f : X \to Y$. Let us restrict (without loss of generality) our attention to the case where the response space is $Y = \mathbb{R}$. We can endow the space of models $C(X; \mathbb{R}) = C(X)$ with the norm $\|f\|_{C(X)} = \sup_{x \in X} |f(x)|$, and it is easy to show that the space of models endowed with this norm $(C(X), \|\cdot\|_{C(X)})$ is a complete normed space (a Banach space).

However, a disturbing fact is that the vector space of all possible models for our data $(C(X), \|\cdot\|_{C(X)})$ is **not** a Hilbert space! A simple way to establish that is to realize that if it was, then the norm $\|\cdot\|_{C(X)}$ would be compatible with an inner product $\langle \cdot, \cdot \rangle$, in the sense that $\|f\|^2 := \|f\|^2_{C(X)} = \langle f, f \rangle$, and then by the properties of the inner product, the parallelogram identity,

$$\|f\|^2 + \|f'\|^2 = \frac{1}{2}\|f + f'\|^2 + \frac{1}{2}\|f - f'\|^2, \ \ \forall \ f, f' \in C(X). \tag{26}$$

should hold. One can find counter examples for the particular case of interest, hence this space is not an inner product space, therefore not a Hilbert space. Trying to change the norm to a different norm will not really help as well, as trying to substitute the supremum norm with the $L^2(X)$ norm will introduce problems with the completeness of the space.

This leads us to the following interesting observation:

> The space of all continuous functions $C(X)$ cannot be realized as a reproducing Hilbert space.

Hence, the class of models that can be realized as elements of reproducing kernel Hilbert spaces is a subspace of the class of continuous functions. The class depends on smoothness and dimensionality of X.

We present some examples of RKHS in the special case where $X \subset \mathbb{R}^n$.

Sobolev Spaces as Reproducing Kernel Hilbert Spaces: An important class of RKHS are Sobolev spaces. In a nutshell, Sobolev spaces consist of functions which are smooth, in a generalized way.

Definition 7.1. The set of absolutely continuous functions $AC(I)$ is defined as

$$AC(I) = \{f : I \subset \mathbb{R} \to \mathbb{R} \ : \ \exists g \in L^1_{loc}(I), \text{ s.t } f(x) = f(c) + \int_c^x g(y)dy, \ \forall \ c \in I\} \tag{27}$$

The set $AC(I)$ consists of functions that admit a derivative in the generalized sense. The generalized derivative of f is in fact the function g (in the above representation), which is defined in the $L^1(I)$ sense. In plain words this means that the function f may not be differentiable everywhere, but may be diffeentiable almost everywhere (in the measure theoretic sense). As an example, consider the function $f : I = [-1, 1] \to \mathbb{R}$, defined as $f(x) = |x|$. This is almost everywhere differentiable, in the sense that it is differentiable everywhere, except the point $x = 0$ (which as a single point is of Lebesgue measure zero). This function is not $C^1(I)$ (i.e, continuously differentialbe everywhere), but it is $AC(I)$.

We will then call g the weak derivative of f, and perceive g not as a function defined pointwise but a function defined as an $L^1_{loc}(I)$ object (i.e, as an object which is defined almost anywhere, that is everywhere apart from a countable set of points of I).

In the case of functions defined on one dimensional sets $I \subset \mathbb{R}$, the set of functions $AC(I)$ coincides with the Sobolev space $W^{1,2}(I)$

Definition 7.2 (Sobolev space $W^{1,2}(I)$)**.** The Sobolev space $W^{1,2}(I)$ is the completion of the space of smooth functions $C^\infty_c(int(I))$ in the norm

$$\|f\|_{W^{1,2}(I)} = \left(\|f\|^2_{L^2(I)} + \left\| \frac{df}{dx} \right\|^2_{L^2(I)} \right)^{1/2}, \tag{28}$$

where for any function $g : I \to \mathbb{R}$ the $L^2(I)$ norm is defined by $\|g\|_{L^2(I)} = (\int_I |g(x)|^2 dx)^{1/2}$.

An equivalent definition for the Sobolev space $W^{1,2}(I)$ can be as

$$W^{1,2}(I) = \{f : I \to \mathbb{R} \ : f \text{ weakly differentiable in the sense of (27) with } f, \frac{df}{dx} \in L^2(I)\}. \tag{29}$$

It can be shown that the Sobolev space $W^{1,2}(I)$ is a Hilbert space when endowed with the inner product

$$\langle f_1, f_2 \rangle_{W^{1,2}(I)} = \langle f_1, f_2 \rangle_{L^2(I)} + \langle \frac{df_1}{dx}, \frac{df_2}{dx} \rangle_{L^2(I)} = \int_I f_1 f_2 dx + \int_I \frac{df_1}{dx} \frac{df_2}{dx} dx \tag{30}$$

Moreover, it is possible to show that Sobolev spaces are related to other function spaces, such as Lebesgue spaces of higher exponent, spaces of continuous functions etc. This is effected through the celebrated Sobolev embedding theorem (see e.g., Evans (2022)) which for Sobolev functions defined on $I \subset \mathbb{R}$ guarantees that they are also continuous functions! This result is stated as $W^{1,2}(I) \hookrightarrow C(I)$ implying that there is an inclusion of the two sets, i.e. that $W^{1,2}(I) \subset C(I)$, and there is a continuous mapping $\iota : W^{1,2}(I) \to C(I)$ such that for any

$f \in W^{1,2}(I)$ its image $\iota(f) = f \in C(I)$ (meaning the image $\iota(f)$ is now considered as a function in the bigger set $C(I)$). This means that for all practical purposes, functions in $W^{1,2}(I)$ can be considered as continuous functions (i.e. functions in $C(I)$), but at the same time endow this set with a norm compatible with an inner product, i.e. turn it into a Hilbert space. This is possible, because we are now only dealing with a strict subset of the space of continuous functions, those that are smoother than their fellows, since they have square integrable weak derivatives.

It turns out that the Sobolev space $W^{1,2}(I)$ is a reproducing kernel Hilbert space (see e.g. Saitoh et al. (2016)).

Theorem 7.3. *Let $I = [a, b]$. The Sobolev space $W^{1,2}(I)$ is a reproducing kernel Hilbert space with kernel*

$$K(x, x') = 1 - a + \min(x, x'). \tag{31}$$

8 Kernelized models

A kernelized model is looking for an ML model in an appropriate RKHS $f \in H_K$, and using as a regularization term a term of the form $R(\|f\|_K)$,

$$\min_{f \in H_K} \mathcal{E}(f) := \min_{f \in H_K} \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{E}_i(y_i, f(x_i)) + R(\|f\|_{H_K}) \right).$$

Then use the representer theorem to transform the problem to the finite dimensional problem

$$\min_{(a_1, \cdots, a_N) \in \mathbb{R}^N} \mathcal{E}(a_1, \cdots, a_N),$$

which is obtained upon substituting $f(x) = \sum_{i=1}^{N} a_i K(x_i, x)$ in the original problem.

Then, this is solved using a standard optimization method, and the $a_1, \cdots, a_N$ are used to recover the model

For more details on kernelization see e.g., Hofmann et al. (2008).

8.1 Kernel ridge regression

The kernel ridge regression model corresponds to the loss function

$$\mathcal{E}(f) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \|f\|_{H_K}^2$$

This is the kernelized generalization of the standard linear ridge regression model.

The choice of K (equiv. H_K) is related to the smoothness of the functional model f.

By the representer theorem the solution of

$$\min_{f \in H_K} \mathcal{E}(f),$$

has to be of the form

$$f(x) = \sum_{i=1}^{N} a_i K(x_i, x),$$

for suitable choice of $(a_1, \cdots, a_N) \in \mathbb{R}^N$.

We will use this representation to transform the original functional (infinite dimensional) problem to a finite dimensional one with respect to the coefficients a_i, $i = 1, \cdots, N$.

We perform some necessary calculations first.

Fidelity term:

$$\begin{aligned}\sum_{i=1}^{N}(y_i - f(x_i))^2 &= \sum_{i=1}^{N}(y_i - \sum_{j=1}^{N} a_j K(x_j, x_i))^2 \\ &= \sum_{i=1}^{N}(y_i - (Ka)_i)^2 = \|y - Ka\|_2^2,\end{aligned}$$

where

$$\begin{aligned}K = (K_{ij})_{i,j=1}^{N}, \quad K_{ij} = k(x_i, x_j), \ \text{Gram matrix}, \ K = K^T, \\ (Ka)_i, \ \text{i-th component of vector} \ Ka\end{aligned}$$

Note the similarity with the standard linear regression problem – Here the linear design matrix is replaced by the nonlinear Gram matrix.

Regularization term:

$$\begin{aligned}\|f\|_{H_K}^2 = \langle f, f\rangle_{H_K} &= \langle \sum_{i=1}^{N} a_i K(x_i, \cdot), \sum_{j=1}^{N} a_j K(x_j, \cdot)\rangle_{H_K} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j \langle K(x_i, \cdot), K(x_j, \cdot)\rangle_{H_K} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j K(x_i, x_j) = \langle a, Ka\rangle_{\mathbb{R}^N}.\end{aligned}$$

In the above we used the fundamental RKHS property that

$$\langle K(x_i, \cdot), K(x_j, \cdot)\rangle_{H_K} = K(x_i, x_j).$$

Note the similarity with the standard linear regression problem – Here the ridge regularization term $\|a\|_2^2$ replaced by the quadratic term $\langle a, Ka\rangle_{\mathbb{R}^N}$, a new norm weighted by the nonlinear Gram matrix.

The original problem

$$\min_{f\in H_K} \mathcal{E}(f) \iff \min_{a=(a_1,\cdots,a_N)^T\in\mathbb{R}^N} \frac{1}{N}\|y-Ka\|_2^2 + \lambda\langle a, Ka\rangle_{\mathbb{R}^N}$$

We redefine λ so as to set $N = 1$.

The first order condition becomes

$$K(y-Ka) - \lambda Ka = 0 \iff K(K+\lambda I)a = Ky$$

which can be solved if K is non singular in terms of

$$a^* = (K+\lambda I)^{-1}y$$

The chosen model is

$$f^*(x) = \sum_{i=1}^{N} a_i^* K(x_i, x).$$

8.2 Kernel SVM

Support vector machines (SVMs) are useful models for classifying binary data y_i, in terms of multidimensional features x_i.

Recall the standard linear SVM:

Given data $(x_i, y_i) \in \mathbb{R}^m \times \{-1, 1\}$, $i = 1, \cdots, N$, find a linear separation boundary given by the hyperplane $b_0 = \sum_{j=1}^m b_j z_j = \langle b, z\rangle_{\mathbb{R}^m}$ such that the data are separated by this boundary.

In particular, if

$$\begin{aligned}\langle b, x_i\rangle > b_0 &\implies y_i = +1,\\ \langle b, x_i\rangle < b_0 &\implies y_i = -1.\end{aligned}$$

More than one hyperplanes can separate the data: We will choose the one that is optimal in terms of maximizing a margin (related to the distance of the boundary of the two populations from this boundary).

This is often expressed as the optimization problem

$$\begin{aligned}&\max_{b\in\mathbb{R}^m, b_0\in\mathbb{R}} M,\\ \langle b, x_i\rangle - b_0 > M &\text{ if } y_i = +1,\\ \langle b, x_i\rangle - b_0 < -M &\text{ if } y_i = -1\end{aligned}$$

or in the equivalent form

$$\begin{aligned}&\max_{b\in\mathbb{R}^m, b_0\in\mathbb{R}} M,\\ y_i(\langle b, x_i\rangle - b_0) > M,\ &i = 1, \cdots, N\end{aligned}$$

Using convex duality methods we can show that the solution to this problem must have the form

$$b = \sum_{i=1}^{N} a_i y_i x_i, \quad a_i \in \mathbb{R},$$

with the a_i to be obtained in terms of the solution of the quadratic optimization problem

$$\max_{(a_1,\cdots,a_N)\in\mathbb{R}_+^N} \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \langle x_i, x_j\rangle_{\mathbb{R}^m} a_i a_j$$

The kernelized SVM method starts from this dual problem by replacing the original dual problem by its kernelized variant

$$\max_{(a_1,\cdots,a_N)\in\mathbb{R}_+^N} \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j K(x_i, x_j) a_i a_j,$$

where K is a chosen kernel function.

Since

$$K(x_i, x_j) = \langle K(x_i, \cdot), K(x_j, \cdot)\rangle_{H_K},$$

this can be interpreted as looking for a separation boundary of the form

$$f(x) = \sum_{i=1}^{N} a_i y_i K(x_i, x),$$

which is a nonlinear separation boundary.

Note that the kernelized version of the dual problem is formally the same as the non kernelized one.

Both are quadratic optimization problems of the form

$$\max_{a\in\mathbb{R}_+^N} \langle \mathbf{1}, a\rangle_{\mathbb{R}^N} - \frac{1}{2}\langle a, Ga\rangle_{\mathbb{R}^N},$$

for some positive (semi) definite matrix $G \in \mathbb{R}^{N\times N}$.

- In the standard case

$$G = (G_{ij})_{i,j=1}^{N}, \quad G_{ij} = y_i \langle x_i, x_j\rangle_{\mathbb{R}^m} y_j,$$

- In the kernelized case

$$G = (G_{ij})_{i,j=1}^{N}, \quad G_{ij} = y_i K(x_i, x_j) y_j = y_i \langle K(x_i, \cdot), K(x_j, \cdot)\rangle_{H_K} y_j,$$

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society 68*(3), 337–404.

Evans, L. C. (2022). *Partial differential equations*, Volume 19. American mathematical society.

Hofmann, T., B. Schölkopf, and A. J. Smola (2008). Kernel methods in machine learning. *The Annals of Statistics 36*(3), 1171–1220.

Saitoh, S., Y. Sawano, et al. (2016). *Theory of reproducing kernels and applications*, Volume 44. Springer.