BGSM/CRM
AL&DNN

Reproducing Kernel Hilbert Spaces

S. Xambó

UPC & IMTech

13/10/2021

Abstract. This session is devoted to the Reproducing Kernel Hilbert Spaces (RKHS) and their relation to algorithmic learning.

References, ordered by the publication year:

[1] (presentation unifying several strands developed earlier, 1950),

[2] (an even more general mathematical theory, 1964),

[3]* (a thorough treatise, with a very extensive bibliography, 2002),

[4, Ch. 2, 3, 9, 10, 11, 12]*, 2004.

[5, Ch. 5, 6] 2008, [6, Ch 16]* 2014, [7] 2015,

[8] 2016, [9, Ch. 5, 6]* 2018, [10, Ch. 5] 2019, [11, Ch. 14] 2020,

[12] (tutorial and survey, 2021).

Index

Kernels

(kernel functions)

The word kernel is used with many meanings, as for example the kernel of a linear map or of a group homomorphism in mathematics, or as a synonym of filter in signal theory, also in convolutional neural networks, and so on.

The kernels, or kernel functions, we are about to study in this session are quite *different notions*.

They were introduced by D. Hilbert and J. Mercer in the early years of last century and have a long, intricate and fascinating history.

> James Mercer (1883-1932) [...] proved [...] that *positive-definite kernels can be expressed as a dot product in a high-dimensional space*. [... which is the] basis of the *kernel trick* [...], which *allows linear algorithms to be easily converted into non-linear algorithms* (`https://en.wikipedia.org/wiki/James_Mercer_(mathematician)`).

Aim of this session: to unravel the main mathematical facts about kernels and their relation to algorithmic learning (*kernel methods*).

- Can be defined on a variety of data, including vectors, text, and strings, among others.
- Can express powerful mappings, relations, and patterns, including classifiers, regressors, clusters, and rankings, among others.
- The algorithms based on kernels can deal with nonlinear phenomena in an efficient and relatively simple way.
- There is a comprehensive mathematical framework, the theory of RKHS, that illuminates the data processing operations and the derived learning algorithms.
- There are convenient software packages that facilitate the computational aspects.

Unless otherwise stated, al vector spaces are real vector spaces.

Generic symbols for some families of vector spaces:

- $\mathbf{R}^n$ ($n \geqslant 1$), or $\mathcal{E}^n$: *Euclidean space*.
- $\mathcal{V}$: *Inner product space*: Vector space endowed with a *bilinear symmetric scalar product*, denoted by $x \cdot x'$ or $\langle x, x' \rangle$.
- $\mathcal{H}$: Hilbert space.
- $\mathcal{B}$: Banach space.

Some basic notions about Normed, Banach and Hilbert spaces will be summarized a little later.

Types of inner product spaces

Property	**Name**	*Alias*
$x \cdot x > 0$ if $x \neq 0$	Positive definite	*Positive*
$x \cdot x \geqslant 0$ for all x	Positive semi-definite	*Semi-positive*
$x \cdot x < 0$ if $x \neq 0$	Negative definite	*Negative*
$x \cdot x \leqslant 0$ for all x	Negative semi-definite	*Semi-negative*
$\exists x, x', x \cdot x > 0, x' \cdot x' < 0$	Indefinite	

A space $\mathcal{V}$ will be denoted $\mathcal{V}_+$ if its inner-product is *semi-positive*.

The types in the table above correspond, with the same names, to types of *real symmetric matrices*. These types can be defined in terms of the signs of the eigenvalues: *positive* (*negative*), if all evs are positive (negative); *semi-positive* (*semi-negative*), if the evs are non-negative (non-positive); *indefinite*, if some evs are positive and some negative.

Kernel functions

Let $\mathcal{X}$ be a set.

A (Mercer) *kernel* on $\mathcal{X}$ is a function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ such that $\kappa(x, x') = \kappa(x', x)$ for all $x, x' \in \mathcal{X}$ (κ is *symmetric*) and with the property that for any $x^1, \cdots, x^m \in \mathcal{X}$, $m \in \mathbf{N}$, and any $\lambda_1, \cdots, \lambda_m \in \mathbf{R}$, the following inequality holds (*Mercer condition*):

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j \kappa(x^i, x^j) \geqslant 0. \tag{1}$$

This means that the symmetric matrix $K = \left(\kappa(x^i, x^j)\right)_{i,j \in [m]}$ (*kernel matrix*, or also *Gram matrix*, of $x^1, \cdots, x^m$) is semi-positive, which in turn is equivalent to say that its eigenvalues are non-negative.

Note. By what we will see, $\kappa(x, x')$ can be thought of as some kind of *similarity* degree of (or *distance* between) x and x'.

Cauchy-Schwarz inequality for kernels

Let κ be a kernel on $\mathcal{X}$. Then $k(x,x) \geqslant 0$ for any $x \in \mathcal{X}$ and

$$\kappa(x,x')^2 \leqslant \kappa(x,x)\kappa(x',x')$$

for any $x, x' \in \mathcal{X}$.

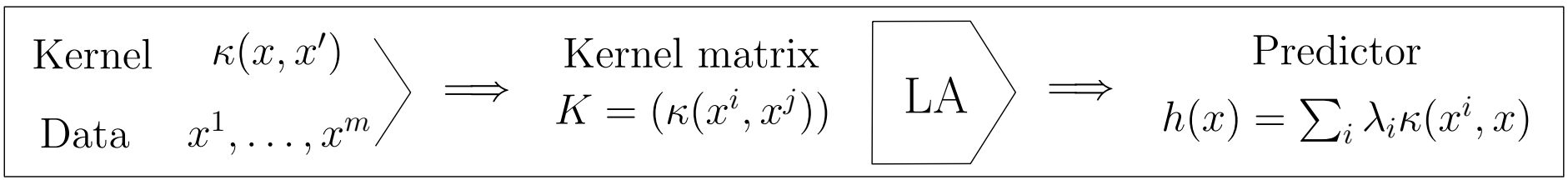
In particular, if $\kappa(x,x) = 0$, then $\kappa(x,x') = 0$ for all $x' \in \mathcal{X}$.

Proof. Since the kernel matrix of x, x',

$$K = \begin{pmatrix} \kappa(x,x) & \kappa(x,x') \\ \kappa(x',x) & \kappa(x',x') \end{pmatrix},$$

is semi-posive, its determinant $\kappa(x,x)\kappa(x',x') - \kappa(x,x')^2 \geqslant 0$. □

Relation to algorithmic learning



Bilinear kernel. Given $\mathcal{V}_+$, $\kappa(x, x') = \langle x, x' \rangle$ is a kernel on $\mathcal{X} = \mathcal{V}$ (we will say that it is the *quadratic* or *(bi)linear kernel* on $\mathcal{V}$).

In fact, in this case the sum in (1) is

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j \langle x^i, x^j \rangle = \left\langle \sum_{i \in [m]} \lambda_i x^i, \sum_{j \in [m]} \lambda_j x^j \right\rangle \geqslant 0.$$

Pull-back of a kernel. If κ is a kernel on $\mathcal{Y}$ and $\phi : \mathcal{X} \to \mathcal{Y}$ is a map, then the function $\kappa_\phi : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ defined by

$$\kappa_\phi(x, x') = \kappa(\phi(x), \phi(x'))$$

is a kernel on $\mathcal{X}$.

An useful special case occurs when $\mathcal{Y} = \mathcal{V}_+$ with κ the bilinear kernel. In this case, $\kappa_\phi(x, x') = \langle \phi(x), \phi(x') \rangle$.

As hinted before, particularly when $\mathcal{V} = \mathbf{R}^N$, $\mathcal{V}$ is usually called a *feature space*, ϕ a *feature map*, and κ_ϕ its *associated kernel*.

Kernels derived from other kernels. Let $\kappa_i : \mathcal{X} \times \mathcal{X}$, $i \in \mathbf{N}$, be a sequence of kernels on $\mathcal{X}$ and λ_i a sequence of non-negative numbers.

Then the following are kernels on $\mathcal{X}$:

(1) If $e_1, \cdots, e_r \in \mathbf{N}$, $\kappa_1^{e_1} \cdots \kappa_r^{e_r}$.
By induction, it is enough to establish that the product $\kappa_1 \kappa_2$ is a kernel if κ_1 and κ_2 are kernels. We will see a proof of this fact below (page 30).

(2) $\lim_{i \to \infty} \kappa_i$, if the limit exists (for any $x, x' \in \mathcal{X}$).

(3) $\sum_{i \geqslant 1} \lambda_i \kappa_i$, if this series converges (for any $x, x' \in \mathcal{X}$). In particular, $\lambda_1 \kappa_1 + \cdots + \lambda_r \kappa_r$.

(4) The exponential $e^{\kappa} = \sum_{j \geqslant 0} \frac{\kappa^j}{j!}$, for any kernel κ.

(5) More generally, let $a_0, \ldots, a_j, \ldots$ be a sequence of *non-negative* real numbers and $a(t) = \sum_{j \geqslant 0} a_j t^j$.

Let κ be a kernel on $\mathcal{X}$ and assume that $-R < \kappa(x, x') < R$ for all $x, x' \in \mathcal{X}$, where R is the radius of convergence of $a(t)$.

Then $a(\kappa) = \sum_{j \geqslant 0} a_j \kappa^j$ is a kernel on $\mathcal{X}$.

(6) If κ is a kernel on $\mathcal{X}$,

$$\bar{\kappa}(x, x') = \begin{cases} 0 & \text{if } \kappa(x, x) = 0 \text{ or } \kappa(x', x') = 0 \\ \frac{\kappa(x, x')}{\sqrt{\kappa(x, x)}\sqrt{\kappa(x', x')}} & \text{otherwise.} \end{cases}$$

This kernel is called the *normalization* of κ. We will prove that it is a kernel later (page 31).

Concrete examples. The following are kernels on the specified $\mathcal{X}$:

(1) For any $\mathcal{X}$, $\kappa(x, x') = a$, $a \in \mathbf{R}_+$ a constant (*constant kernel*).

(2) *Polynomial kernels*: If $\mathcal{X} = \mathcal{V}_+$, in particular if $\mathcal{V} = \mathbf{R}^n$, then for any constant $a \in \mathbf{R}_+$ and any $d \in \mathbf{N}$, $\kappa(x, x') = (a + x \cdot x')^d$. The integer d is the *degree* of the kernel. Note that for fixed x, $\kappa(x, x')$ is a polynomial of degree d in x'.

(3) If $\mathcal{X} = \mathbf{R}^n$, $\alpha \in \mathbf{R}_{++}$, $d \in \mathbf{N}$, $\kappa_{\alpha,d}(x, x') = e^{-\alpha\|x-x'\|^d}$.

The case $d = 1$ is the *Laplacian kernel*, and the case $d = 2$, the *Gaussian kernel* (in this case, α is expressed in the form $1/2\sigma^2$).

That the Gaussian kernel is a kernel will be proved later (page 32).

Remark. The functions of the form $e^{-\alpha\|x-x'\|^d}$ have a *radial property*: They decrease radially around any given x when seen as a function of x', with $\kappa(x, x) = 1$.

Background: Normed, Banach and Hilbert spaces

The notions we are going to recall in this section are also valid for the complex field $\mathbf{C}$, with some mild adaptations, in place of the real field $\mathbf{R}$. For simplicity we focus on the real case.

Main references: [13], [14].

A *norm* on a real vector space $\mathcal{L}$ is a function $\|\cdot\| : \mathcal{L} \to \mathbf{R}$ such that

- $\|x\| \geqslant 0$ for all $x \in \mathcal{L}$, with equality if and only if $x = 0$ (*positivity*);
- $\|\lambda x\| = |\lambda|\|x\|$ for all $x \in \mathcal{L}$ and $\lambda \in \mathbf{R}$ (*homogeneity*); and
- $\|x + x'\| \leqslant \|x\| + \|x'\|$ for all $x, x' \in \mathcal{L}$ (*triangle inequality*).

The pair $(\mathcal{L}, \|\ \|)$ is called a *normed* vector space.

The Euclidean linear spaces are finite-dimensional normed spaces (in $\mathbf{R}^n$, for example, $\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$).

Example: $\ell^p, 1 \leqslant p \leqslant \infty$.
This is the space of real sequences $x = \{x_k\}_{k \geqslant 1}$ such that $\|x\|_p = \left(\sum_{k \geqslant 1} |x_k|^p\right)^{1/p} < \infty$.

The space ℓ^∞ is the space of absolutely bounded sequences x, with norm $\|x\|_\infty = \sup\{|x_k|\}_{k \geqslant 1}$.

In a normed space the notions of convergent and Cauchy sequences make sense.

A sequence $x_1, x_2, \cdots, x_n, \cdots \in \mathcal{L}$ is *convergent* if there exists $v \in \mathcal{V}$ such that $\|x_n - v\| \to 0$ as $n \to \infty$.

If this is the case, v is unique and we write $v = \lim_{n\to\infty} x_n$.

The sequence $\{x_n\}$ is said to be a *Cauchy sequence* if for any $\epsilon > 0$ there exists $N = N_\epsilon$ such that $\|x_i - x_j\| \leqslant \epsilon$ for all $i, j \geqslant N$. Any convergent sequence is a Cauchy sequence.

The normed space $\mathcal{B} = (\mathcal{L}, \| \ \|)$ is said to be *complete*, or a *Banach space*, if any Cauchy sequence is convergent.

Examples. The spaces ℓ^p $(1 \leqslant p \leqslant \infty)$ are Banach spaces.

If X is a compact topological space, then the space $C(X)$ of continuous functions on X, with the norm $\|f\|_{\sup} = \sup_{x\in X} |f(x)|$, is a Banach space.

Any normed space $(\mathcal{L}, \| \|)$ can be enlarged to a Banach space.

This process mimics the construction of $\mathbf{R}$ as a completion of $\mathbf{Q}$.

In fact, the absolute value $|\cdot|$ is a norm on $\mathbf{Q}$ (and on $\mathbf{R}$) and $\mathbf{R}$ can be defined as the quotient ring of the ring of Cauchy sequences in $\mathbf{Q}$ by the ideal of sequences that are convergent to 0 (*null sequences*).

Similarly, we can form the quotient $\bar{\mathcal{L}}$ of the vector space of all Cauchy sequences in $\mathcal{L}$ by the vector subspace of the null sequences.

We note that $\mathcal{L} \subseteq \bar{\mathcal{L}}$, by mapping $x \in \mathcal{L}$ to the class of the constant sequence $x_n = x$ for all n, and that the norm on $\mathcal{L}$ extends to a norm on $\bar{\mathcal{L}}$: the norm $\|\bar{x}\|$ of $\bar{x} \in \bar{\mathcal{L}}$ is defined as $\lim_{n\to\infty} \|x_n\|$ for any Cauchy sequence representing $\bar{x}$.

Example: The Lebesgue spaces L^p (for details, see [14, Ch. 7])

Let $\mathcal{X} = (X, \Sigma, \mu)$ be a *measure space* and $1 \leqslant p < \infty$.

For any measurable function $f : X \to \mathbf{R}$, $\|f\|_p$ is defined as $\left(\int_X |f|^p d\mu\right)^{1/p}$, where $\int_X$ is the Lebesgue integral.

The Lebesgue space $L^p(\mathcal{X})$, or simply $L^p(\mu)$, consists of the measurable functions $f : X \to \mathbf{R}$ such that $\|f\|_p < \infty$, *with the convention that two functions are equal if and only if they agree almost everywhere*.

The space L^p is a normed space, with norm $\| \|_p$ (a consequence of Minkowski inequality), and in fact it is complete (Riesz-Fischer theorem), hence a Banach space.

Remark. The space ℓ^p is a special case: $X = \{1, 2, 3, \cdots\}$ with the counting measure.

Let $(\mathcal{B}, \|\,\|)$ be a Banach space. If $\mathcal{B}$ is finitely-generated, the notion of basis coincides with the notion introduced in elementary linear algebra: a finite set of linearly independent vectors that span the space. Since the results we will state turn out to be obvious in this case, we will assume that $\mathcal{B}$ is not finitely generated.

A *basis* for $\mathcal{B}$ is a sequence $u_1, \cdots, u_k, \cdots \in \mathcal{B}$ such that for each $x \in \mathcal{B}$ there is a *unique sequence* $\lambda_1, \cdots, \lambda_k, \cdots \in \mathbf{R}$ such that $x = \sum_{k\geqslant 1} \lambda_k u_k$, where the convergence of the series is in the sense of the norm: $x = \lim_{n\to\infty} \sum_{k=1}^{n} \lambda_k u_k$. Thus the vectors u_k are linearly independent and span a space (their finite linear combinations) that is dense in $\mathcal{B}$.

Example. Let $u_k \in \ell^p$ be $\{\delta_{k,j}\}_{j\geqslant 1}$ (has 1 in the position k and 0 otherwise). Then $\{u_k\}_{k\geqslant 1}$ is a basis of ℓ^p, for all $p \in [1, \infty)$. □

Given a basis, a subtle result is that the map $\lambda_k : \mathcal{B} \to \mathbf{R}$, $x \mapsto \lambda_k(x)$ is continuous for all k (see [15, Theorem 1.6]).

If for any $x \in \mathcal{B}$ the convergence of $\sum_{k\geqslant 1} \lambda_k u_k$ is unconditional, $\{u_k\}$ is said to be an *unconditional basis*.

The concept of real Hilbert space is a natural extension of the notion of Euclidean space to possibly infinite dimensions.

Let $\mathcal{H}$ be a real vector space endowed with a *inner scalar product* $\langle x, y\rangle \in \mathbf{R}$, for all $x, y \in \mathcal{H}$, that is *bilinear*, *symmetric*, and *positive definite* ($\langle x, x\rangle > 0$ for all $x \neq 0$). In some contexts, particularly if $\mathcal{H}$ has finite dimension, the scalar product is also denoted by $x \cdot y$.

Then $\mathcal{H}$ is a normed space, with $\|x\| = +\sqrt{\langle x, x\rangle}$. In this case the triangle inequality $\|x + x'\| \leqslant \|x\| + \|x'\|$ has the property that it is an equality if an only if either $x = 0$ or $x' = \lambda x$ with $\lambda \in \mathbf{R}_+$.

If this normed space is complete, then we say that $\mathcal{H}$ is a *Hilbert space*.

Obviously, any Hilbert space is a Banach space.

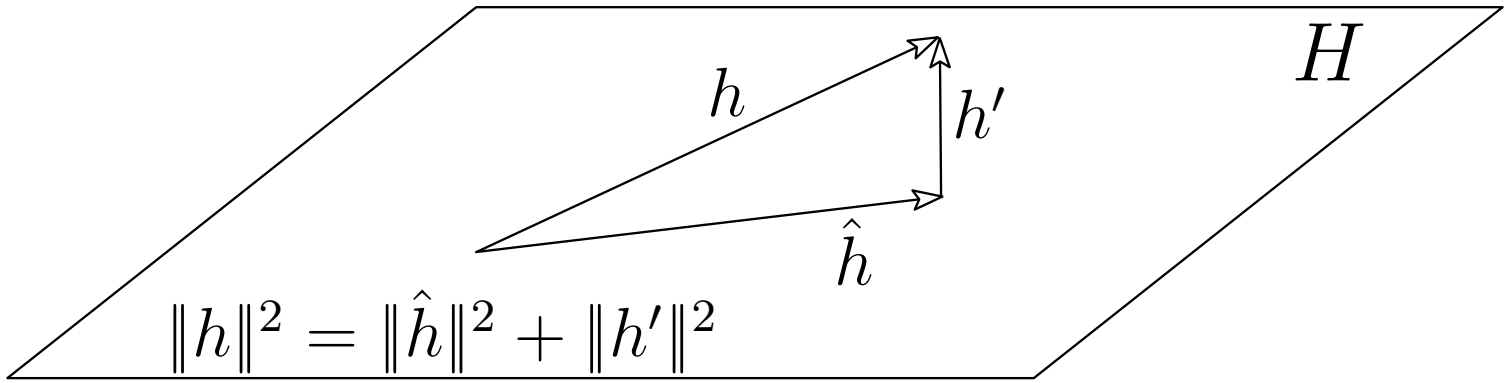
Example. The space $L^2(\mathcal{X})$ is a Hilbert space, with $\langle f, g\rangle = \int_X f(x)g(x)d\mu(x)$, and ℓ^2, with $\langle a, b\rangle = \sum_{k\geqslant 1} a_k b_k$.

Fact. If H is a finite dimensional subspace of a Hilbert space $\mathcal{H}$, then $\mathcal{H} = H \oplus H^\perp$, where $H^\perp = \{h' \in \mathcal{H} \,|\, \langle h, h' \rangle = 0 \ \forall h \in H\}$.

This means that for any $h \in \mathcal{H}$ there is a unique decomposition $h = \hat{h} + h'$ with $\hat{h} \in H$ and $h' \in \mathcal{H}^\perp$.

Since $\langle h, h \rangle = \langle \hat{h}, \hat{h} \rangle + \langle h', h' \rangle$ (as $\langle \hat{h}, h' \rangle = 0$), we have (Pythagoras)

$$\|h\|^2 = \|\hat{h}\|^2 + \|h'\|^2.$$



Remark. The stament about H is valid for any *closed* vector subspace of $\mathcal{H}$, a condition that is automatically satisfied when the dimension of H is finite.

Reproducing Kernel Hilbert Spaces

If $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ is a kernel, and $x \in \mathcal{X}$, we will write $\kappa_x : \mathcal{X} \to \mathbf{R}$ to denote the function defined by $\kappa_x(x') = \kappa(x, x')$.

If $\mathcal{H}$ is a Hilbert space of functions $f : \mathcal{X} \to \mathbf{R}$, we say that κ is a *reproducing kernel* for $\mathcal{H}$ if $\kappa_x \in \mathcal{H}$ for all $x \in \mathcal{X}$ and $f(x) = \langle \kappa_x, f \rangle$ for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$ (*reproducing property*).

In particular we have

$$\kappa(x, x') = \kappa(x', x) = \kappa_{x'}(x) = \langle \kappa_x, \kappa_{x'} \rangle.$$

This means that we can regard $\mathcal{H}$ as a *feature space*, the map $\phi : \mathcal{X} \to \mathcal{H}$, $x \mapsto \kappa_x$, as a *feature map*, and $\kappa = \kappa_\phi$ as the kernel associated to ϕ.

We will often write $\phi(x)$ instead of κ_x, particularly when this helps in the readability of expressions.

Remark. The condition $f(x) = \langle \kappa_x, f \rangle$ implies that the evaluation map $v_x : \mathcal{H} \to \mathbf{R}$, $f \mapsto f(x)$, is continuous for any given $x \in \mathcal{X}$.

Given points $x^1, \dots, x^m \in \mathcal{X}$ and scalars $\lambda_1, \cdots, \lambda_m$, we can consider the function $f = \sum_{i=1}^m \lambda_i \phi(x^i) \in \mathcal{H}$.

If $f' = \sum_{j=1}^{m'} \lambda'_j \phi(x'^j) \in \mathcal{H}$ is another such function ($x'^1, \cdots, x'^{m'} \in \mathcal{X}$), the inner product $\langle f, f' \rangle$ is equal to

$$\sum_{i,j} \lambda_i \lambda'_j \langle \phi(x^i), \phi(x'^j) \rangle = \sum_{i,j} \lambda_i \lambda'_j \kappa(x^i, x'^j).$$

If f is as above and $f' = \kappa_x = \phi(x)$, then

$\langle \phi(x), f \rangle = \sum_i \lambda_i \langle \phi(x), \phi(x^i) \rangle = \sum_i \lambda_i \kappa(x, x^i) = \sum_i \lambda_i \phi(x^i)(x) = f(x)$, which ratifies the reproducing property for this particular $f \in \mathcal{H}$.

Notation. The subspace of $\mathcal{H}$ spanned by the functions $\phi(x) = \kappa_x$, $x \in \mathcal{X}$, will be denoted by $\mathcal{H}_0$.

Note that $\mathcal{H}_0$, and its inner product, only depend on κ.

Theorem. Let κ be a kernel on $\mathcal{X}$. Then there exists a Hilbert space $\mathcal{H}$ and a mapping $\phi : \mathcal{X} \to \mathcal{H}$ such that $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle$ for all $x, x' \in \mathcal{X}$.

Proof. Let $\mathcal{V}_\kappa$ be the vector subspace of $\mathbf{R}^{\mathcal{X}}$ spanned by the functions $\phi(x) = \kappa_x$. The elements of $\mathcal{V}_\kappa$ are finite linear combinations of the form $h = \sum_{x \in F} \lambda_x \phi(x)$, where $F \subseteq \mathcal{X}$ is finite.

Endow $\mathcal{V}_\kappa$ with the inner product $\langle h, h' \rangle$, $h' = \sum_{x' \in F'} \lambda'_{x'} \phi(x')$:

$$(*) \qquad \langle h, h' \rangle = \sum_{x \in F} \sum_{x' \in F'} \lambda_x \lambda'_{x'} \kappa(x, x').$$

This expression only depends on h and h', and not on the linear combinations used to represent them. For example, the right hand side of $(*)$ is equal to
$\sum_{x \in F} \lambda_x \left(\sum_{x' \in F'} \lambda'_{x'} \phi(x')(x) \right) = \sum_{x \in F} \lambda_x h'(x)$, and this shows that it only depends on h'. Similarly, we have
$\sum_{x' \in F'} \lambda'_{x'} \left(\sum_{x \in F} \lambda_x \kappa_x(x') \right) = \sum_{x' \in F'} \lambda'_{x'} h(x')$.

Reproducing property for $\mathcal{V}_\kappa$: $\langle \phi(x), h' \rangle = h'(x)$.

Semi-positivity. Let $x^1, \cdots, x^m \in \mathcal{X}$ and $h = \sum_{i=1}^m \lambda_i \phi(x^i) \in \mathcal{V}_\kappa$. Then $\langle h, h \rangle = \sum_{i,j} \lambda_i \lambda_j \kappa(x^i, x^j)$, which is $\geqslant 0$ by the definition of Mercer kernel.

Let $\mathcal{H}$ be the completion of $\mathcal{V}_\kappa$. It is a Hilbert space.

For any fixed $x \in \mathcal{X}$, the evaluation map $v_x : \mathcal{V}_\kappa \to \mathbf{R}$ is continuous, for $v_x(h) = \langle \kappa_x, h \rangle$ and we can appeal to the Cauchy-Schwarz inequality $|\langle \kappa_x, h \rangle| \leqslant \sqrt{\kappa(x,x)} \|h\|$. So $\langle \kappa_x, h \rangle$ is well defined for any $h \in \mathcal{H}$.

Now if $h_n \in \mathcal{V}_\kappa$ is a Cauchy sequence and $\lim_{n \uparrow \infty} = h \in \mathcal{H}$, then $\langle \kappa_x, h \rangle = \langle \kappa_x, \lim_{n \uparrow \infty} h_n \rangle = \lim_{n \uparrow \infty} \langle \kappa_x, h_n \rangle = \lim_{n \uparrow \infty} h_n(x) = h(x)$, which is the reproducing property. □

The Hilbert space constructed in the above proof is the RKHS associated to κ.

Kernel studio

Product of kernels
Kernel normalization
The Gaussian kernels

The product of two kernels is a kernel. Let κ and κ' be two kernels on $\mathcal{X}$. Let K and K' be the kernel matrices of κ and κ' relative to the same set of m points of $\mathcal{X}$. Then the kernel matrix of $\kappa\kappa'$, relative to the same m points, is $[K_{ij}K'_{ij}]_{ij}$. To prove that $\kappa\kappa'$ is a kernel we have to check that for any $\lambda_1, \cdots, \lambda_m \in \mathbf{R}$ we have $\sum_{ij} \lambda_i\lambda_j K_{ij}K'_{ij} \geqslant 0$.

To see this, we will use that the semi-positive matrix K can be expressed in the form $K = MM^T$ for some $M \in \mathbf{R}(m)$ (Cholesky decomposition; see [16, §I.7]). Then we have

$$\sum_{ij} \lambda_i\lambda_j K_{ij}K'_{ij} = \sum_{ij} \lambda_i\lambda_j \left([\sum_{k=1}^{m} M_{ik}M_{jk}] K'_{ij} \right)$$

$$= \sum_{k=1}^{m} \left(\sum_{i,j} (\lambda_i M_{ik})(\lambda_j M_{jk}) K'_{ij} \right) \geqslant 0.$$

In the last step we use that K' is semi-positive, and hence, with $\lambda'_i = \lambda_i M_{ik}$, $\sum_{ij} \lambda'_i\lambda'_j K'_{ij} \geqslant 0$.

Given $x^1, \cdots, x^m \in \mathcal{X}$, let us check that the matrix $\bar{K} = \left(\bar{\kappa}(x^i, x^j)\right)$ is semi-positive.

For any $\lambda_1, \cdots, \lambda_m \in \mathbf{R}$, we need to show that

$(*) \qquad \sum_{i,j} \lambda_i \lambda_j \bar{\kappa}(x^i, x^j) \geqslant 0.$

If $\kappa(x^i, x^i) = 0$, then $\bar{\kappa}(x^i, x^j) = 0$ for all $j \in [m]$ (by definition of $\bar{\kappa}$). So we may assume that $\kappa(x^i, x^i) \neq 0$ for all $j \in [m]$.

To ease notation, set $h_i = \kappa_{x^i}$. Then $\bar{\kappa}(x^i, x^j) = \langle h_i, h_j \rangle / \|h_i\| \|h_j\|$, where the scalar produt and the norms are relative to the RKHS associated to κ, and the value of $(*)$ is

$$\sum_{i,j} \lambda_i \lambda_j \frac{\langle h_i, h_j \rangle}{\|h_i\| \|h_j\|} = \left\langle \sum_i \lambda_i \frac{h_i}{\|h_i\|}, \sum_j \lambda_j \frac{h_j}{\|h_j\|} \right\rangle \geqslant 0. \qquad \square$$

We have defined the *Gaussian kernel* by the expression $e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$, $x, x' \in \mathbf{R}^n$.

It is a kernel because it is the normalization $\bar{\kappa}$ of the kernel $\kappa(x, x') = e^{\frac{x \cdot x'}{\sigma^2}}$.

Indeed, $\kappa(x, x) = e^{\frac{\|x\|^2}{\sigma^2}}$. Similarly, $\kappa(x', x') = e^{\frac{\|x'\|^2}{\sigma^2}}$. Therefore

$$\bar{\kappa}(x, x') = \frac{e^{\frac{x \cdot x'}{\sigma^2}}}{e^{\frac{\|x\|^2}{2\sigma^2}} e^{\frac{\|x'\|^2}{2\sigma^2}}} = e^{\frac{1}{2\sigma^2}(2x \cdot x' - \|x\|^2 - \|x'\|^2)} = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}.$$

Feature maps studio

Polynomial feature maps
Feature map of the polynomial kernel
Feature map of the Gaussian kernel

Let $\mathcal{J}$ be the set of length n sequences of non-negative integers $J = j_1, \cdots, j_n$, and set $|J| = j_1 + \cdots + j_n$.

If $J \in \mathcal{J}$ and $x \in \mathbf{R}^n$, let $x^J = x_1^{j_1} \cdots x_n^{j_n}$ (a monomial of degree $|J|$).

Let $N = N_{n,d}$ be the dimension of the space of polynomials of degree d in n variables. Then we have a feature map $\phi : \mathbf{R}^n \to \mathbf{R}^N$ such that $\phi(x)_J = x^J$.

Remark. A degree d polynomial map $p : \mathbf{R}^n \to \mathbf{R}$ can be written in the form $p(x) = \sum_{r=0}^{d} \sum_{|J|=r} w_J x^J = w \cdot \phi(x)$, where $w \in \mathbf{R}^N$ is the vector whose components are the scalars w_J.

Remark. By imposing further restrictions on the J used in the definition of $p(x)$, we can get polynomial feature spaces of lower dimension. For example, $p(x) = \sum_{|J|=d} w_J x^J$ lies in the space of *homogeneous polynomials of degree d*. Its dimension is $\binom{d+n-1}{d}$.

In algebraic geometry, this is known as the *Veronese mapping*.

The polynomial kernel of degree d in $\mathbf{R}^n$ is given by $\kappa(x, x') = (1 + x \cdot x')^d$. So we know that there is a feature map $\phi : \mathbf{R}^n \to \mathcal{H}$ such that $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle$.

To find an explicit expression for $\mathcal{H}$ and ϕ, define, for any $x \in \mathbf{R}^n$ and any sequence $J = j_1, \cdots, j_d$ in $\{0, 1, \cdots, n\}$, $x_J = x_{j_1} \cdots x_{j_d}$ with the convention that $x_0 = 1$. Then we can consider the vector $\phi(x) \in \mathbf{R}^{(n+1)^d}$ whose coordinates are the monomials x_J.

Claim. $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle$.

Indeed, $\langle \phi(x), \phi(x') \rangle = \sum_J x_J x'_J = \sum_J x_{j_1} x'_{j_1} \cdots x_{j_d} x'_{j_d}$, while $\kappa(x, x') = (1 + x \cdot x)^d = (x_0 x'_0 + x_1 x'_1 + \cdots + x_n x'_n)^d$, and it is clear that both expressions coincide (by the distributive property of the product).

The Gaussian kernel in $\mathbf{R}^n$ has the form $\kappa(x,x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$.

With similar notations as in the polynomial kernel (preceding slide), let $\phi(x)_J = \frac{1}{\sqrt{k!}} e^{-\frac{\|x\|^2}{2\sigma^2}} \frac{x_J}{\sigma^k}$, now with $J = j_1, \cdots, j_k \in \{1, \cdots, n\}$ (so $k = |J|$). Let $\mathcal{S}$ be the space of real sequences indexed by the J, in some order. Then we have $\phi : \mathbf{R}^n \to \mathcal{S}$.

Claim. For all $x, x' \in \mathbf{R}^n$, $\kappa(x,x') = \sum_J \phi(x)_J \phi(x')_J$.

Indeed, the sum on the right hand side can be expressed as $e^{-\frac{\|x\|^2}{2\sigma^2}} e^{-\frac{\|x'\|^2}{2\sigma^2}} \sum_{k\geqslant 0} \frac{1}{k!} \frac{1}{\sigma^{2k}} \sum_{|J|=k} x_J x'_J$. But $\sum_{|J|=k} x_J x'_J = (x \cdot x')^k$ and therefore $\sum_{k\geqslant 0} \frac{1}{k!} \frac{1}{\sigma^{2k}} \sum_{|J|=k} x_J x'_J = \sum_{k\geqslant 0} \frac{1}{k!} \left(\frac{x\cdot x'}{\sigma^2}\right)^k = e^{\frac{x\cdot x'}{\sigma^2}}$. Finally, $e^{-\frac{\|x\|^2}{2\sigma^2}} e^{-\frac{\|x'\|^2}{2\sigma^2}} e^{\frac{x\cdot x'}{\sigma^2}} = e^{\frac{1}{2\sigma^2}(-\|x\|^2 - \|x\|^2 + 2x\cdot x')} = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$.

Remark. The proof shows that $\phi(x) \in \ell^2$, so that actually $\phi : \mathbf{R}^n \to \ell^2 \subset \mathcal{S}$.

Learning with kernels

A basic scheme
The representer theorem
The kernel trick
Kernel SVM (hard and soft)

The basic scheme to approach a learning task on $\mathcal{X}$ by means of kernels κ can be described as follows:

(1) Choose a *feature space* $\mathcal{H}$ and a *feature map* $\phi : \mathcal{X} \to \mathcal{H}$.

(2) Morph the labeled dataset $\mathcal{D} = \{(x^1, y^1), \cdots, (x^m, y^m)\}$ to the dataset $\bar{\mathcal{D}} = \{(\phi(x^1), y^1), \cdots, (\phi(x^m), y^m)\}$.

(3) Train a (linear) predictor $\bar{h} : \mathcal{H} \to \mathcal{Y}$ over $\bar{\mathcal{D}}$.

(4) Return the predictor $\hat{h} : \mathcal{X} \to \mathcal{Y}$ such that $\hat{h}(x) = \bar{h}(\phi(x))$.

Example. $\mathcal{X} = \mathbf{R}$. *Task*: learning a polynomial $p(t) = \sum_{j=0}^{d} w_j t^j$, $t \in \mathbf{R}$, of degree d from a labeled dataset $\mathcal{D} = \{(t_1, p_1), \cdots, (t_m, p_m)\}$.

Let $\mathcal{H} = \mathbf{R}^{d+1}$ and $\phi(t) = (1, t, \cdots, t^d)$. Learning $p(t)$ is morphed into learning the linear map $\mathcal{H} \to \mathbf{R}$, $x \mapsto w \cdot x$ $(w = (w_0, w_1, \cdots, w_d))$ from the dataset $\bar{\mathcal{D}} = \{(\phi(t_1), p_1), \cdots, (\phi(t_m), p_m)\}$.

Theorem. Let κ be a kernel on $\mathcal{X}$, $\mathcal{H}$ its corresponding RKHS. Let $x^1, \cdots, x^m \in \mathcal{X}$, set $h_i = \phi(x^i) = \kappa_{x^i}$, and $H = \langle h_1, \cdots, h_m \rangle$. Let $\sigma : \mathbf{R} \to \mathbf{R}$ be a non-decreasing function and $L : \mathbf{R}^m \to \mathbf{R} \sqcup \{\infty\}$ an arbitray function. Then the optimization problem

$$\min_{h \in \mathcal{H}} \bar{L}(h) = \sigma(\|h\|) + L(h(x^1), \cdots, h(x^m)) \tag{2}$$

admits a solution of the form $\hat{h} = \sum_{i=1}^m \lambda_i h_i$, $\lambda_1, \cdots, \lambda_m \in \mathbf{R}$. Moreover, if σ is increasing, then any solution has this form.

Proof. For any $h \in \mathcal{H}$, we have a unique decomposition $h = \hat{h} + h'$, with $\hat{h} \in H$ and $h' \in H^\perp$. Then $h(x^i) = \langle h, h_i \rangle = \langle \hat{h}, h_i \rangle = \hat{h}(x^i)$ and hence $L(h(x^1), \cdots, h(x^m)) = L(\hat{h}(x^1), \cdots, \hat{h}(x^m))$. Since σ is non-decreasing, $\sigma(\|\hat{h}\|) \leqslant \sigma(\sqrt{\|\hat{h}\|^2 + \|h'\|^2}) = \sigma(\|h\|)$. We conclude that $\bar{L}(\hat{h}) \leqslant \bar{L}(h)$, and this implies the first part of the theorem.

For the second part, note that if $h' \neq 0$, then $\bar{L}(\hat{h}) < \bar{L}(h)$. So h cannot be a solution unless $h = \hat{h}$. □

In terms of the expression $\hat{h} = \sum_{i=1}^m \lambda_i h_i = \sum_{i=1}^m \lambda_i \kappa_{x^i}$ for minimizers of (2), we have $\hat{h}(x^j) = \sum_{i=1}^m \lambda_i \kappa_{x^i}(x^j) = \sum_{i=1}^m \lambda_i \kappa(x^i, x^j)$. Similarly, $\|\hat{h}\|^2 = \sum_{i,j=1}^m \lambda_i \lambda_j \kappa(x^i, x^j)$ Consequently, solving (2) is equivalent to solving

$$\min_{\boldsymbol{\lambda}} L\left(\sum_{i=1}^m \lambda_i \kappa(x^i, x^1), \cdots, \sum_{i=1}^m \lambda_i \kappa(x^i, x^m)\right) + \sigma\left(\sqrt{\sum_{i,j=1}^m \lambda_i \lambda_j \kappa(x^i, x^j)}\right). \tag{3}$$

Aside from L and σ, this problem *only involves knowledge of the kernel matrix* K, and a solution $\boldsymbol{\lambda}$ yields the predictor

$$\bar{h}(x) = \sum_{j=1}^m \lambda_j \kappa(x^j, x),$$

which is optimal, given the data. It is a weighted sum of the functions $\kappa_{x^j}(x)$, which again only involve knowledge of the kernel.

Recall what we saw on 10-05, part 2, pp. 43 and 44:

The optimization problem was

$$\min_{w,b} \tfrac{1}{2}\|w\|^2 \text{ s.t. } y^j(w \cdot x^j + b) \geqslant 1,\ j \in [m].$$

The constraints $g_j(w,b) = 1 - y^j(w \cdot x^j + b) \leqslant 0$ are affine in w, b, hence qualified. So the problem has a *unique solution*, which was found as an application of KKT theorem to the Lagrangian

$$L(w,b,u) = \tfrac{1}{2}\|w\|^2 - \sum_{j\in[m]} u_j \left(y^j(w \cdot x^j + b) - 1\right):$$

(a) $w = \sum_{j\in[m]} u_j y^j x^j$.

(b) $\sum_{j\in[m]} u_j y^j = 0$.

(c) For all $j \in [m]$, $u_j = 0 \vee y^j(w \cdot x^j + b) = 1$.

The solution w is a *linear combination* of the x^j such that $u_j \neq 0$. They belong to the *marginal hyperplanes*, as $w \cdot x^j + b = y^j = \pm 1$, and are called *support vectors*. They suffice to construct the maximal-margin hyperplane, as we also have, for any support vector x^k, $b = y^k - w \cdot x^k = y^k - \sum_{j=1}^m u_j y^j (x^j \cdot x^k)$.

Duality. On plucking the expression $w = \sum_{j\in[m]} u_j y^j x^j$ into the Lagrangian we find, after some algebra with the KKT conclusions, that

$$L = \sum_{j\in[m]} u_j - \tfrac{1}{2} \sum_{j,j'} u_j u_{j'} y^j y^{j'} (x^j \cdot x^{j'})$$

Thus the dual problem, which we know is equivalent to the primal problem, is:

$$d^* = \max_{u\in\mathbf{R}^m_+} \left(\sum_{j\in[m]} u_j - \tfrac{1}{2} \sum_{j,j'} u_j u_{j'} y^j y^{j'} (x^j \cdot x^{j'}) \right)$$
$$\curlywedge\ u \geqslant 0 \wedge \sum_{j\in[m]} u_j y^j = 0.$$

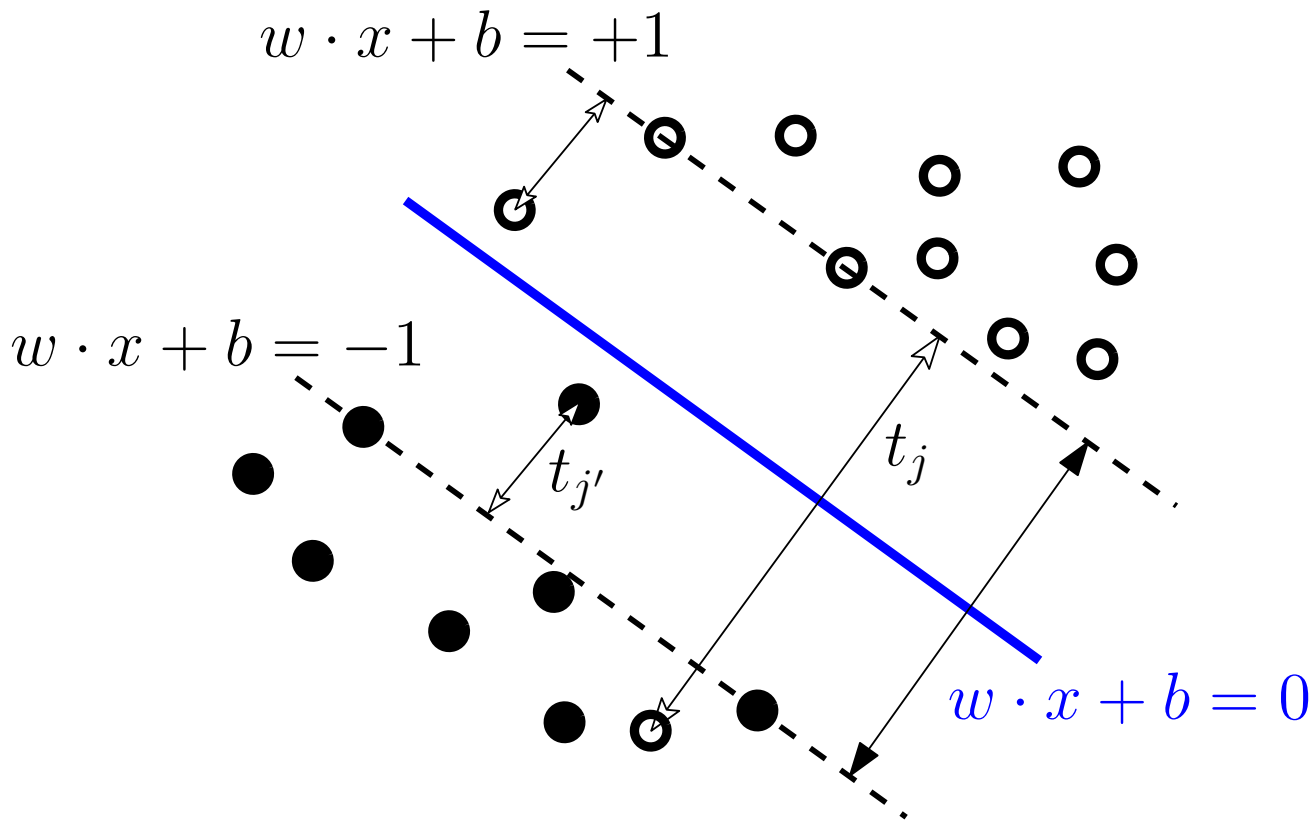
If the data are not linearly separable, it *cannot be assumed* that $y^j(w \cdot x^j + b) \geqslant 1$ for all $j \in [m]$.

In view of applying optimization tools, introduce non-negative *slack variables* $t_1, \ldots, t_n$ and consider the relaxed constraints $y^j(w \cdot x^j + b) \geqslant 1 - t_j$.

In this situation, a convenient function to be minimized is

$$\tfrac{1}{2}\|w\|^2 + \lambda \sum_{j=1}^{m} t_j,$$

where $\lambda \in \mathbf{R}_+$ is a constant that balances the norm $\|w\|$, and hence the margin $1/\|w\|$, and the total slack $t_1 + \cdots + t_m$.



The solution $w \cdot x + b = 0$ defines a *margin* $1/\|w\|$ and the *margin hyperplanes* $w \cdot x + b = +1$ and $w \cdot x + b = -1$. Points x^j satisfying the *hard* constraints $y^j(w \cdot x^j + b) \geqslant 1$ are correctly classified. Otherwise they are *outliers*. Outliers satisfying $0 < y^j(w \cdot x^j + b) < 1$ are correctly classified, but they lie inside the *margin ribbon* (the corresponding slack variable is less than the margin).

The slack, or soft SVM problem is equivalent to the optimization

$$\min_{w,b,t}\left(\tfrac{1}{2}\|w\|^2+\lambda\sum_{j=1}^m t_j\right) \text{ s.t. } y^j(w\cdot x^j+b)\geqslant 1-t_j,\ t_j\geqslant 0, j\in[m],$$

where $t=t_1,\cdots,t_m$.

There are two sets of Lagrange variables: $u=u_1,\cdots,u_m$ for the slack constraints, and $s=s_1,\cdots,s_m$ for the $t_j\geqslant 0$. This leads to the Lagrangian

$$L(w,b,t;u,s)=\frac{1}{2}\|w\|^2+\lambda\sum_{j=1}^m t_j - \sum_{j=1}^m u_j[y^j(w\cdot x^j+b)-1+t_j]-\sum_{j=1}^m s_j t_j.$$

Theorem. The solution of the soft SVM is given by the following equations:

$$\nabla_w L = w - \sum_{j=1}^m u_j y^j x^j = 0 \Rightarrow (1)\ w = \sum_{j=1}^m u_j y^j x^j.$$

$$\nabla_b L = -\sum_{j=1}^m u_j y^j = 0 \Rightarrow (2)\ \sum_{j=1}^m u_j y^j = 0$$

$$\nabla_{t_j} L = \lambda - u_j - s_j = 0 \Rightarrow (3)\ u_j + s_j = \lambda$$

$$u_j[y^j(w \cdot x^j + b) - 1 + t_j] = 0 \Rightarrow (4)\ u_j = 0 \vee y^j(w \cdot x^j + b) = 1 - t_j$$

$$s_j t_j = 0 \Rightarrow (5)\ s_j = 0 \vee t_j = 0$$

Proof. A straightfoward application of the KKT theorem. □

In the expression (1) of w as a linear combination of $x^1, \cdots, x^m$, only the tems with $u_j \neq 0$ matter (*support vectors*) and by (4) these terms satisfy $y^j(w \cdot x^j + b) = 1 - t_j$. If $t_j = 0$, then $y^j(w \cdot x^j + b) = 1$ and x^j lies on the corresponding marginal plane, as in the hard SVM. If $t_j > 0$, then x^j is an outlier. In this case (5) says that $s_j = 0$ and by (3) $u_j = \lambda$. In other words, a support vector x^j is either an outlier, in which case $u_j = \lambda$, or lies on the corresponding marginal hyperplane.

Theorem. The dual of the soft SVM is equivalent to the optimization problem

$$\max_u \sum_{j=1}^m u_j - \frac{1}{2}\sum_{i,j=1}^m u_i u_j y^i y^j (x^i \cdot x^j)$$
$$\text{s.t. } 0 \leqslant u_j \leqslant \lambda \ (j \in [m]), \sum_{j=1}^m u_j y^j = 0.$$

Proof. Plugging the expression $w = \sum_{j=1}^m u_j y^j x^j$ into the Lagrangian, together with algebraic manipulations using the other four KKT equations, we get the same dual Lagrangian as for the hard case,

$$\sum_{j=1}^m u_j - \frac{1}{2}\sum_{i,j=1}^m u_i u_j y^i y^j (x^i \cdot x^j)$$

but in addtion to $u_j \geqslant 0$, we also have to take into account the condition $s_j \geqslant 0$, which is equivalent (by $u_j + s_j = \lambda$) to $u_j \leqslant \lambda$. □

Since in the SVM equations the data $x^1, \cdots, x^m$ only appear in inner products $x^i \cdot x^j$, the kernel trick allows us to replace these inner products by any kernel on $\mathcal{X}$, now *not necessarily of vector type*.

For example, the kernel version of the dual soft SVM is the problem

$$\begin{array}{l} \max_u \sum_{j=1}^m u_j - \frac{1}{2} \sum_{i,j=1}^m u_i u_j y^i y^j \kappa(x^i, x^j) \\ \text{s.t. } 0 \leqslant u_j \leqslant \lambda \ (j \in [m]), \sum_{j=1}^m u_j y^j = 0. \end{array}$$

Implicitely, we run a SVM algorithm (hard or soft) in the feature space of κ.

This algorithm delivers a hyperplane separator in the feature space, which in the original set $\mathcal{X}$ becomes a highly *non linear decision rule*, or *decision boundary*.

The kernel trick guarantees that no operations are performed in the feature space, but only in the original set $\mathcal{X}$.

Remark. The analysis of the SVMs presented above has not applied the kernel trick theorem (representer theorem) to derive that the predictors are linear combinations of the functions $\kappa(x^j, x)$, $j \in [m]$, nor to establish the optimization problem by which they are governed.

Instead, we have followed the (explicit) path offered by the KKT theory.

But we could of course follow the first approach.

In the case of the slack SVM, for example, the σ in the representer theorem would be $\sigma(t) = \frac{1}{2}t^2$, which is strictly increasing for $t \geqslant 0$ and $\sigma(\|h\|) = \frac{1}{2}\|h\|^2$. On the other hand, the functon $L(h(x^1), \cdots, h(x^m))$ would be $\sum_{j=1}^m \max(0, 1 - y^j h(x^j))$ (the *hinge loss*).

For details, including sharp prescriptions of the algorithmic computations, see [6, § 15.2].

[4] —*Kernel methods for pattern analysis*:

- Canonical Correlation Analysis, CCA (p. 169 and § 6.5);
- kernel graph (p. 305);
- kernel perceptron
  (see also WP, Kernel_perceptron and Kernel_method)
- Kernel PCA (p. 150) —there is also a kernel SVD;
- Theorem 7.30 (p. 222) —a bound on the generalization error.

- [17]: *The journey of graph kernels through two decades*

Google Colabs

(In collaboration with Eduardo U. Moya, *WiP*)

- PCA
- Linear regression
- Polynomial regression
- Kernel learning

References I

[1] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the AMS*, vol. 68, no. 3, pp. 337–404, 1950.

[2] L. Schwartz, "Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants)," *Journal d'analyse mathématique*, vol. 13, no. 1, pp. 115–256, 1964.

[3] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*.
MIT Press, 2002.
xviii + 626 pp.

[4] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*.
Cambridge university press, 2004.

References II

[5] S. Haykin, "Neural networks and learning machines," 2009.

xxx + 906 pp.

[6] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*.

Cambridge university press, 2014.

440 pp.

[7] J. H. Manton and P.-O. Amblard, "A primer on reproducing kernel Hilbert spaces," *Foundations and Trends® in Signal Processing*, 2015.

arXiv:1408.0952, iii + 126 pp.

[8] S. Saitoh and Y. Sawano, *Theory of reproducing kernels and applications*.

Springer, 2016.

xviii + 452 pp.

References III

[9] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*.
MIT press, 2018.
xvi + 486 pp.

[10] G. Rebala, A. Ravi, and S. Churiwala, *An Introduction to Machine Learning*.
Springer, 2019.
xxii + 263 pp.

[11] E. Alpaydin, *Introduction to machine learning (fourth edition)*.
Adaptive computation and machine learning, MIT press, 2020.
xxiv + 682 pp. 1st edition: 2004; 2nd, 2010; 3rd, 2014.

References IV

[12] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "Reproducing Kernel Hilbert Space, Mercer's Theorem, Eigenfunctions, Nyström Method, and Use of Kernels in Machine Learning: Tutorial and Survey," 2021.

arXiv:2106.08443.

[13] J. B. Conway, *A course in functional analysis (2nd edition)*, vol. 96 of *Graduate Texts in Mathematics*.

Springer, 2007.

[14] H. L. Royden and P. Fitzpatrick, *Real analysis (4th edition)*.

Printice-Hall, 2010.

[15] E. Hernández and G. Weiss, *A first course on wavelets*.

CRC press, 1996.

References V

[16] G. Strang, *Linear algebra and learning from data*.
Wellesley-Cambridge Press, 2019.

[17] S. Ghosh, N. Das, T. Gonçalves, P. Quaresma, and M. Kundu, "The journey of graph kernels through two decades," *Computer Science Review*, vol. 27, pp. 88–111, 2018.
Elsevier.