Kernel Methods

Clément Lalanne

Objectives of this lecture

- Turn a measure of **similarity** between data points into a learning algorithm
- Prove the **representer theorem**, which reduces an infinite-dimensional problem to n unknowns, and read off the **prediction function** it leaves us with
- Characterise the functions that hide a Hilbert structure, the **positive semi-definite kernels**, in both directions, the second being **Aronszajn's theorem**
- Collect the **rules that build new kernels**, and apply them to the polynomial and Gaussian kernels
- Derive the **support vector machine**, its dual, and its kernel version

Notation

- $\mathcal{X}$ is the set the inputs live in. No structure is assumed on it.
- $\mathcal{H}$ is a real Hilbert space, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\| \cdot \|_{\mathcal{H}}$. A **feature map** is a map $\varphi : \mathcal{X} \to \mathcal{H}$.
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denotes a kernel, $K := (k(x_i, x_j))_{i,j}$ the **Gram matrix** of the sample, and $\mathcal{S}_n^+(\mathbb{R})$ the set of symmetric positive semi-definite matrices.
- $S := ((x_1, y_1), \dots, (x_n, y_n))$ is the training sample and $\widehat{R}_S(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$ the empirical risk.
- Anything sample-dependent carries the sample: $\widehat{R}_S$, the minimizer $\hat{\theta}_S$, its coordinates $\hat{\alpha}_S$. The dependence on the regularisation parameter $\lambda > 0$ is left implicit.
- **Abbreviation.** For $\theta \in \mathcal{H}$ we write $u(\theta) := (\langle \varphi(x_1), \theta \rangle_{\mathcal{H}}, \dots, \langle \varphi(x_n), \theta \rangle_{\mathcal{H}}) \in \mathbb{R}^n$, the vector of the n predictions on the training inputs.

I. Similarities

Similarity, the heuristic

Idea. When we can measure how similar two inputs are, kernels turn linear algorithms into non-linear ones. If (x, y) is in the training set and a new x' resembles x, we want to predict a y' that resembles y.

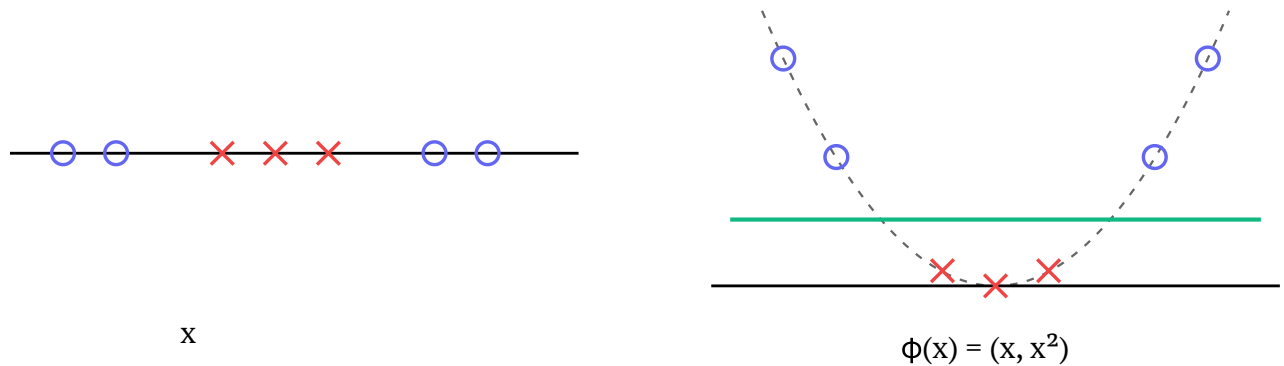
Examples. On $\mathcal{X} \subseteq \mathbb{R}^d$, $s(x, x') = -\|x - x'\|$ or $s(x, x') = x^\top x'$: the larger $s(x, x')$, the more x and x' resemble each other. What "resemble" means depends on the task.

This is a heuristic and nothing more. There is no theorem about similarity functions, and no further use will be made of them. The rest of the lecture isolates the sub-class that is a mathematical object, the positive semi-definite kernels. Keep one sentence from here: kernels retain a linear structure **implicitly**, and we never compute in it.

II. Motivation and the representer theorem

Raising the dimension

Idea. Many problems can be solved linearly by raising the dimension, possibly all the way to infinite dimension.



On the left no threshold separates the two classes. On the right, after the feature map $\varphi(x) = (x, x^2)$, a linear rule does. The price is the dimension, and we will let it become infinite.

The model in infinite dimension

Let $\mathcal{H}$ be a Hilbert space and $\varphi : \mathcal{X} \to \mathcal{H}$ a feature map. We predict with $f_\theta(x) = \langle \varphi(x), \theta \rangle_{\mathcal{H}}$ and we minimize the regularised empirical risk over the whole of $\mathcal{H}$:

$$\hat{\theta}_S \in \arg\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle \varphi(x_i), \theta \rangle_{\mathcal{H}}) + \frac{\lambda}{2} \|\theta\|_{\mathcal{H}}^2$$

This is the regularised empirical risk minimization of the previous lectures, with $\Theta = \mathcal{H}$. Two obstacles: we cannot run a descent algorithm on an infinite number of coordinates, and in general we cannot even write $\varphi(x)$ down.

Representer theorem

The objective will see θ through two things only: the vector of the n predictions it makes on the training inputs,

$$u(\theta) := (\langle \varphi(x_1), \theta \rangle_{\mathcal{H}}, \dots, \langle \varphi(x_n), \theta \rangle_{\mathcal{H}}) = (f_\theta(x_1), \dots, f_\theta(x_n)) \in \mathbb{R}^n,$$

and its norm $\|\theta\|_{\mathcal{H}}^2$.

Theorem. Let $\Psi : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}$ be non-decreasing in its last argument, and let $\mathcal{H}_D := \operatorname{span}\{\varphi(x_1), \dots, \varphi(x_n)\}$. Then

$$\inf_{\theta \in \mathcal{H}} \Psi(u(\theta), \|\theta\|_{\mathcal{H}}^2) = \inf_{\theta \in \mathcal{H}_D} \Psi(u(\theta), \|\theta\|_{\mathcal{H}}^2)$$

If the infimum is attained, it is attained at some $\hat{\theta}_S \in \mathcal{H}_D$. If Ψ is strictly increasing in its last argument, then **every** minimizer lies in $\mathcal{H}_D$.

The role of u

$u(\theta)$ is the vector of the n predictions on the training inputs, and Ψ is an arbitrary function of those predictions and of the norm. Writing the objective as $\Psi(u(\theta), \|\theta\|_{\mathcal{H}}^2)$ is the statement that **nothing else about θ is used**, and that is the only hypothesis the theorem makes.

Every regularised empirical risk is of this form. For any loss ℓ and any $\lambda \geq 0$,

$$\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) + \frac{\lambda}{2}\|\theta\|_{\mathcal{H}}^2 = \Psi(u(\theta), \|\theta\|_{\mathcal{H}}^2), \qquad \Psi(u, t) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, u_i) + \frac{\lambda}{2} t,$$

and Ψ is increasing in t as soon as $\lambda > 0$.

Two instances. Logistic, $\Psi(u,t) = \frac{1}{n}\sum_i \ell_{\text{logistic}}(y_i u_i) + \frac{\lambda}{2}t$. Hinge, $\Psi(u,t) = \frac{1}{n}\sum_i (1 - y_i u_i)_+ + \frac{\lambda}{2}t$, which is the support vector machine of Part VI.

What would break it. An objective looking at θ in any other way, for instance through $\langle \varphi(x_{n+1}), \theta \rangle_{\mathcal{H}}$ at a point outside the sample, or a penalty that is not a function of $\|\theta\|_{\mathcal{H}}$. Monotonicity in t is what lets the orthogonal component be discarded.

Proof

Write $\theta = \theta_D + \theta_D^\perp$ with $\theta_D \in \mathcal{H}_D$ and $\theta_D^\perp \in \mathcal{H}_D^\perp$. Then

$$\langle \varphi(x_i), \theta \rangle_{\mathcal{H}} = \langle \varphi(x_i), \theta_D \rangle_{\mathcal{H}} \qquad \varphi(x_i) \in \mathcal{H}_D$$

$$\|\theta\|_{\mathcal{H}}^2 = \|\theta_D\|_{\mathcal{H}}^2 + \|\theta_D^\perp\|_{\mathcal{H}}^2 \quad \text{Pythagoras}$$

so that, using the monotonicity of Ψ in its last argument,

$$\Psi(u(\theta), \|\theta\|_{\mathcal{H}}^2) = \Psi(u(\theta_D), \|\theta_D\|_{\mathcal{H}}^2 + \|\theta_D^\perp\|_{\mathcal{H}}^2) \geq \Psi(u(\theta_D), \|\theta_D\|_{\mathcal{H}}^2)$$

Every $\theta \in \mathcal{H}$ is beaten by its projection $\theta_D \in \mathcal{H}_D$, hence $\inf_{\mathcal{H}} \Psi \geq \inf_{\mathcal{H}_D} \Psi$. The reverse inequality holds because $\mathcal{H}_D \subset \mathcal{H}$, so the two infima are equal.

If Ψ is strictly increasing in its last argument, the inequality above is strict as soon as $\theta_D^\perp \neq 0$, so no minimizer can have a component orthogonal to $\mathcal{H}_D$. ■

Remark. The theorem lets us look for the solution in a space of dimension at most n, and this is what makes the theoretical method implementable.

Reformulation of the problem

By the theorem we may set $\theta = \sum_{j=1}^n \alpha_j \varphi(x_j)$, a change of variables from $\theta \in \mathcal{H}$ to $\alpha \in \mathbb{R}^n$. Then

$$\langle \varphi(x_i), \theta \rangle_{\mathcal{H}} = \sum_{j=1}^n \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}}, \quad \|\theta\|_{\mathcal{H}}^2 = \sum_{i,j=1}^n \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}}$$

so that, with $K = (\langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}})_{i,j}$,

$$\hat{\alpha}_S \in \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K \alpha$$

Predicting on a new point

Once $\hat{\alpha}_S$ is computed, $\hat{\theta}_S = \sum_{j=1}^{n} \hat{\alpha}_{S,j} \varphi(x_j)$, so the prediction at **any** $x \in \mathcal{X}$ is

$$f_{\hat{\theta}_S}(x) = \langle \varphi(x), \hat{\theta}_S \rangle_{\mathcal{H}} = \sum_{j=1}^{n} \hat{\alpha}_{S,j} \langle \varphi(x), \varphi(x_j) \rangle_{\mathcal{H}} = \sum_{j=1}^{n} \hat{\alpha}_{S,j} k(x, x_j).$$

The trained model is the training set plus n numbers. There is no θ to store. To predict, keep $x_1, \dots, x_n$ and $\hat{\alpha}_S \in \mathbb{R}^n$, and evaluate k at n pairs.

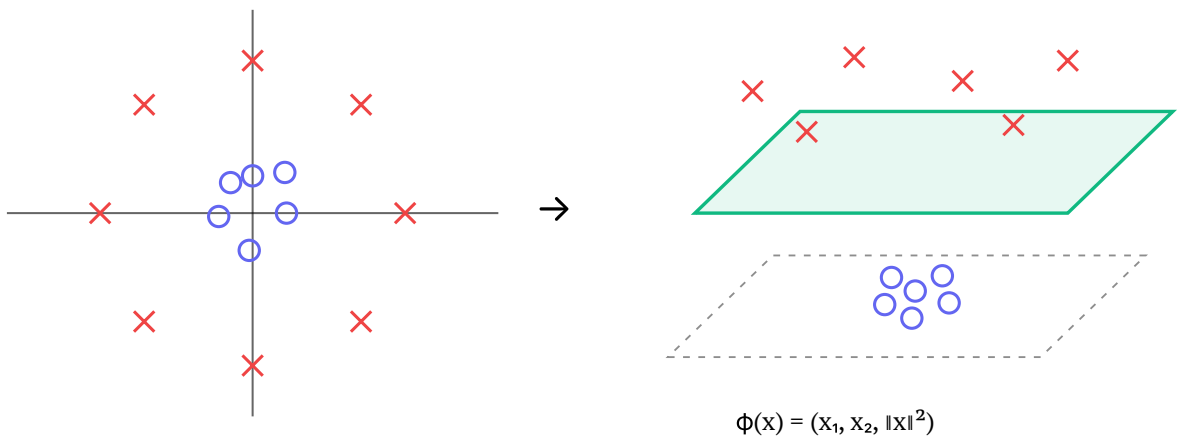
Training and prediction need the same object at two sizes. Training needs the $n \times n$ Gram matrix $K = (k(x_i, x_j))_{i,j}$. Prediction at x needs the vector $k_x := (k(x, x_1), \dots, k(x, x_n)) \in \mathbb{R}^n$, and then $f_{\hat{\theta}_S}(x) = \hat{\alpha}_S^\top k_x$.

The cost is the price of the method. Each prediction costs n kernel evaluations, so it grows with the training set, which a linear model does not. Part VI is the case where many $\hat{\alpha}_{S,j}$ vanish and only the support vectors are kept.

What is actually needed

Look at what the last two slides actually used. Training used the n^2 numbers $\langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}}$, and prediction at x used the n numbers $\langle \varphi(x), \varphi(x_j) \rangle_{\mathcal{H}}$. The feature map is never evaluated and the space $\mathcal{H}$ is never visited.

This turns the question around. Instead of choosing $\mathcal{H}$ and φ, choose a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ directly, and ask which functions are of the form $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$.



III. Positive semi-definite kernels

Definition

Definition. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We say that k is a

positive semi-definite kernel if

$$\forall x, x' \in \mathcal{X}, \quad k(x, x') = k(x', x) \qquad \text{(symmetry)}$$

$$\forall n \geq 1,\ \forall x_1, \dots, x_n \in \mathcal{X},\ \forall \alpha_1, \dots, \alpha_n \in \mathbb{R}, \quad \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \quad \text{(positivity)}$$

Remark. The positivity constraint says exactly that $\big(k(x_i, x_j)\big)_{i,j} \in \mathcal{S}_n^+(\mathbb{R})$ for every finite family of points.

Inner products give kernels

Proposition. If $\mathcal{H}$ is a Hilbert space and $\varphi : \mathcal{X} \to \mathcal{H}$, then $k(x, x') := \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ is a positive semi-definite kernel.

Proof. Symmetry is the symmetry of the inner product. For the positivity, let $n \geq 1$, $x_1, \ldots, x_n \in \mathcal{X}$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$:

$$\begin{aligned}
\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) &= \sum_{i,j=1}^{n} \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^{n} \alpha_i \varphi(x_i), \sum_{j=1}^{n} \alpha_j \varphi(x_j) \right\rangle_{\mathcal{H}} \qquad \text{bilinearity} \\
&= \left\| \sum_{i=1}^{n} \alpha_i \varphi(x_i) \right\|_{\mathcal{H}}^2 \geq 0
\end{aligned}$$

■

Aronszajn's theorem

Theorem (Aronszajn). If $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semi-definite kernel, then there exist a Hilbert space $\mathcal{H}_k$ and a map $\varphi_k : \mathcal{X} \to \mathcal{H}_k$ such that

$$\forall x, x' \in \mathcal{X}, \quad k(x, x') = \langle \varphi_k(x), \varphi_k(x') \rangle_{\mathcal{H}_k}$$

With the proposition of the previous slide, this is an equivalence: **positive semi-definite kernels are exactly the inner products of feature maps**. Choosing a kernel and choosing a pair $(\mathcal{H}, \varphi)$ are the same act, and only the kernel has to be computed.

The proof occupies the next nine slides. It follows Jean-Philippe Vert.

Proof, 1) a pre-Hilbert space

For $x \in \mathcal{X}$ write $k_x := k(x, \cdot) : \mathcal{X} \to \mathbb{R}$, and consider the space of functions $\mathcal{H}_0 := \operatorname{span}\{k_x ; x \in \mathcal{X}\}$, equipped with

$$\left\langle \sum_{i=1}^{m} a_i k_{x_i}, \sum_{j=1}^{m'} b_j k_{z_j} \right\rangle_{\mathcal{H}_0} := \sum_{i=1}^{m} \sum_{j=1}^{m'} a_i b_j k(x_i, z_j)$$

This is well defined: the right-hand side equals $\sum_j b_j f(z_j)$ where $f = \sum_i a_i k_{x_i}$, so it does not depend on the way f is written, and by symmetry it does not depend on the way $g = \sum_j b_j k_{z_j}$ is written either. It is symmetric, bilinear, and positive because k is positive semi-definite.

Proof, 1) the form is definite

Reproducing property. By definition of the bracket on the generators,

$$\forall f \in \mathcal{H}_0,\ \forall x \in \mathcal{X}, \quad f(x) = \langle f, k_x \rangle_{\mathcal{H}_0} \quad (1)$$

Cauchy-Schwarz holds for any symmetric positive bilinear form, definiteness is not needed, so

$$\forall f \in \mathcal{H}_0,\ \forall x \in \mathcal{X}, \quad |f(x)| \leq \sqrt{\langle f, f \rangle_{\mathcal{H}_0}} \sqrt{k(x,x)} \quad (2)$$

Hence $\|f\|_{\mathcal{H}_0} = 0$ forces $f(x) = 0$ for every x, that is $f = 0$. So $(\mathcal{H}_0, \langle \cdot, \cdot \rangle_{\mathcal{H}_0})$ is pre-Hilbert.

Proof, 2) completion, the space

Let $(f_m)_m$ be a Cauchy sequence in $\mathcal{H}_0$. Applying (2) to $f_m - f_{m'}$ shows that for every $x \in \mathcal{X}$ the real sequence $(f_m(x))_m$ is Cauchy, hence converges.

$$\mathcal{H} := \{\text{pointwise limits of Cauchy sequences of } \mathcal{H}_0\}$$

This is a vector space of functions on $\mathcal{X}$, and it contains $\mathcal{H}_0$ through constant sequences. It remains to equip it with a Hilbert structure extending $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$.

Proof, 2) completion, the inner product

Let (f_m) and (g_m) be Cauchy in $\mathcal{H}_0$. By Cauchy-Schwarz,

$$|\langle f_m, g_m\rangle_{\mathcal{H}_0} - \langle f_{m'}, g_{m'}\rangle_{\mathcal{H}_0}| \leq \underbrace{\|f_m\|_{\mathcal{H}_0}}_{\text{bounded}} \underbrace{\|g_m - g_{m'}\|_{\mathcal{H}_0}}_{\to 0} + \underbrace{\|f_m - f_{m'}\|_{\mathcal{H}_0}}_{\to 0} \underbrace{\|g_{m'}\|_{\mathcal{H}_0}}_{\text{bounded}}$$

Cauchy sequences are bounded, so the right-hand side tends to 0 as $\min(m, m') \to \infty$. The sequence $(\langle f_m, g_m\rangle_{\mathcal{H}_0})_m$ is Cauchy in $\mathbb{R}$, hence converges, and we set $\langle f, g\rangle_{\mathcal{H}} := \lim_m \langle f_m, g_m\rangle_{\mathcal{H}_0}$ for the pointwise limits f and g. It remains to check that this does not depend on the chosen sequences.

Proof, 2) a lemma

Lemma. If (f_m) is Cauchy in $\mathcal{H}_0$ and converges pointwise to 0, then $\|f_m\|_{\mathcal{H}_0} \to 0$.

Proof. Let $\epsilon > 0$, let B bound $\|f_m\|_{\mathcal{H}_0}$, and let N be such that $\|f_m - f_N\|_{\mathcal{H}_0} \le \epsilon$ for $m \ge N$. Writing $f_N = \sum_{i=1}^{p} a_i k_{x_i}$,

$$\|f_m\|_{\mathcal{H}_0}^2 = \langle f_m - f_N, f_m\rangle_{\mathcal{H}_0} + \langle f_N, f_m\rangle_{\mathcal{H}_0} \le \epsilon B + \sum_{i=1}^{p} a_i f_m(x_i)$$

using Cauchy-Schwarz and then (1). The sum tends to 0 by pointwise convergence, so $\limsup_m \|f_m\|_{\mathcal{H}_0}^2 \le \epsilon B$ for every ϵ. ∎

Proof, 2) the inner product is well defined

Let $(f_m), (f'_m)$ be Cauchy in $\mathcal{H}_0$ with the same pointwise limit f, and $(g_m), (g'_m)$ with the same pointwise limit g. Then

$$|\langle f_m, g_m\rangle_{\mathcal{H}_0} - \langle f'_m, g'_m\rangle_{\mathcal{H}_0}| \leq \|f_m\|_{\mathcal{H}_0}\|g_m - g'_m\|_{\mathcal{H}_0} + \|g'_m\|_{\mathcal{H}_0}\|f_m - f'_m\|_{\mathcal{H}_0} \xrightarrow[m\to\infty]{} 0$$

since $(g_m - g'_m)$ and $(f_m - f'_m)$ are Cauchy and converge pointwise to 0, so the lemma applies, while the other two factors are bounded.

So $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is well defined. It is symmetric, bilinear and positive by passing to the limit. It is definite: if $\|f\|_{\mathcal{H}} = 0$ then, by (2), $|f(x)| = \lim_m |f_m(x)| \leq \lim_m \|f_m\|_{\mathcal{H}_0}\sqrt{k(x,x)} = 0$ for every x.

Proof, 2) density of $\mathcal{H}_0$

Let $f \in \mathcal{H}$ and let (f_m) be Cauchy in $\mathcal{H}_0$ converging pointwise to f. For fixed m, the sequence $(f_p - f_m)_p$ is Cauchy in $\mathcal{H}_0$ and converges pointwise to $f - f_m$, so by definition of the inner product

$$\lim_{m\to\infty} \|f - f_m\|_{\mathcal{H}} = \lim_{m\to\infty} \lim_{p\to\infty} \|f_p - f_m\|_{\mathcal{H}_0} = 0$$

the last equality because (f_m) is Cauchy in $\mathcal{H}_0$.

Hence $\mathcal{H}_0$ is dense in $\mathcal{H}$, and the norm of $\mathcal{H}$ restricted to $\mathcal{H}_0$ is the norm of $\mathcal{H}_0$.

Proof, 2) completeness

Let (f_m) be Cauchy in $\mathcal{H}$. By density, pick $f'_m \in \mathcal{H}_0$ with $\|f_m - f'_m\|_{\mathcal{H}} \to 0$. Then (f'_m) is Cauchy in $\mathcal{H}_0$:

$$\|f'_m - f'_{m'}\|_{\mathcal{H}_0} = \|f'_m - f'_{m'}\|_{\mathcal{H}} \leq \|f'_m - f_m\|_{\mathcal{H}} + \|f_m - f_{m'}\|_{\mathcal{H}} + \|f_{m'} - f'_{m'}\|_{\mathcal{H}} \leq \epsilon$$

for m, m' large enough.

So (f'_m) converges pointwise to some $f \in \mathcal{H}$, and the previous slide gives $\|f - f'_m\|_{\mathcal{H}} \to 0$. Therefore $\|f - f_m\|_{\mathcal{H}} \leq \|f - f'_m\|_{\mathcal{H}} + \|f'_m - f_m\|_{\mathcal{H}} \to 0$, and $\mathcal{H}$ is complete. ■

Proof, 3) characteristic properties

- $\forall x \in \mathcal{X}, k_x \in \mathcal{H}$.
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, f(x) = \langle f, k_x \rangle_{\mathcal{H}}$. Indeed if $(f_m) \subset \mathcal{H}_0$ is Cauchy and tends pointwise to f, then $f(x) = \lim_m f_m(x) = \lim_m \langle f_m, k_x \rangle_{\mathcal{H}_0} = \langle f, k_x \rangle_{\mathcal{H}}$.
- $\forall x, x' \in \mathcal{X}, k(x, x') = \langle k_x, k_{x'} \rangle_{\mathcal{H}}$.

The last line is Aronszajn's theorem, with $\mathcal{H}_k = \mathcal{H}$ and $\varphi_k : x \mapsto k_x$. ■

$\mathcal{H}$ is called the **reproducing kernel Hilbert space** of k, after the second property. Its elements are functions on $\mathcal{X}$, and evaluation at x is an inner product against k_x.

What the equivalence buys

Positive semi-definite kernels and pairs $(\mathcal{H}, \varphi)$ are the same objects. In one direction a feature map gives a kernel, in the other Aronszajn manufactures a feature map from a kernel. We may therefore design k and forget φ entirely.

The recipe, end to end. Choose a positive semi-definite k. Form $K = (k(x_i, x_j))_{i,j}$, solve $\hat{\alpha}_S \in \arg\min_\alpha \frac{1}{n}\sum_i \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2}\alpha^\top K \alpha$, and predict with $f(x) = \sum_j \hat{\alpha}_{S,j} k(x, x_j)$. The Hilbert space appears in no line of that.

The predictor lives in $\mathcal{H}_k$. With $\varphi_k(x) = k_x = k(x, \cdot)$, the prediction function is $f = \sum_j \hat{\alpha}_{S,j} k_{x_j}$, an element of the reproducing kernel Hilbert space, and the reproducing property $f(x) = \langle f, k_x \rangle_{\mathcal{H}_k}$ is exactly the display above.

What is left to do. Only one thing: recognise which functions k are positive semi-definite. That is the next section.

IV. Building kernels

Three examples

Linear kernel. $k(x, x') = x^\top x'$ on $\mathcal{X} = \mathbb{R}^d$. Positive semi-definite by the proposition, with $\varphi = \mathrm{id}$.

Polynomial kernel. $k(x, x') = (x^\top x')^q$ for $q \geq 1$ an integer.

Gaussian kernel. $k(x, x') = \exp(-a\|x - x'\|_2^2)$ for $a > 0$.

Only the first is visibly an inner product. The other two are the targets of this section: we build the rules first, then apply them to each.

Construction of positive semi-definite kernels

We have seen that positive semi-definite kernels are exactly the inner products of Hilbert spaces. Here are the rules that build new ones from old.

Theorem. The set of positive semi-definite kernels on $\mathcal{X}$ is stable under

- multiplication by a non-negative real number,
- finite sums,
- products,
- pointwise limits, when the limit exists,
- conjugation by a function, $(x, x') \mapsto f(x)k(x, x')f(x')$ for any $f : \mathcal{X} \to \mathbb{R}$.

The first two are immediate from the definition, since a non-negative combination of non-negative quantities is non-negative, and the fourth follows by passing to the limit in the inequality $\sum_{i,j} \alpha_i \alpha_j k_m(x_i, x_j) \geq 0$. The last is the substitution $\beta_i = \alpha_i f(x_i)$, which turns $\sum_{i,j} \alpha_i \alpha_j f(x_i) k(x_i, x_j) f(x_j)$ into $\sum_{i,j} \beta_i \beta_j k(x_i, x_j) \geq 0$. The product is the only real argument.

Proof for the product

Let k_1, k_2 be positive semi-definite, $n \geq 1$ and $x_1, \dots, x_n \in \mathcal{X}$. Write $K_1 = (k_1(x_i, x_j))_{i,j}$ and $K_2 = (k_2(x_i, x_j))_{i,j}$. Both are in $\mathcal{S}_n^+(\mathbb{R})$, so $K_1 = U_1^\top U_1$ and $K_2 = U_2^\top U_2$. Let $u_i \in \mathbb{R}^{r_1}$ and $v_i \in \mathbb{R}^{r_2}$ be the i-th columns of U_1 and U_2, so that $k_1(x_i, x_j) = u_i^\top u_j$ and $k_2(x_i, x_j) = v_i^\top v_j$.

$$
\begin{aligned}
k_1(x_i, x_j) k_2(x_i, x_j) &= (u_i^\top u_j)(v_i^\top v_j) \\
&= (u_i^\top u_j)\mathrm{tr}(v_i v_j^\top) && v_i^\top v_j = v_j^\top v_i = \mathrm{tr}(v_i v_j^\top) \\
&= \mathrm{tr}((u_i^\top u_j) v_i v_j^\top) && \text{the trace is linear} \\
&= \mathrm{tr}(v_i (u_i^\top u_j) v_j^\top) && \text{a scalar commutes with everything} \\
&= \mathrm{tr}((u_i v_i^\top)^\top (u_j v_j^\top)) && (u_i v_i^\top)^\top = v_i u_i^\top \\
&= \langle u_i v_i^\top, u_j v_j^\top \rangle_F && \langle A, B \rangle_F = \mathrm{tr}(A^\top B)
\end{aligned}
$$

The step to watch is the fourth: the brackets are dropped, and $v_i(u_i^\top u_j)v_j^\top$ and $v_i u_i^\top u_j v_j^\top$ are the same product.

Proof for the product, conclusion

Every entry of $\big(k_1(x_i,x_j)k_2(x_i,x_j)\big)_{i,j}$ is an inner product of the family $(u_i v_i^\top)_i$ in $\mathbb{R}^{r_1\times r_2}$, so that matrix is a Gram matrix for the Frobenius inner product, hence in $\mathcal{S}_n^+(\mathbb{R})$. The feature map of the product kernel is $x_i \mapsto u_i v_i^\top$. ■

The same computation without traces. Expanding both inner products in coordinates, for every $\alpha \in \mathbb{R}^n$,

$$\sum_{i,j} \alpha_i \alpha_j (u_i^\top u_j)(v_i^\top v_j) = \sum_{a\le r_1}\sum_{b\le r_2}\left(\sum_{i\le n} \alpha_i u_{ia} v_{ib}\right)^2 \ge 0,$$

where u_{ia} is the a-th coordinate of u_i. The matrix $u_i v_i^\top$ of the first proof is the array $(u_{ia}v_{ib})_{a,b}$ of the second.

Example: polynomial kernels

Claim. For every integer $q \geq 1$, both $(x, x') \mapsto (x^\top x')^q$ and $(x, x') \mapsto (1 + x^\top x')^q$ are positive semi-definite on $\mathbb{R}^d$.

Proof. The linear kernel $x^\top x'$ is positive semi-definite, by the proposition with $\varphi = \mathrm{id}$. Products are stable, so $(x^\top x')^q$ is, by induction on q. The constant kernel $(x, x') \mapsto 1$ is positive semi-definite because $\sum_{i,j} \alpha_i \alpha_j = (\sum_i \alpha_i)^2 \geq 0$, sums are stable, so $1 + x^\top x'$ is, and taking the q-th power gives the second. ■

The feature map, which we will not use. By the multinomial theorem $(x^\top x')^q = \sum_{|\beta| = q} \frac{q!}{\beta!} x^\beta x'^\beta$, so $\varphi(x)$ is the vector of all monomials of degree q in the coordinates of x, scaled by $\sqrt{q!/\beta!}$. It has $\binom{d+q-1}{q}$ entries, and we never build it.

Prediction. $f(x) = \sum_{j=1}^n \hat{\alpha}_{S,j} (x^\top x_j)^q$, a polynomial of degree q in x, obtained without writing a single monomial. The version with $1+$ contains all degrees up to q instead of exactly q, which is usually what is wanted.

Example: the exponential of a kernel

Claim. If k is positive semi-definite then so is $\exp(k)$, meaning $(x, x') \mapsto e^{k(x,x')}$.

Proof. Fix m and consider the partial sum $\sum_{p=0}^{m} \frac{k^p}{p!}$. Each k^p is a product of positive semi-definite kernels, with $k^0 \equiv 1$ the constant kernel, and the coefficients $1/p!$ are non-negative, so the partial sum is a non-negative combination of positive semi-definite kernels and is positive semi-definite. It converges pointwise to e^k as $m \to \infty$, and pointwise limits are stable. ∎

All four of the first rules were used, in that order. The next slide spends them once more.

Example: the Gaussian kernel

Claim. $k(x, x') = \exp(-a\|x - x'\|_2^2)$ is positive semi-definite on $\mathbb{R}^d$ for every $a > 0$.

Proof. Expand the square, $\|x - x'\|_2^2 = \|x\|_2^2 - 2x^\top x' + \|x'\|_2^2$, so that

$$\exp(-a\|x - x'\|_2^2) = f(x)\exp(2ax^\top x')f(x'), \qquad f(x) = \exp(-a\|x\|_2^2).$$

The kernel $2ax^\top x'$ is positive semi-definite, its exponential is by the previous slide, and conjugation by f preserves the property. ∎

Prediction. $f(x) = \sum_{j=1}^{n} \hat{\alpha}_{S,j} \exp(-a\|x - x_j\|_2^2)$, a weighted sum of bumps centred at the training points. Small a gives wide bumps and a smooth predictor, large a gives narrow ones and a predictor that is close to zero away from the data. The scale a is a hyperparameter and is chosen on the validation set.

Bochner's theorem

Theorem (Bochner). Let $q : \mathbb{R}^d \to \mathbb{R}$ be continuous. The kernel $k(x, x') := q(x - x')$ is positive semi-definite if and only if q is the inverse Fourier transform of a finite positive Borel measure ρ:

$$q(z) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{iz^\top \omega} d\rho(\omega)$$

We prove the implication that is used in practice, from the integral representation to positive semi-definiteness. The converse is admitted.

Bochner's theorem, proof of the sufficient condition

Let $n \geq 1, x_1, \dots, x_n \in \mathbb{R}^d$ and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$.

$$\begin{aligned}
\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sum_{i,j} \alpha_i \alpha_j e^{i(x_i - x_j)^\top \omega} d\rho(\omega) && \text{Fubini, } \rho \text{ finite} \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sum_{i,j} (\alpha_i e^{i x_i^\top \omega}) \overline{(\alpha_j e^{i x_j^\top \omega})} d\rho(\omega) && \alpha_j \in \mathbb{R} \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left| \sum_i \alpha_i e^{i x_i^\top \omega} \right|^2 d\rho(\omega) \geq 0 && \rho \geq 0
\end{aligned}$$

Symmetry holds as soon as q is even, which is the case when ρ is invariant under $\omega \mapsto -\omega$, and that is also what makes q real valued. ■

The Gaussian kernel, a second proof

The stability rules already gave this kernel three slides ago. Bochner gives it again and, this time, exhibits the spectral measure. Apply the theorem with the Gaussian density. For every $a > 0$ and $z \in \mathbb{R}^d$,

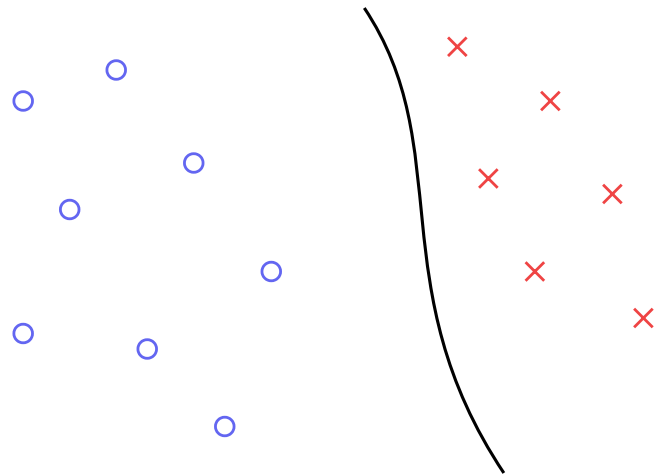
$$e^{-a\|z\|_2^2} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{iz^\top \omega} \underbrace{(\pi/a)^{d/2} e^{-\|\omega\|_2^2/(4a)}}_{\text{density of } \rho} d\omega$$

The measure ρ is positive and finite, and invariant under $\omega \mapsto -\omega$, so $k(x, x') = \exp(-a\|x - x'\|_2^2)$ is a positive semi-definite kernel.

V. Support vector machines

The objective

Separate two classes, with a boundary that need not be linear.



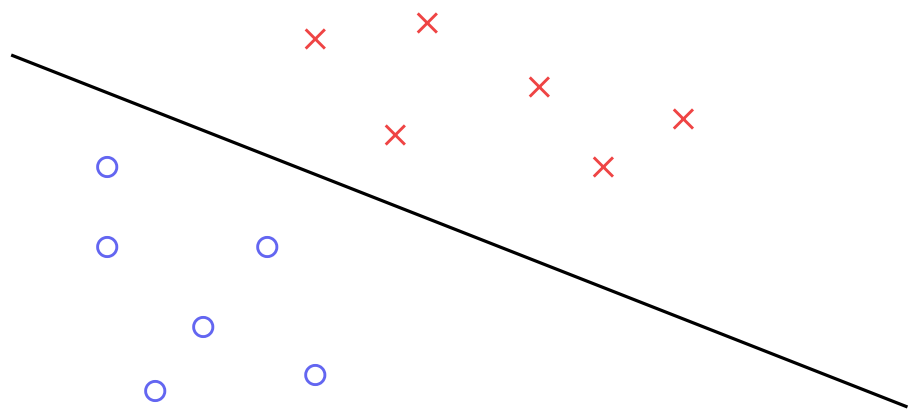
$$\hat{\theta}_S \in \arg\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \max(1 - y_i \langle \theta, \varphi(x_i) \rangle_{\mathcal{H}}, 0) + \frac{\lambda}{2} \|\theta\|_{\mathcal{H}}^2$$

Empirical risk minimization for the **hinge loss**, plus regularisation. The rest of the part builds this problem from geometry, in $\mathcal{H} = \mathbb{R}^d$ with $\varphi = \mathrm{id}$.

1) The linearly separable case

Context. n observations $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$.

Hypothesis. There exists $\theta \in \mathbb{R}^d$ such that $\forall i, y_i(\theta^\top x_i) > 0$, that is, a separating hyperplane exists.



Objective. Maximise the distance from the hyperplane to the closest point.

Distance to a hyperplane

Let $H(\theta)$ be the hyperplane of equation $\theta^\top x = 0$, for $\theta \neq 0$. Then

$$\forall x \in \mathbb{R}^d, \quad \operatorname{dist}(x, H(\theta)) = \frac{|\theta^\top x|}{\|\theta\|_2}$$

Proof. $H(\theta)^\perp = \mathbb{R}\theta$, of which $\theta / \|\theta\|_2$ is an orthonormal basis, so

$$\operatorname{dist}(x, H(\theta))^2 = \|x - P_{H(\theta)}(x)\|_2^2 = \|P_{H(\theta)^\perp}(x)\|_2^2 = (\theta^\top x / \|\theta\|_2)^2$$

■

The largest possible margin

Under the hypothesis, $y_i(\theta^\top x_i) = |\theta^\top x_i|$ for every i, so the distance from the closest point to the hyperplane is $\min_i y_i(\theta^\top x_i)/\|\theta\|_2$, and we solve

$$\max_{\theta \neq 0} \min_{i \in \{1,\ldots,n\}} \frac{y_i(\theta^\top x_i)}{\|\theta\|_2}$$

An equivalent formulation

The objective is invariant under $\theta \mapsto c\theta$ for $c > 0$, and under the hypothesis its value is positive, so we may impose the normalisation $\min_i y_i(\theta^\top x_i) = 1$. Maximising $1/\|\theta\|_2$ under that constraint is minimizing $\frac{1}{2}\|\theta\|_2^2$, and the constraint can be relaxed to an inequality since it is saturated at the optimum:

$$\min_\theta \frac{1}{2}\|\theta\|_2^2 \quad \text{s.t.} \quad \forall i, y_i(\theta^\top x_i) \geq 1$$

A convex quadratic problem with linear constraints, in place of a maximin.

2) The linearly non-separable case

If the problem is not linearly separable, that is if for every θ there is an i with $y_i(\theta^\top x_i) \leq 0$, the constraints above cannot all hold. Introduce variables measuring the sub-optimality of each constraint:

$$\min_{\theta,\xi} \frac{1}{2}\|\theta\|_2^2 + C\sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad \forall i,\ y_i(\theta^\top x_i) \geq 1 - \xi_i, \quad \forall i,\ \xi_i \geq 0$$

$C > 0$ sets the price of violating a constraint. The problem is feasible for every θ, since ξ_i can always be taken large enough.

Back to the hinge loss

At fixed θ the objective is increasing in each ξ_i, so at the optimum $\xi_i = (1 - y_i(\theta^\top x_i))_+$, the smallest value the two constraints allow. Substituting, dividing by nC and setting $\lambda := 1/(nC)$:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} (1 - y_i(\theta^\top x_i))_+ + \frac{\lambda}{2} \|\theta\|_2^2$$

The geometric problem and the regularised empirical risk minimization of the first slide of this part are the same problem, and C large means λ small, that is, little regularisation.

3) The dual problem

$$\textbf{primal:} \quad \min_{\theta,\xi} \frac{1}{2}\|\theta\|_2^2 + C\sum_{i=1}^n \xi_i \quad \text{s.t.} \quad \forall i,\ y_i\theta^\top x_i + \xi_i - 1 \geq 0, \quad \forall i,\ \xi_i \geq 0$$

a convex problem with linear constraints.

Lagrangian, with multipliers $\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n \geq 0$:

$$\mathcal{L}(\theta, \xi, \alpha, \beta) = \frac{1}{2}\|\theta\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i(y_i\theta^\top x_i + \xi_i - 1) - \sum_{i=1}^n \beta_i\xi_i$$

Minimizing the Lagrangian

In ξ. $\mathcal{L}$ is affine in each ξ_i, with coefficient $C - \alpha_i - \beta_i$. The infimum over $\xi \in \mathbb{R}^n$ is $-\infty$ unless $\alpha_i + \beta_i = C$ for every i, and in that case all the terms in ξ disappear.

In θ. $\nabla_\theta \mathcal{L} = \theta - \sum_i \alpha_i y_i x_i$, so $\theta^\star = \sum_{i=1}^n \alpha_i y_i x_i$, and

$$\frac{1}{2}\|\theta^\star\|_2^2 - \sum_i \alpha_i y_i \theta^{\star\top} x_i + \sum_i \alpha_i = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

The dual problem

Eliminating $\beta \geq 0$ through $\alpha_i + \beta_i = C$, which is exactly $\alpha_i \leq C$:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \quad \text{s.t.} \quad \forall i,\ 0 \leq \alpha_i \leq C$$

Duality. Weak duality gives $\text{Dual} \leq \text{Primal}$. The primal has strictly feasible points, for instance $\theta = 0$ and $\xi_i = 2$, so Slater's condition applies and there is strong duality: $\text{Primal} = \text{Dual}$. The dual is a quadratic program with box constraints.

Note where the data sit: only through the inner products $x_i^\top x_j$. That is the opening for part VI.

VI. The kernel version

Setting

Let k be a positive semi-definite kernel, $\mathcal{H}$ the Hilbert space given by Aronszajn's theorem and φ the associated feature map. We look for

$$\hat{\theta}_S \in \arg\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \langle \theta, \varphi(x_i) \rangle_{\mathcal{H}})_+ + \frac{\lambda}{2} \|\theta\|_{\mathcal{H}}^2$$

Warning. This is now an infinite-dimensional problem. It can no longer be solved by stochastic gradient descent on θ.

The representer theorem applies with $\Psi(u, t) = \frac{1}{n} \sum_i (1 - y_i u_i)_+ + \frac{\lambda}{2} t$, which is strictly increasing in t since $\lambda > 0$. So $\hat{\theta}_S \in \operatorname{span}\{\varphi(x_1), \dots, \varphi(x_n)\}$.

Rewriting the problem

Writing $\theta = \sum_j \alpha_j \varphi(x_j)$ and $K = (k(x_i, x_j))_{i,j}$, the problem becomes finite dimensional:

$$\hat{\alpha}_S \in \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} (1 - y_i (K\alpha)_i)_+ + \frac{\lambda}{2} \alpha^\top K \alpha$$

and the classifier is

$$x \longmapsto \operatorname{sgn}\left(\sum_{j=1}^{n} \hat{\alpha}_{S,j} k(x, x_j) \right)$$

Only evaluations of k are needed, at training and at prediction time.

Slack variables and Lagrangian

$$\min_{\alpha,\xi} \frac{1}{n}\sum_{i=1}^{n}\xi_i + \frac{\lambda}{2}\alpha^\top K\alpha \quad \text{s.t.} \quad \forall i,\ \xi_i \geq 1 - y_i(K\alpha)_i, \quad \forall i,\ \xi_i \geq 0$$

With multipliers $\mu, \nu \geq 0$, and using $\sum_i \mu_i y_i (K\alpha)_i = (\mathrm{diag}(y)\mu)^\top K\alpha$,

$$L = \frac{1}{n}\mathbf{1}^\top\xi + \frac{\lambda}{2}\alpha^\top K\alpha - (\mathrm{diag}(y)\mu)^\top K\alpha - (\mu+\nu)^\top\xi + \mu^\top\mathbf{1}$$

Minimizing the Lagrangian

In α. $\nabla_\alpha L = \lambda K\alpha - K\,\mathrm{diag}(y)\mu = 0$, so

$$\alpha = \frac{\mathrm{diag}(y)\mu}{\lambda} + \epsilon, \quad \epsilon \in \mathrm{Ker}(K) \quad (1)$$

Such an ϵ changes neither the objective nor the constraints, since both involve α only through $K\alpha$, so we take $\epsilon = 0$.

In ξ. L is affine in ξ with coefficient $\frac{1}{n}\mathbf{1} - \mu - \nu$. The infimum is $-\infty$ unless $\frac{1}{n}\mathbf{1} - \mu - \nu = 0$, and otherwise all the terms in ξ disappear.

The dual problem

Substituting (1) with $\epsilon = 0$ and eliminating $\nu \geq 0$ through $\mu + \nu = \frac{1}{n}\mathbf{1}$:

$$\max_{\mu} \sum_{i=1}^{n} \mu_i - \frac{1}{2\lambda} \sum_{i,j=1}^{n} y_i y_j \mu_i \mu_j k(x_i, x_j) \quad \text{s.t.} \quad \forall i,\ 0 \leq \mu_i \leq \frac{1}{n}$$

A quadratic program with box constraints, in which the data appear only through the Gram matrix. Strong duality holds for the same reason as before: the primal has strictly feasible points, so Slater's condition applies.

Reformulation with the primal variables

By (1) with $\epsilon = 0$ we have $\mu = \lambda \mathrm{diag}(y)\alpha$, and since $y_i^2 = 1$,

$$\sum_{i=1}^n \mu_i = \lambda y^\top \alpha, \quad \frac{1}{2\lambda} \sum_{i,j=1}^n y_i y_j \mu_i \mu_j k(x_i, x_j) = \frac{\lambda}{2} \alpha^\top K \alpha$$

Dividing the objective by $\lambda/2 > 0$, the dual problem reads

$$\max_\alpha 2\alpha^\top y - \alpha^\top K \alpha \quad \text{s.t.} \quad \forall i,\ 0 \leq y_i \alpha_i \leq \frac{1}{\lambda n}$$

The numerical resolution of this program is the subject of the lab.

Support vectors, and a sparse prediction

The box constraint $0 \le y_i \alpha_i \le \frac{1}{\lambda n}$ comes from $\mu_i \in [0, \frac{1}{n}]$, where μ_i is the multiplier of the constraint $\xi_i \ge 1 - y_i (K\alpha)_i$. Complementary slackness says $\mu_i > 0$ only if that constraint is active.

A well classified point contributes nothing. If $y_i f(x_i) > 1$, the constraint is slack, so $\mu_i = 0$, so $\alpha_i = 0$. Only the points with $y_i f(x_i) \le 1$, those on or inside the margin, have $\alpha_i \neq 0$. They are the **support vectors**.

The prediction is sparse. Writing $V = \{j : \hat{\alpha}_{S,j} \neq 0\}$,

$$f(x) = \sum_{j \in V} \hat{\alpha}_{S,j} k(x, x_j),$$

so a new point costs $|V|$ kernel evaluations rather than n, and the trained model is the support vectors together with their coefficients.

Where the sparsity comes from. The hinge loss is exactly zero for $y_i f(x_i) > 1$. A loss that is positive everywhere, such as the logistic loss, gives no zero coefficient and no sparsity, and its prediction costs the full n evaluations.

What to remember

- A positive semi-definite kernel is exactly an inner product of features: one direction is a two-line computation, the other is Aronszajn's theorem, and the two sit side by side in Part III
- The representer theorem turns a regularised problem posed in $\mathcal{H}$ into a problem over $\alpha \in \mathbb{R}^n$. Its only hypothesis is that the objective sees θ through $u(\theta)$, the n training predictions, and through $\|\theta\|_{\mathcal{H}}$
- **The prediction function is always the same**, $f(x) = \sum_{j=1}^{n} \hat{\alpha}_{S,j} k(x, x_j)$. Training needs the Gram matrix K, prediction at x needs the vector $(k(x, x_j))_j$, and φ appears in neither
- New kernels from old: non-negative combinations, products, pointwise limits, conjugation by a function. Those four rules give the polynomial kernels, $\exp(k)$ and the Gaussian kernel, and Bochner characterises the translation-invariant ones
- Maximal margin, slack variables, and hinge loss plus ridge penalty are three readings of the same problem
- The kernel SVM dual is a box-constrained quadratic program where the data enter only through k