Topics in ML, Lecture 4
Kernel Methods (I)

Lénaïc Chizat*

October 5, 2025

Ref: Chapter 7 in F. Bach "Learning Theory from First Principles" book.

1 Introduction

In today's lecture, we study Empirical Risk Minimization (ERM) for linearly-parameterized predictors

$$f_\theta(x) = \langle \theta, \phi(x) \rangle_{\mathcal{H}}$$

where $\phi : \mathcal{X} \to \mathcal{H}$ is the *feature map* and $\mathcal{H}$ is a Hilbert space (i.e. potentially infinite dimensional) called the *feature space*. The questions we will ask today are:

- How to deal with infinite dimensional features?
- What properties on f_θ does the choice of ϕ entails? How to interpret the norm $\|\theta\|^2_{\mathcal{H}}$?

In this course, $\mathcal{H}$ will typically be $\mathbb{R}^p$, $\ell_2(\mathbb{N})$ or $L^2(\mathbb{R}^d)$ (endowed with their usual Hilbertian structure).

2 Representer theorem: from features to kernels

Consider *penalized* ERM with a linearly-parameterized predictor[1] $f_\theta(x) = \langle \theta, \phi(x) \rangle$ and a training set $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$ (with $\mathcal{X}$ arbitrary and $\mathcal{Y} \subset \mathbb{R}$):

$$\min_{\theta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle \theta, \phi(x) \rangle) + \lambda \|\theta\|^2. \tag{1}$$

Theorem 2.1 (Representer theorem)**.** *Any minimizer (if it exists) is of the form* $\theta = \sum_{i=1}^n \alpha_i \phi(x_i)$ *for some* $\alpha \in \mathbb{R}^n$.

Proof. Consider $\mathcal{H}_D = \{\sum_i^n \alpha_i \phi(x_i), \alpha \in \mathbb{R}^n\}$ which is finite dimensional and thus closed in $\mathcal{H}$. Each $\theta \in \mathcal{H}$ can be decomposed as $\theta = \theta_D + \theta^\perp$ where θ_D is the orthogonal projection on $\mathcal{H}_D$. We have, for $i \in \{1, \dots, n\}$

$$\begin{cases} \langle \theta, \phi(x_i) \rangle = \langle \theta_D, \phi(x_i) \rangle + \langle \theta^\perp, \phi(x_i) \rangle = \langle \theta_D, \phi(x_i) \rangle \\ \|\theta\|^2 = \|\theta_D\|^2 + \|\theta^\perp\|^2 \geq \|\theta_D\|^2 \end{cases} \tag{2}$$

Thus for any θ, its projection θ_D has smaller objective value (notice that we did not need to assume anything on the loss ℓ). □

*EPFL lenaic.chizat@epfl.ch

[1] We drop the index $\mathcal{H}$ on the norm/dot-product whenever there is no ambiguity.

Kernel trick Let us express the objective with θ parameterized as $\theta = \sum_{i=1}^n \alpha_i \phi(x_i)$ where $\alpha \in \mathbb{R}^n$:

$$\begin{cases} f_\theta(x) = \langle \theta, \phi(x) \rangle = \displaystyle\sum_{i=1}^n \alpha_i \langle \phi(x_i), \phi(x) \rangle \\ \|\theta\|^2 = \left\| \displaystyle\sum_{i=1}^n \alpha_i \phi(x_i) \right\|^2 = \displaystyle\sum_{i,j=1}^n \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle \end{cases}$$

We introduce the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $k(x, y) = \langle \phi(x), \phi(y) \rangle$, and the kernel matrix $K \in \mathbb{R}^{n \times n}$ with entries $K_{i,j} = k(x_i, x_j)$. In these notations we have the equivalent ERM problem

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

- For a test point, we have $f_\theta(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$;
- Given the kernel matrix K, we never explicitly manipulate vectors in $\mathcal{H}$ which is advantageous when $\mathcal{H}$ is high dimensional. This fact is called the *kernel trick*;
- The kernel matrix K is different from the covariance matrix. If $\mathcal{H} = \mathbb{R}^d$ and $\Phi \in \mathbb{R}^{n \times d}$ is the matrix of features (a.k.a. design matrix) with i-th row composed of $\phi(x_i)$, then $K = \Phi\Phi^\top$ and the empirical covariance matrix $\hat{\Sigma} = \frac{1}{n}\Phi^\top \Phi \in \mathbb{R}^{d \times d}$. These two matrices however have the same non-zero eigenvalues up to a factor n.

3 From kernels to features and RKHS

Definition 3.1 (Kernel)**.** *Let $\mathcal{X}$ be any set. The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel iff*

1. *it is symmetric : $k(x, y) = k(y, x),\ \forall x, y \in \mathcal{X}$*
2. *all kernel matrices are positive semi-definite (psd), i.e.*

$$\forall x_1, \ldots, x_n \in \mathcal{X}, \forall \alpha \in \mathbb{R}^n, \quad \sum \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

These properties are equivalent to requiring that all matrices K of the form $[K_{i,j}] = k(x_i, x_j)$ are psd (that is, symmetric with nonnegative eigenvalues)[2].

Theorem 3.2 (Aronszajn, 1950)**.** *A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if and only if there exists a Hilbert space $\mathcal{H}$ and a function $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x'$, $k(x, x') = \langle \phi(x), \phi(x') \rangle$.*

Proof. .

- $\Leftarrow$ is simple: if $k : (x, y) \mapsto \langle \phi(x), \phi(y) \rangle$ then k is clearly symmetric and $\forall n \in \mathbb{N}$, $\forall x_1, \ldots, x_n \in \mathcal{X}$, $\forall \alpha \in \mathbb{R}^n$ it holds $\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j) = \| \sum_{i=1}^n \alpha_i \phi(x_i) \|^2 \geq 0$.
- $\Rightarrow$ when $\mathcal{X}$ is finite: if $\mathcal{X} = \{x_1, \ldots, x_N\}$ then any p.d. kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is entirely defined by the $N \times N$ p.s.d. matrix $[K]_{i,j} = k(x_i, x_j)$. This matrix K can be diagonalized

[2]Notice the slightly inconsistent terminology: k is called p.d. (positive definite) when K is p.s.d. (positive semi-definite).

on an orthonormal basis of eigenvectors $u_1, \dots, u_N \in \mathbb{R}^N$ with nonnegative eigenvalues $\lambda_1 \geq \cdots \geq \lambda_N \geq 0$. It follows

$$k(x_i, x_j) = \left[\sum_{l=1}^{N} \lambda_l u_l u_l^\top\right]_{i,j} = \sum_{l=1}^{N} \lambda_l [u_l]_i [u_l]_j = \langle \phi(x_i), \phi(x_j) \rangle_{\mathbb{R}^N} \tag{3}$$

with

$$\phi(x_i) = \begin{pmatrix} \sqrt{\lambda_1}[u_1]_i \\ \vdots \\ \sqrt{\lambda_N}[u_N]_i \end{pmatrix}.$$

(This approach can be generalized to a compact probability space $\mathcal{X}$ and continuous kernels, this is Mercer's theorem; which can be proved via the spectral theory of compact operators).

- $\Rightarrow$ in the general case: let us build the feature space as a Hilbert space of functions $\mathcal{X} \to \mathbb{R}$. Consider $\mathcal{H}_0 = \operatorname{span}(\{k(x, \cdot)\}_{x \in \mathcal{X}})$. For $f = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i)$, $g = \sum_{j=1}^{n} \beta_j k(\cdot, y_j)$ define the dot-product

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i,j} \alpha_i \beta_j k(x_i, y_j). \tag{4}$$

  - This does not depend on the expansion of f and g since

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} \alpha_i g(x_i) = \sum_{j=1}^{n} \beta_j f(y_j)$$

  - This is a symmetric bilinear form.
  - Since k is p.d., $\|f\|_{\mathcal{H}_0}^2 = \sum \alpha_i \alpha_j k(x_i, x_j) \geq 0$. In particular Cauchy-Schwarz is valid.
  - Notice that $\langle f, k(x, \cdot) \rangle_{\mathcal{H}_0} = f(x)$ hence

$$|f(x)| = |\langle f, k(\cdot, x) \rangle_{\mathcal{H}_0}| \leq \|f\|_{\mathcal{H}_0} k(x, x)^{1/2}.$$

    Therefore $\|f\|_{\mathcal{H}_0} = 0 \Rightarrow f = 0$.
  - $\mathcal{H}_0$ is thus a pre-Hilbert space. Adding the pointwise limits of Cauchy sequences, we get a complete space $\mathcal{H}$ with the same properties (details in [Wainwright, 2019, Thm. 12.11]).
  - We thus have constructed $\phi : x \mapsto k(x, \cdot) \in \mathcal{H}$ which is a suitable feature map since $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}} = k(x, y)$. □

In the course of the proof, we have built the following type of space.

Definition 3.3. *The Reproducing Kernel Hilbert Space (RKHS) associated to the kernel k is a Hilbert space $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}$ such that:*

- *$\mathcal{H}$ contains all functions of the form $k(x, \cdot)$*
- *the reproducing property holds $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$*

It can be shown that the RKHS and the kernel are uniquely associated. (However, $x \mapsto k(x, \cdot)$ is in general not the only feature map that one can associate to k).

Proposition 3.4. *If $\mathcal{H}$ is a RKHS of functions $\mathcal{X} \to \mathbb{R}$, then for any $x \in \mathcal{X}$, the evaluation map $e_x : f \mapsto f(x)$ is continuous.*

Proof. Observe that the evaluation map is linear. We have

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

which proves that the evaluation map is bounded and thus continuous. □

In fact, RKHS are exactly the Hilbert spaces of functions $\mathcal{X} \to \mathcal{H}$ for which the evaluation map is continuous at all point (use Riesz representation theorem for the other implication). This is a very natural set-up for ML, since we need to work with predictors for which pointwise evaluation makes sense and is a stable operation (in paticular for the ERM to be well-behaved). In particular, $L^2(\mathbb{R}^d)$ is not a RKHS (the evaluation map is not continuous), and is indeed too big a space for doing ML.

Note: we also have

$$|f(x) - f(x')| = |\langle f, k(x,\cdot) - k(x',\cdot)\rangle \leq \|f\|_{\mathcal{H}} \cdot \underbrace{\|k(x,\cdot) - k(x',\cdot)\|_{\mathcal{H}}}_{=:\mathrm{dist}(x,x')}$$

hence f is Lipschitz continuous with constant $\|f\|_{\mathcal{H}}$ on $(\mathcal{X}, \mathrm{dist})$ (the distance in feature space).

4 First Examples

Linear kernel Let $\mathcal{X} = \mathbb{R}^d$ and $k(x,y) = x^\top y$. The RKHS is the space of linear functions $f_\theta(x) = \theta^\top x$ with norm $\|\theta\|_2^2$. Here the kernel trick is useful when the input have huge dimension d and are sparse so that $x^\top y$ can be computed in time $o(d)$.

Polynomial A natural generalization of the linear kernel on $\mathbb{R}^d$ is the homogeneous polynomial kernel $k(x,y) = (x^\top y)^m$ of degree $m \geq 2$, also defined on $\mathbb{R}^d$. Let us demonstrate that it is p.d. in the special case $m = 2$. We have

$$k(x,x') = \big(\sum_{i=1}^d x_i x_i'\big)^2 = \sum_{i=1}^d x_i^2 (x_i')^2 + 2\sum_{i<j} x_i x_j x_i' x_j'.$$

Setting $p = d + d(d-1)/2$ let us define the mapping $\phi : \mathbb{R}^d \to \mathbb{R}^p$ with entries

$$\phi(x) = \begin{pmatrix} x_j^2, & \text{for } j = 1,\dots,d \\ \sqrt{2}x_i x_j, & \text{for } i < j \end{pmatrix}.$$

We have $k(x,x') = \langle \phi(x), \phi(x')\rangle_{\mathbb{R}^p}$ hence k is pd. Here the RKHS is made of all homogeneous polynomials of degree 2. Another polynomial kernel is given by $k(x,y) = (1 + x^\top y)^m$ which gives all polynomials of degree m or less.

5 Equivalence of the "feature" and "kernel" approaches

So far, we have seen the following:

- (Section 3) Given a p.d. kernel k on $\mathcal{X}$: (i) there exists a (unique) RKHS of functions $\mathcal{F}_k$ associated to this kernel and (ii) there exists (a nonunique) a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that $k(x,y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}$ (in the 2nd proof of Aronszajn's theorem, we built $\mathcal{F}_k = \mathcal{H}$ but in the first proof using a spectral argument we had $\mathcal{F}_k \neq \mathcal{H}$).

- (Section 2): Given a feature map $\phi : \mathcal{X} \to \mathcal{H}$, one can define a space of predictors $\mathcal{F}_\phi \coloneqq \{f : x \mapsto \langle \theta, \phi(x) \rangle_\mathcal{H}\}$, a kernel $k(x, y) = \langle \phi(x), \phi(y) \rangle_\mathcal{H}$

To complete the equivalence between the "feature" and "kernel" approaches, we can ask the following questions:

- are the spaces of functions $\mathcal{F}_\phi$ and $\mathcal{F}_k$ the same?
- if so, can we express the RKHS norm directly in terms of ϕ?

The next proposition answers positively to these questions.

Proposition 5.1. *Given a feature map $\phi : \mathcal{X} \to \mathcal{H}$, the space of predictors $\mathcal{F} \coloneqq \{f : x \mapsto \langle \theta, \phi(x) \rangle_\mathcal{H}\}$ endowed with the norm*

$$\|f\|_\mathcal{F} = \inf\{\|\theta\|_\mathcal{H} \mid f = \langle \theta, \phi(\cdot) \rangle_\mathcal{H}\}$$

is precisely the RKHS associated to the kernel $k(x, y) = \langle \phi(y), \phi(x) \rangle$.

Proof. Consider the linear map $T : \mathcal{H} \to \mathcal{F}$ defined by $(T\theta)(x) = \langle \theta, \phi(x) \rangle_\mathcal{H}$ with null space $\mathcal{K}$. The only subtlety in the proof is to properly deal with the case $\mathcal{K} \neq \{0\}$. When restricted to the orthogonal complement $\mathcal{K}^\perp$, we obtain a bijection $U : \mathcal{K}^\perp \to \mathcal{F}$. We then define a dot-product on $\mathcal{F}$ as $\langle f, g \rangle_\mathcal{F} = \langle U^{-1} f, U^{-1} g \rangle_\mathcal{H}$.

We first show that this defines a RKHS with kernel $k(x, x') = \langle \phi(x), \phi(x') \rangle_\mathcal{H}$. We trivially have $k(\cdot, x) \in \mathcal{F}$ for all $x \in \mathcal{X}$. Moreover, for any $y \in \mathcal{X}$, we have with $p = U^{-1} k(\cdot, y) \in \mathcal{K}^\perp$ that $p - \phi(y) \in \mathcal{K}$ which implies

$$\langle f, k(\cdot, y) \rangle_\mathcal{F} = \langle U^{-1} f, p \rangle_\mathcal{H} = \langle U^{-1} f, \phi(y) \rangle_\mathcal{H} = T(U^{-1} f)(y) = f(y)$$

hence the reproducing property is satisfied and $\mathcal{F}$ is a RKHS. It remains to check that the RKHS norm is indeed the one we have defined. For any $f \in \mathcal{F}$, we have $f = T\theta$ for some $\theta = U^{-1} f + \theta^\perp$, $\theta^\perp \in \mathcal{K}^\perp$. Thus $\|\theta\|_\mathcal{H}^2 = \|U^{-1} f\|_\mathcal{H}^2 + \|\theta^\perp\|_\mathcal{H}^2 = \|f\|_{\text{RKHS}}^2 + \|\theta^\perp\|_\mathcal{H}^2$. This implies $\|\theta\|_\mathcal{H}^2 \geq \|f\|_{\text{RKHS}}^2$ with equality iff $\theta^\perp = 0$. This shows $\|f\|_{\text{RKHS}} = \inf\{\|\theta\|_\mathcal{H} \; ; \; f = T\theta\}$. □

References

Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.