Topics in Math of ML, Lecture 5
Kernel Methods (II)

Lénaïc Chizat*

October 13, 2025

1 Introduction

We continue our investigation of ERM with linearly-parameterized models. We consider learning with a linearly-parameterized predictor $f_\theta : x \to \langle \theta, \phi(x) \rangle_{\mathcal{H}} \in \mathbb{R}^{\mathcal{X}}$, $\theta \in \mathcal{H}$ where $\mathcal{H}$ is a Hilbert space (the feature space) and $\phi : \mathcal{X} \to \mathcal{H}$ is the feature map, and the space of predictors $\mathcal{F}$ is endowed with the Hilbertian norm

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

This space is a RKHS with kernel $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. In today's class, we explore:

- analysis of translation-invariant kernels in $\mathbb{R}^d$;
- combining estimation and approximation error bounds to get excess risk bounds
- approximation of feature maps via random features;

2 Translation invariant kernels on $\mathbb{R}^d$

We consider $\mathcal{X} = \mathbb{R}^d$ and a kernel of the form $k(x, x') = q(x - x')$ with a function $q : \mathbb{R}^d \to \mathbb{R}$.

Theorem 2.1 (Bochner)**.** *The kernel k is positive definite if and only if q is the Fourier transform of a non-negative and symmetric finite Borel measure. As a consequence, if $q \in L^1(\mathbb{R}^d)$, and its Fourier transform is in $L^1(\mathbb{R}^d)$ and nonnegative real-valued, then k is positive definite.*

See for instance [Mallat, 1999, Chap. 2] for a gentle reminder on the Fourier transform in $L^1(\mathbb{R}^d)$ and in $L^2(\mathbb{R}^d)$.

Partial proof. We only give the proof of the consequence (which is the most useful part). Since q is integrable,

$$\hat{q}(\omega) = \int_{\mathbb{R}^d} e^{-i\omega^\top x} q(x) \mathrm{d}x$$

is defined on $\mathbb{R}^d$ and continuous, and we have the inverse Fourier transform formula

$$q(x - x') = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{q}(\omega) e^{i(x-x')^\top \omega} \mathrm{d}\omega.$$

*EPFL lenaic.chizat@epfl.ch

Let $x_1, \dots, x_n \in \mathbb{R}^d$, let $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. We have

$$\begin{aligned}\sum_{s,j} \alpha_s \alpha_j k(x_s, x_j) &= \sum_{s,j} \alpha_s \alpha_j q(x_s - x_j) \\ &= \frac{1}{(2\pi)^d} \sum_{s,j} \alpha_s \alpha_j \int_{\mathbb{R}^d} \hat{q}(\omega) e^{i(x_s - x_j)^\top \omega} \mathrm{d}\omega \\ &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left(\sum_{s,j} \alpha_s \alpha_j e^{i x_s^\top \omega} (e^{i x_j^\top \omega})^* \right) \hat{q}(\omega) \mathrm{d}\omega \\ &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \Big| \sum_{s=1}^n \alpha_s e^{i\omega^\top x_s} \Big|^2 \hat{q}(\omega) \mathrm{d}\omega \geq 0. \end{aligned}$$

□

Associated norm (sketch). We consider the feature space $L^2(\mathbb{R}^d; \mathbb{C})$ and the feature map

$$\phi(x; \omega) = \frac{1}{(2\pi)^{d/2}} \sqrt{\hat{q}(\omega)} e^{i\omega^\top x}$$

(we can directly check that indeed it holds $k(x, y) = \langle \phi(x, \cdot), \phi(y; \cdot) \rangle_{L^2(\mathbb{R}^d)}$). Moreover given $f : \mathbb{R}^d \to \mathbb{R}$ sufficiently regular, we have

$$f(x) = \langle \phi(x), \theta \rangle_{L^2(\mathbb{R}^d)} \quad \text{with} \quad \theta(\omega) = \frac{1}{(2\pi)^{d/2}} \frac{\overline{\hat{f}(\omega)}}{\sqrt{\hat{q}(\omega)}}.$$

The RKHS norm of f is the square norm of θ, that is

$$\|f\|_{\mathcal{F}}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)} \mathrm{d}\omega.$$

This norm is related to the regularity of f. Indeed, recall for $f : \mathbb{R} \to \mathbb{R}$ the link between Fourier and (weak) derivatives for $k \geq 1$,

$$\int_{\mathbb{R}} |f^{(k)}(x)|^2 \mathrm{d}x = \frac{1}{2\pi} \int_{\mathbb{R}} |\omega|^{2k} |\hat{f}(\omega)|^2 \mathrm{d}\omega.$$

For instance, if $\hat{q}(\omega) = (1 + |\omega|^{2k})^{-1}$, then $\|f\|_{\mathcal{H}}^2 = (2\pi)^{-1} \Big(\|f\|_{L^2}^2 + \|f^{(k)}\|_{L^2}^2 \Big)$.

More generally, for $f : \mathbb{R}^d \to \mathbb{R}$ and for any multi-index $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$

$$\int_{\mathbb{R}^d} \Big| \frac{\partial^\alpha f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}(x)} \Big|^2 \mathrm{d}x = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\omega_1^{\alpha_1} \dots \omega_d^{\alpha_d}|^2 |\hat{f}(\omega)|^2 \mathrm{d}\omega.$$

Examples

- **Gaussian kernel:** $k(x, x') = q(x - x') = \exp(-\alpha \|x - x'\|_2^2)$. It has Fourier transform $\hat{q}(\omega) = (\pi/\alpha)^{d/2} \exp(-\|\omega\|_2^2/(4\alpha))$. By expanding in power series

$$\hat{q}(\omega)^{-1} = \Big(\frac{\alpha}{\pi} \Big)^{d/2} \sum_{s=0}^{\infty} \frac{\|\omega\|_2^{2s}}{(4\alpha)^s s!},$$

  this corresponds to a RKHS norm that penalizes all derivatives.

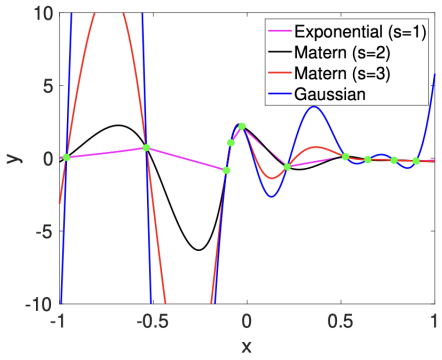

Figure 1: Min-norm interpolator for translation-invariant kernels. The green dots is the training set $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, \ldots, 10$. The functions are the interpolators (that is $f(x_i) = y_i$, $\forall i$) of minimum norm in various RKHS. Notice how the stronger the norm (s large) the stronger the "ringing" effect.

- **Matérn kernels:** $\hat{q}(\omega) \propto (\alpha^2 + \|\omega\|^2)^{-s}$. The associated RKHS is a Sobolev space of order s (the norm controls the L^2 norm of weak derivatives up to order s). One needs $s > d/2$ to ensure integrability (and hence being a RKHS). Indeed, one has by spherical change of coordinates

$$\int |\hat{q}(\omega)| \mathrm{d}\omega = \int_{\mathbb{R}^d} (\alpha^2 + \|\omega\|_2^2)^{-s} \mathrm{d}\omega = \int_0^\infty \int_{\mathbb{S}_{d-1}} (\alpha^2 + r^2)^{-s} r^{d-1} \mathrm{d}r\mathrm{d}\theta \tag{1}$$

  which is finite iff $-2s + d - 1 < -1$, that is $s > d/2$. Matérn kernels have closed form formulas (but we can skip using them, using the technique of random features, see below).

- **Exponential kernel:** $k(x, x') = \exp(-\alpha\|x - x'\|)$. It can be shown to be the Matérn kernel with $s = (d+1)/2$

- These kernels are illustrated on Figure 1.

Remark 2.2. *All these RKHS are dense in $L^2(\mathbb{R}^d)$. Indeed, given any $f \in L^2(\mathbb{R}^d)$, restricting its Fourier transform $\hat{f} \in L^2(\mathbb{R}^d)$ to a compact set, and taking the inverse Fourier transform, gives a function in those RKHS (why?) that can approximate f arbitrarily well in L^2.*

3 Generalization guarantees

Consider our usual statistical learning framework: a training set $(x_i, y_i)_i^n$ iid samples from random variables $(x, y) \sim \rho \in \mathcal{P}(\mathcal{X} \times \mathbb{R})$). Consider a G-Lipchitz continuous loss function ℓ, and the *constrained* ERM problem

$$\hat{f}_D \in \mathrm{argmin}_{\|f\|_{\mathcal{F}} \leq D} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)).$$

We want to derive generalization guarantees by combining estimation and approximation bounds.

General computations Assume that the risk $\mathcal{R}(f) := \mathbf{E}\ell(y, f(x))$ admits a minimizer $f^* \in L^2(\rho_x)$. Our goal is to show bounds on the excess risk :

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}(f^*) &= \mathbf{E}[\ell(y, f(x)) - \ell(y, f^*(x))] \\ &\leq G\mathbf{E}|f(x) - f^*(x)| \\ &\leq G\big(\mathbf{E}|f(x) - f^*(x)|^2\big)^{1/2} = G\|f - f^*\|_{L^2(\rho_x)}.\end{aligned}$$

We consider the constrained problem for theoretical convenience, so that we can directly plug our estimation error bounds based on Rademacher-complexity from Lecture 3[1]. They tell us, assuming that $k(x, x) \leq R^2$ almost surely, that

$$\begin{aligned}\mathbf{E}[\mathcal{R}(\hat{f}_D)] - \mathcal{R}(f^*) &\leq \underbrace{2\mathbf{E}\Big[\sup_{\|f\|_{\mathcal{F}} \leq D} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)|\Big]}_{\text{estimation error}} + \underbrace{\inf_{\|f\|_{\mathcal{F}} \leq D} \mathcal{R}(f) - \mathcal{R}(f^*)}_{\text{approximation error}} \\ &\leq \frac{4GDR}{\sqrt{n}} + G \inf_{\|f\|_{\mathcal{F}} \leq D} \|f - f^*\|_{L^2(\rho_x)}.\end{aligned}$$

We now want to find the optimal D to balance estimation and approximation error. We have, using $a + b \leq 2\sqrt{a^2 + b^2}$ in the last line, that

$$\begin{aligned}\inf_{D \geq 0} \mathbf{E}[\mathcal{R}(\hat{f}_D)] - \mathcal{R}(f^*) &\leq \inf_{D \geq 0} \frac{4GDR}{\sqrt{n}} + G \inf_{\|f\|_{\mathcal{F}} \leq D} \|f - f^*\|_{L^2(\rho_x)} \\ &= G \inf_{f \in \mathcal{F}} \frac{4R\|f\|_{\mathcal{F}}}{\sqrt{n}} + \|f - f^*\|_{L^2(\rho_x)} \\ &\leq 2G\Big(\inf_{f \in \mathcal{F}} \frac{16R^2}{n}\|f\|_{\mathcal{F}}^2 + \|f - f^*\|_{L^2(\rho_x)}^2\Big)^{1/2}.\end{aligned}$$

Setting $\lambda = 16R^2/n$, we thus need to study the quantity

$$A(\lambda, f^*) = \inf_{f \in \mathcal{F}} \{\|f - f^*\|_{L^2(\rho_x)}^2 + \lambda\|f\|_{\mathcal{F}}^2\}$$

- if $f^* \in \mathcal{F}$, then $A(\lambda, f^*) \leq \lambda\|f^*\|_{\mathcal{F}}^2$ and the excess risk is bounded as

$$\inf_{D \geq 0} \mathbf{E}[\mathcal{R}(\hat{f}_D)] - \mathcal{R}(f^*) \leq \frac{8GR\|f^*\|}{\sqrt{n}}.$$

- if $f^* \notin \mathcal{F}$, but $\mathcal{F}$ is dense in $L^2(\rho_x)$ then $\lim_{\lambda \to 0} A(\lambda, f^*) = 0$: we get a *consistent* estimation but a priori no convergence rate (we get rates below under additional assumptions).
- if $\mathcal{F}$ is not dense, there is an incompressible approximation error $A(\lambda, F^*) \geq \|f^* - \Pi_{\mathcal{F}}(f^*)\|_{L^2(\rho_x)}^2$ where $\Pi_{\mathcal{F}}$ is the orthogonal projection on $\mathcal{H}$.

Translation invariant kernels on $\mathbb{R}^d$ Assuming $\|\cdot\|_{L^2(\rho_x)} \leq C\|\cdot\|_{L^2(\mathrm{d}x)}$ (satisfied if ρ_x has a bounded density), we focus now on

$$\tilde{A}(\lambda, f^*) = \inf_{f \in \mathcal{F}} \{\|f - f^*\|_{L^2(\mathrm{d}x)}^2 + \lambda\|f\|_{\mathcal{F}}^2\}.$$

For translation invariant kernels,

$$\|f\|_{\mathcal{F}}^2 = \frac{1}{(2\pi)^d} \int \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)} \mathrm{d}\omega$$

[1] Similar guarantees can be obtained for the regularized problem (see [Bach, 2024, Sec. 7.5.1]).

and thus

$$\tilde{A}(\lambda, f^*) = \inf_{\hat{f} \in L^2(\hat{q}^{-1})} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \Big(|\hat{f}(\omega) - \hat{f}^*(\omega)|^2 + \lambda \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)} \Big) \mathrm{d}\omega.$$

We can optimize pointwisely under the integral and get $0 = 2(\hat{f}(\omega) - \hat{f}^*(\omega)) + 2\lambda \hat{f}(\omega)/\hat{q}(\omega)$ $\Rightarrow \hat{f}_\lambda(\omega) = \hat{f}^*(\omega)/(1 + \lambda \hat{q}(\omega)^{-1})$ and we get (we skip the details):

$$\tilde{A}(\lambda, f^*) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\lambda}{\hat{q}(\omega) + \lambda} |\hat{f}^*(\omega)|^2 \mathrm{d}\omega.$$

Under special decay assumptions We now assume that for some $t > 0$ (regularity of f^*) and $s > d/2$ (regularity of the kernel), it holds

$$\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} (1 + \|\omega\|_2^2)^t |\hat{f}^*(\omega)|^2 \mathrm{d}\omega =: \|f^*\|_{H^t}^2 < \infty \qquad \text{and} \qquad \hat{q}(\omega) \propto (1 + \|\omega\|_2^2)^{-s}. \quad (2)$$

When $t \geq s$, $f^* \in \mathcal{F}$ and this case is discussed above. Otherwise we have

$$\tilde{A}(\lambda, f^*) \leq \frac{1}{(2\pi)^d} \int \frac{(1 + \|\omega\|_2^2)^t}{(1 + \|\omega\|_2^2)^t} |\hat{f}^*(\omega)|^2 \frac{\lambda}{\hat{q}(\omega) + \lambda} \mathrm{d}\omega \quad (3)$$

$$\lesssim \|f^*\|_{H^t}^2 \sup_\omega \frac{\lambda}{\hat{q}(\omega)^{t/s} \lambda^{1-t/s}} \frac{1}{(1 + \|\omega\|_2^2)^t} = O(\lambda^{t/s}) \quad (4)$$

where we used Young's inequality/concavity of the logarithm $(t/s)a + (1 - t/s)b \geq a^{t/s} b^{1-t/s}$ which leads to $\hat{q}(w) + \lambda \geq C \hat{q}(\omega)^{t/s} \lambda^{1-t/s}$ where C only depends on t and s.

- Putting things together, the excess risk is of order

$$\inf_{D \geq 0} \mathbf{E}[\mathcal{R}(\hat{f}_D)] - \mathcal{R}(f^*) \lesssim \sqrt{\tilde{A}(R^2/n, f^*)} = \begin{cases} O(n^{-t/(2s)}) & \text{if } t \leq s, \\ O(n^{-1/2}) & \text{if } t \geq s \end{cases}.$$

- Example: when $t = 1$ and $s = (d+1)/2$ (exponential kernel) then we have a rate of $O(n^{-1/(d+1)})$. This is the curse of dimensionality (no method avoids it). However, these methods are adaptive to extra-level of smoothness: in the best case when $t \geq s$, we have a rate of $O(n^{-1/2})$.

- Here it seems that taking s as small as possible (while satisfying $s > d/2$) is always the best: this is because we only look at the exponent in t. In fact, very smooth kernels are better at learning very smooth functions, see Figure 2.

4 Random feature approximations

Many kernels can be expressed as

$$k(x, x') = \int_{\mathcal{V}} \phi(x, v) \phi(x', v) \mathrm{d}\mu(v) \quad (5)$$

where μ is a probability distribution on some space $\mathcal{V}$ and $\varphi : \mathcal{X} \times \mathcal{V} \to \mathbb{R}$. We can then approximate k by an empirical average: $\hat{k}(x, x') = \frac{1}{m} \sum_{j=1}^m \phi(x, v_j) \phi(x', v_j)$ for $v_1, \dots, v_j \overset{\text{i.i.d}}{\sim} \mu$. This corresponds to an explicit feature representation $\hat{\phi}(x) = (m^{-1/2} \phi(x, v_i))_{i=1}^m$ and we can solve

$$\min_{\theta \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{\phi}(x_i)^\top \theta) + \frac{\lambda}{2} \|\theta\|^2.$$

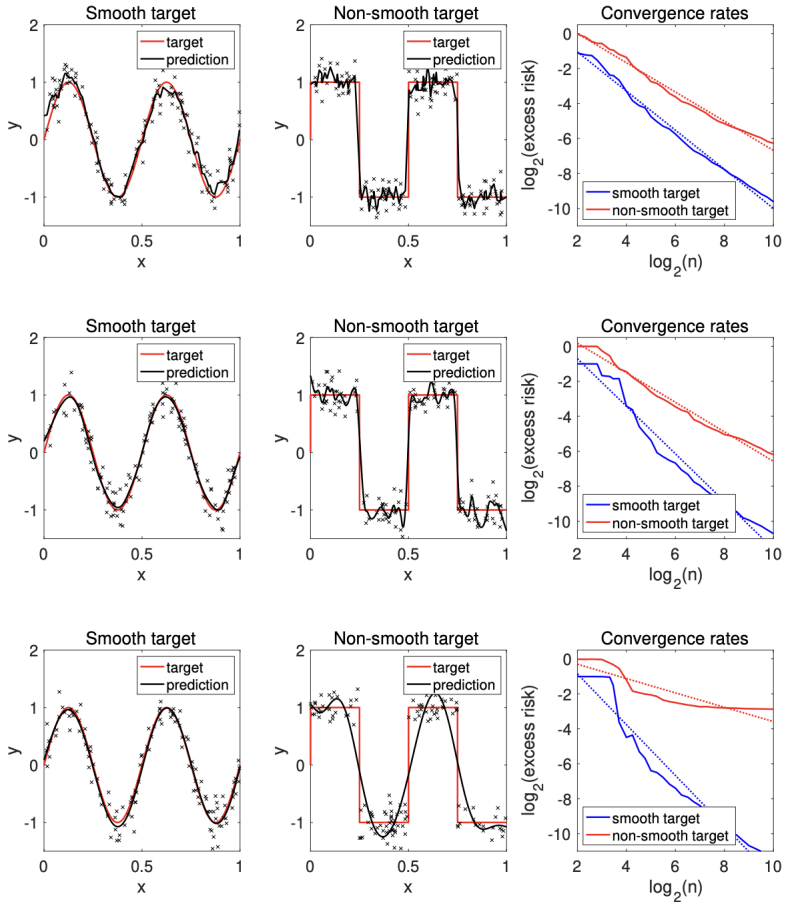

Figure 2: 1D kernel regression (with optimal regularization level): (top) Matérn with $s = 1$ (exponential kernel) (middle) Matérn with $s = 3$ and (bottom) Gaussian (from Bach [2024]).

This trick is useful for large datasets since it allows to replace the kernel matrix of size $n \times n$ by a design matrix of size $m \times m$ by introducing an error of order $m^{-1/2}$ on the kernel (details in Exercices). When $m \leq n$, this error can be small compared to the excess risk $n^{-t/(2s)}$ for "hard" tasks ($t < d/2$).

Translation-invariant kernels For a kernel on $\mathcal{X} = \mathbb{R}^d$ of the form

$$k(x, x') = q(x' - x) = \frac{1}{(2\pi)^d} \int \hat{q}(\omega) e^{i\omega^\top (x - x')} \mathrm{d}\omega \in \mathbb{C}$$

we can take $\phi(x, \omega) = \sqrt{q(0)} e^{i\omega^\top x}$ and $\frac{\mathrm{d}\mu}{\mathrm{d}\omega} = (2\pi)^{-d} \frac{\hat{q}(\omega)}{q(0)}$. Such feature maps are called *random Fourier features*. The measure μ is a probability measure because by the Fourier inversion formula $q(0) = (2\pi)^{-d} \int \hat{q}(\omega) \mathrm{d}\omega$. In the next lectures, we will sometimes interpret neural networks as random feature approximations of kernels.

References

Francis Bach. *Learning theory from first principles.* MIT press, 2024.

Stéphane Mallat. *A wavelet tour of signal processing.* Elsevier, 1999.