Learning theory from first principles

Lecture 6: Kernel methods

Francis Bach

October 30, 2020

Class summary

-Kernels and representer theorems
-Kernels on $\mathbb{R}^d$
-Algorithms
-Analysis of well-specified models
-Sharp analysis of ridge regression
-Universal consistency

1 Introduction

In this lecture, we study empirical risk minimization for linear models, that is, prediction functions $f_\theta : \mathcal{X} \to \mathbb{R}$ which are linear in their parameters θ, that is of the form $f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathcal{H}}$, where $\varphi : \mathcal{X} \to \mathcal{H}$ and $\mathcal{H}$ is a Hilbert space (essentially a Euclidean space with potentially infinite dimension), and $\theta \in \mathcal{H}$. We will often use the notation $\langle \theta, \varphi(x) \rangle$ in this lecture instead of $\langle \theta, \varphi(x) \rangle_{\mathcal{H}}$ when this is not ambiguous.

The key difference with Lecture 2 on least-squares estimation is that, (1) we are not restricted to the square loss (although many of the same concepts with play a role, in particular the analysis of ridge regression), and (2), we will explicitly allow infinite-dimensional models, thus extending the dimension-free bounds from Lecture 2. The notion of kernel $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$ will be particularly fruitful.

Why is this relevant?

- Understanding linear models in finite but very large input dimensions requires tools from infinite-dimensional analysis.
- Kernel methods lead to simple and stable algorithms, with theoretical guarantees, and adaptivity (as opposed to local averaging techniques). They can be applied in high dimensions, with good practical performance (note that for supervised learning problems in domains such as computer vision and natural language processing, they do not achieve the state of the art anymore).
- They can be easily applied when input observations are not vectors.

- They are useful to understand other models such as neural networks (see Lectures 8 amd 9).

The type of kernel we consider here is different from the ones in Lecture 5. The ones here are "positive definite"; the ones from Lecture 5 are "non-negative". See more details in `https://francisbach.com/cursed-kernels/`.

2 Representer theorem

Dealing with infinite-dimensional models seems impossible at first because algorithms cannot be run in infinite dimensions. In this section, we show how the kernel function plays a crucial role to achieve lower-dimensional algorithms.

As a motivation, we consider the optimization problem coming from machine learning with linear models, with data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$:

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

assuming the loss function ℓ is already from $\mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ and not from $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ (e.g., hinge loss, logistic loss or least-squares, see Lecture 3).

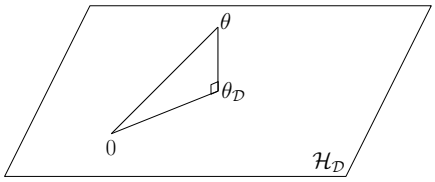
The key property of the objective function in Eq. (1) is that it accesses the input observations $x_1, \dots, x_n \in \mathcal{X}$, only through dot-products $\langle \theta, \varphi(x_i) \rangle$, $i = 1, \dots, n$, and that we penalize using the Hilbert norm $\|\theta\|$. The following theorem is crucial and has a particularly simple proof.

Theorem 1 (Representer theorem [1]) *Let $\varphi : \mathcal{X} \to \mathcal{H}$. Let $(x_1, \dots, x_n) \in \mathcal{X}^n$, and assume that the functional $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ is strictly increasing with respect to the last variable, then the infimum of $\Psi(\langle \theta, \varphi(x_1) \rangle, \cdots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2)$ can be obtained by restricting to θ of the form*

$$\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i),$$

with $\alpha \in \mathbb{R}^n$.

Proof Let $\theta \in \mathcal{H}$, and $\mathcal{H}_{\mathcal{D}} = \Big\{ \sum_{i=1}^{n} \alpha_i \varphi(x_i),\ \alpha \in \mathbb{R}^n \Big\} \subset \mathcal{H}$, the linear span of the feature vectors. Let $\theta_{\mathcal{D}} \in \mathcal{H}_{\mathcal{D}}$ and $\theta_{\perp} \in \mathcal{H}_{\mathcal{D}}^{\perp}$ be such that $\theta = \theta_{\mathcal{D}} + \theta_{\perp}$, a decomposition which is using the Hilbertian structure of $\mathcal{H}$. Then $\forall i \in \{1, \dots, n\}$, $\langle \theta, \varphi(x_i) \rangle = \langle \theta_{\mathcal{D}}, \varphi(x_i) \rangle + \langle \theta_{\perp}, \varphi(x_i) \rangle$ with $\langle \theta_{\perp}, \varphi(x_i) \rangle = 0$.



From Pythagoreas theorem, we get: $\|\theta\|^2 = \|\theta_{\mathcal{D}}\|^2 + \|\theta_{\perp}\|^2$. Therefore we have:

$$\begin{aligned} \Psi(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2) &= \Psi(\langle \theta_{\mathcal{D}}, \varphi(x_1) \rangle, \dots, \langle \theta_{\mathcal{D}}, \varphi(x_n) \rangle, \|\theta_{\mathcal{D}}\|^2 + \|\theta_{\perp}\|^2) \\ &\geqslant \Psi(\langle \theta_{\mathcal{D}}, \varphi(x_1) \rangle, \dots, \langle \theta_{\mathcal{D}}, \varphi(x_n) \rangle, \|\theta_{\mathcal{D}}\|^2). \end{aligned}$$

Thus

$$\inf_{\theta \in \mathcal{H}} \Psi(\langle \theta, \varphi(x_1) \rangle, \cdots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2) = \inf_{\theta \in \mathcal{H}_{\mathcal{D}}} \Psi(\langle \theta, \varphi(x_1) \rangle, \dots, \langle \theta, \varphi(x_n) \rangle, \|\theta\|^2),$$

which is exactly the desired result. ■

This implies that the minimizer of Eq. (1) can be looked among the vectors of the form $\theta = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$:

Corollary 1 (Representer theorem for supervised learning) *For $\lambda > 0$,*

$$\inf_{\theta\in\mathcal{H}} \frac{1}{n}\sum \ell(y_i, \langle\theta, \varphi(x_i)\rangle) + \frac{\lambda}{2}\|\theta\|^2 = \inf_{\alpha\in\mathbb{R}^n} \frac{1}{n}\sum \ell(y_i, \langle\theta, \varphi(x_i)\rangle) + \frac{\lambda}{2}\|\theta\|^2 \text{ such that } \theta = \sum_{i=1}^n \alpha_i\varphi(x_i).$$

- It is important to note that there is no assumption on the loss function ℓ. In particular no convexity is assumed. This is to be contrasted to the use of duality in Section 4, where convexity will play a major role and similar α's will be defined (but with some notable differences).
- We have: $\forall j \in \{1, \dots, n\}$, $\langle\theta, \varphi(x_j)\rangle = \sum_{i=1}^n \alpha_i k(x_i, x_j) = (K\alpha)_j$ where $K \in \mathbb{R}^{n\times n}$ is the *kernel matrix*, such that $K_{ij} = \langle\varphi(x_i), \varphi(x_j)\rangle = k(x_i, x_j)$, and

  $$\|\theta\|^2 = \sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j\langle\varphi(x_i), \varphi(x_j)\rangle = \sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j K_{ij} = \alpha^\top K\alpha.$$

  We can then write:

  $$\min_{\theta\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle\theta, \varphi(x_i)\rangle) + \frac{\lambda}{2}\|\theta\|^2 = \min_{\alpha\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2}\alpha^\top K\alpha.$$

  For a test point $x \in \mathcal{X}$, we have $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$.

  Thus, the input observations are summarized in the kernel matrix and the kernel function, regardless of the dimension of $\mathcal{H}$. This is the *kernel trick*.

The kernel trick allows to:

- replace $\mathcal{H}$ by $\mathbb{R}^n$; this is interesting computationally when the dimension of $\mathcal{H}$ is very large (see more details in Section 4),
- separate the representation problem (design of kernels on a set $\mathcal{X}$) and algorithms and analysis (which only use the kernel matrix K); this is interesting because a wide range of kernels can be defined for many data types (see more details in Section 3) .

3 Kernels

In the section above, we have introduced the kernel function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ as obtained from a dot product $k(x, y) = \langle\varphi(x), \varphi(y)\rangle$. The associated kernel matrix is then a matrix of dot-products (often called a Gram matrix), and is thus symmetric positive semi-definite, that is, all of its eigenvalues are non-negative, or $\forall\alpha \in \mathbb{R}^n$, $\alpha^\top K\alpha \geqslant 0$. It turns out that this simple property is enough to impose the existence of a feature function.

⚠ In this section y in an element of $\mathcal{X}$, just like $x \in \mathcal{X}$, and not a label.

⚠ If $\mathcal{H} = \mathbb{R}^d$, and $\Phi \in \mathbb{R}^{n \times d}$ is the matrix of features (design matrix in the context of regression) with i-rh row composed of $\varphi(x_i)$, then $K = \Phi\Phi^\top \in \mathbb{R}^{n \times n}$ is the kernel matrix, while $\frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d \times d}$ is the empirical covariance matrix.

This thus leads to the definition:

Definition 1 *a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel if and only if all kernel matrices are symmetric positive semi-definite.*

The important following theorem that dated back to Aronszajn (1950), with a constructive proof.

Theorem 2 ([2]) *k is a positive definite kernel if and only if there exists a Hilbert space $\mathcal{H}$, and a function $\varphi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, y$, $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$.*

Proof We only give a proof sketch. One direction is straightforward. For the other direction we consider a positive-definite kernel, and we will construct explicitly a space of functions from $\mathcal{X}$ to $\mathbb{R}$ with a dot-product. We define the set $\mathcal{H}'$ as the set of linear combinations of kernel functions $\sum_{i=1}^n \alpha_i k(\cdot, x_i)$ for any integer n, any set of n points and any $\alpha \in \mathbb{R}^n$. This is a vector space, on which we can define a dot-product through

$$\Big\langle \sum_{i=1}^n \alpha_i k(\cdot, x_i), \sum_{j=1}^m \beta_j k(\cdot, y_j) \Big\rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, y_j).$$

One can check that this is a well-defined function on $\mathcal{H}' \times \mathcal{H}'$ (the value does not depend on the chosen representation as linear combination of kernel functions), that it is a dot-product on $\mathcal{H}'$, which satisfies the two properties for any $f \in \mathcal{H}'$, $x, y \in \mathcal{X}$:

$$\langle k(\cdot, x), f \rangle = f(x) \text{ and } \langle k(\cdot, x), k(\cdot, y) \rangle = k(x, y).$$

These are called reproducing properties, and corresponds to $\varphi(x) = k(\dot{,}x)$.

The space $\mathcal{H}'$ is called "pre-Hilbertian", because it is not complete. It can be "completed" into a Hilbert space $\mathcal{H}$ with the same reproducing property. See [2, 3] for more details.

■

We can make the following observations:

- $\mathcal{H}$ is called the "feature space", and φ the "feature map".
- No assumptions are needed about the input space $\mathcal{X}$, and no regularity assumptions are needed for k. Up to isomorphisms, the feature map and space happen to be unique. The particular space of functions, we built is called the reproducing kernel Hilbert space (RKHS), associated to $\mathcal{H}$, for which $\varphi(x) = k(\cdot, x)$.
- A classical intuitive interpretation of the identity $\langle k(\cdot, x), f \rangle = f(x)$ is that the function evaluation is the dot-product with a function. If $L_2(\mathbb{R}^d)$ was an RKHS, this would mean that there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\int_{\mathbb{R}^d} k(x, y) f(y) dy = f(x)$. In other words, $k(x, y)dy$ would be a Dirac measure at x, which is impossible (as Dirac measures have no density with respect to the Lebesgue measures). Thus $L_2(\mathbb{R}^d)$ is a Hilbert space which is too large to be an RKHS.

- Given a positive-definite kernel k, we can thus associate it to some feature map φ such that $k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$, but also to a *space of functions on $\mathcal{X}$ with a given norm*, either directly through the RKHS above, or by looking at all functions f_θ of the form $f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathcal{H}}$, with a regularization term $\|\theta\|_{\mathcal{H}}^2$.

  ⚠ From now on, we will denote elements of the Hilbert space $\mathcal{H}$ through the notation $f \in \mathcal{H}$ to highlight the fact that we are considering a space of functions from $\mathcal{X}$ to $\mathbb{R}$, except for optimization algorithms in Section 4, where will use the notation $\langle \theta, \varphi(x) \rangle_{\mathcal{H}}$ instead of $f(x)$.

- **Exercise**: the sum and (pointwise) product of kernels are kernels. What are their associated feature spaces and feature maps?

- Sometimes, the feature map is easy to find, sometimes it is not. In the next section, we will look at the main examples, and describe the associated spaces of functions (and the corresponding norms).

We now look at different ways of building the kernels, by starting first from the feature vector (e.g., linear kernels), from the kernel and explicit feature map (polynomial kernel), from the norm (translation-invariant kernel on $[0,1]$), or from the kernel without explicit features (translation-invariant kernel on $\mathbb{R}^d$).

3.1 Linear and polynomial kernels

We start with the most obvious kernels on $\mathcal{X} = \mathbb{R}^d$, for which feature maps are easily found.

- Linear kernel: $k(x,y) = x^\top y$. It corresponds to linear functions $f_\theta(x) = \theta^\top x$, with an ℓ_2-penalty $\|\theta\|_2^2$. The kernel trick can be useful when the input data have huge dimension d, but are quite sparse (many zeros), such as in text processing, so that the dot-product $x^\top y$ can be computed in time $o(d)$.

- Polynomial kernel: for r a postive integer, the kernel $k(x,y) = (x^\top y)^r$ can be expanded as:

$$k(x,y) = \Big(\sum_{i=1}^d x_i y_i \Big)^r = \sum_{\alpha_1 + \cdots + \alpha_d = r} \binom{r}{\alpha_1, \dots, \alpha_d} \underbrace{(x_1 y_1)^{\alpha_1} \cdots (x_d y_d)^{\alpha_d}}_{(x_1^{\alpha_1} \cdots x_d^{\alpha_d})(y_1^{\alpha_1} \cdots y_d^{\alpha_d})}.$$

  We have an explicit feature map: $\varphi(x) = \Big(\binom{r}{\alpha_1, \dots, \alpha_d}^{\frac{1}{2}} x_1^{\alpha_1} \cdots x_d^{\alpha_d} \Big)_{\alpha_1 + \cdots + \alpha_d = r}$, and the set of functions is the set of homogeneous polynomials on $\mathbb{R}^d$, which has dimension $\binom{d+r-1}{r}$.

  When d and r grows, and the dimension grows as d^r, an explicit representation is not desirable, and the kernel trick can be advantageous. Note however, that the associated norm (which penalizes coefficients of the polynomials), is hard to interpret.

  Exercise: how can we go beyond homogeneous polynomials, and consider all monomials $x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ such that $\alpha_1 + \cdots + \alpha_d \leqslant r$?

3.2 Translation-invariant kernels on $[0,1]$

We consider $\mathcal{X} = [0,1]$, and a kernel of the form $k(x,y) = q(x-y)$ with a function $q : [0,1] \to \mathbb{R}$, which is 1-periodic. Squared integrable functions which are 1-periodic can be expanded in Fourier series, that is, $q(x) = \sum_{m\in\mathbb{Z}} e^{2im\pi x}\hat{q}_m$, with $\hat{q}_m = \int_0^1 q(x)e^{-2im\pi x}dx$, for $m \in \mathbb{Z}$.

Given a 1-periodic function f decomposed into its Fourier series $f(x) = \sum_{m\in\mathbb{Z}} e^{2im\pi x}\hat{f}_m$, we consider the penalty

$$\sum_{m\in\mathbb{Z}} c_m|\hat{f}_m|^2,$$

with $c \in \mathbb{R}_+^{\mathbb{Z}}$; this corresponds to the feature vector $\varphi(x)_m = \dfrac{e^{2im\pi x}}{\sqrt{c_m}}$, and $\theta \in \mathbb{C}^{\mathbb{Z}}$, such that $\theta_m = \hat{f}_m\sqrt{c_m}$ (we can easily consider complex-valued features instead of real-valued features if Hermitian dot-products are considered), so that $\sum_{m\in\mathbb{Z}}|\theta_m|^2$ is equal to the norm $\sum_{m\in\mathbb{Z}} c_m|\hat{f}_m|^2$.

Thus the kernel is

$$k(x,y) = \sum_{m\in\mathbb{Z}} \varphi(x)_m\varphi(x)_m^* = \sum_{m\in\mathbb{Z}} \frac{e^{2im\pi x}}{\sqrt{c_m}}\frac{e^{-2im\pi y}}{\sqrt{c_m}} = \sum_{m\in\mathbb{Z}} \frac{1}{c_m}e^{2im\pi(x-y)} = q(x-y).$$

It is thus natural to consider functions q which are 1-periodic, and such that the Fourier series has non-negative real values $\hat{q}_m = c_m^{-1}$.

Penalization of derivatives. For certain penalties based on c, there is a natural link with penalties on derivatives, as, if f is s-times differentiable with squared integrable derivative, we have $f^{(s)}(x) = \sum_{m\in\mathbb{Z}}(2im\pi)^s e^{2im\pi x}\hat{f}_m$, and thus, from Parseval's theorem:

$$\int_0^1 |f^{(s)}(x)|^2dx = (2\pi)^{2s}\sum_{m\in\mathbb{Z}} m^{2s}|\hat{f}_m|^2.$$

In this lecture we will consider penalizing such derivatives, leading to Sobolev spaces on $[0,1]$. The following examples are often considered:

- Bernoulli polynomials: $c_0 = 1$ and $c_m = |m|^{2s}$ for $m \neq 0$, for which we have $k(x,y) = 1 + (-1)^{s-1}\frac{(2\pi)^{2s}}{(2s)!}B_{2s}(\{x-y\})$, where B_{2s} the $(2s)$-th Bernoulli polynomial[1], and $\{x-y\} \in [0,1)$ is the fractional part of $x-y$. The associated norm is $\|f\|_{\mathcal{H}}^2 = \dfrac{1}{(2\pi)^{2s}}\displaystyle\int_0^1 |f^{(s)}(x)|^2dx + \Big(\int_0^1 |f(x)|^2dx\Big)$.
- Periodic exponential kernel: we can consider $c_m = 1 + \alpha^2|m|^2$, for which we have also a closed-form formula, with the penalty $\|f\|_{\mathcal{H}}^2 = \dfrac{\alpha^2}{(2\pi)^2}\displaystyle\int_0^1 |f^{(s)}(x)|^2dx + \int_0^1 |f(x)|^2dx$.

  Exercise (♦♦♦♦) : Give a closed-form for the kernel $k(x,y) = \displaystyle\sum_{m\in\mathbb{Z}} \frac{e^{2im\pi(x-y)}}{1+\alpha^2|m|^2}$. Hint: use the Cauchy residue formula.

[1] See `https://en.wikipedia.org/wiki/Bernoulli_polynomials`.

These kernels are mostly used for their simplicity and their explicit feature map, which are simpler than the kernels which are most used below (with similar links with Sobolev spaces). Note also, that for the uniform distribution on $[0,1]$, the Fourier basis will be an orthogonal eigenbasis of the covariance operator with eigenvalues c_m^{-1} (see Section **??**).

From the kernel $q(x-y)$ with Fourier series $\hat{q}_m$ for q, the associated norm is $\sum_{m\in\mathbb{Z}} \frac{|\hat{f}_m|^2}{\hat{q}_m}$. We now extend this to Fourier transforms (instead of Fourier series).

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

We consider $\mathcal{X} = \mathbb{R}^d$, and a kernel of the form $k(x,y) = q(x-y)$ with a function $q : \mathbb{R}^d \to \mathbb{R}$. The following theorem gives conditions under which we obtain a positive definite kernel.

Theorem 3 (Böchner [4]) *The kernel k is positive definite if and only if q is the Fourier transform of a non-negative Borel measure. As a consequence, if $q \in L^1(dx)$ and its Fourier transform only has non-negative real values, then k is positive definite.*

Proof We only give the proof of the consequence, which is the only one that we need. Since q is integrable, $\hat{q}(\omega) = \int_{\mathbb{R}^d} e^{-i\omega^\top x} q(x)dx$ is defined on $\mathbb{R}^d$ and continuous, and we have through the inverse Fourier transform formula:

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

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

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

which shows the positive-definiteness. ■

Construction of the norm. We give an intuitive (non-rigorous) reasoning: if q is in $L^1(dx)$, then $\hat{q}(\omega)$ exists and, we have an explicit representation as

$$k(x,y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \langle \sqrt{\hat{q}(\omega)} e^{i\omega^\top x}, \sqrt{\hat{q}(\omega)} e^{i\omega^\top y} \rangle d\omega = \int_{\mathbb{R}^d} \langle \varphi(x)_\omega, \varphi(y)_\omega \rangle d\omega,$$

which is of the form $\langle \varphi(x), \varphi(y) \rangle$, with $\varphi(x)_\omega = \frac{1}{(2\pi)^{d/2}} \sqrt{\hat{q}(\omega)} e^{i\omega^\top x}$. If we consider $f(x) = \int_{\mathbb{R}^d} \varphi(x)_\omega \theta_\omega d\omega = \langle \varphi(x), \theta \rangle$, then $\theta_\omega = \frac{1}{(2\pi)^{d/2}} \hat{f}(\omega) / \sqrt{\hat{q}(\omega)}$, and the squared norm of θ is equal to $\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(w)|^2}{\hat{q}(\omega)} d\omega$, where

$\hat{f}$ denotes the Fourier transform of f. Therefore, we norm of a function $f \in \mathcal{H}$ is (for a formal proof, see [5]):

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

Note the similarity with the penalty for the kernel on $[0,1]$ (see more similarity below).

Link with derivatives. When f has partial derivatives, then the Fourier transform of $\frac{\partial f}{\partial x_j}$ is equal to $i\omega_j$ times the Fourier transform of f. This leads to, using Parseval's theorem, $\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\omega_j|^2 |\hat{f}(w)|^2 d\omega = \int_{\mathbb{R}^d} \big|\frac{\partial f}{\partial x_j}(x)\big|^2 dx$, which extends to higher order derivatives:

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

This will allow us to find corresponding norms, by expanding $\hat{q}(\omega)^{-1}$ with polynomials.

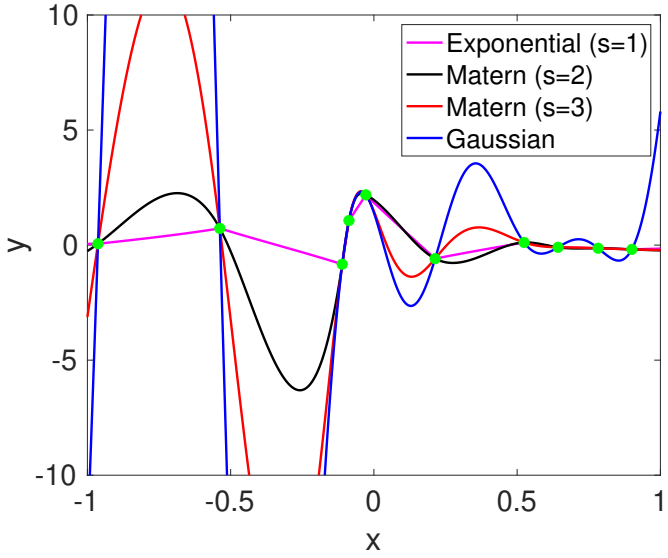
Classical examples. The main examples are the following ones:

- Exponential kernel $q(x-y) = \exp(-\alpha\|x-y\|_2)$, for which $\hat{q}(\omega) = 2^d \pi^{(d-1)/2} \Gamma((d+1)/2) \frac{\alpha}{(\alpha^2 + \|\omega\|_2^2)^{(d+1)/2}}$. See [6, page 84]. This corresponds to an RKHS norm which is penalizing all derivatives up to total order $(d+1)/2$, that is for all $\alpha \in \mathbb{N}^d$ such that $\alpha_1 + \cdots + \alpha_d = (d+1)/2$, which is a Sobolev space (fractional for d even).
- Gaussian kernel $q(x-y) = \exp(-\alpha\|x-y\|^2)$, for which $\hat{q}(\omega) = \big(\frac{\pi}{\alpha}\big)^{d/2} \exp(-\|\omega\|_2^2/(4\alpha))$. By expanding $\hat{q}(\omega)$ through its entire series as $\hat{q}(\omega) = \big(\frac{\pi}{\alpha}\big)^{d/2} \sum_{s=0}^{\infty} (-1)^s \frac{\|\omega\|_2^{2s}}{(4\alpha)^s s!}$, this corresponds to an RKHS norm which is penalizing all derivatives. Note that all members of this RKHS are infinitely differentiable.
- More generally, one can define a series of kernels so that $\hat{q}(\omega) \propto \frac{1}{(\alpha^2 + \|\omega\|_2^2)^s}$ for $s > d/2$, to ensure integrability of the Fourier transform. These so-called Matern kernels all correspond to Sobolev spaces of order s. See [6, page 84]. A key fact is that to be an RKHS, a Sobolev space has to have many derivatives when d grows.

  For $s = \frac{d+3}{2}$, we have $k(x,y) \propto (1 + \sqrt{3}\alpha\|x-y\|_2) \exp(-\sqrt{3}\alpha\|x-y\|_2)$, and for $s = \frac{d+5}{2}$, we have $k(x,y) \propto (1 + \sqrt{5}\alpha\|x-y\|_2 + \frac{5}{3}\alpha^2\|x-y\|_2^2) \exp(-\sqrt{5}\alpha\|x-y\|_2)$.
- For all the kernels below, the set $\mathcal{H}$ is dense in $L_2(dx)$ or $L_2(dp(x))$, meaning that all functions in $L_2(dx)$ or in $L_2(dp(x))$ can be approached (with respect to their corresponding norm) by a function in $\mathcal{H}$. This is made quantitative in Section 5.2.

Examples of members of RKHS. Below, we sampled $n = 10$ random points in $[-1, 1]$ with 10 random responses, and we look for the function $f \in \mathcal{H}$ such that $f(x_i) = y_i$ for all $i \in \{1 \dots, n\}$ and with minimum norm. Given the representer theorem, we can write $f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$, and the interpolation condition implies that $K\alpha = y$, and thus $y = K^{-1}\alpha$.

We consider several kernels below, going from close to piecewise affine interpolation to infinitely differentiable functions (for the Gaussian kernel).



3.4 Beyond (♦)

While the theoretical analysis of kernel methods focuses a lot on kernels on $\mathbb{R}^d$ and their link with differentiability properties of the target function, kernels can be applied to a wide variety of problems, with various input types. We give below classical examples (see more details in [7]).

- Set of subsets of a given set V: for example, the function k defined as $k(A, B) = \frac{|A \cap B|}{|A \cup B|}$ is a positive definite kernel.
- Text documents / web pages: with the usual "bag of words" assumption, we represent a text document or a web page by considering a vocabulary of "words" (this could be group of letters, single original words, or groups of words), and counting the number of occurrences of this word in the corresponding document. This gives a typically high-dimensional features $\varphi(x)$ (with dimension the size of the vocabulary). Using linear functions on this feature provide a cheap and stable predictors on such data types (better models that take into account the word order can be obtained, such as neural networks, at the expense of significantly more computational resources). See, e.g., [8] for examples.
- Sequences: given some finite alphabet $\mathcal{A}$, we consider the set $\mathcal{X}$ of finite sequences in $\mathcal{A}$ with arbitrary length. A classical infinite-dimensional feature space is indexed by $\mathcal{X}$ itself, and for $y \in \mathcal{X}$, $\varphi(x)_y$ is equal to 1 is y is a subsequence of x (we could also count the number of times the subsequence y

appears in x, or we could add a weight that depend on y, e.g., to penalize longer subsequences). This kernel has an infinite-dimensional feature space, but for two sequences x and x', we can enumerate all subsequences of x and x' and compare them in polynomial time (there exists much faster algorithms, see [9]). These kernels have many applications in bioinformatics.

The same techniques can be extended to more general combinatorial objects such as trees, graphs (see [7]).

- Images: before neural networks took over in the years 2010s with the use of large amounts of data, several kernels were designed for images, with often a "bag-of-words" assumption which provides for free invariance by translation. The key is what to consider as "words", i.e., presence of certain local patterns in the image, as well as the regions under which this assumption is made. See [10] for details.

4 Algorithms

In this section, we briefly mention algorithms aimed at solving

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \tag{2}$$

for ℓ being convex with respect to its second variable. We assume that for all $i \in \{1,\dots,n\}$, $k(x_i,x_i) = \|\varphi(x_i)\|^2 \leqslant R^2$.

Representer theorem. We can directly apply the representer theorem and try solve

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

which is a convex optimization problem.

In the special case of the square loss (ridge regression), this leads to

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{2n}\|y - K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^\top K\alpha,$$

and setting the gradient to zero, we get $(K^2 + n\lambda K)\alpha = Ky$, with a solution $\alpha = (K + n\lambda I)^{-1}y$.

However, in general (beyond square loss), it is a ill-conditioned optimization problem because K has often very small eigenvalues (more on this later), and when the loss is smooth, the Hessians are equal to $\frac{1}{n}K\,\mathrm{Diag}(h)K + \lambda K$, where $h \in \mathbb{R}^n$ is a vector of second-order derivatives of ℓ, so that the Hessians are ill-conditioned.

A better alternative is to first compute a square root of K as $K = \Phi\Phi^\top$, where $\Phi \in \mathbb{R}^{n\times m}$, and m the rank of K, and solve

$$\min_{\beta\in\mathbb{R}^m} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, (\Phi\beta)_i) + \frac{\lambda}{2}\|\beta\|_2^2,$$

with $\alpha = \Phi^\top \beta$. Note that this correspond to an explicit feature space representation (that is, the rows of Φ corresponds to features in $\mathbb{R}^n$ for the corresponding data point).

Computing a square root can be done in several ways (through Cholesky decomposition or SVD) [11], in running time $O(m^2 n)$.

Column sampling. Approximate square roots are a very useful tool, and among various algorithms, approximating $K \in \mathbb{R}^{n \times n}$ from a subset of its columns can be done as $K \approx K(V,I)K(I,I)^{-1}K(I,V)$, where $K(A,B)$ is the sub-matrix of K obtained by taking rows from the set $A \subset \{1,\dots,n\}$ and columns from $B \subset \{1,\dots,n\}$, and $V = \{1,\dots,n\}$. See below for an illustration when $I = \{1,\dots,m\}$ and a partition of the kernel matrix.

$K(I,I)$	$K(I,J)$
$K(J,I)$	$K(J,J)$

This corresponds to an approximate square root $\Phi = K(V,I)K(I,I)^{-1/2} \in \mathbb{R}^{n \times m}$, with $m = |I|$, and it can be computed in time $O(m^2 n)$ (not even the need to compute the entire kernel matrix). Then, the complexity is typically $O(m^2 n)$ instead of $O(n^3)$ (e.g., when using matrix inversion for ridge regression, for faster algorithms, see below), and is thus linear in n.

- **Exercise**: Show that this corresponds to approximating optimally each $\varphi(x_j)$, $j \notin I$, by a linear combination of $\varphi(x_i)$, $i \in I$.

This approximation technique, often called "Nyström approximation", can be analyzed when the columns are chosen randomly [12].

Random features. Some kernels have a special form that leads to specific approximation schemes, that is,

$$k(x,x') = \int_{\mathcal{V}} \varphi(x,v)\varphi(x',v) d\mu(v),$$

where $d\mu$ is a probability distribution on some space $\mathcal{V}$ and $\varphi(x,v) \in \mathbb{R}$. We can then approximate the expectation by an empirical average

$$\hat{k}(x,x') = \frac{1}{m}\sum_{i=1}^m \varphi(x,v_i)\varphi(x',v_i),$$

where the v_i's are sampled i.i.d. from $d\mu$. We can thus use an explicit feature representation $\hat{\varphi}(x) = (\frac{1}{\sqrt{m}}\varphi(x,v_i))_{i \in \{1,\dots,m\}}$, and solve

$$\min_{\beta \in \mathbb{R}^m} \frac{1}{n}\sum_{i=1}^n \ell(y_i, \hat{\varphi}(x_i)^\top \beta) + \frac{\lambda}{2}\|\beta\|_2^2.$$

For this scheme to makes sense, the number m of random features has to be significantly smaller than n, which is often sufficient in practice (see an analysis in [13]).

The two classical examples are:

- Translation-invariant kernels: $k(x,y) = q(x-y) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \hat{q}(\omega) e^{i\omega^\top(x-y)} d\omega$, for which we can take $\varphi(x,\omega) = \sqrt{q(0)} e^{i\omega^\top x} \in \mathbb{C}$, where ω is sampled from the distribution with density $\frac{1}{(2\pi)^d}\frac{q(\omega)}{q(0)}$, which is a Gaussian distribution for the Gaussian kernel. Alternatively, one can use a real-valued feature (instead of a complex-valued one) by using $\sqrt{2}\cos(\omega^\top x + b)$ with b sampled uniformly in $[0, 2\pi]$ [14].
- Neural networks with random weights: we can start from an expectation, for which the sampled features are common, e.g., $\varphi(x,v) = \sigma(v^\top x)$ for some function $\sigma : \mathbb{R} \to \mathbb{R}$. For the "rectified linear unit", that is, $\sigma(\alpha) = \max\{0,\alpha\}$, and for v sampled uniform on the sphere, we have $k(x,x') = \frac{\|x\|_2\|x'\|_2}{2(d+1)\pi}\big[(\pi-\eta)\cos\eta + \sin\eta\big]$, where $\cos\eta = \frac{x^\top x'}{\|x\|_2\cdot\|x'\|_2}$ [15]. Therefore, we can see a neural network with a large number of hidden neurons, with input weights which are random and not optimized as a kernel method. See a thorough discussion in Lectures 8 and 9.

Dual algorithms (♦). For the next two algorithms, we go back to the notation $f(x) = \langle \varphi(x), \theta\rangle$ with $\theta \in \mathcal{H}$ because it is more adapted (and is a direct infinite-dimensional extension of the algorithms from Lecture 4). To solve $\min_{\theta\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle\varphi(x_i),\theta\rangle) + \frac{\lambda}{2}\|\theta\|^2$, for a loss which is convex with respect to the second variable, we can derive a Lagrange dual in the following way (for an introduction to Lagrange duality, see [16]):

$$
\begin{aligned}
& \min_{\theta\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle\varphi(x_i),\theta\rangle) + \frac{\lambda}{2}\|\theta\|^2 \\
= & \min_{\theta\in\mathcal{H},\ u\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \ell(y_i,u_i) + \frac{\lambda}{2}\|\theta\|^2 \text{ such that } \forall i \in \{1,\dots,n\},\ \langle\varphi(x_i),\theta\rangle = u_i \\
= & \max_{\alpha\in\mathbb{R}^n} \min_{\theta\in\mathcal{H},\ u\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \ell(y_i,u_i) + \frac{\lambda}{2}\|\theta\|^2 + \lambda\sum_{i=1}^n \alpha_i\big(u_i - \langle\varphi(x_i),\theta\rangle\big) \\
= & \max_{\alpha\in\mathbb{R}^n} \Big\{ \frac{1}{n}\sum_{i=1}^n \min_{u_i\in\mathbb{R}} \big\{\ell(y_i,u_i) + n\lambda\alpha_i u_i]\big\} + \min_{\theta\in\mathcal{H}} \big\{\frac{\lambda}{2}\|\theta\|^2 - \lambda\sum_{i=1}^n \alpha_i\langle\varphi(x_i),\theta\rangle\big\}\Big\} \\
= & \max_{\alpha\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \min_{u_i\in\mathbb{R}} \big\{\ell(y_i,u_i) + n\lambda\alpha_i u_i\big\} - \frac{1}{2\lambda}\Big\|\sum_{i=1}^n \alpha_i\varphi(x_i)\Big\|^2 \\
= & \max_{\alpha\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \min_{u_i\in\mathbb{R}} \big\{\ell(y_i,u_i) + n\lambda\alpha_i u_i\big\} - \frac{1}{2\lambda}\alpha^\top K\alpha,
\end{aligned}
$$

with $\theta = \sum_{i=1}^n \alpha_i\varphi(x_i)$ at optimum. Since the functions $\alpha_i \mapsto \min_{u_i\in\mathbb{R}}\big\{\ell(y_i,u_i) + n\lambda\alpha_i u_i\big\}$ are concave (as minima of affine functions), this is a concave maximization problem.

Note the similarity with the representer theorem (existence of $\alpha \in \mathbb{R}^n$ such that $\theta = \sum_{i=1}^n \alpha_i\varphi(x_i)$) and the dissimilarity (one is a minimization problem, one is maximization problem). Moreover, when the loss

is smooth, one can show that the function $\min_{u_i \in \mathbb{R}} \{\ell(y_i, u_i) + n\lambda\alpha_i u_i\}$ is a strongly concave function, and thus relatively easy to optimize (in other words, the associated condition numbers are smaller),

- **Exercise**: (a) for ridge regression, compute the dual problem and compare the condition number of the primal problem and the condition number of the primal problem; (b) compare the two formulations to using normal equations like in Lecture 2, and relate the two using the matrix inversion lemma $(\Phi\Phi^\top + n\lambda I)^{-1}\Phi = \Phi(\Phi^\top\Phi + n\lambda I)^{-1}$.

SGD (♦). When minimizing an expectation

$$\min_{\theta \in \mathcal{H}} \mathbb{E}\big[\ell(y, \langle \varphi(x), \theta\rangle)\big] + \frac{\lambda}{2}\|\theta\|^2$$

like in Lecture 4, the stochastic gradient algorithm leads to the recursion

$$\theta_t = \theta_{t-1} - \gamma_t\big[\ell'(y_t, \langle\varphi(x_t), \theta_{t-1}\rangle)\varphi(x_t) + \lambda\theta_{t-1}\big],$$

where (x_t, y_t) is an i.i.d. sample from the distribution defining the expectation, and ℓ' is the derivative with respect to the second variable.

When initializing at $\theta_0 = 0$, then θ_t is a linear combination of all $\varphi(x_i)$, $i = 1, \dots, t$, and thus we can write

$$\theta_t = \sum_{i=1}^{t} \alpha_i^{(t)} \varphi(x_i),$$

with $\alpha^{(0)} = 0$, and the recursion in α as

$$\alpha_i^{(t)} = (1 - \gamma_t\lambda)\alpha_i^{(t-1)} \text{ for } i \in \{1, \dots, t-1\}, \text{ and } \alpha_t^{(t)} = -\gamma_t \ell'\Big(y_t, , \sum_{i=1}^{t-1} \alpha_i^{(t-1)} k(x_t, x_i)\Big).$$

The complexity after t iterations is $O(t^2)$ kernel evaluations. The convergences rates from Lecture 4 apply. More precisely, if the loss is G-Lipschitz continuous, then, for $F(\theta) = \mathbb{E}\big[\ell(y, \langle\varphi(x), \theta\rangle)\big] + \frac{\lambda}{2}\|\theta\|^2$, we have, for the averaged iterate θ_t,

$$\mathbb{E}\big[F(\bar{\theta}_t)\big] - \inf_{\theta \in \mathcal{H}} F(\theta) \leqslant \frac{G^2R^2}{\lambda t}.$$

When doing a single pass with $t = n$, then $F(\theta)$ is the regularized expected risk, and we obtain a generalization bound, leading to $\mathbb{E}\big[\mathcal{R}(f_{\bar{\theta}_t})\big] \leqslant \frac{G^2R^2}{\lambda n} + \inf_{f \in \mathcal{H}} \big\{\mathcal{R}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2\big\}$. These bounds are similar than the ones below.

"Kernelization" of linear algorithms. Beyond supervised learning, many unsupervised learning algorithms can be "kernelized", such as principal component analysis or canonical correlation analysis. See [5, 7] for details.

5 Generalization guarantees - Lipschitz-continuous losses

In this section, we consider a G-Lipschitz-continuous loss function, and consider a minimizer $\hat{f}_D^{(c)}$ of the constrained problem

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) \text{ such that } \|f\|_{\mathcal{H}} \leqslant D, \tag{3}$$

and the unique minimizer $\hat{f}_\lambda^{(r)}$ of the regularized problem

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2. \tag{4}$$

We denote by $\mathcal{R}(f) = \mathbb{E}\big[\ell(y, f(x))\big]$ the expected risk, and by f^* one of its minimizers (which we assume to be square integrable). We assume $k(x,x) \leqslant R^2$ almost surely.

Note that we have

$$\mathcal{R}(f)-\mathcal{R}(f^*) \leqslant \mathbb{E}[|\ell(y,f(x))-\ell(y,f^*(x))|] \leqslant G\mathbb{E}[|f(x)-f^*(x)|] \leqslant G\sqrt{\mathbb{E}[|f(x)-f^*(x)|^2]} = G\|f-f^*\|_{L_2(dp(x))},$$

that is, the excess risk is dominated by the ℓ_2-norm of $f - f^*$.

5.1 Risk decomposition

Constrained problem. Dimension-free results from Lecture 3, based on Rademacher complexities immediately apply, and we obtain that the estimation error is bounded from above by $\frac{2GDR}{\sqrt{n}}$, leading to:

$$\mathbb{E}\big[\mathcal{R}(\hat{f}_D^{(c)})\big] - \mathcal{R}(f^*) \leqslant \frac{2GDR}{\sqrt{n}} + G \inf_{\|f\|_{\mathcal{H}}\leqslant D} \|f - f^*\|_{L_2(dp(x))},$$

(the first term is the estimation error, the second term is the approximation error).

In order to find the optimal D (to balance estimation and approximation error), we can minimize the bound with respect to D, leading to, using Lagrange duality:

$$\begin{aligned}
\inf_{D\geqslant 0} \frac{2GRD}{\sqrt{n}} + G \inf_{\|f\|_{\mathcal{H}}\leqslant D} \|f-f^*\|_{L_2(dp)} &= \inf_{D\geqslant 0} \frac{2GBD}{\sqrt{n}} + G\sup_{\lambda\geqslant 0}\inf_{f\in\mathcal{H}} \|f-f^*\|_{L_2(dp(x))} + \sqrt{\lambda}(\|f\|_{\mathcal{H}} - D) \\
&\leqslant \sup_{\lambda\geqslant 0}\inf_{D\geqslant 0} GD\big[\frac{2R}{\sqrt{n}} - \sqrt{\lambda}\big] + 2G\sqrt{\inf_{f\in\mathcal{H}}\big\{\|f-f^*\|_{L_2(dp(x))}^2 + \lambda\|f\|_{\mathcal{H}}^2\big\}} \\
&\leqslant \sup_{\lambda\geqslant 0} G\sqrt{\inf_{f\in\mathcal{H}}\big\{\|f-f^*\|_{L_2(dp(x))}^2 + \lambda\|f\|_{\mathcal{H}}^2\big\}} \text{ such that } \lambda \leqslant \frac{2R}{\sqrt{n}}, \\
&\leqslant 2G\sqrt{\inf_{f\in\mathcal{H}}\big\{\|f-f^*\|_{L_2(dp(x))}^2 + \frac{4R^2}{n}\|f\|_{\mathcal{H}}^2\big\}},
\end{aligned}$$

with $\lambda^* = \frac{4R^2}{n}$ (note that this is not the a regularization parameter to be used in an algorithm).

Overall, we need to undertand how the deterministic quantity

$$A(\lambda, f^*) = \inf_{f \in \mathcal{H}} \{ \|f - f^*\|^2_{L_2(dp)} + \lambda \|f\|^2_{\mathcal{H}} \}$$

goes to zero when λ goes to zero. This can only be possible if the space $\mathcal{H}$ is dense in $L_2(d\mu)$, which will be the case for translation-invariant kernels in Section 5.2 below.

Otherwise, denoting $\Pi_{\bar{\mathcal{H}}}(f^*)$ the orthogonal projection in $L_2(dp(x))$ of f^* on the closure of $\mathcal{H}$, by Pythagoreas theorem, $A(\lambda, f^*) = A(\lambda, \Pi_{\bar{\mathcal{H}}}(f^*)) + \|f^* - \Pi_{\bar{\mathcal{H}}}(f^*)\|^2_{L_2(dp(x))}$.

Regularized problem (⧫). For the regularized problem, we can use the bound from Lecture 3:

$$\mathbb{E}\big[\mathcal{R}(\hat{f}^{(r)}_\lambda)\big] - \mathcal{R}(f^*) \quad \leqslant \quad \frac{32 G^2 R^2}{\lambda n} + \inf_{f \in \mathcal{H}} \big\{ G \|f - f^*\|_{L_2(dp(x))} + \frac{\lambda}{2} \|f\|^2_{\mathcal{H}} \big\}.$$

We can now minimize the bound with respect to λ as $\lambda^* = \frac{8RG}{\sqrt{n}}$, to obtain the bound:

$$G \inf_{f \in \mathcal{H}} \big\{ \|f - f^*\|_{L_2(dp(x))} + \frac{8R}{\sqrt{n}} \|f\|_{\mathcal{H}} \big\} \leqslant 2G \sqrt{\inf_{f \in \mathcal{H}} \big\{ \|f - f^*\|^2_{L_2(dp)} + \frac{64 R^2}{n} \|f\|^2_{\mathcal{H}} \big\}},$$

which is the same bound as for constrained problem, but on a more commonly used optimization problem in practice.

5.2 Approximation error for translation-invariant kernels on $\mathbb{R}^d$

We first start with the analysis of the approximation error of kernel methods for translation invariant kernels. Given a distribution $dp(x)$, the goal is to compute

$$A(\lambda, f^*) = \inf_{f \in \mathcal{H}} \|f - f^*\|^2_{L_2(dp(x))} + \lambda \|f\|^2_{\mathcal{H}},$$

where f^* is the target function (e.g., the minimizer of the test risk), which we assume squared-integrable. If $A(\lambda, f^*)$ tends to zero when λ tends to zero for any fixed f^*, then kernel-based supervised learning leads to universally consistent algorithms.

We assume that $\|f - f^*\|^2_{L_2(dp(x))} \leqslant C \|f - f^*\|^2_{L_2(dx)}$ (e.g., with $C = \|p\|_\infty$ where p is the density of $dp(x)$). Moreover, for simplicity, we assume that $\|f^*\|_{L_2(dx)}$ is finite (that is, f^* is not allowed to explode at infinity). We now give bounds on

$$\widetilde{A}(\lambda, f^*) = \inf_{f \in \mathcal{H}} \|f - f^*\|^2_{L_2(dx)} + \lambda \|f\|^2_{\mathcal{H}},$$

Explicit approximation. We have, for translation-invariant kernels, $\|f\|^2_{\mathcal{H}} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)} d\omega$, and thus

$$\widetilde{A}(\lambda, f^*) \quad = \quad \inf_{\hat{f} \in L_2(dx)} \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \Big[|\hat{f}(\omega) - \hat{f}^*(\omega)|^2 + \lambda \frac{|\hat{f}(\omega)|^2}{\hat{q}(\omega)} \Big] d\omega,$$

with solution $\hat{f}_\lambda(\omega) = \frac{\hat{f}^*(\omega)}{1+\lambda\hat{q}(\omega)^{-1}}$, and value:

$$\widetilde{A}(\lambda, f^*) \quad = \quad \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\Big[|\hat{f}^*(\omega)|^2\Big(1-\frac{1}{1+\lambda\hat{q}(\omega)^{-1}}\Big)\Big]d\omega = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\Big[|\hat{f}^*(\omega)|^2\frac{\lambda}{\hat{q}(\omega)+\lambda}\Big]d\omega.$$

When λ goes to zero, we see that for each ω, $\hat{f}_\lambda(\omega)$ tends to $\hat{f}(\omega)$. By the dominated convergence theorem, $\widetilde{A}(\lambda, f^*)$ goes to zero, when λ goes to zero.

Without further assumptions it is not possible to obtain a rate of convergence (otherwise the no-free lunch theorem from Lecture 1 would be invalidated). However, this is possible when assuming regularity properties for f^*.

Sobolev spaces (♦). If we assume that $\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}(1+\|\omega\|_2^2)^t|\hat{f}^*(\omega)|^2d\omega$ is finite for some $t>0$, that is, for f^* with squared integrable partial derivatives up to order t, then we can further bound:

$$\widetilde{A}(\lambda, f^*) \quad \leqslant \quad \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}(1+\|\omega\|_2^2)^t|\hat{f}^*(\omega)|^2d\omega \times \sup_{\omega\in\mathbb{R}^d}\Big\{\frac{\lambda}{\hat{q}(\omega)+\lambda}\frac{1}{(1+\|\omega\|_2^2)^t}\Big\}.$$

If we now assume $\hat{q}(\omega) \propto (1+\|\omega\|_2^2)^{-s}$ (Matern kernels), with $s > d/2$ to get an integrable function, then with $t \geqslant s$, $f^* \in \mathcal{H}$, and have $\widetilde{A}(\lambda, f^*) = \lambda\|f^*\|_{\mathcal{H}}^2$. With $t < s$, that is the function is not inside the RKHS $\mathcal{H}$, then we get a bound proportional to

$$\sup_{\omega\in\mathbb{R}^d}\Big\{\frac{\lambda}{\hat{q}(\omega)+\lambda}\frac{1}{(1+\|\omega\|_2^2)^t}\Big\} \leqslant \sup_{\omega\in\mathbb{R}^d}\Big\{\frac{\lambda}{\hat{q}(\omega)^{t/s}\lambda^{1-t/s}}\frac{1}{(1+\|\omega\|_2^2)^t}\Big\} = O(\lambda^{t/s}).$$

- **Exercise** (♦): Find an upper-bound of $\widetilde{A}(\lambda, f^*)$ for the same assumption on f^* but with the Gaussian kernel.

There are two regularities: $t \geqslant 0$ for the target function, and $s > d/2$ for the kernel.

Putting things together. Thus, for Lipschitz-continuous losses, we get an expected excess risk of the order $\sqrt{\widetilde{A}(R^2/n, f^*)} = O(\frac{1}{n^{t/(2s)}})$, when $t \leqslant s$.

Approximation bounds (♦). Note that a bound in leads to bounds on the quantity $A(\lambda, f^*) = \inf_{f\in\mathcal{H}}\{\|f-f^*\|_{L_2(dp)}^2 + \lambda\|f\|_{\mathcal{H}}^2\}$ of the form $c\lambda^\alpha$ for $\alpha \in (0,1)$ leads to the following bound:

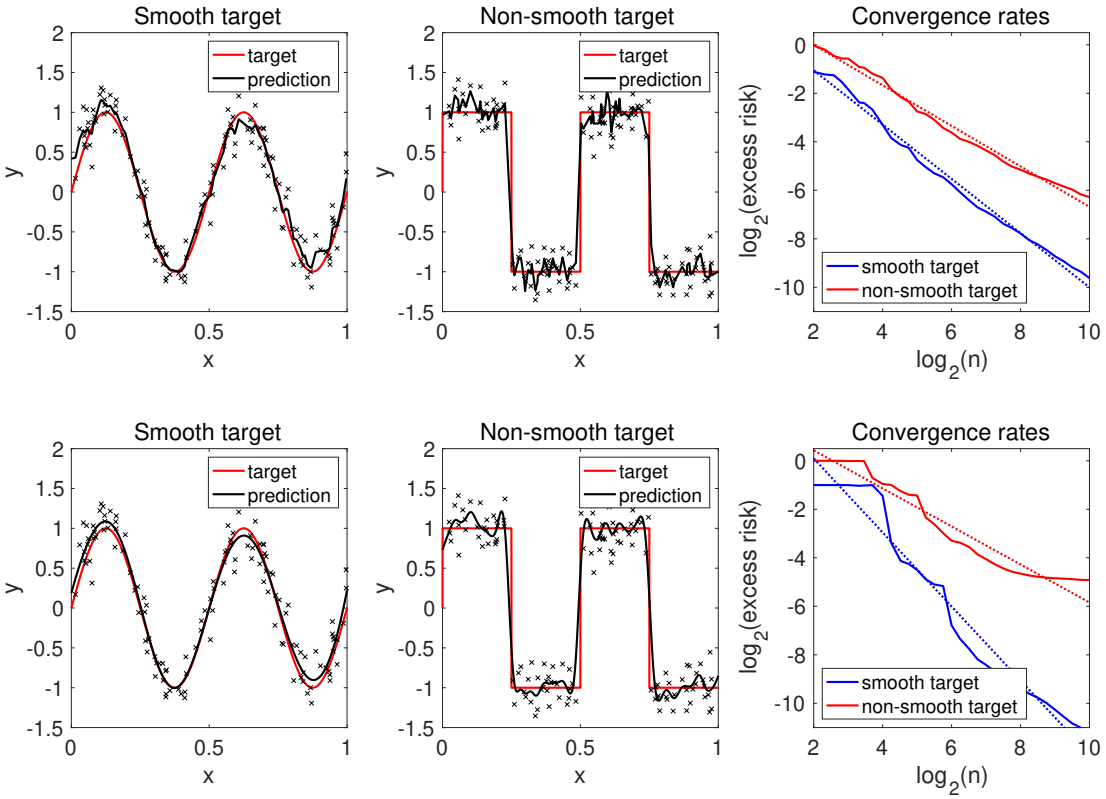
$$\begin{aligned}
\inf_{f\in\mathcal{H}}\|f\|_{\mathcal{H}}^2 \text{ such that } \|f-f^*\|_{L_2(dx)} \leqslant \varepsilon \quad &= \quad \inf_{f\in\mathcal{H}}\|f\|_{\mathcal{H}}^2 + \mu(\|f-f^*\|_{L_2(dx)}^2 - \varepsilon^2) \\
&= \quad \inf_{f\in\mathcal{H}}\sup_{\mu\geqslant 0}\|f\|_{\mathcal{H}}^2 + \mu(\|f-f^*\|_{L_2(dx)}^2 - \varepsilon^2) \\
&= \quad \sup_{\mu\geqslant 0}\mu A(\mu^{-1}, f^*) - \mu\varepsilon^2 \\
&\leqslant \quad \sup_{\mu\geqslant 0}\mu c\mu^{-\alpha} - \mu\varepsilon^2.
\end{aligned}$$

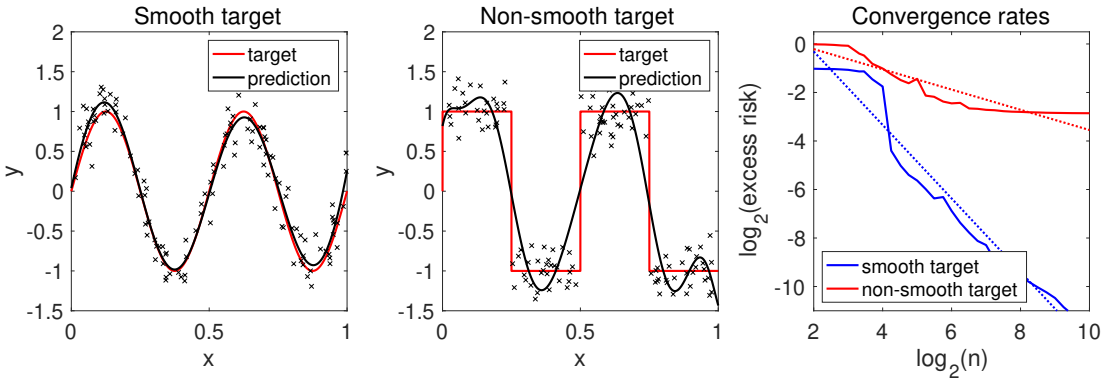
The optimal μ is such that $(1-\alpha)c\mu^{-\alpha} = \varepsilon^2$, leading to an approximation bound proportion to $\varepsilon^{2(1-1/\alpha)} = \varepsilon^{-2(1-\alpha)/\alpha}$.

Applied to $\alpha = t/s$ like before, this leads to an $RKHS$ norm proportional to $\varepsilon^{-(1-\alpha)/\alpha}$ to get an error less than $\|f - f^*\|_{L_2(dx)}$. So where $t = 1$ (single derivative for the target function), and $s > d/2$ (for the Sobolev kernel), we get a norm of the order $\varepsilon^{-(1/\alpha-1)} = \varepsilon^{-(s-1)} \geqslant \varepsilon^{-d/2+1}$, which explodes exponentially in dimension, which is another way of formulating the curse of dimensionality.

6 Experiments

We consider one-dimensional problems to highlight the adaptivity of kernel methods to the regularity of the target function, with one smooth target and one non-smooth target, and two kernels: exponential kernel corresponding to the Sobolev space of order 1 (top), Matern kernel corresponding to the Sobolev space of order 3 (middle), and Gaussian kernel (bottom). In the right plots, dotted lines are affine fits to the log-log learning curves. The regularization parameter for ridge regression is selected to minimize expected risk, and learning curves are obtained by averaging over 20 replications.





We observe adaptivity for the three kernels: learning is possible even with irregular function, and the rates are better for the smooth target function. We also note that for kernels with smaller feature spaces (Matern and Gaussian), the performance on the non-smooth target function is worse than for the large feature space (exponential kernel).

Acknowledgements

These class notes have been adapted from the notes of many colleagues I have the pleasure to work with, in particular Lénaïc Chizat, Pierre Gaillard, Alessandro Rudi and Simon Lacoste-Julien. Special thanks to Alessandro for his help.

References

[1] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Jour. Math. Anal. Appl.*, 33:82–95, 1971.

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68:337 – 404, 1950.

[3] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, volume 3. Springer, 2004.

[4] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics, Volume 2.* Academic press, 1978.

[5] B. Schölkopf and A. J. Smola. *Learning with Kernels.* MIT Press, 2001.

[6] Christopher K. I. Williams and Carl Edward Rasmussen. *Gaussian Processes for Machine Learning.* MIT press Cambridge, MA, 2006.

[7] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

[8] Armand Joulin, Édouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, 2017.

[9] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[10] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

[11] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[12] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

[13] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.

[14] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

[15] Nicolas Le Roux and Yoshua Bengio. Continuous neural networks. In *Artificial Intelligence and Statistics*, pages 404–411, 2007.

[16] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[17] Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

[18] Zaid Harchaoui, Francis Bach, and Eric Moulines. Testing for homogeneity with kernel fisher discriminant analysis. Technical Report 0804.1026, arXiv, 2008.

[19] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.