4.3. Approximation error

Further decomposition

- **Reminder:** approximation error is defined as

$$\inf_{f \in \mathcal{H}} \mathcal{R}(f) - \mathcal{R}^\star .$$

- deterministic, small if function class is large
- let us focus on parametric models, in particular $\mathcal{H} = \{f_\theta, \theta \in \Theta\}$
- $\theta^\star$ parameter corresponding to $f^\star$
- typically does not belong to Θ!
- further decomposition of the approximation error:

$$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \mathcal{R}^\star = \left(\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^p} \mathcal{R}(f_\theta) \right) + \left(\inf_{\theta \in \mathbb{R}^p} \mathcal{R}(f_\theta) - \mathcal{R}^\star \right) .$$

- **Remark:** both positive terms

Incompressible approximation error

- **Recall:**

$$\inf_{\theta\in\Theta} \mathcal{R}(f_\theta) - \mathcal{R}^\star = \left(\inf_{\theta\in\Theta} \mathcal{R}(f_\theta) - \inf_{\theta\in\mathbb{R}^p} \mathcal{R}(f_\theta)\right) + \left(\inf_{\theta\in\mathbb{R}^p} \mathcal{R}(f_\theta) - \mathcal{R}^\star\right) .$$

- let us start with the second term
- for rich model class, this **goes to zero**

Upper bounds

- now focus on $\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^p} \mathcal{R}(f_\theta)$
- this term is typically upper bounded by a **distance** between the best candidate in Θ and the best candidate in $\mathbb{R}^d$
- **Example:** $f_\theta(x) = \theta^\top \varphi(x)$, $\Theta = \{\theta \in \mathbb{R}^d, \|\theta\| \leq D\}$
- for a L-Lipschitz loss, we write

$$\begin{aligned}\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^p} \mathcal{R}(f_\theta) &= \mathbb{E}\left[\ell(\theta_1^\top \varphi(X), Y) - \ell((\theta^\star)^\top \varphi(X), Y)\right] \\ &\leq L\mathbb{E}\left[\|\varphi(X)\| \cdot \|\theta_1 - \theta^\star\|\right] \\ &\leq L\mathbb{E}\left[\|\varphi(X)\|\right] \cdot (\|\theta^\star\| - D)_+ \,.\end{aligned}$$

- **Remark:** equal to zero if $\|\theta^\star\| \leq D$ (well-specified model)

5. Kernel methods

5.1. Positive semi-definite kernels

Representation of the data

- **What we have seen so far:** linear classification / linear regression
- works well if the data is linearly separable
- **Problem:** that is not always the case!
- what if we could transport the data to another space where it is well-behaved?
- for instance a very high-dimensional space
- first we define a (positive-definite) *kernel*
- **many** definitions in maths, introduced in machine learning by Aizerman, Braverman, and Rozonoer in the 60s[7]

[7] Aizerman, Braverman, Rozonoer, *Theoretical foundations of the potential function method in pattern recognition learning*, Automation and Remote Control, 1964

Positive semi-definite kernels

Definition: a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *positive semi-definite kernel* if $k(x, x') = k(x', x)$ for any $x, x' \in \mathcal{X}$, and

$$\forall x_1, \ldots, x_n \in \mathcal{X}, \forall c_1, \ldots, c_n \in \mathbb{R}, \quad \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \,.$$

- in other words, the Gram matrix $K = (k(x_i, x_j))_{i,j=1}^{n}$ is positive definite for any input data $x_1, \ldots, x_n$
- *kernel methods* take this K as input
- **Remark:** this is *costly*, $\mathcal{O}\left(n^2\right)$ whatever we do, with possible dependency in the dimensionality of the data
- Beware: unlike the name suggests, k has no reason to be *positive*

Fundamental example

- suppose that $\mathcal{X} = \mathbb{R}$
- then $k(x, y) := xy$ is a positive definite kernel
- **Why?** first, we check that $k(x, y) = k(y, x)$
- second, let $n \geq 1$, $x_1, \dots, x_n \in \mathbb{R}^d$, and $c_1, \dots, c_n \in \mathbb{R}$, then

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j x_i x_j \\ &= \left(\sum_{i=1}^{n} c_i x_i\right)^2 \\ &\geq 0\,.\end{aligned}$$

Fundamental example, ctd.

- we can extend this example: set $k(x, y) := x^\top y$ on $\mathcal{X} = \mathbb{R}^d$
- let $n \geq 1$, $x_1, \ldots, x_n \in \mathbb{R}^d$, and $c_1, \ldots, c_n \in \mathbb{R}$, then

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j x_i^\top x_j \\ &= \left\| \sum_{i=1}^{n} c_i x_i \right\|^2 \\ &\geq 0 \,. \end{aligned}$$

- $k(x, y) := x^\top y$ is usually called the **linear kernel**
- **Intuition:** kernels are a generalization of inner product

Other examples

- **Polynomial kernel:**

$$\mathcal{X} = \mathbb{R}^d, \qquad k(x, y) = (x^\top y + c)^k .$$

- **min kernel:**

$$\mathcal{X} = \mathbb{R}, \qquad k(x, y) = \min(x, y) .$$

- **Gaussian kernel:**

$$\mathcal{X} = \mathbb{R}^d, \qquad k(x, y) = \exp\left(\frac{-\|x - y\|^2}{2\nu^2}\right) .$$

- **Exponential kernel:**

$$\mathcal{X} = \mathbb{R}^d, \qquad k(x, y) = \exp\left(\frac{-\|x - y\|}{2\nu}\right) .$$

- ...

Choosing the bandwidth

- Gaussian and Laplace kernel: one has to choose the bandwidth parameter ν
- indeed, if ν is *too large* with respect to the typical value of $\|x_i - x_j\|$, then $K \approx I_n$
- in the other direction, if ν is *too small*, then $K \approx \mathbf{1}\mathbf{1}^\top$
- both cases are degenerate: whatever we do with K is not going to work very well
- one possible solution: **median heuristic**[8]

$$\nu = \mathrm{Med}\{\|x_i - x_j\|\,,\quad 1 \leq i,j \leq n\}\,.$$

- preferable to the mean (too sensitive to extreme values)
- we can pick other quantiles

[8] Garreau, Jitkrittum, Kanagawa, *Large sample analysis of the median heuristic*, 2017

Hilbert spaces

Definition: A *Hilbert space* is a real or complex vector space which is also a complete metric space with respect to the distance function induced by the inner product.

- **Remark:** recall the linear kernel, all we used were properties of inner product
- let $\Phi : \mathcal{X} \to \mathcal{H}$ be some mapping, $\mathcal{H}$ a Hilbert space with scalar product $\langle \cdot, \cdot \rangle$
- then $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ is positive definite:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \langle \Phi(x), \Phi(y) \rangle = \left\| \sum_{i=1}^{n} c_i \Phi(x_i) \right\|^2 \geq 0 \,,$$

by linearity of the inner product.

Kernel as inner products

- **Remarkable fact:** the converse statement is true!

Theorem:[9] For any kernel k on $\mathcal{X}$, there exists a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that

$$\forall x, y \in \mathcal{X}, \qquad k(x, y) = \langle \Phi(x), \Phi(y) \rangle \,.$$

- **Reminder:** Hilbert space = inner product + *complete* for the associated norm (Cauchy sequences converge in $\mathcal{H}$)
- **Consequence:** we can think of any kernel as a dot product in the *feature space*
- **Main idea:** forget about Φ and work only with kernel evaluations (more on that later)

[9] Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society, 1950

Proof in the finite case

- assume that $\mathcal{X} = \{x_1, \ldots, x_N\}$ is finite of size N
- any kernel k is entirely defined by the $N \times N$ positive semi-definite matrix $K := (k(x_i, x_j))_{i,j=1}^{N}$
- we can diagonalize K in an orthonormal basis $(u_1, \ldots, u_N)$ with associated (non-negative) eigenvalues $\lambda_1, \ldots, \lambda_N$: $K = U \Lambda U^\top$, with $U_{:,i} = u_i$, $\Lambda = \operatorname{diag}(\lambda)$, $UU^\top = U^\top U = \mathrm{I}$
- then we write

$$k(x_i, x_j) = \left(\sum_{\ell=1}^{N} \lambda_\ell u_\ell u_\ell^\top \right)_{i,j}$$

$$= \sum_{\ell=1}^{N} \lambda_\ell (u_\ell)_i (u_\ell)_j = \langle \Phi(x_i), \Phi(x_j) \rangle ,$$

with

$$\Phi(x_i) := \left(\sqrt{\lambda_1} (u_1)_i, \cdots, \sqrt{\lambda_n} (u_N)_i \right)^\top .$$

□

5.2. Reproducing kernel Hilbert spaces

Function spaces

- among all spaces in the previous statement, one of them has interesting properties
- in particular, it is a **space of functions**
- *i.e.*, we can map each point $x \in \mathcal{X}$ to a *function* $\Phi(x) = k_x \in \mathcal{H}$
- **Example:** $\mathcal{X} = \mathbb{R}$, we map each x to the function $t \mapsto xt$
- $\rightarrow$ space of linear functions
- more complicated in general...

Reproducing Kernel Hilbert Space (RKHS)

Definition: let $\mathcal{X}$ be a set and $\mathcal{H}$ be a function space forming a Hilbert space with inner product $\langle \cdot, \cdot \rangle$. The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if

- $\mathcal{H}$ contains all functions of the form $k_x : t \mapsto k(x, t)$
- for every $x \in \mathcal{X}$ and $f \in \mathcal{H}$, the *reproducing property* holds:

$$f(x) = \langle f, k_x \rangle \, .$$

- if a reproducing kernel exists, then $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS)

Equivalent definition

Theorem: the Hilbert space $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ is a RKHS if, and only if, for any $x \in \mathcal{X}$, the mapping $f \mapsto f(x)$ is continuous.

- *Proof of* $\Rightarrow$*:* let k be a reproducing kernel, $x \in \mathcal{X}$ and $f_n \to f$ in $\mathcal{H}$
- we write

$$\begin{aligned} |f_n(x) - f(x)| &= |\langle f_n - f, k_x \rangle| \\ &\leq \|f_n - f\| \cdot \|k_x\| \end{aligned}$$

  by Cauchy-Schwarz inequality.
- $\|f_n - f\| \to 0$ and we can conclude
- **Remark:** $\|k_x\|^2 = \langle k_x, k_x \rangle = k(x, x)$, thus $|f(x)| \leq \|f\| \cdot k(x, x)^{1/2}$

Continuity ctd.

- *Proof of* $\Leftarrow$*:* let $x \in \mathcal{X}$
- by the reproducing property, $L : x \mapsto f(x)$ is a *linear functional*
- Riesz theorem: there exists ℓ_x such that $L(x) = \langle f, \ell_x \rangle$
- define $k(x, y) := \ell_y(x)$
- one can check readily the RKHS properties. $\square$

Uniqueness

Theorem: if $\mathcal{H}$ is a RKHS, then it has a unique reproducing kernel. Conversely, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be the reproducing kernel of at most one RKHS.

- we talk about *the* RKHS associated to k
- *Proof:* let k and k' be two reproducing kernels
- then for all $x \in \mathcal{X}$,

$$\begin{aligned} \|k_x - k'_x\|^2 &= \langle k_x - k'_x, k_x - k'_x \rangle \\ &= k_x(x) - k'_x(x) - k_x(x) + k'_x(x) \\ &= 0 \end{aligned}$$

□

Equivalence psd / RKHS

Theorem: a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if, and only if, it is a reproducing kernel.

- **Idea:** build $\mathcal{H}$ as the completion of

$$\mathcal{H}_0 := \left\{ \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_i \in \mathcal{X} \right\}$$

- **Remark:** showing that a kernel is positive definite is enough to get Φ and $\mathcal{H}$ with the reproducing property "for free"

Example

- **Example:** polynomial kernel of degree 2:

$$k(x,y) = (x^\top y)^2 \,.$$

- proved during the exercise session:

$$k(x,y) = \langle xx^\top, yy^\top \rangle_F \,,$$

  thus k is positive definite
- **Question:** what is the RKHS?
- we know that $\mathcal{H}$ contains all the functions

$$f(x) = \sum_i a_i k(x_i, x) = \sum_i a_i \langle x_i x_i^\top, xx^\top \rangle = \langle \sum_i a_i x_i x_i^\top, xx^\top \rangle$$

Example, ctd.

- spectral theorem: any symmetric matrix can be decomposed as $\sum_i a_i x_i x_i^\top$
- candidate RKHS: set a quadratic functions

$$f_S(x) = \langle S, xx^\top \rangle = x^\top S x\,,$$

  with S symmetric matrix of size $d \times d$
- inner product on $\mathcal{H}$:

$$\langle f_S, f_{S'} \rangle = \langle S, S' \rangle_F\,.$$

- we can check that $\mathcal{H}$ is a Hilbert space (isomorphic to $\mathcal{S}^{d\times d}$)
- finally, we check the reproducing property

5.3. More examples

Elementary properties

Proposition: Let $k_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a (potentially infinite) family of p.d. kernels. Then

- for any $\lambda_1, \ldots, \lambda_p \geq 0$, the sum $\sum_{i=1}^{p} \lambda_i k_i$ is positive definite
- for any $a_1, \ldots, a_p \in \mathbb{N}$, the product $k_1^{a_1} \cdots k_p^{a_p}$ is positive definite
- if it exists, the limit $k = \lim_{p \to +\infty} k_p$ is positive definite

Moreover, let $\mathcal{X}_i$ be a sequence of sets and k_i positive kernels on each $\mathcal{X}_i$. Then

$$k((x_1, \ldots, x_p), (y_1, \ldots, y_p)) := \prod_{i=1}^{p} k_i(x_i, y_i)$$

and

$$k((x_1, \ldots, x_p), (y_1, \ldots, y_p)) := \sum_{i=1}^{p} k_i(x_i, y_i)$$

are positive definite kernels.

Taking the exponential

Theorem: if k is a positive definite kernel, then e^k as well.

- *Proof:* we write
$$\mathrm{e}^{k(x,y)} = \lim_{n\to+\infty} \sum_{p=0}^{n} \frac{k(x,y)^p}{p!} \,,$$
then reason step by step.
- by the product property, $k(x,y)^p$ is a kernel for any $p \geq 0$
- as a positive linear combination of kernels, $\sum_{p=0}^{n} \frac{k(x,y)^p}{p!}$ is a kernel for all $n \geq 1$
- finally, e^k is a kernel as a limit of kernels. □

5.4. The kernel trick and applications

The kernel trick

- input data $x_1, \ldots, x_n \in \mathcal{X}$
- $k : \mathcal{X} \times \mathcal{X}$ kernel with associated RKHS $\mathcal{H}$
- we call $\Phi : \mathcal{X} \to \mathcal{H}$ the feature map
- **Idea:** imagine that our algorithm only depends on scalar products $x_i^\top x_j$
- then we can map the x_i to $\mathcal{H}$ and replace the inner products by kernel evaluations, since

$$\langle \Phi(x_i), \Phi(x_j) \rangle = k(x_i, x_j)\,.$$

- simple “trick” with many, many applications

Example

- **Example:** computing distances
- suppose that our algo relies on distance computation
- that is, $\|x - y\|^2$
- we can write

$$\begin{aligned}\|\Phi(x) - \Phi(y)\|^2 &= \langle \Phi(x) - \Phi(y), \Phi(x) - \Phi(y) \rangle \\ &= \langle \Phi(x), \Phi(x) \rangle - 2\langle \Phi(x), \Phi(y) \rangle + \langle \Phi(y), \Phi(y) \rangle \\ \|\Phi(x) - \Phi(y)\|^2 &= k(x,x) - 2k(x,y) + k(y,y)\,.\end{aligned}$$

- in other words,

$$d_{\mathcal{H}}(x,y) = \sqrt{k(x,x) - 2k(x,y) + k(y,y)}\,.$$

- as promised, **we do not need to know Φ!**

5.5. The representer theorem

Motivation

- let us imagine that we take $\mathcal{H}$ as hypothesis class
- starting from regularized ERM, our optimization problem will look like

$$\arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \left\| f \right\|^2 \right\} . \qquad (\star)$$

- we penalize by the norm because it is an indicator of the *smoothness* of f
- **Why?** Cauchy-Schwarz + exercise:

$$|f(x) - f(y)| = |\langle f, k_x - k_y \rangle| \leq \|f\| \cdot \|k_x - k_y\| = \|f\| \cdot d_{\mathcal{H}}(x, y) .$$

- Eq. $(\star)$ is a complicate problem, potentially *infinite-dimensional*
- **Question:** how to solve it in practice?

The representer theorem

Theorem: let $\mathcal{H}$ be the RKHS associated to k defined on $\mathcal{X}$. Let $S = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ be a finite set of points. Let $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ be a function, increasing in the last variable. Then any solution to the minimization problem

$$\arg\min_{f \in \mathcal{H}} \Psi(f(x_1), \ldots, f(x_n), \|f\|)$$

admits a representation of the form

$$\forall x \in \mathcal{X}, \qquad f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)\,.$$

- **Main consequence:** Eq. $(\star)$ is actually a finite-dimensional problem (!)

Practical use

- recall that we defined $K := (k(x_i, x_j))_{i,j=1}^n$
- before turning to concrete examples, we notice that we can simply express the key quantities
- for instance, for any $1 \leq j \leq n$,

$$f(x_j) = \sum_{i=1}^n \alpha_i k(x_i, x_j) = (K\alpha)_j \, .$$

- in the same way,

$$\|f\|^2 = \left\| \sum_{i=1}^n \alpha_i k(x_i, \cdot) \right\|^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) = \alpha^\top K \alpha \, .$$

5.6. Kernel ridge regression

Kernel Ridge Regression[10] (KRR)

- regression setting: $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$
- $\mathcal{Y} \subseteq \mathbb{R}$, but $\mathcal{X}$ could be anything
- we have a kernel k on $\mathcal{X}$
- same idea than with ridge regression:

$$\hat{f} \in \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \left\| f \right\|^2 \right\} .$$

- here effect of the regularization is to make $\hat{f}$ smoother

[10] Cristianini and Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, 2000

Solving KRR

- representer theorem $\Rightarrow$

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)\,,$$

  for some $\alpha \in \mathbb{R}^n$

- as per the previous remark, we know that

$$(\hat{f}(x_1), \ldots, \hat{f}(x_n))^\top = K\alpha\,,$$

  and

$$\|\hat{f}\|^2 = \alpha^\top K \alpha\,.$$

- thus KRR can be re-written as

$$\hat{\alpha} \in \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n}(K\alpha - y)^\top (K\alpha - y) + \lambda \alpha^\top K \alpha \right\}\,.$$

Solving KRR, ctd.

- convex, smooth objective $\Rightarrow$ set the gradient to zero
- $\hat{\alpha}$ has to be solution of

$$0 = \frac{-2}{n} K(y - K\alpha) + 2\lambda K\alpha = \frac{2}{n} K \left[(K + n\lambda I_n)\alpha - y \right]$$

- since $\lambda > 0$, $K + n\lambda I_n$ is invertible
- a solution is given by

$$\hat{\alpha} = (K + n\lambda I_n)^{-1} y \,.$$

- **Remark:** not unique if K is singular
- why? $K + \lambda n \, \mathrm{I}$ and $(K + \lambda n \, \mathrm{I})^{-1}$ both leave $\ker K$ stable, can add ε such that $K\varepsilon = 0$
- but same element in the RKHS...

5.7. Kernel logistic regression

Kernel Logistic Regression[11] (KLR)

- classification setting: $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$
- $\mathcal{Y} = \{0, 1\}$, but $\mathcal{X}$ could be anything
- we have a kernel k on $\mathcal{X}$
- kernelized version of logistic regression:

$$\hat{f} \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i f(x_i)}\right) + \lambda \left\| f \right\|^2 \right\} .$$

- same regularization effect

[11] Green, Yandell, *Semi-parametric generalized linear models*, Generalized linear models, 1985

Solving KLR

- no explicit solution, but convex and smooth
- again, we can use the representer theorem:

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

  for some $\alpha \in \mathbb{R}^n$
- again, $(\hat{f}(x_1), \ldots, \hat{f}(x_n))^\top = K\alpha$ and $\|\hat{f}\|^2 = \alpha^\top K \alpha$
- we can rewrite KLR as

$$\hat{\alpha} \in \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \left\{ \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{-y_i (K\alpha)_i}\right) + \lambda \alpha^\top K \alpha \right\} .$$

- this can be solved (approximately) by gradient descent

Illustration