Learning theory from first principles

Lecture 3: Empirical risk minimization

Francis Bach

October 2, 2020

Class summary

-Convexification of the risk
-Risk decomposition
-Estimation error: finite number of hypotheses and covering numbers
-Rademacher complexity
-Penalized problems

Given a joint distribution $dp(x, y)$, and n independent and identically distributed observations from $dp(x, y)$, our goal is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ with minimum risk $\mathcal{R}(f) = \mathbb{E}[\ell(y, f(x))]$, or equivalently minimum excess risk:

$$\mathcal{R}(f) - \mathcal{R}^* = \mathcal{R}(f) - \inf_{g \text{ measurable}} \mathcal{R}(g).$$

1 Convexification of the risk

In this section, for simplicity, we focus on binary classification where $\mathcal{Y} = \{-1, 1\}$ with the 0-1 loss, but many of the concepts extend to the more general structured prediction set-up (see [1, 2] and the many references therein).

As our goal is to estimate a binary-valued function, the first idea that comes into mind is to minimize the empirical risk over a hypothesis space of binary-valued functions (or equivalently, space of subsets of $\mathcal{X}$). However, this approach leads to a combinatorial problem which can be computationally intractable and moreover, it is not clear how to control the capacity (i.e., how to regularize) for these type of hypothesis spaces. Learning a real-valued function instead through the framework of convex surrogates simplifies and overcomes this problem as it convexifies the problem and classical penalty-based regularization techniques can be used.

Instead of learning $f : \mathcal{X} \to \{-1, 1\}$, we will thus learn a function $g : \mathcal{X} \to \mathbb{R}$ and define $f(x) = \mathrm{sign}(g(x))$ where

$$\mathrm{sign}(a) = \begin{cases} 1 & \text{si } a \geqslant 0 \\ -1 & \text{si } a < 0. \end{cases}$$

Note here, that the value at 0 could also be chosen to be -1. Within our context, this corresponds for maximally ambiguous observations to choose one of the two labels which are equally likely (and thus equally bad in expectation, with a 50% chance of being incorrect).

The risk of the function $f = \mathrm{sign} \circ g$, still denoted $\mathcal{R}(g)$ (⚠ slight overloading $\mathcal{R}(g) = \mathcal{R}(\mathrm{sign} \circ g)$), is then equal to:

$$\mathcal{R}(g) = \mathbb{P}(\mathrm{sign}(g(x)) \neq y) = \mathbb{E}(1_{\mathrm{sign}(g(x)) \neq y}) = \mathbb{E}(1_{yg(x)<0}) = \mathbb{E}\Phi_{0-1}(yg(x)),$$

where $\Phi_{0-1} : \mathbb{R} \to \mathbb{R}$, with $\Phi_{0-1}(u) = 1_{u<0}$ is called the 0-1 loss function.

⚠ Note the slightly overloaded denomination above where the 0-1 loss function is defined on $\mathbb{R}$, with the 0-1 loss function defined in Lecture 1 on $\{-1, 1\} \times \{-1, 1\}$.

In practice, for empirical risk minimization, we then minimize with respect to $g : \mathcal{X} \to \mathbb{R}$ the corresponding empirical risk $\frac{1}{n} \sum_{i=1}^{n} \Phi_{0-1}(y_i g(x_i))$. The function Φ_{0-1} is not continuous (and thus also non-convex) and leads to difficult optimization problems.

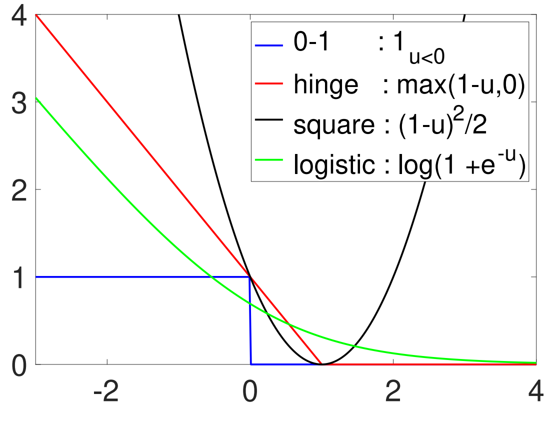
A key concept in machine learning is the use of *convex surrogates*, where we replace Φ_{0-1} by another function Φ with better numerical properties (all will be convex). See examples below.

Instead of minimizing the classical risk $\mathcal{R}(g)$ or its empirical version, one then minimizes the Φ-risk (and its empirical version) defined as

$$\mathcal{R}_\Phi(g) = \mathbb{E}[\Phi(yg(x))].$$

In this context, the function g is sometimes called the *score function*.

The key question is: does it make sense to simply convexify the problem? In other words, does it lead to good predictions for the 0-1 loss?



Examples of convex surrogates

- Quadratic loss: $\Phi(u) = (u-1)^2$, leading to, since $y^2 = 1$: $\Phi(yg(x)) = (y - g(x))^2 = (g(x) - y)^2$. We get back least-squares, and we simply ignore the fact that the labels have to belong to $\{-1, 1\}$.

- Logistic loss: $\Phi(u) = \log(1 + e^{-u})$, leading to $\Phi(yg(x)) = \log(1 + e^{-yg(x)}) = -\log(\frac{1}{1+e^{-yg(x)}}) = -\log(\sigma(yg(x))$, where: $\sigma(v) = \frac{1}{1+e^{-v}}$ is the sigmoid function.

  Note the link with maximum likelihood estimation, where we define the model through

  $$\mathbb{P}(y = 1|x) = \sigma(f(x)) \text{ and } \mathbb{P}(y = -1|x) = \sigma(-f(x)) = 1 - \sigma(f(x)).$$

  The risk is then the negative conditional log-likelihood $\mathbb{E}[-\log p(y|x)]$. See `https://en.wikipedia.org/wiki/Logistic_regression` for details. It is also often called the cross-entropy loss.

- Hinge loss: $\Phi(u) = \max(1 - u, 0)$. With linear predictors, this leads to the support vector machine, and $yf(x)$ is often called the "margin" in this context. See `https://en.wikipedia.org/wiki/Support_vector_machine` for details.

- Squared hinge loss: $\Phi(u) = \max(1 - u, 0)^2$. This is a smooth counterpart to the regular hinge loss.

Conditional Φ-risk and classification calibration

Most of the convex surrogates are upper-bounds on the 0-1 loss and all can be made so with rescaling. Using this as the sole justification of the good performance of a convex surrogate is a misleading justification, with the exception of problems with uniform zero loss (which is only possible when the Bayes risk is zero).

If we denote $\eta(x) = \mathbb{P}(y = 1|x) \in [0, 1]$, then we have, $\mathbb{E}[y|x] = 2\eta(x) - 1$, and, as seen in Lecture 1:

$$\mathcal{R}(g) = \mathbb{E}[\Phi_{0-1}(yg(x))] = \mathbb{E}[\mathbb{E}(1_{(g(y))\neq y})|x)] \geqslant \mathbb{E}[\min(\eta(x), 1 - \eta(x))] = \mathcal{R}^*,$$

and one best classifier is $f^*(x) = \text{sign}(2\eta(x) - 1)$. Note that there are **many** potential other functions $g(x)$ than $2\eta(x) - 1$ so that $f^*(x) = \text{sign}(g(x))$ is optimal. The first (minor) reason is the arbitrary choice of prediction for $\eta(x) = 1/2$. The other reason is that $f(x)$ simply has to have the same sign as $2\eta(x) - 1$, which leads to many possibilities beyond $2\eta(x) - 1$.

In order to study the impact of using the Φ-risk, we first look at the conditional risk, for a given x (as for the 0-1 loss, the function that g that will minimize the Φ-risk can be determined by looking at each x separately).

Definition 1 *Let $g : \mathcal{X} \to \mathbb{R}$, we define the conditional Φ-risk as*

$$\mathbb{E}[\Phi(yg(x))|x] = \eta(x)\Phi(g(x)) + (1 - \eta(x))\Phi(-g(x)) \textit{ which we denote } C_{\eta(x)}(g(x)),$$

with

$$C_\eta(\alpha) = \eta\Phi(\alpha) + (1 - \eta)\Phi(-\alpha).$$

The least we can expect from a convex surrogate is that in the population case, where all x's decouple, the optimal $g(x)$ obtained by minimizing the conditional Φ-risk exactly leads to the same prediction as the

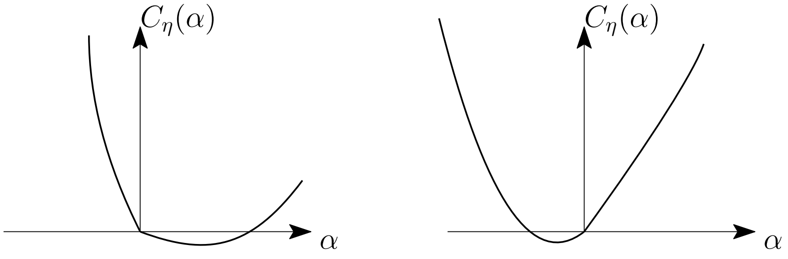
Bayes predictor (at least when this prediction is unique). In other words, since the prediction is $\text{sign}(g(x))$, we want that for any $\eta \in [0,1]$:

$$\text{(positive optimal prediction)} \qquad \eta > 1/2 \quad \Leftrightarrow \quad \arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subset \mathbb{R}_+^* \tag{1}$$
$$\text{(negative optimal prediction)} \qquad \eta < 1/2 \quad \Leftrightarrow \quad \arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subset \mathbb{R}_-^*. \tag{2}$$

A function Φ that satisfies these two statements is said *classification-calibrated*, or simply *calibrated.* It turns out that when Φ is convex, a simple sufficient and necessary condition is available:

Proposition 1 *[3] let $\Phi : \mathbb{R} \to \mathbb{R}$ convex. Φ calibrated $\Leftrightarrow$ Φ is differentiable at 0 and $\Phi'(0) < 0$.*

Proof Since Φ is convex, so is C_η for any $\eta \in [0,1]$, and thus we simply consider left and right derivatives at zero to obtain conditions about location of minimizers, with the two possibilities below.



$$\arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subset \mathbb{R}_+^* \quad \Leftrightarrow \quad (C_\eta)_+(0)' = \eta\Phi'_+(0) - (1-\eta)\Phi'_-(0) < 0$$
$$\arg\min_{\alpha\in\mathbb{R}} C_\eta(\alpha) \subset \mathbb{R}_+^* \quad \Leftrightarrow \quad (C_\eta)_-(0)' = \eta\Phi'_-(0) - (1-\eta)\Phi'_+(0) > 0.$$

- Assume Φ is calibrated. By letting η tend to $1/2$ in Eq. (1), this leads to $(C_{1/2})_+(0)' = \frac{1}{2}\big[\Phi'_+(0) - \Phi'_-(0)\big] \leqslant 0$. Since Φ is convex, we always have $\Phi'_+(0) - \Phi'_-(0) \geqslant 0$. Thus the left and right derivatives are equal, which implies that Φ is differentiable at 0. Then $C'_\eta(0) = (2\eta - 1)\Phi'(0)$, and from Eq. (1) and Eq. (2), we need to have $\Phi'(0) < 0$.
- Assume Φ is differentiable at 0 and $\Phi'(0) < 0$, then $C'_\eta(0) = (2\eta - 1)\Phi'(0)$; Eq. (1) and Eq. (2) are then direct consequences.

■

The proposition above excludes the convex surrogate $u \mapsto (-u)+ = \max\{-u, 0\}$, which is not differentiable at zero.

We now assume that Φ is calibrated and convex, that is, Φ is convex, Φ differentiable in 0, and $\Phi'(0) < 0$.

Relationship between risk and Φ-risk (♦)

Now that we know that for any $x \in \mathcal{X}$, minimizing $C_{\eta(x)}(g(x))$ with respect to $g(x)$ leads to the optimal prediction through $\text{sign}(g(x))$, we would like to make sure that an explicit control of the excess Φ-risk (which we aim to do with empirical risk minimization using tools from later sections) leads to an explicit control of the excess risk. In other words, we are looking for a monotonic function $H : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\mathcal{R}(g) - \mathcal{R}^* \leq H\big[\mathcal{R}_\Phi(g) - \mathcal{R}_\Phi^*\big]$, where $\mathcal{R}_\Phi^*$ is the minimum possible Φ-risk. The function H is often called the *calibration function*.

As opposed to the least-squares regression case, where the loss function used for testing is directly the one used within empirical risk minimization, there are two notions here: the testing error $\mathcal{R}(g)$, which is obtained after thresholding at zero the function g, and the quantity $\mathcal{R}_\Phi(g)$, which is sometimes called the testing loss.

We first start with a simple lemma expressing the excess risk, as well as an upper bound (adapted from Theorem 2.2 from [4]), that we will need for comparison inequalities below:

Lemma 1 *For any function $g : \mathcal{X} \to \mathbb{R}$, and for a Bayes predictor g^*:*

$$\mathcal{R}(g) - \mathcal{R}(g^*) = \mathbb{E}[1_{g(x)g^*(x)<0} \cdot |2\eta(x) - 1|].$$

Moreover, we have $\mathcal{R}(g) - \mathcal{R}(g^*) \leqslant \mathbb{E}[|2\eta(x) - 1 - g(x)|]$.

Proof We express the excess risk as:

$$\mathcal{R}(g) - \mathcal{R}(g^*) \quad = \mathbb{E}[\mathbb{E}[1_{\mathrm{sign}(g(x)) \neq y} - 1_{\mathrm{sign}(g^*)(x) \neq y}|x]] \text{ by definition of the 0-1 loss.}$$

For any given $x \in \mathcal{X}$, we can look at the two possible cases for the signs of $\eta(x) - 1/2$ and $g(x)$ that lead to different predictions for g and g^*, namely (a) $\eta(x) > 1/2$ and $g(x) < 0$, and (b) $\eta(x) < 1/2$ and $g(x) > 0$ (equality cases are irrelevant). For the first case the expectation with respect to x is $\eta(x) - (1 - \eta(x)) = 2\eta(x) - 1$, while for the second case, we get $1 - 2\eta(x)$. By combining these two cases into the condition $g(x)g^*(x) < 0$ and the condional expectation $|2\eta(x) - 1|$, we get the first result.

For the second result, we simply use the fact that if $g(x)g^*(x) < 0$, then, by splitting the cases in two (the first one being $\eta(x) > 1/2$ and $g(x) < 0$, the second one being $\eta(x) < 1/2$ and $g(x) > 0$), we get $|2\eta(x) - 1| \leqslant |2\eta(x) - 1 - g(x)|$, and thus the second result.

Note that for any function $b : \mathbb{R} \to \mathbb{R}$ that preserves the sign (that is $b(\mathbb{R}_+^*) \subset \mathbb{R}_+^*$ and $b(\mathbb{R}_-^*) \subset \mathbb{R}_-^*$), we have $\mathcal{R}(g) - \mathcal{R}(g^*) \leqslant \mathbb{E}[|2\eta(x) - 1 - b(g(x))|]$. ■

We see that the excess risk is the expectation of a quantity $|2\eta(x) - 1)| \cdot 1_{g(x)g^*(x)<0}$, which is equal to 0 if the classification is the same as the Bayes predictor and equal to $|2\eta(x) - 1|$ otherwise. The excess conditional Φ-risk is the quantity

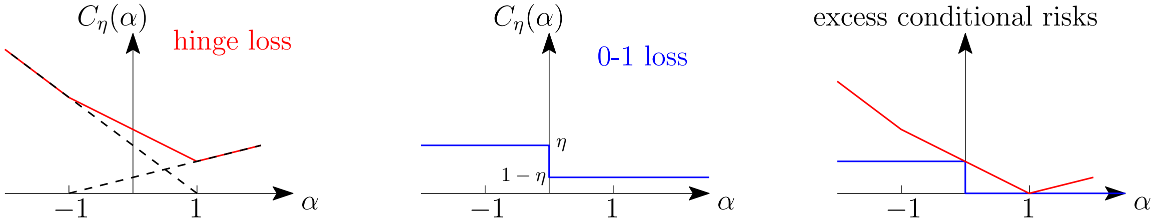
$$\eta(x)\Phi(g(x)) + (1 - \eta(x))\Phi(-g(x)) - \inf_\alpha \big\{\eta(x)\Phi(\alpha) + (1 - \eta(x))\Phi(-\alpha)\big\},$$

which, as a function of $g(x)$, is the deviation between a convex function and its minimum value. We simply need to relate it to the quantity $|2\eta(x) - 1)| \cdot 1_{g(x)g^*(x)<0}$ above for any $x \in \mathcal{X}$ and take expectations.

In [3], a general framework is proposed. We will only consider the hinge loss and smooth losses for simplicity.

- For the hinge loss $\Phi(\alpha) = (1 - \alpha)_+ = \max\{1 - \alpha, 0\}$, we can easily compute the minimizer of the conditional Φ-risk (which leads to the minimizer of the Φ-risk). Indeed, we need to minimize

$\eta(x)(1-\alpha)_+ + (1-\eta(x))(1+\alpha)_+$, which is a piecewise affine function with kinks at -1 and 1, with a minimizer attained at $u=1$ for $\eta(x)>1/2$ (see below), and symmetrically at $u=-1$ for $\eta(x)>1/2$, with a minimum conditional Φ-risk equal to $2\min\{1-\eta(x),\eta(x)\}$. The two excess risks are plotted below for the hinge loss and the 0-1 loss, for $\eta(x)>1/2$, showing pictorially that the conditionl excess Φ-risk is greater than than the excess risk.



This leads to the calibration function $H(\sigma)=\sigma$ for the hinge loss.

Note that when the Bayes risk is zero, that is, $\eta(x)\in\{0,1\}$ almost surely, then using the fact that the hinge loss is an upper-bound on the $0-1$ loss is enough to show that the excess risk is less than the excess Φ-risk (indeed, the two optimal risks $\mathcal{R}^*$ and $\mathcal{R}^*_\Phi$ are equal to zero).

- We consider smooth losses of the form (up to additive and multiplicative constants) $\Phi(v)=a(v)-v$, where $a(v)=\frac{1}{2}v^2$ for the quadratic loss, $a(v)=2\log(e^{v/2}+e^{-v/2})$ for the logistic loss. We assume that a is even, $a(0)=0$, a is β-smooth (that is, $a''(v)\leqslant\beta$ for all v). This implies[1] that for all $v\in\mathbb{R}$, $a(v)-\alpha v-\inf_{w\in\mathbb{R}}\big\{a(w)-\alpha w\big\}\geqslant\frac{1}{2\beta}|\alpha-a'(v)|^2$, leading to:

$$\begin{aligned}
\mathcal{R}_\Phi(g)-\mathcal{R}^*_\Phi &= \mathbb{E}\big[a(g(x))-(2\eta(x)-1)g(x)-\inf_{w\in\mathbb{R}}\big\{a(w)-(2\eta(x)-1)w\big\}\big]\\
&\geqslant \frac{1}{2\beta}\mathbb{E}\big[\big|2\eta(x)-1-a'(g(x))\big|^2\big]\\
&\geqslant \frac{1}{2\beta}\big(\mathbb{E}\big[\big|2\eta(x)-1-a'(g(x))\big|\big]\big)^2 \text{ by Jensen's inequality,}\\
&= \frac{1}{2\beta}\big(\mathcal{R}(g)-\mathcal{R}^*\big)^2 \text{ using Lemma 1.}
\end{aligned}$$

  This leads to the calibration function $H(\sigma)=\sqrt{\sigma}$ for the square loss and $H(\sigma)=\sqrt{2\sigma}$ for the logistic loss.

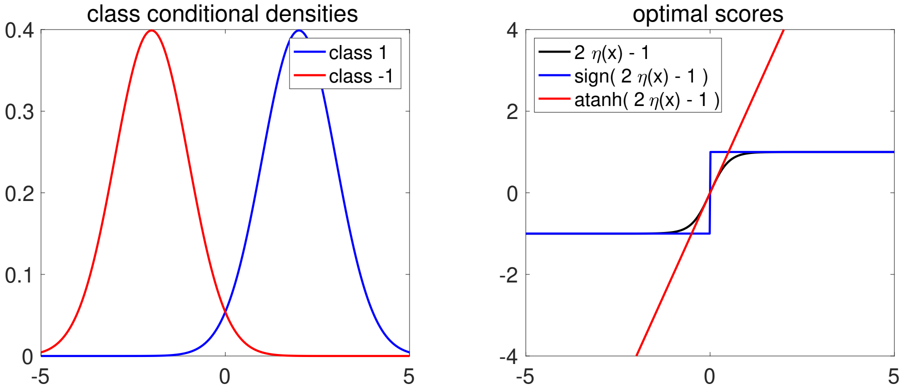
  Exercise: show that the function a^* satisfies $a^*\big(\mathcal{R}(g)-\mathcal{R}^*\big)\leqslant\mathcal{R}_\Phi(g)-\mathcal{R}^*_\Phi$ for any $g:\mathcal{X}\to\mathbb{R}$.

- For the non-smooth hinge loss, the calibration function is identity, so if the excess Φ-risk goes to zero at a certain rate, the excess risk goes to zero at the same rate, as for the smooth loss, the upper-bound only ensures a (worse) rate with a square root. Therefore, when going from the excess Φ-risk to the excess risk, that is, after thresholding the function g at zero, the observed rates may be worse.

- Note that the noiseless case where $\eta(x)\in\{0,1\}$ (zero Bayes risk) leads to stronger calibration function, as well as a series of intermediate "low-noise" conditions (see [3] for details).

[1] Using the Fenchel conjugate $a^*:\mathbb{R}\to\mathbb{R}$ which is $1/(2\beta)$-strongly convex (see Lecture 4), we have: $a(v)-\alpha v-\inf_{w\in\mathbb{R}}\big\{a(w)-\alpha w\big\}=a(v)-\alpha v+a^*(\alpha)=a^*(\alpha)-a^*(a'(v))-(\alpha-a'(v))(a^*)'(a(v))\geqslant\frac{1}{2\beta}|\alpha-a'(v)|^2$, where a^* is the Fenchel conjugate of a [5].

- Impact on approximation errors: for the same classification problem, several convex surrogates can be used. While the Bayes classifier is always the same, that is, $f^*(x) = \mathrm{sign}(2\eta(x)-1)$, the minimizer of the testing Φ-risk will be different. For example, for the hinge loss, the minimizer g is exactly $\mathrm{sign}(2\eta(x)-1)$, while for losses of the form like above $\Phi(v) = a(v) - v$, we have $a'(g(x)) = 2\eta(x) - 1$, and thus for the square loss $g(x) = 2\eta(x) - 1$, while for the logistic loss, one can check that $g(x) = \mathrm{atanh}(2\eta(x) - 1)$ (hyperbolic arc tangent). See example below, with $\mathcal{X} = \mathbb{R}$ and Gaussian class conditional densities.



The choice of surrogates will have an impact since to attained the minimal Φ-risk, different assumptions are needed on the class of functions used for empirical risk minimization, that is, $\mathrm{sign}(2\eta(x)-1)$ has to be in the class of functions we use (for the hinge loss), or $2\eta(x) - 1$ for the square loss, or $\mathrm{atanh}(2\eta(x) - 1)$ for the logistic loss.

- **Exercise**: for the logistic loss, show that for data generated as $x|y = 1$ and $x|y = -1$ Gaussians with the same covariance matrix, then the function $g(x)$ minimizing the expected logistic loss is affine in x (this model is often referred to as linear discriminant analysis).

2 Risk minimization decomposition

We consider a family $\mathcal{F}$ of prediction functions $f : \mathcal{X} \to \mathcal{Y}$. Empirical risk minimization aims at finding

$$\hat{f} \arg\min_{f\in\mathcal{F}} \ \widehat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)).$$

We can decompose the risk as follows into two terms:

$$\begin{aligned}\mathcal{R}(\hat{f}) - \mathcal{R}^* &= \Big\{\mathcal{R}(\hat{f}) - \inf_{f'\in\mathcal{F}} \mathcal{R}(f')\Big\} + \Big\{\inf_{f'\in\mathcal{F}} \mathcal{R}(f') - \mathcal{R}^*\Big\} \\ &= \text{estimation error} \qquad + \text{ approximation error}\end{aligned}$$

A classical example is the situation where the family of functions is parameterized by a subset of $\mathbb{R}^d$, that is, $\mathcal{F} = \{f_\theta,\ \theta \in \Theta\}$, for $\Theta \subset \mathbb{R}^d$. This includes neural networks (Lecture 8) and the simplest case of linear models of the form $f_\theta(x) = \theta^\top \varphi(x)$, for a certain feature vector $\varphi(x)$ (such as in Lecture 2). We will use linear models with Lipschitz-continuous loss functions as a motivating example, most often with constraints or penalties on the ℓ_2-norm $\|\theta\|_2$.

We now turn separately to the approximation and estimation errors.

3 Approximation error

- This means bounding $\inf_{f\in\mathcal{F}} \mathcal{R}(f) - \mathcal{R}^*$ and requires assumptions on the target function f^* (and hence on the testing distribution) to achieve non-trivial learning rates.
- In this section, we will focus on $\mathcal{F} = \{f_\theta,\ \theta \in \Theta\}$, for $\Theta \subset \mathbb{R}^d$ (we will consider infinite-dimensions in Lecture 6) and convex Lipschitz-continuous losses, assuming that θ_* is the minimizer of $\mathcal{R}(f_\theta)$ over $\theta \in \mathbb{R}^d$ (typically, it does not belong to Θ).
- This implies that the approximation error decomposes into

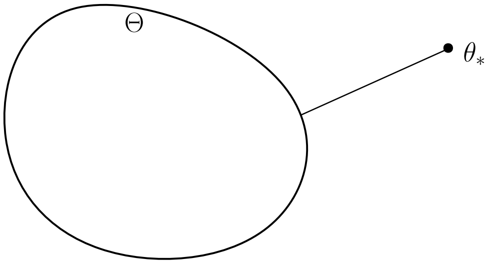
$$\inf_{\theta\in\Theta} \mathcal{R}(f_\theta) - \mathcal{R}^* = \Big(\inf_{\theta\in\Theta} \mathcal{R}(f_\theta) - \inf_{\theta\in\mathbb{R}^d} \mathcal{R}(f_\theta)\Big) + \Big(\inf_{\theta\in\mathbb{R}^d} \mathcal{R}(f_\theta) - \mathcal{R}^*\Big).$$

- The term $\mathcal{R}(f_\theta) - \mathcal{R}^*$ is the incompressible approximation error coming from the chosen model.
- The function $\theta \mapsto \mathcal{R}(f_\theta) - \inf_{\theta\in\mathbb{R}^d} \mathcal{R}(f_\theta)$ is a positive function on $\mathbb{R}^d$, which can be typically upperbounded by a certain norm (or its square) $\Omega(\theta - \theta_*)$, and we can see the term $\inf_{\theta\in\Theta} \mathcal{R}(f_\theta) - \inf_{\theta\in\mathbb{R}^d} \mathcal{R}(f_\theta)$ as a "distance" between θ_* and Θ.

  For example, if the loss which is considered is G-Lipschitz-continuous with respect to the second variable, with have,

$$\mathcal{R}(f_\theta) - \mathcal{R}(f_{\theta'}) \ = \ \mathbb{E}\big[\ell(y, f_\theta(x)) - \ell(y, f_{\theta'}(x))\big] \leqslant G\mathbb{E}\big[|f_\theta(x) - f_{\theta'}(x)|\big],$$

  and thus the approximation error is upper bounded by G times the distance between f_{θ_*} and $\mathcal{F} = \{f_\theta,\ \theta \in \Theta\}$, for a particular distance $d(\theta, \theta') = \mathbb{E}\big[|f_\theta(x) - f_{\theta'}(x)|\big]$.



A classical example will be $f_\theta(x) = \theta^\top \varphi(x)$, and $\Theta = \{\theta \in \mathbb{R}^d,\ \|\theta\|_2 \leqslant D\}$, leading to the upper bound $G\mathbb{E}\big[\|\varphi(x)\|_2\big](\|\theta_*\|_2 - D)_+$, which is equal to zero if $\|\theta_*\|_2 \leqslant D$ (well-specified model).

- **Exercise**: perform the same computation for the ℓ_1-norm on Θ.

4 Estimation error

We will consider general techniques, and all apply them to linear models with bounded ℓ_2-norm by D, and G-Lipschitz-losses for illustration.

- The estimation error is often decomposed as, using $g \in \arg\min_{g \in \mathcal{F}} \mathcal{R}(g)$ and $\hat{f} \in \arg\min_{f \in \mathcal{F}} \widehat{\mathcal{R}}(f)$:

$$\begin{aligned} \mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{F}} \mathcal{R}(f) = \mathcal{R}(\hat{f}) - \mathcal{R}(g) &= \Big\{\mathcal{R}(\hat{f}) - \widehat{\mathcal{R}}(\hat{f})\Big\} + \Big\{\widehat{\mathcal{R}}(\hat{f}) - \widehat{\mathcal{R}}(g)\Big\} + \Big\{\widehat{\mathcal{R}}(g) - \mathcal{R}(g)\Big\} \\ &\leqslant \sup_{f \in \mathcal{F}} \Big\{\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\Big\} + \Big\{\widehat{\mathcal{R}}(\hat{f}) - \widehat{\mathcal{R}}(g)\Big\} + \sup_{f \in \mathcal{F}} \Big\{\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\Big\} \\ &\leqslant \sup_{f \in \mathcal{F}} \Big\{\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\Big\} + 0 + \sup_{f \in \mathcal{F}} \Big\{\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\Big\} \text{ by definition of } \hat{f}. \end{aligned}$$

  This is often upper-bounded by $2 \sup_{f \in \mathcal{F}} \Big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\Big|$.

  When $\hat{f}$ is not the global minimizer but simply satisfies $\widehat{\mathcal{R}}(\hat{f}) \leqslant \inf_{f \in \mathcal{F}} \widehat{\mathcal{R}}(f) + \varepsilon$, then the *optimization error* ε has to be aded to the bound above (see more details in Lecture 3).

- The uniform deviation grows with the "size" of $\mathcal{F}$, and usually decays with n. See examples below.

  The key issue that we need a uniform control of for all $f \in \mathcal{F}$: with a single f, we could apply any concentration inequality to the random variable $\ell(y, f(x))$ to obtain a bound in $O(1/\sqrt{n})$; however, when controlling the maximal deviations over many values of f, there is always a small chance that one of these deviations get large. We thus need an explicit control of this phenomenon, which we now tackle, by first showing that we can focus on the expectation alone.

4.1 Application of Mac Diarmid inequality

Let $H(z_1, \dots, z_n) = \sup_{f \in \mathcal{F}} \Big\{\mathcal{R}(f) - \widehat{\mathcal{R}}(f)\Big\}$, where the random variables $z_i = (x_i, y_i)$ are independent and identically distributed, and $\widehat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$. We let ℓ_∞ denote the maximal absolute value of the loss functions for all (x, y) in the support of the data generating distribution and $f \in \mathcal{F}$.

When changing a single $z_i \in \mathcal{X} \times \mathcal{Y}$ into $z_i' \in \mathcal{X} \times \mathcal{Y}$, the deviation in H is almost surely at most $\frac{2}{n}\ell_\infty$. Thus, applying Mac Diarmid inequality (see Lecture 1), with probability greater than $1-\delta$, we have:

$$H(z_1, \dots, z_n) - \mathbb{E}[H(z_1, \dots, z_n)] \leqslant \frac{\ell_\infty \sqrt{2}}{\sqrt{n}} \sqrt{\log \frac{1}{\delta}}.$$

We thus only need to bound the expectation of $\sup_{f \in \mathcal{F}} \left\{ \mathcal{R}(f) - \widehat{\mathcal{R}}(f) \right\}$ and of $\sup_{f \in \mathcal{F}} \left\{ \widehat{\mathcal{R}}(f) - \mathcal{R}(f) \right\}$ (which will typically have the same bound).

4.2 Easy case I: quadratic functions

We will show what happens with a quadratic loss function and an ℓ_2-ball constraint. We remember that in this case $\ell(y, \theta^\top \varphi(x)) = (y - \theta^\top \varphi(x))^2$. From that we get

$$\begin{aligned} \widehat{\mathcal{R}}(f) - \mathcal{R}(f) &= \theta^\top \left(\frac{1}{n} \sum_{i=1}^n \varphi(x_i)\varphi(x_i)^\top - \mathbb{E}\left[\varphi(x)\varphi(x)^\top \right] \right) \theta \\ &\quad -2\theta^\top \left(\frac{1}{n} \sum_{i=1}^n y_i \varphi(x_i) - \mathbb{E}\left[y\varphi(x) \right] \right) + \left(\frac{1}{n} \sum_{i=1}^n y_i^2 - \mathbb{E}\left[y^2 \right] \right). \end{aligned}$$

Hence, the supremum can be upper bounded in closed form as

$$\begin{aligned} \sup_{\|\theta\|_2 \leqslant D} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)| &\leqslant D^2 \left\| \frac{1}{n} \sum_{i=1}^n \varphi(x_i)\varphi(x_i)^\top - \mathbb{E}\left[\varphi(x)\varphi(x)^\top \right] \right\|_{\text{op}} \\ &\quad +2D \left\| \frac{1}{n} \sum_{i=1}^n y_i \varphi(x_i) - \mathbb{E}\left[y\varphi(x) \right] \right\|_2 + \left| \frac{1}{n} \sum_{i=1}^n y_i^2 - \mathbb{E}\left[y^2 \right] \right|, \end{aligned}$$

where $\|M\|_{\text{op}}$ is the operator norm of the matrix M defined as $\sup_{\|u\|_2 = 1} \|Mu\|_2$.

Thus, in order to get a uniform bound, we simply need to upper-bond the three *non-uniform* expectations of deviations, and thus of order $O(1/\sqrt{n})$, and we get an overall uniform deviation bound. This particular case gives the impression that it should be possible to get such a rate in $O(1/\sqrt{n})$ for other types of losses than the quadratic loss. However, closed form calculations are not possible, so we need to introduce new tools.

- **Exercise**: provide an explicit bound on $\sup_{\|\theta\|_2 \leqslant D} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)|$ above, and compare it to the use of Rademacher complexities below.

⚠ Note that in this section, we do not require the loss to be convex.

4.3 Easy case II: Finite number of models

We have, if the loss functions are bounded between $-\ell_\infty$ and ℓ_∞, using the upper-bound $2\sup_{f\in\mathcal{F}}\Big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\Big|$ on the estimation error, and the union bound:

$$\begin{aligned}\mathbb{P}\Big(\mathcal{R}(\hat f) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) \geqslant t\Big) &\leqslant \mathbb{P}\Big(2\sup_{f\in\mathcal{F}}\Big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\Big| \geqslant t\Big)\\ &\leqslant \sum_{f\in\mathcal{F}} \mathbb{P}\Big(2\Big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\Big| \geqslant t\Big).\end{aligned}$$

We have, for $f \in \mathcal{F}$ fixed, $\widehat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(y_i))$, and we can apply Hoeffding's inequality to bound each $\mathbb{P}\Big(2\Big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\Big| \geqslant t\Big)$, leading to

$$\mathbb{P}\Big(2|\mathcal{R}(\hat f) - \mathcal{R}(f)| \geqslant t\Big) \leqslant \sum_{f\in\mathcal{F}} 2\exp(-nt^2/2\ell_\infty^2) = 2|\mathcal{F}|\exp(-nt^2/2\ell_\infty^2).$$

Thus, by setting $\delta = 2|\mathcal{F}|\exp(-nt^2/2\ell_\infty^2)$, and finding the corresponding t, with probability greater than $1-\delta$, we get:

$$\mathcal{R}(\hat f) - \mathcal{R}(f) \leqslant \frac{2\ell_\infty}{\sqrt{n}}\sqrt{\log\frac{2|\mathcal{F}|}{\delta}} = \frac{2\ell_\infty}{\sqrt{n}}\sqrt{\log(|\mathcal{F}|) + \log\frac{2}{\delta}} \leqslant 2\ell_\infty\sqrt{\frac{\log(|\mathcal{F}|)}{n}} + \frac{2\ell_\infty}{\sqrt{n}}\sqrt{\log\frac{2}{\delta}}.$$

- **Exercise**: in terms of expectation, we get (using the proof of the max of random variables from Lecture 1):
$$\mathbb{E}\big[\mathcal{R}(\hat f) - \inf_{f\in\mathcal{F}}\mathcal{R}(f)\big] \leqslant 2\mathbb{E}\big[\sup_{f\in\mathcal{F}}\Big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\Big|\big] \leqslant \ell_\infty\sqrt{\frac{2\log(2|\mathcal{F}|)}{n}}.$$

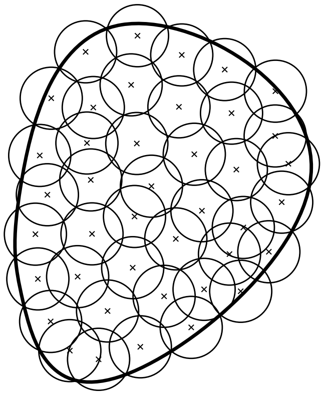
Thus, according to the bound, when the logarithm of the number of models is small compared to n, learning is possible. This is a first generic control of the uniform deviations.

⚠ Note that this is only an upper-bound and learning is possible with infinitely many models (which is the most classical scenario). See below.

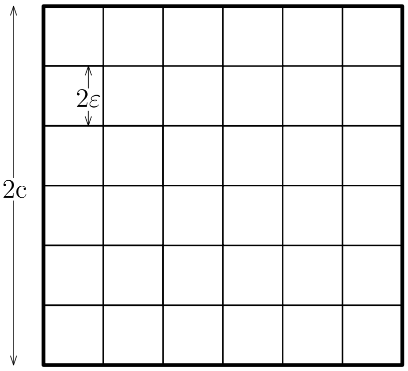
4.4 Beyond finite number models through covering numbers (♦)

The simple idea behind covering numbers is to deal with function spaces with infinitely many elements by approximating them through a finite number of elements. This is often referred to as an ε-net argument.

- We first need to assume that the risks $\mathcal{R}$ and $\widehat{\mathcal{R}}$ are regular, for example, that they are G-Lipschitz-continuous with respect to some distance d on $\mathcal{F}$.
- Covering numbers: We assume there exists $m = m(\varepsilon)$ elements $f_1, \dots, f_m$ such that for any $f \in \mathcal{F}$, $\exists i \in \{1, \dots, n\}$ such that $d(f, f_i) \leqslant \varepsilon$. The minimal possible number $m(\varepsilon)$ is the *covering number* of $\mathcal{F}$ at precision ε. See an example below in two dimensions of a covering with Euclidean balls.



- Property: the covering number is a non-increasing function of ε.
- Typically, $m(\varepsilon)$ grows with ε as a power ε^{-d} when $\varepsilon \to 0$, where d is the underlying dimension. Indeed, for the ℓ_∞-metric, if (in a certain parameterization) $\mathcal{F}$ is included in a ball of radius c in the ℓ_∞-ball of dimension d, it can be easily covered by $(c/\varepsilon)^d$ cubes of length 2ε. See below.



Given that all norms are equivalent in dimension d, we get the same dependence in d for all bounded subsets of a finite-dimensional vector space.

For some sets (e.g, all Lipschitz-continuous functions in d dimensions) $\log m(\varepsilon)$ grows fasterm such as ε^{-d}. See, e.g., [6].

- Then, given the cover, for all $f \in \mathcal{F}$, and f_i the associated cover elements,

$$\begin{aligned}\Big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\Big| &\leqslant \Big|\widehat{\mathcal{R}}(f) - \widehat{\mathcal{R}}(f_i)\Big| + \Big|\widehat{\mathcal{R}}(f_i) - \mathcal{R}(f_i)\Big| + \Big|\mathcal{R}(f_i) - \mathcal{R}(f)\Big| \\ &\leqslant 2G\varepsilon + \sup_{i\in\{1,\dots,m(\varepsilon)\}} \Big|\widehat{\mathcal{R}}(f_i) - \mathcal{R}(f_i)\Big|.\end{aligned}$$

- This implies that

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}} \Big|\widehat{\mathcal{R}}(f) - \mathcal{R}(f)\Big|\Big] \leqslant 2G\varepsilon + \mathbb{E}\Big[\sup_{i\in\{1,\dots,m(\varepsilon)\}} \Big|\widehat{\mathcal{R}}(f_i) - \mathcal{R}(f_i)\Big|\Big] \leqslant 2G\varepsilon + \frac{1}{2}\sqrt{\frac{2\log(2m(\varepsilon)))}{n}}.$$

Therefore, if $m(\varepsilon) \sim \varepsilon^{-d}$, we need to balance $\varepsilon + \sqrt{d \log(1/\varepsilon)/n}$, which leads to, with a choice of ε proportional to $1/\sqrt{n}$, $\sqrt{(d/n)\log(n/d)}$, a rate essentially proportional to $d/\sqrt{n}$.

However, this typically leads to a dependence on dimension because for the unit ball of some normed space, the covering number of the unit ball is grows as ε^{-d} (see examples above).

- One very powerful tool that avoids these undesired dependences on dimension is Rademacher complexities [7] or Gaussian complexities [8]. In this lecture, we will focus on Rademacher complexity.

4.5 Rademacher complexity

We consider n independent and identically distributed random variables $z_1, \dots, z_n \in \mathcal{Z}$, and a class $\mathcal{H}$ of functions from $\mathcal{Z}$ to $\mathbb{R}$. In our context, the space of functions is related to the learning problem as: $\mathcal{H} = \{(x, y) \mapsto \ell(y, f(x)),\ f \in \mathcal{F}\}$.

Our goal in this section is to provide an upper-bound on $\sup_{f \in \mathcal{F}} \mathcal{R}(f) - \widehat{\mathcal{R}}(f)$, which happens to be equal to

$$\sup_{h \in \mathcal{H}} \mathbb{E}[h(z)] - \frac{1}{n} \sum_{i=1}^{n} h(z_i),$$

where $\mathbb{E}[h(z)]$ denotes the expectation with respect to a variable having the same distribution as all z_i's.

We denote $\mathcal{D} = \{z_1, \dots, z_n\}$ the data. We define the *Rademacher complexity* of the class of functions $\mathcal{H}$ from $\mathcal{X}$ to $\mathbb{R}$:

$$R_n(\mathcal{H}) = \mathbb{E}_{\varepsilon, \mathcal{D}} \Big(\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(z_i) \Big),$$

where $\varepsilon \in \mathbb{R}^n$ is a vector of independent Rademacher random variable (that is taking values -1 or 1 with equal probabilities), which is also independent of $\mathcal{D}$. It is a deterministic quantity that only depends on n and $\mathcal{H}$.

In words, the Rademacher complexity is equal to the expectation of the maximal dot-product between values of a function h at the observations x_i and random labels. It is a measure of the "capacity" of the set of functions $\mathcal{H}$. We will see later that it can be computed in many interesting cases and leads to interesting bounds.

Symmetrization

First, we relate it to the uniform deviation through this symmetrization property.

Proposition 2

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} h(z_i) - \mathbb{E}[h(z)] \right)\right] \leqslant 2 R_n(\mathcal{H}) \textit{ and } \mathbb{E}\left[\sup_{h \in \mathcal{H}} \left(\mathbb{E}[h(z)] - \frac{1}{n} \sum_{i=1}^{n} h(z_i) \right)\right] \leqslant 2 R_n(\mathcal{H}).$$

Proof Let $\mathcal{D}' = \{z_1', \dots, z_n'\}$ an independent copy of the data $\mathcal{D} = \{z_1, \dots, z_n\}$. Let $(\varepsilon_i)_{i \in \{1, \dots, n\}}$ be i.i.d Rademacher random variables, which are also independent of $\mathcal{D}$ and $\mathcal{D}'$. Using that for all i, $\mathbb{E}[h(z_i') | \mathcal{D}] =$

$\mathbb{E}[h(z)]$, we have:

$$\begin{aligned}
\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}[h(z)] - \frac{1}{n}\sum_{i=1}^n h(z_i)\right)\right] &= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[h(z_i')|\mathcal{D}] - \frac{1}{n}\sum_{i=1}^n h(z_i)\right)\right] \\
&= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[h(z_i') - h(z_i)|\mathcal{D}\right]\right)\right] \\
&\qquad\qquad \text{by definition of the independent copy } \mathcal{D}', \\
&\leqslant \mathbb{E}\left[\mathbb{E}\Big(\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \left[h(z_i') - h(z_i)\right]\right)\Big|\mathcal{D}\Big)\right] \\
&\text{using that the supremum of the expectation is less than expectation of the supremum,} \\
&= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \left[h(z_i') - h(z_i)\right]\right)\right] \text{ by the towering law of expectation,} \\
&= \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(h(z_i') - h(z_i)\right)\right)\right] \text{ by symmetry of the law of } \varepsilon_i \in \{-1,1\}, \\
&\leqslant \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(h(z_i)\right)\right)\right] + \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(-h(z_i)\right)\right)\right] \\
&= 2\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i h(z_i)\right)\right] = 2R_n(\mathcal{H}).
\end{aligned}$$

The reasoning is essentially indentical for $\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n h(z_i) - \mathbb{E}[h(z)]\right)\right] \leqslant 2R_n(\mathcal{H})$. ■

We thus see that the Rademacher complexity directly controls the uniform deviation.

- **Exercise**: If $\mathcal{H}$ is finite, and so that, for all $h \in H$ and almost all z, $|h(z)| \leqslant \ell_\infty$, compute an upperbound on $R_n(\mathcal{H})$. Solution: $R_n(\mathcal{H}) \leqslant \ell_\infty\sqrt{2\log(2|\mathcal{H}|)}$. We recover the same result as in Section 4.3.

Lipschitz-continuous losses

4.6 Lipschitz-continuous losses

A particularly appealing property in our context is the following property, sometimes called the "contraction principle", using a simple proof from [9, Lemma 5].

Proposition 3 (Contraction principle - Lipschitz-continuous functions) *Given any functions b, $a_i : \Theta \to \mathbb{R}$ (no assumption) and $\varphi_i : \mathbb{R} \to \mathbb{R}$ any 1-Lipschitz-functions, for $i = 1, \ldots, n$, we have, for $\varepsilon \in \mathbb{R}^n$ a vector of independent Rademacher random variables:*

$$\mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\ b(\theta) + \sum_{i=1}^n \varepsilon_i\varphi_i(a_i(\theta))\right] \leqslant \mathbb{E}_\varepsilon\left[\sup_{\theta\in\Theta}\ b(\theta) + \sum_{i=1}^n \varepsilon_i a_i(\theta)\right].$$

Proof (♦) We consider a proof by induction on n. The case $n=0$ is trivial, and we show how to go from $n \geqslant 0$ to $n+1$. We thus consider $\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_{n+1}}\Big[\sup_{\theta\in\Theta} \ b(\theta)+\sum_{i=1}^{n+1}\varepsilon_i\varphi_i(a_i(\theta))\Big]$ and compute the expectation with respect to ε_{n+1} explicitly, by considering the two potential values with probability $1/2$:

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_{n+1}}\Big[\sup_{\theta\in\Theta} b(\theta)+\sum_{i=1}^{n+1}\varepsilon_i\varphi_i(a_i(\theta))\Big]\\
&= \frac{1}{2}\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\Big[\sup_{\theta\in\Theta} b(\theta)+\sum_{i=1}^{n}\varepsilon_i\varphi_i(a_i(\theta))+\varphi_{n+1}(a_{n+1}(\theta))\Big]+\frac{1}{2}\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\Big[\sup_{\theta\in\Theta} b(\theta)+\sum_{i=1}^{n}\varepsilon_i\varphi_i(a_i(\theta))-\varphi_{n+1}(a_{n+1}(\theta))\Big]\\
&= \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\Big[\sup_{\theta,\theta'\in\Theta}\frac{b(\theta)+b(\theta')}{2}+\sum_{i=1}^{n}\varepsilon_i\frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2}+\frac{\varphi_{n+1}(a_{n+1}(\theta))-\varphi_{n+1}(a_{n+1}(\theta'))}{2}\Big],
\end{aligned}$$

by assembling the term together. By taking the supremum over (θ,θ') and (θ',θ), we get

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\Big[\sup_{\theta,\theta'\in\Theta}\frac{b(\theta)+b(\theta')}{2}+\sum_{i=1}^{n}\varepsilon_i\frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2}+\frac{|\varphi_{n+1}(a_{n+1}(\theta))-\varphi_{n+1}(a_{n+1}(\theta'))|}{2}\Big]\\
&\leqslant \ \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\Big[\sup_{\theta,\theta'\in\Theta}\frac{b(\theta)+b(\theta')}{2}+\sum_{i=1}^{n}\varepsilon_i\frac{\varphi_i(a_i(\theta))+\varphi_i(a_i(\theta'))}{2}+\frac{|a_{n+1}(\theta)-a_{n+1}(\theta')|}{2}\Big],
\end{aligned}$$

Using Lipschitz-continuity. We can redo the exact same sequence of equalities, to obtain

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n}\mathbb{E}_{\varepsilon_{n+1}}\Big[\sup_{\theta\in\Theta} b(\theta)+\varepsilon_{n+1}a_{n+1}(\theta)+\sum_{i=1}^{n}\varepsilon_i\varphi_i(a_i(\theta))\Big]\\
&\leqslant \ \mathbb{E}_{\varepsilon_1,\dots,\varepsilon_n,\varepsilon_{n+1}}\Big[\sup_{\theta\in\Theta} b(\theta)+\varepsilon_{n+1}a_{n+1}(\theta)+\sum_{i=1}^{n}\varepsilon_i a_i(\theta)\Big] \text{ by recursion,}
\end{aligned}$$

which leads to the desired result. ■

We can apply the contraction principle above to supervised learning situations where $u_i \mapsto \ell(y_i,u_i)$ is G-Lipschitz-continuous for all i almost surely (which is possible for regression or when using a convex surrogate for binary classification as presented earlier), leading to:

$$\mathbb{E}_\varepsilon\Big(\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\ell(y_i,f(x_i)) \ \Big| \ \mathcal{D}\Big) \ \leqslant \ G\cdot\mathbb{E}_\varepsilon\Big(\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i) \ \Big| \ \mathcal{D}\Big) \text{ by the contraction principle,}$$

which leads to

$$\mathcal{R}_n(\mathcal{H}) \leqslant G\cdot\mathcal{R}_n(\mathcal{F}). \qquad (3)$$

Thus the Rademacher complexity of the class of prediction functions controls the uniform deviations of the empirical risk. We now consider simple examples.

Ball-constrained linear predictions

We now assume that $\mathcal{F} = \{f_\theta(x) = \theta^\top \varphi(x),\ \Omega(\theta) \leqslant D\}$ where Ω is a norm on $\mathbb{R}^d$. We denote by $\Phi \in \mathbb{R}^{n\times d}$ the design matrix. We have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{\Omega(\theta)\leqslant D}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \theta^\top \varphi(x_i)\right)\right] = \mathbb{E}\left[\sup_{\Omega(\theta)\leqslant D}\frac{1}{n}\varepsilon^\top \Phi\theta\right] \\
&= \frac{D}{n}\mathbb{E}\left[\Omega^*(\Phi^\top \varepsilon)\right],
\end{aligned}$$

where $\Omega^*(u) = \sup_{\Omega(\theta)\leqslant 1} u^\top\theta$ is the *dual norm* of Ω. For example, when Ω is the ℓ_p-norm, with $p \in [1,\infty]$, then Ω^* is the ℓ_q-norm, where q is such that $\frac{1}{p}+\frac{1}{q} = 1$, e.g., $\|\cdot\|_2^* = \|\cdots\|_2$, $\|\cdot\|_1^* = \|\cdot\|_\infty$, and $\|\cdot\|_\infty^* = \|\cdot\|_1$. For more details, see [5].

Thus, computing Rademacher complexities is equivalent to computing expectation of norms. When $\Omega = \|\cdot\|_2$, we get:

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \frac{D}{n}\mathbb{E}\left[\|\Phi^\top\varepsilon\|_2\right] \\
&\leqslant \frac{D}{n}\sqrt{\mathbb{E}\left[\|\Phi^\top\varepsilon\|_2^2\right]} \text{ by Jensen's inequality,} \\
&= \frac{D}{n}\sqrt{\mathbb{E}\left[\mathrm{tr}[\Phi^\top\varepsilon\varepsilon^\top\Phi]\right]} \\
&= \frac{D}{n}\sqrt{\mathbb{E}\left[\mathrm{tr}[\Phi^\top\Phi]\right]} \text{ using that } \mathbb{E}[\varepsilon\varepsilon^\top] = I, \\
&= \frac{D}{n}\sqrt{\sum_{i=1}^n \mathbb{E}(\Phi^\top\Phi)_i} = \frac{D}{n}\sqrt{\sum_{i=1}^n \mathbb{E}\|x_i\|_2^2} = \frac{D}{\sqrt{n}}\sqrt{\mathbb{E}\|x\|_2^2}. \qquad (4)
\end{aligned}$$

We thus obtain a dimension-independent Rademacher complexity that we can use in the summary below.

- **Exercise**: Upper-bound the Rademacher complexity for $\Omega = \|\cdot\|_1$.

4.7 Putting things together (linear predictions)

With all the elements above, we can now propose the following general result (where no convexity is assumed).

Proposition 4 (Estimation error) *Assume a G-Lipschitz-continuous loss function, linear prediction functions with* $\mathcal{F} = \{f_\theta(x) = \theta^\top\varphi(x),\ \|\theta\|_2 \leqslant D\}$, *where* $\mathbb{E}\|\varphi(x)\|_2^2 \leqslant R^2$. *Let* $\hat{f} = f_{\hat{\theta}} \in \mathcal{F}$ *be the minimizer of the empirical risk, then:*

$$\mathbb{E}\big[\mathcal{R}(f_{\hat{\theta}})\big] \leqslant \inf_{\|\theta\|_2\leqslant D} \mathcal{R}(f_\theta) + \frac{2GRD}{\sqrt{n}}.$$

Proof Using Prop. 2, Eq. (3) and Eq. (4), we get the desired result. ■

If we assume that there exists a minimizer θ_* of $\mathcal{R}(f_\theta)$ over $\mathbb{R}^d$, the the approximation error is upper-bounded by

$$\begin{aligned}\inf_{\|\theta\|_2\leqslant D}\mathcal{R}(f_\theta)-\mathcal{R}(f_{\theta_*}) &\leqslant G\inf_{\|\theta\|_2\leqslant D}\mathbb{E}\big[|f_\theta(x)-f_{\theta_*}(x)|\big]\\ &= G\inf_{\|\theta\|_2\leqslant D}\mathbb{E}\big[|\varphi(x)^\top(\theta-\theta_*)|\big]\\ &\leqslant G\inf_{\|\theta\|_2\leqslant D}\|\theta-\theta_*\|_2\mathbb{E}\big[\|\varphi(x)\|_2^2\big]\leqslant GR\inf_{\|\theta\|_2\leqslant D}\|\theta-\theta_*\|_2.\end{aligned}$$

This leads to

$$\mathbb{E}\big[\mathcal{R}(f_{\hat\theta})\big]\leqslant GR\inf_{\|\theta\|_2\leqslant D}\|\theta-\theta_*\|_2+\frac{2GRD}{\sqrt{n}}=GR(\|\theta_*\|_2-D)_++\frac{2GRD}{\sqrt{n}}.$$

We see that for $D=\|\theta_*\|_2$, we obtain the bound $\frac{2GRD}{\sqrt{n}}$, but this setting requires to know $\|\theta_*\|_2$ which is not possible in practice. If D is too large, the estimation error gets larger (overfitting), while if D is too small, the approximation error can quickly kick in (with a value that does not go to zero when n tends to infinity), leading to underfitting.

4.8 From constrained to regularized estimation (♦)

In practice, it is preferable to penalize by the norm $\Omega(\theta)=\|\theta\|_2$ instead of constraining (the main reasons being that the hyperparameter is easier to find and the optimization is easier). For simplicity, we consider only the ℓ_2-norm in this section.

We now denote $\hat\theta_\lambda$ the minimizer of

$$\hat{\mathcal{R}}(f_\theta)+\frac{\lambda}{2}\|\theta\|_2^2. \qquad (5)$$

If the loss is always positive, then

$$\frac{\lambda}{2}\|\hat\theta\|_2^2\leqslant\hat{\mathcal{R}}(f_{\hat\theta})+\frac{\lambda}{2}\|\hat\theta\|_2^2\leqslant\widehat{\mathcal{R}}(f_0),$$

leading to a bound $\|\hat\theta\|_2=O(1/\sqrt{\lambda})$. Thus, with $D=O(1/\sqrt{\lambda})$ in the bound above, this leads to a deviation of $O(1/\sqrt{\lambda}n)$, which is not optimal.

We now cite without proof an interesting result using the strong convexity of the squared ℓ_2-norm.

Proposition 5 (Fast rates for regularized objectives [10]) *Assume a G-Lipschitz-continuous* ***convex*** *loss function, linear prediction functions with $\mathcal{F}=\{f_\theta(x)=\theta^\top\varphi(x),\ \|\theta\|_2\leqslant D\}$, where $\mathbb{E}\|\varphi(x)\|_2^2\leqslant R^2$. Let $\hat\theta_\lambda\in\mathbb{R}^d$ be the minimizer of the regularized empirical risk in Eq. (5), then:*

$$\mathbb{E}\big[\mathcal{R}(f_{\hat\theta_\lambda})\big]\leqslant\inf_{\theta\in\mathbb{R}^d}\big\{\mathcal{R}(f_\theta)+\frac{\lambda}{2}\|\theta\|_2^2\big\}+\frac{32G^2R^2}{\lambda n}.$$

Note that we obtain a "fast rate" in $O(R^2/(\lambda n)$, which has a better dependence in n, but depends on λ, which can be very small in practice.

One classical choice of λ that we have seen in Lecture 2 also applies here, as $\lambda \propto \frac{GR}{\sqrt{n}\|\theta_*\|}$, leading to the slow rate

$$\mathbb{E}\big[\mathcal{R}(f_{\hat{\theta}_\lambda})\big] \leqslant \mathcal{R}(f_{\theta_*}) + O\Big(\frac{GR}{\sqrt{n}}\|\theta_*\|_2\Big).$$

This is the similar result as for Lecture 2, but not for all Lipschitz-continuous losses.

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 Alex Nowak-Vila for his help on calibration functions.

References

[1] Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. A general theory for structured prediction with smooth convex surrogates. Technical Report 1902.01958, arXiv, 2019.

[2] Alex Nowak-Vila, Francis Bach, and Alessandro Rudi. Consistent structured prediction with Max-Min Margin Markov Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[3] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[4] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 1996.

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

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

[7] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[8] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[9] Ron Meir and Tong Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

[10] Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems*, 2009.