Supervised Learning (COMP0078)

6. Learning Theory (Part I)

Carlo Ciliberto

University College London
Department of Computer Science

Previous Classes

In the previous classes we:

- Formulated the supervised learning problem.
- Focused on Empirical Risk Minmization (ERM).
- Discussed overfitting and how to tackle it (heuristically).

Previous Classes

We studied a few different algorithms (in particular focusing on classification):

- Nearest neighbor
- Least-squares
- Support Vector Machines
- Logistic Regression
- Decision Trees
- Ensemble methods (Bagging, Boosting)

We observed that many of these methods can be formulated as ERM problems.

Outline

Equipped with our shiny new tool-set of supervised learning algorithms, we will now go back to asking ourselves the question:

> *When/why/how do these methods "work"?*

Outline:

- Emprical Risk Minimization (Again!)
- Generalization error
- Excess Risk Decomposition
- Regularization & Bias-Variance Trade-off (Again!)

Refresher on the Learning Problem

The goal of Supervised Learning is to find a "good" estimator $f_n : \mathcal{X} \to \mathcal{Y}$, approximating the lowest expected risk

$$\inf_{f:\mathcal{X}\to\mathcal{Y}} \mathcal{E}(f), \qquad \mathcal{E}(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y)\, d\rho(x, y)$$

given only a finite number of (training) examples $(x_i, y_i)_{i=1}^n$ sampled independently from the *unknown* distribution ρ.

A Wishlist

What do we mean by... “good”?

A Wishlist

What do we mean by... “good”?

To have a low *excess risk* $\mathcal{E}(f_n) - \mathcal{E}(f_*)$

- **Consistency**. Does $\mathcal{E}(f_n) - \mathcal{E}(f_*) \to 0$
  - in Expectation?
  - in Probability?

  as the size of the training set $S = (x_i, y_i)_{i=1}^n$ of points randomly sampled from ρ grows, $n \to +\infty$.
- **Learning Rates**. How “fast” is consistency achieved? Nonasymptotic bounds: finite sample complexity, tail bounds, error bounds...

Empirical Risk as a Proxy

If ρ is unknown, can we say anything about $\mathcal{E}(f_n) - \inf_{f \in \mathcal{F}} \mathcal{E}(f)$?

We only "glimpse" ρ via the samples $(x_i, y_i)_{i=1}^n$. Can we use them to gather some information about ρ (or better, on $\mathcal{E}(f)$)?

Empirical Risk as a Proxy

If ρ is unknown, can we say anything about $\mathcal{E}(f_n) - \inf_{f \in \mathcal{F}} \mathcal{E}(f)$?

We only "glimpse" ρ via the samples $(x_i, y_i)_{i=1}^n$. Can we use them to gather some information about ρ (or better, on $\mathcal{E}(f)$)?

Consider function $f : \mathcal{X} \to \mathcal{Y}$ and its *empirical risk*

$$\mathcal{E}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Is this a good idea?

Empirical Risk as a Proxy

If ρ is unknown, can we say anything about $\mathcal{E}(f_n) - \inf_{f \in \mathcal{F}} \mathcal{E}(f)$?

We only "glimpse" ρ via the samples $(x_i, y_i)_{i=1}^n$. Can we use them to gather some information about ρ (or better, on $\mathcal{E}(f)$)?

Consider function $f : \mathcal{X} \to \mathcal{Y}$ and its *empirical risk*

$$\mathcal{E}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Is this a good idea? A simple calculation shows that

$$\mathbb{E}_{S \sim \rho^n}[\mathcal{E}_n(f)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{(x_i,y_i) \sim \rho}[\ell(f(x_i), y_i)] = \frac{1}{n} \sum_{i=1}^{n} \mathcal{E}(f) = \mathcal{E}(f)$$

The expectation of $\mathcal{E}_n(f)$ is the expected risk $\mathcal{E}(f)$!

Empirical Risk as a Proxy

If ρ is unknown, can we say anything about $\mathcal{E}(f_n) - \inf_{f \in \mathcal{F}} \mathcal{E}(f)$?

We only "glimpse" ρ via the samples $(x_i, y_i)_{i=1}^n$. Can we use them to gather some information about ρ (or better, on $\mathcal{E}(f)$)?

Consider function $f : \mathcal{X} \to \mathcal{Y}$ and its *empirical risk*

$$\mathcal{E}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Is this a good idea? A simple calculation shows that

$$\mathbb{E}_{S \sim \rho^n}[\mathcal{E}_n(f)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{(x_i,y_i) \sim \rho}[\ell(f(x_i), y_i)] = \frac{1}{n} \sum_{i=1}^{n} \mathcal{E}(f) = \mathcal{E}(f)$$

The expectation of $\mathcal{E}_n(f)$ is the expected risk $\mathcal{E}(f)$!

Empirical Vs Expected

Nice! But... how close is $\mathcal{E}_n(f)$ to $\mathcal{E}(f)$ with respect to n?

Empirical Vs Expected

Nice! But... how close is $\mathcal{E}_n(f)$ to $\mathcal{E}(f)$ with respect to n?

Let X and $(X_i)_{i=1}^n$ be i.i.d. random variables, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Then

$$\mathbb{E}[(\bar{X}_n - \mathbb{E}(X))^2] = \mathrm{Var}(\bar{X}_n)$$

Empirical Vs Expected

Nice! But... how close is $\mathcal{E}_n(f)$ to $\mathcal{E}(f)$ with respect to n?

Let X and $(X_i)_{i=1}^n$ be i.i.d. random variables, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Then

$$\mathbb{E}[(\bar{X}_n - \mathbb{E}(X))^2] = \mathrm{Var}(\bar{X}_n) = \frac{\mathrm{Var}(X)}{n}$$

Therefore, the expected (squared) distance between the empirical mean of the X_i and their expectation $\mathbb{E}(X)$ goes to zero as $O(1/n)$... (Assuming X has finite variance).

Empirical Vs Expected Risk

If $X_i = \ell(f(x_i), y_i)$, we have $\bar{X}_n = \mathcal{E}_n(f)$ and therefore

$$\mathbb{E}[(\mathcal{E}_n(f) - \mathcal{E}(f))^2] = \frac{V_f}{n}$$

Where $V_f = \mathrm{Var}(\ell(f(x), y))$. In particular

$$\mathbb{E}[|\mathcal{E}_n(f) - \mathcal{E}(f)|] \leq \sqrt{\frac{V_f}{n}}$$

Empirical Vs Expected Risk

If $X_i = \ell(f(x_i), y_i)$, we have $\bar{X}_n = \mathcal{E}_n(f)$ and therefore

$$\mathbb{E}[(\mathcal{E}_n(f) - \mathcal{E}(f))^2] = \frac{V_f}{n}$$

Where $V_f = \mathrm{Var}(\ell(f(x), y))$. In particular

$$\mathbb{E}[|\mathcal{E}_n(f) - \mathcal{E}(f)|] \leq \sqrt{\frac{V_f}{n}}$$

Exercise. Can we similarly get a "tail" bound?

$$\mathbb{P}\Big(|\mathcal{E}_n(f) - \mathcal{E}(f)| \leq \varepsilon(n, \delta)\Big) \geq 1 - \delta$$

Empirical Vs Expected

Assume there exists a minimizer $f_* : \mathcal{X} \to \mathcal{Y}$ of the expected risk

$$\mathcal{E}(f_*) = \inf_{f \in \mathcal{F}} \mathcal{E}(f)$$

Then, for any $f : \mathcal{X} \to \mathcal{Y}$ we can **decompose** the excess risk as

$$\mathcal{E}(f) - \mathcal{E}(f_*) = \mathcal{E}(f) - \mathcal{E}_n(f) + \mathcal{E}_n(f) - \mathcal{E}_n(f_*) + \mathcal{E}_n(f_*) - \mathcal{E}(f_*),$$

We can therefore leverage on the statistical relation between $\mathcal{E}_n$ and $\mathcal{E}$ to *study the expected risk in terms of the empirical risk*.

This leads naturally to **Empirical Risk Minimization**

Empirical Risk Minimization (ERM)

Let f_n be the minimizer of the *empirical risk*

$$f_n = \arg\min_{f \in \mathcal{F}} \mathcal{E}_n(f)$$

We automatically have $\mathcal{E}_n(f_n) - \mathcal{E}_n(f_*) \leq 0$ (for **any** training set).

Then...

$$\mathbb{E}\left[\mathcal{E}(f_n) - \mathcal{E}(f_*)\right] \leq \mathbb{E}\left[\mathcal{E}(f_n) - \mathcal{E}_n(f_n)\right] \qquad \textbf{(Question. Why?)}$$

We can "just" focus on studying only the *generalization error*

$$\mathbb{E}\left[\mathcal{E}(f_n) - \mathcal{E}_n(f_n)\right]$$

Generalization Error

How can we control the generalization error

$$\mathcal{E}_n(f_n) - \mathcal{E}(f_n)$$

with respect to the number n of examples?

This question is far from trivial (a key one in SLT, in fact):

In general... $\mathbb{E}[\mathcal{E}_n(f_n) - \mathcal{E}(f_n)]$

Generalization Error

How can we control the generalization error

$$\mathcal{E}_n(f_n) - \mathcal{E}(f_n)$$

with respect to the number n of examples?

This question is far from trivial (a key one in SLT, in fact):

In general... $\mathbb{E}[\mathcal{E}_n(f_n) - \mathcal{E}(f_n)] \neq 0$ (**Question.** Why?)

Generalization Error

How can we control the generalization error

$$\mathcal{E}_n(f_n) - \mathcal{E}(f_n)$$

with respect to the number n of examples?

This question is far from trivial (a key one in SLT, in fact):

In general... $\mathbb{E}[\mathcal{E}_n(f_n) - \mathcal{E}(f_n)] \neq 0$ (**Question.** Why?)

$\mathcal{E}_n$ and f_n *both* depend on the sampled training data. Therefore, we **cannot use the result**

$$\mathbb{E}[|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)|] \leq \sqrt{\frac{\mathrm{Var}(\ell(f_n(x), y))}{n}}$$

which indeed is **not true** in general...

Issues with ERM

Let$\mathcal{X} = \mathcal{Y} = \mathbb{R}$, ρ with dense support[1] and $\ell(y, y) = 0 \; \forall y \in \mathcal{Y}$.

For any $S = (x_i, y_i)_{i=1}^n$ with distinct inputs x_i, let $f_n : \mathcal{X} \to \mathcal{Y}$ be

$$f_n(x) = \begin{cases} y_i & \text{if } x = x_i \quad \text{for some} \quad i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Then, for any number n of training points:

- $\mathbb{E}\left[\mathcal{E}_n(f_n)\right] = 0$
- $\mathbb{E}\left[\mathcal{E}(f_n)\right] = \mathcal{E}(0)$, which is greater than $\mathcal{E}(f^*)$ (unless $f^* \equiv 0$)

Therefore $\mathbb{E}\left[\mathcal{E}(f_n) - \mathcal{E}_n(f_n)\right] = \mathcal{E}(0) \not\to 0$ as n increases!

[1] and such that every pair (x, y) has measure zero according to ρ

Overfitting

An estimator f_n is said to *overfit* the training data if $\forall n \in \mathbb{N}$:

- $\mathbb{E}\left[\mathcal{E}(f_n) - \mathcal{E}(f_*)\right] > C$ for a constant $C > 0$, and
- $\mathbb{E}\left[\mathcal{E}_n(f_n) - \mathcal{E}_n(f_*)\right] \leq 0$

According to this definition ERM overfits...

ERM on Finite Hypotheses Spaces

Is ERM hopeless? Consider the case $\mathcal{X}$ and $\mathcal{Y}$ finite.

Then, $\mathcal{F} = \mathcal{Y}^{\mathcal{X}} = \{f : \mathcal{X} \to \mathcal{Y}\}$ is finite as well (albeit possibly very large), and therefore:

$$\begin{aligned}\mathbb{E}|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| &\leq \mathbb{E} \sup_{f \in \mathcal{F}} |\mathcal{E}_n(f) - \mathcal{E}(f)| \\ &\leq \sum_{f \in \mathcal{F}} \mathbb{E}|\mathcal{E}_n(f) - \mathcal{E}(f)| \leq |\mathcal{F}| \sqrt{V_{\mathcal{F}}/n}\end{aligned}$$

where $V_{\mathcal{F}} = \sup_{f \in \mathcal{F}} V_f$ and $|\mathcal{F}|$ denotes the cardinality of $\mathcal{F}$.

ERM on Finite Hypotheses Spaces

Is ERM hopeless? Consider the case $\mathcal{X}$ and $\mathcal{Y}$ finite.

Then, $\mathcal{F} = \mathcal{Y}^{\mathcal{X}} = \{f : \mathcal{X} \to \mathcal{Y}\}$ is finite as well (albeit possibly very large), and therefore:

$$\begin{aligned}\mathbb{E}|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| &\leq \mathbb{E} \sup_{f \in \mathcal{F}} |\mathcal{E}_n(f) - \mathcal{E}(f)| \\ &\leq \sum_{f \in \mathcal{F}} \mathbb{E}|\mathcal{E}_n(f) - \mathcal{E}(f)| \leq |\mathcal{F}| \sqrt{V_{\mathcal{F}}/n}\end{aligned}$$

where $V_{\mathcal{F}} = \sup_{f \in \mathcal{F}} V_f$ and $|\mathcal{F}|$ denotes the cardinality of $\mathcal{F}$.

Here ERM works! $\lim_{n \to +\infty} \mathbb{E}|\mathcal{E}(f_n) - \mathcal{E}(f_*)| = 0$

ERM on Finite Hypotheses (Sub) Spaces

The same argument holds in general: let $\mathcal{H} \subset \mathcal{F}$ be a *finite* space of hypotheses (even if $\mathcal{F}$ is not). Then,

$$\mathbb{E}|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| \leq |\mathcal{H}|\sqrt{V_{\mathcal{H}}/n}$$

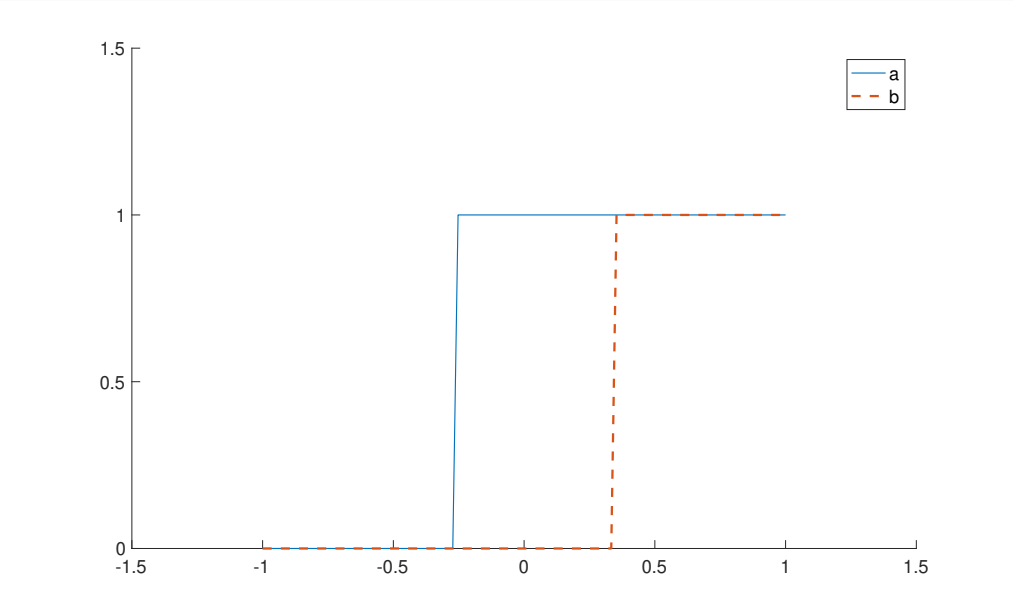
In particular, if $f_* \in \mathcal{H}$, then

$$\mathbb{E}|\mathcal{E}(f_n) - \mathcal{E}(f_*)| \leq |\mathcal{H}|\sqrt{V_{\mathcal{H}}/n}$$

and ERM is a good estimator for the problem considered.

Example: Threshold functions

Consider a binary classification problem $\mathcal{Y} = \{0, 1\}$. Someone has told us that the minimizer of the risk is a "threshold function" $f_{a^*}(x) = \mathbf{1}_{[a^*, +\infty)}$ with $a^* \in [-1, 1]$.



We can learn on $\mathcal{H} = \{f_a | a \in \mathbb{R}\} = [-1, 1]$. However on a computer we can only represent real numbers *up to a given precision*.

Example: Threshold Functions (with precision p)

Discretization: given a $p > 0$, we can consider

$$\mathcal{H}_p = \{f_a \mid a \in [-1, 1],\ a \cdot 10^p = [a \cdot 10^p]\}$$

with $[a]$ denoting the integer part (i.e. the closest integer) of a scalar a. The value p can be interpreted as the "precision" of our space of functions $\mathcal{H}_p$. Note that $|\mathcal{H}_p| = 2 \cdot 10^p$

If $f^* \in \mathcal{H}_p$, then we have automatically that

$$\mathbb{E}|\mathcal{E}(f_n) - \mathcal{E}(f_*)| \leq |\mathcal{H}_p|\sqrt{V_{\mathcal{H}}/n} \leq 10^p/\sqrt{n}$$

(This is because $V_{\mathcal{H}} \leq 1/4$ **Question.** Why?)

Rates in Expectation Vs Probability

In practice, even for small values of p

$$\mathbb{E}|\mathcal{E}(f_n) - \mathcal{E}(f_*)| \leq 10^p/\sqrt{n}$$

will need a very large n in order to have a meaningful bound on the expected error.

Interestingly, we can get much better constants (not rates though!) by working in probability...

Hoeffding's Inequality

Let $X_1, \ldots, X_n$ independent random variables s.t. $X_i \in [a_i, b_i]$.

Let $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Then,

$$\mathbb{P}\left(\left|\overline{X} - \mathbb{E}\,\overline{X}\right| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Applying Hoeffding's inequality

Assume that $\forall f \in \mathcal{H}, x \in \mathcal{X}, y \in \mathcal{Y}$ the loss is bounded $|\ell(f(x), y)| \leq M$ by some constant $M > 0$. Then, for any $f \in \mathcal{H}$ we have

$$\mathbb{P}\left(|\mathcal{E}_n(f) - \mathcal{E}(f)| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2M^2}\right)$$

Controlling the Generalization Error

We would like to control the generalization error $\mathcal{E}_n(f_n) - \mathcal{E}(f_n)$ of our estimator *in probability*. One possible way to do that is by controlling the generalzation error of the whole set $\mathcal{H}$.

$$\mathbb{P}\left(|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| \geq \epsilon\right) \leq \mathbb{P}\left(\sup_{f \in \mathcal{H}} |\mathcal{E}_n(f) - \mathcal{E}(f)| \geq \epsilon\right)$$

The latter term is the probability that *least one* of the events $|\mathcal{E}_n(f) - \mathcal{E}(f)| \geq \epsilon$ occurs for $f \in \mathcal{H}$. In other words the probability of the *union* of such events. Therefore

$$\mathbb{P}\left(\sup_{f \in \mathcal{H}} |\mathcal{E}_n(f) - \mathcal{E}(f)| \geq \epsilon\right) \leq \sum_{f \in \mathcal{H}} \mathbb{P}\left(|\mathcal{E}_n(f) - \mathcal{E}(f)| \geq \epsilon\right)$$

by the so-called *union bound*.

Hoeffding the Generalization Error

By applying Hoeffding's inequality,

$$\mathbb{P}\left(|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| \geq \epsilon\right) \leq 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2M^2}\right)$$

Or, equivalently, that for any $\delta \in (0, 1]$,

$$|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| \leq \sqrt{\frac{2M^2 \log(2|\mathcal{H}|/\delta)}{n}}$$

with probability at least $1 - \delta$.

Example: Threshold Functions (in Probability)

Going back to $\mathcal{H}_p$ space of threshold functions...

$$|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| \leq \sqrt{\frac{4 + 6p - 2\log\delta}{n}}$$

since $M = 1$ and
$\log 2|\mathcal{H}| = \log 4 \cdot 10^p = \log 4 + p \log 10 \leq 2 + 3p$.

For example, let $\delta = 0.001$. We can say that

$$|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| \leq \sqrt{\frac{6p + 18}{n}}$$

holds at least 99.9% of the times.

Bounds in Expectation Vs Probability

Comparing the two bounds

$$\mathbb{E}\,|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| \leq 10^p/\sqrt{n} \qquad \text{(Expectation)}$$

While, with probability greater than 99.9%

$$|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| \leq \sqrt{\frac{6p + 18}{n}} \qquad \text{(Probability)}$$

Although we cannot be 100% sure of it, we can be quite confident that the generalization error will be much smaller than what the bound in expectation tells us...

Rates: note however that the rates of convergence to 0 are the same (i.e. $O(1/\sqrt{n})$).

Improving the bound in Expectation

Exploiting the bound in probability and the knowledge that on $\mathcal{H}_p$ the excess risk is bounded by a constant, we can improve the bound in expectation...

Let X be a random variable s.t. $|X| < M$ for some constant $M > 0$. Then, for any $\epsilon > 0$ we have

$$\mathbb{E}\,|X| \;\leq\; \epsilon\,\mathbb{P}\left(|X| \leq \epsilon\right) + M\mathbb{P}\left(|X| > \epsilon\right)$$

Applying to our problem: for any $\delta \in (0, 1]$

$$\mathbb{E}\,|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| \leq (1-\delta)\sqrt{\frac{2M^2\log(2|\mathcal{H}_p|/\delta)}{n}} + \delta M$$

Therefore only $\log|\mathcal{H}_p|$ appears (no $|\mathcal{H}_p|$ alone).

Infinite Hypotheses Spaces

What if $f_* \in \mathcal{H} \setminus \mathcal{H}_p$ for any $p > 0$?

ERM on $\mathcal{H}_p$ will never minimize the expected risk. There will always be a gap for $\mathcal{E}(f_{n,p}) - \mathcal{E}(f_*)$.

For $p \to +\infty$ it is natural to expect such gap to decrease... **BUT** if p increases too fast (with respect to the number n of examples) we cannot control the generalization error anymore!

$$|\mathcal{E}_n(f_n) - \mathcal{E}(f_n)| \leq \sqrt{\frac{6p + 18}{n}} \to +\infty \qquad \text{for } p \to +\infty$$

Therefore we need to increase p gradually as a function $p(n)$ of the number of training examples. This approach is known as *regularization*.

Approximation Error for Threshold Functions

Consider $f_p = 1_{[a_p, +\infty)} = \arg\min_{f \in \mathcal{H}_p} \mathcal{E}(f)$ with $a_p \in [-1, 1]$.

We decompose the excess risk $\mathcal{E}(f_n) - \mathcal{E}(f_*)$:

$$\mathcal{E}(f_n) - \mathcal{E}_n(f_n) + \underbrace{\mathcal{E}_n(f_n) - \mathcal{E}_n(f_p)}_{\leq 0} + \mathcal{E}_n(f_p) - \mathcal{E}(f_p) + \mathcal{E}(f_p) - \mathcal{E}(f_*)$$

We already know how to control the generalization of f_n (via the supremum over $\mathcal{H}_p$) and f_p (since it is a single function).

Can we say anything about the approximation error

$$\mathcal{E}(f_p) - \mathcal{E}(f_*) \leq \quad ?$$

- Note that this quantity does **not** depend on training data.
- However, it depends on other characteristics of the problem, such as the unknown distribution ρ. How?

Approximation Error - Regularity Assumption

Unfortunately, we cannot study $\mathcal{E}(f_p) - \mathcal{E}(f_*)$ further until we make more assumptions on the problem.

Therefore, in the following we will assume that the underlying (marginal) distribution $\rho_{\mathcal{X}}$ has density $r_\rho : \mathcal{X} \to \mathbb{R}$ with support equal to $\mathcal{X} = [-1, 1]$. Namely, for any integrable $g : \mathcal{X} \to \mathbb{R}$

$$\int_{-1}^{1} g(x)\, d\rho_{\mathcal{X}}(x) = \int_{-1}^{1} g(x) r(x)\, dx$$

In other words, we are asking the learning problem (through ρ) to satisfy some **regularity** assumption (i.e. it cannot be any arbitrarily "bad" learning problem).

What can this restriction buy us?

Apprixiamation Error -for Threshold Functions (II)

Given the above regularity assumption, we can show that

$$\mathcal{E}(f_p) - \mathcal{E}(f_*) \leq |a_p - a_*| \leq \|r_\rho\|_\infty 10^{-p} \qquad \text{(why?)}$$

where we have introduced the sup norm

$$\|r_\rho\|_\infty = \sup_{x \in \mathcal{X}} |r_\rho(x)|$$

Approximation Error for Threshold Functions III

Putting everything together: for any $\delta \in (0, 1]$ and $p \geq 0$,

$$\mathcal{E}(f_n) - \mathcal{E}(f_*) \leq 2\sqrt{\frac{4 + 6p - 2\log\delta}{n}} + \|r_\rho\|_\infty 10^{-p} = \phi(n, \delta, p)$$

holds with probability greater or equal to $1 - \delta$.

So, for any n and δ, we can choose the best precision as

$$p(n, \delta) = \arg\min_{p \geq 0} \phi(n, \delta, p)$$

which leads to an error bound $\epsilon(n, \delta) = \phi(n, \delta, p(n, \delta))$ holding with probability larger or equal than $1 - \delta$.

Regularization

Most hypotheses spaces are “too” large and therefore prone to overfitting. *Regularization* is the process of controlling the “freedom” of an estimator *as a function on the number of training examples*.

Idea. Parametrize $\mathcal{H}$ as a union $\mathcal{H} = \cup_{\gamma>0}\mathcal{H}_\gamma$ of hypotheses spaces $\mathcal{H}_\gamma$ that are not prone to overfitting (e.g. finite spaces). γ is known as the *regularization parameter* (e.g. the precision p in our examples). Assume $\mathcal{H}_\gamma \subset \mathcal{H}_{\gamma'}$ if $\gamma \leq \gamma'$.

Regularization Algorithm. Given n training points, find an estimator $f_{\gamma,n}$ on $\mathcal{H}_\gamma$ (e.g. ERM on $\mathcal{H}_\gamma$). Let $\gamma = \gamma(n)$ increase as $n \to +\infty$.

Regularization and Decomposition of the Excess Risk

Let $\gamma > 0$ and $f_\gamma = \operatorname*{argmin}_{f \in \mathcal{H}_\gamma} \mathcal{E}(f)$

We can decompose the excess risk $\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_*)$ as

$$\underbrace{\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_\gamma)}_{\text{Sample error}} + \underbrace{\mathcal{E}(f_\gamma) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)}_{\text{Approximation error}} + \underbrace{\inf_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}(f_*)}_{\text{Irreducible error}}$$

Irreducible Error

$$\inf_{f \in \mathcal{H}} \mathcal{E}(f) - \mathcal{E}(f_*)$$

Recall: $\mathcal{H}$ is the "largest" possible Hypotheses space we are considering.

If the irreducible error is zero, $\mathcal{H}$ is called *universal* (e.g. the RKHS induced by the Gaussian kernel is a universal Hypotheses space).

Approximation Error

$$\mathcal{E}(f_\gamma) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)$$

- Does not depend on the dataset (deterministic).
- Does depend on the distribution ρ.
- Also referred to as *bias*.

Convergence of the Approximation Error

Under mild assumptions,

$$\lim_{\gamma \to +\infty} \mathcal{E}(f_\gamma) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) = 0$$

Density Results

$$\lim_{\gamma \to +\infty} \mathcal{E}(f_\gamma) - \mathcal{E}(f_*) = 0$$

- Convergence of Approximation error

  +

- Universal Hypotheses space

Note: It corresponds to a density property of the space $\mathcal{H}$ in $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$

Approximation error bounds

$$\mathcal{E}(f_\gamma) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) \leq \mathcal{A}(\rho, \gamma)$$

- No rates without assumptions – related to the so-called *No Free Lunch* Theorem.
- Studied in Approximation Theory using tools such as Kolmogorov n-width, K-functionals, interpolation spaces...

Prototypical result:

If f_* is *s-"regular"*[2] parameter equal to s, then

$$\mathcal{A}(\rho, \gamma) = c\gamma^{-s}.$$

[2] Some abstract notion of regularity parametrizing the class of target functions. Typical example: f_* in a Sobolev space $W^{s,2}$.

Sample Error

$$\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_\gamma)$$

Random quantity depending on the data.

Two main ways to study it:

- Capacity/Complexity estimates on $\mathcal{H}_\gamma$.
- Stability.

Sample Error Decomposition

We have seen how to decompose $\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_\gamma)$ in

$$\underbrace{\mathcal{E}(f_{\gamma,n}) - \mathcal{E}_n(f_{\gamma,n})}_{\text{Generalization error}} + \underbrace{\mathcal{E}_n(f_{\gamma,n}) - \mathcal{E}_n(f_\gamma)}_{\text{Excess empirical Risk}} + \underbrace{\mathcal{E}_n(f_\gamma) - \mathcal{E}(f_\gamma)}_{\text{Generalization error}}$$

Generalization Error(s)

As we have observed,

$$\mathcal{E}(f_{\gamma,n}) - \mathcal{E}_n(f_{\gamma,n}) \qquad \text{and} \qquad \mathcal{E}_n(f_\gamma) - \mathcal{E}(f_\gamma)$$

Can be controlled by studying the *empirical process*

$$\sup_{f \in \mathcal{H}_\gamma} |\mathcal{E}_n(f) - \mathcal{E}(f)|$$

Example: we have already observed that for a finite space $\mathcal{H}_\gamma$

$$\mathbb{E}\left[\sup_{f \in \mathcal{H}_\gamma} |\mathcal{E}_n(f) - \mathcal{E}(f)|\right] \leq |\mathcal{H}_\gamma| \sqrt{\frac{V_{\mathcal{H}_\gamma}}{n}}$$

ERM on Finite Spaces and Computational Efficiency

The strategy used for threshold functions can be generalized to any $\mathcal{H}$ for which it is possible to find a *finite* discretization $\mathcal{H}_p$ with respect to the $L_1(\mathcal{X}, \rho_{\mathcal{X}})$ norm (e.g. $\mathcal{H}$ compact with respect to such norm).

However, it could be computationally very expensive to find the empirical risk minimizer on a discretization $\mathcal{H}_p$, since in principle it could be necessary to evaluate $\mathcal{E}_n(f)$ for any $f \in \mathcal{H}_p$.

ERM on Convex Spaces?

As we have seen in previous classes, ERM on *convex* (thus dense) spaces is often much more amenable to computations.

In principle, we have observed that on infinite hypotheses spaces it is difficult to control the generalization error but...

...we might be able to leverage the **discretization argument** used for threshold functions to control the generalization error of ERM for a larger family of hypotheses spaces.

Example: Risks for Continuous functions

Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact (i.e., closed and bounded) space and $C(\mathcal{X})$ be the space of continuous functions. Let $\|\cdot\|_\infty$ be defined for any $f \in C(\mathcal{X})$ as $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$.

If the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is such that $\ell(\cdot, y)$ is uniformly Lipschitz with constant $L > 0$, for any $y \in \mathcal{Y}$, we have that

1) $|\mathcal{E}(f_1) - \mathcal{E}(f_2)| \leq L\|f_1 - f_2\|_{L^1(\mathcal{X}, \rho_\mathcal{X})} \leq L\|f_1 - f_2\|_\infty$, and

2) $$|\mathcal{E}_n(f_1) - \mathcal{E}_n(f_2)| \leq \frac{1}{n}\sum_{i=1}^{n} |\ell(f_1(x_i), y_i) - \ell(f_2(x_i), y_i)| \leq L\|f_1 - f_2\|_\infty$$

Therefore, “close” functions in $\|\cdot\|_\infty$ will have similar expected and empirical risks!

Example: Covering numbers

We define the *covering number* of $\mathcal{H}$ of radius $\eta > 0$ as the cardinality of a minimal cover of $\mathcal{H}$ with balls of radius η.

$$\mathcal{N}(\mathcal{H}, \eta) = \inf\{m \mid \mathcal{H} \subseteq \bigcup_{i=1}^{m} B_\eta(h_i) \;\; h_i \in \mathcal{H}\}$$

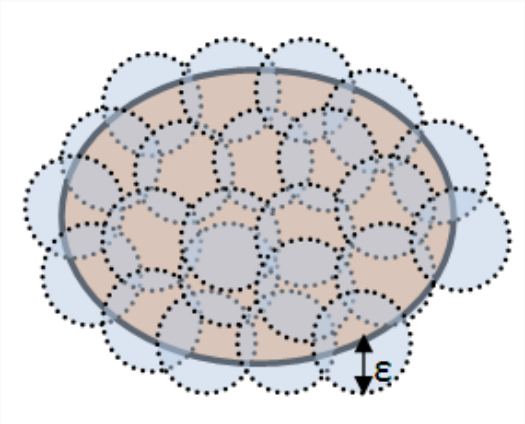


Image credits: Lorenzo Rosasco.

Example. If $\mathcal{H} \cong B_R(0)$ is a ball of radius R in $\mathbb{R}^d$:

$$\mathcal{N}(B_R(0), \eta) = (4R/\eta)^d$$

Example: Covering numbers (continued)

Note that the resolution η is related to the precision in the sense of a distance between two hypothesis. We can apply the same reasoning used for threshold functions

$$\mathbb{E}\left[\sup_{f\in\mathcal{H}} |\mathcal{E}(f_n) - \mathcal{E}(f)|\right] \leq 2L\eta + \mathcal{N}(\mathcal{H},\eta)\sqrt{\frac{V_{\mathcal{H}}}{n}}$$

For $\eta \to 0$ the covering number $\mathcal{N}(\mathcal{H},\eta) \to +\infty$. However, for $n \to +\infty$ the bound tends to zero.

It is typically possible to find an $\eta(n)$ for which the bound tends to zero as $n \to +\infty$.

(**Exercise.** find such an $\eta(n)$ for $\mathcal{H}$ a ball of radius R)

Example: Covering numbers (continued)

Same argument can be reproduced for bounds in probability, namely for any $\delta \in [0, 1)$,

$$\sup_{f \in \mathcal{H}} |\mathcal{E}(f_n) - \mathcal{E}(f)| \leq 2L\eta + \sqrt{\frac{2M^2 \log(2\mathcal{N}(\mathcal{H}, \eta)/\delta)}{n}}$$

holds with probability at least $1 - \delta$.

Complexity Measures

In general, the error

$$\sup_{f \in \mathcal{H}_\gamma} |\mathcal{E}_n(f) - \mathcal{E}(f)|$$

Can be controlled via capacity/complexity measures:

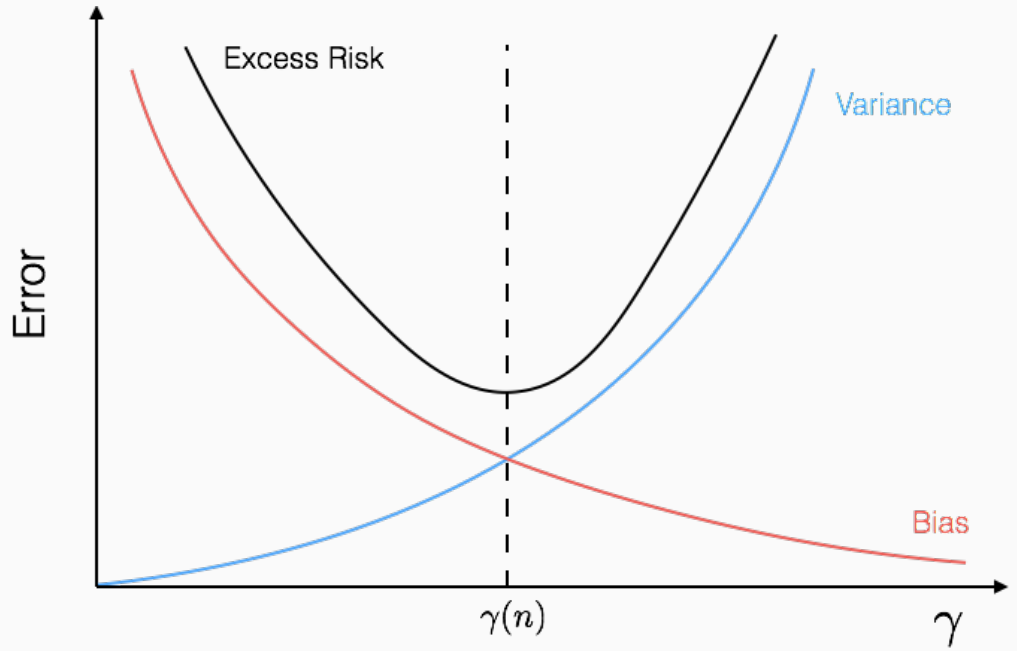
- Covering numbers,
- combinatorial dimension, e.g. VC-dimension, fat-shattering dimension
- Rademacher complexities
- Gaussian complexities
- …

Prototypical Results

A prototypical result (under suitable assumptions, e.g. regularity of f_*):

$$\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f^*) \leq \underbrace{\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_\gamma)}_{\lesssim \gamma^\beta n^{-\alpha}} + \underbrace{\mathcal{E}(f_\gamma) - \mathcal{E}(f^*)}_{\lesssim \gamma^{-\tau}}$$

(Variance) (Bias)



Choosing $\gamma(n)$ in practice

The best $\gamma(n)$ depends on the unknown distribution ρ. So how can we choose such parameter in practice?

Problem known as *model selection*. Possible approaches:

- Cross validation,
- complexity regularization/structural risk minimization,
- balancing principles.
- ...

Abstract Regularization

We got a new perspective on the concept of *regularization*: controlling the expressiveness of the hypotheses space according to the number of training examples in order to guarantee good prediction performance and consistency.

There are many ways to implement this strategy in practice:

- Tikhonov (and Ivanov) regularization
- Spectral filtering
- Early stopping
- Random sampling
- ...

Wrapping Up

Building on a few (reasoble?) assumptions on the learning problem, we:

- Have shown how the empirical risk can be a proxy for the expected.
- Identified the main reasons behind overfitting and discussed how to counteract it (in a more principle way!).
- Highlighted the key role played by the choice of the hypotheses space and how their “complexity” affect performance.

Next class we will focus on one such measure of complexity and derive upper bounds on the generalization error.

Recommended Reading

Chapter 4, 5 and 6 of Shalev-Shwartz, Shai, and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.