Class 2
Overfitting & Regularization

Carlo Ciliberto
Department of Computer Science, UCL

October 13, 2017

Last Class

The goal of Statistical Learning Theory 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.

Last Class: The SLT Wishlist

What does "good" estimator mean? Low *excess risk* $\mathcal{E}(f_n) - \mathcal{E}(f_*)$

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

  with respect to a training set $S = (x_i, y_i)_{i=1}^n$ of points randomly sampled from ρ.

- **Learning Rates**. How "fast" is consistency achieved? Nonasymptotic bounds: finite sample complexity, tail bounds, error bounds...

Last Class (Expected Vs Empirical Risk)

Approximate the expected risk of $f : \mathcal{X} \to \mathcal{Y}$ via its empirical risk

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

- Expectation: $\mathbb{E}|\mathcal{E}_n(f) - \mathcal{E}(f)| \leq \sqrt{\frac{V_f}{n}}$
- Probability (e.g. using Chebyshev):

$$\mathbb{P}\left(|\mathcal{E}_n(f) - \mathcal{E}(f)| \geq \epsilon\right) \leq \frac{V_f}{n\epsilon^2} \qquad \forall \epsilon > 0$$

where $V_f = \mathrm{Var}_{(x,y)\sim\rho}(\ell(f(x), y))$.

Last Class (Empirical Risk Minimization)

Idea: if $\mathcal{E}_n$ is a good approximation to $\mathcal{E}$, then we could use

$$f_n = \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{E}_n(f)$$

to approximate f_*. This is known as *empirical risk minimization (ERM)*

Note: If we sample the points in $S = (x_i, y_i)_{i=1}^n$ independently from ρ, the corresponding $f_n = f_S$ is a random variable and we have

$$\mathbb{E}\ \mathcal{E}(f_n) - \mathcal{E}(f_*) \leq \mathbb{E}\ \mathcal{E}(f_n) - \mathcal{E}_n(f_n)$$

Question: does $\mathbb{E}\ \mathcal{E}(f_n) - \mathcal{E}_n(f_n)$ go to zero as n increases?

Issues with ERM

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

For any set $(x_i, y_i)_{i=1}^n$ s.t. $x_i \neq x_j \ \forall i \neq j$ let $f_n : \mathcal{X} \to \mathcal{Y}$ be such that

$$f_n(x) = \begin{cases} y_i & \text{if } x = x_i \ \exists i \in \{1, \dots n\} \\ 0 & \text{otherwise} \end{cases}$$

Then, for any number n of training points:

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

Therefore $\mathbb{E}\ \mathcal{E}(f_n) - \mathcal{E}_n(f_n) = \mathcal{E}(0) \nrightarrow 0$ 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 for any $n \in \mathbb{N}$:

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

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 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 $\mathcal{F}$.

Then ERM works! Namely: $\lim_{n \to +\infty} \mathbb{E}|\mathcal{E}(f_n) - \mathcal{E}(f)| = 0$

ERM on Finite Hypotheses (Sub) Spaces

The same argument holds in general: let $\mathcal{H} \subset \mathcal{F}$ be a *finite* space of hypotheses. Then,

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

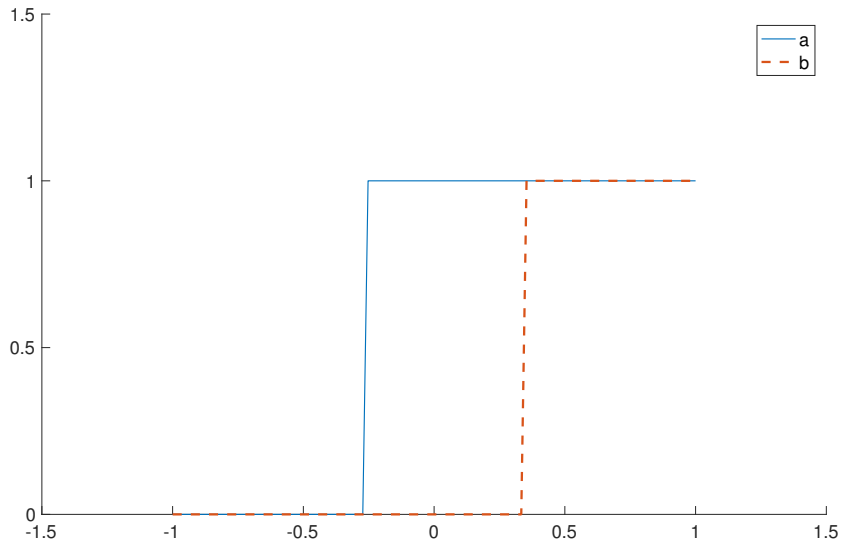
In particular, if $f_* \in \mathcal{H}$, then

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

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

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

($V_{\mathcal{H}} \leq 1$ since ℓ is the 0-1 loss and therefore $|\ell(f(x), y)| \leq 1$ for any $f \in \mathcal{H}$)

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

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

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,

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

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!

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

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

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 point, 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_* has *"smoothness"*[2].

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

[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 the sample error $\mathcal{E}(f_{\gamma,n}) - \mathcal{E}(f_\gamma)$ 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 $\mathcal{H}_\gamma$

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

Example: Covering numbers

We already observed that for a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, such that $\ell(\cdot, y)$ is uniformly Lipschitz with constant $C > 0$, for any $f_1, f_2 : \mathcal{X} \to \mathcal{Y}$ we have

$$\mathcal{E}(f_1) - \mathcal{E}(f_2) \leq C \int_{\mathcal{X}} |f_1(x) - f_2(x)| \, d\rho_{\mathcal{X}}(x) = C\|f_1 - f_2\|_{L^1(\mathcal{X}, \rho_{\mathcal{X}})}$$

Therefore, functions in $\mathcal{H} \subset L^1(\mathcal{X}, \rho_{\mathcal{X}})$ that are sufficiently close in L_1 norm, will have similar risk!

Let consider a cover of the hypotheses space $\mathcal{H}$ with balls of radius ϵ.

Example: Covering numbers (continued)

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

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

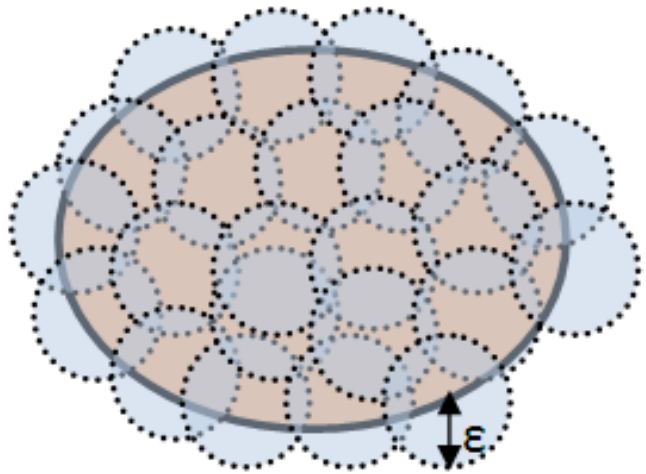


Image credits: Lorenzo Rosasco.

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

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

Example: Covering numbers (continued)

Putting the two together

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{f\in\mathcal{H}} |\mathcal{E}_n(f) - \mathcal{E}(f)| \geq \epsilon \Big) \\
&\qquad\qquad \leq \mathbb{P}\left(\sup_{i=1,\ldots,\mathcal{N}(\mathcal{H},\frac{\epsilon}{2C})} |\mathcal{E}_n(h_i) - \mathcal{E}(h_i)| \geq \epsilon/2 \right) \\
&\qquad\qquad \leq 2\, \mathcal{N}(\mathcal{H}, \frac{\epsilon}{2C}) \exp\left(-\frac{n\epsilon^2}{8M^2} \right)
\end{aligned}$$

For $\epsilon \to 0$ the covering number $\mathcal{N}(\mathcal{H}, \epsilon) \to +\infty$. However for $n \to +\infty$ the bound tends to zero. It is typically possible to show that there exists an $\epsilon(n)$ for which the bound tends to zero as $n \to +\infty$.

Complexity Measures

In general, the error

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

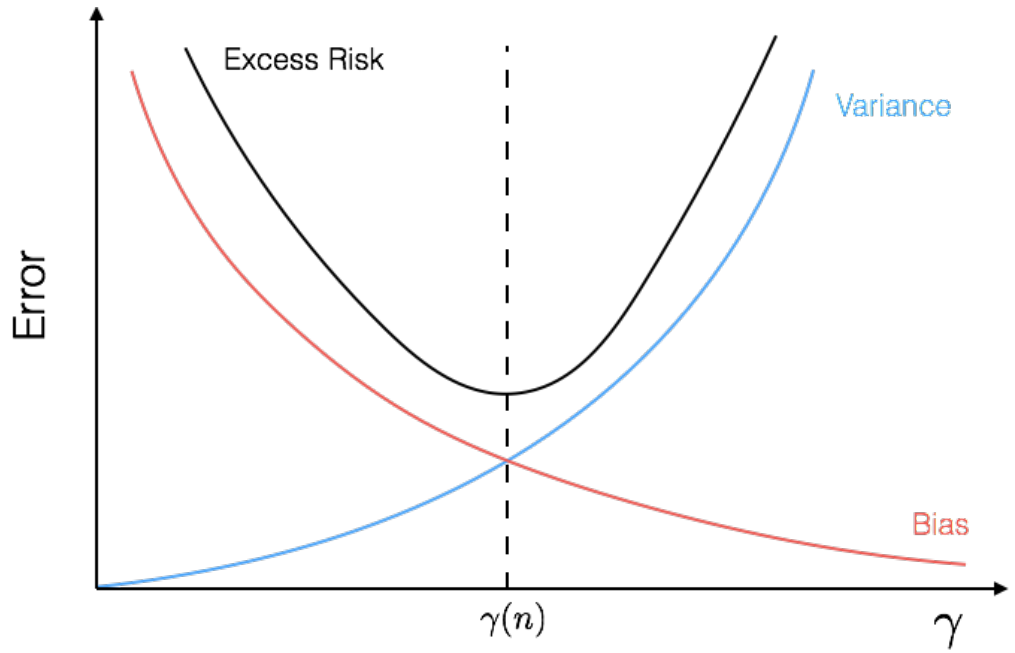
Can be controlled via capacity/complexity measures:

- covering numbers,
- combinatorial dimension, e.g. VC-dimension, fat-hattering dimension
- Rademacher complexities
- Gaussian complexities
- ...

Prototypical Results

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

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



Goal: find the $\gamma(n)$ achieving the best Bias - Variance

Choosing $\gamma(n)$

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 just got our first introduction to the concept of *regularization*: controlling the expressiveness of the hypotheses space according to the number of training examples in order to guarantee good prediciton performance and consistency.

There are many ways to implement this strategy in practice (we will see some of them in this course):

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

Wrapping Up

This class:

- Overfitting
- Controlling the Generalization error
- Abstract Regularization

Next class: Tikhonov Regularization