Machine learning - Master ICFP 2019-2020

Statistical Learning Theory

Francis Bach

January 24, 2020

1 Announcements

- Don't forget to register online (if not done so already).
- Send outcomes of practical sessions to `lenaicfrancisml@gmail.com` (all other emails to our own email addresses).
- Not all results are proved in class.

2 Introduction - review

- Training set: observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, of inputs/outputs, features/variables are independent and identically distributed (i.i.d.) random variables with common distribution P.
- $\mathcal{X}$ can de diverse, $\mathcal{Y}$ is typically $\{0, 1\}$ (binary classification) or $\mathbb{R}$ (regression).
- A machine learning algorithm $\mathcal{A}$ is a function that goes from $\bigcup_{n>0} (\mathcal{X} \times \mathcal{Y})^n$ to a function $\hat{f}_n$ from $\mathcal{X}$ to $\mathcal{Y}$, called an "estimator". We will use the notation $\hat{f}_n = \mathcal{A}(\mathcal{D}_n)$, and often only $\hat{f}_n$.
- We consider a fixed (testing) distribution P on $\mathcal{X} \times \mathcal{Y}$ and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$; $\ell(y, z)$ is the loss of predicting z while the true label is y. **We assume that the testing distribution is the same as the training distribution.**
- Risk, or generalization performance of a prediction function $f : \mathcal{X} \to \mathcal{Y}$:

$$\mathcal{R}(f) = \mathbb{E}\Big[\ell(Y, f(X))\Big].$$

  Be careful with the randomness or lack thereof of f: $\hat{f}_n$ depends on the training data and not on the testing data, and thus $\mathcal{R}(\hat{f}_n)$ is random because of the dependence on the training data $\mathcal{D}_n$.

  The function $\mathcal{R}$ depends on the distribution P on (X, Y). We sometimes use the notation $\mathcal{R}_P(f)$ to make it explicit.

- Binary classification: $\mathcal{Y} = \{0, 1\}$ (or often $\mathcal{Y} = \{-1, 1\}$), and $\ell(y, z) = 1_{y \neq z}$ ("0-1" loss). Then $\mathcal{R}(f) = \mathbb{P}(f(X) \neq Y)$.
- Regression: $\mathcal{Y} = \mathbb{R}$ and $\ell(y, z) = (y - z)^2$ (square loss). Then $\mathcal{R}(f) = \mathbb{E}(Y - f(X))^2$.

- Target function = Bayes predictor $f^* \in \arg\min \mathcal{R}(f) = \mathbb{E}\Big[\ell(Y, f(X))\Big]$.

Proposition 1 (Bayes predictor) *The risk is minimized at a* Bayes predictor $f^* : \mathcal{X} \to \mathcal{Y}$ *satisfying for all* $x \in \mathcal{X}$, $f^*(x) \in \arg\min_{z \in \mathcal{Y}} \mathbb{E}(\ell(Y, z)|X = x)$. *The Bayes risk* $\mathcal{R}^*$ *is the risk of all Bayes predictors and is equal to*

$$\mathcal{R}^* = \mathbb{E}_{x \sim P_X} \inf_{z \in \mathcal{Y}} \mathbb{E}(\ell(Y, z)|X = x).$$

Note that (a) the Bayes predictor is not unique, but that all lead to the same Bayes risk, and (b) that the Bayes risk is usually non zero (unless the dependence between x and y is deterministic).

- For binary classification: $\mathcal{Y} = \{0, 1\}$ and $\ell(y, z) = 1_{y \neq z}$, the Bayes predictor is $f^*(x) \in \arg\max_{i \in \{1, \dots, k\}} \mathbb{P}(Y = i|X = x)$.
- For regression: $\mathcal{Y} = \mathbb{R}$ and $\ell(y, z) = (y - z)^2$, the Bayes predictor is $f^*(x) = \mathbb{E}(Y|X = x)$.

- Goal of supervised machine learning: estimate f^*, knowing only the data $\mathcal{D}_n$ *and* the loss ℓ.

Definition 1 (Excess risk) *The excess risk of a function from* $f : \mathcal{X} \to \mathcal{Y}$ *is equal to* $\mathcal{R}(f) - \mathcal{R}^*$ *(it is always non-negative).*

3 Statistical learning theory

3.1 Consistency and learning rates

- One aim of learning theory is to prove "consistency' of learning algorithms. Essentially, we want $\mathcal{R}(\hat{f}_n) - \mathcal{R}^* = \mathcal{R}(\mathcal{A}(\mathcal{D}_n)) - \mathcal{R}^*$ to go to zero when the number of observations are i.i.d. Since $\mathcal{R}(\mathcal{A}(\mathcal{D}_n)) - \mathcal{R}^*$ is a random variable (due to the randomness of the data) whose distribution depends on P, several criteria exist (in expectation, almost surely, in probability)

Definition 2 (Consistency) *An algorithm* $\mathcal{A}$ *is said "consistent" is for a probability distribution* P *generating the i.i.d. data* $\mathcal{D}_n$, *if*

$$\lim_{n \to +\infty} \mathbb{E}\big[\mathcal{R}(\mathcal{A}(\mathcal{D}_n)) - \mathcal{R}^*\big] = 0.$$

It is said "strongly consistent" if $\lim_{n \to +\infty} \big[\mathcal{R}(\mathcal{A}(\mathcal{D}_n)) - \mathcal{R}^*\big] = 0$ *almost surely (i.e., with probability one).*

The algorithm is said "probably approximately correct" (PAC), if for any $\varepsilon > 0$, $\lim_{n \to +\infty} \mathbb{P}\big[\mathcal{R}(\mathcal{A}(\mathcal{D}_n)) - \mathcal{R}^*\big] > \varepsilon\big] = 0$. *This corresponds to convergence in probability.*

- Consistency is the first requirement we can expect from a learning algorithm. However, the speed of convergence is also important, and we call these "learning rates".

  Definition 3 (Learning rates) *The sequence (ε_n) is learning rate in expectation for an algorithm $\mathcal{A}$ and for the probability distribution P generating the i.i.d. data $\mathcal{D}_n$, if*

  $$\forall n > 0,\ \mathbb{E}\big[\mathcal{R}(\mathcal{A}(\mathcal{D}_n)) - \mathcal{R}^*\big] \leqslant \varepsilon_n.$$

  High probabiity learning rates can also be defined.

3.2 No free lunch theorems

"Learning is not possible without assumptions." See [1, Chapter 7] for details.

The following theorem shows that for any algorithm, for a fixed n, there is a data distribution that makes the algorithm useless.

Theorem 1 (no free lunch - fixed n) *Consider the binary classification with $0-1$ loss, with $\mathcal{X}$ infinite. Let $\mathcal{P}$ denote the set of all probability distributions on $\mathcal{X} \times \{0,1\}$. For any $n > 0$ and learning algorithm $\mathcal{A}$,*

$$\sup_{P \in \mathcal{P}} \mathbb{E}\Big[\mathcal{R}_P(\mathcal{A}(\mathcal{D}_n(P)))\Big] - \mathcal{R}^*_P \geqslant 1/2.$$

Proof Let k be a positive integer. Without loss of generality, we can assume that $\mathbb{N} \subset \mathcal{X}$.

Given $r \in \{0,1\}^k$, we define the distribution P such that $\mathbb{P}(X = j, Y = r_j) = 1/k$; that is, we choose one of the first k elements uniformly at random, and then Y is selected deterministically as $Y = R_X$. Thus the Bayes risk is zero: $\mathcal{R}^*_P = 0$.

Denoting $\hat{f}_{\mathcal{D}_n} = \mathcal{A}(\mathcal{D}_n(P))$, and $S(r) = \mathbb{E}\Big[\mathcal{R}_P(\hat{f}_{\mathcal{D}_n})\Big]$ the expected risk, we want to maximize $S(r)$ with respect to $r \in \{0,1\}^k$; the maximum is greater than the expectation of $S(r)$ for any distribution R on r, in particular the uniform distribution (each r_j an independent unbiased Bernoulli variable). Then

$$\begin{aligned}
\max_{r \in \{0,1\}^k} S(r) &\geqslant \mathbb{E}_{r\sim R} S(r) \\
&= \mathbb{P}(\hat{f}_{\mathcal{D}_n}(X) \neq Y) = \mathbb{P}(\hat{f}_{\mathcal{D}_n}(X) \neq r_X)
\end{aligned}$$

because X is almost surely in $\{1, \dots, k\}$ and $Y = r_X$ almost surely. Then

$$\begin{aligned}
\mathbb{E}_{r\sim R} S(r) &= \mathbb{E}\Big[\mathbb{P}\big(\hat{f}_{\mathcal{D}_n}(X) \neq r_X \big| X_1, \dots, X_n, r_{X_1}, \dots, r_{X_n}\big)\Big] \\
&\geqslant \mathbb{E}\Big[\mathbb{P}\big(\hat{f}_{\mathcal{D}_n}(X) \neq r_X \ \&\ X \notin \{X_1, \dots, X_n\} \big| X_1, \dots, X_n, r_{X_1}, \dots, r_{X_n}\big)\Big] \\
&= \mathbb{E}\Big[\frac{1}{2}\mathbb{P}\big(X \notin \{X_1, \dots, X_n\} \big| X_1, \dots, X_n, r_{X_1}, \dots, r_{X_n}\big)\Big],
\end{aligned}$$

by the law of total expectation on using monotonicity of probabilities, and because we have $\mathbb{P}\big(\hat{f}_{\mathcal{D}_n}(X) \neq r_X \big| X \notin \{X_1, \dots, X_n\}, X_1, \dots, X_n, r_{X_1}, \dots, r_{X_n}\big) = 1/2$. Thus,

$$\mathbb{E}_{r\sim R} S(r) = \frac{1}{2}\mathbb{P}\big(X \notin \{X_1, \dots, X_n\}\big) = \frac{1}{2}\mathbb{E}\Big[\prod_{i=1}^{n} \mathbb{P}(X_i \neq X | X)\Big] = \frac{1}{2}\big(1 - 1/k\big)^n.$$

Given n, we can let k tend to infinity to conclude. ■

A caveat is that the hard distribution may depend on n. The following theorem is given without proof and is much stronger [1, Theorem 7.2], as it more convincingly shows that learning can be arbitrarily slow without assumption.

Theorem 2 (no free lunch - sequence of errors) *Consider the binary classification with $0-1$ loss, with $\mathcal{X}$ infinite. Let $\mathcal{P}$ denote the set of all probability distributions on $\mathcal{X} \times \{0,1\}$. For any decreasing sequence a_n tending to zero and such that $a_1 \leqslant 1/16$, for any learning algorithm $\mathcal{A}$, there exists $P \in \mathcal{P}$, such that for all $n \geqslant 1$:*

$$\mathbb{E}\Big[\mathcal{R}_P(\mathcal{A}(\mathcal{D}_n(P)))\Big] - \mathcal{R}_P^* \geqslant a_n.$$

Therefore, we need to impose some restrictions on the learning problem to reach non-trivial rates.

3.3 Classes of learning problems

- The goal is to provide some guarantees of performance on unseen data. A common assumption is that the data $\mathcal{D}_n(P) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is obtained as independent and identically distributed (i.i.d.) observations, from a distribution P **within a class $\mathcal{P}$ of distributions satisfying some regularity properties** (e.g., the inputs live in a compact space, or the dependence between y and x is at most of some complexity).

- An algorithm $\mathcal{A}$ is a mapping from $\mathcal{D}_n(P)$ (for any n) to a function from $\mathcal{X}$ to $\mathcal{Y}$. The risk depends on the probability distribution $P \in \mathcal{P}$, as $\mathcal{R}_P(f)$. The goal is to find $\mathcal{A}$ such that the risk

  $$\mathcal{R}_P(\mathcal{A}(\mathcal{D}_n(P)))$$

  is small, assuming $\mathcal{D}_n(P)$ is sampled from P, but without knowing which $P \in \mathcal{P}$ is considered. Moreover, the risk is random because $\mathcal{D}_n$ is random. There are several ways of dealing with the randomness to obtain a criterion.

  The simplest one is to take the expectation, and we thus aim at finding an algorithm $\mathcal{A}$ such that

  $$\sup_{P \in \mathcal{P}} \mathbb{E}\Big[\mathcal{R}_P(\mathcal{A}(\mathcal{D}_n(P)))\Big]$$

  is as small as possible (note the supremum with respect to $P \in \mathcal{P}$). This is typically a function of the sample size n and of properties of $\mathcal{X}$, $\mathcal{Y}$ and the allowed set of problems $\mathcal{P}$ (e.g., dimension of $\mathcal{X}$, number of parameters). The algorithm is said consistent over $\mathcal{P}$ if the quantity above goes to zero when n tends to infinity.

- Lower-bounding the optimal performance: in some set-ups, it is possible to show that the infimum over all algorithms is greater than a certain quantity. Machine learners are happy when upper-bounds and lower-bounds match.

- PAC learning: for a given $\delta \in (0,1)$ and $\varepsilon > 0$, for any $P \in \mathcal{P}$:

  $$\mathbb{P}\Big(\Big[\mathcal{R}_P(\mathcal{A}(\mathcal{D}_n(P)))\Big] \leqslant \varepsilon\Big) \geqslant 1 - \delta.$$

The crux is to find ε which is as small as possible (typically as a function of δ and n).

- The analysis can be "non-asymptotic", with an upper-bound with explicit dependence on all quantities; the bound is then valid for all n, even if sometimes vacuous (e.g., a bound greater than 1 for a loss uniformly bounded by 1).

  The analysis can also be "asymptotic", where for examples n goes to infinity and limits are taken (alternatively, several quantities can be made to grow simultaneously).

4 Empirical risk minimization

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

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

The most classical example is linear least-squares regression, where we minimize

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - w^\top \Phi(x_i))^2,$$

where f is linear in some feature vector $\Phi(x) \in \mathbb{R}^d$ (no need for $\mathcal{X}$ to be a vector space). The vector $\Phi(x)$ can be quite large (or even implicit, like in kernel methods). Other examples include neural networks.

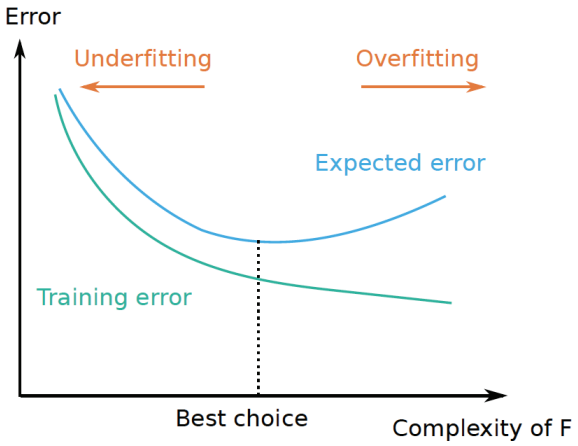
- Risk decomposition in estimation error + approximation error : given any $\hat{f} \in \mathcal{F}$,

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

  When the "number" of models in $\mathcal{F}$ grows, the approximation error is deterministic and goes down, while the estimation error is random and goes up.

  Examples of approximations by polynomials in one-dimensional regression.

  Typical curve:



5 Approximation error

- This means bounding $\inf_{f \in \mathcal{F}} \mathcal{R}(f) - \mathcal{R}^*$ and requires assumptions on the target function f^* to achieve non-trivial learning rates.
- The simplest example is regression on $\mathcal{X} = [0, 1]$, which can be generalized "semi-easily" to multiple dimensions and other losses.
- We consider only the uniform distribution on $[0, 1]$, we can then represent functions by their Fourier series

$$f(x) = \sum_{k \in \mathbb{Z}} c_k(f) e^{i2k\pi x}.$$

  The functions $\left(x \mapsto e^{i2k\pi x}\right)_{k \in \mathbb{Z}}$ form an orthonormal basis of $L^2([0, 1])$, and we are going to consider classes of functions $\mathcal{F}_{\alpha,C}$ defined as

$$\mathcal{F}_{\alpha,C} = \left\{ f \in L^2([0, 1]),\ \sum_{k \in \mathbb{Z}} |c_k(f)|^2 (1 + k^{2\alpha}) \leqslant C^2 \right\}.$$

  These are functions for which $\int_0^1 |f(x)|^2 + \frac{1}{(2k\pi)^{2\alpha}} \int_0^1 |f^{(\alpha)}(x)|^2 dx \leqslant c^2$. This is the usual Sobolev space (it happens to be a Hilbert space, see lecture on kernel methods).
- Given our assumptions: $\mathcal{R}(f) - \mathcal{R}^* = \int_0^1 |f(x) - f^*(x)|^2 dx = \sum_{k \in \mathbb{Z}} |c_k(f) - c_k(f^*)|^2$, and we are thus looking at approximations in ℓ_2.
- No assumptions on f^*: the Sobolev spaces are known to be dense in $L^2([0, 1])$, therefore

$$\inf_{f \in \mathcal{F}_{\alpha,C}} \mathcal{R}(f) - \mathcal{R}^*$$

  tends to zero when C tends to infinity, but no explicit rates can be given.
- Well-specified assumption: if $f^* \in \mathcal{F}_{\alpha^*,C^*}$, then, *if we choose the correct* $\alpha = \alpha^*$:

$$\inf_{f \in \mathcal{F}_{\alpha^*,C}} \mathcal{R}(f) - \mathcal{R}^*$$

  is equal to zero for $C \geqslant C^*$ (but this requires to know α^* in advance).
- As an example of Poincaré inequalities, $\sum_{k \in \mathbb{Z}} |c_k(f)|^2 (1 + k^{2\alpha})$ is increasing in α, thus if $f^* \in \mathcal{F}_{\alpha^*,C^*}$ then $f^* \in \mathcal{F}_{\alpha,C^*}$ for $\alpha \leqslant \alpha^*$. Thus, for approximation purposes, choosing a small α for a class a function is advantageous. However, since estimation error will grow with the size of the space, and thus decrease with α, we can also choose larger α, which could then be larger than α^*. Then we have an approximation error as follows:

$$\begin{aligned} & \sup_{f^* \in \mathcal{F}_{\alpha^*,C^*}} \inf_{f^* \in \mathcal{F}_{\alpha,C}} \mathcal{R}(f) - \mathcal{R}^* \\ = & \sup_{\sum_{k \in \mathbb{Z}} |c_k(f^*)|^2 (1 + k^{2\alpha^*}) \leqslant (C^*)^2} \quad \inf_{\sum_{k \in \mathbb{Z}} |c_k(f)|^2 (1 + k^{2\alpha}) \leqslant C^2} \quad \sum_{k \in \mathbb{Z}} |c_k(f) - c_f(f^*)|^2. \end{aligned}$$

We can now upper bound this quantity. For example, we can consider $c_k(f) = c_k(f_*)$ for $|k| \leqslant K$ and 0 otherwise, for which we have, for $\alpha \geqslant \alpha^*$:

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} |c_k(f) - c_f(f^*)|^2 &= \sum_{|k|>K} c_f(f^*)|^2 \\
&\leqslant \sum_{|k|>K} |c_f(f^*)|^2 \frac{1+k^{2\alpha^*}}{1+K^{2\alpha^*}} \leqslant \frac{(C^*)^2}{K^{2\alpha^*}} \\
\sum_{k\in\mathbb{Z}} |c_k(f)|^2(1+k^{2\alpha}) &= \sum_{|k|\leqslant K} |c_k(f^*)|^2(1+k^{2\alpha}) \\
&\leqslant \sum_{|k|\leqslant K} |c_k(f^*)|^2(1+k^{2\alpha^*})K^{2\alpha-2\alpha^*} \leqslant (C^*)^2 K^{2\alpha-2\alpha^*}.
\end{aligned}$$

Thus, for $K = (C/C^*)^{1/(2\alpha-2\alpha^*)}$, we get an approximation error of

$$\frac{(C^*)^2}{K^{2\alpha^*}} = (C^*)^2(C^*/C)^{\alpha^*/(\alpha-\alpha^*)} = \frac{(C^*)^{2+\alpha^*/(\alpha-\alpha^*)}}{C^{\alpha^*/(\alpha-\alpha^*)}}.$$

which is decreasing in C and increasing in C^*.

6 Estimation error

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

$$\begin{aligned}
\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f) = \mathcal{R}(\hat{f}) - \mathcal{R}(g) &= \Big\{\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})\Big\} + \Big\{\hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(g)\Big\} + \Big\{\hat{\mathcal{R}}(g) - \mathcal{R}(g)\Big\} \\
&\leqslant 2\sup_{f\in\mathcal{F}} \Big|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\Big| + \text{ empirical optimization error.}
\end{aligned}$$

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

6.1 Preliminaries on probabilities

- Union bound: given events indexed by $f \in \mathcal{F}$ (which can have an infinite number of elements), we have:

$$\mathbb{P}\Big(\bigcup_{f\in\mathcal{F}} A_f\Big) \leqslant \sum_{f\in\mathcal{F}} \mathbb{P}(A_f).$$

- Supremum of random variables (proof: direct application of the union bound):

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} Z_f > t\right) \leqslant \sum_{f\in\mathcal{F}} \mathbb{P}(Z_f > t)$$

- Hoeffding's inequality: if $Z_1, \dots, Z_n$ are independent random variables such that $Z_i \in [0,1]$ almost surely, then, for any $t \geqslant 0$,

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n Z_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i] \geqslant t\Big) \leqslant \exp(-2nt^2).$$

  - Corollary (by just applying to Z_i's and $1 - Z_i$'s and using the union bound):

$$\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n Z_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i]\Big| \geqslant t\Big) \leqslant 2\exp(-2nt^2).$$

  Note the differencee with the central limit theorem, which is more precise (as it involves the variance of Z_i's, but is asymptotic). Bernstein inequalities are in between.

- Proof of Hoeffding inequality:

  (a) If $Z \in [0,1]$ almost surely, then $\mathbb{E}\big[\exp(s(Z - \mathbb{E}[Z]))\big] \leqslant \exp(s^2/8)$.
  Proof: Proof: compute the first two derivative of $s \mapsto \log(\mathbb{E}\big[\exp(s(Z - \mathbb{E}[Z]))\big])$. We have

$$\begin{aligned}
\varphi'(s) &= \frac{\mathbb{E}\Big((Z - \mathbb{E}[Z])e^{s(Z-\mathbb{E}[Z])}\Big)}{\mathbb{E}\Big(e^{s(Z-\mathbb{E}[Z])}\Big)} \\
\varphi''(s) &= \frac{\mathbb{E}\Big((Z - \mathbb{E}[Z])^2 e^{s(Z-\mathbb{E}[Z])}\Big)}{\mathbb{E}\Big(e^{s(Z-\mathbb{E}[Z])}\Big)} - \Big[\frac{\mathbb{E}\Big((Z - \mathbb{E}[Z])e^{s(Z-\mathbb{E}[Z])}\Big)}{\mathbb{E}\Big(e^{s(Z-\mathbb{E}[Z])}\Big)}\Big]^2.
\end{aligned}$$

  We have that $\varphi'(0) = 0$ and $\varphi''(s)$ is the variance of some $\tilde{Z} \in [0,1]$, with distribution proportional to $e^{s(z-\mathbb{E}[Z])}d\mu(z)$ where $d\mu(z)$ is the distribution of Z.
  We have $\mathrm{var}(\tilde{Z}) = \inf_{\mu\in[0,1]} \mathbb{E}(\tilde{Z} - \mu)^2 \leqslant \mathbb{E}(\tilde{Z} - 1/2)^2 = \frac{1}{4}\mathbb{E}(2\tilde{Z} - 1)^2 \leqslant \frac{1}{4}$. Thus, by Taylor's formula, $\varphi(s) \leqslant \frac{s^2}{8}$.

  (b) By Markov inequality, for any $t \geqslant 0$, and denoting $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$:

$$\begin{aligned}
\mathbb{P}\Big(\bar{Z} - \mathbb{E}[\bar{Z}] \geqslant t\Big) &= \mathbb{P}\Big(\exp(s(\bar{Z} - \mathbb{E}[\bar{Z}])) \geqslant \exp(st)\Big) \\
&\leqslant \exp(-st)\mathbb{E}\big[\exp(s(\bar{Z} - \mathbb{E}[\bar{Z}]))\big] \\
&\leqslant \exp(-st)\prod_{i=1}^n \mathbb{E}\big[\exp(\frac{s}{n}(\bar{X}_i - \mathbb{E}[X_i]))\big] \text{ by independence} \\
&\leqslant \exp(-st)\prod_{i=1}^n \exp(\frac{s^2}{n^2}/8) = \exp(-st + \frac{s^2}{8n})
\end{aligned}$$

  which is minimized for $s = 4nt$, to get to the result.

- Expectation of the maximum: if $Z_1, \dots, Z_n$ are (potentially dependent) random variables such that $Z_i \in [0,1]$ almost surely, then

$$\mathbb{E}[\max\{Z_1 - \mathbb{E}[Z_1], \dots, Z_n - \mathbb{E}[Z_n]\}] \leqslant \frac{\sqrt{2\log n}}{2}.$$

Proof:

$$\begin{aligned}
\mathbb{E}[\max\{Z_1 - \mathbb{E}[Z_1], \dots, Z_n - \mathbb{E}[Z_n]\}] &\leqslant \frac{1}{t}\log \mathbb{E}[e^{t\max\{Z_1-\mathbb{E}[Z_1],\dots,Z_n-\mathbb{E}[Z_n]\}}] \text{ by Jensen's inequality,} \\
&= \frac{1}{t}\log \mathbb{E}[\max\{e^{tZ_1-\mathbb{E}[Z_1]}, \dots, e^{tZ_n-\mathbb{E}[Z_n]}\}] \\
&\leqslant \frac{1}{t}\log \mathbb{E}[e^{tZ_1-\mathbb{E}[Z_1]} + \cdots + e^{tZ_n-\mathbb{E}[Z_n]}] \\
&\leqslant \frac{1}{t}\log(ne^{t^2/8}) = \frac{\log n}{t} + \frac{t}{8} = \frac{\sqrt{2\log n}}{2} \text{ with } t = 2\sqrt{2\log n},
\end{aligned}$$

using the step (a) in Hoeffding inequality proof.

6.2 Finite number of models

We have, if the loss functions are bounded between 0 and 1:

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

We have, for $f \in \mathcal{F}$ fixed, $\hat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i))$, and we can apply Hoeffding's inequality, leading to

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

Thus, with probability greater than $1-\delta$,

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

- Exercise: in terms of expectation, we get (using the proof of the max of random variables above):

$$\mathbb{E}\big[\mathcal{R}(\hat{f}) - \inf_{f\in\mathcal{F}} \mathcal{R}(f)\big] \leqslant 2\mathbb{E}\big[\sup_{f\in\mathcal{F}} \Big|\hat{\mathcal{R}}(f) - \mathcal{R}(f)\Big|\big] \leqslant \sqrt{\frac{2\log(2|\mathcal{F}|)}{n}}.$$

WARNING: the order of $\mathbb{E}$ and sup matters a lot!

Thus, according to the bound, when the logarithm of the umber of models is small compared to n, learning is possible. Note that this is only an upper-bound and learning is possible with infinitely many models (which is the most classical scenario). See below.

6.3 Beyond finite number models

- Covering numbers: assume that the risks $\mathcal{R}$ and $\hat{\mathcal{R}}$ are r-Lipschitz-continuous with respect to some distance d on $\mathcal{F}$, and that 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 number $m(\varepsilon)$ is the covering number of $\mathcal{F}$ at precision ε.

- Typically, $m(\varepsilon)$ grows with ε as a power ε^{-d} when $\varepsilon \to 0+$, where d is the underlying dimension (example of cubes for the ℓ_∞ distance. For some sets (e.g, all Lipschitz-continuous functions in d dimensions) $\log m(\varepsilon)$ grows as ε^{-d}. For Sobolev functions in one-dimension, it grows as $(C/\varepsilon)^{1/\alpha}$.

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

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

- This implies that

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

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

  Alternatively, if $\log m(\varepsilon) \sim (C/\varepsilon)^{1/\alpha}$, we need to balance $\varepsilon + \sqrt{(C/\varepsilon)^{2/\alpha}/n}$, which leads to a slower power of n which is of the form $C^{\beta'}/n^{\beta''}$.

- Another very powerful tool is Rademacher complexity [2]

- Putting things together: If the approximation error goes down with C as $C^{-\beta}$ and the estimation error goes up with C as $C^{\beta'}/n^{\beta''}$, for some positive powers β, β', β'', then we have some balance. [3]

References

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

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

[3] Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.