Mathematical aspects in machine learning

Lorenz Richter
BTU Cottbus-Senftenberg

October 27, 2017

Contents

1 Statistical learning theory

We consider data from an input space $\mathcal{X}$ and an output space $\mathcal{Y}$, specifically the sample $S_n = ((X_i, Y_i))_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$. The goal is to learn a (prediction) function $f : \mathcal{X} \to \mathcal{Y}$ that maps input to output data.

Example 1.1. $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$ *(binary classification),* $\mathcal{Y} = \mathbb{N}$, $\mathcal{Y} = \mathbb{R}$ *(regression).*

In order to develop a proper theory we need to make some assumptions:

- There exists an unknown probability distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$.
- The data S_n are i.i.d. from $\mathbb{P}$, i.e. $(X_i, Y_i) \sim \mathbb{P}$ for every $i = 1, \dots, n$.
- The future data (sometimes called test data) also come from $\mathbb{P}$.

In order to measure how good we learn the prediction function f we consider a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ that indicates deviations from predictions and true values. Being stochastically more meaningful, we consider for any function f the expected loss

$$L(f) := \mathbb{E}_{(X,Y)\sim\mathbb{P}}[l(f(X), Y)],$$

which in statistics is sometimes called "(Bayes-)Risk". The goal is to come up with a learning algorithm $\mathcal{A} : S_n \mapsto \hat{f}_n$ (i.e. $\hat{f}_n := \mathcal{A}(S_n)$, indicating that the function depends on the n training data) s.t. $L(\hat{f}_n)$ is small. We therefore consider $\mathbb{E}[l(\hat{f}_n(X), Y)|S_n]$, i.e. the expectation conditioned on the samples, which is still a random quantity (as it depends on the sample data)[1].

Definition 1.2. *A predictor f^B is called Bayes-optimal if it minimizes the expected loss, i.e.*

$$L(f^B) = \inf_f L(f) =: L^B.$$

Remark 1.3. *Bayes-optimality depends on $\mathbb{P}, S_n$ and l.*

Example 1.4 (Zero noise or function learning)**.** *One could consider the case that our targets Y are deterministically prescribed by a function g, i.e. $\mathbb{P}(Y = g(X)|X = x) = 1$. However, this is a rather unrealistic case.*

[1] We will continuously omit the measure $\mathbb{P}$ in the expected value.

Example 1.5 (Binary classification)**.** *Consider $\mathcal{Y} = \{0,1\}$ and the loss $l(y',y) = \mathbb{1}\{y' \neq y\}$. One can show that the Bayes-optimal predictor (in this case classifier) is $f^B(x) = \mathbb{1}\{\eta(x) > \frac{1}{2}\}$ with $\eta(x) = \mathbb{P}(Y=1|X=x)$. We have*

$$\begin{aligned}\mathbb{P}(f(X) \neq Y|X=x) &= 1 - \mathbb{P}(f(X) = Y|X=x) \\ &= 1 - \left(\mathbb{1}_{\{f(x)=1\}}\eta(x) - \mathbb{1}_{\{f(x)=-1\}}(1-\eta(x))\right).\end{aligned}$$

This yields

$$\begin{aligned}\mathbb{P}(f(X) \neq Y|X=x) - \mathbb{P}(f^B(X) \neq Y|X=x) &= \eta(x)\left(\mathbb{1}_{\{f^B(x)=1\}} - \mathbb{1}_{\{f(x)=1\}}\right) + (1-\eta(x))\left(\mathbb{1}_{\{f^B(x)=-1\}} - \mathbb{1}_{\{f(x)=-1\}}\right) \\ &= (2\eta(x)-1)\left(\mathbb{1}_{\{f^B(x)=1\}} - \mathbb{1}_{\{f(x)=1\}}\right) \geq 0\end{aligned}$$

and by integration w.r.t. x the statement

Example 1.6 (Regression)**.** *We consider $\mathcal{Y} = \mathbb{R}$ and $l(y',y) = (y'-y)^2$. Then the Bayes-optimal predictor is $f^B(x) = \mathbb{E}[Y|X=x]$, which we can see by noting that*

$$\begin{aligned}L(f) &= \mathbb{E}\left[(f(X)-Y)^2\right] \\ &= \mathbb{E}\left[(f(X)-f^B(X)+f^B(X)-Y)^2\right] \\ &= \mathbb{E}\left[(f(X)-f^B(X))^2\right] + \mathbb{E}\left[(f^B(X)-Y)^2\right] - 2\mathbb{E}\left[(f(X)-f^B(X))(f^B(X)-Y)\right]\end{aligned}$$

is minimized by $f = f^B$ since the last term is equal to

$$\mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[(f(X)-f^B(X))(f^B(X)-Y)\right]\right] = \mathbb{E}_X\left[(f(X)-f^B(X))\mathbb{E}_{Y|X}\left[(f^B(X)-Y)\right]\right] = 0.$$

Let us come back to finding a good function f. Somehow we need to work with our sample S_n. We therefore define the empirical risk to be

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(X_i), Y_i).$$

We see that $\mathbb{E}[L_n(f)] = L(f)$ and can therefore consider empirical risk minimization (ERM), namely

$$\hat{f}_n := \arg\inf_f L_n(f)$$

as a reasonable strategy to learn the function f. However, the following example shows that this is not always a good idea.

Example 1.7. *Consider $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}, l(y',y) = \mathbb{1}\{y' \neq y\}$ and $\mathbb{P}(Y=1|x) = \frac{1}{2}$ for all $x \in \mathcal{X}$. Define the predictor as*

$$\hat{f}_n(x) = \begin{cases} y & \text{if } (x,y) \in S_n \text{ for some value } y \\ 0 & \text{otherwise} \end{cases}.$$

Note again that this predictor is not fixed, but depends on the data. We can see that it performs very badly: $L_n(\hat{f}_n) = 0$, but $L(\hat{f}_n) = \frac{1}{2}$. This is a stereotypical example of what is described as overfitting.

Theorem 1.8 (No-free-lunch)**.** *Let $\mathcal{A}$ be any learning algorithm for binary classification. Let S_n be a sample s.t. its size $n < \frac{|\mathcal{X}|}{2}$. Then there exists a probability distribution $\mathbb{P}$ s.t.*

(i) there exists a function $f : \mathcal{X} \to \mathcal{Y}$ with $L(f) = 0$,

(ii) with probability $\geq \frac{1}{7}$ we have $L(\mathcal{A}(S_n)) > \frac{1}{8}$.

Proof. See chapter 5 in [5][2]. □

Remark 1.9. *This means that there is no prefect algorithm that works for all distributions. Bayes optimality cannot be obtained independent of $\mathbb{P}$.*

[2]The proof is similar to the concept of VC dimension, which we will discuss later.

In order to address the problem of overfitting[3], we consider the function class $\mathcal{F}$ (usually a class chosen with prior knowledge regarding the prediction problem) and restrict our predictors to come from this class, i.e. $\hat{f}_n \in \mathcal{F}$. We can then decompose our expected error as

$$L(\hat{f}_n) - L(f^B) = \underbrace{L(\hat{f}_n) - \inf_{f\in\mathcal{F}} L(f)}_{\text{estimation error (variance)}} + \underbrace{\inf_{f\in\mathcal{F}} L(f) - L(f^B)}_{\text{approximation error (bias)}} .$$

Here, $\hat{f}_n$ shall be the predictor that minimizes the empirical error in our hypothesis class, i.e. $L_n(\hat{f}_n) = \inf_{f\in\mathcal{F}} L_n(f)$. In the sequel, we will only study the estimation error, i.e. we aim at statements of the form

$$\mathbb{P}\{L(\hat{f}_n) - L(f^*) > \epsilon\} \leq B(\epsilon, n, \mathcal{F}),$$

where B decreases to 0 as $n \to \infty$ for any ϵ and $f^* = \arg\inf_{f\in\mathcal{F}} L(f)$, i.e. the best predictor in the class $\mathcal{F}$ (we will not study the approximation error here, details can for instance be found in [3]). Equivalently, we will consider the following kind of deviation inequality: With probability $1-\delta$ for $\delta \in (0,1)$ we want

$$L(\hat{f}_n) - L(f^*) \leq D(\delta, n, \mathcal{F}).$$

We will particularly be interested in statements for all $f \in \mathcal{F}$, i.e. we want uniform and not only pointwise bounds. Note that we have $L(\hat{f}_n) - L_n(\hat{f}_n) \leq \sup_{f\in\mathcal{F}}\{L(f) - L_n(f)\}$, so we will then also investigate the difference between expected and empirical loss, $L(\hat{f}_n) - L_n(\hat{f}_n)$, which we call generalization error[4].

More abstractly, we consider the following learnability definition.

Definition 1.10 (PAC learnability [6])**.** *A hypothesis class $\mathcal{F}$ is "probably approximately correctly" (PAC) learnable if there is an algorithm $\mathcal{A}$ such that for all $\epsilon > 0$ and $\delta > 0$ there is a sample size $n(\epsilon, \delta) \in \mathbb{N}$ such that for all distributions $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$*

$$\mathbb{P}(L(\mathcal{A}(S_n)) - L(f^*) < \epsilon) \geq 1 - \delta.$$

Remark 1.11. *One interesting aspect in introducing hypothesis classes $\mathcal{F}$ is model selection. Consider a family of hypothesis classes $\{\mathcal{F}_\alpha\}_{\alpha\in A}$ with $\mathcal{F} = \cup_\alpha \mathcal{F}_\alpha$ (e.g. $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$). Let $\hat{f}_n^\alpha$ be the predictor corresponding to the ERM over $\mathcal{F}_\alpha$. One considers $\alpha_n := \alpha(S_n)$, i.e. a data-dependent selection of the possible classes. We are then interested in $L(\hat{f}_n^{\alpha_n}) - \inf_{\alpha\in A}\inf_{f\in\mathcal{F}_\alpha} L(f)$.*

Remark 1.12. *If we consider only one possible function, i.e. $\mathcal{F} = \{f_1\}$, then nothing will be learnt and we know from the law of large numbers that*

$$L_n(\hat{f}_n) - L(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^n l(f_1(X_i), Y_i) - \mathbb{E}[l(f_1(X), Y)] \to 0$$

for $n \to \infty$ at rate $\frac{1}{\sqrt{n}}$.

So in the rather pathological case of $\mathcal{F} = \{f_1\}$ we have convergence of $L(\hat{f}_n) - L(f^*)$. But does it also hold with N predictors at hand, and if yes, how fast does it happen? To answer these questions the following so-called concentration inequalities will be of help.

1.1 Concentration inequalities

Concentration inequalities are statements of the form[5]

$$\mathbb{P}(|\bar{Z}_n - \mathbb{E}[Z]| < \epsilon) \geq 1 - \delta.$$

Theorem 1.13 (Markov inequality)**.** *Consider a random variable Z with $Z \geq 0$. Then for all $\epsilon > 0$ we have*

$$\mathbb{P}(Z > \epsilon) \leq \frac{\mathbb{E}[Z]}{\epsilon}.$$

[3]Another way to address overfitting is to modify the criterion to be minimized by for instance adding a penalty term for too "complicated" functions – see structural risk minimization and (normalized) regularization.

[4]Compare to the uniform law of large numbers.

[5]Sometimes the notation $Pf = \mathbb{E}[f(X,Y)]$ and $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i, Y_i)$ for the empirical measure is used. Compare this to the theory of empirical processes, where uniform deviations of averages from their expectations are investigated, i.e. $\sup_{f\in\mathcal{F}}\{Pf - P_n f\}$.

Proof. See [4]. □

Theorem 1.14 (Chebysheff inequality)**.** *Consider a random variable Z. Then for all $\epsilon > 0$ we have*

$$\mathbb{P}(|Z - \mathbb{E}[Z]| > \epsilon) \leq \frac{\mathrm{Var}(Z)}{\epsilon^2}.$$

Proof. See [4]. □

We motivate the next bound with

Example 1.15. *Consider $X \sim \mathcal{N}(\mu, \sigma^2)$, then we have*

$$\mathbb{P}(|X - \mu| > \epsilon) = \int_{\epsilon}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \mathrm{d}x \leq \int_{\epsilon}^{\infty} \frac{x}{\epsilon\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \mathrm{d}x$$

and therefore $\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma}{\epsilon}\sqrt{\frac{2}{\pi}} e^{-\frac{\epsilon^2}{2\sigma^2}}$. We will see that this is a special case of the following theorem, when replacing σ^2 with the maximal possible variance of a random variable in $[0, 1]$ (i.e. $\frac{1}{4}$). It is also intuitive to look at $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, i.e. $\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) = \mathbb{P}(|Z| > \sqrt{n}\epsilon) \leq \frac{\sigma}{\sqrt{n}\epsilon}\sqrt{\frac{2}{\pi}} e^{-\frac{n\epsilon^2}{2\sigma^2}}$, which decreases exponentially for large n.

Lemma 1.16 (Hoeffding's lemma)**.** *For a random variable $Z \in [a, b]$ with $\mathbb{E}[Z] = 0$ we can bound the moment generating function for all $\lambda \in \mathbb{R}$ as*

$$\mathbb{E}[e^{\lambda Z}] \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right).$$

Proof. Since $e^{\lambda Z}$ is convex we have for all $a \leq Z \leq b$

$$e^{\lambda Z} \leq \frac{b - Z}{b - a} e^{\lambda a} + \frac{Z - a}{b - a} e^{\lambda b}$$

and therefore

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda Z}\right] &\leq \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b} \\
&= \left(-\frac{a}{b-a}\right) e^{\lambda a} \left(-\frac{b}{a} + e^{\lambda b - \lambda a}\right) \\
&= \left(-\frac{a}{b-a}\right) e^{\lambda a} \left(-\frac{b-a}{a} - 1 + e^{\lambda(b-a)}\right) \\
&= \left(1 - c + c e^{d}\right) e^{-cd} \\
&=: e^{L(d)}
\end{aligned}$$

with $c = -\frac{a}{b-a} > 0$, $d = \lambda(b - a)$ and

$$L(d) = -cd + \log(1 - c + ce^d).$$

Taking derivatives w.r.t. d brings

$$L(0) = L'(0) = 0$$

and

$$L''(d) = \frac{ce^d\left(1 - c + ce^d\right) - c^2 e^{2d}}{\left(1 - c + ce^d\right)^2} = \frac{ce^d}{1 - c + ce^d}\left(1 - \frac{ce^d}{1 - c + ce^d}\right) = t(1 - t) \leq \frac{1}{4}$$

with $t = \frac{ce^d}{1 - c + ce^d} > 0$. By Taylor's theorem there exists an $e \in [0, d]$ s.t.

$$L(d) = L(0) + dL'(0) + \frac{1}{2} d^2 L''(e) \leq \frac{1}{8} d^2 = \frac{\lambda^2}{8}(b - a)^2.$$

This implies the bound (by convexity of the logarithm). □

Theorem 1.17 (Hoeffding's inequality)**.** *Let $Z_1, \dots, Z_n$ be independent (but not necessarily identically distributed). Assume w.l.o.g. that $Z_i \in [0,1]$ for $i = 1, \dots, n$. Then with $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$*

$$\mathbb{P}\left(\left|\bar{Z}_n - \mathbb{E}\left[\bar{Z}_n\right]\right| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Proof. Consider $Z_i \in [a_i, b_i]$. With Markov's inequality we have for $\lambda > 0$

$$\begin{aligned}
\mathbb{P}\left(\bar{Z}_n - \mathbb{E}\left[\bar{Z}_n\right] > \epsilon\right) &\leq \frac{\mathbb{E}\left[e^{\lambda(\bar{Z}_n - \mathbb{E}[\bar{Z}_n])}\right]}{e^{\lambda\epsilon}} \\
&= \frac{\mathbb{E}\left[e^{\frac{\lambda}{n}\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i])}\right]}{e^{\lambda\epsilon}} \\
&= \frac{\prod_{i=1}^n \mathbb{E}\left[e^{\frac{\lambda}{n}(Z_i - \mathbb{E}[Z_i])}\right]}{e^{\lambda\epsilon}} \\
&\leq \frac{\prod_{i=1}^n e^{\frac{\lambda^2}{8n^2}(b_i - a_i)^2}}{e^{\lambda\epsilon}} \\
&= \exp\left(\frac{\sum_{i=1}^n (b_i - a_i)^2}{8n^2}\lambda^2 - \lambda\epsilon\right),
\end{aligned}$$

where in the last inequality we used Hoeffding's lemma. This general proof technique is called Chernoff bound. We now minimize w.r.t λ and get $\lambda^* = \frac{4n^2\epsilon}{\sum_{i=1}^n (b_i - a_i)^2}$ and therefore

$$\mathbb{P}\left(\bar{Z}_n - \mathbb{E}\left[\bar{Z}_n\right] > \epsilon\right) \leq \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

which yields the desired result with $a_i = 0$, $b_i = 1$ and by noting that $\mathbb{P}(|Z|) \leq \mathbb{P}(Z \leq t) + \mathbb{P}(-Z \leq t)$. □

Remark 1.18. *Note that this theorem also holds for functions of random variables. One can for instance consider $Z = g(X, Y)$ with $g(X, Y) = l(f(X), Y)$. However, also note that it only holds for a fixed function and not uniformly for all $f \in \mathcal{F}$. For a fixed sample one finds an $f \in \mathcal{F}$ that yields a very large error, as for instance seen in example 1.7.*

Remark 1.19. *We can formulate Hoeffding's bound as a deviation inequality by realizing the equivalence of the statements $\mathbb{P}(|Z| > \epsilon) \leq \delta$ and $\mathbb{P}(|Z| < \epsilon) \geq 1 - \delta$. Set $\delta := 2e^{-2n\epsilon^2}$, then $\epsilon = \sqrt{\frac{\log\frac{2}{\delta}}{2n}}$, and we have with probability at least $1 - \delta$ that*

$$\left|\bar{Z}_n - \mathbb{E}\left[\bar{Z}_n\right]\right| \leq \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Hoeffding's inequality does not use any knowledge about the distribution of variables. Therefore one can consider Bernstein's inequality, which uses the variance of the distribution to get a tighter bound.

Theorem 1.20 (Bernstein's inequality)**.** *Let $Z_1, \dots, Z_n$ be independent (but not necessarily identically distributed). Assume w.l.o.g. $|Z_i| < 1$. Then with probability $1 - \delta$ we have*

$$\left|\bar{Z}_n - \mathbb{E}\left[\bar{Z}_n\right]\right| \leq \sqrt{2\,\mathrm{Var}\left(\bar{Z}_n\right)\log\frac{2}{\delta}} + \frac{2\log\frac{2}{\delta}}{3n}.$$

Proof. See [2]. □

Remark 1.21. *If the variance is very small then the first term on the right hand side becomes negligible.*

1.2 Error bounds in the finite dimensional case

Let us first consider the finite-dimensional hypothesis space $\mathcal{F} = \{f_1, \dots, f_N\}$.

Theorem 1.22 (Realizable case)**.** *Consider $\mathcal{Y} = \{0, 1\}$ in the zero-noise case, i.e. $L(f^B) = 0$ with $f^B \in \mathcal{F}$ (as in example 1.4). As before we call such an f^B, since it is in $\mathcal{F}$, just f^*. Then with probability $\geq 1 - \delta$ we have*

$$L(\hat{f}_n) - L(f^*) = L(\hat{f}_n) \leq \frac{\log N + \log\left(\frac{1}{\delta}\right)}{n}.$$

Proof. Let us try to bound $\mathbb{P}(\text{ERM selects } f \text{ with } L(f) > \epsilon)$. By definition we know that $L_n(\hat{f}_n) = 0$, since ERM selects only these predictors f for which $L_n(f) = 0$. Now take any $f \in \mathcal{F}$ s.t. $L(f) = \mathbb{P}(f(X) \neq Y) > \epsilon$, i.e. consider the functions that are expected to make some error. That implies that $\mathbb{P}(L(f) = 0) < 1 - \epsilon$ and therefore for n samples we get

$$\mathbb{P}(L_n(f) = 0) < (1 - \epsilon)^n \leq e^{-n\epsilon},$$

where the last step follows from $1 + x \leq e^x$. Now we consider

$$\begin{aligned}
\mathbb{P}(\text{ERM selects } \hat{f}_n \text{ with } L(\hat{f}_n) > \epsilon) &\leq \mathbb{P}\left(\exists f \in \mathcal{F} \text{ s.t. } L(f) > \epsilon, L_n(f) = 0\right) \\
&= \mathbb{P}\left(\cup_{\substack{f \in \mathcal{F} \\ L(f) > \epsilon}} L_n(f) = 0\right) \\
&\leq \sum_{\substack{f \in \mathcal{F} \\ L(f) > \epsilon}} \mathbb{P}(L_n(f) = 0) \\
&\leq \sum_{\substack{f \in \mathcal{F} \\ L(f) > \epsilon}} e^{-n\epsilon} \\
&\leq N e^{-n\epsilon} =: \delta.
\end{aligned}$$

This gives $\epsilon = \frac{\log N + \log\left(\frac{1}{\delta}\right)}{n}$. So with probability at least $1 - \delta$ we have the complement event, i.e.

$$\mathbb{P}(\text{ERM selects } f \text{ with } L(f) \leq \epsilon) \geq 1 - \delta.$$

The event $\{\text{ERM selects } f \text{ with } L(f) \leq \epsilon\}$ is just $\{L(\hat{f}_n) \leq \epsilon\}$ and therefore we get the desired bound. □

Remark 1.23. *One can show that given our assumptions this bound is actually the best possible one.*

Let us now drop the assumption $L(f^*) = 0$.

Theorem 1.24 (Agnostic case)**.** *We consider again $\mathcal{F} = \{f_1, \ldots, f_N\}$. Let l be bounded. Then with probability $\geq 1 - \delta$ we have*

$$L(\hat{f}_n) - L(f^*) \leq \sqrt{2\frac{\log(2N) + \log\left(\frac{1}{\delta}\right)}{n}}.$$

Proof.

$$\begin{aligned}
L(\hat{f}_n) - L(f^*) &= L(\hat{f}_n) - L_n(\hat{f}_n) + \underbrace{L_n(\hat{f}_n) - L_n(f^*)}_{\leq 0} + L_n(f^*) - L(f^*) \\
&\leq |L(\hat{f}_n) - L_n(\hat{f}_n)| + |L_n(f^*) - L(f^*)| \\
&\leq 2 \sup_{f \in \mathcal{F}} |L(f) - L_n(f)|.
\end{aligned}$$

Since our $\mathcal{F}$ is finite dimensional, we can actually consider the maximum and compute

$$\begin{aligned}
\mathbb{P}\left(\max_{i=1,\ldots,N} |L(f_i) - L_n(f_i)| > \epsilon\right) &= \mathbb{P}\left(\cup_{i=1}^N \{|L(f_i) - L_n(f_i)| > \epsilon\}\right) \\
&\leq \sum_{i=1}^N \mathbb{P}(|L(f_i) - L_n(f_i)| > \epsilon) \\
&\leq 2N e^{-2n\epsilon^2} =: \delta,
\end{aligned}$$

where the last step follows from Hoeffding's inequality, which also holds when taking functions of the random variables. If we solve for ϵ we get $\epsilon = \sqrt{\frac{\log \frac{2N}{\delta}}{2n}}$. Considering again the complement event (as in the proof of theorem 1.22) yields our desired statement. □

We have now seen the two extreme cases: The rather unrealistic realizible setting ($L(f^*) = 0$) and the totally agnostic setting ($L(f^*) = \frac{1}{2}$). The first one has convergence rate $\mathcal{O}(\frac{1}{n})$, the latter one $\mathcal{O}(\frac{1}{\sqrt{n}})$. Let us now look at cases in between those extremes ("from slow to fast rates").

Theorem 1.25. *We again consider $\mathcal{F} = \{f_1, \dots, f_N\}$ and our loss being bounded. With probability $\geq 1 - \delta$ we have*

$$L(\hat{f}_n) - L(f^*) \leq C \left(\sqrt{L(f^*) \frac{\log 4N + \log\left(\frac{1}{\delta}\right)}{n}} + \frac{\log\left(\frac{4N}{\delta}\right)}{n} \right),$$

where C is a constant.

Proof. First note that

$$\mathrm{Var}(\mathbb{1}_{\{f(X) \neq Y\}}) = \mathbb{E}[\mathbb{1}^2] - \mathbb{E}[\mathbb{1}]^2 \leq \mathbb{E}[\mathbb{1}] = L(f).$$

Now take Bernstein's inequality with $\delta \to \frac{\delta}{N}$:

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \underbrace{|L_n(f) - L(f)| \geq \sqrt{2L(f) \frac{\log \frac{2N}{\delta}}{n}} + \frac{2 \log \frac{2N}{\delta}}{3n}}_{=:E(f)} \right) = \mathcal{P}(\cup_{f \in \mathcal{F}} E(f)) \leq \sum_{f \in \mathcal{F}} \mathbb{P}(E) \leq \sum_{f \in \mathcal{F}} \frac{\delta}{N} = \delta.$$

Forming the complement yields

$$\mathbb{P}\left(\forall f \in \mathcal{F} : |L_n(f) - L(f)| \leq \sqrt{2L(f) \frac{\log \frac{2N}{\delta}}{n}} + \frac{2 \log \frac{2N}{\delta}}{3n} \right) \geq 1 - \delta.$$

We can therefore pick $f = \hat{f}_n$ and get that with probability at least $1 - \delta$

$$|L_n(\hat{f}_n) - L(\hat{f}_n)| \leq \sqrt{2L(\hat{f}_n) \frac{\log \frac{2N}{\delta}}{n}} + \frac{2 \log \frac{2N}{\delta}}{3n}.$$

Using another time Bernstein's inequality for f^* yields

$$L_n(\hat{f}_n) \leq L_n(f^*) \leq L(f^*) + \sqrt{2L(f^*) \frac{\log \frac{2}{\delta}}{n}} + \frac{2 \log \frac{2}{\delta}}{3n}$$

We can now combine the preceding two equations to get

$$\frac{L(\hat{f}_n) - L(f^*)}{\sqrt{L(\hat{f}_n)}} \leq \sqrt{2 \frac{\log \frac{4N}{\delta}}{n}}$$

and use the fact $\frac{A-B}{\sqrt{A}} \leq C \Rightarrow A \leq B + C^2 + \sqrt{B}C$ for $A, B, C \geq 0$ to get the result.

□

Remark 1.26. *This bound is again tight. Note that if we assume $n < \frac{1}{L(f^*)} =: \gamma$ we are in a fast-rates-regime, i.e. $\mathcal{O}(\frac{1}{n})$.*

Remark 1.27. *Note that we often do not use ERM in practice, as the minimization can be NP-hard.*

2 Bibliography

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

[2] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

[3] Cucker, F. and Zhou, D. X. (2007). *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press.

[4] Hartmann, C. (2017). *Wahrscheinlichkeitstheorie*. Lecture script, BTU Cottbus-Senftenberg.

[5] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[6] Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.