Mathematical aspects in machine learning

Lorenz Richter
BTU Cottbus-Senftenberg

November 21, 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 $\mathbb{P}, \mathcal{X} \times \mathcal{Y}$ 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 and let $n < \frac{|\mathcal{X}|}{2}$ be a sample size. 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 $\geq \frac{1}{7}$ over the sample $S_n \sim \mathbb{P}^n$ we have $L(\mathcal{A}(S_n)) > \frac{1}{8}$.

Proof. See chapter 5 in [9][2]. □

Remark 1.9. *This means that there is no perfect 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 [4]). 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 [10])**.** *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 [6]. □

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 [6]. □

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 $\mathbb{P}(|Z| \leq \epsilon) \leq \mathbb{P}(Z \leq \epsilon) + \mathbb{P}(-Z \leq \epsilon)$. □

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.*

Theorem 1.22 (McDiarmid's inequality)**.** *Consider a function $g : \mathcal{Z}^n \to \mathbb{R}$. Assume that there exist constants $c_1, \dots, c_n$ s.t.*

$$\sup_{Z_1, \dots, Z_n, Z'} |g(Z_1, \dots, Z_n) - g(Z_1, \dots, Z_{i-1}, Z', Z_{i+1}, \dots, Z_n)| \leq c_i$$

for all $i = 1, \dots, n$. Let $Z_1, \dots, Z_n \sim \mathbb{P}$ i.i.d. with values in $\mathcal{Z}$ and let $g := g(Z_1, \dots, Z_n)$. Then with probability $\geq 1 - \delta$ we have

$$|g - \mathbb{E}[g]| \leq \sqrt{\frac{1}{2}\left(\sum_{i=1}^n c_i^2\right)\log\frac{2}{\delta}}.$$

Proof. First note that

$$\mathbb{P}\left(|g - \mathbb{E}[g]| \geq \epsilon\right) = \mathbb{P}\left(g - \mathbb{E}[g] \geq \epsilon\right) + \mathbb{P}\left(g - \mathbb{E}[g]\right] \leq -\epsilon\right),$$

from which we will show the first inequality – the second one works analogously. Define for $i = 2, \ldots, n$ the martingale difference sequence

$$V_i := \mathbb{E}[g|Z_1, \ldots, Z_i] - \mathbb{E}[g|Z_1, \ldots, Z_{i-1}] \qquad \text{and} \qquad V_1 := \mathbb{E}[g|Z_1] - \mathbb{E}[g],$$

then we have

$$g(Z_1, \ldots, Z_n) - \mathbb{E}[g(Z_1, \ldots, Z_n)] = \sum_{i=1}^{n} V_i.$$

With a similar argument as in Lemma 1.16 (see [2], section 6.1) we get

$$\mathbb{E}\left[e^{\lambda V_i}|Z_1, \ldots, Z_{i-1}\right] \leq \exp\left(\frac{\lambda^2 c_i^2}{8}\right).$$

Then we have for all $\lambda > 0$

$$\begin{aligned}
\mathbb{P}\left(g(Z_1, \ldots, Z_n) - \mathbb{E}[g(Z_1, \ldots, Z_n)] \geq \epsilon\right) &= \mathbb{P}\left(\sum_{i=1}^{n} V_i \geq \epsilon\right) \\
&= \mathbb{P}\left(e^{\lambda \sum_{i=1}^{n} V_i} \geq e^{\lambda\epsilon}\right) \\
&\leq e^{-\lambda\epsilon}\mathbb{E}\left[e^{\lambda \sum_{i=1}^{n} V_i}\right] \\
&= e^{-\lambda\epsilon}\mathbb{E}\left[e^{\lambda \sum_{i=1}^{n-1} V_i}\mathbb{E}\left[e^{\lambda V_n}|Z_1, \ldots, Z_{n-1}\right]\right] \\
&\leq e^{-\lambda\epsilon}e^{\frac{\lambda^2 c_n^2}{8}}\mathbb{E}\left[e^{\lambda \sum_{i=1}^{n-1} V_i}\right] \\
&\leq e^{-\lambda\epsilon}e^{\frac{\lambda^2}{8}\sum_{i=1}^{n} c_i^2}.
\end{aligned}$$

Taking $\lambda = \frac{4\epsilon}{\sum_{i=1}^{n} c_i^2}$ brings the statement. □

Remark 1.23. *Note that with $g(Z_1, \ldots, Z_n) = \frac{1}{n}\sum_{i=1}^{n} Z_i$ we rediscover Hoeffdings inequality.*

1.2 Error bounds for finite classes

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

Theorem 1.24 (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 $\mathbb{P}$(ERM selects f 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 {ERM selects f with $L(f) \leq \epsilon$} is just $\{L(\hat{f}_n) \leq \epsilon\}$ and therefore we get the desired bound. □

Remark 1.25. *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 1.26 (Agnostic case)**.** *We consider again $\mathcal{F} = \{f_1, \dots, f_N\}$. 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, we can actually consider the maximum and compute

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

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 1.24) 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 1.27. *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 1.28. *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 1.29. *Note that we often do not use ERM in practice, as the minimization can be NP-hard.*

1.3 Error bounds for infinite classes

We now want to consider function classes $\mathcal{F}$ that have infinitely many elements. The former bounds will then be meaningless (as we have $N = \infty$). When looking into the proofs, we realize that our union bounds as in theorem 1.26 will yield an infinite sum that has to be bounded in order to be usefull. If the function class is even uncountable we need a completely different approach. For that, we make the following definition.

Definition 1.30 (Rademacher complexity)**.** *Consider a function class $\mathcal{F}$ and i.i.d. random variables $Z_1, \ldots, Z_n \sim \mathbb{P}$. Let $\sigma = \{\sigma_1, \ldots, \sigma_n\}$ be a set of i.i.d. random variables s.t. $\mathbb{P}(\sigma_i = \pm 1) = \frac{1}{2}$. We then define the empirical Rademacher complexity[6] as*

$$\hat{R}_n(\mathcal{F}, S_n) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right| \middle| S_n\right].$$

For the Rademacher complexity we average over all samples $S_n = (Z_1, \ldots, Z_n)$, i.e.

$$R_n(\mathcal{F}) = \mathbb{E}_{S_n}\left[\hat{R}_n(\mathcal{F})\right].$$

Remark 1.31. *The supremum can be interpreted as finding the function f out of $\mathcal{F}$ that makes the samples correlate the most with some random values σ. Loosely speaking, the idea is to find a function that looks like random noise the most. The Rademacher complexity then gives a measure of how well we can find a function out of the function class $\mathcal{F}$ that behaves like random noise. Note that $\hat{R}_n(\mathcal{F}) = 0$ if $\mathcal{F}$ consists of only one function and $\hat{R}_n(\mathcal{F}) = 1$ if $\mathcal{F}$ consists of all functions. In fact, we have $\hat{R}_n(\mathcal{F}) \in [0,1]$. Note that in general computing Rademacher complexities can be difficult (actually as difficult as computing empirical risk minimizers).*

Remark 1.32. *For $\mathcal{F} \subset \mathcal{G}$ we have $R_n(\mathcal{F}) \leq R_n(\mathcal{G})$.*

[6] In the context of empirical processes $R_n(f) = \frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)$ is sometimes called Rademacher process.

Theorem 1.33. *Consider any prediction task $(\mathbb{P}, \mathcal{X}, \mathcal{Y}, l, \mathcal{F})$ s.t. $l(f(X), Y) \in [0,1]$. Then with probability $\geq 1-\delta$*

$$\sup_{f\in\mathcal{F}} |L(f) - L_n(f)| \leq 2R_n(l\circ\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}},$$

where $l\circ\mathcal{F} = \{l(f(X),Y) : f\in\mathcal{F}, (X,Y)\in\mathcal{X}\times\mathcal{Y}\}$.

Proof. Comparing to theorem 1.22, let $g(Z_1,\dots,Z_n) = \sup_{f\in\mathcal{F}} |L_n(f) - L(f)|$. If we change one datum Z_i to Z_i' we get

$$\begin{aligned}\left|\sup_{f\in\mathcal{F}} |L_n(f) - L(f)| - \sup_{f\in\mathcal{F}} \left|L_n^i(f) - L(f)\right|\right| &\leq \left|\frac{1}{n}\sup_{f\in\mathcal{F}} |l(f(X_i),Y_i) - l(f(X_i'),Y_i')|\right| \\ &\leq \frac{1}{n},\end{aligned}$$

where L_n^i is the empirical risk with the changed datum. We can therefore use McDiarmid's inequality

$$\sup_{f\in\mathcal{F}} |L_n(f) - L(f)| \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}} |L_n(f) - L(f)|\right] + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

The idea now is to introduce a ghost sample S_n' that is identically distributed as S_n. We then have

$$\begin{aligned}\mathbb{E}_{S_n}\left[\sup_{f\in\mathcal{F}} |L_n(f) - L(f)|\right] &= \mathbb{E}_{S_n}\left[\sup_{f\in\mathcal{F}} \left|\mathbb{E}_{S_n'}\left[L_n(f) - L_n'(f)\right]\right|\right] \\ &\leq \mathbb{E}_{S_n,S_n'}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \left(l(f(X_i),Y_i) - l(f(X_i'),Y_i')\right)\right|\right] \\ &= \mathbb{E}_{S_n,S_n',\sigma}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i\left(l(f(X_i),Y_i) - l(f(X_i'),Y_i')\right)\right|\right] \\ &\leq \mathbb{E}_{S_n,\sigma}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i l(f(X_i),Y_i)\right|\right] + \mathbb{E}_{S_n',\sigma}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i l(f(X_i'),Y_i')\right|\right] \\ &= 2R_n(l\circ\mathcal{F}),\end{aligned}$$

where we used Jensen's inequality in the second step. □

Remark 1.34. *Note by looking at the proof of theorem 1.26 that we actually have*

$$L(\hat{f}_n) - L(f^*) \leq \sup_{f\in\mathcal{F}}(L(f) - L_n(f)) + \sup_{f\in\mathcal{F}}(L_n(f) - L(f))$$

and can therefore bound the estimation error.

Let us now investigate how to control the Rademacher complexity. First, we rediscover the finite class case.

Lemma 1.35. *Assume $\mathcal{F} = \{f_1,\dots,f_N\}$. Then*

$$\hat{R}_n(\mathcal{F}) \leq \frac{\sqrt{2\log N}}{n}\max_{j=1,\dots,N}\sqrt{\sum_{i=1}^n f_j^2(Z_i)}$$

and therefore

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

Proof. We have for any $\lambda > 0$

$$\begin{aligned}
\exp\left(\lambda \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f_j(Z_i)\right]\right) &\leq \mathbb{E}\left[\exp\left(\lambda \max_{j=1,\dots,N} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f_j(Z_i)\right)\right] \\
&= \mathbb{E}\left[\max_{j=1,\dots,N} \exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\sigma_i f_j(Z_i)\right)\right] \\
&\leq \sum_{j=1}^{N}\prod_{i=1}^{n}\mathbb{E}\left[\exp\left(\frac{\lambda}{n}\sigma_i f_j(Z_i)\right)\right] \\
&\leq \sum_{j=1}^{N}\prod_{i=1}^{n}\exp\left(\frac{\lambda^2}{2n^2} f_j^2(Z_i)\right) \\
&= \sum_{j=1}^{N}\exp\left(\frac{\lambda^2}{2n^2}\sum_{i=1}^{n} f_j^2(Z_i)\right) \\
&\leq N \max_{j=1,\dots,N}\exp\left(\frac{\lambda^2}{2n^2}\sum_{i=1}^{n} f_j^2(Z_i)\right),
\end{aligned}$$

where we used Hoeffding's lemma. We get

$$\hat{R}_n(\mathcal{F}) \leq \frac{1}{\lambda}\left(\log N + \max_{j=1,\dots,N}\frac{\lambda^2}{2n^2}\sum_{i=1}^{n} f_j^2(Z_i)\right).$$

Minimizing w.r.t. λ brings $\lambda^* = \sqrt{\frac{\log N 2n^2}{\max_{j=1,\dots,N}\sum_{i=1}^{n} f_j^2(Z_i)}}$ and therefore the result. □

Remark 1.36. *Sometimes the Rademacher complexity is defined without the absolute values. Note, however, that we have* $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right|\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}\cup-\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right]$.

Let us now consider the infinite class case, i.e. a general $\mathcal{F}$ for binary classification with $l(y', y) = \mathbb{1}\{y' \neq y\}$. From theorem 1.33 we know that

$$L(\hat{f}_n) - L(f^*) \leq 4R_n(l\circ\mathcal{F}) + \sqrt{\frac{2\log\frac{1}{\delta}}{n}}.$$

Still, in R_n we take the supremum over all $f \in \mathcal{F}$, thus we need to find a strategy how to deal with that in the infinite class case. The idea is to argue that we can divide $\mathcal{F}$ into finitely many equivalence classes. Note that in the computation of R_n we consider the sum over n summands, where each of them has the value 0 or 1. We can say that each function f induces an n-dimensional vector with the losses of the sample,

$$f \mapsto (l(f(X_1), Y_1), \dots, l(f(X_n), Y_n)) \in \{0,1\}^n,$$

and we realize that we can have at most 2^n such vectors, i.e. no matter how big the function class is, we have to consider at most 2^n cases. Therefore, we can divide $\mathcal{F}$ into 2^n equivalence classes and identify functions that lead to the same loss vector with one another. Unfortunately, however, this is not enough, as plugging in 2^n instead of N in lemma 1.35 just yields a constant bound that is not decreasing with n. The next idea is to consider only function classes $\mathcal{F}$ that yield (much) less than 2^n possible permutations. Let us therefore define the following set of vectors, the projections of the predictors on the sample,

$$\mathcal{F}_{S_n} := \{(l(f(X_1), Y_1), \dots, l(f(X_n), Y_n)), f \in \mathcal{F}\} \subset \{0,1\}^n,$$

and ask what the cardinatlity of the vector over all possible data sets S_n is. For this we define the so called growth function (or shattering number)

$$S_{\mathcal{F}}(n) = \sup_{S_n}|\mathcal{F}_{S_n}| \leq 2^n,$$

which is the worst case cardinality of our equivalence classes. The hope is that the growth function will be indeed much smaller than 2^n. Note that $S_{\mathcal{F}}(n)$ does not depend on the sample distribution $\mathbb{P}$, but only on $\mathcal{F}$.

Now we introduce a crucial trick.

Lemma 1.37 (Symmetrization). *For all* $\epsilon \geq \sqrt{\frac{2}{n}}$

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}|L_n(f)-L(f)|>\epsilon\right)\leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}|L_n(f)-L'_n(f)|>\frac{\epsilon}{2}\right),$$

where L'_n *is the empirical error of a ghost sample.*

Proof. We again consider a ghost sample $S'_n = (Z'_1, \ldots, Z'_n)$ that is distributed as the original sample S_n and denote its empirical error as L'_n. Let $f_n \in \mathcal{F}$ maximize $|L_n(f) - L(f)|$. Note that if $|L_n(f_n) - L(f_n)| > \epsilon$ and $|L(f_n) - L'_n(f_n)| \leq \frac{\epsilon}{2}$, then

$$\epsilon < |L_n(f_n)-L(f_n)| = |L_n(f_n)-L'_n(f_n)+L'_n(f_n)-L(f_n)| \leq |L_n(f_n)-L'_n(f_n)|+|L'_n(f_n)-L(f_n)| \leq \frac{\epsilon}{2}+|L'_n(f_n)-L(f_n)|$$

and therefore $|L'_n(f_n) - L(f_n)| > \frac{\epsilon}{2}$. This then yields

$$\begin{aligned}\mathbb{1}\Big\{|L_n(f_n)-L(f_n)|>\epsilon\Big\}\mathbb{1}\left\{|L(f_n)-L'(f_n)|\leq\frac{\epsilon}{2}\right\} &= \mathbb{1}\left\{|L_n(f_n)-L(f_n)|>\epsilon, |L(f_n)-L'(f_n)|\leq\frac{\epsilon}{2}\right\}\\ &\leq \mathbb{1}\left\{|L(f_n)-L'_n(f_n)|>\frac{\epsilon}{2}\right\}.\end{aligned}$$

Taking the expectation w.r.t. S'_n brings

$$\mathbb{1}\Big\{|L_n(f_n)-L(f_n)|>\epsilon\Big\}\mathbb{P}'\left(|L(f_n)-L'(f_n)|\leq\frac{\epsilon}{2}\right)\leq\mathbb{P}'\left(|L_n(f_n)-L'_n(f_n)|>\frac{\epsilon}{2}\right).$$

Note that by Chebyshev's inequality and with $X \in [0,1] \Rightarrow \mathrm{Var}(X) \leq \frac{1}{4}$ we get

$$\mathbb{P}'\left(|L(f_n)-L'(f_n)|\leq\frac{\epsilon}{2}\right)\leq\frac{4\,\mathrm{Var}'(f_n)}{n\epsilon^2}\leq\frac{1}{n\epsilon^2}\leq\frac{1}{2},$$

which then yields

$$\mathbb{1}\Big\{\sup_{f\in\mathcal{F}}|L_n(f)-L(f)|>\epsilon\Big\}\leq 2\mathbb{P}'\left(\sup_{f\in\mathcal{F}}|L_n(f)-L'_n(f)|>\frac{\epsilon}{2}\right).$$

Taking expectation w.r.t. S_n brings the desired result. □

Remark 1.38. *Note that* $L_n(f) - L(f)$ *can take any real value, whereas* $L_n(f) - L'_n(f)$ *can only take finitely many values, i.e. we can hope to apply the union bound in the uncountable case.*

Theorem 1.39 (Vapnik-Chervonenkis bound). *We again consider* $\mathcal{Y} = \{0, 1\}$. *Then with probability* $\geq 1 - \delta$

$$L(\hat{f}_n) - L(f^*) \leq 2\sqrt{\frac{4}{n}\log\left(\frac{4S_{\mathcal{F}}(2n)}{\delta}\right)}.$$

Proof. With the preceding lemma we have

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}}|L_n(f)-L(f)|\geq\epsilon\right) &\leq 2\mathbb{P}\left(\sup_{f\in\mathcal{F}}|L_n(f)-L'_n(f)|>\frac{\epsilon}{2}\right)\\ &= 2\mathbb{P}\left(\max_{l\in\mathcal{F}_{S_n,S'_n}}|\tilde{L}_n(l)-\tilde{L}'_n(l)|>\frac{\epsilon}{2}\right)\\ &\leq 2\sum_{l\in\mathcal{F}_{S_n,S'_n}}\mathbb{P}\left(|\tilde{L}_n(l)-\tilde{L}'_n(f)|>\frac{\epsilon}{2}\right)\\ &\leq 2\sum_{l\in\mathcal{F}_{S_n,S'_n}}2e^{-\frac{n}{4}\epsilon^2}\\ &\leq 4S_{\mathcal{F}}(2n)e^{-\frac{n}{4}\epsilon^2},\end{aligned}$$

where with some abuse of notation $\tilde{L}_n(l) - \tilde{L}'_n(l) = \frac{1}{n}\left(\sum_{i=1}^{n} l_i - \sum_{i=n+1}^{2n} l_i\right)$ with $l_i = l(f(X_i), Y_i)$, for which we used Hoeffding's inequality. □

Still we have to answer the question when this bound goes to zero for $n \to \infty$, or in other words, what the geometry of $\mathcal{F}$ has to be s.t. $S_{\mathcal{F}}(n)$ does not grow too fast. The crucial tool that shall answer these questions is the following.

Definition 1.40 (VC dimension)**.** *The VC dimension of a function class $\mathcal{F}$ is defined as*

$$VC(\mathcal{F}) = \sup\{n : S_{\mathcal{F}}(n) = 2^n\}.$$

If such an n does not exist we set $\mathrm{VC}(\mathcal{F}) = \infty$.

Remark 1.41. $\mathrm{VC}(\mathcal{F}) = d$ *implies that there exists a sample $X_1, \dots, X_d \in \mathcal{X}$ s.t. for any $Y_1, \dots, Y_d \in \mathcal{Y}$ there exists an $f \in \mathcal{F}$ with $f(X_i) = Y_i$. We say that the sample is "shattered".*

Remark 1.42. *The VC dimension can be understood as measuring the effective size of a function class. Note that it is distribution independent. One can consider a concept called VC entropy to include properties of $\mathbb{P}$ (see for instance chapter 4.3 in [3]).*

Class $\mathcal{F}$	$\mathrm{VC}(\mathcal{F})$
$\{f_1, \dots, f_N\}$	$\leq \log_2 N$
Intervals on the real line	2
Discs in $\mathbb{R}^2$	3
Rectangles in $\mathbb{R}^d$	$2d$
Half-spaces in $\mathbb{R}^d$	$d+1$
Convex polygons in $\mathbb{R}^2$	∞
$\{\mathrm{sgn}(\sin(\alpha x)) : \alpha \in \mathbb{R}\}$	∞

Figure 1: Examples of VC dimensions.

Example 1.43 (VC dimensions)**.** *The last example in figure 1 is interesting as it shows that the VC dimension is not necessarily linked to the number of free parameters. To show this take for instance $X_i = 10^{-i}$ and $\alpha = \pi\left(1 + \sum_{i=1}^n \frac{1}{2}(1 - Y_i)10^i\right)$. Then*

$$\begin{aligned} f(X_j) &= \mathrm{sgn}\left(\sin\left(10^{-j}\pi + \sum_{i=1}^n \frac{\pi}{2}(1-Y_i)10^{i-j}\right)\right) \\ &= \mathrm{sgn}\left(\sin\left(10^{-j}\pi + \sum_{i\leq j, Y_i=-1} \frac{\pi}{2}(1-Y_i)10^{i-j}\right)\right) \\ &= \mathrm{sgn}\left(\sin\left(\frac{\pi}{2}(1-Y_j) + \pi\left(\underbrace{10^{-j} + \sum_{i<j, Y_i=-1} 10^{i-j}}_{<1}\right)\right)\right), \end{aligned}$$

after the properties of the geometric series. Now, if $Y_j = 1$ the sine function takes values between 0 and π and if $Y_j = -1$ it takes values between π and 2π – therefore the classification is correct.

The VC dimension brings the following theorem.

Theorem 1.44 (Sauer, Shelah, Perles, Vapnik, Chervonenkis)**.** *Assume* $\mathrm{VC}(\mathcal{F}) = d < \infty$. *Then for all n*

$$S_{\mathcal{F}}(n) \leq \sum_{i=0}^d \binom{n}{i}$$

and for all $n \geq d$ we get

$$S_{\mathcal{F}}(n) \leq \left(\frac{en}{d}\right)^d.$$

Proof. For $n = d = 1$ we have

$$S_{\mathcal{F}}(1) \leq \binom{1}{0} + \binom{1}{1} = 2,$$

which clearly holds. By induction assume that the statement holds for $n-1$ and $d-1$ as well as for $n-1$ and d. We want to show that it also holds for n and d. Let $S_1 = (Z_1, \ldots, Z_n)$ and $S_2 = (Z_2, \ldots, Z_n)$, let $\mathcal{F}_{S_1}$ and $\mathcal{F}_{S_2}$ be the corresponding projections onto $\mathcal{F}$. For $f, g \in \mathcal{F}$ write

$$f \sim g \qquad \text{if} \qquad f(Z_1) = 1 - g(Z_1) \quad \text{and} \quad f(Z_j) = g(Z_j) \qquad \text{for } j = 2, \ldots, n.$$

and let

$$\mathcal{G} = \{f \in \mathcal{F} : \exists g \in \mathcal{F} \text{ s.t. } f \sim g\}.$$

Now consider $\mathcal{G}_{S_2}$, then

$$|\mathcal{F}_{S_1}| = |\mathcal{F}_{S_2}| + |\mathcal{G}_{S_2}|.$$

Note that $\mathrm{VC}(\mathcal{F}_{S_2}) \leq d$ and $\mathrm{VC}(\mathcal{G}_{S_2}) \leq d-1$. The latter follows since, if $\mathcal{G}_{S_2}$ shatters a set, we can add Z_1 to create a set that is shattered by $\mathcal{F}_{S_1}$. Therefore we have by assumption that

$$\begin{aligned}
|\mathcal{F}_{S_1}| &\leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\
&= \sum_{i=0}^{d} \left(\binom{n-1}{i} + \binom{n-1}{i-1} \right) \\
&= \sum_{i=0}^{d} \binom{n}{i}.
\end{aligned}$$

To prove the second statement we note that for $n \geq d$

$$\begin{aligned}
\sum_{i=0}^{d} \binom{n}{i} &\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \\
&\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^i \\
&= \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \\
&\leq \left(\frac{n}{d}\right)^d e^d.
\end{aligned}$$

□

In analogy to the Vapnik-Chervonenkis bound one can also bound the Rademacher complexity. For this we need

Lemma 1.45. *Let $X_1, \ldots, X_n$ be random variables with $\mathbb{E}\left[e^{\lambda X_i}\right] \leq e^{\frac{\lambda^2 \xi^2}{2}}$ for all $\lambda > 0$ and $\xi > 0$. Then*

$$\mathbb{E}\left[\max_{1 \leq i \leq n} X_i\right] \leq \xi \sqrt{2 \log n}.$$

Proof. With Jensen's inequality we have

$$\begin{aligned}
\exp\left(\lambda \mathbb{E}\left[\max_{1 \leq i \leq n} X_i\right]\right) &\leq \mathbb{E}\left[\exp\left(\lambda \max_{1 \leq i \leq n} X_i\right)\right] \\
&= \mathbb{E}\left[\max_{1 \leq i \leq n} \exp\left(\lambda X_i\right)\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right] \\
&\leq n e^{\frac{\lambda^2 \xi^2}{2}}.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\max_{1\le i\le n} X_i\right] \le \frac{\log n}{\lambda} + \frac{\lambda\xi^2}{2}$$

and minimizing w.r.t. λ gives the desired bound. □

Theorem 1.46. *Let $\mathcal{F}$ be a set of binary functions. Then*

$$R_n(\mathcal{F}) \le \sqrt{\frac{2\log S_{\mathcal{F}}(n)}{n}}.$$

Proof. We have

$$\begin{aligned} R_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right|\right] \\ &= \mathbb{E}_{S_n}\left[\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i f(Z_i)\right|\middle| S_n\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\max_{v_i\in\mathcal{F}_{S_n}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i v_i\right|\middle| S_n\right]\right]. \end{aligned}$$

Note that $\frac{\sigma_i v_i}{n}$ has mean zero and $-\frac{1}{n} \le \frac{\sigma_i v_i}{n} \le \frac{1}{n}$. Therefore, with lemma 1.16 we have $\mathbb{E}\left[e^{\lambda\sigma_i v_i}\right] \le e^{\frac{\lambda^2}{2n^2}}$ and $\mathbb{E}\left[e^{\lambda\sum_{i=1}^n \frac{\sigma_i v_i}{n}}\right] \le e^{\frac{\lambda^2}{2n}}$ and thus with the preceding lemma 1.45 we get

$$\mathbb{E}\left[\max_{v\in\mathcal{F}_{S_n}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i v_i\right|\middle| S_n\right] \le \sqrt{\frac{2\log|\mathcal{F}_{S_n}|}{n}} \le \sqrt{\frac{2\log S_{\mathcal{F}}(n)}{n}}.$$

□

Remark 1.47. *For $n \ge d$ the growth function $S_{\mathcal{F}}(n)$ only grows polynomially and therefore gives a helpful bound when plugging it in to theorem 1.39, namely $\le C\sqrt{\frac{d\log n}{n}}$. In other words, a finite VC dimension brings learnability. In fact, using techniques called "chaining" and "covering numbers" one can even get a bound $\le C\sqrt{\frac{d}{n}}$ – details can for instance be found in [5], chapter 4.3.*

In analogy to theorem 1.27 one can show

Theorem 1.48. *With probability at least $1-\delta$ we have*

$$L(\hat{f}_n) - L(f^*) \le 2\sqrt{L(f^*)\frac{\log S_{\mathcal{F}}(2n) + \log\frac{4}{\delta}}{n}} + \frac{\log S_{\mathcal{F}}(2n) + \log\frac{4}{\delta}}{n}.$$

Proof. See chapter 6.2 in [3]. □

Lemma 1.49 (Contraction principle)**.** *Assume $\varphi : \mathbb{R}\to\mathbb{R}, \varphi(0) = 0$ and φ is L_φ-Lipschitz. Then for all function classes $\mathcal{F}$*

$$R_n(\varphi\circ\mathcal{F}) \le L_\varphi R_n(\mathcal{F}).$$

Proof. Let $g_i(z)$ and $h_i(z)$ be functions s.t. for any z, z'

$$|g_i(z) - g_i(z')| \le |h_i(z) - h_i(z')|.$$

Let us show

$$\mathbb{E}_{\sigma_1,\dots,\sigma_n,x}\left[\sup_z\left\{c(x,z) + \sum_{i=1}^n \sigma_i g_i(z)\right\}\right] \le \mathbb{E}_{\sigma_1,\dots,\sigma_n,x}\left[\sup_z\left\{c(x,z) + \sum_{i=1}^n \sigma_i h_i(z)\right\}\right]$$

by induction. Obviously it is true for $n = 0$. Let us assume it holds for n, then

$$\begin{aligned}
\mathbb{E}_{\sigma_1,\dots,\sigma_{n+1},x}\left[\sup_z\left\{c(x,z)+\sum_{i=1}^{n+1}\sigma_i g_i(z)\right\}\right] &= \mathbb{E}_{\sigma_1,\dots,\sigma_n,x}\left[\sup_{z,z'}\left\{\frac{c(x,z)+c(x,z')}{2}+\sum_{i=1}^{n}\sigma_i\frac{g_i(z)-g_i(z')}{2}+\frac{g_{n+1}(z)-g_{n+1}(z')}{2}\right\}\right] \\
&\le \mathbb{E}_{\sigma_1,\dots,\sigma_n,x}\left[\sup_{z,z'}\left\{\frac{c(x,z)+c(x,z')}{2}+\sum_{i=1}^{n}\sigma_i\frac{g_i(z)-g_i(z')}{2}+\frac{h_{n+1}(z)-h_{n+1}(z')}{2}\right\}\right] \\
&= \mathbb{E}_{\sigma_1,\dots,\sigma_{n+1},x}\left[\sup_y\left\{c(x,z)+\sum_{i=1}^{n}\sigma_i g_i(y)+\sigma_{n+1}h_{n+1}(z)\right\}\right] \\
&\le \mathbb{E}_{\sigma_1,\dots,\sigma_{n+1},x}\left[\sup_z\left\{c(x,z)+\sum_{i=1}^{n+1}\sigma_i h_i(z)\right\}\right].
\end{aligned}$$

One can show the same also with absolute values. Now take $c(x,z) = 0$, $f(z) = \varphi(f(x)y)$ and $h(z) = L_\varphi f(x)y$ to get the result. □

Remark 1.50. *This lemma brings that Rademacher complexity is bounded whenever the loss is Lipschitz and therefore also brings learning statements for the regression setting. To give some examples, hinge loss given by $l(y,y') = \max\{1-yy',0\}$ is 1-Lipschitz, logistic loss given by $l(y,y') = \log(1+e^{-yy'})$ is 1-Lipschitz, squared loss $l(y,y') = (y'-y)^2$ is $4B$-Lipschtiz when $|y|,|y'| \le B$, absolute loss $l(y,y') = |y-y'|$ is 1-Lipschitz.*

2 Algorithms

2.1 Adaptive boosting

The idea of adaptive boosting is to have multiple maybe not so well performing classifiers, and combine them to a single classifier that performs well [8, 7, 1]. We again consider binary classification, i.e. $\mathcal{Y} = \{-1,1\}$ and choose K multiple base classifiers $\{f_k\}_{k=1}^K$ with $f_k \in \mathcal{F}$. We then consider the composition classifier $\tilde{f}_K(x) = \operatorname{sign}\left\{\sum_{k=1}^K \alpha_k f_k(x)\right\}$, where $\alpha_i \in \mathbb{R}$ are weights – so the overall classification can be interpreted as a "weighted majority rule": each datum x gets mapped to a vector $(f_1(x),\dots,f_K(x)) \in \{-1,1\}^T$ and the weights tell us how to combine these multiple predictions. We now want to study the generalization and therefore estimation error. We have

$$\begin{aligned}
L_n(\tilde{f}_K) &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\tilde{f}_K(X_i) \ne Y_i\} \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{Y_i\sum_{k=1}^K \alpha_k f_k(X_i) \le 0\} \\
&\le \frac{1}{n}\sum_{i=1}^n \exp\left(-Y_i\sum_{k=1}^K \alpha_k f_k(X_i)\right) =: \mathcal{L}_n(\tilde{f}_K),
\end{aligned}$$

where the last line follows from $\mathbb{1}\{z \le 0\} \le e^{-z}$. Instead of minimizing $L_n(\tilde{f}_K)$ directly, which can be difficult as it is a non-continuous function, we can minimize the upper bound $\mathcal{L}_n(\tilde{f}_K)$ and hope that we will find a "similar" minimium. Note that we still consider a minimization over $f_1,\dots,f_K$ as well as $\alpha_1,\dots,\alpha_K$. Since this is still computationally heavy we consider an iterative approach, a greedy optimitazion: Imagine we already tuned $f_1,\dots,f_K$ and $\alpha_1,\dots,\alpha_K$, how do we then choose f_{K+1} and α_{K+1}? Let us change the writing a bit:

$$\begin{aligned}
\mathcal{L}_n(\tilde{f}_{K+1}) &= \sum_{i=1}^n \underbrace{\frac{1}{n}\exp\left(-Y_i\sum_{k=1}^K \alpha_k f_k(X_i)\right)}_{=:w_i^{(K)}} \exp\left(-Y_i\alpha_{K+1}f_{K+1}(X_i)\right) \\
&= \sum_{i=1}^n w_i^{(K)} \exp\left(-Y_i\alpha_{K+1}f_{K+1}(X_i)\right).
\end{aligned}$$

Let $\tilde{w}_i^{(K)} := \frac{w_i^{(K)}}{\sum_{i=1}^n w_i^{(K)}}$, then

$$\begin{aligned}\frac{\mathcal{L}_n(\tilde{f}_{K+1})}{\mathcal{L}_n(\tilde{f}_K)} &= \sum_{i=1}^n \tilde{w}_i^{(K)} \exp\left(-Y_i \alpha_{K+1} f_{K+1}(X_i)\right) \\ &= e^{-\alpha_{K+1}} \sum_{i: f_{K+1}(X_i)=Y_i} \tilde{w}_i^{(K)} + e^{\alpha_{K+1}} \sum_{i: f_{K+1}(X_i)\neq Y_i} \tilde{w}_i^{(K)} \\ &= \left(e^{\alpha_{K+1}} - e^{-\alpha_{K+1}}\right) \sum_{i=1}^n \tilde{w}_i^{(K)} \mathbb{1}\{f_{K+1}(X_i) \neq Y_i\} + e^{-\alpha_{K+1}}.\end{aligned}$$

Note that $f_{K+1} \in \mathcal{F}$ only appears in $\epsilon_K := \sum_{i=1}^n \tilde{w}_i^{(K)} \mathbb{1}\{f_{K+1}(X_i) \neq Y_i\}$. Minimization of the above expression gives

$$\alpha_{K+1}^* = \frac{1}{2} \log \frac{1-\epsilon_K}{\epsilon_K},$$

which is positive if $\epsilon_K < \frac{1}{2}$, which we expect since $\epsilon_K = \frac{1}{2}$ would just correspond to random guessing. Note that

$$w_i^{(K+1)} = w_i^{(K)} e^{-Y_i f_{K+1}(X_i) \alpha_{K+1}},$$

which explains the name of "adaptive boosting". While learning we concentrate on those predictors that have made mistakes in the previous training rounds. We now make the assumption:

$$\epsilon_k < \frac{1}{2} - \gamma, \quad \gamma > 0$$

for every $k = 1, \ldots, K$. Note that this assumption depends on choosing a good function class $\mathcal{F}$. We then have

$$\begin{aligned}\mathcal{L}_n(\tilde{f}_{K+1}) &= \mathcal{L}_n(\tilde{f}_K) \left(\left(e^{\alpha_{K+1}^*} - e^{-\alpha_{K+1}^*}\right) \epsilon_K + e^{-\alpha_{K+1}^*}\right) \\ &= \mathcal{L}_n(\tilde{f}_K) 2\sqrt{\epsilon_K(1-\epsilon_K)} \\ &\leq \mathcal{L}_n(\tilde{f}_K) 2\sqrt{\left(\frac{1}{2} - \gamma\right)\left(\frac{1}{2} + \gamma\right)} \\ &= \mathcal{L}_n(\tilde{f}_K)\sqrt{1-4\gamma^2} \\ &\leq \mathcal{L}_n(\tilde{f}_K) e^{-2\gamma^2}\end{aligned}$$

and therefore

$$L_n(\tilde{f}_K) \leq \mathcal{L}_n(\tilde{f}_K) \leq (e^{-2\gamma^2})^K \mathcal{L}(\tilde{f}_0) = e^{-2K\gamma^2}.$$

Since $L_n(\tilde{f}_K) \in \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ we now know that AdaBoost achieves zero empirical loss in finite steps. For investigating the generalization error let us say $\tilde{f}_K \in C(\mathcal{F}, K)$ and note that we have $C(\mathcal{F}, K) \subset C(\mathcal{F}, K+1)$. This means that the approximation error gets smaller, however at the same time the estimation error could get worse. Theorems 1.39 and 1.44 give us

$$L(\tilde{f}_K) \leq L_n(\tilde{f}_K) + \mathcal{O}\left(\sqrt{\frac{\mathrm{VC}(C(\mathcal{F}, K)) \log n + \log \frac{1}{\delta}}{n}}\right),$$

where the first summand goes to zero, however, one can show that $\mathrm{VC}(C(\mathcal{F}, K)) \approx K \,\mathrm{VC}(\mathcal{F})$ (see lemma 10.3 in [9]). So the error should grow sublinearly in K. Still, in practice, one can observe that the loss on a test set keeps decreasing even after the empirical risk has gone to zero. In order to understand why this is the case another approach is necessary that shall be summarized with the following theorem.

Lemma 2.1. *For the Rademacher complexity of convex hulls of function classes we have*

$$\hat{R}_n(\mathrm{conv}(\mathcal{F})) = \hat{R}_n(\mathcal{F}).$$

Proof.

$$\begin{aligned}
\hat{R}_n(\text{conv}(\mathcal{F})) &= \mathbb{E}_\sigma \left[\sup_{f_k \in \mathcal{F}, \alpha \geq 0, \|\alpha\|_1 = 1} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \sum_{k=1}^K \alpha_k f_k(x_i) \right| \right] \\
&= \mathbb{E}_\sigma \left[\sup_{f_k \in \mathcal{F}} \sup_{\alpha \geq 0, \|\alpha\|_1 = 1} \left| \frac{1}{n} \sum_{k=1}^K \alpha_k \sum_{i=1}^n \sigma_i f_k(x_i) \right| \right] \\
&= \mathbb{E}_\sigma \left[\sup_{f_k \in \mathcal{F}} \max_{1 \leq k \leq K} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f_k(x_i) \right| \right] \\
&= \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right| \right] \\
&= \hat{R}_n(\mathcal{F}).
\end{aligned}$$

□

Remark 2.2. *We have noted before that this is not valid for the VC dimension.*

Theorem 2.3. *Let $\mathcal{Y} = \{-1, 1\}$, $l(y', y) = \mathbb{1}\{y' \neq y\}$. We consider the base class $\mathcal{F}$ with $\text{VC}(\mathcal{F}) < \infty$ and classifiers of the form $\tilde{f}(x) = \text{sign}\{g(x)\}$, where $g \in \text{conv}(\mathcal{F}) := \{g(x) = \sum_{k=1}^K \alpha_k f_k(x), K \in \mathbb{N}, f_1, \ldots, f_K \in \mathcal{F}, \alpha_k > 0, \sum_{k=1}^K \alpha_k = 1\}$. Then for all $\eta > 0$ with probability $\geq 1 - \delta$*

$$L(\tilde{f}) \leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{g(X_i) Y_i \leq \eta\} + \frac{2}{\eta} \sqrt{\frac{2 \,\text{VC}(\mathcal{F}) \log(n+1)}{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{n}}.$$

Proof. Let

$$L(\tilde{f}) = \mathbb{E}[\mathbb{1}\{\tilde{f}(X) \neq Y\}] = \mathbb{E}[\mathbb{1}\{g(X)Y \leq 0\}].$$

Take any function $\psi : \mathbb{R} \to [0, 1]$ that is L_ψ-Lipschitz and for that $\psi(x) \geq \mathbb{1}\{x \leq 0\}$. Then with theorem 1.33

$$\begin{aligned}
L(\tilde{f}) &\leq \mathbb{E}[\psi(g(X)Y)] \\
&\leq \frac{1}{n} \sum_{i=1}^n \psi(g(X_i)Y_i) + 2R_n(\psi) + \sqrt{\frac{\log \frac{1}{\delta}}{n}}
\end{aligned}$$

with

$$\begin{aligned}
R_n(\psi) &= \mathbb{E}_{S_n, \sigma} \left[\sup_{g \in \text{conv}(\mathcal{F})} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \psi(g(X_i)Y_i) \right| \right] \\
&= \mathbb{E}_{S_n, \sigma} \left[\sup_{g \in \text{conv}(\mathcal{F})} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i \left(\psi(g(X_i)Y_i) - \psi(0) \right) \right| \right] \\
&\leq L_\psi \mathbb{E}_{S_n, \sigma} \left[\sup_{g \in \text{conv}(\mathcal{F})} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g(X_i)Y_i \right| \right] \\
&= L_\psi \mathbb{E}_{S_n, \sigma} \left[\sup_{g \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i g(X_i) \right| \right] \\
&\leq L_\psi \sqrt{\frac{2 \,\text{VC}(\mathcal{F}) \log n}{n}},
\end{aligned}$$

where we used lemmas 1.49 and 2.1, theorem 1.46 and the fact that $\sigma_i \overset{d}{=} \sigma_i Y_i$. Setting

$$\psi(x) = \begin{cases} 1 & \text{if } x \leq 0 \\ 0 & \text{if } x \geq \eta \\ 1 - \frac{x}{\eta} & \text{in between} \end{cases}$$

yields $L_\psi = \frac{1}{\eta}$ and completes the proof. □

Remark 2.4. *$|g(X_i)Y_i|$ can be interpreted as a margin or a confidence of a prediction $g(X_i)$. Then, the indicator function not only counts the number of wrong classifications, but also the correct classifications that have however not been confident enough (depending on the parameter η).*

3 Bibliography

[1] Bartlett, P. L. and Traskin, M. (2007). Adaboost is consistent. *Journal of Machine Learning Research*, 8(Oct):2347–2368.

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

[3] Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to statistical learning theory. pages 169–207.

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

[5] Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation*. Springer, New-York.

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

[7] Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, pages 1–50.

[8] Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S., et al. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26(5):1651–1686.

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

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