Lecture Notes on Statistical Learning Theory

Marius Kloft

June 20, 2013

1 Introduction

Problem setting. In statistical learning theory (SLT), the goal is to find a classifier $g : \mathbb{R}^d \to \{0, 1\}$, predicting the correct class y of an observation $x \in \mathbb{R}^d$, based on data $(x_1, y_1), \ldots, (x_n, y_n)$. Clearly, we cannot learn any reasonable classifier, if no assumption is imposed on the relationship between the data and the test observation (x, y). To this end, we assume in SLT that the data pairs $D_n := (x_i, y_i)_{i=1}^n$ and the test observation (x, y) are independently drawn from one and the same probability distribution $\mathbb{P}$. Correspondingly, we denote the random variables associated to (x_i, y_i) and (x, y) by capital letters, i.e., (X_i, Y_i) and (X, Y), respectively. Thus a classifier errs if $g(X) \neq Y$ so that $L(g) := \mathbb{P}(g(X) \neq Y | D_n)$ is the probability of error of g.

The Bayes classifier and the ERM approximation. The most accurate — the best! — classifier is the *Bayes classifier*, given by

$$g^*(x) := \operatorname*{argmin}_g L(g) = \begin{cases} 1, & \text{if } \mathbb{P}(Y = 1 | X = x) > \frac{1}{2} \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

i.e., in average no classifier errs less than the Bayes classifier. If the distribution $\mathbb{P}$ is known the Bayes classifier may be computed.

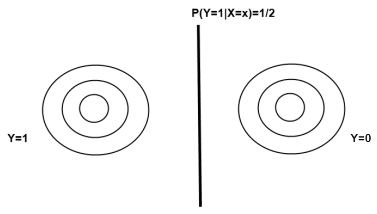

Figure 1: Illustration of the Bayes classifier. A data point is assigned $y = 1$ when $\mathbb{P}(Y = 1|X = x) > \frac{1}{2}$ and $y = 0$ otherwise.

However, most often $\mathbb{P}$ is unknown in practice and needs to be approximated on

base of the data. To this end, note that the quantity $\hat{L}_n(g) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{g(X_i)\neq Y_i\}}$ (called *empirical error* or *training error*), which counts the number of false predictions on the data, has in expectation: $\mathbb{E}\hat{L}_n(g) = L(g)$. Thus we may use $\hat{L}_n(g)$ as a rough estimate/approximation for $L(g)$ so that (1) becomes

$$g^* := \operatorname*{argmin}_{g\in\mathcal{C}} \hat{L}_n(g)\,, \tag{2}$$

where we additionally allow restricting the minimization to a some predefined class $\mathcal{C}$ of classifiers. The minimization procedure (2) is called *empirical risk minimization* (ERM). In comparison to the Bayes classifier, we thus employ two approximations:

1. in the minimization, we replace $L(g)$ by $\hat{L}_n(g)$
2. we restrict the computation of the minimum to a class of functions $\mathcal{C}$.

The basic decomposition. Of course, we are interested in what is "lost" by the above ERM approximation, given by (2). To this end, denoting the best classifier in the class $\mathcal{C}$ by $g^*_{\mathcal{C}} := \operatorname{argmin}_{g\in\mathcal{C}} L(g)$, it holds

$$\underbrace{L(g_n^*) - L(g^*)}_{\substack{\text{difference between Bayes classifier and ERM}}} \;=\; \underbrace{L(g_n^*) - L(g^*_{\mathcal{C}})}_{\text{"estimation error"}} \;+\; \underbrace{L(g^*_{\mathcal{C}}) - L(g^*)}_{\text{"approximation error"}}\,.$$

This is the basic inequality of all of SLT and the starting point for all that is coming up. A bad message is appropriate here: although we may be able to show that the approximation error decreases to zero in n for many learning algorithms (such as ERM), the *speed* of convergence can be arbitrarily slow — unless make rather strong assumptions on the "smoothness" of the Bayes classifier.

Therefore, the majority of literature on learning theory focuses on bounding the *estimation error*, for which we are able to derive precise convergence speeds. The following Lemma is very helpful in this respect:

Lemma 1.

$$L(g_n^*) - L(g^*_{\mathcal{C}}) \;\leq\; 2\sup_{g\in\mathcal{C}} \left|\hat{L}_n(g) - L(g)\right|\,.$$

Table 1: Definitions and names of some basic "players" of statistical learning theory.

$L(g) := \mathbb{P}(g(X) \neq Y \vert D_n)$	probability of error of g
$\hat{L}_n(g) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{g(X_i)\neq Y_i\}}$	error of g on data ("empirical/training error")
$g^* := \text{argmin}_g\, L(g)$	"Bayes classifier"
$g_n^* := \text{argmin}_{g\in\mathcal{C}}\, \hat{L}(g)$	"empirical risk minimization" (ERM)
$g_{\mathcal{C}}^* := \text{argmin}_{g\in\mathcal{C}}\, L(g)$	best classifier in class

Proof. It holds

$$\begin{aligned} & L(g_n^*) - L(g_{\mathcal{C}}^*) \\ &= L(g_n^*) - \hat{L}_n(g_n^*) + \Big(\underbrace{\hat{L}_n(g_n^*)}_{\overset{(*)}{\leq} \hat{L}(g_{\mathcal{C}}^*)} - \hat{\mathrm{L}}(g_{\mathcal{C}}^*) \Big) \\ &\leq 2 \sup_{g\in\mathcal{C}} \left| \hat{L}_n(g) - L(g) \right| , \end{aligned}$$

where $(*)$ holds because, by (2), $\hat{L}_n(g_n^*) \leq \hat{L}_n(g_{\mathcal{C}}^*)$. □

The above lemma states that upper bounds on $\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|$ automatically provide us with upper bounds on the sub-optimality of g_n^* within $\mathcal{C}$, that is, a bound for the estimation error $L(g_n^*) - L(g_{\mathcal{C}}^*)$. We can thus summarize:

The classical aim of classical statistical learning theory is to derive upper bounds on $\sup_{g\in\mathcal{C}} \left\vert \hat{L}_n(g) - L(g) \right\vert$.

The various forms of convergence. Warning: there are two forms of convergence that should not be confused:

1. pointwise convergence: $\forall g \in \mathcal{C} : |\hat{L}_n(g) - L(g)| \to 0$ when $n \to \infty$
2. uniform convergence: $\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \to 0$ when $n \to \infty$.

To understand the difference in the two ways of convergence, consider the function

$$f : \begin{array}{ccc} [0,1[& \to & \mathbb{R} \\ x & \mapsto & x^n \end{array}$$

For any fixed $x \in [0,1[$ the function $f(x) = x^n$ converges to zero when $n \to \infty$. In contrast, $\sup_{x\in[0,1[} x^n = 1$ for any $n \in \mathbb{N}$, which thus does not converge to zero.

As Lemma (1) shows, what we need in SLT is *uniform* convergence. The underlying reason is that the function g is not fixed beforehand, but computed on base of the data; it is thus itself a random quantity.

1.1 What is coming up

We approach bounding $P(\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \geq t)$ in the following two steps:

1. showing that $\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|$ is concentrated around its mean $\mathbb{E}\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|$
2. showing that $\mathbb{E}\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \to 0$ when $n \to \infty$ at an adequate rate.

This is justified by the following decomposition

$$\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \leq \underbrace{\left| \sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| - \mathbb{E}\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \right|}_{\leq \text{ bound } (\textsc{Step } 1)} + \underbrace{\mathbb{E}\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)|}_{\leq \text{ bound } (\textsc{Step } 2)} \tag{3}$$

The first required result, $P(\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \geq t)$, directly follows from a very powerful concentration inequality known as *McDiarmid's inequality*. Afterwards, we address bounding $\mathbb{E}\sup_{g\in\mathcal{C}} |\hat{L}_n(g) - L(g)| \to 0$ using Vapnik-Chervonenkis theory.

2 Basic Concentration Inequalities

The aim of this section is to prove McDiarmid's inequality, a powerful *uniform* concentration inequality. On the way, we need to first show a couple of pointwise inequalities.

2.1 Pointwise Concentration Inequalities

For the moment, we will study the convergence of $\hat{L}_n(g) - L(g)$ pointwise, i.e., for fixed g because this is the starting point for the uniform convergence analysis to be carried out in the next section.

What is a *concentration inequality*? To study the convergence of random variables to its means, probability theory offers a powerful machinery, called *concentration inequalities*, that is, inequalities of the form:

Definition 2. *A bound of the following form is called* concentration inequality*: for a random variable Z and any real number $t > 0$,*

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq \text{bound}(t, n)\,.$$

Note that we can analogously study left-sided concentration inequality, i.e., $\mathbb{P}(Z - \mathbb{E}Z \leq t) \geq \text{bound}(t, n)$, as well as bounds on the absolute deviation, $\mathbb{P}(|Z - \mathbb{E}Z| \leq t) \leq \text{bound}(t, n)\,.$

The starting point of everything: Markov's inequality. A simple, yet very useful inequality is Markov's inequality:

Proposition 3 (Markov's inequality)**.** *For any positive random variable Z and any real number $t > 0$,*

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}Z}{t}\,. \tag{4}$$

Proof. The core idea of the proof is to consider the random variable $Z_t := t\mathbb{I}_{\{Z \geq t\}}$. Note that Z_t is positive and it holds $Z_t \leq Z$ with probability one as well as, per construction, $EZ_t = t\mathbb{E}\mathbb{I}_{\{Z \geq t\}} = t\mathbb{P}(Z \geq t)$. Thus it follows

$$\mathbb{P}(Z \geq t) = \frac{EZ_t}{t} \leq \frac{\mathbb{E}Z}{t}\,,$$

which was to show. □

Our first concentration inequality: Chebychef's inequality. From Markov's inequality we can derive our first concentration inequality (i.e., an inequality on the sense of Definition 2), namely, Chebyshev's inequality.

Proposition 4 (Chebyshev's inequality)**.** *For any random variable Z with $\operatorname{var}(Z) = \sigma^2 < \infty$ and any $t > 0$,*

$$\mathbb{P}\left(|Z - \mathbb{E}Z| \geq t\right) \leq \frac{\sigma^2}{t^2}\,. \tag{5}$$

Proof. Since $|Z - \mathbb{E}Z|^2$ is a positive random variable, by Markov's inequality,

$$\mathbb{P}\left(|Z - \mathbb{E}Z| \geq t\right) = \mathbb{P}(|Z - \mathbb{E}Z|^2 \geq t^2) \leq \frac{\mathbb{E}\,|Z - \mathbb{E}Z|^2}{t^2} = \frac{\operatorname{var}(Z)}{t^2}\,,$$

which was to show. □

Alternative formulation of concentration inequalities. Besides the standard form of Definition 2, there is an alternative formulation of concentration inequalities, as we illustrate by means of Chebychef's inequality:

Proposition 5 (Cheb.'s inequ., alternative formulation)**.** *For any random variable Z with $\operatorname{var}(Z) = \sigma^2 < \infty$ and any $t > 0$, it holds with probability $1 - \delta$ over the draw of Z that*

$$|Z - \mathbb{E}Z| \leq \sigma\sqrt{\delta}\,.$$

Proof. Denote the right-hand side of (5) by δ. Hence, with probability δ, it holds $|Z - \mathbb{E}Z| \geq \sigma\sqrt{\delta}$ and thus, by inversion, with probability $1 - \delta$, it holds $|Z - \mathbb{E}Z| \leq \sigma\sqrt{\delta}$, which was to show. □

Application of Chebychef's inequality: the weak law of large numbers. As a very nice and exciting application of Chebyshev's inequality, we can — surprisingly easily — prove one of the fundamental results of probability theory: the *weak law of large numbers*.

Proposition 6 (Weak law of large numbers)**.** *Let $(Z_i)_{i\in\mathbb{N}}$ be an i.i.d. family with expected value $\mathbb{E}Z_i = \mu$ and $\operatorname{var}(Z_i) = \sigma^2 < \infty$. Then, when $n \to \infty$,*

$$\mathbb{P}\left(\left|\bar{Z}_n - \mu\right| \geq t\right) \to 0\,,$$

where $\bar{Z}_n := \frac{1}{n}\sum_{i=1}^{n} Z_i$ is the nth sample mean.

Proof. By Chebychef's inequality,

$$\mathbb{P}\left(\left|\bar{Z}_n \geq t\right|\right) \leq \frac{\operatorname{var}\left(\bar{Z}_n\right)}{t^2}$$

An easy calculation now shows:

$$\begin{aligned} \operatorname{var}\left(\bar{Z}_n\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}\frac{1}{n}\sum_{i=1}^{n} Z_i\right)^2 = \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}Z_i)\right)^2 \\ &= \frac{1}{n^2}\sum_{i,j=1}^{n}\operatorname{cov}(Z_i, Z_j) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}(Z_i) = \frac{\sigma^2}{n}\,, \end{aligned} \tag{6}$$

because, by the iidness of the variables $(Z_i)_{i\in\mathbb{N}}$, all covariances $\operatorname{cov}(Z_i, Z_j) = 0$ unless $i = j$. Thus

$$\mathbb{P}\left(\left|\bar{Z}_n - \mathbb{E}Z\right| \geq t\right) \leq \frac{\sigma^2}{nt^2}\,, \tag{7}$$

which converges to zero when $n \to \infty$. □

Can we do better? Chernoff's inequality. The following result by Chernoff (1952) improves on Chebychef's inequality.

Proposition 7 (Chernoff's inequality)**.** *For any random variable Z and any $t > 0$,*

$$\mathbb{P}(Z \geq t) \leq \min_{s\in\mathbb{R}} \frac{M_Z(s)}{e^{st}}\,,$$

where $M_Z(s) = \mathbb{E}e^{sZ}$ is the moment-generating function of Z.

Hint : *the moment-generating function of a random variable Z is defined as $M_Z(s) := \mathbb{E}e^{sZ}$, $s \in \mathbb{R}$.*

Proof. Note that by Markov's inequality $\mathbb{P}(Z \geq t) = \mathbb{P}(e^{sZ} \geq e^{st}) \leq \frac{\mathbb{E}e^{sZ}}{e^{st}}$, which was to show. □

On the first view, the occurrence of the moment-generating function in Chernoff's inequality may look a little complicated, but the moment-generating function of many prominent random variable is very well known from the classical probability literature:

Lemma 8. *A Gaussian random variable Z with expected value μ and variance σ^2 has the moment-generating function $M_Z(s) = e^{\mu s + \frac{1}{2}\sigma^2 s^2}$.*

Regardless of whether or not the moment-generating function is easy to compute, the efforts are definitely worth the price(!): Chernoff's inequality is usually much tighter than Chebychef's inequality, as the following example illustrates.

Example 9. Let $(Z_i)_{i=1,\dots,n}$ be an i.i.d. family if Gaussian random variables with expected value $\mathbb{E}(Z_i) = \mu$ and variance $\mathrm{var}(Z_i) = \sigma^2$. Thus, the sample mean $\bar{Z}_n := \frac{1}{n}\sum_{i=1}^n Z_i$ is a Gaussian random variable with expected value $\mathbb{E}(\bar{Z}_n) = \mu$ and, by (6), variance $\mathrm{var}(\bar{Z}_n) = \sigma^2/n$. Thus its centered version $\bar{Z}_n - \mathbb{E}\bar{Z}_n$ has the moment-generating function

$$M_{\bar{Z}_n - \mathbb{E}\bar{Z}_n}(s) = e^{\frac{\sigma^2 s^2}{2n}}$$

so that Chernoff's inequality gives

$$\mathbb{P}(\bar{Z}_n - \mathbb{E}\bar{Z}_n \geq t) \leq \min_{S\in\mathbb{R}} e^{-st+\sigma^2 s^2/2n}\,, \tag{8}$$

which is minimized for $s = \frac{nt}{\sigma^2}$. Thus, (8) translates into

$$\mathbb{P}(\bar{Z}_n - \mathbb{E}\bar{Z}_n \geq t) \leq e^{-\frac{nt^2}{2\sigma^2}}\,,$$

which — remarkably — is of exponential order in n! (I.e., the above bound decreases "exponentially fast") Contrast this to the bound of Eq. (7), derived by Chebychef's inequality, which just decreases by a linear rate in n.

Moreover, the random variable $\mathbb{E}\bar{Z}_n - \bar{Z}_n$ has, by symmetry, the same distribution as $\bar{Z}_n - \mathbb{E}\bar{Z}_n$ and thus enjoys the very same bound. We thus obtain the following *two-sided* concentration inequality:

$$\mathbb{P}(\left|\bar{Z}_n - \mathbb{E}\bar{Z}_n\right| \geq t) \leq\ 2e^{-\frac{nt^2}{2\sigma^2}}\,. \tag{9}$$

■

<u>Bounded</u> random variables: Höffding's Lemma and Höffding's inequality. A random variable Z is bounded, if there exist constants $a, b > 0$ such that $\mathbb{P}(a \leq Z \leq b) = 1$. In fact, the bounded random variables behave, in many ways, similar to Gaussian ones:

1. the moment-generating function of a bounded random variable can be bounded by the one of a Gaussian distribution *(this statement is known as* "Hoeffding's lemma"*)*

2. similar to Gaussian families (cf. Example 2.1), the tail of the sample mean of a family of bounded random variables decreases exponentially fast in n. *(This statement is known as* "Hoeffding's inequality"*.)*

The first result, Hoeffding's lemma, can be proved using elementary arguments; the proof is shown in Appendix A.

Lemma 10 (HÖFFDING'S LEMMA)**.** *Let Z be a random variable with $\mathbb{E}Z = 0$ and $\mathbb{P}(a \leq Z \leq b) = 1$. Then for any $s \in \mathbb{R}$,*

$$M_Z(s) \,=\, \mathbb{E}e^{sZ} \,\leq\, e^{s^2(b-a)^2/8}\,.$$

I.e., the moment-generating function of Z can be uniformly bounded by the one of a zero mean Gaussian random variable with variance $\sigma^2 = (b-a)^2/4$ (see Lemma 8). We call such a random variable sub-Gaussian with parameter σ^2.

The second result, Höffding's inequality, follows from the above Lemma in combination with Chernoff's inequality.

Lemma 11 (HÖFFDING'S INEQUALITY)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables with $\mathbb{P}(a_i \leq Z_i \leq b_i) = 1$. Denote their mean by $\bar{Z}_n := \frac{1}{n}\sum_{i=1}^n Z_i$. Then, for any $t > 0$, we have the following concentration inequality*

$$\mathbb{P}(\left|\bar{Z}_n - \mathbb{E}\bar{Z}_n\right| \geq t) \leq 2e^{-2n^2\epsilon^2/\sum_{i=1}^n (b_i-a_i)^2}\,.$$

Proof. By Chernoff's inequality,

$$\begin{aligned}
&\mathbb{P}(\bar{Z}_n - \mathbb{E}\bar{Z}_n \geq t) \\
&\quad\leq \min_{s\in\mathbb{R}} e^{-st}\mathbb{E}e^{s(\bar{Z}_n - \mathbb{E}\bar{Z}_n)} \;=\; \min_{s\in\mathbb{R}} e^{-st}\mathbb{E}e^{\frac{s}{n}\sum_{i=1}^n (Z_i - \mathbb{E}Z_i)} \\
&\quad\overset{\text{iid}}{=} \min_{s\in\mathbb{R}} e^{-st}\prod_{i=1}^n \mathbb{E}e^{s\frac{Z_i - \mathbb{E}Z_i}{n}} \;\overset{\text{Lem. 11}}{\leq}\; \min_{s\in\mathbb{R}} e^{-st}\prod_{i=1}^n e^{\frac{s^2(b_i-a_i)^2}{8n^2}} \\
&\quad= e^{\frac{s^2}{n^2}\sum_{i=1}^n (b_i-a_i)^2 - st}.
\end{aligned}$$

The above term is minimized for $s = 4tn^2/\sum_{i=1}^n (b_i - a_i)^2$, so that the above bound translates into

$$\mathbb{P}(\bar{Z}_n - \mathbb{E}\bar{Z}_n \geq t) \leq e^{-2t^2n^2/\sum_{i=1}^n (b_i-a_i)^2}.$$

Furthermore, by the same argumentation, we obtain the following left-sided version of the above bound:

$$\mathbb{P}(\bar{Z}_n - \mathbb{E}\bar{Z}_n \leq -t) \;=\; \mathbb{P}(-(\bar{Z}_n - \mathbb{E}\bar{Z}_n) \geq t) \;\leq\; e^{-2t^2n^2/\sum_{i=1}^n (b_i - a_i)^2}.$$

Combining the above two bounds gives the required result. □

Example 12. If $(Z_i)_{i\in\mathbb{N}}$ is, for example, an i.i.d. Bernoulli family with parameter p, we have $\mathbb{P}(0 \leq Z_i \leq 1) = 1$, so Höffding's inequality gives

$$\mathbb{P}(|\bar{Z}_n - p| \geq t) \leq 2e^{-2nt^2}\,, \tag{10}$$

which is the same, exponential bound as (9) for i.i.d. Gaussian random variables with variance $\sigma^2 = \frac{1}{4}$. This may be contrasted to the rather "slow" bounds obtained by Markov's and Chebychef's inequalities, which are just of linear and quadratic order, respectively. ■

Corollary 13. *Let $\mathcal{C}$ be a class of functions. Then, for any $g \in \mathcal{C}$ and any $t > 0$,*

$$P(|\hat{L}_n(g) - L(g)| \geq t) \leq 2e^{-2nt^2}\,.$$

Proof. Follows from the previous example because $Z_i := \mathbb{I}_{\{g(X_i)\neq Y_i\}}$ is, for any fixed g, a Bernoulli variable with parameter $p := L(g)$. □

How tight is the above result inequality? Again, consider the i.i.d. Bernoulli family $(Z_i)_{i\in\mathbb{N}}$, $Z_i = \mathbb{I}_{\{g(X_i)\neq Y_i\}}$. Then by the central limit theorem, $\bar{Z}_n$ is, for large n, approximately normally distributed with variance $\sigma^2 = p(1-p)/n$. Hence, using $n = 1$ in (9), we would expect the following bound to be approximately valid:

$$P(|\bar{Z}_n - p| \geq t) \leq 2e^{\frac{-t^2}{2\sigma^2}} \leq 2e^{-2nt^2}\,,$$

which is exactly the same bound of Corollary 13, obtained by Höffding's inequality. We can thus conclude that Höffding's inequality is just about of of the right order.

2.2 Uniform Concentration

A very powerful concentration inequality, proved by McDiarmid (1989), is based on the following assumption.

Assumption 14 (Bounded difference assumption)**.** *Let A be some set; a function $f : A^n \to \mathbb{R}$ satisfies the* bounded difference assumption*, if there exist real numbers $c_1, \ldots, c_n > 0$ so that for all $i = 1, \ldots, n$,*

$$\sup_{z_1,\ldots,z_n,z_i' \in A} |f(z_1, \ldots, z_n) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c_i \,.$$

In words, we assume that if we change the ith variable of g while keeping all the others fixed, then the value of the function does not change by more than c_i. The following theorem is powerful generalization of Höffding's inequality.

Theorem 15 (McDiarmid's inequality)**.** *Under the bounded difference assumption, i.e., Assumption 2.2, it holds, for all $t > 0$,*

$$\mathbb{P}(|f(Z_1, \ldots, Z_n) - \mathbb{E}f(Z_1, \ldots, Z_n)| \geq t) \leq 2e^{-2t^2 / \sum_{i=1}^n c_i^2} \,.$$

Proof. The proof is a clever martingale analogue of the proof of Höffding's inequality. In the proof, we use the shorthand

$$f \equiv f(Z_1, \ldots, Z_n) \,.$$

<u>Hint :</u> *A* martinale *is a family of random varibles $(Z_i)_{i\in\mathbb{N}}$ such that $\mathbb{E}[Z_i|Z_1, \ldots, Z_{i-1}] = Z_i$.*

To bound $V := f - \mathbb{E}f$, we write $V = \sum_{i=1}^n V_i$, where

$$V_i := \mathbb{E}[f|Z_1, \ldots, Z_i] - \mathbb{E}[f|Z_1, \ldots, Z_{i-1}] \,,$$

where $\mathbb{E}[f|Z_1, \ldots, Z_i]$ denotes the <u>conditional</u> expectation operator (which itself is a random variable depending on the draw of $Z_1, ..., Z_i$). Let us now think of the variables $Z_1, \ldots, Z_{i-1}$ as being fixed. Then, by the bounded difference assumption, changing the value of Z_i can change the value of V_i by at most c_i. Furthermore, $\mathbb{E}[V_i|Z_1, \ldots, Z_{i-1}] = 0$. Thus, by Höffding's lemma,

$$\mathbb{E}[e^{sV_i}|Z_1, \ldots, Z_{i-1}] \leq e^{s^2 c_i^2/8} \,. \tag{11}$$

Hence, by Chernoff's inequality,

$$\begin{aligned}
&\mathbb{P}(f - \mathbb{E}f \geq t) \\
&\quad \leq \min_{s\in\mathbb{R}} e^{-st}\mathbb{E}e^{s(f-\mathbb{E}f)} = \min_{s\in\mathbb{R}} e^{-st}\mathbb{E}e^{s\sum_{i=1}^n V_i} \\
&\quad = \min_{s\in\mathbb{R}} e^{-st}\mathbb{E}\,\mathbb{E}[e^{s\sum_{i=1}^n V_i}|Z_1,\ldots,Z_{n-1}] \\
&\quad = \min_{s\in\mathbb{R}} e^{-st}\mathbb{E}\,\mathbb{E}[e^{s\sum_{i=1}^{n-1} V_i}\mathbb{E}[e^{sV_n}|Z_1,\ldots,Z_{n-1}]|Z_1,\ldots,Z_{n-1}] \\
&\quad \overset{(11)}{\leq} \min_{s\in\mathbb{R}} e^{s^2c_i^2/8-st}\mathbb{E}\,\mathbb{E}[e^{s\sum_{i=1}^{n-1} V_i}|Z_1,\ldots,Z_{n-1}] \\
&\quad \leq \quad \ldots \qquad (\text{REPEATING THE ARGUMENT } (n-1) \text{ TIMES}) \\
&\quad \leq \min_{s\in\mathbb{R}} e^{ns^2\sum_{i=1}^n c_i^2/8-st},
\end{aligned}$$

which is minimized for $s := 4t/\sum_{i=1}^n c_i^2$, giving

$$\mathbb{P}(f - \mathbb{E}f \geq t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

Analogously, repeating the argument for the function $-f$, we obtain the corresponding left-sided inequality

$$\mathbb{P}(f - \mathbb{E}f \leq -t) = \mathbb{P}(-f - \mathbb{E}(-f) \geq t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2}.$$

Combining both results gives the claimed result. □

Corollary 16. *Let $\mathcal{C}$ be a class of functions. Then, for any $t > 0$,*

$$P\left(\left|\sup_{g\in\mathcal{C}}|\hat{L}_n(g) - L(g)| - \mathbb{E}\sup_{g\in\mathcal{C}}|\hat{L}_n(g) - L(g)|\right| \geq t\right) \leq 2e^{-2nt^2}.$$

Proof. Put $Z_i := (X_i, Y_i)$, $i \in \mathbb{N}$, and $f(Z_1,\ldots,Z_n) := \sup_{g\in\mathcal{C}}|\hat{L}_n(g) - L(g)|$. Then f satisfies the bounded difference assumption with $c_i = 1/n$ for all $n \in \mathbb{N}$. The claimed inequality thus follows from McDiarmid's inequality. □

2.3 The Big Picture

Recall that our overall aim is to bound $\sup_{g\in\mathcal{C}}\left|\hat{L}_n(g) - L(g)\right|$, since, by Lemma 1, this would imply a bound on the estimation error of empirical risk minimization. By the decomposition (3), we can achieve this goal in two steps:

1. bounding the deviation of $\sup_{g\in\mathcal{C}} \left|\hat{L}_n(g) - L(g)\right|$ from its expected value (STEP 1)
2. bounding $\mathbb{E}\sup_{g\in\mathcal{C}} \left|\hat{L}_n(g) - L(g)\right|$ (STEP 2).

As Corollary 16 shows, Step 1 is achieved by a simple application of McDiarmid's inequality (and the resulting bound is also on the right, exponential order)! However, to prove this sophisticated inequality, we had to go quite a long way via Markov, Chernoff, and Höffding's inequalities. Anyway, what we are left with at this point is just Step 2, that is, bounding the expected value $\mathbb{E}\sup_{g\in\mathcal{C}} \left|\hat{L}_n(g) - L(g)\right|$. To do so, we now introduce Vapnik-Chervonenkis theory.

3 Vapnik-Chervonenkis Theory

In order to bound the performance of ERM, we are left with bounding $\mathbb{E}\sup_{g\in\mathcal{C}} \left|\hat{L}_n(g) - L(g)\right|$. To this end, we proceed in three steps:

1. relating $\mathbb{E}\sup_{g\in\mathcal{C}} \left|\hat{L}_n(g) - L(g)\right|$ with $\mathfrak{R}_n(\mathcal{C})$, the so-called *Rademacher complexity* of the class $\mathcal{C}$
2. relating $\mathfrak{R}_n(\mathcal{C})$ with the so-called *VC shattering coefficient* $\mathbb{S}_n(\mathcal{C})$
3. relating $\mathbb{S}_n(\mathcal{C})$ with the *VC dimension* V
4. computing V for specific classes $\mathcal{C}$.

We start our analysis with step 1, i.e., introducing the Rademacher complexity.

Definition 17. *(*RADEMACHER COMPLEXITY*) The* (empirical) Rademacher complexity *of a function class $\mathcal{C}$ is defined as*

$$\mathfrak{R}_n(\mathcal{C}) := \mathbb{E}_{\boldsymbol{\varsigma}} \sup_{g\in\mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^{n} \varsigma_i \mathbb{I}_{\{g(X_i)\neq Y_i\}} \right| ,$$

where $\boldsymbol{\varsigma} = (\varsigma_i)_{i=1,\dots,n}$ is an i.i.d. family of Rademacher variables*, i.e., $\mathbb{P}(\varsigma_i = +1) = \mathbb{P}(\varsigma_i = -1)$.*

The Rademacher complexity, intuitively, measures how well the empirical error can, when optimized over $g \in \mathcal{C}$, match with random signs.

Proposition 18. *Let $\mathcal{C}$ be a class of functions. Then*

$$\mathbb{E} \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right| \leq 2\mathbb{E}_S \mathfrak{R}_n(\mathcal{C}) \,.$$

Proof. The core idea of the proof is to introduce $X'_1, \ldots, X'_n$ and $Y'_1, \ldots, Y'_n$, an independent copy of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, respectively (called *ghost sample*), as well as $\boldsymbol{\varsigma} = (\varsigma_i)_{i=1}^n$, an i.i.d. family of *Rademacher variables* that are independent of the sample and the ghost sample. Then, denoting $\hat{L}'_n(g) := \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{g(X'_i) \neq Y'_i\}}$, we have

<u>Hint :</u> *Jensen's inequality states that, for any random variable Z and any convex function $\varphi : \mathbb{R} \to \mathbb{R}$, it holds $\varphi(\mathbb{E}Z) \leq \mathbb{E}\varphi(Z)$. An example of a convex function is the absolute value function $|\cdot| : z \mapsto |z|$. Furthermore, if, for all $i \in I$, h_i is a convex function, then the pointwise supremum $z \mapsto \sup_{i \in I} h_i(z)$ is convex as well.*

$$\begin{aligned}
&\mathbb{E} \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right| \\
&\quad = \mathbb{E}_S \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - \mathbb{E}_{S'} \hat{L}'_n(g) \right| \\
&\quad \leq \mathbb{E}_S \mathbb{E}_{S'} \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - \hat{L}'_n(g) \right| \\
&\qquad \text{(by Jensen's inequality)} \\
&\quad = \mathbb{E}_S \mathbb{E}_{S'} \mathbb{E}_{\boldsymbol{\varsigma}} \sup_{g \in \mathcal{C}} \left| \sum_{i=1}^n \varsigma_i \left(\mathbb{I}_{\{g(X'_i) \neq Y'_i\}} - \mathbb{I}_{\{g(X'_i) \neq Y'_i\}} \right) \right| \\
&\qquad \text{(by the symmetry of the Rademacher variables)} \\
&\quad \leq 2\mathbb{E}_S \underbrace{\mathbb{E}_{\boldsymbol{\varsigma}} \sup_{g \in \mathcal{C}} \left| \frac{1}{n} \sum_{i=1}^n \varsigma_i \mathbb{I}_{\{g(X_i) \neq Y_i\}} \right|}_{= \mathfrak{R}_n(\mathcal{C})} ,
\end{aligned} \tag{12}$$

which was to show. □

We now proceed with step 2, i.e., introducing the VC shatter coefficient and relating it to the Rademacher complexity.

Definition 19. *(*VC shatter coefficient*) The* VC shatter coefficient *of a function class $\mathcal{C}$ is defined as*

$$\mathbb{S}_n(\mathcal{C}) = \max_{x_i \in \mathbb{R}^d,\, y_i \in \mathbb{R},\, i=1,\ldots,n} \left| \left\{ \left(\mathbb{I}_{\{g(x_1) \neq y_1\}}, \ldots, \mathbb{I}_{\{g(x_n) \neq y_n\}} \right) : g \in \mathcal{C} \right\} \right| .$$

Let the above maximum be attained for certain points $x_1, \ldots, x_n$. Then, we say $\mathcal{C}$ shatters *$x_1, \ldots, x_n$ if and only if $\mathbb{S}_n(\mathcal{C}) = 2^n$.*

The core idea in the proof of the following classical result by Vapnik and Chervonenkis (1971) is that the supremum in the definition of the Rademacher complexity essentially depends is only over $\mathbb{S}_n(\mathcal{C})$ many elements (even if $\mathcal{C}$ has infinite cardinality)!

Theorem 20 (Vapnik-Chervonenkis inequality)**.** *Let $\mathcal{C}$ be a class of functions. Then*

$$\mathfrak{R}_n(\mathcal{C}) \leq \sqrt{\frac{2\log(2\,\mathbb{S}_n(\mathcal{C}))}{n}}\,.$$

Proof. The core idea consists in investigating $\mathfrak{R}_n(\mathcal{C})$ for fixed values of the variables $X_i, Y_i, i = 1, \ldots, n$, i.e., $\mathfrak{R}_n(\mathcal{C})$ only randomly depending on $\varsigma_1, \ldots, \varsigma_n$. To this end, note that $\varsigma_i \mathbb{I}_{\{g(X_i)\neq Y_i\}}$, $i = 1, \ldots, n$ has zero mean and ranges in $[-1, 1]$. Thus, by Höffding's Lemma, i.e., Lemma 11, it is sub-Gaussian with unit variance, i.e., $\mathbb{E}e^{s\varsigma_i \mathbb{I}_{\{g(X_i)\neq Y_i\}}} \leq e^{s^2/2}$. Clearly it follows

$$\mathbb{E}e^{\frac{s}{n}\sum_i \varsigma_i \mathbb{I}_{\{g(X_i)\neq Y_i\}}} = \prod_{i=1}^{n} \mathbb{E}e^{\frac{s}{n}\varsigma_i \mathbb{I}_{\{g(X_i)\neq Y_i\}}} \leq \prod_{i=1}^{n} e^{\frac{s^2}{2n^2}} = e^{\frac{s^2}{2n}}\,,$$

i.e., the variable $\frac{1}{n}\sum_i \varsigma_i \mathbb{I}_{\{g(X_i)\neq Y_i\}}$ is sub-Gaussian with parameter $\sigma^2 = 1/n$. Hence, by Lemma 22,

$$\mathfrak{R}_n(\mathcal{C}) \leq \sqrt{\frac{2\log(2\mathbb{S}_n(\mathcal{C}))}{n}}$$

because, for fixed X_i, Y_i, $i = 1, \ldots, n$, the sup in the definition of $\mathfrak{R}_n(\mathcal{C})$ is effectively only over $\mathbb{S}_n(\mathcal{C})$ many values. □

The above proof builds on the following useful lemma from probability theory.

Definition 21. *A random variable Z is called* sub-Gaussian with parameter σ^2*, if its moment-generating function can be bounded by the one of a Gaussian random variable with variance σ^2, i.e., $M_Z(s) = \mathbb{E}e^{sZ} \leq e^{\frac{\sigma^2 s^2}{2}}$.*

Lemma 22. *Let $n > 1$ and $Z_1, \ldots, Z_n$ be sub-Gaussian with parameter σ^2. Then*

$$\mathbb{E}\max_{i=1,\ldots,n} |Z_i| \leq \sigma\sqrt{2\log(2n)}\,.$$

Proof. By Jensen's inequality,

$$e^{s\mathbb{E}\max_{i=1,\dots,n} Z_i} \overset{\text{JENSEN}}{\leq} \mathbb{E}e^{s\max_{i=1,\dots,n} Z_i} = \mathbb{E}\max_{i=1,\dots,n} e^{sZ_i}$$
$$\leq \sum_{i=1}^{n} \mathbb{E}e^{sZ_i} \leq ne^{s^2\sigma^2/2}.$$

Thus, $\mathbb{E}\max_{i=1,\dots,n} Z_i \leq \log(n)/s + s\sigma^2/2$, which is minimized for $s := \sqrt{2\log(n)/\sigma^2}$. Resubstitution gives

$$\mathbb{E}\max_{i=1,\dots,n} Z_i \leq \sigma\sqrt{2\log(n)}. \tag{13}$$

Finally, note that

$$\max_{i=1,\dots,n} |Z_i| = \max(Z_1, -Z_1, \dots, Z_n, -Z_n).$$

Thus, by (13),

$$\max_{i=1,\dots,n} |Z_i| = \max(Z_1, -Z_1, \dots, Z_n, -Z_n) \leq \sigma\sqrt{2\log(2n)}.$$

□

Observe that, in order for the Vapnik-Chervonenkis inequality to converge to zero for $n \to \infty$, the quantity $\log(\mathbb{S}_n(\mathcal{C}))$ needs to decrease sublinearly in n. This motivates the following definition.

Definition 23. *The* V-C dimension V *is the smallest integer* n *such that* $\mathbb{S}_n(\mathcal{C}) = 2^n$.

An interesting phase transition occurs for the VC shattering coefficient $\mathbb{S}_n(\mathcal{C})$ when $n > V$, as the following classical lemma, proved by Sauer (1972), shows.

Lemma 24 (Sauer's lemma)**.** *For any* $n > V$,

$$\mathbb{S}_n(\mathcal{C}) \leq (n+1)^V.$$

Proof. Fix the variables $x_i, y_i, i = 1, \dots, n$ and consider the resulting table of values $\left\{\left(\mathbb{I}_{\{g(x_1)\neq y_1\}}, \dots, \mathbb{I}_{\{g(x_n)\neq y_n\}}\right) : g \in \mathcal{C}\right\}$. E.g., for $n = 5$, this could look as follows: Each row corresponds to one possible evaluation of a

$T :=$

	x_1	x_2	x_3	x_4	x_5
g_1	0	1	0	1	1
g_2	1	0	0	1	1
g_3	1	1	1	0	1
g_4	0	1	1	0	0
g_5	0	0	0	1	0

function in $\mathcal{C}$ on the sample, and the cardinality

$$\left|\left\{\left(\mathbb{I}_{\{g(x_1)\neq y_1\}}, \ldots, \mathbb{I}_{\{g(x_n)\neq y_n\}}\right) : g \in \mathcal{C}\right\}\right|$$

equals the number of rows.

We translate the table by *shifting*, for each $i = 1, \ldots, n$, column i, that is, for each row, we replace a 1 in column i by a 0, unless this would produce a row that is already contained in the table.

After applying the shifting operation in order from x_1 to x_n, we get the following table, which contains not so many 1s anymore. From the example,

$T^* :=$

	x_1	x_2	x_3	x_4	x_5
g_1	0	1	0	0	0
g_2	0	0	0	1	1
g_3	0	0	0	0	1
g_4	0	0	0	0	0
g_5	0	0	0	0	0

we can make the following observations:

1. The size of the table is unchanged because the rows are still distinct.
2. The shifted table T^* exhibits the is *closed below*, i.e., replacing any of the 1 in the table would produce a duplicate row in the table.
3. The VC dimension of the original table is at least as high as the one of the shifted table, i.e., $\mathrm{VC}(T) \geq \mathrm{VC}(T^*)$. To see this, consider a subset of columns that is shattered in T^*; the same subset must also be shattered in T.

We conclude from 2. and 3. that T^* cannot have more than V 1s in a row and thus has $\leq \sum_{i=0}^{n} \binom{n}{i}$ rows (imagine assigning, for each $i = 0, \ldots, V$, i many 1s to the positions $1, \ldots, n$) and, by 1., the same holds for T.

Moreover, by the binomial theorem,

$$\sum_{i=0}^{V} \binom{n}{i} = \sum_{i=0}^{V} \frac{n!}{((n-i)!i!} \leq \sum_{i=0}^{n} \frac{n^i}{i!}$$
$$\leq \sum_{i=0}^{V} \frac{n^i V!}{i!(V-i)!} = \sum_{i=0}^{V} n\binom{V}{i} \overset{\text{Bin.}}{=} (n+1)^V$$

□

Thus, we obtain the following bound

$$\mathbb{E} \sup_{g \in \mathcal{C}} \left| \hat{L}_n(g) - L(g) \right| \leq 2\sqrt{\frac{2(V \log(n+1) + \log 2)}{n}}, \tag{14}$$

which we can combine with (3) and (12), obtaining the following main result.

Corollary 25. *With probability* $1 - \delta$*,*

$$\sup_{g \in \mathcal{C}} \left| \hat{L}(g) - L(g) \right| \leq \sqrt{\frac{\log(1/\delta)}{2n}} + 2\sqrt{\frac{2(V \log(n+1) + \log 2)}{n}}$$

Proof. The result is obtained from (3), (12), and (14) by setting $\epsilon := \sqrt{\frac{\log(2/\delta)}{2n}}$. □

It just remains to compute the VC dimension for specific classes. For example, since any three points in the plane can be separated, in any label configuration, by a simple linear function, the VC dimension of linear functions in $\mathbb{R}^2$ is equal to 3. More generally, the following result holds.

Proposition 26. *The VC dimension of the class of linear function in* $\mathbb{R}^d$ *is* $V = d + 1$*.*

Proof. For any non-colinear set of points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and any choice of labels $y_1, \ldots, y_n \in \{0, 1\}$, there is an affine-linear function, separating the two classes without any error, if and only if $n = d + 1$. Thus $V = d + 1$. □

In combination with Corollary 25, we immediately obtain a performance bound for ERM using linear functions:

Corollary 27. *The estimation error of ERM with linear functions in $\mathbb{R}^d$, is, with probability $1-\delta$, bounded by*

$$L(g_n^*) - L(g^*) \;\leq\; 2\sqrt{\frac{\log(1/\delta)}{2n}} + 4\sqrt{\frac{2(d+1)\log(n+1)+2\log 2)}{n}}\,.$$

4 Conclusion

Recall the basic decomposition from the beginning of this lecture:

$$L(g_n^*) - L(g^*) = \underbrace{L(g_n^*) - L(g_{\mathcal{C}}^*)}_{\text{“estimation error”}} + \underbrace{L(g_{\mathcal{C}}^*) - L(g^*)}_{\text{“approximation error”}}\,.$$

The approximation error is, in general, not controllable. However, we have just shown that the *estimation error* converges to zero at a rate of $O(\sqrt{V/n})$, where V is the VC dimension of the class $\mathcal{C}$ and n is the sample size.

We observe that, regarding the choice of the class $\mathcal{C}$, there is trade-off between estimation and approximation error: the larger the size of the class, the smaller the approximation error, but also the higher the VC dimension and thus the estimation error.

A Further Proofs

*Proof of Lemma 11 (*Höffding's Lemma*).* Since $z \mapsto e^{sz}$ is a convex function, we have, for all $a \leq x \leq b$,

$$e^{sx} \leq \frac{b-z}{b-a}e^{sa} + \frac{z-a}{b-a}e^{sb}$$

and thus

$$\mathbb{E}\left[e^{sZ}\right] \leq \frac{b-\mathbb{E}Z}{b-a}e^{sa} + \frac{\mathbb{E}Z-a}{b-a}e^{sb}\,.$$

Put $h = s(b-a)$, $p = \frac{-a}{b-a}$ and $L(h) = -hp + ln(1-p+pe^h)$. Then, since $\mathbb{E}Z = 0$,

$$\frac{b-\mathbb{E}Z}{b-a}e^{sa} + \frac{\mathbb{E}Z-a}{b-a}e^{sb} = e^{L(h)}\,.$$

Taking derivative of $L(h)$, it holds $L(0) = L'(0) = 0$ and $L''(h) \leq \frac{1}{4}$. Hence, by Tayor's expansion,

$$L(h) \leq \frac{1}{8}h^2 = \frac{1}{8}s^2(b-a)^2\,,$$

i.e., $\mathbb{E}\left[e^{sZ}\right] \leq e^{\frac{1}{8}s^2(b-a)^2}$. $\square$

References

H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sums of observations. *Annals of Mathematical Statistics*, 23: 409–507, 1952.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989.

N. Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13 (1):145–147, 1972.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.