ETH Zurich

Empirical Risk Minimization

Author:
Peter Hinz

Professor:
Prof. Sara van de Geer

June 25, 2016

Contents

1 Introduction **2**
- 1.1 Risk minimization problem . . . 2
- 1.2 Empirical Risk Minimization . . . 2
- 1.3 Empirical and Rademacher Processes . . . 3

2 Symmetrization Inequalities **3**

3 Bounded Difference Inequality **6**
- 3.1 Definitions and Proof . . . 6
- 3.2 Applications . . . 8
  - 3.2.1 Hoeffding's Inequality . . . 8
  - 3.2.2 Glivenko Cantelli Theorem . . . 9

4 Vapnik Chernovenkis Theory **13**
- 4.1 Shattering and Vapnik-Chernovenkis classes . . . 14
- 4.2 Examples . . . 15
  - 4.2.1 The class of half lines . . . 15
  - 4.2.2 The class of intervals . . . 15
- 4.3 The Vapnik Chernovenkis Inequality . . . 16
- 4.4 Selection of 0-1-Classifiers . . . 17
  - 4.4.1 Notation and Definitions . . . 17
  - 4.4.2 Interesting Theorems . . . 19

A Essential Supremum and Essential Infimum **20**

B A bound on expectation for exponentially decaying tails **20**

C Notation **21**

1 Introduction

Let a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable space $(\mathcal{S}, \mathcal{A})$ be given. Let $X_1, \ldots, X_n$ be IID random variables from Ω to $\mathcal{S}$, each with distribution P on $(\mathcal{S}, \mathcal{A})$.

Notation For a measurable $f : \mathcal{S} \to \mathbb{R}$ and a probability measure Q on $(\mathcal{S}, \mathcal{A})$, we define $Qf := \mathbb{E}_Q[f] := \int_{\mathcal{S}} f(s)\, dQ(s)$ if the expectation exists.

1.1 Risk minimization problem

Let $\mathscr{F}$ be a set of Borel-measurable functions from $(\mathcal{S}, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ which are integrable with respect to the distribution P of the $(X_i)_{i\in\mathbb{N}}$. The functions $f \in \mathscr{F}$ represent losses for each value $s \in \mathcal{S}$. The goal is to find a minimizer of the function

$$\Phi : \begin{cases} \mathscr{F} & \to \mathbb{R} \\ f & \mapsto Pf \end{cases},$$

assuming that such a minimizer exists. It can be seen as a weighted loss with weights given by P. For known P, this is a classical minimization problem, but here we assume that P is unknown. The only information we have are n observations $x_1, \ldots, x_n$ from the random variables $X_1, \ldots, X_n$. Since the $(X_i)_{i\in\mathbb{N}}$ are independent and all have the distribution P, $(x_1, \ldots, x_n)$ can be seen as one sample from P^n or as n independent samples from P. This is the only information we have on the distribution P and the standard approach to find a minimizer of Φ is to find a minimizer of the empirical risk.

1.2 Empirical Risk Minimization

Definition 1 (empirical distribution)**.** For $n \in \mathbb{N}$, the empirical distribution $\hat{P}_n$ is defined as

$$\hat{P}_n : \begin{cases} (\Omega, \mathcal{A}) & \to [0,1] \\ (\omega, A) & \mapsto \frac{1}{n}\sum_{k=1}^n \mathbb{1}_A(X_k(\omega)) \end{cases}.$$

For $\omega \in \Omega$, $\hat{P}_n(\omega, \cdot)$ is the probability measure giving mass $\frac{1}{n}$ on every $X_i(\omega)$, $i \in \{1, \ldots, n\}$. The empirical distribution is a stochastic kernel.

Theorem 2. *For $n \in \mathbb{N}$, $\hat{P}_n$ is a stochastic kernel.*

Proof. Just check the two properties for a stochastic kernel

- For $A \in \mathcal{A}$, $\omega \mapsto \frac{1}{n}\sum_{k=1}^n \mathbb{1}_A(X_k(\omega)) = \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{X_k^{-1}(A)}(\omega)$ is $(\mathcal{F}/\mathcal{B}(\mathbb{R}))$-measurable since the $(X_i)_{i\in\mathbb{N}}$ are measurable.
- For $\omega \in \Omega$, $A \mapsto \frac{1}{n}\sum_{k=1}^n \mathbb{1}_A(X_k(\omega)) = \frac{1}{n}\sum_{k=1}^n \delta_{\{X_k(\omega)\}}(A)$ is a probability measure

□

With this definition, one can formulate a different minimization problem. Since P is unknown, one uses the n observations $(x_1, \ldots, x_n) = (X_1(\omega), \ldots, X_n(\omega))$ for some $\omega \in \Omega$. Instead or P itself, $\hat{P}_n(\omega, \cdot)$ is used for risk minimization. So the goal of empirical risk minimization is to find a minimizer of the function

$$\Theta : \begin{cases} \mathscr{F} & \to \mathbb{R} \\ f & \mapsto \hat{P}_n(\omega, \cdot)f = \frac{1}{n}\sum_{k=1}^n f(x_k) \end{cases},$$

assuming that a minimizer exists. Note that each f is evaluated at most at n different points which implies that the empirical expectation always exists, in other words, each f is automatically integrable with respect to $\hat{P}_n(\omega, \cdot)$.

The idea behind empirical risk minimization is that one hopes that for $f \in \mathscr{F}$ the property of having a "close to minimal $\Theta(f)$ value" implies having a "close to minimal $\Phi(f)$ value". In order to relate the true and the empirical risk minimization problem, one tries to find bounds on

$$\|\hat{P}_n - P\|_{\mathscr{F}} := \sup_{f \in \mathscr{F}} |\hat{P}_n f - Pf|,$$

for example in probability or for the expectation. For a finite or countable $\mathscr{F}$, we obviously get measurability, but in general, this is a problem which has to be solved in some way, certain measurability assumptions are required, see [2], section 2. These bounds exploit the structure of $\mathscr{F}$ and P.

1.3 Empirical and Rademacher Processes

The empirical process

$$Z_n := \sqrt{n}(\hat{P}_n - P), \quad n \in \mathbb{N}$$

can be seen as a random measure, but often, each Z_n is seen as a stochastic process, indexed by the function class $\mathscr{F}$.

Definition 3. For $n \in \mathbb{N}$, define the stochastic process $(Z_n(f))_{f \in \mathscr{F}}$ by

$$Z_n(f) : \begin{cases} \Omega & \to \mathbb{R} \\ \omega & \mapsto \frac{1}{\sqrt{n}}(\hat{P}_n(\omega, \cdot)f - Pf) \end{cases}$$

A random variable, which has distribution $\frac{1}{2}(\delta_{\{-1\}} + \delta_{\{1\}})$ is called a *Rademacher random variable*. Without loss of generality assume that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is sufficiently large to allow for countably many Rademacher random variables $(\varepsilon_i)_{i \in \mathbb{N}}$ from Ω to $\mathbb{R}$ which are independent of each other and independent of the X_i, $i \in \mathbb{N}$.

Definition 4. For $n \in \mathbb{N}$, define the Rademacher process $(R_n(f))_{f \in \mathscr{F}}$ by

$$R_n(f) : \begin{cases} \Omega & \to \mathbb{R} \\ \omega & \mapsto \frac{1}{n} \sum_{k=1}^{n} \varepsilon_k(\omega) f(X_k(\omega)) \end{cases}$$

As above, define

$$\|R_n\|_{\mathscr{F}} := \sup_{f \in \mathscr{F}} |R_n|$$

Here we assume again that $\|R_n\|_{\mathscr{F}}$ is measurable. The Rademacher process is an important tool for the investigation of Z_n because of the following symmetrization inequalities

2 Symmetrization Inequalities

The following theorem is adapted from [2].

Theorem 5 (Symmetrization Inequality)**.** *For any class $\mathscr{F}$ of P-integrable functions and for any convex increasing function $\Phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$*

$$\mathbb{E}\left[\Phi\left(\|\hat{P}_n - P\|_{\mathscr{F}}\right)\right] \leq \mathbb{E}\left[\Phi\left(2\|R_n\|_{\mathscr{F}}\right)\right]$$

Proof. Firs let us introduce some notation. We will identify $\sigma \in \{-1,1\}^n$ with

$$\sigma = (\sigma_1, \ldots, \sigma_n), \text{ where } \sigma_i \in \{-1,1\} \text{ for } i \in \{1,\ldots,n\}.$$

In the same way identify $s \in \mathcal{S}^n$ with

$$s = (s_1, \ldots, s_n), \text{ where } s_i \in \mathcal{S} \text{ for } i \in \{1,\ldots,n\}.$$

Next, observe that if μ is the uniform measure on $\{-1,1\}^n$, we have

- μ^n is the joint distribution of $(\varepsilon_1, \ldots, \varepsilon_n)$ defined above.
- P^n is the joint distribution of $(X_1, \ldots, X_n)$.
- $\mu^n \otimes P^n$ is the joint distribution of $(\varepsilon_1, \ldots, \varepsilon_n, X_1, \ldots, X_n)$ because of independence.

Now it follows

$$\begin{aligned}
& \mathbb{E}\left[\Phi\left(\|\hat{P}_n - P\|_{\mathscr{F}}\right)\right] \\
= & \int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n f(s_i) - Pf|\right) dP^n(s) \\
= & \int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \left(f(s_i) - Pf\right)|\right) dP^n(s) \\
= & \int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \left(f(s_i) - \int_{\mathcal{S}} f(s_i')dP(s_i')\right)|\right) dP^n(s) \\
= & \int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \left(f(s_i) - \int_{\mathcal{S}^n} f(s_i')dP^n(s')\right)|\right) dP^n(s) \\
= & \int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\int_{\mathcal{S}^n} \frac{1}{n}\sum_{i=1}^n \left(f(s_i) - f(s_i')\right) dP^n(s')|\right) dP^n(s) \\
\overset{\Phi \text{ increasing}}{\leq} & \int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} \int_{\mathcal{S}^n} |\frac{1}{n}\sum_{i=1}^n \left(f(s_i) - f(s_i')\right)| dP^n(s')\right) dP^n(s) \\
\overset{\Phi \text{ increasing}}{\leq} & \int_{\mathcal{S}^n} \Phi\left(\int_{\mathcal{S}^n} \sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \left(f(s_i) - f(s_i')\right)| dP^n(s')\right) dP^n(s) \\
\overset{\Phi \text{ convex}}{\leq} & \int_{\mathcal{S}^n}\int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \left(f(s_i) - f(s_i')\right)|\right) dP^n(s')dP^n(s)
\end{aligned}$$

Now observe that for each $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{-1,1\}^n$ it holds

$$\begin{aligned}
& \int_{\mathcal{S}^n}\int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \left(f(s_i) - f(s_i')\right)|\right) dP^n(s')dP^n(s) \\
= & \int_{\mathcal{S}^n}\int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i\left(f(s_i) - f(s_i')\right)|\right) dP^n(s')dP^n(s)
\end{aligned}$$

because of the symmetric structure of the integral. Therefore, we can write

$$
\begin{aligned}
& \mathbb{E}\left[\Phi\left(\|\hat{P}_n - P\|_{\mathscr{F}}\right)\right] \\
\leq\ & \int_{\mathcal{S}^n}\int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \left(f(s_i) - f(s_i')\right)|\right) dP^n(s')dP^n(s) \\
=\ & \frac{1}{2^n}\sum_{\sigma\in\{-1,1\}^n}\int_{\mathcal{S}^n}\int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i\left(f(s_i) - f(s_i')\right)|\right) dP^n(s')dP^n(s) \\
=\ & \int_{\{-1,1\}^n}\int_{\mathcal{S}^n}\int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i\left(f(s_i) - f(s_i')\right)|\right) dP^n(s')dP^n(s)d\mu(\sigma) \\
\overset{\Phi \text{ increasing}}{\leq}\ & \int_{\{-1,1\}^n}\int_{\mathcal{S}^n}\int_{\mathcal{S}^n} \Phi\left(\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(s_i)| + \sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(s_i')|\right) dP^n(s')dP^n(s)d\mu(\sigma) \\
\overset{\Phi \text{ convex}}{\leq}\ & \int_{\{-1,1\}^n}\int_{\mathcal{S}^n}\int_{\mathcal{S}^n} \frac{1}{2}\Phi\left(2\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(s_i)|\right) + \\
& \frac{1}{2}\Phi\left(2\sup_{f\in} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(s_i')|\right) dP^n(s')dP^n(s)d\mu(\sigma) \\
=\ & \frac{1}{2}\int_{\{-1,1\}^n}\int_{\mathcal{S}^n} \Phi\left(2\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(s_i)|\right) dP^n(s)d\mu(\sigma) + \\
& \frac{1}{2}\int_{\{-1,1\}^n}\int_{\mathcal{S}^n} \Phi\left(2\sup_{f\in} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(s_i')|\right) dP^n(s')d\mu(\sigma) \\
=\ & \int_{\{-1,1\}^n}\int_{\mathcal{S}^n} \Phi\left(2\sup_{f\in\mathscr{F}} |\frac{1}{n}\sum_{i=1}^n \sigma_i f(s_i)|\right) dP^n(s)d\mu(\sigma) \\
=\ & \mathbb{E}\left[\Phi\left(2\|R_n\|_{\mathscr{F}}\right)\right]
\end{aligned}
$$

□

There is an inequality from below for $\mathbb{E}\left[\Phi\left(\|\hat{P}_n - P\|_{\mathscr{F}}\right)\right]$, see [2]:

Theorem 6 (Desymmetrization inequality)**.** *For any class $\mathscr{F}$ of P-integrable functions and for any convex increasing function $\Phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$*

$$\mathbb{E}\left[\Phi(\frac{1}{2}\|R_n\|_{\mathscr{F}_c})\right] \leq \mathbb{E}\left[\Phi\left(\|\hat{P}_n - P\|_{\mathscr{F}}\right)\right]$$

where $\mathscr{F}_c := \{f - Pf | f \in \mathscr{F}\}$ are the centered functions of $\mathscr{F}$.

In particular, it follows from Theorems 5 and 6 with $\Phi = \mathrm{Id}$ that

$$\frac{1}{2}\mathbb{E}\left[\|R_n\|_{\mathscr{F}_c}\right] \leq \mathbb{E}\left[\|\hat{P}_n - P\|_{\mathscr{F}}\right] \leq 2\mathbb{E}\left[\|R_n\|_{\mathscr{F}}\right]$$

These inequalities show why it makes sense to study $\mathbb{E}\left[\|R_n\|_{\mathscr{F}}\right]$, the expectation of sup-norms of Rademacher processes.

3 Bounded Difference Inequality

3.1 Definitions and Proof

The bounded difference inequality is a type of concentration inequality that holds under a condition defined below.

Definition 7. A function $g : \mathbb{R}^n \to \mathbb{R}$ is said to fulfill the *bounded difference condition* if it is $(\mathcal{B}(\mathbb{R})^n/\mathcal{B}(\mathbb{R}))$-measurable and allows constants $c_1, \ldots, c_i$ such that for all $i \in \{1, \ldots, n\}$

$$\sup_{x_1,\ldots,x_n,x_i' \in \mathbb{R}} |g(x_1, \ldots, x_{i-1}, x_i, x_{i-1}, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i.$$

The following inequality is an important ingredient for the Bounded Difference Inequality. It can be derived by a Taylor series argument.

Lemma 8. *For all $s > 0$, $a < b$ it holds*

$$-\frac{a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa} < e^{s^2/8}$$

We use this to prove the next result

Lemma 9. *Let V, Z be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Assume that Z is $(\mathcal{G}/\mathcal{B}(\mathbb{R}))$-measurable and that*

- $\mathbb{E}[V|\mathcal{G}] = 0$ *almost surely*
- *There is a measurable map $\psi : \mathbb{R} \to \mathbb{R}$ and a constant $c > 0$ such that almost surely $\psi(Y) \leq V \leq \psi(Z) + c$.*

Then

$$\forall s > 0 : \quad \mathbb{E}\left[e^{sV}|\mathcal{G}\right] \leq e^{\frac{s^2c^2}{8}} \text{ almost surely.}$$

Proof. Define the random variable

$$\alpha := \frac{\psi(Z) + c - V}{c}.$$

Then we can write V as a convex combination of $\psi(Z)$ and $\psi(Z) + c$

$$V = \alpha\psi(Z) + (1-\alpha)(\psi(Z) + c)$$

Since almost surely $0 \leq \alpha \leq 1$, we can exploit the convexity of $v \mapsto e^{sv}$ to write

$$\begin{aligned} & \mathbb{E}\left[e^{sV}|\mathcal{G}\right] \\ \leq\ & \mathbb{E}\left[\alpha e^{s\psi(Z)} + (1-\alpha)e^{s(\psi(Z)+c)}|\mathcal{G}\right] \\ =\ & e^{s\psi(Z)}\mathbb{E}[\alpha|\mathcal{G}] + e^{s(\psi(Z)+c)}\mathbb{E}[1-\alpha|\mathcal{G}] \\ =\ & e^{s\psi(Z)}\frac{\psi(Z)+c}{c} - e^{s(\psi(Z)+c)}\frac{\psi(Z)}{c} \end{aligned}$$

almost surely. If we define the random variables $a := \psi(Z)$ and $b := \psi(Z) + c$, we can arrive at

$$\mathbb{E}\left[e^{sV}|\mathcal{G}\right] \leq -\frac{a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa}$$

almost surely and we can use Lemma 8 pointwise. □

We are now able to prove the Bounded Difference Inequality.

Theorem 10 (McDiarmid)**.** *Let $X_1, \dots, X_n$ be independent random variables. If g fulfills the bounded difference condition with constants $c_1, \dots, c_n$ then*

$$\mathbb{P}\left[g(X_1, \dots, X_n) - \mathbb{E}\left[g(X_1, \dots, X_n)\right] \geq \varepsilon\right] \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{k=1}^n c_i^2}\right).$$

Proof. Use the convention $\mathbb{E}\left[Y|X_1, \dots, X_{k-1}\right] := \mathbb{E}\left[Y\right]$ for $k = 0$. Define

$$V_i \quad := \quad \mathbb{E}\left[g(X_1, \dots, X_n)|X_1, \dots, X_i\right] - \mathbb{E}\left[g(X_1, \dots, X_n)|X_1, \dots, X_{i-1}\right] \text{ for } i \in \{1, \dots, n\}$$

Note the following two things

- $g(X_1, \dots, X_n) - \mathbb{E}\left[X_1, \dots, X_n\right] = \sum_{k=1}^n V_k$ almost surely
- $\mathbb{E}\left[V_k|\sigma(X_1, \dots, X_{k-1})\right] = 0$ almost surely for $k \in \{1, \dots, n\}$.

Let μ_k be the distribution of X_k for $k \in \{1, \dots, n\}$. Then $\mu_1 \otimes \dots \otimes \mu_n$ is the distribution of $(X_1, \dots, X_n)$ because of independence. Let

$$\begin{aligned} U_k &:= \operatorname*{ess\,inf}_{x_k \in \mathbb{R}} \int_{\mathbb{R}^{n-k}} g(X_1, \dots, X_{k-1}, x_k, x_{k+1}, \dots, x_n) d\mu_{k+1} \otimes \dots \otimes \mu_n(x_{k+1}, \dots, x_n) \\ Z_k &:= U_k - \mathbb{E}\left[g(X_1, \dots, X_n)|X_1, \dots, X_{k-1}\right] \end{aligned}$$

Here ess sup and ess inf denote the essential supremum and the essential infimum respectively, see Appendix A. Hence, U_k and W_k are $\sigma(X_1, \dots, X_{k-1})$ measurable[1]! This is a detail that was missing in many proofs I have read. Usually, just the infimum and the supremum is taken, which means that these proofs are all wrong. In order to fix the measurability issue, I came up with the idea to use the essential supremum and infimum.

Because of independence, we obviously have almost surely

$$\mathbb{E}\left[g(X_1, \dots, X_n)|X_1, \dots, X_{k-1}\right] = \int_{\mathbb{R}^{n-k+1}} g(X_1, \dots, X_{k-1}, x_k, \dots, x_n) d\mu_k \otimes \dots \otimes \mu_n(x_k, \dots, x_n)$$

If we introduce the notation

$$S_{k,x_k} := \int_{\mathbb{R}^{n-k}} g(X_1, \dots, X_{k-1}, x_k, \dots, x_n) d\mu_{k+1} \otimes \dots \otimes \mu_n(x_{k+1}, \dots, x_n)$$

Now, we can write

- $U_k = \operatorname*{ess\,inf}_{x_k \in \mathbb{R}} S_{k,x_k}$
- $\mathbb{E}\left[g(X_1, \dots, X_n)|X_1, \dots, X_k\right] = \int_{\mathbb{R}} S_{k,x_k} d\mu_k(x_k)$ almost surely

Therefore we have for all $k \in \{1, \dots, n\}$ and all $x_k \in \mathbb{R}$ by the bounded difference condition

$$\operatorname*{ess\,inf}_{x'_k \in \mathbb{R}} S_{k,x'_k} \leq S_{k,x_k} \leq \operatorname*{ess\,inf}_{x'_k \in \mathbb{R}} S_{k,x'_k} + c_k.$$

But this implies

$$U_k \leq \mathbb{E}\left[g(X_1, \dots, X_n)|X_1, \dots, X_k\right] \leq U_k + c_k$$

[1] Use the convention that $\sigma(X_1, \dots, X_{k-1})$ is the trivial σ-algebra for $k = 1$.

and therefore

$$Z_k \leq V_k \leq Z_k + c_k.$$

Now we can use Lemma 9 with $\mathcal{G} = \sigma(X_1, \ldots, X_{k-1})$, $V = V_k$, $Z = Z_k$, $c = c_k$ and $\psi : x \mapsto x$:

$$\forall s > 0 : \mathbb{E}\left[e^{sV_k} | X_1, \ldots, X_{k-1}\right] \leq e^{\frac{s^2 c_k^2}{8}} \tag{1}$$

We are now in the position to finish the proof. For any $s > 0$

$$\begin{aligned}
& \mathbb{P}\left[g(X_1, \ldots, X_n) - \mathbb{E}\left[g(X_1, \ldots, X_n) \geq \varepsilon\right]\right] \\
= \; & \mathbb{P}\left[\sum_{k=1}^{n} V_k \geq \varepsilon\right] \\
= \; & \mathbb{P}\left[\exp\left(s \sum_{k=1}^{n} V_k\right) \geq \exp\left(s\varepsilon\right)\right] \\
\leq \; & e^{-s\varepsilon} \mathbb{E}\left[e^{s \sum_{k=1}^{n} V_k}\right] \\
= \; & e^{-s\varepsilon} \mathbb{E}\left[e^{s \sum_{k=1}^{n-1} V_k} \mathbb{E}\left[e^{sV_n} | X_1, \ldots, X_{n-1}\right]\right] \\
\overset{(1)}{=} \; & e^{-s\varepsilon} e^{\frac{s^2 c_k^2}{8}} \mathbb{E}\left[e^{s \sum_{k=1}^{n-1} V_k}\right] = \ldots \\
= \; & e^{-s\varepsilon + \frac{s^2}{8} \sum_{k=1}^{n} c_k^2}
\end{aligned}$$

by iteratively using the tower property for conditional expectation and using equation (1). Taking $s = \frac{4\varepsilon}{\sum_{k=1}^{n} c_k^2}$ finishes the proof □

By using the above theorem with $-g$ instead of g, we get

$$\mathbb{P}\left[-g(X_1, \ldots, X_n) + \mathbb{E}\left[g(X_1, \ldots, X_n)\right] \geq \varepsilon\right] \leq \exp\left(-\frac{2e^w}{\sum_{k=1}^{n} c_k^2}\right)$$

which is equivalent to

$$\mathbb{P}\left[g(X_1, \ldots, X_n) - \mathbb{E}\left[g(X_1, \ldots, X_n)\right] \leq -\varepsilon\right] \leq \exp\left(-\frac{2e^w}{\sum_{k=1}^{n} c_k^2}\right)$$

Together with the original inequality of Theorem 10, we get a bound for the absolute value of the difference of $g(X_1, \ldots, X_n)$ from its expectation:

$$\mathbb{P}\left[\left|g(X_1, \ldots, X_n) - \mathbb{E}\left[g(X_1, \ldots, X_n)\right]\right| \geq \varepsilon\right] \leq 2\exp\left(-\frac{2e^w}{\sum_{k=1}^{n} c_k^2}\right)$$

3.2 Applications

3.2.1 Hoeffding's Inequality

The Hoeffding's Inequality is a special case of the Bounded Difference Inequality.

Theorem 11. *Let $X_1, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_j] = 0$ for $j \in \{1, \ldots, n\}$. Suppose $X_k \in [a_k, b_k]$ for $k \in \{1, \ldots, n\}$ and constants $a_k < b_k$, $k \in \{1, \ldots, n\}$. Then*

$$\mathbb{P}\left[\sum_{k=1}^{n} X_k \geq t\right] \leq e^{-\frac{2t^2}{\sum_{k=1}^{n} (b_j - a_j)^2}}$$

Proof. This immediately follows from Theorem 10 because the bounded difference condition is satisfied with constants $c_k = b_k - a_k$ for $k \in \{1, \dots, n\}$. □

Theorem 12. *Assume that $\mathscr{F}$ is uniformly bounded by the constant $U > 0$:*

$$\forall f \in \mathscr{F} : \forall s \in \mathcal{S} : \quad |f(s)| \leq U$$

Then for all $\varepsilon > 0$

$$\mathbb{P}\left[\|\hat{P}_n - P\|_{\mathscr{F}} - \mathbb{E}\left[\|\hat{P}_n - P\|_{\mathscr{F}}\right] \geq \frac{tU}{\sqrt{n}}\right] \leq e^{\frac{t^2}{2}}.$$

Proof. For each $f \in \mathscr{F}$, we have by assumption $|f(s)| \leq U$ for all $s \in \mathcal{S}$. It follows

$$\hat{P}_n f - Pf = \frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n} f(X_k)\right] = g_f(X_1, \dots, X_n)$$

with

$$g_f : \begin{cases} \mathbb{R}^n & \to \mathbb{R} \\ (x_1, \dots, x_n) & \mapsto \frac{1}{n}\sum_{k=1}^{n} f(x_k) - \mathbb{E}\left[g(X_1, \dots, X_n)\right] \end{cases}$$

satisfies the bounded difference condition with parameter $c_k = 2U/n$, since all $f \in \mathscr{F}$ are bounded by U in absolute value. This implies that also

$$g : \begin{cases} \mathbb{R}^n & \to \mathbb{R} \\ (x_1, \dots, x_n) & \mapsto \sup_{f \in \mathscr{F}} |g_f(x_1, \dots, x_n)| \end{cases}$$

satisfies the bounded difference condition with the same parameters[2]. Now note that

$$g(X_1, \dots, X_n) = \|\hat{P}_n - P\|_{\mathscr{F}}$$

and use Theorem 10. □

3.2.2 Glivenko Cantelli Theorem

The Glivenko-Cantelli Theorem is one of the most important theorems of mathematical statistics. It states that we are able to learn a distribution function from IID samples. The following proof of the Glivenko-Cantelli Theorem uses the Hoeffding Inequality. It is taken from [4] but since the part "Step 1. First Symmetrization by a ghost sample" on pages 193 and 194 seems to be wrong, the "Step 1" of the following proof is my own work.

Theorem 13 (Glivenko-Cantelli)**.** *Let $(X_i)_{i \in \mathbb{N}}$ be IID real-valued random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ each having distribution P on $\mathbb{R}$. Then the distribution functions*

$$F : \begin{cases} \mathbb{R} & \to [0, 1] \\ x & \mapsto \mathbb{P}\left((-\infty, x]\right) \end{cases}$$

$$\hat{F}_n : \begin{cases} \Omega \times \mathcal{B}(\mathbb{R}) & \to [0, 1] \\ (\omega, x) & \mapsto \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{(-\infty, x]}(X_k(\omega)) \end{cases}$$

satisfy the following:[3]

[2] Note that here again, we assume measurability

[3] Because of right-continuity of distribution functions, we have $\sup_{z \in \mathbb{R}} |F(z) - \hat{F}_n(z)| = \sup_{z \in \mathbb{Q}} |F(z) - \hat{F}_n(z)|$ and therefore, we do not have any measurability problems here.

1. $\mathbb{P}\left[\sup_{z\in\mathbb{R}} |F(z) - \hat{F}_n(z)| > \varepsilon\right] \leq 8(n+1)e^{-n\varepsilon^2/32}$

2. $\sup_{x\in\mathbb{R}} |F(z) - \hat{F}_n(z)| \to 0$ *almost surely.*

Proof. The second claim follows from the first claim: By the first Borel-Cantelli lemma,

$$\sum_{n=1}^{\infty} 8(n+1)e^{-n\varepsilon^2/32} < \infty$$

implies that for every $\varepsilon > 0$

$$\limsup_{n\to\infty} \sup_{z\in\mathbb{R}} |F(z) - F_n(z)| \leq \varepsilon,$$

which implies that

$$\lim_{n\to\infty} \sup_{z\in\mathbb{R}} |F(z) - F_n(z)| = 0.$$

The first part will be proved in 4 steps.

Step 1: Symmetrization with ghost sample For $n\varepsilon^2 < 2$, the inequality is obvious, since in this case the bound evaluates to

$$8(n+1)e^{-n\varepsilon^2/32} \geq 8(n+1)e^{-1/16} \geq 8e^{-1} > 1,$$

hence assume $n\varepsilon^2 \geq 2$. Without loss of generality, assume that $(\Omega, \mathcal{F}, \mathbb{P})$ admits random variables $(X'_i)_{i\in\mathbb{N}}$ independent of each other and of all $(X_i)_{i\in\mathbb{N}}$ with the same distribution P. Define the empirical distribution functions

$$\hat{P}_n : \begin{cases} \Omega \times \mathcal{B}(\mathbb{R}) & \to [0,1] \\ (\omega, A) & \mapsto \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_A(X_k(\omega)) \end{cases}$$
$$\hat{P}'_n : \begin{cases} \Omega \times \mathcal{B}(\mathbb{R}) & \to [0,1] \\ (\omega, A) & \mapsto \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_A(X'_k(\omega)). \end{cases}$$

Next, define $\mathcal{A} := \{(-\infty, x] \subset \mathbb{R} |\ x \in \mathbb{R}\}$. With this definition, we have

$$\sup_{A\in\mathcal{A}} |\hat{P}_n(A) - P(A)| = \sup_{z\in\mathbb{R}} |F(z) - F_n(z)|$$

In this step, we want to prove

$$\mathbb{P}\left(\left\{\sup_{A\in\mathcal{A}} |\hat{P}_n(A) - P(A)| > \varepsilon/2\right\}\right) \leq 2\mathbb{P}\left(\left\{\sup_{A\in\mathcal{A}} |\hat{P}_n(A) - \hat{P}'_n(A)| > \varepsilon/2\right\}\right). \tag{2}$$

This result is achieved by a conditioning argument. First observe that for $a, b, c \in \mathbb{R}$, we have the implication

$$|a-b| > \varepsilon \wedge |c-b| < \frac{\varepsilon}{2} \implies |a-c| > \frac{\varepsilon}{2}.$$

Therefore, we have for every $A \in \mathcal{A}$

$$\left\{|\hat{P}_n(A) - P(A)| > \varepsilon\right\} \bigcap \left\{|\hat{P}'_n(A) - P(A)| < \varepsilon/2\right\} \subset \left\{|\hat{P}_n(A) - \hat{P}'_n(A)| > \varepsilon/2\right\},$$

which justifies the step from the fourth to the fifth line in the following calculation[4][5]:

$$
\begin{aligned}
& \mathbb{E}\left[\mathbb{1}_{\left\{\sup_{A\in\mathcal{A}}|\hat{P}_n(A)-\hat{P}'_n(A)|>\varepsilon/2\right\}}|X_1,\ldots,X_n\right] \\
\overset{\text{a.s.}}{=} & \mathbb{E}\left[\sup_{A\in\mathcal{A}}\mathbb{1}_{\left\{|\hat{P}_n(A)-\hat{P}'_n(A)|>\varepsilon/2\right\}}|X_1,\ldots,X_n\right] \\
\overset{\text{a.s.}}{=} & \int_{\mathbb{R}^n}\sup_{A\in\mathcal{A}}\mathbb{1}_{\left\{|\hat{P}_n(A)-\frac{1}{n}\sum_{k=1}^n\mathbb{1}_A(x'_k)|>\varepsilon/2\right\}}\,dP^n(x'_1,\ldots,x'_n) \\
\geq & \sup_{A\in\mathcal{A}}\int_{\mathbb{R}^n}\mathbb{1}_{\left\{|\hat{P}_n(A)-\frac{1}{n}\sum_{k=1}^n\mathbb{1}_A(x'_k)|>\varepsilon/2\right\}}\,dP^n(x'_1,\ldots,x'_n) \\
\geq & \sup_{A\in\mathcal{A}}\int_{\mathbb{R}^n}\mathbb{1}_{\left\{|\hat{P}_n(A)-P(A)|>\varepsilon\right\}}\mathbb{1}_{\left\{|\frac{1}{n}\sum_{k=1}^n\mathbb{1}_A(x'_k)-P(A)|<\varepsilon/2\right\}}\,dP^n(x'_1,\ldots,x'_n) \\
= & \sup_{A\in\mathcal{A}}\mathbb{1}_{\left\{|\hat{P}_n(A)-P(A)|>\varepsilon\right\}}\int_{\mathbb{R}^n}\mathbb{1}_{\left\{|\frac{1}{n}\sum_{k=1}^n\mathbb{1}_A(x'_k)-P(A)|<\varepsilon/2\right\}}\,dP^n(x'_1,\ldots,x'_n) \\
= & \sup_{A\in\mathcal{A}}\mathbb{1}_{\left\{|\hat{P}_n(A)-P(A)|>\varepsilon\right\}}\underbrace{\mathbb{E}\left[\mathbb{1}_{\left\{|\hat{P}'_n(A)-P(A)|<\varepsilon/2\right\}}\right]}_{\geq\frac{1}{2}} \\
\geq & \frac{1}{2}\sup_{A\in\mathcal{A}}\mathbb{1}_{\left\{|\hat{P}_n(A)-P(A)|>\varepsilon\right\}} = \frac{1}{2}\mathbb{1}_{\left\{\sup_{A\in\mathcal{A}}|\hat{P}_n(A)-P(A)|>\varepsilon\right\}}
\end{aligned}
$$

where we used that

$$
\mathbb{1}_{\left\{\sup_{A\in\mathcal{A}}|\hat{P}_n(A)-\hat{P}'_n(A)|>\varepsilon/2\right\}} = \sup_{A\in\mathcal{A}}\mathbb{1}_{\left\{|\hat{P}_n(A)-\hat{P}'_n(A)|>\varepsilon/2\right\}}
$$

and

$$
\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\left\{|\hat{P}'_n(A)-P(A)|<\varepsilon/2\right\}}\right] &= \mathbb{E}\left[\mathbb{1}_{\left\{(\hat{P}'_n(A)-P(A))^2<\varepsilon^2/4\right\}}\right] = \mathbb{E}\left[1-\mathbb{1}_{\left\{(\hat{P}'_n(A)-P(A))^2\geq\varepsilon^2/4\right\}}\right] \\
&\geq 1-\frac{Var(\frac{1}{n}\sum_{k=1}^n\mathbb{1}_A(X'_k))}{\varepsilon^2/4} = 1-\frac{nP(A)(1-P(A))}{n^2\varepsilon^2/4} \\
&\geq 1-\frac{1}{n\varepsilon^2}\geq 1-\frac{1}{2}=\frac{1}{2}.
\end{aligned}
$$

The second inequality above follows from $x(1-x)\leq\frac{1}{4}$ for $x\in[0,1]$ and the third one follows from the assumption $n\varepsilon^2\geq 2$. We now have

$$
\begin{aligned}
& \mathbb{P}\left(\left\{\sup_{A\in\mathcal{A}}|\hat{P}_n(A)-\hat{P}'_n(A)|>\varepsilon/2\right\}\right) \\
= & \mathbb{E}\left[\mathbb{1}_{\left\{\sup_{A\in\mathcal{A}}|\hat{P}_n(A)-\hat{P}'_n(A)|>\varepsilon/2\right\}}\right] \\
= & \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{\left\{\sup_{A\in\mathcal{A}}|\hat{P}_n(A)-\hat{P}'_n(A)|>\varepsilon/2\right\}}|X_1,\ldots,X_n\right]\right] \\
\geq & \mathbb{E}\left[\frac{1}{2}\mathbb{1}_{\left\{\sup_{A\in\mathcal{A}}|\hat{P}_n(A)-P(A)|>\varepsilon\right\}}\right] \\
= & \frac{1}{2}\mathbb{P}\left(\left\{\sup_{A\in\mathcal{A}}|\hat{P}_n(A)-P(A)|>\varepsilon\right\}\right),
\end{aligned}
$$

[4]What we need here is that $\sup_{A\in\mathcal{A}}|\hat{P}_n(A)-\hat{P}'_n(A)|$ (first line) and $\sup_{A\in\mathcal{A}}|\hat{P}_n(A)-P(A)|$ (last line) are measurable. However, this is fulfilled here, since we can replace the uncountable supremum by a countable one, see previous footnote. We do not need the expression in the fourth line of the calculation to be measurable.

[5]The expression in the third line is a version of the conditional expectation: Because of independence, this means we integrate out the $x'_1,\ldots,x'_n$, whereas the whole integral is still a function from Ω to $\mathbb{R}$ (via $\hat{P}_n$, which means it is $\sigma(X_1,\ldots,X_n)$-measurable, see the definition $\hat{P}_n$ above)

which shows equation (2).

Step 2: Symmetrization with Rademacher random variables Let $\sigma_1, \dots, \sigma_n$ be IID Rademacher random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ independent of $(X_i)_{i\in\mathbb{N}}$ and $(X'_i)_{i\in\mathbb{N}}$. Then, since $(X_1, \dots, X_n)$ has the same distribution as $(X'_1, \dots, X'_n)$, we have that

$$\sup_{A\in\mathcal{A}} \Big| \sum_{k=1}^{n} (\mathbb{1}_A(X_i) - \mathbb{1}_A(X'_i)) \Big| \text{ and } \sup_{A\in\mathcal{A}} \Big| \sum_{k=1}^{n} \sigma_i (\mathbb{1}_A(X_i) - \mathbb{1}_A(X'_i)) \Big|$$

have the same distribution. Hence, we get by equation (2)

$$\begin{aligned}
& P\left(\left\{ \sup_{A\in\mathcal{A}} \big| \hat{P}_n(A) - P(A) \big| > \varepsilon \right\} \right) \\
\leq\ & 2P\left(\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \big| \sum_{k=1}^{n} (\mathbb{1}_A(X_k) - \mathbb{1}_A(X'_i)) \big| > \varepsilon/2 \right\} \right) \\
=\ & 2P\left(\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \big| \sum_{k=1}^{n} \sigma_k (\mathbb{1}_A(X_k) - \mathbb{1}_A(X'_i)) \big| > \varepsilon/2 \right\} \right) \\
\leq\ & 2P\left(\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \big| \sum_{k=1}^{n} \sigma_k \mathbb{1}_A(X_k) \big| > \varepsilon/4 \right\} \right) + 2P\left(\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \big| \sum_{k=1}^{n} \sigma_k \mathbb{1}_A(X'_k) \big| > \varepsilon/4 \right\} \right) \\
\leq\ & 4P\left(\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \big| \sum_{k=1}^{n} \sigma_k \mathbb{1}_A(X_k) \big| > \varepsilon/4 \right\} \right),
\end{aligned}$$

where we used the union bound in the third step. In Summary, we now have

$$P\left(\left\{ \sup_{A\in\mathcal{A}} \big| \hat{P}_n(A) - P(A) \big| > \varepsilon \right\} \right) \leq 4P\left(\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \big| \sum_{k=1}^{n} \sigma_k \mathbb{1}_A(X_k) \big| > \varepsilon/4 \right\} \right) \tag{3}$$

Step 3: Conditioning For $(x_1, \dots, x_n) \in \mathbb{R}^n$, define

$$M_{(x_1,\dots,x_n)} = \{ (\mathbb{1}_A(x_1), \dots, \mathbb{1}_A(x_n)) | A \in \mathcal{A} \} \subset \{0,1\}^n$$

so we have obviously

$$\begin{aligned}
\left\{ \sup_{A\in\mathcal{A}} \frac{1}{n} \Big| \sum_{k=1}^{n} \sigma_k \mathbb{1}_A(x_k) \Big| > \varepsilon/4 \right\} &= \left\{ \sup_{\alpha\in M_{(x_1,\dots,x_n)}} \frac{1}{n} \Big| \sum_{k=1}^{n} \sigma_k \alpha_k \Big| > \varepsilon/4 \right\} \\
&= \bigcup_{\alpha\in M_{(x_1,\dots,x_n)}} \left\{ \frac{1}{n} \big| \sum_{k=1}^{n} \sigma_k \alpha_k \big| > \varepsilon/4 \right\}
\end{aligned}$$

This implies for P^n almost every $(x_1, \dots, x_n) \in \mathbb{R}^n$

$$\begin{aligned}
& P\left(\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\Big|\sum_{k=1}^{n} \sigma_k \mathbb{1}_A(X_k)\Big| > \varepsilon/4\right\} | X_1 = x_1, \dots, X_n = x_n\right) \\
= \;& P\left(\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\Big|\sum_{k=1}^{n} \sigma_k \mathbb{1}_A(x_k)\Big| > \varepsilon/4\right\} | X_1 = x_1, \dots, X_n = x_n\right) \\
= \;& P\left(\bigcup_{\alpha\in M_{(x_1,\dots,x_n)}} \left\{\frac{1}{n}\Big|\sum_{k=1}^{n} \sigma_k \alpha_k\Big| > \varepsilon/4\right\} | X_1 = x_1, \dots, X_n = x_n\right) \\
\leq \;& (n+1) \underbrace{\max_{\alpha\in M_{(x_1,\dots,x_n)}} P\left(\left\{\frac{1}{n}\Big|\sum_{k=1}^{n} \sigma_k \alpha_k\Big| > \varepsilon/4\right\} | X_1 = x_1, \dots, X_n = x_n\right)}_{\leq 2\exp(-n\varepsilon^2/32)} \\
\leq \;& 2(n+1)\exp(-n\varepsilon^2/32)
\end{aligned}$$

The last two steps will be explained below but first note that the calculation above completes the proof when combined with equation (3):

$$\begin{aligned}
P\left(\left\{\sup_{A\in\mathcal{A}} \left|\hat{P}_n(A) - P(A)\right| > \varepsilon\right\}\right) &\leq 4P\left(\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\Big|\sum_{k=1}^{n} \sigma_k \mathbb{1}_A(X_k)\Big| > \varepsilon/4\right\}\right) \\
&\leq 8(n+1)e^{-n\varepsilon^2/32}
\end{aligned}$$

For the last two missing steps, note that

- the first one follows from the union bound because the number of elements in $M_{(x_1,\dots,x_n)}$ is bounded by $n+1$ for every $(x_1, \dots, x_n) \in \mathbb{R}^n$:

$$\begin{aligned}
\left|\left\{(\mathbb{1}_A(x_1), \dots, \mathbb{1}_A(x_n)) | A \in \mathcal{A}\right\}\right| &= \left|\left\{(\mathbb{1}_{(-\infty,x]}(x_1), \dots, \mathbb{1}_{(-\infty,x]}(x_n)) | x \in \mathbb{R}\right\}\right| \\
&\leq n+1
\end{aligned}$$

- the second one follows from Hoeffding's inequality: For $(x_1, \dots, x_n) \in \mathbb{R}^n$ and $\alpha \in M_{(x_1,\dots,x_n)}$ let $N_\alpha = \sum_{k=1}^{n} \alpha_k$. Then $\sum_{k=1}^{n} \sigma_k \alpha_k$ is the sum of N_α independent identically distributed random variables which are almost surely in $[-1, 1]$. Since the $(\sigma_i)_{i\in\{1,\dots,n\}}$ are independent of $(X_k)_{k\in\mathbb{N}}$, conditioning on $X_1 = x_1, \dots, X_k = x_k$ doesn't change this. In consequence, we have for P^n almost every $(x_1, \dots, x_n) \in \mathbb{R}^n$ by Hoeffding's inequality

$$\begin{aligned}
& \max_{\alpha\in M_{(x_1,\dots,x_n)}} P\left(\left\{\frac{1}{n}\Big|\sum_{k=1}^{n} \sigma_k \alpha_k\Big| > \varepsilon/4\right\} | X_1 = x_1, \dots, X_n = x_n\right) \\
\leq \;& \max_{\alpha\in M_{(x_1,\dots,x_n)}} 2\exp\left(-2\frac{(\frac{n\varepsilon}{4})^2}{\sum_{k=1}^{N_\alpha} 2^2}\right) = \max_{\alpha\in M_{x_1,\dots,x_n}} 2\exp\left(-\frac{n^2\varepsilon^2}{32N_\alpha}\right) \leq 2e^{-\frac{n\varepsilon^2}{32}}
\end{aligned}$$

□

4 Vapnik Chernovenkis Theory

The exposition of this section is adapted from [4], chapters 12 and 13. Fix $d \in \mathbb{N}$ and let $\mathcal{A} \subset \mathcal{B}(\mathbb{R})^d$ be a collection of Borel-measurable subsets of $\mathbb{R}^d$.

4.1 Shattering and Vapnik-Chernovenkis classes

Definition 14 (Shatter Coefficient)**.** For $x_1, \ldots, x_n \in \mathbb{R}^n$, let $N_{\mathcal{A}}$ be the number of different possible intersections of $\{x_1, \ldots, x_n\}$ with an $A \in \mathcal{A}$:

$$N_{\mathcal{A}}(x_1, \ldots, x_n) := \big|\, \{\{x_1, \ldots, x_n\} \cap A | A \in \mathcal{A}\} \,\big|$$

The n-th *shatter coefficient* of $\mathcal{A}$ is defined as

$$s(\mathcal{A}, n) = \max_{x_1, \ldots, x_n \in \mathbb{R}^d} N_{\mathcal{A}}(x_1, \ldots, x_n)$$

The shatter coefficient is a measure for the complexity of the set $\mathcal{A}$. We always have

$$N_{\mathcal{A}}(x_1, \ldots, x_n) = \big|\, \{\{x_1, \ldots, x_n\} \cap A | A \in \mathcal{A}\} \,\big| \leq \mathcal{P}(\{x_1, \ldots, x_n\}) = 2^n,$$

where $\mathcal{P}$ denotes the power set, i.e. the set of all subsets. In particular, we get

$$s(\mathcal{A}, n) \leq 2^n.$$

Note that if $s(\mathcal{A}, n_0) < 2^{n_0}$ for some $n_0 \in \mathcal{N}$ then obviously $s(\mathcal{A}, n) < 2^n$ for every $n > n_0$ in $\mathbb{N}$. We can thus make the following definition:

Definition 15 (Vapnik-Chernovenkis dimension)**.** Assume that $|\mathcal{A}| \geq 2$. Then

$$V_{\mathcal{A}} := \sup \{n \in \mathbb{N} | s(\mathcal{A}, n) = 2^n\}$$

is called *Vapnik-Chernovenkis dimension* (*VC dimension*) of the class $\mathcal{A}$. If it is finite, then $\mathcal{A}$ is called a *Vapnik-Chernovenkis class* (*VC class*).

The Sauer Lemma gives a bound on $s(\mathcal{A}, n)$ in terms of of the VC-dimension $V_{\mathcal{A}}$. For a proof, see [4], Theorem 13.2.

Lemma 16 (Sauer)**.** *For every n we have*

$$s(\mathcal{A}, n) \leq \sum_{k=0}^{V_{\mathcal{A}}} \binom{n}{k}$$

We can immediately conclude that

$$\begin{aligned} s(\mathcal{A}, n) &\leq \sum_{k=0}^{V_{\mathcal{A}}} \binom{n}{k} \leq \sum_{k=0}^{V_{\mathcal{A}}} \binom{n}{k} \underbrace{k! \binom{V_{\mathcal{A}}}{k}}_{\geq 1} \leq \sum_{k=0}^{V_{\mathcal{A}}} \underbrace{\binom{n}{k} k!}_{\leq n^k} \binom{V_{\mathcal{A}}}{k} \\ &\leq \sum_{k=0}^{V_{\mathcal{A}}} \binom{V_{\mathcal{A}}}{k} n^k 1^{n-k} = (n+1)^{V_{\mathcal{A}}} \end{aligned}$$

This property is very interesting, since it implies that

- either (if $V_{\mathcal{A}} = \infty$) the growth rate is of exponential order, more precisely

$$\forall n \in \mathbb{N}: \quad s(\mathcal{A}, n) = 2^n$$

- or (if $V_{\mathcal{A}} < \infty$), the growth rate is of order

$$s(\mathcal{A}, n) = \mathcal{O}(n^{V_{\mathcal{A}}}).$$

Nothing else is possible, for example $s(\mathcal{A}, n) \sim 2^{\sqrt{n}}$ for $n \to \infty$ is impossible. Note that there are also many other sharper bounds on $\sum_{k=0}^{V_{\mathcal{A}}} \binom{n}{k}$ than $(n+1)^{V_{\mathcal{A}}}$.

4.2 Examples

In this section, we will derive the shatter coefficients and the VC-dimension of some classes $\mathcal{A}$ for illustrational purposes.

4.2.1 The class of half lines

Let $d = 1$, i.e. we have classes $\mathcal{A}$ of Borel measurable subsets of $\mathbb{R}$. Let $\mathcal{A} = \left\{(-\infty, x] \middle| x \in \mathcal{R}\right\}$. Then for $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathbb{R}$ we have that

$$N_{\mathcal{A}}(x_1, \ldots, x_n) = \left| \{\{x_1, \ldots, x_n\} \cap (-\infty, x] \, | x \in \mathbb{R}\} \right| \leq n + 1$$

with equality if the points are pairwise different, because this is the number of possible intersections of $\{x_1, \ldots, x_n\}$ with all sets of the form $(-\infty, x]$ for $x \in \mathbb{R}$ (just like the number of values attained by the empirical distribution function of n samples in $\mathbb{R}$). Hence the shatter coefficient is

$$s(\mathcal{A}, n) = \sup_{x_1, \ldots, x_n \in \mathbb{R}} N_{\mathcal{A}}(x_1, \ldots, x_n) = n + 1$$

and the VC dimension is

$$V_{\mathcal{A}} = \sup \{n \in \mathbb{N} | s(\mathcal{A}, n) = 2^n\} = \sup \{n \in \mathbb{N} | n + 1 = 2^n\} = 1.$$

Note that

$$n + 1 = \binom{n}{0} + \binom{n}{1} = \sum_{k=0}^{V_{\mathcal{A}}} \binom{n}{k},$$

so the bound from the Sauer Lemma 16 is sharp in this case.

4.2.2 The class of intervals

Let again $d = 1$ and this time let $\mathcal{A} = \{[a, b] | a, b \in \mathbb{R}, a < b\}$. We now have that

$$\begin{aligned} s(\mathcal{A}, n) &= \sup \left\{x_1, \ldots, x_n \in \mathbb{R} \middle| \{x_1, \ldots, x_n\} \cap [a, b] \, | a, b \in \mathbb{R}, a < b\right\} \\ &= 1 + \sum_{k=1}^{n} k = 1 + \frac{n(n+1)}{2} \end{aligned}$$

because for n points in $\mathbb{R}$, we can have one empty set, one interval containing all n points, at most two intervals containing $n - 1$ points,..., at most n intervals containing one point. In total, we get the bound in the second line above which is attained if all points are different. Hence we have

$$V_{\mathcal{A}} = \sup \{n \in \mathbb{N} | s(\mathcal{A}, n) = 2^n\} = \sup \left\{n \in \mathbb{N} | 1 + \tfrac{n(n+1)}{2} = 2^n\right\} = 2.$$

Like before, the bound from the Sauer Lemma 16 is sharp here:

$$s(\mathcal{A}, n) = 1 + \frac{n(n+1)}{2} = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} = \sum_{k=0}^{V_{\mathcal{A}}} \binom{n}{k}$$

4.3 The Vapnik Chernovenkis Inequality

The following theorem is an important result by Vapnik and Chernovenkis from 1971. The proof shown here is almost the same as the one shown before for the Glivenko Cantelli Theorem 13. Note that it is a distribution independent theorem, just like the Theorem 13. Further note that we do not deal with measurability issues here. Unlike in Theorem 13, where we could replace every supremum by a countable one, we here assume that the suprema are measurable where it is necessary. For further information, see [4], page 198 or [2], page 17.

Like before, assume that $(X_i)_{i\in\mathcal{N}}$ are random variables from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathbb{R}^d$. Note that the target space is now $\mathbb{R}^d$ and not $\mathbb{R}$. Again, define the empirical distribution $\hat{P}_n$ by

$$\hat{P}_n : \begin{cases} \Omega \times \mathcal{B}(\mathbb{R})^d & \to [0,1] \\ (\omega, A) & \mapsto \frac{1}{n}\sum_{k=1}^n \mathbb{1}_A(X_k(\omega)) \end{cases}$$

Theorem 17 (Vapnik Chernovenkis inequality)**.** *For any probability measure P on $\mathbb{R}^d$, $n \in \mathbb{N}$ and $\varepsilon > 0$ we have*

$$P\left(\left\{\sup_{A\in\mathcal{A}} \left|\hat{P}_n(A) - P(A)\right| > \varepsilon\right\}\right) \leq 8s(\mathcal{A}, n)e^{-n\varepsilon^2/32} \tag{4}$$

Proof. The proof is split up into three steps. **Step 1** and **Step 2** of the proof are completely the same as in the proof of the Glivenko-Cantelli Theorem 13 except for the measurability issues. There it is proven (see equation (3))

$$P\left(\left\{\sup_{A\in\mathcal{A}} \left|\hat{P}_n(A) - P(A)\right| > \varepsilon\right\}\right) \leq 4P\left(\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\Big|\sum_{k=1}^n \sigma_k \mathbb{1}_A(X_k)\Big| > \varepsilon/4\right\}\right) \tag{5}$$

Step 3: Conditioning For $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$, define

$$M_{(x_1,\ldots,x_n)} = \{(\mathbb{1}_A(x_1), \ldots, \mathbb{1}_A(x_n)) | A \in \mathcal{A}\} \subset \{0,1\}^n$$

so we have obviously

$$\begin{aligned} \left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\Big|\sum_{k=1}^n \sigma_k \mathbb{1}_A(x_k)\Big| > \varepsilon/4\right\} &= \left\{\sup_{\alpha\in M_{(x_1,\ldots,x_n)}} \frac{1}{n}\Big|\sum_{k=1}^n \sigma_k\alpha_k\Big| > \varepsilon/4\right\} \\ &= \bigcup_{\alpha\in M_{(x_1,\ldots,x_n)}} \left\{\frac{1}{n}\Big|\sum_{k=1}^n \sigma_k\alpha_k\Big| > \varepsilon/4\right\} \end{aligned}$$

This implies for P^n almost every $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$

$$\begin{aligned}
& P\left(\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\Big|\sum_{k=1}^{n} \sigma_k \mathbb{1}_A(X_k)\Big| > \varepsilon/4\right\} | X_1 = x_1, \dots, X_n = x_n\right) \\
= \; & P\left(\left\{\sup_{A\in\mathcal{A}} \frac{1}{n}\Big|\sum_{k=1}^{n} \sigma_k \mathbb{1}_A(x_k)\Big| > \varepsilon/4\right\} | X_1 = x_1, \dots, X_n = x_n\right) \\
= \; & P\left(\bigcup_{\alpha\in M_{(x_1,\dots,x_n)}} \left\{\frac{1}{n}\Big|\sum_{k=1}^{n} \sigma_k \alpha_k\Big| > \varepsilon/4\right\} | X_1 = x_1, \dots, X_n = x_n\right) \\
\leq \; & s(\mathcal{A}, n) \underbrace{\max_{\alpha\in M_{(x_1,\dots,x_n)}} P\left(\left\{\frac{1}{n}\Big|\sum_{k=1}^{n} \sigma_k \alpha_k\Big| > \varepsilon/4\right\} | X_1 = x_1, \dots, X_n = x_n\right)}_{\leq 2\exp(-n\varepsilon^2/32), \text{ see proof of Theorem 13}} \\
\leq \; & 2s(\mathcal{A}, n)\exp(-n\varepsilon^2/32)
\end{aligned}$$

The first inequality follows from the union bound because there are $N_{\mathcal{A}}(x_1, \dots, x_n)$ different elements in $M_{(x_1,\dots,x_n)}$, which is bounded by $s(\mathcal{A}, n)$. The calculation above completes the proof when combined with equation (5): □

Remark Note that the estimate of

$$\mathbb{P}\left[\sup_{z\in\mathbb{R}} |F(z) - \hat{F}_n(z)| > \varepsilon\right] \leq 8(n+1)e^{-n\varepsilon^2/32}$$

in the Glivenko-Cantelli Theorem 13 is exactly an application of the above theorem since there $s(\mathcal{A}, n) = n + 1$, see section 4.2.1.

Remark The above Theorem is only useful if $s(\mathcal{A}, n)$ does not grow too fast when n increases. As we have seen in Section 4.2, it can happen that $s(\mathcal{A}, n) = 2^n$ for every $n \in \mathbb{N}$ and in this case, the bound is useless since

$$8s(\mathcal{A}, n)e^{-n\varepsilon^2/32} = 8e^{n(\ln(2)-\varepsilon^2/32)} > 8 > 1.$$

However, we have seen in section 4.1 that we have

- either $s(\mathcal{A}, n) = 2^n$ for every n (then the bound is useless)
- or (if VC dimension $V_{\mathcal{A}}$ is finite) we have that $s(\mathcal{A}, n)$ grows of order $\mathcal{O}(n^{V_{\mathcal{A}}})$.

Since a polynomial times an exponential decay is still an exponential decay, we have either a useless bound (if A is not a VC class) or an exponentially decaying bound (if $\mathcal{A}$ is a VC class) in equation (4).

4.4 Selection of 0-1-Classifiers

4.4.1 Notation and Definitions

In this section, we will derive some results for empirical risk minimization over a set of 0-1 classifiers. For this entire section, fix $d \in \mathbb{N}$ and assume that C is a set of Borel-measurable functions $\Phi : \mathbb{R}^d \to \{0, 1\}$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with IID random variables $(Z_k)_{k\in\mathbb{N}}$ which map to $\mathbb{R}^d \times \{0, 1\} \subset \mathbb{R}^{d+1}$. For notational convenience,

we will write $Z_k = (X_k, Y_k)$ where X_k maps to $\mathbb{R}^d$ and Y_k maps to $\{0,1\}$ for $k \in \mathbb{N}$. Denote the distribution of each of the Z_k, $k \in \mathbb{N}$ by P. We will use the 0-1 loss function

$$l : \begin{cases} \{0,1\} \times \{0,1\} & \to \mathbb{R} \\ (a,b) & \mapsto \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b. \end{cases} \end{cases}$$

The set C takes the role of the set of decision rules and the risk of such a decision rule is therefore

$$L : \begin{cases} C & \to \mathbb{R} \\ \Phi & \mapsto \int_{\mathbb{R}^d \times \mathbb{R}} l(\Phi(x), y) dP(x,y) = P\left(\{\Phi(X) \neq Y\}\right) \end{cases}$$

Like before define the empirical distribution by

$$\hat{P}_n : \begin{cases} \Omega \times \mathcal{B}(\mathbb{R})^{d+1} & \to [0,1] \\ (\omega, A) & \mapsto \frac{1}{n} \sum_{k=1}^n \mathbb{1}_A(Z_k(\omega)). \end{cases}$$

Therefore the empirical risk function is

$$\hat{L}_n : \begin{cases} \Omega \times C & \to \mathbb{R} \\ (\omega, \Phi) & \mapsto \frac{1}{n} \sum_{k-1}^n \mathbb{1}_{\{\Phi(X_k(\omega)) \neq Y_k(\omega)\}}. \end{cases}$$

Note that for every $\Phi \in C$, $\hat{L}_n(\cdot, \Phi)$ is a random variable. Again we do not deal with measurability issues here.

Remark For $\Phi \in C$, it is useful to define $l \circ \Phi$ by

$$l \circ \Phi : \begin{cases} \mathbb{R}^d \times \{0,1\} & \to \mathbb{R} \\ (x,y) & \mapsto l(\Phi(x), y) \end{cases}$$

With this definition, this risk minimization problem fits in the framework of Section 1:

$$\mathscr{F} := \left\{ l \circ \Phi \middle| \, \Phi \in C \right\}, \quad P\, l \circ \Phi = L(\Phi), \quad \hat{P}_n l \circ \Phi = \hat{L}(\cdot, \Phi)$$

The next lemma gives a bound on the excess loss of the classifier $\Phi_n^*(\omega)$ which is selected as a minimizer the empirical loss $\hat{L}_n(\omega, \cdot)$ on the basis of the observations $(X_1(\omega), Y_1(\omega)), \ldots, (X_n(\omega), Y_n(\omega))$.

Definition 18. For the class C of decision functions defined above, we define the set

$$\mathcal{A}_C := \left\{ \{x \in \mathbb{R}^d | \, \Phi(x=1)\} \times \{0\} \bigcup \{x \in \mathbb{R}^d | \, \Phi(x) = 0\} \times \{1\} \, \middle| \, \Phi \in C \right\}$$

and the shatter coefficient $\mathcal{S}(C,n)$ as

$$\mathcal{S}(C,n) := s(\mathcal{A}_C, n), \quad n \in \mathbb{N}.$$

Furthermore define the VC dimension V_C of C as

$$V_C := V_{\mathcal{A}_C}$$

4.4.2 Interesting Theorems

Theorem 19. *It holds*

$$\mathbb{P}\left(\left\{\sup_{\Phi\in C}\left|\hat{L}_n(\cdot,\Phi)-L(\Phi)\right|>\varepsilon\right\}\right)\leq 8\mathcal{S}(C,n)e^{-n\varepsilon^2/32}$$

Proof.

$$\begin{aligned}
&\mathbb{P}\left(\left\{\sup_{\Phi\in C}\left|\hat{L}_n(\cdot,\Phi)-L(\Phi)\right|>\varepsilon\right\}\right)\\
=\;&\mathbb{P}\Big(\Big\{\sup_{\Phi\in C}\Big|\hat{P}_n\left(\left\{(x,y)\in\mathbb{R}^d\times\{0,1\}\,\middle|\,\Phi(x)\neq y\right\}\right)\\
&\qquad -P\left(\left\{(x,y)\in\mathbb{R}^d\times\{0,1\}\,\middle|\,\Phi(x)\neq y\right\}\right)\Big|>\varepsilon\Big\}\Big)\\
=\;&\mathbb{P}\left(\left\{\sup_{A\in\mathcal{A}_C}\left|\hat{P}_n(A)-P(A)\right|>\varepsilon\right\}\right)
\end{aligned}$$

Here, you see that the set $\mathcal{A}_C$ was exactly chosen for that purpose. Now an application of Theorem 17 proves the result. □

The difficult part is always to estimate $\mathcal{S}(C,n)$ for a given set of classifiers. Now we prove a lemma that allows us to make use of the above Theorem 19.

Lemma 20. *For all $n\in\mathbb{N}$ and all $\omega\in\Omega$*

$$L(\Phi_n^*(\omega))-\inf_{\Phi\in C}\leq 2\sup_{\Phi\in C}\left|\hat{L}_n(\omega,\Phi)-L(\Phi)\right|.$$

Proof.

$$\begin{aligned}
L(\Phi_n^*(\omega))-\inf_{\Phi\in C}L(\Phi) &= L(\Phi_n^*(\omega))-\hat{L}_n(\omega,\Phi_n^*(\omega))+\hat{L}_n(\Phi_n^*(\omega))-\inf_{\Phi\in C}L(\Phi)\\
&= L(\Phi_n^*(\omega))-\hat{L}_n(\omega,\Phi_n^*(\omega))+\sup_{\Phi\in C}\Big(\underbrace{\hat{L}_n(\Phi_n^*(\omega))}_{\leq\hat{L}_n(\Phi)\text{ for all }\Phi}-L(\Phi)\Big)\\
&\leq L(\Phi_n^*(\omega))-\hat{L}_n(\omega,\Phi_n^*(\omega))+\sup_{\Phi\in C}\left(\hat{L}_n(\Phi))-L(\Phi)\right)\\
&\leq \left|L(\Phi_n^*(\omega))-\hat{L}_n(\omega,\Phi_n^*(\omega))\right|+\sup_{\Phi\in C}\left|\hat{L}_n(\Phi))-L(\Phi)\right|\\
&\leq 2\sup_{\Phi\in C}\left|\hat{L}_n(\omega,\Phi)-L(\Phi)\right|
\end{aligned}$$

□

We are now in the position to prove the following theorem:

Theorem 21. *It holds*

$$\mathbb{P}\left(\left\{L(\Phi_n^*(\cdot))-\inf_{\Phi\in C}L(\Phi)>\varepsilon\right\}\right)\leq 8\mathcal{S}(C,n)e^{-n\varepsilon^2/128}$$

Proof. From Lemma 20 and Theorem 19 we get

$$\begin{aligned}
\mathbb{P}\left(\left\{L(\Phi_n^*(\cdot))-\inf_{\Phi\in C}L(\Phi)>\varepsilon\right\}\right) &\leq \mathbb{P}\left(\left\{2\sup_{\Phi\in C}\left|\hat{L}_n(\cdot,\Phi)-L(\Phi)\right|>\varepsilon\right\}\right)\\
&\leq 8\mathcal{S}(C,n)e^{-n\varepsilon^2/128}.
\end{aligned}$$

□

Remark This Theorem is very important because in case that C is a VC class we know that the bound is exponentially decaying in n which means by Borel-Cantelli lemma (see the proof of the second statement of the Glivenko-Cantelli Theorem 13) that

$$L(\Phi_n^*(\cdot)) \to \inf_{\Phi \in C} L(\Phi) \quad \text{a.s.}$$

In words: we know that with probability one the sequence of losses of the classifiers selected by empirical risk minimization converges to the infimum over all losses. As a corollary of Theorem 21 we get

Corollary 22. *For every $n \in \mathbb{N}$ and every $\omega \in \Omega$*

$$\mathbb{E}\left[L(\Phi_n^*(\cdot))\right] - \inf_{\Phi \in C} L(\Phi) \leq 16\sqrt{\frac{\log(8e\mathcal{S}(C,n))}{2n}}.$$

Proof. This follows immediately from Theorem 21 by using Lemma 25 with $c = 8\mathcal{S}(C,n)$ and $b = n/128$. □

The Corollary gives a bound on the average excess risk. If C is a VC class (i.e. $S(C,n) = \mathcal{O}(n^{V_C})$) then obviously the bound in Corollary 22 is of order $\mathcal{O}(\sqrt{\frac{\log(n)}{n}})$

A Essential Supremum and Essential Infimum

If we have a collection $\mathscr{F}$ of measurable functions from a measure space into $\mathbb{R}$, the pointwise supremum of $\mathscr{F}$ need not be measurable. However there is the essential supremum which is a measurable supremum up to null sets. For reference of the following theorem, see [3], chapter five, section 18.

Theorem 23. *Let $\mathscr{F}$ be a class of measurable functions from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into $\mathbb{R}$. There is a measurable function g, uniquely determined up to null sets by the properties*

- $\forall f \in \mathscr{F}: \quad g \geq f$ *a.s. (It is almost surely a supremum)*
- *For each measurable g' satisfying the above condition, $g \leq g'$ almost surely. (It is up to null sets the lowest such function)*

Denote the function g from theorem by $g = \operatorname{ess\,sup}_{f \in \mathscr{F}}$. It is determined up to null sets. Obviously, the above theorem also proves the existence of $h = \operatorname{ess\,inf}_{f \in \mathscr{F}}$ satisfying

- $\forall f \in \mathscr{F}: \quad h \leq f$ a.s.
- For each measurable h' satisfying the above condition, $h' \geq h$ a.s.

B A bound on expectation for exponentially decaying tails

Lemma 24. *Assume we have a non-negative random variable Z and we know that $P(Z \geq t) \leq ce^{-bt}$. Then*

$$\mathbb{E}[Z] \leq \frac{\log(ce)}{b}.$$

Proof. For all $u \geq 0$ it holds

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}\left[\int_0^\infty \mathbb{1}_{[0,Z]}(t)\, dt\right] = \mathbb{E}\left[\int_0^\infty \mathbb{1}_{[t,\infty)}(Z)\, dt\right] \overset{\text{Fubini}}{=} \int_0^\infty \mathbb{E}\left[\mathbb{1}_{[t,\infty)}(Z)\right]\, dt \\ &= \int_0^\infty P(Z \geq t)\, dt \leq \int_0^u 1\, dt + \int_u^\infty ce^{-bt} dt = u + \frac{c}{b} e^{-bu}.\end{aligned}$$

This implies that

$$\mathbb{E}[Z] \leq \inf\left\{u + \tfrac{c}{b} e^{-bu} \middle| u \geq 0\right\} = \frac{\log(ce)}{b},$$

where the calculation of the infimum is an easy analytic minimization problem. □

We can use this lemma to bound the expectation, where we have an even faster decaying tail.

Lemma 25. *Assume we have a non-negative random variable Z and we know that $P(Z \geq t) \leq ce^{-bt^2}$. Then*

$$\mathbb{E}[Z] \leq \sqrt{\frac{\log(ce)}{b}}.$$

Proof.

$$\mathbb{E}[Z] \leq \sqrt{\mathbb{E}[Z^2]} \leq \sqrt{\frac{\log(ce)}{b}},$$

where the first step follows from Hölder inequality and the second one from Lemma 24 because $P(Z^2 \geq t) = P(Z \geq \sqrt{t}) \leq ce^{-bt}$. □

C Notation

Symbol	Explanation	Example
$\mathcal{B}$	This symbol denotes the Borel σ-algebra,	$\mathcal{B}(\mathbb{R})$
$\circ$	Composition of two functions	$f \circ g$
$\mathbb{1}$	Indicator function of a set for example of A	$\mathbb{1}_A$
a.s.	abbreviation for almost surely (with probability one)	

References

[1] David Pollard, *Empirical Processes: Theory and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, 1990.

[2] Vladimir Koltchinskii, *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, Springer, 2011.

[3] Joseph L. Doob, *Measure Theory*, Springer, 1992.

[4] L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, corrected second printing, Springer, 1997.