Empirical Process Theory

MEYRIGNAC PIERRE-GABRIEL

May 6, 2025

Contents

Preface

The central theme of this bachelor project is **Empirical Process Theory**, a rich field that blends tools from probability, statistics, and functional analysis. This area is pivotal for understanding the behavior of random processes, especially in high dimensions, where classical techniques often fall short.

This project begins with a recap of key concepts in probability theory, setting the stage for more advanced topics. Next, we delve into concentration inequalities, such as Hoeffding's and Chernoff's bounds, and look at some applications. In the pursuit of extending these results, we explore random vectors in high dimensions and study their properties. We then introduce a broad class of random variables, known as sub-Gaussian random variables, with the aim of extending the concentration inequalities established in the previous chapter. These results are instrumental in understanding the behavior of sums of random variables and their deviations from expected values. In the subsequent chapters, our goal is to analyze empirical processes, which requires introducing tools such as Dudley's inequality and the Vapnik-Chervonenkis (VC) dimension. These tools enable us to address key questions in statistical learning theory, such as bounding the risk of classification errors.

Chapter 1

Probability Theory: A Quick Recap

In this section, we recall some probability results that may be used in the next chapters. We assume that all our random variables live in a well-defined probability space (Ω, T, P). The capital letter X will always denote a real random variable (rv) unless otherwise stated. When we write iid for a collection of random variables, we mean that the random variables are independent and identically distributed. The expectation of X will be denoted by $\mathbb{E}(X)$. The p-th moment of a random variable is defined by: $\mathbb{E}(X^p)$. The covariance of random variables X and Y will be denoted $cov(X, Y)$ and $cov(X, Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right)$. We define the variance of X as $Var(X) = cov(X, X)$. When we use the change of variable/transformation theorem, we mean the following:

$$\mathbb{E}(f(X)) = \int_{\mathbb{R}} f(t) f_X(t)\, dt$$

where f is measurable and f_X is the density of X, provided the above expression is well-defined.

An important way of rewriting the expectation uses the tail of the random variable. If X is a non-negative integrable random variable, it is easy to verify that: **(Expectation with tail)**

$$\mathbb{E}(X) = \int_0^{+\infty} P(X > t)\, dt.$$

This result follows from rewriting X as:

$$X = \int_0^X 1\, dt = \int_0^{+\infty} 1_{t \leq X}\, dt,$$

and applying *Fubini-Tonelli's* theorem to interchange the expectation and the integral. This formula can be extended to any integrable random variable X, and not only non negative one:

$$\mathbb{E}(X) = \int_0^{+\infty} P(X > t)\, dt - \int_{-\infty}^{0} P(X < t)\, dt.$$

Indeed, for non-positive random variables, $-X$ is a non-negative random variable and writing the formula for $-X$, we obtain:

$$\mathbb{E}(-X) = -\mathbb{E}(X) = \int_0^{\infty} P(-X > t)\, dt = \int_0^{\infty} P(X < -t)\, dt.$$

Using the substitution $u = -t$, we deduce:

$$\mathbb{E}(X) = \int_{-\infty}^{0} P(X < t)\, dt.$$

Since $X = X^+ + X^-$ (the positive and negative parts of X), the formula for expectation using tails follows by the linearity of $\mathbb{E}$.

We can use this result to derive another useful expression for the expectation of $|X|^p$, which will reappear when discussing sub-Gaussian random variables. If $p \in (0, \infty)$ and X^p is integrable, then:

$$\mathbb{E}(|X|^p) = \int_0^{\infty} pt^{p-1} P(|X| > t)\, dt.$$

Exercise 1.0.1:

Let X be a random variable with finite p-th moment where $p \in (0, \infty)$. Show that:

$$\mathbb{E}(|X|^p) = \int_0^{\infty} pt^{p-1} P(|X| > t) dt,$$

whenever the right-hand side is finite.

Solution. By the formula for expectation via tails, we have:

$$\mathbb{E}(|X|^p) = \int_0^{\infty} P(|X|^p > t)\, dt = \int_0^{\infty} P(|X| > t^{1/p})\, dt.$$

The equivalence of the events $\{|X|^p > t\}$ and $\{|X| > t^{1/p}\}$ follows from the monotonicity of $x \mapsto x^{1/p}$ for $p > 0$. Substituting $u = t^{1/p}$, we have $du = \frac{1}{p} t^{\frac{1}{p}-1} dt$, so $dt = pu^{p-1} du$. Thus:

$$\int_0^{\infty} P(|X| > t^{1/p})\, dt = \int_0^{\infty} pu^{p-1} P(|X| > u)\, du.$$

Hence:

$$\mathbb{E}(|X|^p) = \int_0^\infty p t^{p-1} P(|X| > t)\, dt.$$

□

Chapter 2

Concentration inequalities and their applications

Concentration inequalities quantify how much a random variable deviates around it's mean. They usually take the form:

$$P(|X - \mu| > t) \leq f$$

where f is a small quantity that depends on t and μ. They can give precise bounds for the "tail" of the distribution. In this chapter, we will derive several concentration inequalities using both analytical and probabilistic methods. We will begin by presenting Hoeffding's concentration inequality, which applies to Rademacher random variables, followed by Chernoff's concentration inequality, which applies to general Bernoulli random variables. While these inequalities have valuable applications, they are somewhat limited, as they only address Bernoulli random variables. To overcome this restriction and extend their applicability, we will introduce a broader class of random variables, enabling us to derive more general results.

We begin this chapter by introducing two classical inequalities commonly covered in any introductory probability course.

Theorem 1 (Markov's inequality)**.** *Let X be a non-negative random variable with finite expectation. For any $a > 0$:*

$$P\left(X \geq a\right) \leq \frac{\mathbb{E}(X)}{a}.$$

Proof. Fix $a > 0$. Then we have the function inequality:

$$a \cdot 1_{X \geq a} \leq X.$$

Using monotonicity and linearity of the expectation completes the proof. □

Chebyshev's inequality is our first concentration inequality and it applies to any random variable with finite second moment.

Theorem 2 (Chebyshev's inequality)**.** *Let X be a random variable with finite second moment i.e.* $\mathbb{E}|X|^2 < \infty$. *Show that:*

$$P(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

with $\sigma^2 = Var(X)$.

Proof. The events $\{|X-\mu| \geq t\}$ and $\{|X-\mu|^2 \geq t^2\}$ are equivalent since $x \mapsto x^2$ is increasing on $[0,\infty)$. Using Markov's inequality on the random variable $|X-\mu|^2$, which is non-negative, we have:

$$P(|X-\mu| \geq t) = P(|X-\mu|^2 \geq t^2) \leq \frac{\mathbb{E}(|X-\mu|^2)}{t^2} = \frac{\sigma^2}{t^2}.$$

□

2.1 Hoeffding's inequality

Hoeffding's inequality provides a sharp exponential bound on the probability that the sum (or mean) of independent symmetric random variables deviates from 0.

Theorem 3. *(Hoeffding's inequality) :Let $X_1, \dots, X_N$ be independent symmetric Bernoulli random variables, and $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then, for any $t \geq 0$, we have:*

$$P\left\{\sum_{i=1}^{N} a_i X_i \geq t\right\} \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

Proof. Without loss of generality, we can assume that $\|a\|_2^2 = 1$ even if it means dividing by $\|a\|_2^2$. Let $\lambda \in \mathbb{R}$ be a parameter.

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{N} a_i X_i \geq t\right\} &= \mathbb{P}\left\{\exp\left(\lambda \sum_{i=1}^{N} a_i X_i\right) \geq \exp(\lambda t)\right\}, \quad \text{exponentiating} \\ &\leq e^{-\lambda t}\, \mathbb{E} \exp\left(\lambda \sum_{i=1}^{N} a_i X_i\right) \quad \text{using Markov's inequality after}\end{aligned}$$

But we also have:

$$\mathbb{E}\exp\left(\lambda\sum_{i=1}^{N} a_iX_i\right) = \prod_{i=1}^{N}\mathbb{E}\exp(\lambda a_iX_i), (*) \quad \text{using the independence of the} X_i's.$$

Let us fix i. Since X_i takes values -1 and 1 with probabilities $1/2$ each, we have

$$\mathbb{E}\exp(\lambda a_iX_i) = \frac{\exp(\lambda a_i) + \exp(-\lambda a_i)}{2} = \cosh(\lambda a_i).$$

Using the inequality : $\cosh(x) \le \exp(x^2/2) \quad$ for all $x \in \mathbb{R}$, we derive :

$$\mathbb{E}\exp(\lambda a_iX_i) \le \exp\left(\frac{\lambda^2 a_i^2}{2}\right).$$

Substituting back in $(*)$, we get:

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{N} a_iX_i \ge t\right\} &\le e^{-\lambda t}\prod_{i=1}^{N}\exp\left(\frac{\lambda^2 a_i^2}{2}\right)\\ &= \exp\left(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^{N} a_i^2\right)\\ &= \exp\left(-\lambda t + \frac{\lambda^2}{2}\right).\end{aligned}$$

Optimizing over λ gives us $\lambda = t$ as an optimal value which is exactly what we wanted to prove. □

In the proof of Hoeffding's inequality, we used the following inequality that we prove as an exercise:

Exercise 2.1.1: A useful inequality

Show that

$$\cosh(x) \le \exp(x^2/2) \quad \text{for all } x \in \mathbb{R}.$$

Solution. We want to show that:

$$\forall x \in \mathbb{R} \quad cosh(x) \le \exp(x^2/2)$$

. We use the Taylor expansions of both cosh and exp:

$$cosh(x) = \sum_{n=0}^{\infty}\frac{x^{2n}}{(2n)!}$$

and

$$\exp(x^2/2) = \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n n!}$$

It is easy to see that:

$$\forall n \in \mathbb{N} \quad 2^n n! \leq (2n)!$$

because $(2n)!$ is just $(2n) \cdot (2n-1) .... (n+1) \cdot n!$ and each of these first n terms or greater than 2 in the product. Therefore taking the inverse and noticing we have positive quantities, we get the result. □

We now give an extension of Hoeffding's inequality to bounded random variables.

Exercise 2.1.2: Hoeffding's inequality for bounded rvs

Let $X_1, \dots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have

$$\mathbb{P}\left\{ \sum_{i=1}^{N} (X_i - \mathbb{E}X_i) \geq t \right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N} (M_i - m_i)^2} \right).$$

Solution. We first show the following lemma: Suppose $X \in [m, M]$ almost surely, then we have :

$$\mathbb{E}\left(e^{s(X-EX)}\right) \leq e^{(1/8s^2(M-m)^2)}$$

Indeed, without loss of generality, by replacing $\mathbf{X}$ by $\mathbf{X} - \mathbb{E}[\mathbf{X}]$, we can assume $\mathbb{E}[\mathbf{X}] = 0$, so that $a \leq 0 \leq b$.

Since $e^{\lambda x}$ is a convex function of x, we have that for all $x \in [a, b]$,

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

So,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda \mathbf{X}}\right] &\leq \frac{b - \mathbb{E}[\mathbf{X}]}{b-a} e^{\lambda a} + \frac{\mathbb{E}[\mathbf{X}] - a}{b-a} e^{\lambda b} \\
&= \frac{b}{b-a} e^{\lambda a} + \frac{-a}{b-a} e^{\lambda b} \\
&= e^{L(\lambda(b-a))}.
\end{aligned}$$

where

$$L(h) = \frac{ha}{b-a} + \ln\left(1 + \frac{a - e^h a}{b-a}\right).$$

By computing derivatives, we find

$$L(0) = L'(0) = 0 \quad \text{and} \quad L''(h) = -\frac{abe^h}{(b-ae^h)^2}.$$

From the AMGM inequality, we thus see that $L''(h) \leq \frac{1}{4}$ for all h, and thus, from Taylor's theorem, there is some $0 \leq \theta \leq 1$ such that

$$L(h) = L(0) + hL'(0) + \frac{1}{2}h^2 L''(h\theta) \leq \frac{1}{8}h^2. \tag{2.1}$$

Thus,

$$\mathbb{E}\left[e^{\lambda \mathbf{X}}\right] \leq e^{\frac{1}{8}\lambda^2(b-a)^2}. \tag{2.2}$$

We are now ready to prove the result:

$$\begin{aligned} P\left(S_n - \mathbb{E}[S_n] \geq s\right) &= P\left(\exp(t(S_n - \mathbb{E}[S_n])) \geq \exp(ts)\right) \\ &\leq \exp(-ts)E\left[\exp(t(S_n - \mathbb{E}[S_n]))\right] \\ &= \exp(-ts)\prod_{i=1}^{n} \mathbb{E}\left[\exp(t(X_i - \mathbb{E}[X_i]))\right] \\ &\leq \exp(-ts)\prod_{i=1}^{n} \exp\left(\frac{t^2(b_i-a_i)^2}{8}\right) \\ &= \exp\left(-ts + \frac{1}{8}t^2\sum_{i=1}^{n}(b_i-a_i)^2\right) \end{aligned}$$

Here we used Markov's inequality and independence.Taking then $t = \frac{4s}{\sum_{i=1}^{n}(b_i-a_i)^2}$,which minimizes the quantity above as a function of t gets us the desired bound. □

We now give a first application of Hoeffding's inequality

Exercise 2.1.3: Boosting randomized algorithms

Imagine we have an algorithm for solving some decision problem (e.g., is a given number p a prime?). Suppose the algorithm makes a decision at random and returns the correct answer with probability $\frac{1}{2}+\delta$ with some $\delta > 0$, which is just a bit better than a random guess. To improve the performance, we run the algorithm N times and take the majority vote. Show that, for any $\epsilon \in (0,1)$, the answer is correct with probability at least $1-\epsilon$, as long as

$$N \geq \frac{1}{2\delta^2}\ln\left(\frac{1}{\epsilon}\right).$$

Solution. Let $X_1, ..., X_N$ be the indicator of the events i.e. X_i follows a Bernouilli law of parameter $1/2 + \delta$. Let

$$S_N := \sum_{i=1}^{N} X_i.$$

Let $\varepsilon > 0$. We take a majority vote that is we look at the event $\{S_N > N/2\}$. We want to bound

$$P(S_N > N/2) = P\left(S_N - \mathbb{E}(S_N) > N/2 - \mathbb{E}(S_n)\right).$$

Since

$$\mathbb{E}(S_N) + N/2 = N\delta$$

using Hoeffding's inequality for bounded random variables, we get :

$$P(S_N > N/2) \leq e^{-2N\delta^2}$$

Now we want this probability to be less than ε this means

$$N > 1/2\delta^2 ln(1/\varepsilon)$$

which is the inequality we wanted to have. □

A second application of Hoeffding's inequality is the following bound:

Exercise 2.1.4: Small ball probability

Let $X_1, \ldots, X_N$ be *non-negative* independent random variables with continuous distributions. Assume that the densities of X_i are uniformly bounded by 1.

(a) Show that the MGF of X_i satisfies

$$\mathbb{E}\exp(-tX_i) \leq \frac{1}{t} \quad \text{for all } t > 0.$$

(b) Deduce that, for any $\varepsilon > 0$, we have

$$\mathbb{P}\left\{\sum_{i=1}^{N} X_i \leq \varepsilon N\right\} \leq (e\varepsilon)^N.$$

Solution. (a) Using the transformation formula for the expectation, we have :

$$\mathbb{E}(\exp(-tX_i)) = \int_0^{+\infty} \exp(-tx) f_{X_i}(x) dx \leq \int_0^{+\infty} \exp(-tx) dx = \frac{1}{t}$$

Since X_i are almost surely non negative and their density is bounded by 1.

(b)

$$\begin{aligned}
P\left(\sum_{i=1}^{N} X_i \leq \epsilon \cdot N\right) &= P\left(\sum_{i=1}^{N} \frac{-X_i}{\epsilon} \geq -N\right) && (2.3)\\
&= P\left(\exp(\sum_{i=1}^{N} \frac{-X_i}{\epsilon} \geq e^{-N}\right) && (2.4)\\
&\leq \mathbb{E}\left(\exp(\sum_{i=1}^{N} \frac{-X_i}{\epsilon})e^{N}\right) && (2.5)\\
&= \prod_{i=1}^{N} \mathbb{E}\left(\exp(\frac{-X_i}{\epsilon})\right) e^{N} && (2.6)\\
&\leq \left(\prod_{i=1}^{N} \frac{1}{\frac{1}{\epsilon}}\right) .e^{N} && (2.7)\\
&= (e \cdot \epsilon)^{N} && (2.8)
\end{aligned}$$

Here we used Markov's inequality once (in the 2.3 line) and the question (a) to maximize the expectation of the exponential (line 2.3 to 2.4). □

In the next section, we present an other concentration inequality.

2.2 Chernoff's inequality

We saw that Hoeffding's inequality already finds some nice applications. In this section, we present an other concentration inequality that will apply to general Bernouilli random variables.

Theorem 4. *[Chernoff's inequality] Let X_i be independent Bernoulli random variables with parameters p_i. Consider their sum $S_N = \sum_{i=1}^{N} X_i$ and denote its mean by $\mu = \mathbb{E} S_N$. Then, for any $t > \mu$, we have*

$$P\{S_N \geq t\} \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t .$$

Proof. The idea of the proof is the same as the proof of Hoeffding's inequality. Let $\lambda \in \mathbb{R}$ be a parameter. As in Hoeffding's inequality proof, using Markov's inequality and exponentiating, we get:

$$\mathbb{P}\left\{S_N \geq t\right\} \leq e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E} \exp(\lambda X_i). \qquad (*)$$

Now, it remains to compute $\mathbb{E}\exp(\lambda X_i)$: Since X_i takes value 1 with probability p_i and value 0 with probability $1 - p_i$, we have

$$\mathbb{E}\exp(\lambda X_i) = e^{\lambda} p_i + (1 - p_i) = 1 + (e^{\lambda} - 1) p_i \leq \exp\left[(e^{\lambda} - 1) p_i\right] .$$

Using the numerical identity $(1 + x) \leq e^x$:

$$\prod_{i=1}^{N} \mathbb{E}\exp(\lambda X_i) \leq \exp\left[(e^{\lambda} - 1) \sum_{i=1}^{N} p_i\right] = \exp\left[(e^{\lambda} - 1)\mu\right] .$$

Now going back to $(*)$ and choosing $\lambda = \mu/t$, we get the claim. □

We now give in an exercise a small modification of Chernoff's inequality proof to bound the lower tail of S_N.

Exercise 2.2.1:

Modify the proof of Theorem 4 to obtain the following bound on the lower tail. For any $t < \mu$, we have

$$\mathbb{P}\{S_N \leq t\} \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t .$$

Solution. Let

$$S_N := \sum_{i=1}^{N} X_i \ \textit{where}\ X_i \sim \textit{Bern}(p_i).$$

Let $\lambda > 0$ and $t < \mu$ where μ is the mean value of S_N.

$$P(S_N \leq t) = P(e^{-\lambda S_n} \geq e^{-\lambda t}) \leq \frac{\mathbb{E}(e^{-\lambda S_N})}{e^{-\lambda t}} \quad \textit{Using Markov's inequality}$$

Then,

$$\mathbb{E}(e^{-\lambda S_N}) = \prod_{i=1}^{n} \mathbb{E}(e^{-\lambda X_i}) \quad \textit{Using independence}$$

.

But

$$\mathbb{E}(e^{-\lambda X_i}) = p_i e^{-\lambda} + (1 - p_i) \leq e^{p_i(e^{-\lambda}-1)}$$

using the inequality $1 + x \leq e^x$. From this, we get :

$$P(S_N \leq t) \leq e^{\lambda t} \prod_{i=1}^{n} \mathbb{E}(e^{-\lambda X_i}) = e^{\lambda t} e^{(e^{-\lambda}-1)\sum_{i=1}^{n} p_i} = e^{\lambda t} e^{(e^{-\lambda}-1)\mu}.$$

This inequality holds for all $\lambda > 0$, thus plugging for $\lambda = ln(\mu/t)$, we get the desired result (indeed $\mu/t > 1$ by assumption). □

An interesting result is that we can extend Chernoff's inequality to Poisson's random variable using limit theorem approximation.

Exercise 2.2.2: Poisson's tail

Let $X \sim \text{Pois}(\lambda)$. Show that for any $t > \lambda$, we have:

$$\mathbb{P}\{X \geq t\} \leq e^{-\lambda} \left(\frac{e\lambda}{t}\right)^t.$$

Solution. Let $X \sim \textit{Pois}(\lambda)$. Let $t > \lambda$. We know by Poisson's limit theorem 16 (appendix) that there exists a sequence $X_{i,n} \sim \textit{Bern}(p_{i,n})$ such that $S_n := \sum_{i=1}^{n} X_{i,n}$ converges in law to X. Moreover, $\mu_n = \sum_{i=1}^{n} p_{i,n} = \mathbb{E}(S_n)$ converges to λ. In this context, Using Chernoff's inequality that holds for all n and $t > \lambda > \mu_n$, we have:

$$P(S_n \geq t) \leq e^{-\mu_n} \left(\frac{e\mu_n}{t}\right)^t,$$

Now taking the limit as n goes to infinity (using the convergence in law to take the limit inside the probability) and using the fact that μ_n increases to λ, we get the desire result. □

We now give a generalisation of Chernoff's inequality:

Exercise 2.2.3: General Chernoff's inequality

Show that, in the setting of Theorem 4, for $\delta \in (0,1]$ we have

$$\mathbb{P}\{|S_N - \mu| \geq \delta\mu\} \leq 2e^{-c\mu\delta^2}$$

where $c > 0$ is an absolute constant.

Solution. Let us be in the context of Theorem 4 (Chernoff's inequality). Let $\delta \in (0,1]$:

$$P(|S_n - \mu| \geq \delta\mu) = P(S_n \geq (\delta+1)\mu) + P(S_n \leq (1-\delta)\mu)$$

Using theorem 4 (with $t = (\delta+1)\mu$, $t > \mu$):

$$P(S_n \geq (\delta+1)\mu) \leq e^{-\mu}(\frac{e\mu}{(1+\delta)\mu})^{(1+\delta)\mu}$$

Using exercise 2.2.1 (with $t = (1-\delta)\mu$, $t < \mu$):

$$P(S_n \leq (1-\delta)\mu) \leq e^{-\mu}(\frac{e\mu}{(1-\delta)\mu})^{(1-\delta)\mu})$$

. Now we want a bound on :

$$\begin{aligned} A &= e^{-\mu}\left(\frac{e\mu}{(1-\delta)\mu}\right)^{(1-\delta)\mu} + e^{-\mu}(\frac{e\mu}{(1+\delta)\mu})^{(1+\delta)\mu} \\ &= e^{-\mu}((\frac{e}{(1-\delta)})^{(1-\delta)\mu}) + (\frac{e}{(1+\delta)})^{(1+\delta)\mu}) \leq e^{\delta\mu} + e^{-\delta\mu} \\ &\leq 2e^{(\delta\mu)^2/2)}. \end{aligned}$$

□

Exercise 2.2.4: Poisson's distribution near the mean

Let $X \sim \text{Pois}(\lambda)$. Show that for $t \in (0, \lambda]$, we have

$$\mathbb{P}\{|X - \lambda| \geq t\} \leq 2\exp\left(-c\frac{t^2}{\lambda}\right).$$

Solution. Let $X \sim Pois(\lambda)$ where $\lambda > 0$,let $t \in (0, \lambda]$ and $c > 0$ of the 2.2 In the setting of exercise 2.3.3, we have :

$$P(|S_n - \mu_n| \geq t)$$

converges to

$$P(|X - \lambda| \geq t)$$

because (S_n) converges in law to X and μ_n increases to λ. Now using exercise 2.2, with $\delta = t/\mu_n \in (0, 1]$, we get :

$$P(|S_n - \mu_n| \geq t) \leq 2\exp(-c\mu_n(\frac{t}{\mu_n})^2) = 2\exp(-ct^2/\mu_n)$$

Now taking the limit as n goes to infinity and using continuity of exp, we get the desired inequality :

$$P(|X - \lambda| \geq t) \leq 2\exp(-ct^2/\lambda).$$

□

Exercise 2.2.5: Normal approximation to Poisson

Let $X \sim \text{Pois}(\lambda)$. Show that, as $\lambda \to \infty$, we have

$$\frac{X - \lambda}{\sqrt{\lambda}} \to N(0, 1) \quad \text{in distribution.}$$

Solution. Let

$$S_n = \sum_{i=1}^{n} X_i \quad \text{where} \quad X_i \sim Poi(\lambda/n).$$

We have

$$\mathbb{E}(X_i) = \lambda/n \quad \text{and} \quad Var(X_i) = \sqrt{\lambda/n}$$

$S_n \sim Poi(\lambda)$ and central limit theorem (see appendix) gives us:

$$\frac{1}{\sqrt{n}} \cdot \frac{\frac{1}{n}S_n - \lambda/n}{\sqrt{\lambda/n}} = \frac{S_n - \lambda}{\sqrt{\lambda}}$$

converges in distribution to Z where $Z \sim N(0, 1)$. Since $S_n \overset{\text{law}}{=} X$, we have the result. □

Some more applications of Chernoff's inequality.

Computing π

Exercise 2.2.6: Computing π

Choose N points $(x_1, y_1), \dots, (x_N, y_N)$ from $[-1,1] \times [-1,1]$ uniformly and independently at random, and let

$$\tilde{\pi} := \frac{4}{N} \sum_{i=1}^{N} ic(x_i, y_i)$$

where

$$ic(x,y) = \begin{cases} 1, & \text{if } x^2 + y^2 \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Fix $\delta, \varepsilon > 0$, show that if $N \geq 4 \cdot \varepsilon^{-2} \cdot \log(2/\delta)$, then the probability that $\tilde{\pi}$ is in the interval $[(1-\varepsilon)\pi, (1+\varepsilon)\pi]$ is at least $1 - \delta$.

Solution. At first, we define the sets $S := [-1,1] \times [-1,1]$ and $X := \{(x,y) \mid x^2 + y^2 \leq 1\}$ and observe that $\|S\| = 4$ and $\|X\| = \pi$. (Here we use the notation $\|S\|$ to denote the area of S.)
Note that $Y_i = ic(x_i, y_i)$ are independent Bernoulli-distributed random variables, with parameter $p = \pi/4$. We can apply the inequality of Chernoff and estimate the probability as follows:

$$P\left[\left|\sum_{i=1}^{N} Y_i - N\frac{\pi}{4}\right| \geq \varepsilon \cdot N\frac{\pi}{4}\right] \leq 2e^{-N\frac{\pi}{4}\varepsilon^2/3}.$$

Since $N \geq 4\varepsilon^{-2}\log(2/\delta)$, a simple calculation shows that the probability that $\tilde{\pi}$ is in the set $[(1-\varepsilon)\pi, (1+\varepsilon)\pi]$ is at least $1 - \delta$.

□

Erdös Rényi Random graphs

In this section, we look at random graphs $G(n,p)$ that follow the Erdös Rényi model. Every pair of the n vertices are connected to each other with probability p. In applications, the Erdös-Renyi model often appears as the simplest stochastic model for large, real-world networks. The degree of a vertex is the number of edges incident to that vertex. The expected degree of each vertex in $G(n, p)$ clearly equals:

$$(n-1)p =: d$$

Theorem 5 (Dense graphs are almost regular)**.** *There is an absolute constant C such that the following holds. Consider a random graph $G \sim G(n, p)$ with expected degree satisfying $d \geq C \log n$. Then, with high probability (for example, 0.9), the following occurs: all vertices of G have degrees between $0.9d$ and $1.1d$.*

Proof. The argument is a combination of Chernoff's inequality with a *union bound*. Let us fix a vertex i of the graph. The degree of i, which we denote d_i, is a sum of $n-1$ independent $\mathrm{Ber}(p)$ random variables (the indicators of the edges incident to i). Thus we can apply Chernoff's inequality, which yields

$$P\{|d_i - d| \geq 0.1d\} \leq 2e^{-cd}.$$

This bound holds for each fixed vertex i. Next, we can "unfix" i by taking the union bound over all n vertices. We obtain

$$P\{\exists i \leq n : |d_i - d| \geq 0.1d\} \leq \sum_{i=1}^{n} P\{|d_i - d| \geq 0.1d\} \leq n \cdot 2e^{-cd}.$$

If $d \geq C \log n$ for a sufficiently large absolute constant C, the probability is bounded by 0.1. This means that with probability 0.9, the complementary event occurs, and we have

$$P\{\forall i \leq n : |d_i - d| < 0.1d\} \geq 0.9.$$

This completes the proof. □

Exercise 2.2.7:

Consider a random graph $G \sim G(n, p)$ with expected degrees $d = O(\log n)$. Show that with high probability (say, 0.9), all vertices of G have degrees $O(\log n)$.

Solution. Let d_i, d denote the same quantity as in the proof of the above theorem. We assume that $d = O(\log n)$, that is there exists $N \in \mathbb{N}, C > 0$ such that $d \leq C \log n$ for $n \geq N$. Using Chernoff's inequality as in the theorem above, we get :

$$P\{|d_i - d| \geq \log n\} \leq 2e^{-cd} = O(\frac{1}{n}.)$$

Now by "unfixing i", we get with a union bound, we have:

$$P\left\{\exists i \leq n : |d_i - d| \geq \log n\right\} \leq \sum_{i=1}^{n} P\left\{|d_i - d| \geq \log n\right\} \leq n \cdot 2e^{-cd} = O(1).$$

This means that for n large enough the probability of the above event is very small say less that 0.1, so for n large enough the probability that all vertices have degree $O(\log n)$ is greater than 0.9. □

While Hoeffding's and Chernoff's inequalities have notable applications, their scope is limited as they apply exclusively to Bernoulli random variables. Furthermore, real-world data is often represented with high-dimensional vectors, making univariate concentration inequalities like Hoeffding's and Chernoff's less practical. The next chapter aims to extend these inequalities to accommodate a broader class of random variables.

Chapter 3

Random vectors in high dimension

In this chapter, we build upon the foundations laid in Chapter 2 by extending our analysis of concentration inequalities to a broader and more powerful framework. While Chapter 2 focused on classical inequalities such as Hoeffding's and Chernoff's bounds, which are primarily limited to independent and bounded random variables, Chapter 3 introduces tools and concepts that allow us to analyze more complex random structures such as random vectors. This is a first step to the world of high dimensional probability. We will first start by introducing the notion of random vectors and look at certain classical distributions like the multivariate normal distribution and the uniform multivariate distribution. After gaining some understanding of random vectors, we introduce a new class of random variables known as sub-Gaussian random variables that will help us extending our concentration results. By exploring these generalized distributions, we not only gain a deeper understanding of the behavior of random variables but also develop techniques that are crucial for analyzing high-dimensional random processes. These results serve as the foundation for subsequent chapters, where we apply these advanced concentration tools to empirical processes theory and statistical learning problems.

This chapter provides a bridge between the more restrictive results of Chapter 2 and the sophisticated methods of empirical process analysis in later sections, making it an essential step in our progression towards understanding ways to work with empirical processes.

3.1 Analysis of random vectors

We observed that the expectation and variance of random variables served as powerful tools for deriving the concentration inequalities we established. This naturally leads to the question: how can we extend the concepts of expectation and variance to random vectors?

The expectation of a matrix is just the matrix of the coordinates-wise expectation.

Definition 1. *(Variance-covariance matrix) The variance/covariance matrix of a random vector $X \in \mathbb{R}^n$ is :*

$$\operatorname{cov}(X) = \mathbb{E}(X-\mu)(X-\mu)^\top = \mathbb{E}XX^\top - \mu\mu^\top, \quad \textit{where } \mu = \mathbb{E}X.$$

Definition 2. *(Isotropic random vectors). A random vector X in $\mathbb{R}^n$ is called isotropic if*

$$\Sigma(X) = \mathbb{E}XX^\top = I_n$$

where I_n denotes the identity matrix in $\mathbb{R}^n$.

How can we transform a vector to be isotropic ?

Exercise 3.1.1: Reduction to isotropy

(a) Let Z be a mean zero, isotropic random vector in $\mathbb{R}^n$. Let $\mu \in \mathbb{R}^n$ be a fixed vector and Σ be a fixed $n \times n$ symmetric positive semidefinite matrix. Check that the random vector

$$X := \mu + \Sigma^{1/2} Z$$

has mean μ and covariance matrix $\operatorname{cov}(X) = \Sigma$.

(b) Let X be a random vector with mean μ and invertible covariance matrix $\Sigma = \operatorname{cov}(X)$. Check that the random vector

$$Z := \Sigma^{-1/2}(X - \mu)$$

is an isotropic, mean zero random vector.

Solution. Let Z be a mean 0 isotropic vector.

(a) We show that: $X := \mu + \Sigma^{\frac{1}{2}} Z$ has expectation μ and covariance Σ. First we prove the fact that if A is a matrix and X a vector (with right dimensions) then :

$$\mathbb{E}(AX) = A\mathbb{E}(X)$$

But $AX_i = \sum_{k=1}^n a_{i,k}x_k$, so

$$\mathbb{E}(AX)_i = \sum_{k=1}^n a_{i,k}\mathbb{E}(x_k) = (A\mathbb{E}(X))_i.$$

Therefore,

$$\mathbb{E}(X) = \mathbb{E}(\mu) + \Sigma^{\frac{1}{2}}\mathbb{E}(Z) = \mu.$$

Now we compute the variance-covariance matrix of X :

$$\begin{aligned}\mathbb{E}((X-\mu)(X-\mu)^T) &= \mathbb{E}\left((\Sigma^{\frac{1}{2}}Z)(\Sigma^{\frac{1}{2}}Z)^T\right) \\ &= \mathbb{E}\left(\Sigma^{\frac{1}{2}}ZZ^T\Sigma^{\frac{1}{2}T}\right) \\ &= \Sigma^{\frac{1}{2}}\mathbb{E}(ZZ^T)\Sigma^{\frac{1}{2}} \\ &= \Sigma\end{aligned}$$

(b) Now we let

$$Z := \Sigma^{\frac{-1}{2}}(X-\mu).$$

By linearity of expectation with see that Z has expectation 0. Computing we get:

$$\mathbb{E}((Z-\mu)(Z-\mu)^T) = \Sigma^{\frac{-1}{2}}\mathbb{E}(X-\mu)(X-\mu)^T(\Sigma^{\frac{-1}{2}})^T - \mathbb{E}(\mu\mu^T) = I_n.$$

So Z is an isotropic vector.

□

We next give a useful characterization of the notion of isotropy.

Lemma 1. *(Characterization of isotropy) A random vector X in $\mathbb{R}^n$ is isotropic if and only if*

$$\mathbb{E}\langle X, x\rangle^2 = \|x\|_2^2 \quad \textit{for all } x \in \mathbb{R}^n.$$

Proof. Recall that two symmetric $n \times n$ matrices A and B are equal if and only if $x^\top Ax = x^\top Bx$ for all $x \in \mathbb{R}^n$. (Check this!) Thus X is isotropic if and only if

$$x^\top\left(\mathbb{E}XX^\top\right)x = x^\top I_n x \quad \text{for all } x \in \mathbb{R}^n.$$

The left side of this identity equals $\mathbb{E}\langle X, x\rangle^2$ and the right side is $\|x\|_2^2$. This completes the proof. □

Here is a nice application of this result:

The symmetric Bernoulli distribution is *isotropic*. This can be easily seen again by checking $\mathbb{E}[\langle X, e_i\rangle^2] = 1$ for all $i = 1, \dots, n$, or for any $x \in \mathbb{R}^n$ by checking that

$$\mathbb{E}[\langle X, x\rangle^2] = \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 x_i^2\right] + \mathbb{E}\left[2 \sum_{1 \le i < j \le n} X_i X_j x_i x_j\right] = \|x\|_2^2.$$

Lemma 2. *: Let X be an isotropic random vector in $\mathbb{R}^n$. Then*

$$\mathbb{E}\|X\|_2^2 = n.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left(\|X\|_2^2\right) &= \mathbb{E}\left(X^\top X\right) = \mathbb{E}\left(\operatorname{tr}(X^\top X)\right) && \text{(viewing } X^\top X \text{ as a } 1 \times 1 \text{ matrix)}\\
&= \mathbb{E}\left(\operatorname{tr}(XX^\top)\right) && \text{(by the cyclic property of trace)}\\
&= \operatorname{tr}\left(\mathbb{E}\left(XX^\top\right)\right) && \text{(by linearity)}\\
&= \operatorname{tr}(I_n) && \text{(by isotropy)}\\
&= n.
\end{aligned}$$

□

Exercise 3.1.2: Distance between isotropic vectors

Let X and Y be independent, mean zero, isotropic random vectors in $\mathbb{R}^n$. Check that

$$\mathbb{E}\|X - Y\|_2^2 = 2n.$$

Using this lemma, we can compute the distance between isotropic random vectors with respect to the L^2 norm.

Solution. By linearity, we compute the distance between isotropic random vectors using the previous lemma: $\mathbb{E}(\|X - Y\|_2^2) = \mathbb{E}(X^T X) + \mathbb{E}(Y^T Y) - \mathbb{E}(X^T Y) + \mathbb{E}(Y^T X) = 2n$ Because $\mathbb{E}(X^T X) = \mathbb{E}(\|X\|_2^2)$ and $\mathbb{E}(Y^T X) = \mathbb{E}(X^T Y)$ because transposed real numbers are the same.

□

We are now ready to study some classical high dimensional distribution.

3.2 Multivariate uniform distribution

We say that a random vector $X \sim Unif\left(\sqrt{n}S^{n-1}\right)$ when X is uniformly distributed on the Euclidean sphere of $\mathbb{R}^n$ with radius $\sqrt{n}$ centered at the origin. We now check that the uniform distribution is isotropic.

Exercise 3.2.1: Uniform distribution on the sphere

Show that the spherically distributed random vector X is isotropic. Argue that the coordinates of X are not independent.

Solution. We use the characterization of isotropy of lemma 1. By the rotational invariance of the distribution of X, we have that projections $\mathbb{E}\left(\langle X, x\rangle^2\right)$ depends only on the length of x. Let $\{x = x_1, \ldots, x_n\}$ be orthogonals vectors of norm $\|x\|$. Then

$$n\mathbb{E}\left(\langle X, x\rangle^2\right) = \mathbb{E}\left(\sum_{k=1}^{n} \langle X, x_i\rangle\right) = \mathbb{E}\left(\|X\|^2\|x\|^2\right) = n\|x\|^2$$

And so the conclusion of lemma 1 tells us that X is isotropic. □

3.3 Multivariate normal distribution

We say that $g = (g_1, \ldots, g_n)$ has the *standard normal distribution* in $\mathbb{R}^n$, denoted

$$g \sim N(0, I_n),$$

if the coordinates g_i are independent standard normal random variables $N(0,1)$. The density of Z is then the product of the n standard normal densities (1.6), which is

$$f(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|x\|_2^2/2}, \quad x \in \mathbb{R}^n. \tag{3.4}$$

The standard normal distribution is isotropic. Indeed:
We can use lemma 2: X is isotropic if and only if for all unit vector u: $\mathbb{E}(\langle X, u\rangle^2) = 1$ But this is true because a sum of mean zero gaussians is a mean zero gaussian.

Exercise 3.3.1: Rotation invariance

Deduce the following properties from the rotation invariance of the normal distribution.

(a) Consider a random vector $g \sim N(0, I_n)$ and a fixed vector $u \in \mathbb{R}^n$. Then

$$\langle g, u \rangle \sim N(0, \|u\|_2^2).$$

(b) Consider independent random variables $X_i \sim N(0, \sigma_i^2)$. Then

$$\sum_{i=1}^n X_i \sim N(0, \sigma^2) \quad \text{where } \sigma^2 = \sum_{i=1}^n \sigma_i^2.$$

(c) Let G be an $m \times n$ Gaussian random matrix, i.e., the entries of G are independent $N(0,1)$ random variables. Let $u \in \mathbb{R}^n$ be a fixed unit vector. Then

$$Gu \sim N(0, I_m).$$

Solution. (a) Consider a random vector $g \sim N(0, I_n)$ and fix a vector u in $\mathbb{R}^n$, if we take $(e_1, ..., e_n)$ the canonical basis of $\mathbb{R}^n$ then :

$$\langle g, u \rangle = \sum_{k=1}^n u_i \langle g, u_i \rangle.$$

Now we know that each $u_i \langle g, u_i \rangle$ follow a $N(0, u_i^2)$ law so we have by sum of gaussians:

$$X \sim N(0, \|u\|_2^2$$

(b) If $X_i \sim N(0, \sigma_i^2)$ then $\sum_{i=1}^n \sim N(0, \sigma^2)$.
This result can easily be recover by computing the MGF of the sum and use independance. Then we know that MGF completely characterises the random variable.

(c) Let $u \in \mathbb{R}^n$ be a unit vector and G be a random $m \times n$ Gaussian-matrix. We have $Gu \sim N(0, I_m)$ Indeed,

$$(Gu)_i = \sum_{j=1}^n G_{i,j} u_j = \langle G_i, u \rangle$$

where G_i is the i-th row of G. Now since $G_i \sim N(0, I_n)$ by definition, we have that:

$$(Gu)_i \sim N(0, 1)$$

by question (a) (because u is a unit random vector). So this concludes the proof.

□

We next give a characterisation of gaussian random vectors.

Exercise 3.3.2: Characterisation of gaussians (via projections)

Let Y be a random vector in $\mathbb{R}^n$. Suppose that for all $u \in \mathbb{R}^n \quad u^T X$ has a uni variate Gaussian distribution. Let μ be the mean of Y and Σ be the covariance matrix of Y. Then the moment generating function of Y is

$$M_Y(u) = e^{u^T \mu + \frac{1}{2} u^T \Sigma u}$$

Solution. Let $v \in \mathbb{R}^d$ be arbitrary. Then $v^\top Y$ is scalar Gaussian with mean $v^\top \mu$ and variance $v^\top \Omega v$. Hence it has moment generating function:

$$M_{v^\top Y}(t) = \mathbb{E}\left(e^{tv^\top Y}\right) = \exp\left\{t(v^\top \mu) + \frac{t^2}{2}(v^\top \Omega v)\right\}.$$

Now take $t = 1$ and observe that

$$M_{v^\top Y}(1) = \mathbb{E}\left(e^{v^\top Y}\right) = M_Y(v).$$

Combining the two, we conclude that

$$M_Y(v) = \exp\left(v^\top \mu + \frac{1}{2} v^\top \Omega v\right), \quad v \in \mathbb{R}^d.$$

The moment generating function of Y is the same as the moment generating function of a $N(\mu, \Sigma)$ random variable and since MGFs completely determinate a random variable, we see that $Y \sim N(\mu, \sigma)$. □

The following exercise builds upon the framework developed so far to compute the distances of projections of a mean-zero Gaussian vector with covariance I_n.

Exercise 3.3.3:

Let $X \sim N(0, I_n)$.

(a) Show that, for any fixed vectors $u, v \in \mathbb{R}^n$, we have

$$\mathbb{E}\left(\langle X, u\rangle\langle X, v\rangle\right) = \langle u, v\rangle.$$

(b) Given a vector $u \in \mathbb{R}^n$, consider the random variable $X_u := \langle X, u\rangle$. From Exercise 3.3.1 we know that $X_u \sim N(0, \|u\|_2^2)$. Check that

$$\|X_u - X_v\|_{L_2} = \|u - v\|_2$$

for any fixed vectors $u, v \in \mathbb{R}^n$. (Here $\|\cdot\|_{L_2}$ denotes the norm in the Hilbert space L^2 of random variables)

Solution. (a) Let $X \sim N(0, I_n)$.

$$\mathbb{E}\left(\langle X, v\rangle\langle X, U\rangle\right) = \mathbb{E}(v^T X X^T u) = v^T \mathbb{E}(X X^T)u = v^t I_n u = \langle u, v\rangle.$$

(b)

$$\|X_u - X_v\|_{L_2}^2 = \mathbb{E}(X_u^2) + \mathbb{E}(X_v^2) - 2\mathbb{E}(X_u X_v) = \|u\|_2^2 - 2\langle u, v\rangle + \|v\|_2^2 = \|u - v\|_2^2.$$

By using question (a).Taking the root on both sides finished the claim.

□

We look at a last property of Gaussian random vectors.

Exercise 3.3.4:

Let G be an $m \times n$ Gaussian random matrix, i.e., the entries of G are independent $N(0,1)$ random variables. Let $u, v \in \mathbb{R}^n$ be unit orthogonal vectors. Prove that Gu and Gv are independent $N(0, I_m)$ random vectors.

Solution. We already proved that G_u and G_v are $N(0, I_m)$ vectors in exercise 3.3.3. Now we want to show independence. We remember that in the case of gaussian random vectors, uncorrelatedness and independence is the same.

So we need to show that $Cov(Gu, Gv) = 0$ But this comes naturally as:

$$Cov(Gu, Gv) = \mathbb{E}(Guv^T G^T) = 0$$

because u and v and orthogonal vectors. We now introduce a new class of random variable in order □

3.4 Sub-gaussian random variables

Although Hoeffding's inequality and Chernoff's inequality have valuable applications, they are limited to a very restricted class of random variables. Sub-Gaussian random variables provide a framework for extending concentration inequalities to a broader class of random variables, beyond just Bernoulli random variables. This class of random variables will prove to be highly effective for obtaining general results when dealing with empirical processes.

We first see that the p-norm of a random variable Z such that $Z \sim N(0,1)$ is bounded when divided by $\sqrt{p}$. This will turn out to be a characteristic property of sub-gaussians random variables.

Exercise 3.4.1:

Show that for $p \geq 1$, the random variable $X \sim N(0,1)$ satisfies

$$\|X\|_{L_p} = (\mathbb{E}|X|^p)^{1/p} = \sqrt{2}\left[\frac{\Gamma((1+p)/2)}{\Gamma(1/2)}\right]^{1/p}.$$

Deduce that

$$\|X\|_{L_p} = O(\sqrt{p}) \quad \text{as } p \to \infty.$$

Proof. 1. We first compute $\mathbb{E}(|X|^p)$. Using transformation formula:

$$\mathbb{E}(|X|^p) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} |t|^p e^{-t^2/2} dt.$$

Now since the integrand is even the integral can be written as

$$\frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} t^p e^{-t^2/2} dt.$$

Using the change of variable $u = t^2/2$, we get $du = tdt = \sqrt{2u}dt$ and $t^p = (2u)^{p/2}$, the integral becomes :

$$\mathbb{E}(|X|^p) = \frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} (2u)^{p/2} e^{-u}/(\sqrt{2u}) du = \frac{2^{p/2}}{\sqrt{\pi}}\int_{0}^{+\infty} u^{(p-1)/2} e^{-u} du.$$

Now taking the p-th root, we get the wanted result :

$$\mathbb{E}(|X|^p) = \sqrt{2}(\frac{\Gamma((1+p)/2)}{\Gamma(1/2)})^{1/p}.$$

2. Using Stirling's approximation for $\Gamma(x) \approx \left(\frac{x}{e}\right)^x \sqrt{\frac{2\pi}{x}}$ we have:

$$\Gamma\left(\frac{p+1}{2}\right) \approx \left(\frac{p+1}{2e}\right)^{(p+1)/2} \sqrt{\frac{4\pi}{p+1}} = 2^{(1-p)/2} e^{-(p+1)/2} (p+1)^{p/2} \sqrt{\pi}$$

Using this and simplifying $\sqrt{\pi}$ with $\Gamma(1/2)$, we get:

$$E(|X|^p)^{1/p} = \sqrt{2} \left(\frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma(1/2)}\right)^{1/p} \approx \sqrt{2} \cdot 2^{(1-p)/2p} \cdot e^{-(p+1)/2p} (p+1)^{1/2}$$

Now since the limit quantity is clearly bounded when dividing by $\sqrt{p}$, we get the wanted result .

□

The following theorem give some conditions that characterize sub-gaussian random variables.

Theorem 6. *(sub-gaussian properties) Let X be a random variable . Then the following properties are equivalent;*

(i) There exists $K_1 > 0$ such that the tails of X satisfy

$$\mathbb{P}\{|X| \geq t\} \leq 2 \exp\left(-t^2/K_1^2\right) \quad \textit{for all } t \geq 0.$$

(ii) There exists $K_2 > 0$ such that the moments of X satisfy

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p} \quad \textit{for all } p \geq 1.$$

(iii) There exists $K_3 > 0$ such that the MGF of X^2 satisfies

$$\mathbb{E}\exp\left(\lambda^2 X^2\right) \leq \exp\left(K_3^2 \lambda^2\right) \quad \textit{for all } \lambda \textit{ such that } |\lambda| \leq \frac{1}{K_3}.$$

(iv) There exists $K_4 > 0$ such that the MGF of X^2 is bounded at some point, namely

$$\mathbb{E}\exp\left(X^2/K_4^2\right) \leq 2.$$

The parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.

Moreover, if $\mathbb{E}X = 0$ then properties (i)–(iv) are also equivalent to the following one.

(v) There exists $K_5 > 0$ such that the MGF of X satisfies

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2) \quad \text{for all } \lambda \in \mathbb{R}.$$

We have already seen that a random variables X that follows a $N(0, 1)$ is a sub-gaussian random variable as it satisfy property **(ii)** of **Theorem 6** (see 3.4.1). Also, looking at property **(iv)** it can easily be checked that any bounded random variable is sub-gaussian.

The next exercise shows some property when satisfying property **(v)** of **Theorem 6**.

Exercise 3.4.2:

(a) Show that if $X \sim N(0, 1)$, the function $\lambda \mapsto \mathbb{E}\exp(\lambda^2X^2)$ is only finite in some bounded neighborhood of zero.

(b) Suppose that some random variable X satisfies $\mathbb{E}\exp(\lambda^2X^2) \leq \exp(K\lambda^2)$ for all $\lambda \in \mathbb{R}$ and some constant K. Show that X is a bounded random variable, i.e., $\|X\|_\infty < \infty$.

Proof. (a) Suppose $X \sim N(0, 1)$. Using the transformation formula, we get :

$$\mathbb{E}(\exp(\lambda^2X^2)) = \frac{2}{\sqrt{2\pi}}\int_0^{+\infty} e^{\lambda^2x^2}e^{-x^2/2}dx = \frac{2}{\sqrt{2\pi}}\int_0^{+\infty} e^{(\lambda^2-1/2)x^2}dx,$$

which is finite if and only if $\lambda \in (-\sqrt{2}/2, \sqrt{2}/2)$. Thus $\mathbb{E}(\exp(\lambda^2X^2))$ is finite only on an open bounded neighborhood of 0.

(b) Let $t \geq 0$ and $\lambda > 0$. Suppose we meet the condition stated in the exercise.

$$P(|X| \geq t) = P\left(\exp(\lambda^2X^2 \geq \lambda^2t^2)\right) \leq \frac{\mathbb{E}(\exp(X^2\lambda^2))}{\lambda^2t^2} \leq \frac{\exp(K^2\lambda^2)}{\lambda^2t^2},$$

where we used Markov's inequality and the bound on the MGF of X^2. This quantity goes to 0 as t goes to infinity. This show the set of X being infinity is of probability measure 0 thus X is P-almost surely bounded.

□

We now put a norm on the space of sub-gaussian random variables.

Definition 3. *(Sub-gaussian rvs and sub-gaussian norm) A random variable X that satisfies one of the equivalent properties (i)–(iv) 6 in Proposition 2.5.2 is called a sub-gaussian random variable. The sub-gaussian norm of X, denoted $\|X\|_{\psi_2}$, is defined to be the smallest K_4 in property (iv). In other words, we define*

$$\|X\|_{\psi_2} = \inf \left\{ t > 0 : \mathbb{E} \exp(X^2/t^2) \leq 2 \right\}.$$

One can check that the space of sub-gaussian random variable is a subspace of any $L^p(\Omega, T, P), p \geq 1$.

Exercise 3.4.3: Sub-gaussian norm

Check that $\|\cdot\|_{\psi_2}$ is indeed a norm on the space of sub-gaussian random variables.

Proof. Here we check that $||X||_{\psi_2}$ is a norm on the space of sub-gaussian random variables. First, we see that this norm is well defined thanks to property 4 of sub-gaussian random-variables. Now if X is such that $||X||_{\psi_2} = 0$. That means by definition that there exists a sequence of real number (t_n) such that $\lim_{n\to\infty} t_n = 0$ and such that $\forall n \in \mathbb{N}$ $E(\exp(X^2/t_n^2)) \leq 2$. By contradiction, let's assume that X is non zero on a non null set. That is the expectation of X^2 is not zero as well. We have :

$$\mathbb{E}(\exp(X^2/t_n^2)) = \mathbb{E}(\sum_{k=0}^{\infty} X^{2k}/t_n^{2k}k!) \tag{3.1}$$

$$= \sum_{k=0}^{\infty} \mathbb{E}(X^{2k})/t_n^{2k}k! \tag{3.2}$$

Using monotone convergence theorem to interchange the expectation with the infinite sum. Letting n go to infinity, we get that this quantity goes to infinity as we can interchange limit and sums for power-series like the exponential. This is our contradiction, indeed we stated that the MGF of X^2 was bounded by 2 for all t_n.

Now, we check

$$||aX||_{\psi_2} = |a|||X||_{\psi_2}, \forall a \in \mathbb{R}.$$

Let $f : x \mapsto e^{x^2}$. Pick $a \in \mathbb{R}$.It suffices to realize that

$$|a|||X||_{\psi_2}$$

is in the set

$$\{t : Ef\left(X/||X||_{\psi_2} t\right) \leq 2\}$$

Now triangular inequality: Let X, Y be sub-gaussians random variables. we notice that f is a convex increasing map therefore by Jensen's inequality, we have :

$$\mathbb{E}\left(f(\frac{|X+Y|)}{a+b}\right) \leq \frac{a}{a+b}\mathbb{E}\left(f(|X|/a)\right) + \frac{a}{a+b}\mathbb{E}\left(f(|Y|/b)\right)$$

Now letting

$$a = ||X||_{\psi_2}, b = ||Y||_{\psi_2},$$

we get that $a+b$ is in the set of:

$$\{t : \mathbb{E}\left(\exp(\frac{(X+Y)^2}{t^2})\right) \leq 2\},$$

so we get the result. □

This next exercise shows that we can get a sharp bound on the maximum of sub gaussian random variables using a union bound.

Exercise 3.4.4: Max of sub-gaussians

Let $X_1, X_2, \ldots$ be a sequence of sub-gaussian random variables, which are not necessarily independent. Show that

$$\mathbb{E} \max_i \frac{|X_i|}{\sqrt{1+\log i}} \leq CK,$$

where $K = \max_i \|X_i\|_{\psi_2}$. Deduce that for every $N \geq 2$, we have

$$\mathbb{E} \max_{i \leq N} |X_i| \leq CK\sqrt{\log N}.$$

Solution. Let $Y_i = \frac{X_i}{CK(1+log(i))}$ and let $t \geq 1$.

$$P\left(\exists i : Y_i \geq t\right) \leq \sum_{n=0}^{\infty} P(Y_i \geq t) \lesssim e^{-t^2} \quad \text{(by sub-gaussianity)}$$

using sigma-sub-additivity. then we use that $P(Y_i \geq t) \leq$ Therefore,we can compute the formula of expectation with tail for Y_i and get

$$\mathbb{E}(\max_i Y_i) = \int_0^\infty P(\exists i : Y_i \geq t)dt \leq \int_0^1 P(\exists i : Y_i \geq t)dt + \int_1^\infty e^{-t^2}dt = C$$

Thus, we get the result for question 1. By using linearity of the expectation and the fact that log is an increasing function, we get the desired bound for question 2. □

Lemma 3 (Centering)**.** *Let X be a sub-gaussian random variable. Then $X - \mathbb{E}(X)$ is also sub-gaussian and:*

$$\|X - \mathbb{E}X\|_{\psi_2} \leq C\|X\|_{\psi_2},$$

where C is a positive constant.

Proof. Using the fact that any constant is sub-gaussian (in particular $\mathbb{E}(X)$ is sub-gaussian), we get the claim using triangular inequality. □

We are now ready to state a more general Hoeffding's inequality.

3.5 General form of Hoeffding's inequality

The following theorem is a generalization of Hoeffding's inequality where we assumed that random variables had a bounded second moment. It will turn out to be useful for proving results about empirical processes in chapter 4.

Theorem 7. *(General Hoeffding's inequality) Let $X_1, \dots, X_N$ be independent, mean zero, sub-gaussian random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^N X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{K^2}\right)$$

where $K = \max_i \|X_i\|_{\psi_2}$.

Proof. First, from triangle inequality it holds that

$$\|\sum_{i=1}^N X_i\|_{\psi_2} \leq \sum_{i=1}^N \|X_i\|_{\psi_2} \leq \infty,$$

therefore $\sum_{i=1}^N X_i$ is a sub-Gaussian random variable. It suffices to show that:

$$\|\sum_{i=1}^N X_i\|_{\psi_2}^2 \lesssim \sum_{i=1}^N \|X_i\|_{\psi_2}^2.$$

To prove this, we need to use the independence of the random variables X_i. From the assumption that the random variables are mean zero, the sum is also from linearity of the expectation, hence, using **(v)** of Theorem 6, we have that for some absolute constant $C > 0$ that

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{N} X_i\right)\right] = \prod_{i=1}^{N}\mathbb{E}\left[\exp(\lambda X_i)\right] \leq \prod_{i=1}^{N}\exp(C\lambda\|X_i\|_{\psi_2}^2) = \exp\left(C\lambda\sum_{i=1}^{N}\|X_i\|_{\psi_2}^2\right),$$

where we used the independence to split the expectation into a product. □

3.6 Concentration of the norm

We start this section with a strong results that answers the question of where a random vector is likely to be located according to the $\|\cdot\|_{\psi_2}$-norm.

Theorem 8. *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, sub-Gaussian coordinates X_i that satisfy $\mathbb{E}X_i^2 = 1$. Then*

$$\big\| \|X\|_2 - \sqrt{n} \big\|_{\psi_2} \leq CK^2,$$

where $K = \max_i \|X_i\|_{\psi_2}$ and C is an absolute constant.

For the proof one can check the proof given in the book high-dimensional probability page 42. We give some consequences of this result in the form of exercises.

Exercise 3.6.1:

Show by using Theorem 8 that:

$$\sqrt{n} - CK^2 \leq \mathbb{E}\|X\|_2 \leq \sqrt{n} + CK^2.$$

Solution. The trick is to use the inequality $e^x \geq 1 + x$ to get

$$\mathbb{E}(exp(\frac{(\|X\|_2 - \sqrt{n})^2}{(\mathbb{E}\,\|X\|_2 - \sqrt{n})^2})) \geq 2$$

Thus we get,

$$|\mathbb{E}\,\|X\|_2 - \sqrt{n}| \leq \big\|\|X\|_2 - \sqrt{n}\big\|_{\psi_2} \leq CK^2$$

by definition of the $\|.\|_{\psi_2}$.

□

Exercise 3.6.2: Variance of the norm

Show by using Theorem 8 that:

$$\mathrm{Var}(\|X\|_2) \leq CK^4.$$

Solution. We now want to show that:

$$Var(\|X\|_2) \leq CK^4.$$

For this, we can expand the variance for the norm 2 of X and apply the previous exercise:

$$Var(\|X\|_2) = \mathbb{E}(\|X\|_2^2 - \mathbb{E}(\|X\|_2)^2 \leq n - (\sqrt{n} + CK^2)^2 \leq C^2K^4$$

□

Exercise 3.6.3: Variance of the norm under finite moment assumption

Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent coordinates X_i that satisfy $\mathbb{E}X_i^2 = 1$ and $\mathbb{E}X_i^4 \leq K^4$. Show that

$$\mathrm{Var}(\|X\|_2) \leq CK^4.$$

Solution. Here, we show that the conclusion of exercise 3.6.2 hold under the weaker assumptions that X_i's only have a finite moment of order 4. For that, we first expand and use linearity of the expectation :

$$\mathbb{E}(\|X\|_2^2 - n)^2 = \sum_{i,j} \mathbb{E}(X_i^2 X_j^2) + n^2 - 2n \sum_{i=1}^{n} \mathbb{E}(X_i^2)$$

Now using that $\mathbb{E}(X_i^2) = 1$, we get :

$$\mathbb{E}(\|X\|_2^2 - n)^2 \leq n(n-1) + K^4 n + n^2 - 2n^2 \leq K^4 n$$

Now this leads to:

$$\mathbb{E}(\|X\|_2 - \sqrt{n})^2 \leq K^4$$

To conclude, we need to argue that :

$$Var\left(\|X\|_2\right) \leq \mathbb{E}(\|X\|_2 - \sqrt{n})^2$$

i.e. can we replace $\sqrt{n}$ with $\mathbb{E}(\|X_2\|)$. But this is straight forward using exercise 3.6.1.

□

Exercise 3.6.4: Small ball probabilities

Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent coordinates X_i with continuous distributions. Assume that the densities of X_i are uniformly bounded by 1. Show that, for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(\|X\|_2 \leq \varepsilon\sqrt{n}\right) \leq (C\varepsilon)^n.$$

Solution. Here we prove the inequality using the same technique that is bounding the MGF of X^2 and using independance.

$$\begin{aligned} P\left(\|X\|_2^2 \leq \varepsilon^2 n\right) &= P\left(-\|X\|_2^2 \geq -\varepsilon^2 n\right) \\ &\leq e^{\lambda\varepsilon^2 n}\mathbb{E}\left[\exp\left(-\lambda\|X\|_2^2\right)\right] \\ &= e^{\lambda\varepsilon^2 n}\prod_{i=1}^{n}\mathbb{E}\left[\exp\left(-\lambda X_i^2\right)\right], \end{aligned}$$

Inserting

$$\mathbb{E}\left[\exp\left(-\lambda X_i^2\right)\right] \leq \int_{\mathbb{R}} e^{-\lambda x_i^2} f_i(x)\,dx \leq \int_{\mathbb{R}} e^{-\lambda x_i^2}\,dx = \sqrt{\frac{\pi}{\lambda}},$$

we derive

$$P\left(\|X\|_2^2 \leq \varepsilon^2 n\right) \leq \exp\left(\lambda\varepsilon^2 n - \frac{n}{2}\log\left(\frac{\lambda}{\pi}\right)\right) \leq (C\varepsilon)^n,$$

where the last inequality holds by optimizing over λ and choosing $\lambda = \frac{1}{2\varepsilon^2}$. □

We are now equipped to delve into empirical processes; however, we must first introduce the concept of random processes and additional tools to effectively handle them.

Chapter 4

Foundations of Empirical Processes

In this chapter, we start by introducing the definition of **random processes**, which lays the groundwork for discussing empirical processes and bounds in the latter part of the chapter. Next, we present the concept of $\varepsilon-$**coverings of a set**, a fundamental tool in understanding metric spaces and their structure. We follow with **Dudley's inequality**, a powerful inequality that plays a pivotal role in bounding the behavior of empirical processes using $\varepsilon-$coverings of a set. Dudley's inequality will prove particularly effective when applied to empirical processes indexed by Lipschitz functions, offering valuable insights into their convergence properties. After establishing these tools, we introduce the notion of VC dimension which provides a critical framework for understanding the complexity of boolean function classes and their generalization capabilities in statistical learning. After gaining a clear understanding of how the VC dimension operates, we proceed to its direct application in the context of empirical processes. By combining the insights from VC dimension with Dudley's inequality, we will derive effective bounds on the risk of error when classifying data.

4.1 Random processes

A random process (also known as a stochastic process) is a collection

$$(X_t)_{t \in T}$$

of random variables indexed by time or some other parameter.

There are lots of classical random processes:

1. $(\langle g, t \rangle)_{t \in T}$ where g is a gaussian random vector in $\mathbb{R}^n$ and T is a subset of $\mathbb{R}^n$
2. The standard Brownian motion: $(X_t)_{t \geq 0}$ where $(X_t)t \geq 0$ are independant, almost surely continuous random variables and the difference $X_t - X_s \sim N(0, t-s)\ \forall t > s > 0$.

We can generalise the notion of **covariance** with the **covariance function**:

$$\Sigma(t, s) := cov(X_t, X_s) = \mathbb{E}[X_t X_s], \quad t, s \in T \quad \text{where we assume for simplicity that } \mathbb{E}(X_t) = 0\ \forall t$$

We also define the increments of the random process as:

$$d(t, s) := \|X_t - X_s\|_{L^2} = \left(\mathbb{E}(X_t - X_s)^2\right)^{1/2}, \quad t, s \in T.$$

We can identify for example the increments of a random-walk $Z_n = \sum_{k=1}^{n} X_k$ where X_k have mean zero and variance one:

$$d(m, n)^2 = E(Z_n - Z_m)^2 = E(\sum_{k=n}^{m} X_k)^2 = (m-n)E(X_1)^2 = m - n$$

Thus we get the result for the increments of random-walks with random variable that have mean zero and variance 1.

Exercise 4.1.1: Covariances vs increments

We consider a random process $(X_t)_{t\in t}$.

1. Fix $s > t > 0$.

$$\begin{aligned} d(t,s)^2 = \mathbb{E}(X_t - X_s)^2 &= \Sigma(t,t) + \Sigma(s,s) - 2\Sigma(t,s) \\ &= (\Sigma(t,t) - \Sigma(s,s))^2 \end{aligned}$$

   Thus we get a formulation

2. If we assume that 0 is in the random process then we get that that each element of the process has mean 0 so the formula above allows us to isolate $\Sigma(t,s)$ and get :

$$\Sigma(t,s) = \frac{1}{2}\left(\Sigma(t,t) + \Sigma(s,s) - \|X_t - X_s\|_{L^2}^2\right).$$

4.1.1 Gaussian processes

Definition 4. *A random process $(X_t)_{t\in T}$ is called a* ***Gaussian process*** *if, for any finite subset $T_0 \subset T$, the random vector $(X_t)_{t\in T_0}$ has normal distribution. Equivalently, $(X_t)_{t\in T}$ is Gaussian if every finite linear combination $\sum_{t\in T_0} a_t X_t$ is a normal random variable. (This equivalence is due to standard properties of normal distributions we showed in 3.3.2.)*

Lemma 4. *Let Y be a mean zero Gaussian random vector in $\mathbb{R}^n$. Then there exist points $t_1, \ldots, t_n \in \mathbb{R}^n$ such that*

$$Y \equiv (\langle g, t_i\rangle)_{i=1}^n, \quad \textit{where } g \sim N(0, I_n).$$

Here "$\equiv$" means that the distributions of the two random vectors are the same.

Solution. Let Σ denote the covariance matrix of Y. Then we may realize

$$Y \equiv \Sigma^{1/2} g \quad \text{where } g \sim N(0, I_n)$$

. Next, the coordinates of the vector $\Sigma^{1/2}g$ are $\langle t_i, g\rangle$ where t_i denote the rows of the matrix $\Sigma^{1/2}$. This completes the proof. □

4.2 Tools to work with empirical processes

4.2.1 Nets covering

Definition 5. *Let (T, d) be a metric space, $K \subset T$, and $\varepsilon > 0$.*

*(a) ε-**net:** A subset $\mathcal{N} \subset K$ is an ε-net of K if every point of K is within a distance ε of some point of $\mathcal{N}$,*

$$\forall x \in K \; \exists x_0 \in \mathcal{N} : \; d(x, x_0) \leq \varepsilon.$$

*(b) **Covering number:** The smallest possible cardinality of an ε-net of K is the covering number of K and is denoted $\mathcal{N}(K, d, \varepsilon)$. Equivalently, $\mathcal{N}(K, d, \varepsilon)$ is the smallest number of closed balls with centres in K and radii ε whose union covers K.*

*(c) ε-**separated sets:** A subset $\mathcal{N} \subset T$ is ε-separated if $d(x, y) > \varepsilon$ for all distinct points $x, y \in \mathcal{N}$.*

*(d) **Packing number:** The largest possible cardinality of an ε-separated subset of K is called the packing number of K and denoted $\mathcal{P}(K, d, \varepsilon)$.*

Definition 6. *(Hamming cube). The Hamming cube $\{0,1\}^n$ consists of all binary strings of length n. The Hamming distance $d_H(x,y)$ between two binary strings is defined as the number of bits where x and y disagree, i.e.*

$$d_H(x,y) := \#\{i : x(i) \neq y(i)\}, \quad x, y \in \{0,1\}^n.$$

Endowed with this metric, the Hamming cube is a metric space $(\{0,1\}^n, d_H)$, which is sometimes called the Hamming space.

Exercise 4.2.1: Hamming cube distance is a distance

1. **Non-negativity and definit:** By definition, $d_H(x,y)$ is the number of positions at which x and y differ. Since this is a count, we have $d_H(x,y) \geq 0$. Moreover, $d_H(x,y) = 0$ if and only if $x = y$, as there are no differing positions in this case.$d_H(x,y) = 0$ if and only if $x = y$. This follows directly from the definition, as $d_H(x,y)$ counts the number of positions where x and y differ. If they do not differ at any position, then $x = y$.

2. **Symmetry:** The Hamming distance $d_H(x,y)$ is symmetric because the number of positions where x and y differ is the same as the number of positions where y and x differ. Thus, $d_H(x,y) = d_H(y,x)$ for all $x, y \in \{0,1\}^n$.

3. **Triangle Inequality:** For any $x, y, z \in \{0,1\}^n$, the Hamming distance satisfies

$$d_H(x,z) \leq d_H(x,y) + d_H(y,z).$$

This is because each differing position between x and z must either differ between x and y or between y and z. Thus, the total number of differing positions between x and z cannot exceed the sum of those between x and y and between y and z.

4.2.2 Dudley's inequality and chaining

We introduce sub-gaussian increments and present Dudley's inequality.

Definition 7 (Sub-gaussian increments). *Consider a random process $(X_t)_{t\in T}$ on a metric space (T,d). We say that the process has sub-gaussian increments if there exists $K \geq 0$ such that*

$$\|X_t - X_s\|_{\psi_2} \leq Kd(t,s) \quad \textit{for all } t,s \in T.$$

We are now ready to present Dudley's inequality. It gives a bound on the magnitude of a mean zero sub-gaussian random process using only the covering number of the set of indexation of the process.

Theorem 9. *(Dudley's inequality) Let $(X_t)_{t\in T}$ be a mean zero random process on a metric space (T,d) with sub-gaussian increments. Then*

$$\mathbb{E}\sup_{t\in T} X_t \leq CK\int_0^\infty \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon.$$

Proof. We first prove the discrete version of Dudley's inequality.

$$\mathbb{E}\sup_{t\in T} X_t \leq CK\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}(T,d,2^{-k})}.$$

We may assume that T is finite and the constant K of our increment is 1. Indeed, applying Dudley's inequality to $\frac{X_t}{K}$ that has sub-gaussian increment constant K=1 gives us the claim. Also, we may assume that T is finite because by definition of the supremum, having the claim for all subset $T_0 \subseteq T$ that is finite, means having the claim for T. Then, if we have the discrete version of Dudley's inequality automatically gives us the continuous version by expressing $2^{-k} = 2\int_{2^{-k-1}}^{2^{-k}} d\varepsilon$ and summing over $\mathbb{Z}$.

Step 1: The first step of the proof is the chaining set up. Let $\varepsilon_k = 1/2^k, \quad k \in \mathbb{Z}$, and let T_k be ε_k-nets of T with $|T_k| = \mathcal{N}(T,d,\varepsilon_k)$.
Since T is finite, there exists integers k_1 and k_2 such that

$$T_{k_1} = \{t_0\}(\text{ for some } t_0 \in T) \text{ and } T_{k_2} = T. \quad \textbf{(*)}$$

For a point $t \in T$, let $\pi_k(t)$ denote a closest point in T_k, so we have

$$d(t,\pi_k(t)) \leq \varepsilon_k.$$

Since $\mathbb{E}X_{t_0} = 0$, we have:

$$\mathbb{E}\sup_{t\in T} X_t = \mathbb{E}\sup_{t\in T}(X_t - X_{t_0}).$$

We can express $X_t - X_{t_0}$ as a telescoping sum; imagine moving from t_0 to t through a sequence of points $\pi_k(t)$, each providing increasingly precise approximations of t:

$$X_t - X_{t_0} = (X_{\pi_\kappa(t)} - X_{t_0}) + (X_{\pi_{\kappa+1}(t)} - X_{\pi_\kappa(t)}) + \cdots + (X_t - X_{\pi_\kappa(t)}),$$

Now, using **(*)**, we get :

$$X_t - X_{t_0} = \sum_{k=\kappa+1}^{K} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \right).$$

Now, by taking the supremum and summing, we get:

$$\mathbb{E} \sup_{t \in T} (X_t - X_{t_0}) \leq \sum_{k=\kappa+1}^{K} \mathbb{E} \sup_{t \in T} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \right).$$

It now remains to bound the increments this can be derived from subgaussianity of the random variables:

$$\begin{aligned}
\|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\|_{\psi_2} &\leq d(\pi_k(t), \pi_{k-1}(t)) \text{ since } K = 1 \\
&\leq d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \qquad \text{(by triangle inequality)} \\
&\leq \varepsilon_k + \varepsilon_{k-1} \\
&\leq 2\varepsilon_{k-1}.
\end{aligned}$$

We can now conclude from using Exercise 3.4.4 that a maximum of N subgaussian random variables is bounded by $CL\log(N)$ for some $C > 0$ which gives us:

$$\mathbb{E} \sup_{t \in T} \left(X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \right) \leq C\varepsilon_{k-1}\sqrt{\log |T_k|}.$$

and the proof is completed.

□

We now present an example of an empirical process indexed by L-Lipschitz function and look at some application of Dudley's inequality to it.

If we take $\mathcal{F} = \{f : \mathbb{R} \mapsto \mathbb{R} : \ \|f\|_{Lip} \leq L\}$ the set of all Lipschitz-continuous functions than have a Lipschitz norm less than L, then the random process :

$$X_f = \frac{1}{n} \sum_{k=1}^{n} f(X_k) - E(f(X))$$

where $f \in \mathcal{F}$ and X_i are independent identically distributed random variables with the same law as X is an empirical process.

Here, we give an application of Dudley's inequality to the empirical process we just presented.

Theorem 10 (Uniform law of large numbers). *Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. random variables taking values in $[0,1]$ and $\mathcal{F}$ be the class of L-lipschitz functions. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \leq \frac{CL}{\sqrt{n}}.$$

Proof. We may assume that L=1 as we can consider $\frac{f}{L}$ for $f \in \mathcal{F}$ and then apply the result of the theorem with the function $\frac{f}{L}$ that has Lipschitz constant 1. we then recover the claim by multiplying by L.

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \leq \frac{CL}{\sqrt{n}}.$$

Our goal is to apply Dudley's inequality to the empirical process:

$$X_f = \frac{1}{n} \sum_{k=1}^{n} f(X_k) - E(f(X))$$

In order to bound the order of magnitude.
Now we need to check that the empirical process we defined just above has sub-gaussian increment.

$$\|X_f - X_s\|_{\psi_2} = \frac{1}{n} \left\| \sum_{k=1}^{n} Z_i \right\|_{\psi_2}$$

where $Z_i = (f-g)(X_i) + E(f-g)(X)$
Using that Z_i have mean 0 and are independent: we have:

$$\|X_f - X_g\|_{\psi_2} \lesssim \frac{1}{n} \left(\sum_{i=1}^{n} \|Z_i\|_{\psi_2}^2 \right)^{1/2}.$$

Now using centering lemma we proved in chapter 2, we get:

$$\|Z_i\|_{\psi_2} \lesssim \|(f-g)(X_i)\|_{\psi_2} \lesssim \|f-g\|_\infty.$$

So finely, we have :

$$\|X_f - X_g\|_{\psi_2} \lesssim \frac{1}{n} \cdot n^{1/2} \|f-g\|_\infty = \frac{1}{\sqrt{n}} \|f-g\|_\infty.$$

Our random variables have sub-gaussian increments with respect to the L^∞ distance, we can now use Dudley's inequality to the random process in order to bound the order of magnitude but first we can see that functions have L^∞ norm bounded by 1 so we can just stop the integral at 1 since the covering number for $\varepsilon > 1$ will just be 1.

$$\mathbb{E}\sup_{f\in\mathcal{F}}|X_f| = \mathbb{E}\sup_{f\in\mathcal{F}}|X_f - X_0| \lesssim \frac{1}{\sqrt{n}}\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)}\, d\varepsilon.$$

Now, in the next exercise, we will prove that

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$$

is bounded by $\left(\frac{2}{\varepsilon}\right)^{\frac{2}{\varepsilon}}$ for $\varepsilon \in (0,1)$ which concludes the proof. □

Exercise 4.2.2: Bound on the covering number of Theorem 10

We want to show the bound

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq \left(\frac{2}{\varepsilon}\right)^{\frac{2}{\varepsilon}} \qquad \forall\varepsilon \in (0,1)$$

Solution. Lets fix $\varepsilon \in (0,1)$. To outline the proof, we construct a grid with spacing ϵ along the y-axis and ϵ along the x-axis. Next, we consider functions defined on this grid that are piecewise linear, where each segment has a slope of either $+1$ or -1. There are approximately $1/\epsilon$ initial points, and from each of these points, there are $2^{1/\epsilon}$ possible slope configurations. It follows that this collection forms an $O(\epsilon)$ covering which concludes the proof. □

In the next section, we present a new tool to obtain effective bounds on the class of boolean functions.

4.2.3 VC dimension

The **VC dimension (Vapnik-Chervonenkis dimension)** is a fundamental concept in statistical learning theory that measures the complexity or capacity of a class of functions (or hypothesis space).

Definition 8 (VC dimension)**.** *Consider a class $\mathcal{F}$ of Boolean functions on some domain Ω. We say that a subset $\Lambda \subseteq \Omega$ is shattered by $\mathcal{F}$ if any function $g : \Lambda \to \{0,1\}$ can be obtained by restricting some function $f \in \mathcal{F}$ onto Λ.*

The VC dimension of $\mathcal{F}$, denoted $\mathrm{vc}(\mathcal{F})$, is the largest cardinality of a subset $\Lambda \subseteq \Omega$ shattered by $\mathcal{F}$.

Exercise 4.2.3: VC dimension of intervals

Let $\mathcal{F}$ be the class of indicators of all closed intervals in $\mathbb{R}$, that is

$$\mathcal{F} := \left\{ 1_{[a,b]} : a, b \in \mathbb{R},\, a \leq b \right\}.$$

We claim that there exists a two-point set $\Lambda \subset \mathbb{R}$ that is shattered by $\mathcal{F}$, and thus

$$\mathrm{vc}(\mathcal{F}) \geq 2.$$

Solution. take $\Omega = \{0,1\}$ be our two points of $\mathbb{R}$. We verify that this two points set is shattered by $\mathcal{F}$. Consider any boolean function from Ω. If this boolean function takes value 1,1 (resp 0,0) then it suffices to take the indicator of $[-1,1]$ (resp $[100,200]$). Else, without loss of generality we can assume that it takes value 1 on 0 and 0 on value 1 (Even if it means symmetrizing). Then take the interval $[\frac{1}{2}, 1]$. Then the above function is the restriction of this indicator function to Ω. So this proves that Ω is shattered by $\mathcal{F}$ thus the result.

□

Exercise 4.2.4: VC dimension of pairs of interval

Let $\mathcal{F}$ be the class of indicators of sets of the form $[a,b] \cup [c,d]$ in $\mathbb{R}$. Show that

$$\mathrm{vc}(\mathcal{F}) = 4.$$

Solution. It is essentially the same idea as presented in the exercise just above. We show that any boolean function from 4 ordered points $x_1 < \cdots < x_4$ is the indicator of $\mathcal{F}$ by doing a disjunction of case. Then we show that no boolean function from 5 ordered points is shattered by $\mathcal{F}$. □

Exercise 4.2.5: VC dimension of Circles

Let $\mathcal{F}$ be the class of indicators of all circles in $\mathbb{R}^2$. Show that

$$\mathrm{vc}(\mathcal{F}) = 3.$$

Solution. First we prove that

$$\mathrm{vc}(\mathcal{F} \geq 3)$$

Indeed, if we pick for example 3 points x_1, x_2, x_3 of the plane $\mathbb{R}^2$ that are the vertices of some equilateral triangle, then it is easy to check that any boolean function from these three points can be written as the indicator of some circle. For example, take the function which is 0 on x_1 and 1 on x_2, x_3. Then the circle that has radius $\frac{1}{2}\|x_3 - x_2\|$ and is centered at $\frac{x_3+x_1}{2}$ contains the points x_2 and x_3 and doesn't contain x_1 by the way we chose x_1, x_2, x_3. This proves that the boolean function we just picked is this indicator of some circle. Thus we proved that:

$$\mathrm{vc}(\mathcal{F} \geq 3).$$

Now, we show that $\mathrm{vc}(\mathcal{F} < 4)$ which will conclude on the result. Indeed, lets pick 4 points $x_1, \ldots, x_4$ in $\mathbb{R}^2$. Then take the two points that are furthest to each other. Without loss of generality, these two points are x_1 and x_3. Now, we want to show that boolean functions from $\Omega = \{x_1, \ldots, x_4\}$ cannot be restriction of indicators of circle. Indeed, take the function that takes value 1 for x_1, x_3. This means that x_1, x_3 are in some circle. But now this would mean that this circle also contains x_2 or x_4, by the way we chose points x_1 and x_3 .This concludes that there is no function that is the indicator of a circle that can take value 0 at both x_2 and x_4.

□

Exercise 4.2.6:

Give the definition of the VC dimension for a class of subsets of Ω without mentioning any functions.

Solution. The **VC dimension** of a class $\mathcal{C}$ of subsets of a set Ω is the largest integer d such that there exists a subset of Ω of size d, denoted $\{x_1, x_2, \ldots, x_d\}$, that is **shattered** by $\mathcal{C}$.

A subset $\{x_1, x_2, \ldots, x_d\}$ is shattered by $\mathcal{C}$ if, for every subset $S \subseteq \{x_1, x_2, \ldots, x_d\}$, there exists a set $C \in \mathcal{C}$ such that

$$C \cap \{x_1, x_2, \ldots, x_d\} = S.$$

If no such d exists, the VC dimension is infinite. □

Lemma 5 (Pajor's lemma)**.** *Let $\mathcal{F}$ be a class of Boolean functions on a finite set Ω. Then*

$$|\mathcal{F}| \leq |\{\Lambda \subseteq \Omega : \Lambda \textit{ is shattered by } \mathcal{F}\}| .$$

We include the empty set $\Lambda = \emptyset$ in the counting on the right side.

Proof. The proof is done by induction on the cardinality of Ω for more details, see [1, Chapter 8]. □

Lemma 6 ((Sauer-Shelah Lemma))**.** *Let $\mathcal{F}$ be a class of Boolean functions on an n-point set Ω. Then*

$$|\mathcal{F}| \leq \sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{en}{d}\right)^d$$

where $d = \mathrm{vc}(\mathcal{F})$.

Proof. Pajor's Lemma states that $|\mathcal{F}|$ is bounded by the number of subsets $\Lambda \subseteq \Omega$ that are shattered by $\mathcal{F}$. The cardinality of each such set Λ is bounded by $d = \mathrm{vc}(\mathcal{F})$, according to the definition of VC dimension. Thus,

$$|\mathcal{F}| \leq |\{\Lambda \subseteq \Omega : |\Lambda| \leq d\}| = \sum_{k=0}^{d} \binom{n}{k},$$

since the sum on the right-hand side gives the total number of subsets of an n-element set with cardinalities at most k. This proves the first inequality of Sauer-Shelah. □

Theorem 11 (Covering number via the VC dimension)**.** *Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ). Then, for every $\epsilon \in (0, 1)$, we have*

$$\mathcal{N}(\mathcal{F}, L^2(\mu), \epsilon) \leq \left(\frac{2}{\epsilon}\right)^{Cd}$$

where $d = \mathrm{vc}(\mathcal{F})$.

The proof can be found in [1, Chapter 8].

4.3 Empirical processes via the VC dimension

In a previous section, we demonstrated how to bound an empirical process, focusing on a class of functions that are Lipschitz continuous. In this section, our objective is to establish an alternative bound for the empirical process, this time utilizing a bound specific to Boolean functions. To achieve this, we will leverage the concept of VC dimension as a new tool to derive the desired bound.

Theorem 12 (Empirical process via VC dimension)**.** *Let $\mathcal{F}$ be a class of Boolean functions on a probability space (Ω, Σ, μ) with finite VC dimension* $\mathrm{vc}(\mathcal{F}) \geq 1$. *Let $X, X_1, X_2, \ldots, X_n$ be independent random points in Ω distributed according to the law μ. Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \leq C \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}.$$

Proof. The general idea is to apply Dudley's inequality to our random process and then use the theorem of covering number via the VC dimension (for more details see ”empirical process via VC dimension” [1, Chapter 8]). It is useful to pre-process our random process using symmetrization. □

Exercise 4.3.1: Symmetrization of the random process

Let $\mathcal{F}$ be a class of functions on a probability space (Ω, Σ, μ). Let $X, X_1, X_2, \ldots, X_n$ be random points in Ω distributed according to the law μ. Prove that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| \leq 2 \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|,$$

where $\varepsilon_1, \varepsilon_2, \ldots$ are independent symmetric Bernoulli random variables (which are also independent of $X_1, X_2, \ldots$).

Solution. Let $Z_i = \frac{1}{n}(f(X_i)) - f(X))$ Let (Z_i') be an independent copy of the random vectors (Z_i). Since $\sum_i Z_i'$ has zero mean, we have

$$p := \mathbb{E} \left\| \sum_i Z_i \right\| \leq \mathbb{E} \left\| \sum_i Z_i - \sum_i Z_i' \right\| = \mathbb{E} \left\| \sum_i (Z_i - Z_i') \right\|.$$

The inequality here is an application of the following version of Lemma 6.1.2 [1] for independent random vectors Y and Z:

$$\text{if } \mathbb{E}Z = 0 \quad \text{then} \quad \mathbb{E}\|Y\| \leq \mathbb{E}\|Y + Z\|.$$

Next, since $(Z_i - Z_i')$ are symmetric random vectors, they have the same distribution as $\varepsilon_i(Z_i - Z_i')$ (see Exercise 6.4.1) [1]. Then

$$\begin{aligned} p &\leq \mathbb{E}\left\| \sum_i \varepsilon_i (Z_i - Z_i') \right\| \\ &\leq \mathbb{E}\left\| \sum_i \varepsilon_i Z_i \right\| + \mathbb{E}\left\| \sum_i \varepsilon_i Z_i' \right\| \quad \text{(by triangle inequality)} \\ &= 2\mathbb{E}\left\| \sum_i \varepsilon_i Z_i \right\| \quad \text{(since the two terms are identically distributed).} \end{aligned}$$

□

We now give an important application of theorem 12, which is the Glivenko-Cantelli theorem. This theorem answers a basic question we may ask ourselves: How to estimate the distribution of a random variable X given some sample $X_1, \ldots, X_n$.

Suppose we have a collection of i.i.d. random variables $X_1, \ldots, X_n$, all following the same distribution as X. It is natural to estimate $F(x)$ by calculating the proportion of sample points satisfying $X_i \leq x$. This leads to the empirical distribution function, defined as:

$$F_n(x) := \frac{|\{i \in [n] : X_i \leq x\}|}{n}, \quad x \in \mathbb{R}.$$

Observe that $F_n(x)$ is a random function as it depend on the random variables $X_1, \ldots, X_n$

Theorem 13 (Glivenko-Cantelli-Theorem). *Let $X_1, \ldots, X_n$ be independent random variables with a common cumulative distribution function F. Then,*

$$\mathbb{E}\|F_n - F\|_\infty = \mathbb{E}\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \frac{C}{\sqrt{n}}.$$

Proof. We consider the class of functions : $\mathcal{F} = \{1_{(-\infty, a]}, a \in \mathbb{R}\}$ and let μ be the measure on $\mathbb{R}$ to be the law of X_i. Then we know from exercise 4.2.3 that the VC dimension of $\mathcal{F}$ is 2. Thus we can immediately conclude from theorem 12.

□

4.4 Application: Statistical learning theory

Statistical learning theory is a mathematical framework that explores how machines can learn patterns from data, focusing on understanding the performance of learning algorithms in predicting unseen data by analyzing the trade-off between fitting the training data and ensuring generalization to new data, using concepts like hypothesis spaces, risk minimization, and complexity measures such as VC dimension to guide the design and evaluation of these algorithms. A typical problem in statistical learning can be stated as follows. Consider a function $T : \Omega \mapsto \mathbb{R}$ on some set Ω, which we call a target function. Suppose that T is unknown. We want to learn T from its values on a finite sample of points $X_1 \ldots X_n \in \Omega$. Thus our training data is:

$$(x_i, T(x_i) \quad i = 1, \ldots, n$$

our goal is to make some good prediction on some new value $T(X)$ for X outside of our training sample based on the data we have. In this section, we focus on a specific class of functions defined on a set $\mathcal{F}$, where T is a Boolean function (taking values 0 or 1). The function T classifies points within the set Ω. This important category of learning tasks is referred to as **classification problems.**

4.4.1 Risk fit and complexity

In a learning problem, the solution can be represented as a function $f : \Omega \to \mathbb{R}$. Ideally, we want f to approximate the target function T as closely as possible. To achieve this, we aim to select f that minimizes the **risk**, defined as

$$R(f) := \mathbb{E}\left[(f(X) - T(X))^2\right].$$

Here, X is a random variable that follows the distribution $\mathbb{P}$, which is the same as the distribution of the sample points $X_1, \ldots, X_n \in \Omega$.

In classification tasks, both T and f are Boolean-valued functions. In this case, the risk becomes

$$R(f) = \mathbb{P}(f(X) \neq T(X)).$$

Given that the set of functions f can be extremely large, it is common practice to restrict our search to a more constrained subset, referred to as the **Hypothesis Space**.

4.4.2 Empirical Risk

Based on a given data, what would be the function that best approximates

Definition 9 (Empirical risk)**.** *The empirical risk for a function $f : \Omega \mapsto \mathbb{R}$ is defined as:*

$$R_n(f) = \frac{1}{n}\sum_{k=1}^{n}(f(X_i) - T(X_i))^2$$

We will denote by $f_n^* := \arg\min_{f\in\mathcal{F}} R_n(f)$ the function that minimizes the empirical risk over the hypothesis space.f_n^* and $R_n(f)$ are quantities that can be computed from a given data. Now, we would like to quantify how big the *excess risk* is that is how big is

$$R(f_n^*) - R(f^*)$$

We give an answer to this question in the next section.

4.4.3 Bounding the Excess Risk by the VC dimension

Theorem 14 (Excess Risk via VC dimension)**.** *Assume that the target T is a Boolean function, and the hypothesis space $\mathcal{F}$ is a class of Boolean functions with finite VC dimension* $\mathrm{vc}(\mathcal{F}) \geq 1$*. Then*

$$\mathbb{E}R(f_n^*) \leq R(f^*) + C\sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}.$$

We deduce this theorem from a uniform deviation inequality that we proved in Theorem12. The following elementary observation will help us connect these two results.

Lemma 7 (Excess risk via uniform deviations)**.**

$$R(f_n^*) - R(f^*) \leq 2\sup_{f\in\mathcal{F}}|R_n(f) - R(f)|$$

Proof. Denote $\varepsilon := \sup_{f\in\mathcal{F}}|R_n(f) - R(f)|$. Then,

$$\begin{aligned} R(f_n^*) &\leq R_n(f_n^*) + \varepsilon && \text{(since } f_n^* \in \mathcal{F} \text{ by construction)} \\ &\leq R_n(f^*) + \varepsilon && \text{(since } f_n^* \text{ minimizes } R_n \text{ in the class } \mathcal{F}) \\ &\leq R(f^*) + 2\varepsilon && \text{(since } f^* \in \mathcal{F} \text{ by construction).} \end{aligned}$$

Subtracting $R(f^*)$ from both sides, we obtain the desired inequality. □

Bounding the excess risk. By Lemma 4.4.3, it suffices to show that

$$\mathbb{E} \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| \lesssim \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}.$$

Recalling the definitions of the empirical and true (population) risk, we can rewrite the left-hand side as

$$\mathbb{E} \sup_{\ell \in \mathcal{L}} \left| \frac{1}{n} \sum_{i=1}^{n} \ell(X_i) - \mathbb{E}\ell(X) \right|,$$

where $\mathcal{L}$ is the class of Boolean functions defined as

$$\mathcal{L} = \{(f - T)^2 : f \in \mathcal{F}\}.$$

let us recall that in the proof of Theorem 12, we first bounded the term by

$$\frac{1}{\sqrt{n}} \mathbb{E} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{L}, L^2(\mu_n), \epsilon)} \, d\epsilon,$$

up to an absolute constant factor. It is not hard to see that the covering numbers of $\mathcal{L}$ and $\mathcal{F}$ are related by the inequality

$$\mathcal{N}(\mathcal{L}, L^2(\mu_n), \epsilon) \le \mathcal{N}(\mathcal{F}, L^2(\mu_n), \epsilon) \quad \text{for any } \epsilon \in (0,1).$$

Thus, we may replace $\mathcal{L}$ by $\mathcal{F}$ in the bound, paying the price of an absolute constant factor. Following the remaining steps of the proof of Theorem 12, we conclude that is bounded by

$$\sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}$$

as desired. □

Appendix

We gather in this section a collection of result that can be interesting when studying probabilities.

Appendix

Limit theorems

Theorem 15. *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and finite variance. Show that:*

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| = O\left(\frac{1}{\sqrt{N}}\right) \quad \text{as } N \to \infty.$$

Proof. By the *Central Limit Theorem*

$$\frac{1}{\sqrt{N}}\left(\sum_{i=1}^{N} X_i - N\mu\right) \to Z \quad \text{where } Z \sim N(0, \sigma^2).$$

Thus:

$$\sqrt{N} \cdot \left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \to |Z|.$$

Since $\mathbb{E}(|Z|)$ is finite, we conclude that:

$$\mathbb{E}\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| = O\left(\frac{1}{\sqrt{N}}\right).$$

□

Theorem 16 (Poisson limit theorem)**.** *Let $X_{N,i}$, $1 \leq i \leq N$, be independent random variables $X_{N,i} \sim Ber(p_{N,i})$, and let*

$$S_N = \sum_{i=1}^{N} X_{N,i}.$$

Assume that, as $N \to \infty$,

$$\max_{i \leq N} p_{N,i} \to 0 \quad and \quad \mathbb{E} S_N = \sum_{i=1}^{N} p_{N,i} \to \lambda < \infty.$$

Then, as $N \to \infty$,

$$S_N \to Pois(\lambda) \quad in\ distribution.$$

Lemma 8 (tail of the normal distribution)**.** *Let $g \sim N(0,1)$. Then for all $t > 0$, we have*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}\{g \geq t\} \leq \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Exercise 4.4.1: Truncated normal distribution

Let $g \sim N(0,1)$. Show that for all $t \geq 1$, we have

$$\mathbb{E}\, g^2 \mathbf{1}_{\{g>t\}} = t \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} + \mathbb{P}\{g > t\} \leq \left(t + \frac{1}{t}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Proof. Let $g \sim N(0,1)$, let $t \geq 1$. Using lemma 8 about tails of the normal distribution, we have that $P(g \geq t) \leq \frac{1}{t} \cdot \frac{1}{\sqrt{2pi}} e^{-t^2/2}$ So we have the result for the inequality. Now we want to show that the expectation matches the identity. Using transfer formula, we get:

$$\mathbb{E}(g^2 1_{\{g>t\}}) = \int_R x^2 1_{x>t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Now we integrate by part by integrating $x.e^{-x^2}$ and differentiating $x \cdot 1_{x>t}$ to get the result. □

The following useful lemma helps us to understand the link between sub-gaussian and sub-exponential random variables:

Lemma 9 (Product of sub-gaussians is sub-exponential)**.** *Let X and Y be sub-gaussian random variables. Then XY is sub-exponential. Moreover,*

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}.$$

Proof. This lemma follows easily from the important Young's inequality and the monotonicity of the expectation. □

Orlicz spaces

In this small section, we show that Sub-gaussian and sub-exponential can be embedded in a more general notion that is Orlicz space. An Orlicz function is a map $\psi : [0,\infty) \mapsto [0,\infty)$ that is increasing and convex such that : $\psi(0) = 0$, $\lim_{x\to\infty} \psi(x) = \infty$ and In the same way as before , we define the Orlicz norm:

$$||X||_\psi = \inf\{t > 0, \mathbb{E}(\psi(|X|/t) \leq 2)\}$$

The set of all random variable with finite Orlicz norm will be denoted L_ψ. We will only verify the triangular inequality:

Exercise 4.4.2: $\|X\|_\psi$ norm

Show that $\|X\|_\psi$ is indeed a norm on the space L_ψ.

Proof.

$$\mathbb{E}\left(\psi(\frac{|X+Y|)}{a+b}\right) \leq \frac{a}{a+b}\mathbb{E}\left(\psi(|X|/a)\right) + \frac{b}{a+b}\mathbb{E}\left(\psi(|Y|/b)\right).$$

Now letting:

$$a = ||X||_\psi, b = ||Y||_\psi$$

, we get that $a+b$ is in the set of

$$\{t : \mathbb{E}(\psi(X+Y)/t)) \leq 1\}$$

so we get the result. □

What's interesting that Orlicz space is that it generalizes the definitions of L^p spaces, sub-gaussians and sub-exponential random. Any Orlicz space is a Banach space.

Frames

For a highly discrete distribution example, let us define a coordinate random vector X that is uniformly distributed over the set

$$\{\sqrt{n}e_i\}_{i=1}^n$$

where $\{e_i\}_{i=1}^n$ represents the canonical basis of $\mathbb{R}^n$.

This distribution can be expressed as

$$X \sim \text{Unif}\left(\{\sqrt{n}e_i : i = 1, \ldots, n\}\right).$$

It follows that X is an isotropic random vector in $\mathbb{R}^n$.

Definition 10. *A frame is a set of vectors* $\{u_i\}_{i=1}^N$ *in* $\mathbb{R}^n$ *which obeys an approximate Parseval's identity, i.e., there exist numbers* $A, B > 0$ *called frame bounds such that*

$$A\|x\|_2^2 \leq \sum_{i=1}^{N} \langle u_i, x\rangle^2 \leq B\|x\|_2^2 \quad \textit{for all } x \in \mathbb{R}^n.$$

If $A = B$*, the set* $\{u_i\}_{i=1}^N$ *is called a tight frame.*

Exercise 4.4.3: Tight frames condition

Show that $\{u_i\}_{i=1}^N$ is a tight frame in $\mathbb{R}^n$ with bound A if and only if

$$\sum_{i=1}^{N} u_i u_i^\top = AI_n.$$

Solution. $(u_i)_{i=1}^n$ is a tight frame with bound A if and only if :

$$\forall x \in \mathbb{R}^n \sum_{i=1}^{n} (\langle u_i, x\rangle^2 - Ax_i^2) = 0$$

Taking $x = \langle x, u_i\rangle e_i$ for all $i \in \mathbb{R}^n$ we get the first equality. Now if we assume that : $\sum_{i=1}^n u_i u_i^T = AI_n$ then by multiplying both sides by x^T and x, we get:

$$\sum_{i=1}^{n} \langle x, u_i\rangle\langle u_i, x\rangle = A\,\|x\|_2^2$$

which exactly means that $(u_i)_{i=1,..n}$ is a tight frame.

□

Bibliography

[1] R. Vershynin. *High-dimensional probability: An introduction with applications in data science.* Cambridge University Press, 2008.